

МЕЖГОСУДАРСТВЕННАЯ ОБРАЗОВАТЕЛЬНАЯ  
ОРГАНИЗАЦИЯ ВЫСШЕГО ОБРАЗОВАНИЯ  
КЫРГЫЗСКО-РОССИЙСКИЙ СЛАВЯНСКИЙ УНИВЕРСИТЕТ  
имени первого Президента Российской Федерации Б.Н. Ельцина  
МЕДИЦИНСКИЙ ФАКУЛЬТЕТ

Кафедра физики, медицинской информатики и биологии

# МЕДИЦИНСКАЯ ИНФОРМАТИКА И МЕДИЦИНСКАЯ СТАТИСТИКА

## Учебно-методическое пособие

*Посвящается 30-летию медицинского факультета  
Кыргызско-Российского Славянского  
университета им. Б.Н. Ельцина*



Бишкек 2025

УДК 61:004

ББК 51.1

М 42

**Рецензенты:**

*Т.Б. Атанаев*, канд биол. наук, профессор Кыргызского государственного университета им. И. Арабаева,  
*Н.Т. Карашева*, канд. пед. наук, доцент Кыргызской государственной медицинской академии им. И.К. Ахунбаева,  
*Б.П. Губанов*, канд. мед. наук, доцент Кыргызско-Российского Славянского университета им. Б.Н. Ельцина

**Авторы-составители:**

*Е.И. Кондратьева,*  
*Ю.М. Курманбакиев, А.А. Сорокин*

Рекомендовано к изданию Учёным советом МОО ВО  
Кыргызско-Российский Славянский университет им. Б.Н. Ельцина

М 42 МЕДИЦИНСКАЯ ИНФОРМАТИКА И МЕДИЦИНСКАЯ СТАТИСТИКА: учебно-методическое пособие / сост. Е.И. Кондратьева, Ю.М. Курманбакиев, А.А. Сорокин. – Бишкек: Издательский дом КРСУ, 2025. – 180 с.

ISBN 978-9967-36-062-4

Данное учебно-методическое пособие предназначено для освоения дисциплины «Медицинская информатика» студентов второго курса медицинских вузов различных специальностей.

Настоящее учебно-методическое пособие имеет четкую структуру изложения материала: приводятся примеры различных расчетов и соответствующие выводы; дается перечень контрольных вопросов, тестовых заданий и ситуационных задач. Это позволит студентам правильно интерпретировать полученные результаты.

УДК 61:004

ББК 51.1

ISBN 978-9967-36-062-4

© МОО ВО КРСУ, 2025

# СОДЕРЖАНИЕ

ВВЕДЕНИЕ .....	5
<b>Глава 1. ИНФОРМАЦИЯ И ИНФОРМАТИКА .....</b>	<b>6</b>
1.1. Что такое информация и ее свойства? .....	6
1.2. Особенности медицинской информации и ее виды .....	9
1.3. Что такое медицинская информатика? .....	15
1.4. Тестовые вопросы.....	17
<b>Глава 2. МЕДИЦИНСКАЯ СТАТИСТИКА.....</b>	<b>20</b>
2.1. Этапы развития медицинской статистики.....	20
2.2. Организация статистических исследований .....	24
2.3. Программы для статистической обработки медико-биологических данных .....	26
2.4. Программа SPSS. Описание полей. Типы переменных .....	30
2.5. Алгоритм для создания файла в программе SPSS .....	38
2.6. Задания для самостоятельной работы .....	46
2.7. Тестовые вопросы.....	47
<b>Глава 3. ОПИСАТЕЛЬНЫЕ СТАТИСТИКИ</b>	
<b>ЧАСТЬ 1 .....</b>	<b>52</b>
3.1. Описательные статистики.....	52
3.2. Частоты .....	55
3.3. Описательные таблицы .....	58
3.4. Частоты + Выборки .....	59
3.5. Задачи для закрепления материала .....	67
3.6. Задание для самостоятельной работы.....	67
3.7. Тестовые вопросы.....	68
<b>Глава 4. ОПИСАТЕЛЬНЫЕ СТАТИСТИКИ</b>	
<b>ЧАСТЬ 2.....</b>	<b>74</b>
4.1. Что такое описательные статистики?.....	74
4.2. Виды статистических величин .....	76

4.3. Нормальное распределение случайной величины.....	79
4.4. Описательные статистики, нормальное распределение .....	82
4.5. Описательные статистики, распределение, отличное от нормального .....	85
4.6. Задачи для закрепления материала .....	89
4.7. Задание для самостоятельной работы.....	90
4.8. Тестовые вопросы.....	91
<b>Глава 5. СРАВНИТЕЛЬНЫЕ СТАТИСТИКИ .....</b>	<b>101</b>
5.1. Что такое сравнительные статистики?.....	101
5.2. Сравнение средних с применением парного t-критерия Стьюдента .....	104
5.3. Сравнение средних с применением непарного t-критерия .....	109
5.4. Сравнение средних с применением парного непараметрического критерия Вилкоксона .....	112
5.5. Сравнение средних с применением непараметрического критерия Манна – Уитни .....	118
5.6. Задачи для закрепления материала .....	123
5.7. Задание для самостоятельной работы.....	125
5.9. Тестовые вопросы.....	127
<b>Глава 6. КОРРЕЛЯЦИОННЫЕ СТАТИСТИКИ.....</b>	<b>139</b>
6.1. Что такое корреляция?.....	139
6.2. Критерии корреляции .....	142
6.3. Виды корреляции .....	146
6.4. Задачи для закрепления материала .....	158
6.5. Задание для самостоятельной работы.....	160
6.6. Тестовые вопросы.....	162
<b>СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ .....</b>	<b>171</b>
<b>Приложение 1. РЕФЕРАТ СРС .....</b>	<b>173</b>
<b>Приложение 2. ФОРМА СРС .....</b>	<b>175</b>

## ВВЕДЕНИЕ

Предлагаемое учебно-методическое пособие разработано преподавателями кафедры физики, медицинской информатики и биологии Кыргызско-Российского Славянского университета им. Б.Н. Ельцина и предназначено для студентов второго курса медицинского факультета.

Данное учебно-методическое пособие посвящено медицинской информатике, медицинской статистике, одному из основных разделов профессиональной подготовки, и предназначено для освоения дисциплины «Медицинская информатика» студентами-медиками специальностей «Лечебное дело», «Педиатрия» и «Стоматология».

В учебно-методическом пособии в сжатой и очень доступной форме содержится базовая информация и определение основных принципов медицинской информатики, описание и сущность статистических методов обработки результатов медицинских исследований.

Основная цель учебно-методического пособия – научить студентов-медиков грамотно использовать специальное программное обеспечение для решения конкретных задач в своей будущей профессиональной деятельности.

Настоящее учебное пособие отвечает современным требованиям, каждая глава имеет четкую структуру изложения материала, примеры расчета и выводы, что позволяет студентам правильно интерпретировать полученные результаты. В учебном пособии есть контрольные вопросы, тестовые задания и ситуационные задачи для самостоятельного решения студентами.

Учебное пособие предназначено для практических занятий и самостоятельной внеаудиторной работы студентов, для облегчения процесса усвоения общепрофессиональных компетенций.

## **Глава 1. ИНФОРМАЦИЯ И ИНФОРМАТИКА**

### **1.1. Что такое информация и ее свойства?**

### **1.2. Особенности медицинской информации и ее виды.**

### **1.3. Что такое медицинская информатика?**

### **1.4. Тестовые вопросы.**

#### **1.1. Что такое информация и ее свойства?**

Несмотря на то, что с понятием информации мы сталкиваемся ежедневно, но общепризнанного ее определения до сих пор не существует. Поэтому вместо определения информации обычно используют понятие «информация». Понятия в отличие от определений не даются однозначно, а вводятся на примерах. При этом типична ситуация, когда понятие об информации, введенное в рамках одной научной дисциплины, может опровергаться конкретными примерами и фактами, полученными в рамках другой дисциплины [1]. Вот несколько примеров толкования понятия информации.

1. Информация – это сведения (сообщения) об объектах и явлениях окружающей среды, их параметрах, свойствах и состоянии, которые уменьшают имеющуюся о них степень неопределенности, неполноты знаний.

2. Информация – это количественная мера устранения неопределенности в отношении исхода того или другого события.

3. Информация – это продукт взаимодействия данных и адекватных им методов.

Научные исследования показывают, что свыше 80 % информации, получаемой человеком из внешнего мира, приходится на зрение, около 10 % – на тактильные ощущения, и лишь 7 % составляет информация, воспринимаемая в текстовой форме.

Информацию передают с помощью сообщений, сообщения передают с помощью последовательности сигналов, среда, в которой происходит распространение сигнала, называется каналом передачи [2].

### **Свойства информации**

Под качеством информации понимают степень её соответствия потребностям пользователей. Свойства информации являются относительными, так как зависят от потребностей пользователя информации. Выделяют следующие свойства, характеризующие качество информации [3]:

- объективность;
- полнота;
- достоверность;
- адекватность;
- доступность;
- актуальность;
- дуализм.

**Объективность и субъективность информации.** Понятие объективности информации является относительным. Это понятно, если учесть, что методы являются субъективными, поскольку являются свойствами субъекта – человека. Более объективной принято считать ту информацию, в которую методы вносят меньший субъективный элемент. Так, например, принято считать, что в результате наблюдения фотоснимка объекта или явления образуется более объективная информация, чем в результате наблюдения рисунка того же объекта [3].

**Полнота информации.** Полнота информации во многом характеризует качество информации и определяет достаточность данных для принятия решений. Чем полнее информация, тем шире диапазон методов, которые можно использовать, тем проще подобрать метод, вносящий минимум погрешностей в ход информационного процесса.

**Достоверность информации.** Данные возникают в момент регистрации сигналов, но не все сигналы являются «полезными» – всегда присутствует какой-то уровень посторонних

сигналов, в результате чего полезные данные сопровождаются определенным уровнем «информационного шума». Если полезный сигнал зарегистрирован более четко, чем посторонние сигналы, достоверность информации может быть более высокой. При увеличении уровня шумов достоверность информации снижается. В этом случае для передачи того же количества информации требуется использовать либо больше данных, либо более сложные методы [3].

**Адекватность информации** – это определенный уровень соответствия образа, создаваемого с помощью полученной информации реальному объекту, процессу, явлению и т. п. Неадекватная информация может образоваться при создании новой информации на основе неполных или недостоверных данных. Однако и полные, и достоверные данные могут приводить к созданию неадекватной информации в случае применения к ним неадекватных методов. В реальной жизни вряд ли возможна ситуация, когда можно рассчитывать на полную адекватность информации. Всегда присутствует некоторая степень неопределенности. От степени адекватности информации реальному состоянию объекта или процесса зависит правильность принятия решений человеком [3].

**Доступность информации** – мера возможности получить ту или иную информацию. На степень доступности информации влияют одновременно как доступность данных, так и доступность адекватных методов для их интерпретации.

**Актуальность информации** – это степень соответствия информации текущему моменту времени. Поскольку информационные процессы растянуты во времени, то достоверная и адекватная, но устаревшая информация может приводить к ошибочным решениям. Необходимость поиска или разработки адекватного метода для работы с данными может приводить к такой задержке в получении информации, что она становится неактуальной и ненужной [3].

Дуализм информации характеризует ее двойственность. С одной стороны, информация объективна в силу объективности

данных, с другой – информация субъективна в силу субъективности применяемых методов, то есть методы могут вносить в большей или меньшей степени субъективный фактор и таким образом влиять на информацию. Например, два человека читают одну и ту же книгу и получают подчас весьма разную информацию, хотя прочитанный текст, то есть данные, были одинаковы [3].

Информацию по одному и тому же вопросу можно излагать кратко (сжато, без излишних деталей) или пространно (подробно, многословно). Краткость информации наглядно демонстрируется в справочниках, энциклопедиях, инструкциях.

## **1.2. Особенности медицинской информации и ее виды**

**Медицинская информация**, в широком смысле, – это любая информация, относящаяся к медицине. В более узком – это информация, относящаяся непосредственно к человеку как пациенту, то есть информация о его здоровье, особенностях организма, перенесенных заболеваниях и др.

Медицинская информация отражает данные и результаты медицинских научных исследований и медицинской практики. С одной стороны, она отражает процессы и явления в системе здравоохранения (то есть является средством, которое используется врачами во время медицинской практики), с другой – она может быть результатом работы информационно-вычислительных центров, специалистов организационно-методического отдела и т. д. [4].

Основные требования, предъявляемые к медицинской информации, – постоянная ее обновляемость и конфиденциальность. Персонализированная информация (то есть информация о конкретном пациенте с указанием данных, по которым его можно идентифицировать) должна быть доступна только на уровне лечебно-профилактического учреждения, то есть лечащему врачу, заинтересованному среднему медицинскому персоналу, консультантам из других отделений и служб клиники, а также

руководителю учреждения. При передаче информации в выше-стоящие организации и слиянии информационных потоков из разных лечебных учреждений в один необходимо обеспечить деперсонализацию информации [4].

### **Виды медицинской информации**

1. Алфавитно-цифровая информация:

2. Визуальная информация:

а) статическая;

б) динамическая;

3. Звуковая информация;

4. Комбинированные виды информации.

**Алфавитно-цифровая информация** является основой почти всех форм печатных и рукописных документов (кроме случаев, когда документ представляет собой график или схему). Она составляет большую содержательную часть медицинской информации. К этому типу относится большая часть медицинской информации – фактически это содержимое истории болезни (как «бумажной», так и электронной).

Алфавитная, или символьная, информация может быть представлена либо в виде свободного текста (дневниковая запись лечащего врача о текущем состоянии больного), либо в более строгом и структурированном виде, подразумевающим конечное количество используемых терминов (например, лист первичного осмотра врача в приемном отделении) [4].

**Визуальная информация.** К этому виду информации относятся изображения отдельных частей, органов, тканей или даже клеток организма, получаемые с помощью диагностического медицинского оборудования.

**Статическая визуальная информация.** К этой категории медицинской информации относятся различные изображения (рентгенограммы, эхокардиограммы и т. д.). В зависимости от технических средств и других особенностей полученная информация может быть серошкальной (например, рентгеновское изображение, состоящее из различных оттенков серого) или цветной (например, эндоскопическое изображение) [4].

**Динамическая визуальная информация.** Примерами подобной информации являются походка пациента, мимика или судороги, сухожильные рефлексы, реакция зрачка на свет, генерируемое диагностическим оборудованием динамическое изображение [4].

**Звуковая информация** включает речь, усиленные техническим способом естественные звуки человеческого организма и звуковые сигналы, генерируемые медицинским оборудованием. Примерами речевой информации являются комментарий лечащего врача, речь пациента с неврологической или психической патологией, речь пациента с патологией гортани. Примерами усиленных техническим способом звуковых сигналов являются тоны, шумы, хрипы и другие элементы аускультации, слышимые с помощью фонендоскопа. Примерами звуковых сигналов, генерируемых медицинским оборудованием, являются доплеровские сигналы кровотока при эхокардиографии, флоуметрические сигналы, сигналы от фетальных мониторов и т. п. Некоторые виды или отдельные случаи звуковой информации могут входить в состав комбинированных видов медицинской информации (например, в сочетании с визуально-графической информацией) [4].

**Комбинированной** называется медицинская информация, представляющая собой любую комбинацию алфавитно-цифровой, визуально-графической и звуковой информации. Наиболее популярным комбинированным видом информации является сочетание динамической визуальной информации со звуковой. Однако на практике широко применяются и другие сочетания: например, статическая визуальная информация со звуковой, статическая визуальная информация с алфавитно-цифровой и прочие [4].

Следует отметить, что в большинстве случаев медикам приходится иметь дело с комбинированной информацией.

Как и в любой другой области знаний, информация в медицине имеет свои особенности. В медицинской практике часто используются выражения «сбор данных» или «получение информации». Эти выражения могут трактоваться неверно на основе предположения, что медицинская информация содержится

в реальном мире в состоянии доступности для использования её в диагностических или лечебных целях.

На самом деле некоторые объективные параметры, такие как биологическая дозировка, могут интерпретироваться или, другими словами, становиться информацией только в контексте, например, мотивации назначения, условий получения образца крови, используемого для измерений метода, и т. д. Симптом клинический или радиологический признак – это результат комплексного процесса принятия решений. Следовательно, медицинская информация как таковая существует только в интерпретируемой среде и должна постоянно обновляться, чтобы избежать диагностических и терапевтических ошибок. Возникающие у врачей гипотезы определяют направление сбора данных и критерии оценки «полезности» информации. Субъективность играет преобладающую роль в медицине. Эта ситуация частично объясняет неисчерпаемую природу медицинской информации. Информация может отсутствовать потому, что пациенту не был задан вопрос, или потому, что ответ пациента не был записан. Считается, что до 40 % проблем, идентифицированных в ходе исследований, были связаны с тем, что медицинская информация не была корректно сохранена. Следовательно, оценка качества медицинских данных очень важна и должна в первую очередь позволять оценить их информационное значение [5].

Характерной особенностью медицинской информации является её конфиденциальность. Права граждан на конфиденциальность информации о факте обращения за медицинской помощью и иных передаваемых ими при обращении за медицинской помощью сведений, на информированное добровольное согласие как предварительное условие для медицинского вмешательства и отказ от него установлены законом Кыргызской Республики «Об информации персонального характера» от 12.07.22 № 61 (Постановление № 5488-1). Этические нормы и правила обращения с медицинской информацией также регламентируются «Кодексом профессиональной этики медицинского работника Кыргызской Республики» от 16 января 2004 года № 2-4. Ст. 9 «Врачебная

тайна». Этой статьей установлено, что пациент вправе рассчитывать на то, что медицинский работник сохранит в тайне всю медицинскую и доверенную ему личную информацию. Медицинский работник не вправе разглашать без разрешения пациента или его законного представителя сведения, полученные в ходе обследования и лечения, включая и сам факт обращения за медицинской помощью. Медицинский работник должен принять меры, препятствующие разглашению медицинской тайны [6].

Другая особенность медицинской информации – неоднозначность. Она зависит от методов измерения характеристик пациента, человеческого фактора, особенностей интерпретации полученных данных, так как зависит от точки зрения врача на каждый конкретный клинический случай или схемы обследования и лечения.

Несомненно, в последнее время важным источником информации стал Интернет.

Описать ресурсы Интернета в какой-либо области – крайне сложная задача. Во-первых, они огромны и практически неисчерпаемы, а во-вторых, быстро обновляются и изменяются. Это базы данных, мультимедийные учебные серверы, виртуальные атласы и учебники, демонстрации клинических случаев, медицинские библиотеки, электронные версии журналов, описания научно-исследовательских проектов, программное обеспечение для обработки изображений и многое другое.

Необходимо отметить основные характеристики ресурсов Интернета, отличающие их от других источников информации. Любой пользователь Интернета может создать свою Web-страницу и наполнить ее любой информацией. При этом отсутствует какая-либо система контроля за содержанием (контентом) сайтов. Врач, пользующийся Интернетом, должен сам выбрать, каким Web-страницам он доверяет, какую информацию он может использовать для принятия правильных клинических решений.

Существующие в Интернете информационные ресурсы можно разделить на следующие виды: издательства и журналы, базы данных аннотаций и сервисные службы, обеспечивающие одновременный поиск статей на нескольких сайтах [7].

Среди сайтов, предоставляющих доступ к базам данных, безусловным лидером является всем известный PubMed MEDLINE (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed>). Поиск статей с помощью MEDLINE может осуществляться: по словам, встречающимся в названии статьи, в рефератах, по фамилиям авторов, названиям журналов. Данный вариант поиска является общим для большинства систем поиска информации в Интернете. Кроме того, может использоваться тезаурус, содержащий более 15 тыс. медико-биологических рубрик, включающих анатомическую номенклатуру, классификатор болезней, методы диагностики и лечения и многое другое [7].

В состав Научной электронной библиотеки (<http://www.elibrary.ru>) входят более миллиона полнотекстовых электронных версий статей по всем областям науки почти из 2000 журналов (около 500 Гб информации) с 1991 г. (издательства «Elsevier», «Springer»). Кроме того, в библиотеке представлены электронные полнотекстовые версии книг и медицинские базы данных (MEDLINE, Cancerlit, AIDSLINE). Для работы в библиотеке необходимо зарегистрироваться. Права бесплатного доступа к полным текстам статей предоставляются зарегистрированным организациям и ограничиваются определенными диапазонами IP-адресов, приписанных к каждой из зарегистрированных организаций [7].

Особого внимания заслуживают бесплатные онлайн-версии рецензируемых журналов, таких как British Medical Journal (<http://www.bmj.com>) и New England Journal of Medicine (<http://www.nejm.org>). Помимо данных журналов одними из наиболее достоверных источников информации считаются метаанализы, представленные на сайте Кокрановского сообщества (<http://www.cochrane.org>) и ACP Journal Club (<http://www.acpjournals.org>), а также клинические рекомендации The National Guideline Clearinghouse (<http://www.guidelines.gov>) [7].

Из российских сайтов в данном аспекте представляют интерес сайты некоторых журналов, таких как Русский медицинский журнал (<http://www.rmj.ru>) и Consilium Medicum (<http://www.consilium-medicum.com>).

Стоит отметить, что, используя достоверные источники информации при различных расчетах, можно получить достоверные результаты, при применении адекватных методов для их обработки [7].

### **1.3. Что такое медицинская информатика?**

Перед тем, как погрузиться в мир медицинской статистики, давайте разберемся, что же такое на самом деле медицинская информатика?

При определении медицинской информатики как самостоятельной области знаний можно выделить несколько аспектов. В одном случае делается акцент на общие позиции информатики, в другом – на специфику информационных процессов в медицине и в ее организационной части – здравоохранении, либо на роль информации и информатизации в решении проблем отрасли.

Один из ведущих разработчиков задач медицинской информатики Е. Shortliffe (1995) предложил ее расширенную трактовку «Медицинская информатика – это быстро развивающаяся область науки, которая ориентирована на биомедицинскую информацию, данные и знания, их хранение, передачу и оптимальное использование для решения проблем или принятия решений». Далее автор отметил, что появление медицинской информатики как новой дисциплины и причины роста ее популярности стало возможным, главным образом, благодаря прогрессу в компьютерных и коммуникационных технологиях [8].

По определению В.Г. Кудриной (1999) «Медицинская информатика – это научная дисциплина, представляющая собой систему знаний об информационных процессах в медицине, здравоохранении и смежных дисциплинах, обосновывающая и определяющая способы и средства рациональной организации и использования информационных ресурсов в целях охраны здоровья населения» [8].

В.К. Гасников (1997) определил «информатику как область научно-технической деятельности, занимающуюся исследованием

процессов получения, передачи, обработки, хранения, распространения, представления и использования информационной техники и технологии во всех сферах общественной жизни». Соответствующим образом, подчеркивает автор, трактуется и медицинская информатика [8].

Аналогичным образом определяет предмет и В.Я. Гельман (2001), подчеркивая, что медицинская информатика является результатом перекрестного взаимодействия медицины и информатики. Данная трактовка с точки зрения конкретности содержания представляется наиболее приемлемой [8].

Поэтому **медицинская информатика** – это научная дисциплина, занимающаяся исследованием процессов получения, передачи, обработки, хранения, распространения, представления информации с использованием информационной техники и технологии в медицине и здравоохранении.

Предметом изучения медицинской информатики при этом являются информационные процессы, сопряженные с медико-биологическими, клиническими и профилактическими проблемами [8].

Объектом изучения медицинской информатики являются информационные технологии, реализуемые в медицине и здравоохранении.

Информационные технологии – это преимущественно компьютеризированные способы выработки, хранения, передачи и использования информации.

Хотя, строго говоря, понятие информационные технологии шире, чем компьютерные технологии, но применительно к современной медицинской информатике они практически совпадают.

Основной целью медицинской информатики является оптимизация информационных процессов в медицине за счет использования компьютерных технологий, обеспечивающая повышение качества охраны здоровья населения.

Учитывая, что медицинская информатика является одним из прикладных видов информатики, ее можно представить состоящей из двух разделов: общей, базовой информатики и собственно медицинской информатики [8].

Общая информатика, как фундаментальная наука, занимается разработкой методологии создания информационного обеспечения процессов управления любыми объектами на базе компьютерных информационных систем. Как прикладная дисциплина, информатика изучает закономерности информационных процессов, занимается созданием информационных моделей коммуникаций, разработкой информационных систем и технологий в конкретных областях.

Собственно **медицинская информатика** рассматривает медицинские приложения информационных технологий. При этом изучаются как использование стандартных, универсальных средств информатики для решения медицинских задач, так и специальные медицинские информационные технологии и системы [8].

## **1.4. Тестовые вопросы**

### **1. С помощью чего передают информацию?**

- a) звуков;
- b) сигналов;
- c) картинки;
- d) каналов;
- e) сообщений.

### **2. Перечислите основные свойства информации:**

- a) объективность, субъективность, достоверность, полнота, актуальность, адекватность, доступность, полезность;
- b) полезность, ценность, понятность, защищенность, однозначность, правдивость, полнота, своевременность, доступность;
- c) ценность, краткость, эргономичность, своевременность, объективность, точность, актуальность, правдивость;
- d) доступность, понятность, актуальность, ценность, краткость, точность, своевременность, нужность, полнота;
- e) точность, понятность, ценность, актуальность, доступность, краткость, своевременность, однозначность.

**3. Определите, что называется наукой о получении, обработке, распространении, хранении, передаче и представлении медицинской информации на основе компьютерных технологий?**

- a) медицинская информатика;
- b) экономическая информатика;
- c) социальная информатика;
- d) политологическая информатика;
- e) математическая информатика.

**5. Информация о здоровье пациента, особенностях организма, перенесенных заболеваниях и др. – это:**

- a) медицинская информатика;
- b) медицинский анамнез;
- c) медицинская информация;
- d) медицинская статистика;
- e) медико-биологическая информатика.

**6. Перечислите виды медицинской информации:**

- a) алфавитно-цифровая, визуальная, звуковая, комбинированная;
- b) амбулаторная карта, рентген-снимки, ЭКГ, выписки, справки;
- c) зрительная, обонятельная, вкусовая, тактильная, звуковая;
- d) электронная карта больного, результаты анализов;
- e) медицинские журналы, медицинские учебники, монографии.

**7. Перечислите, что относится к алфавитно-цифровой информации:**

- a) все печатные и рукописные медицинские документы;
- b) различные статистические и динамические изображения;
- c) речь пациента, перкутанные звуки, звуки при аускультации;
- d) полный анамнез жизни пациента, результаты ЭКГ, снимки;
- e) амбулаторные карты, справки, выписки, анализы, снимки.

**8. Перечислите, что относится к визуальной информации:**

- a) различные статистические и динамические изображения;
- b) все печатные и рукописные медицинские документы;
- c) речь пациента, перкутанные звуки, звуки при аускультации;
- d) полный анамнез жизни пациента, результаты ЭКГ, снимки;
- e) амбулаторные карты, справки, выписки, анализы, снимки.

**9. Перечислите, что относится к звуковой информации:**

a) речь пациента, перкутанные звуки, звуки при аускультации;

- b) все печатные и рукописные медицинские документы;
- c) различные статистические и динамические изображения;
- d) полный анамнез жизни пациента, результаты ЭКГ, снимки;
- e) амбулаторные карты, справки, выписки, анализы, снимки.

**10. Перечислите, что относится к комбинированной информации:**

a) амбулаторные карты, выписки, анализы, снимки, речь пациента;

- b) все печатные и рукописные медицинские документы;
- c) различные статистические и динамические изображения;
- d) речь пациента, перкутанные звуки, звуки при аускультации;
- e) полный анамнез жизни пациента, результаты ЭКГ, снимки.

**11. Как называется сообщение, не несущее полезной информации?**

- a) шум;
- b) реклама;
- c) мусор;
- d) спам;
- e) шлак.

## **Глава 2. МЕДИЦИНСКАЯ СТАТИСТИКА**

**2.1. Этапы развития медицинской статистики.**

**2.2. Организация статистических исследований.**

**2.3. Программы для статистической обработки медико-биологических данных.**

**2.4. Программа SPSS. Описание полей. Типы переменных.**

**2.5. Алгоритм для создания файла в программе SPSS**

**2.6. Задания для самостоятельной работы**

**2.7. Тестовые вопросы**

### **2.1. Этапы развития медицинской статистики**

Начало статистических наблюдений можно отнести к глубокой древности, когда государственные органы собирали различные сведения о состоянии государства. В Египте, Персии, Греции и Риме велись переписи населения, записи рождений и смертей, списки совершеннолетних.

В средние века, особенно в первую половину, статистический интерес могли представлять в основном записи церковных треб, которые и послужили материалом для проведения исследований о движении населения [9].

В конце средних веков, в связи с развитием торговли, в Италии, Бельгии и Голландии начинают появляться сведения о государствах и странах, с которыми велись торговые отношения. В сборниках, относящихся к средним векам, включая первую половину XVII столетия можно отметить различные заметки исторического, государственного, а также этнографического характера, которые объединены без какой-либо взаимной связи. Однако понимая под статистикой научное объединение собранных

наблюдений, следует отметить, что в этот период времени статистики ещё не было [9].

Как наука статистика возникла во второй половине XVII века в Англии и Германии почти в одно и то же время. Её возникновение связано с развитием государственного управления и практики учётно-статистических работ.

Первое направление, которое известно, как описательная школа, возникло в Германии. Основной задачей этого направления было описание территории, населения, политического устройства государства, ведения его хозяйства, торговли. Основателем описательной школы был немецкий учёный Герман Конринг, который в 1666 году ввёл преподавание государственного управления в университетах, в то время как в Италии и Голландии ограничилось составлением сборников описаний государств [9].

Само понятие «Statistica» произошло от двух латинских слов «stato» – государство и «status» – положение вещей, политическое состояние. Причем в большинстве своем, изначально, как уже упоминалось, под понятием подразумевалось совокупность общих сведений о государстве и его достопримечательностях. Немецкий учёный Готфрид Ахенваль в 1746 году предложил преподавать статистику как курс в университетах Германии, заменив преподававшееся в университетах Германии с 1666 года государственное управление, что в свою очередь способствовало её развитию и формированию как научной и учебной дисциплины [9].

Второе направление развития статистики, известное как «политическая арифметика» возникло в Англии. Представителями этого направления были Уильям Петти и Джон Граунт. Особенность данного направления заключается в том, что его представители считали основной задачей выявление взаимосвязей и закономерностей на основе большого числа наблюдений [9].

Д. Граунт исследовал в основном закономерности воспроизводства населения. Он в течение многих лет изучал данные бюллетеней смертности, издававшиеся с 1603 года, в которых каждую неделю публиковались данные о числе родившихся и умерших в Лондоне. Изучая эти данные, он установил следующие

закономерности: соотношение численности родившихся мальчиков и девочек, которое составляло 14:13, из числа родившихся до 6 лет доживало только 64 %, а до 16 лет доживало 40 %. Кроме того, он доказал, что население Лондона пополняется за счет пришлого населения, поскольку смертность в нем превышает рождаемость. Д. Граунт впервые рассчитал кривую дожития, а также составил таблицу смертности [9].

В России статистика начинает формироваться к середине XVIII века. Развитие санитарной статистики было тесно связано с изучением этнографии, топографии, географии, экономики. Первые попытки собрать централизованно сведения о заболеваемости относятся к середине XVIII века, а сбор и централизация материалов о рождаемости и смертности – к концу XVIII столетия [9].

Большое внимание вопросам движения населения уделял М.В. Ломоносов, которого глубоко волновали высокая смертность и особенно высокая детская смертность.

С середины XVIII века по инициативе П.З. Кондоиди, который в то время руководил «Медицинской канцелярией», при проведении медико-топографических описаний отдельных местностей включались и статистические данные о заболеваемости населения, рождаемости и смертности [9].

Передовым деятелем второй половины XVIII века был С.Г. Зыбелин. В своих работах он уделял большое внимание высокой детской смертности, где среди многих причин особенно выделял неправильное кормление детей. Основной причиной высокой смертности населения С.Г. Зыбелин считал заболеваемость, которая, несмотря на высокую рождаемость, приводила к медленному росту населения [9].

С начала XIX века постепенно расширяется программа санитарно-статистических исследований, в ней отводится значительное место санитарно-статистической характеристике здоровья населения.

Однако значительный подъем санитарной статистики был связан с развитием земской медицины – это произошло

в последней четверти XIX века. Наличие на селе земских лечебных учреждений способствовало развитию статистики общей заболеваемости, которая стала основным направлением исследований земских врачей [9].

Основоположником земской санитарной статистики был Е.А. Осипов, руководивший работой Московского губернского санитарного бюро (1875–1895 гг.). Большую роль в земской санитарной статистике сыграли Ф.Ф. Эрисман – профессор гигиены Московского университета, работавший с Е.А. Осиповым в земском санитарном бюро, а также один из лучших организаторов и руководителей земской санитарной статистики П.И. Куркин, преемник Е.А. Осипова [9].

В начале XX века С.А. Новосельский вместе с учениками опубликовал ряд работ, посвященных анализу санитарно-демографических материалов, среди которых особого внимания заслуживает книга «Смертность и продолжительность жизни в России» (Петроград, 1916 г.), которая была удостоена премии Российской академии наук. Она содержала первые полные таблицы смертности населения Европейской России, которые были построены на основании единственной переписи населения 1897 г. и приуроченным к ним данным о смертности [9].

После 1917 года статистика в медицине стала развиваться по двум направлениям – статистика народного здоровья и статистика здравоохранения. Эти два направления были тесно связаны по содержанию, организации, методике [9].

В.Ю. Урбах отмечал, что при изучении методов математической статистики необходимо уделять особое внимание границам применимости каждого метода, поскольку неправильное применение статистических методов приводит к ошибочным выводам [9].

В развитие отечественной медицинской статистики внесли свой вклад выдающиеся ученые – А.М. Мерков, Е.А. Садвокасова, Л.Е. Поляков, И.С. Случанко, Г.Ф. Церковный, Е.Л. Ноткин, Н.А. Плохинский, Е.Н. Шиган [9].

В настоящее время большое значение в современном медицинском образовании имеет обучение методам статистического анализа показателей.

Знание медицинской статистики позволяет объективно оценить проблемы заболеваемости населения, деятельность медицинских организаций и учреждений здравоохранения, правильно понимать и рационально использовать специальную и научную медицинскую литературу, что необходимо для повышения профессиональной квалификации, а также для оптимального решения вопросов планирования, экономики, маркетинга и менеджмента в здравоохранении [9].

## **2.2. Организация статистических исследований**

Медицинская статистика необходима для оценки, планирования, прогноза результатов работы и её качества. В этих целях вычисляются и анализируются показатели объема и качества деятельности врача и среднего медицинского работника [9].

Важно, что, зная основные понятия и статистические термины можно грамотно проанализировать собственную работу любого медицинского сотрудника и наметить мероприятия по её совершенствованию.

Неоценима роль понимания и практического использования профессиональных публикаций в сфере своей специальности и методик применения инновационных разработок и/или современных новейших технологий, материалов в процессе обслуживания пациентов.

Медицинская статистика при её систематическом использовании, несомненно, развивает профессиональное мышление, на что указывают Всемирная организация здравоохранения и ведущие специалисты-медики. [9].

### **Этапы развития медицинской статистики**

**I этап.** Определяется цель (разработать мероприятия, например, по улучшению качества диагностических услуг); задачи по достижению цели (проработка приказов, инструкций,

методических материалов, профессиональных публикаций); оценка реальной работы, ее положительных и отрицательных характеристик; изучение отечественного и зарубежного опыта с оптимальными результатами. Составляется план работы (сроки, объем, кадры, материальное и финансовое обеспечение, методика сбора данных).

В этот же первый этап входит разработка перечня собираемых сведений и макетов («скелетов») статистических таблиц, программного обеспечения [9].

**II этап.** «Сбор статистических материалов» включает определение единицы наблюдения (пациент, случай заболевания или др.). Определяется объект изучения (лечебно-диагностическая клиника).

Используются содержащие информацию медицинские учетные и отчетные документы или разрабатываются специальные анкеты (например, для опроса пациентов об удовлетворенности обслуживанием).

На специальные карты собирается информация о пациенте (пол, возраст, профессия и др.), заболевании (диагноз, продолжительность заболевания), методах лечения и др. [9].

**III этап.** Включает следующие элементы сводки и обработки статистических данных, в том числе:

➤ Контроль собранного материала (проверка правильности и точности собранных статистических данных).

➤ Шифровка (для удобства компьютерной и ручной обработки).

➤ Группировка количественная (по возрасту, стажу работы, продолжительности заболеваний и др.) и качественная (социальная группа, профессия, методы лечения и др.) [9].

- подсчет данных по выделенным группам;
- заполнение статистических таблиц;
- вычисление показателей и индексов (например, частота распространенности кариеса зубов, патологии пародонта, показателей качества деятельности и др.);

- использование в расчетах метода оценки достоверности разности средних и относительных величин, вычисление коэффициента корреляции, показателей динамического ряда и др.;
- построение графических изображений (диаграмм) [9].

**IV этап.** Включает анализ и оценку разработанных статистических показателей (показателей, индексов и др.). Определяется уровень достижения цели, выполнения намеченных на первом этапе задач, плана и программы.

В итоге формулируются выводы и предложения, приемлемые в практических и научных целях [9].

### **2.3. Программы для статистической обработки медико-биологических данных**

Одним из обязательных этапов любого научного исследования является статистический анализ данных. Продолжительное время анализ медицинских данных был уделом специалистов, так как это требовало серьезной предварительной подготовки. С появлением и совершенствованием современных программ обработки данных статистическая обработка поднялась на новый уровень. Теперь исследователь-медик может и не иметь математической подготовки. Достаточно оперировать статистическими понятиями и, самое главное, правильно выбрать метод анализа. Все осуществимо благодаря компьютеру и новейшим программам [10].

Все программы статистической обработки данных можно разделить на профессиональные, полупрофессиональные (популярные) и специализированные. Статистические программы относятся к наукоемкому программному обеспечению, цена их часто недоступна индивидуальному пользователю. Профессиональные пакеты имеют большое количество методов анализа, популярные пакеты – количество функций, достаточное для универсального применения. Специализированные же пакеты ориентированы на какую-либо узкую область анализа данных. Отсутствие у большинства исследователей времени для освоения

нескольких программ, делает непростым ее выбор. В данном разделе приведена базовая информация о присутствующих основных полупрофессиональных программных пакетах пригодных для статистической обработки биомедицинских данных [10].

**MS Excel.** Самой часто упоминаемым (и используемым) в отечественных статьях является приложение MS Excel из пакета офисных программ компании Microsoft – MS Office. Причины этого кроются в широком распространении этого программного обеспечения, наличии русскоязычной версии, тесной интеграцией с MS Word и PowerPoint. Однако MS Excel – это электронная таблица с достаточно мощными математическими возможностями, где некоторые статистические функции являются просто дополнительными встроенными формулами. Расчеты, сделанные при ее помощи, не признаются авторитетными биомедицинскими журналами. Также в MS Excel невозможно построить качественные научные графики. Безусловно, MS Excel хорошо подходит для накопления данных, промежуточного преобразования, предварительных статистических прикидок, для построения некоторых видов диаграмм. Однако окончательный статистический анализ необходимо делать в программах, которые специально созданы для этих целей. Существует макрос-дополнение XLSTAT-Pro для MS Excel, который включает в себя более 50 статистических функций, включая анализ выживаемости, которых в основных случаях достаточно для обычного применения. Пробную версию макроса можно взять на сайте производителя [10].

**STADIA.** Программа отечественной разработки с 16-летней историей. Включает в себя все необходимые статистические функции. Она прекрасно справляется со своей задачей – статистическим анализом. Но программа внешне фактически не изменяется с 1996 года. Графики и диаграммы, построенные при помощи STADIA, выглядят в современных презентациях архаично. Цветовая гамма программы (красный шрифт на зеленом) очень утомляет в работе. К положительным качествам программы можно отнести русскоязычный интерфейс и наличие книг, описывающих работу [10].

**STATA.** Профессиональный статистический программный пакет с datamanagement system, который может применяться для биомедицинских целей. Один из самых популярных в образовательных и научных учреждениях США наряду с SPSS. Программа хорошо документирована, издается специальный журнал для пользователей системы. Однако возможности предварительного ознакомления с демоверсией нет [10].

**STATISTICA.** Производителем программы является фирма StatSoft Inc. (США), которая выпускает статистические приложения, начиная с 1985 года. STATISTICA включает большое количество методов статистического анализа (более 250 встроенных функций) объединенных следующими специализированными статистическими модулями: Основные статистики и таблицы, Непараметрическая статистика, Дисперсионный анализ, Множественная регрессия, Нелинейное оценивание, Анализ временных рядов и прогнозирование, Кластерный анализ, Факторный анализ, Дискриминантный функциональный анализ, Анализ длительностей жизни, Каноническая корреляция, Многомерное шкалирование, Моделирование структурными уравнениями и др. Несложный в освоении этот статистический пакет может быть рекомендован для биомедицинских исследований любой сложности. Сайт компании содержит много информации по статистической обработке медицинских данных, учебник по статистике на русском языке. Сам пакет STATISTICA описан в нескольких книгах, одна из которых, для медицинских работников: О.Ю. Реброва «Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA» [10].

**JMR.** Один из мировых лидеров в анализе данных. Развивает этот статистический пакет SAS Institute, который выкупил в конце 2002 года известную статистическую программу StatView. Однако особых преимуществ для медикобиологической статистики этот программный продукт не имеет [10].

**SYSTAT.** Статистическая система для персональных компьютеров. Последняя версия обладает неплохим интуитивно понятным интерфейсом. Компания Systat Software

также разрабатывает популярные у отечественных исследователей SigmaStat и SigmaPlot, которые являются соответственно, программой статистической обработки и программой построения диаграмм. При совместной работе становятся единым пакетом для статистической обработки и визуализации данных [10].

**NCSS.** Программа развивается с 1981 года и рассчитана на непрофессионалов в области статистической обработки. Интерфейс системы многооконный и, как следствие этого явления, – немного непривычный в использовании. Все действия пользователя сопровождаются подсказками. Сейчас доступна версия 2004 года [10].

**MINITAB 14.** Статистический пакет MINITAB в настоящее время выпускается в версии 14. Это достаточно удобный в работе программный пакет, имеющий хороший интерфейс пользователя, хорошие возможности по визуализации результатов работы. Имеет подробную справку.

**STATGRAPHICS PLUS.** Довольно мощная статистическая программа. Содержит более 250 статистических функций, генерирует понятные, настраиваемые отчеты. Следует отметить, что ранние версии этой программы были весьма популярны у отечественных исследователей [10].

**PRISM.** Эта программа создавалась специально для биомедицинских целей. Интуитивно понятный интерфейс позволяет в считанные минуты проанализировать данные и построить качественные графики. Программа содержит основные часто применяемые статистические функции, которых в большинстве исследований будет достаточно. Однако, как отмечают сами разработчики, программа не может полностью заменить серьезных статистических пакетов [10].

**SPSS (Statistical Package for Social Science).** Самый часто используемый пакет статистической обработки данных с более чем 30-летней историей. Отличается гибкостью, мощностью, применим для всех видов статистических расчетов, применяемых в биомедицине. Существует русскоязычное представительство компании, которое предлагает полностью русифицированную

версию SPSS для Windows. Появился учебник на русском языке, позволяющий шаг за шагом освоить возможности SPSS, репетитор по статистике на русском языке, помогающий в выборе нужной статистической или графической процедуры для конкретных данных и задач, а также справка по SPSS Base и SPSS Tables. Российский офис SPSS регулярно проводит учебные курсы по анализу данных при помощи программного обеспечения SPSS. На русский язык переведена книга по SPSS, которая вышла в свет в 2002 году в Киевском издательстве «Диасофт» под названием «SPSS 10: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей» [10].

На какой программе остановить свой выбор? Безусловно, дороговизна программ не позволяет их менять. Поэтому имеет смысл посмотреть демо-версии, разобраться с работой и потом делать окончательный вывод. Русскоязычные версии (с документацией) имеют только SPSS и STATISTICA. Что касается возможных рекомендаций, то они следующие:

- Если нужен мощный, общепризнанный пакет с простым и понятным даже начинающим пользователям интерфейсом, то лучше воспользоваться SPSS.

- Для начинающих и профессионалов, которым нужна подсказка и развитая документация на русском языке, можно рекомендовать STATISTICA. Это мощное приложение с профессиональными возможностями.

- Для непритязательных пользователей, которые ограничиваются в своих исследованиях стандартными статистическими методами можно рекомендовать англоязычную программу Prism [10].

## **2.4. Программа SPSS. Описание полей.**

### **Типы переменных**

Программа SPSS является самой распространённой программой для статистической обработки медико-биологической информации.

Два студента Норман Най (Norman Nie) и Дейл Вент (Dale Bent), специализировавшиеся в области политологии, в 1965 году пытались отыскать в Стенфордском университете Сан-Франциско компьютерную программу, подходящую для анализа статистической информации. Вскоре они разочаровались в своих попытках, так как имеющиеся программы оказывались более или менее непригодными, неудачно построенными или не обеспечивали наглядность представления обработанной информации. К тому же принципы пользования менялись от программы к программе [11].

Так, недолго думая, они решили разработать собственную программу, со своей концепцией и единым синтаксисом. В их распоряжении тогда был язык программирования FORTRAN и вычислительная машина типа IBM 7090. Уже через год была разработана первая версия программы, которая еще через год, в 1967, могла работать на IBM 360. К этому времени к группе разработчиков присоединился Хэдлай Халл (Hadlai Hull) [11].

Как известно из истории развития информатики, программы тогда представляли собой пакеты перфокарт. Как раз на это указывает и исходное название программы, которое авторы дали своему продукту: SPSS – это аббревиатура от Statistical Package for the Social Science.

В 1970 году работа над программой была продолжена в Чикагском университете, а Норман Най основал соответствующую фирму.

С появлением персональных компьютеров была разработана также и PC-версия SPSS, с 1983 года появилась PC-версия SPSS/PC+, рассчитанная на MS-DOS. Позже, с момента основания в 1984 году европейского торгового представительства в Горингеме в Нидерландах, SPSS стал широко применяться и в Европе. В настоящее время это самое распространённое программное обеспечение для статистического анализа во всём мире [11].

Основу программы SPSS составляет SPSS Base (базовый модуль), предоставляющий разнообразные возможности доступа к данным и управления данными. Он содержит методы анализа, которые применяются чаще всего. Он включает все процедуры

ввода, отбора и корректировки данных, а также большинство предлагаемых в SPSS статистических методов.

Наряду с простыми методиками статистического анализа, такими как частотный анализ, расчет статистических характеристик, таблиц сопряженности, корреляций, построения графиков, этот модуль включает t-тесты и большое количество других непараметрических тестов, а также усложненные методы, такие как многомерный линейный регрессионный анализ, дискриминантный анализ, факторный анализ, кластерный анализ, дисперсионный анализ, анализ пригодности (анализ надежности) [11].

### Описание полей

Для того чтобы начать создавать файл необходимо установить программу SPSS на Вашем компьютере. Только после полного завершения установки можно открыть программу SPSS: ПУСК→ВСЕ ПРОГРАММЫ→SPSS 16.0. Внешний вид программы показан на рисунке 1.

Как видно из рисунка 1, столбцы в этой таблице – это измеряемые параметры (переменные), а строки – это наблюдения. Если рассматривать результаты различных исследований, каждая строка, как правило, включает набор данных об одном участнике исследования.

Если посмотреть в левый нижний угол рисунка 1, то можно увидеть, что программа SPSS состоит из двух вкладок «*DATA*

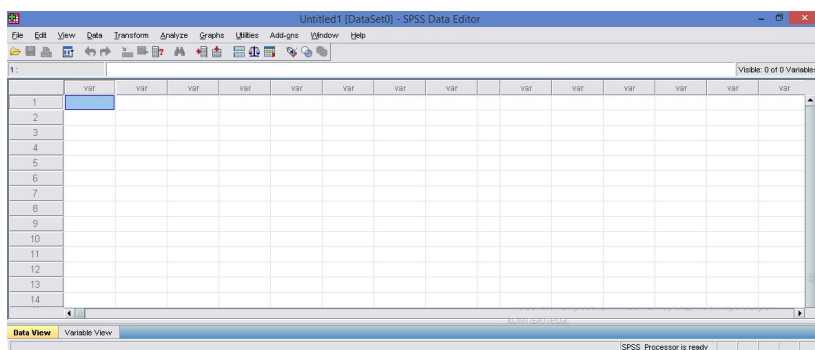


Рисунок 1 – Общий вид окна данных (Data view) программы SPSS 16.0

*VIEW*» – редактирование и просмотр данных, и «*VARIABLE VIEW*» – редактирование и просмотр переменных.

Для того чтобы начать создавать переменные, необходимо использовать заранее заполненную анкету, затем перейти во вкладку «*VARIABLE VIEW*», где можно увидеть следующие столбцы (рисунок 2):

Давайте разберем, что означает каждый из столбцов:

**NAME** – включает в себя сокращенное имя новой переменной. **ВАЖНО ОТМЕТИТЬ** – при задании имени переменной нельзя использовать: «пробел», «точку» и «запятую», а само имя переменной обязательно должно начинаться с буквы.

**TYPE** – тип данных для создаваемой переменной. Если требуется изменить тип переменной, щелкните в ячейке на кнопке с тремя точками (рисунок 3).

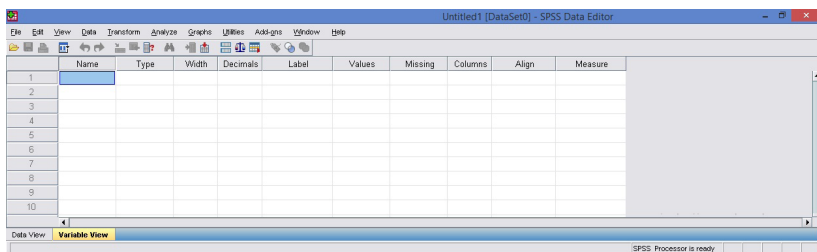


Рисунок 2. Общий вид вкладки Variable View

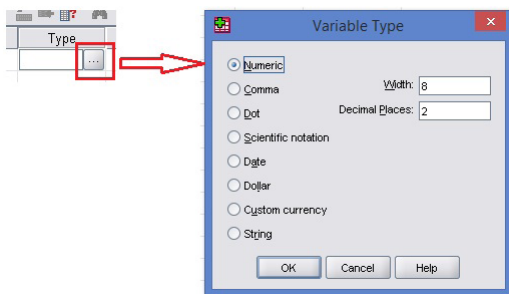


Рисунок 3 – Определение типа переменной

Из всех существующих типов данных (см. рисунок 3), чаще всего используются следующие:

- *Numeric* – числовые данные (возраст, рост, вес, давление и т. д.);
- *Data* – дата и/или время (дата и/или время рождения, дата поступления/выписки и т. д.);
- *String* – строчные или буквенные переменные (Ф.И.О., адрес, диагноз, тяжесть заболевания и т. д.), их достаточно сложно использовать для анализа, поэтому их желательно перекодировать в числовые.

**DECIMALS** – указывается количество знаков после десятичной запятой, то есть нужно знать, имеют ли данные создаваемой переменной десятые или сотые доли (рисунок 4). При изменении этого свойства меняется именно отображение самого числа, округления в расчетах не происходит.

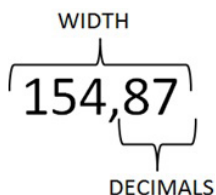


Рисунок 4 – Определение количества знаков

**WIDTH** – указывается количество знаков до и после десятичной запятой, то есть указывается общее количество знаков (см. рисунок 4).

**LABEL** – здесь указывается любое название или пояснение к переменной без ограничений в символах – в том виде, в каком оно будет представлено в таблицах после обработки данных. В данном случае это название, которое позволяет описать переменную более подробно.

**VALUES** – указываются варианты кодирования буквенных данных в числовые. Если требуется внести кодировку для создаваемой переменной, необходимо щелкнуть в ячейке на кнопке с тремя точками (рисунок 5).

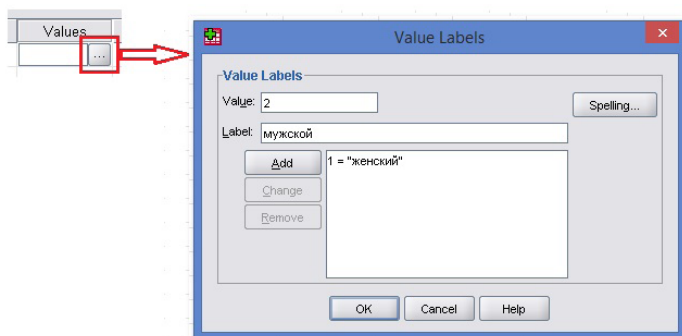


Рисунок 5 – Вспомогательное окно для кодирования буквенных данных в числовые

**Пояснение.** Например, в учебном файле имеется следующая переменная – «пол пациента», которая имеет два значения – женский и мужской; необходимо буквы перевести в цифры следующим образом – принимаем женский пол за 1, а мужской за 2 (см. рисунок 5). Теперь, когда мы будем заполнять базу, мы будем использовать цифры 1 или 2.

**MISSING** – задаются параметры пропущенных значений.

В SPSS допускаются два вида пропущенных значений:

1 – *пропущенные значения, определяемые системой* (System-defined missing values). Если в матрице данных есть незаполненные численные ячейки, программа SPSS самостоятельно идентифицирует их как пропущенные значения;

2 – *пропущенные значения, задаваемые пользователем* (User-defined missing values). Если в определенных случаях у переменных отсутствуют значения, например, если на вопрос не был дан ответ, ответ неизвестен или существуют другие причины, пользователь может с помощью кнопки Missing объявить эти значения как пропущенные. Пропущенные значения можно исключить из последующих вычислений. Обычно, пропущенным значением, определяемым пользователем объявляется вариант ответа «0» (нет данных). Наличие двух возможностей кодирования пропусков связано с отношением к ним исследователя.

Если исследователю безразлично почему произошел пропуск, то используется системный пропуск. Если же существует необходимость фиксировать, почему произошел пропуск (например, пропуск потому, что человек умер или потому что человек уехал – это разные по значимости пропуски), лучше использовать пропуски, задаваемые пользователем.

Обычно для учебных целей используется первый вариант, то есть программа SPSS сама определяет незаполненные данные.

**COLUMNS** – определяется ширина столбца для значений данной переменной.

**ALIGN** – задается выравнивание данных при их вводе (по левому краю, по центру, по правому краю).

**MEASURE** – шкала измерения для переменной, она может быть метрической, порядковой или номинальной (рисунок 6).

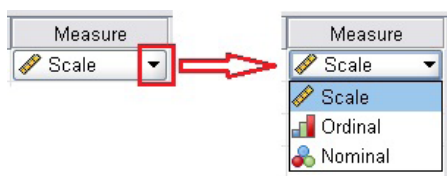


Рисунок 6 – Шкала переменной

- *scale* – метрическая шкала, указывается для тех данных переменной, которые можно измерить с помощью чего-либо или над которыми можно произвести арифметические действия – сложить или вычесть;
- *ordinal* – порядковая шкала, указывается для тех типов данных, которые не допускают арифметических действий, но которые можно сравнить – «больше» или «меньше»;
- *nominal* – номинальная шкала, указывается во всех остальных случаях, для которых нельзя применить «*scale*» или «*ordinal*».

### Типы переменных

В статистике переменной называют характеристику (признак), которая описывает свойства изучаемого объекта (пациента, лабораторного животного, клеточной культуры и т. д.) и может

принимать различные значения (меняться или варьироваться). Выбор переменных зависит от целей исследования. Факторы риска и исходы с точки зрения статистики также являются переменными. Примерами переменных являются пол, возраст, концентрация биомаркера, число беременностей в анамнезе, стадия злокачественного новообразования, анамнез курения и др. Первым этапом анализа любого эксперимента является выбор переменных и оценка их шкалы измерения [12]. Существуют два основных типа данных: количественные и качественные (рисунок 7).

Количественные переменные можно измерить в числовом выражении, а их величину можно представить с помощью метрической шкалы. Например, их количество можно посчитать или измерить в метрах, килограммах, ммоль/л или других единицах. Среди количественных переменных выделяют непрерывные и дискретные.

Непрерывные переменные могут принимать бесконечное число значений в промежутке между двумя любыми значениями, включая целые числа и доли. Примерами могут служить масса тела, концентрация билирубина или температура.

Дискретные переменные могут принимать ограниченное число значений, которые обычно представлены целыми числами. Например, число обострений заболевания и количество клеток в поле зрения микроскопа являются дискретными величинами [12].

Качественные переменные (категориальные) невозможно измерить, вместо этого их разделяют на группы (категории). Например, цвет волос нельзя измерить по метрической шкале, однако



Рисунок 7 – Типы переменных

можно выделить несколько групп (категорий) этого признака: темные волосы, светлые, рыжие и т. д. Качественные переменные разделяют на порядковые (ординарные) и номинальные [12].

Порядковые переменные могут быть ранжированы в соответствии со своим рангом. В частности, стадия опухолевого процесса не измеряется по метрической шкале, хотя более высокие значения характеризуют более тяжелое течение (поздние стадии) заболевания.

Номинальные переменные не могут быть ранжированы, как, например, место жительства, профессия и некоторые другие признаки. Одним из наиболее распространенных типов номинальных переменных являются бинарные (дихотомические) переменные, которые могут принимать только два значения (исход наступил или нет, пациент выжил или умер, мужской пол или женский) (см. рисунок 7) [12].

После того, как были разобраны все столбцы программы SPSS и все типы переменных, перейдем непосредственно к созданию переменных в самом учебном файле.

## **2.5. Алгоритм для создания файла в программе SPSS**

Для того чтобы начать создавать любой файл, нужно выбрать какую-либо нозологию (иными словами заболевание).

Давайте рассмотрим вариант создания файла на примере такой распространенной нозологии как Гипертоническая болезнь (ГБ). Необходимо воспользоваться ресурсами Интернета и посмотреть, что представляет собой данная нозология и какие показатели (то есть переменные) характерны для нее. Для работы необходима информация следующего типа: определение названия заболевания, краткое ее описание, классификация, лабораторная и/или функциональная диагностика и лечение. Данная информация нужна для того, чтобы начать формировать базовый файл данных СРС, то есть первую его часть – реферат.

Исходя из заданной нозологии – Гипертоническая болезнь, был сформирован короткий реферат (приложение 1).

Было найдено достаточно много переменных, для создания базового файла. Конечно, необходимы не все переменные, но обязательно нужно включить такие переменные, как: пол, возраст, вредные привычки и виды заболевания. Остальные возьмем из тех, которые были найдены в ходе создания реферата. Чтобы данные можно было полноценно анализировать и смотреть динамику выздоровления, разобьем все переменные (кроме пола, возраста, вредных привычек и вида заболевания) на две переменные – переменная «до лечения» и переменная «после лечения».

Далее, для облегчения создания базового учебного файла, предварительно составим таблицу переменных (таблица 1):

После определения всех необходимых переменных и уточнения их показателей, открываем программу SPSS, вкладку Variable View и начинаем ввод переменных.

**Пол:**

- name – «пол»;
- type – numeric, поскольку необходимо использовать цифровые данные, вместо буквенных;
- decimals – 0, переменная пол имеет только одну позицию – 1 или 2, то есть это целые числа, нет десятых и сотых (см. таблицу 1);
- width – 1, по причине, описанной выше (см. таблицу 1);
- label – «пол пациента»;
- values – необходимо провести кодировку из буквенных данных в числовые, то есть закрепить за конкретными цифрами буквенную информацию – 1 – женский, 2 – мужской (см. рисунок 5);
- measure – «nominal», для данной переменной мы не можем применить арифметических действий или сравнить, то есть нельзя сложить «мужчину» с «женщиной» и получить какое-то число или нельзя сказать, что «мужчина» – это нечто большее чем «женщина».

Таблица 1 – Перечень переменных

№ п/п	Переменная/имя (Name)	Кодировка (Value)	Нормальные значения	Шкала измерения (Measure)
1	Пол/пол	1 – женщина 2 – мужчина	-	nominal
2	Возраст/возраст	-	-	scale
3	Курение/курение	1 – курит 2 – не курит	-	ordinal
4	Диагноз/диагноз	1 – эссенциальная гипертензия I ст; 2 – эссенциальная гипертензия II ст; 3 – реноваскулярная гипертензия; 4 – ренопривная гипертензия	-	ordinal
5	Систолическое давление (до и после лечения)/сад	-	Категории САД Оптимальное < 120 Нормальное 120 – 129 Высокое нормальное 130 – 139 АГ I степени 140 – 159 АГ II степени 160 – 179 АГ III степени ≥ 180	scale
6	Диагностическое давление (до и после лечения)/даддо	-	Категории ДАД Оптимальное < 80 Нормальное 80 – 84 Высокое нормальное 85 – 89 АГ I степени 90 – 99 АГ II степени 100 – 109 АГ III степени ≥ 110	scale

7	Общий холестерин до лечения/охдо	-	Норма 3,5 – 5,0 ммоль/л	scale
8	Липопротеиды высокой плотности до лечения/лвнпдо	-	Норма 1 – 1,55 ммоль/л	scale
9	Липопротеиды низкой плотности до лечения/лпнпдо	-	Норма 2,6 – 4,1 ммоль/л	scale
10	Триглицериды до лечения/тгдо	-	Норма 0,15 – 1,80 ммоль/л	scale
11	Сахар сыворотки крови до лечения/сск	-	Норма 3,3 – 5,5 ммоль/л	scale
12	индекс массы миокарда левого желудочка (до и после лечения)/иммлж	-	Женщины: 70 – 100 Мужчины: 90 – 120	scale
13	толщина интим-медиа/тим	-	Норма – до 0,900 мм	scale
14	альбумины мочи/ам	-	До 60 мг/г	scale
15	креатинин мочи/км	-	5,5–17,7 ммоль/сут.	scale
16	креатинин крови/кк	-	55–115 мкмоль/л	scale
17	скорость клубочковой фильтрации/скф	-	80–150 мл/мин	scale

### **Возраст:**

- name – «возраст»;
- type – numeric, возраст всегда записывается в числах;
- decimals – 0, всегда в медицинских документах указывается полное количество лет (кроме детей до 1 года);
- width – 3, есть почтенные пациенты старше 100 лет;
- label – «возраст пациента»;
- values – остается незаполненным, так как возраст вводится числом;
- measure – «scale», данную переменную мы можем посчитать. Например: можно сложить возраст всех пациентов и разделить на их количество, то есть рассчитать среднее арифметическое.

### **Вредные привычки:**

- name – «курение»;
- type – numeric, поскольку необходимо использовать цифровые данные, вместо буквенных, примем следующие позиции: 1 – курит, 2 – не курит;
- decimals – 0, переменная привычки имеет только одну позицию – 1 или 2, то есть это опять целые числа, нет десятых и сотых (см. таблицу 1);
- width – 1 (см. таблицу 1);
- label – «курение пациента»;
- values – необходимо провести кодировку из буквенных данных в числовые, то есть закрепить за конкретными цифрами буквенную информацию: 1 – курит, 2 – не курит;
- measure – «ordinal», можно сравнить, то есть мы можем сказать, лучше не иметь вредных привычек, чем бороться с ними.

По такому же принципу необходимо заполнить переменную «диагноз».
------------------------------------------------------------------

### **Систолическое артериальное давление до лечения:**

- name – «систолическое»;
- type – numeric, используются числа при определении систолического давления;

- decimals – 0, систолическое давление записывается целыми числами (см. таблицу 1);
- width – 3, общее количество знаков составляет 3, от 131 и выше (см. таблицу 1);
- label – «систолическое артериальное давление до лечения»;
- values – остается незаполненным;
- measure – «scale», существует тонометр, который может замерить систолическое артериальное давление.

По такому же принципу необходимо заполнить переменные «садпос», «даддо» и «дадпос».

#### **Общий холестерин до лечения:**

- name – «охдо»;
- type – numeric, используются числа при определении общего холестерина;
- decimals – 1, после запятой в уровнях общего холестерина есть цифры (см. таблицу 1);
- width – 2, общее количество знаков составляет 2, один знак до запятой и один после (см. таблицу 1);
- label – «общий холестерин до лечения»;
- values – остается незаполненным;
- measure – «scale», существует электронный микроскоп, который может посчитать общий холестерин.

По такому же принципу необходимо заполнить переменные «лпвпдо», «лпппдо», «тгдо», «сск» и другие.

Вот что у Вас должно в результате получиться (рисунок 8):

Теперь остается открыть и заполнить вкладку DATA VIEW на вымышленных пациентов, но с данными, взятыми из таблицы 1. При вводе данных не забывайте использовать, если надо, десятичную «запятую», а не «точку» (рисунок 9):

Не забывайте периодически сохранять введенные данные, экономьте время.

Иванов_ГБ.sav [DataSet1] - SPSS Data Editor										
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	возраст	Numeric	2	0	возраст	None	None	8	Right	Scale
2	пол	Numeric	1	0	пол пациента	{1, женщина}...	None	8	Right	Nominal
3	курение	Numeric	1	0	курение пациента	{1, курит}...	None	8	Right	Nominal
4	диагноз	Numeric	1	0	диагноз пациента	{1, эссенция...	None	8	Right	Ordinal
5	систолическое артериальное давление до лечения	Numeric	3	0	систолическое артериальное давление до лечения	None	None	8	Right	Scale
6	диастолическое артериальное давление до лечения	Numeric	3	0	диастолическое артериальное давление до лечения	None	None	8	Right	Scale
7	общий холестерин ммоль/л до лечения	Numeric	3	1	общий холестерин ммоль/л до лечения	None	None	8	Right	Scale
8	липопротеиды высокой плотности до лечения	Numeric	2	1	липопротеиды высокой плотности до лечения	None	None	8	Right	Scale
9	липопротеиды низкой плотности до лечения	Numeric	2	1	липопротеиды низкой плотности до лечения	None	None	8	Right	Scale
10	триглицериды до лечения	Numeric	2	1	триглицериды до лечения	None	None	8	Right	Scale
11	сахар сыворотки крови до лечения	Numeric	3	1	сахар сыворотки крови до лечения	None	None	8	Right	Scale
12	индекс массы миокарда левого желудочка до лечения	Numeric	3	0	индекс массы миокарда левого желудочка до лечения	None	None	8	Right	Scale
13	толщина интимы-медиа	Numeric	2	1	толщина интимы-медиа	None	None	8	Right	Scale
14	альбумины мочи в сутки	Numeric	3	0	альбумины мочи в сутки	None	None	8	Right	Scale
15	креатинин мочи	Numeric	3	1	креатинин мочи	None	None	8	Right	Scale
16	скорость клубочковой фильтрации	Numeric	2	0	скорость клубочковой фильтрации	None	None	8	Right	Scale
17	креатинин сыворотки крови	Numeric	3	1	креатинин сыворотки крови	None	None	8	Right	Scale
18	систолическое артериальное давление после лечения	Numeric	3	0	систолическое артериальное давление после лечения	None	None	8	Right	Scale
19	диастолическое артериальное давление после лечения	Numeric	3	0	диастолическое артериальное давление после лечения	None	None	8	Right	Scale
20	индекс массы миокарда левого желудочка после лечения	Numeric	3	0	индекс массы миокарда левого желудочка после лечения	None	None	8	Right	Scale

Рисунок 8 – Учебный файл со всеми заполненными переменными

Муанов_16.sav [DataSet1] - SPSS Data Editor																				
SS:																				
	Диагноз	сэдэд	дараа	оходо	лүгдэд	лүгдэд	түдэд	сүх	иммунгэд	түм	эм	им	сүф	күк	сэдүс	дэдүс	иммүнгүс	v/e		
1	1	160	60	6,2	1,5	3,5	2,1	5,5	110	1,0	10	4,4	60	115,0	150	80	110			
2	2	180	110	5,0	1,4	6,3	3,0	4,2	132	1,5	10	5,1	64	120,0	170	90	126			
3	3	190	110	1,6	1,5	2,9	1,6	5,2	115	1,0	160	1,5	40	130,0	180	90	115			
4	4	180	100	1,6	1,0	3,0	1,6	3,6	120	2,0	200	2,5	50	141,0	170	90	115			
5	1	180	110	5,0	1,0	4,0	1,7	6,0	125	0,8	30	5,0	61	107,0	170	100	110			
6	2	160	100	4,6	1,0	2,5	1,6	5,4	120	0,7	10	4,6	60	116,0	150	90	117			
7	3	195	95	6,5	0,3	6,0	2,3	6,9	110	1,2	245	2,0	30	100,0	185	95	110			
8	1	150	90	3,3	1,0	3,0	1,5	5,7	120	0,8	20	5,1	63	132,0	140	80	120			
9	1	176	100	4,9	0,9	2,9	1,4	4,8	111	0,6	25	6,2	60	117,0	160	80	111			
10	3	160	100	7,2	0,3	5,4	3,2	7,1	130	2,0	309	1,0	40	90,0	150	90	125			
11	4	170	110	6,2	0,8	5,1	2,5	4,0	127	1,9	400	3,3	45	70,0	160	90	120			
12	2	179	50	5,2	0,7	2,8	0,9	5,5	130	0,9	30	5,6	60	131,0	165	80	118			
13	3	190	100	5,0	1,3	2,7	4,1	7,2	330	5,3	402	3,1	30	80,0	180	90	200			
14	1	140	80	5,0	1,1	2,5	1,1	5,9	110	0,8	15	6,5	60	118,0	120	80	110			
15	2	160	100	4,0	0,8	3,0	1,2	5,6	115	0,9	25	7,1	70	130,0	150	90	115			
16	4	200	130	6,4	0,7	4,6	2,5	3,5	130	0,9	207	2,5	40	50,0	180	100	127			
17	3	190	70	7,4	0,3	4,7	2,0	6,9	150	1,5	310	3,2	30	75,0	170	70	135			
18	2	160	105	5,0	1,5	3,0	1,4	3,3	178	0,8	60	7,8	60	120,0	155	80	170			
19	3	170	50	3,1	1,0	3,0	1,9	6,8	200	1,6	40	3,4	50	64,0	140	60	159			
20	1	150	90	5,1	1,5	2,7	1,4	5,5	330	0,7	35	8,1	60	119,0	140	90	205			
21	2	160	103	6,5	0,3	3,5	1,7	5,2	125	0,8	40	10,1	65	132,0	139	90	125			

Рисунок 9 – Учебный файл с заполненными данными

Надеемся, что при создании учебного файла, у Вас было не так много вопросов. Но если они все-таки есть, мы их обязательно обсудим.

## **2.6. Задания для самостоятельной работы**

Для выполнения первого самостоятельного задания необходимо:

1. Попросить у преподавателя определить для Вас болезнь (нозологию).

2. Изучить выбранную болезнь, которая будет являться Вашей темой для выполнения самостоятельной работы. Для получения актуальной информации по выбранной теме советуем использовать <https://www.elibrary.ru/>, <https://www.ncbi.nlm.nih.gov/>. В качестве переводчика можно использовать функцию «перевести на русский», которую Вы можете запустить, щелкнув правой клавишей мыши. Сделать краткое описание болезни, не более 1–2 страниц, с указанием ссылок на источники используемой литературы (не больше 5) (оформление СРС смотри в приложении 1).

3. Составить таблицу переменных (не больше 20), характерных для предложенного заболевания (на примере таблицы 1, см. выше). На основе полученных данных и таблицы создать файл в программе SPSS. Все переменные должны быть связаны с описываемой болезнью, поэтому соответствующие ссылки на эти переменные должны быть в описании. Для лабораторных показателей создать показатели «до» и «после лечения». Тип измерения (measure) для данных показателей должен быть «scale», иначе невозможно будет сделать сравнение со связанными выборками. Среди 20 переменных должны быть минимум 3 переменных типа «nominal» или «ordinal», такие как: 1) пол, 2) вредные привычки или курение и др., 3) вид заболевания, иначе не получится решить задачу расчета корреляции по второму типу.

4. Заполнить полученный файл данными, согласно предложенному заболеванию, то есть представить, что Вы описываете

данные на 50 вымышленных пациентов, но опираясь на реальные показатели.

5. Обязательно сохранить полученный файл на компьютере (продублировать на флеш-карте).

Имя файла должно быть следующего формата (рисунок 10):



Рисунок 10 – Пример названия зачетного файла SPSS

6. Вводную часть СРС и файл предоставить Вашему преподавателю для проверки и необходимой корректировки.

## 2.7. Тестовые вопросы

**1. Перечислите основные вкладки, из которых состоит редактор данных?**

- a) data view;
- b) variable view;
- c) scale;
- d) nominal;
- e) ordinal.

**2. Вкладка, предназначенная для создания и редактирования переменных, называется:**

- a) variable view;
- b) data view;
- c) scale;
- d) nominal;
- e) ordinal.

**3. Вкладка, предназначенная для создания и редактирования данных, называется:**

- a) data view;
- b) variable view;
- c) scale;
- d) nominal;
- e) ordinal.

**4. Перечислите, что категорически нельзя использовать при написании имени переменной:**

- a) пробел;
- b) точку;
- c) запятую;
- d) цифру;
- e) букву.

**5. Укажите столбец, в котором задается имя новой переменной:**

- a) name;
- b) decimals;
- c) type;
- d) width;
- e) label.

**6. Укажите столбец, отвечающий за тип конкретной переменной:**

- a) type;
- b) decimals;
- c) name;
- d) width;
- e) label.

**7. Укажите столбец, отвечающий за количество десятичных знаков конкретной переменной:**

- a) decimals;
- b) name;
- c) type;
- d) width;
- e) label.

**8. Укажите столбец, отвечающий за общее количество знаков конкретной переменной:**

- a) width;
- b) decimals;
- c) name;
- d) type;
- e) label.

**9. Укажите столбец, в котором дается текстовая информация конкретной переменной:**

- a) label;
- b) width;
- c) decimals;
- d) name;
- e) type.

**10. Укажите столбец, отвечающий за кодирование пропущенных данных конкретной переменной:**

- a) missing;
- b) width;
- c) decimals;
- d) type;
- e) label.

**11. Укажите столбец, отвечающий за ширину столбца конкретной переменной:**

- a) columns;
- b) missing;
- c) width;
- d) decimals;
- e) label.

**12. Укажите столбец, отвечающий за шкалу измерения конкретной переменной:**

- a) measure;
- b) columns;
- c) missing;
- d) decimals;
- e) label.

**13. Укажите, какая шкала измерения используется для таких переменных, значения которых можно измерить с помощью чего-либо, либо применить математические действия – сложение или вычитание:**

- a) scale;
- b) columns;
- c) nominal;
- d) decimals;
- e) ordinal.

**14. Укажите, какая шкала измерения используется для таких переменных, значения которых можно сравнить – больше или меньше:**

- a) ordinal;
- b) scale;
- c) columns;
- d) nominal;
- e) decimals.

**15. Укажите, какой знак используется при вводе десятичных данных в файл данных:**

- a) запятая;
- b) точка;
- c) точка с запятой;
- d) двоеточие;
- e) пробел.

**16. Сопоставьте параметры с типом измерения. Укажите соответствие для всех 5 вариантов ответа:**

- |             |                             |
|-------------|-----------------------------|
| a) scale;   | ___ пол;                    |
| b) ordinal; | ___ возраст;                |
| c) nomina.  | ___ лейкоциты крови;        |
|             | ___ систолическое давление; |
|             | ___ стадия заболевания.     |

**17. Сопоставьте параметры с типом измерения. Укажите соответствие для всех 5 вариантов ответа:**

- |             |              |
|-------------|--------------|
| a) scale;   | ___ пол;     |
| b) ordinal; | ___ возраст; |

- c) nominal.                      \_\_\_ сахар крови;  
                                         \_\_\_ сахар мочи;  
                                         \_\_\_ тяжесть заболевания.

**18. Сопоставьте параметры с типом измерения. Укажите соответствие для всех 5 вариантов ответа:**

- a) scale;                            \_\_\_ пол;  
b) ordinal;                        \_\_\_ возраст;  
c) nominal.                        \_\_\_ класс;  
                                         \_\_\_ средняя оценка;  
                                         \_\_\_ хобби.

## **Глава 3. ОПИСАТЕЛЬНЫЕ СТАТИСТИКИ**

### **ЧАСТЬ 1**

**3.1. Описательные статистики.**

**3.2. Частоты.**

**3.3. Описательные таблицы.**

**3.4. Частоты + Выборки.**

**3.5. Задачи для закрепления материала.**

**3.6. Задания для самостоятельной работы.**

**3.7. Тестовые вопросы.**

#### **3.1. Описательные статистики**

Наука всегда имеет дело с разнообразием действительности, но свою задачу она видит в обнаружении порядка вещей, некоторой устойчивости внутри наблюдаемого разнообразия. Например, антропология изучает физические особенности строения тела человека. По признакам строения тела (цвет кожи, глаз, волос, форма черепа и т. д.) она выделяет определенные типы – человеческие расы. Но ее не интересуют такие признаки, как количество глаз, ушей или пальцев, поскольку это константы: люди (в норме) не различаются между собой по этим признакам. Наука имеет дело с переменными, однако она стремится выделить такие их значения, которые лучше всего объясняют наблюдаемое разнообразие проявлений. Статистика снабжает нас удобными приемами такого анализа [15].

Один из наиболее известных и влиятельных статистиков XX века Рональд Фишер писал в своей работе «О математических основах теоретической статистики» о том, что «...целью статистической обработки является сокращение объема данных. Большой объем данных ... должен быть сведен к относительно небольшому числу параметров, которые адекватно описывают целое». На наш

взгляд, эта цитата превосходно передает сущность описательной статистики – раздела статистики, посвященного обобщению и описанию данных, который представляет инструменты для исследования массивов данных и их представления. Медицинские исследования могут объединять данные сотен или тысяч единиц наблюдения (отдельных пациентов, клеток или других объектов), которые невозможно понять, изучая их по отдельности. Однако исследователей обычно интересует ограниченное число параметров, таких как, например, возраст или анамнез курения, которые можно обобщить и представить с использованием методов описательной статистики и таким образом получить представление обо всей выборке. Другими словами, необходимо стремиться описать свойства всей выборки несколькими числами или одним графиком [12].

Для использования статистики требуются два основных условия:

- 1) мы должны иметь данные о группе (выборке) людей;
- 2) эти данные должны быть представлены в формализованном (кодифицированном) виде.

Задачи описательной статистики в целом совпадают с задачами качественного анализа: это сжатие информации и представление ее в удобном для обозрения виде.

Наиболее простой случай возникает тогда, когда наши данные могут быть представлены одной переменной. Используемые для этого методы объединяются в группу под названием одномерный статистический анализ. В зависимости от типа данных и задач, которые ставит исследователь, выбирается тот или иной конкретный способ анализа [15].

Достаточно простым и удобным приемом анализа количественных данных является построение распределения частот. Проиллюстрируем его на примере. Допустим, проводится исследование, в котором каждому участнику, среди прочих, задается вопрос о его семейном положении. Ответы фиксируются, а общие результаты сводятся в таблицу, где против фамилии каждого участника указано его семейное положение. Далее эти первичные данные

мы хотим представить в более сжатой и удобной для анализа форме. Для этого мы разбиваем их на категории, подсчитывая количество людей в каждой группе. Сами категории выбираются в зависимости от целей исследования и могут быть более широкими или более узкими. Если используется вопрос закрытого типа, то мы сразу фиксируем ответы в кодированной форме. Если же используется вопрос открытого типа, то кодирование осуществляет сам исследователь. Отметим, что более узкие категории всегда можно укрупнить. Но если информация структурируется прямо на этапе ее сбора, то потом нельзя будет разбить данные на большее число категорий. Часть информации оказывается потерянной [15].

Анализируя таблицу, мы сравниваем между собой категории и видим, как они представлены в нашей выборке. Нас чаще всего интересуют не абсолютные значения, а относительные. Относительные единицы позволяют сравнивать не только аналогичные показатели, полученные на разных выборках, но и качественно различные показатели между собой. Например, для характеристики имущественного расслоения в обществе используют следующий прием: подсчитывают, какой процент всей собственности сосредоточен в руках десяти процентов самых богатых людей. Понятно, что в случае полного имущественного равенства этот показатель будет равен десяти процентам. Чем он больше отличается от этого значения, тем выше степень имущественного неравенства в обществе [15].

До сих пор мы разбирали случай, когда исходные данные были качественными, то есть образовывали шкалу наименований. Но мы можем строить распределение частот и для шкал порядка или отношений. Правда, в последнем случае обычно данные представляют в сгруппированном виде, так как иначе образуется очень много классов. Например, мы исследуем возрастной состав той же группы из тридцати человек. Если она не однородна в этом отношении, данные окажутся «размазанными». Тогда мы их группируем, выбирая определенный шаг (например, пять или десять лет), и вносим в таблицу обобщенные данные. Шаг выбирается с учетом характера данных и задач анализа. Снова отметим, что

группировка данных приводит к потере части информации, но зато мы добиваемся ее лучшей обозримости [15].

### 3.2. Частоты

Первым этапом статистического анализа данных, как правило, является частотный анализ, то есть отбор наблюдений, который осуществляется несколькими способами.

Таблицы частот представляют собой простейший метод анализа категориальных (номинальных) переменных. Часто их используют как одну из процедур разведочного анализа, чтобы просмотреть, каким образом различные группы данных распределены в выборке.

Практически каждый исследовательский проект начинается с построения таблиц частот. Например, в медицинских исследованиях табулируют пациентов по полу, с определенными симптомами, заболеваниями и т. д.

В этом параграфе мы проведем частотный анализ с помощью команды «*Frequencies*» на примере созданного учебного файла «Иванов\_ГБ.sav». Напомним, что данный файл содержит результаты лечения пациентов с гипертонической болезнью (см. параграф 2.5). Файл содержит такие данные как: значения систолического и диастолического давления, уровень холестерина, сахар сыворотки крови, креатинин и альбумины мочи и т. д. Кроме того, имеются такие данные, как пол, возраст, курение и диагноз. Но необходимо не забывать, что с командой «Частоты» используются только переменные «*Nominal*» и «*Ordinal*».

Для того чтобы начать проводить частотный анализ, то есть «Кого и сколько в данном файле имеется?», необходимо предельно следующие действия:

1. Открыть файл «Иванов\_ГБ.sav».
2. Поставить соответствующую **задачу**, например – Посчитать, сколько в данном файле мужчин и женщин?
3. Для решения поставленной задачи необходимо воспользоваться командой «*Frequencies*», то есть заходим в меню:

## Analyze → Descriptive Statistics → Frequencies

появится диалоговое окно «*Frequencies*» (рисунок 11, а). Переносим в правую часть переменную «пол», так как именно данная переменная содержит информацию о поле респондентов. Для этого необходимо найти переменную «пол» в левой части окна в списке переменных, щелкнуть левой клавишей мыши на ней и нажать стрелочку между двумя частями окна. Нажать кнопку «OK» (рисунок 11, б).

4. Результаты наших вычислений появятся в отдельном файле, который называется «*Output 1*», или файл выдачи. Он отражает результаты всех возможных действий в виде текста, таблиц и графиков (рисунок 12).

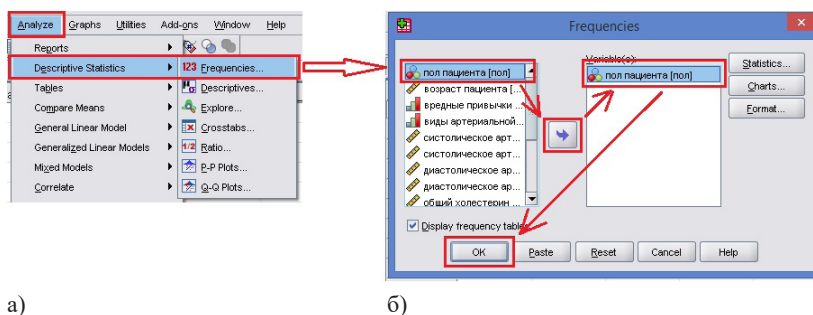


Рисунок 11 – Алгоритм (а) и рабочее окно (б) команды «Frequencies»

Statistics

пол пациента

N	Valid	100
	Missing	0

пол пациента

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	женщина	55	55,0	55,0	55,0
	мужчина	45	45,0	45,0	100,0
Total		100	100,0	100,0	

Рисунок 12 – Таблицы результатов вычисления команды «Frequencies»

5. Далее остается только прочитать эти таблицы и дать итоговый вывод.

Первая таблица показывает обзор допустимых и отсутствующих значений.

Для интерпретации результатов надо обратить внимание на вторую таблицу, в которой имеются все необходимые данные. Давайте разберем, что означает каждый столбец (см. рисунок 12).

Первый столбец показывает, какие варианты имеет переменная «пол»:

- женщина, мужчина;
- total (всего заполненных данных).

Во втором столбце, который называется «*Frequency*» приведены данные для двух вариантов ответа на вопрос пол. Мы видим, что в данном файле имеются женщин 55 и мужчин 45, всего заполнены данные у 100 пациентов, пропущенных или незаполненных данных (*Missing system*) нет.

В третьем столбце, который называется «*Percent*» приведены процентные данные, по каждому ответу: женщин – 55,0 %, мужчин – 45,0 %, всего заполнено у 100,0 %, пропущенных или незаполненных данных нет.

В четвертом столбце, который называется «*Valid Percent*» приведен расчет процентов, только по заполненным данным: женщин – 55,0 %, мужчин – 45,0 %, всего 100,0 %. Так как для данной переменной нет пропущенных или незаполненных данных, графы «*Percent*» и «*Valid Percent*» содержат идентичные данные.

В пятом столбце, который называется «*Cumulative Percent*» приведены накопленные процентные данные, то есть сумма процентных частот допустимых ответов. В нашем случае женщин и мужчин 100 %, это число определилось выражением:  $55,0 \% + 45,0 \% = 100,0 \%$ . Мы полностью разобрали полученную таблицу, осталось дать правильный ответ.

6. Ответ – в данном файле имеются 55 женщин (55,0 %) и 45 мужчин (45,0 %) из 100 пациентов.

### 3.3. Описательные таблицы

Описательные таблицы или таблицы сопряженности также дают ответ на вопрос «Кого и сколько в данном файле имеется?», но в данном случае это можно узнать путем использования двух переменных. Также необходимо не забывать, что с данной командой используются только переменные «*Nominal*» и «*Ordinal*».

Давайте посмотрим, как работает данная команда. Необходимо:

1. Открыть файл «Иванов\_ГБ.sav».
2. Поставить соответствующую **задачу**, например – Посчитать, сколько в данном файле курящих и некурящих женщин и мужчин?
3. Для решения поставленной задачи необходимо убедиться, что данные переменные являются переменными «*Nominal*» и «*Ordinal*». Затем воспользуемся командой «*Crosstabs*»:

*Analyze* → *Descriptive Statistics* → *Crosstabs*

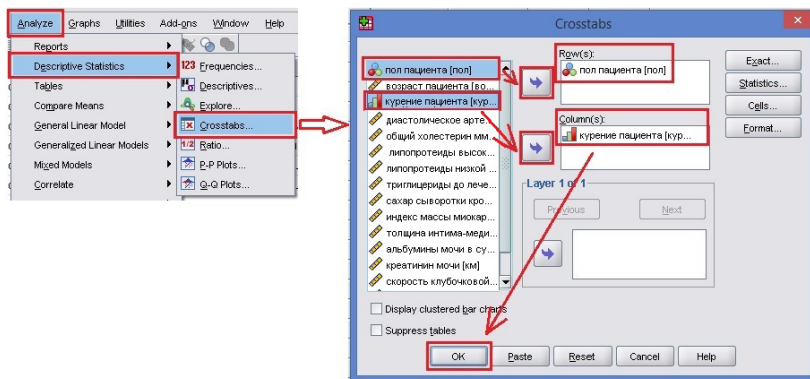
появится диалоговое окно «*Crosstabs*» (рисунок 13). Переносим переменную «курение» в блок *Row(s)* – строки, а переменную «пол» в блок *Column(s)* – столбцы и нажимаем кнопку «ОК». Можно переносить и наоборот – «пол» в *Row(s)*, а «курение» в *Column(s)*, принципиальной разницы в этом нет (см. рисунок 13).

4. Результаты наших вычислений появились в виде таблиц (рисунок 14).

Для интерпретации результатов надо обратить внимание на вторую таблицу, в которой имеются все необходимые данные.

5. **Ответ:** в данном файле 35 женщин курят и 20 – не курят; 30 мужчин курят, 15 – не курят.

Также мы можем получить и дополнительную информацию о мужчинах и женщинах – курящих и некурящих, то есть еще можно сказать, что в данном файле 55 женщин и 45 мужчин; 65 – курящих, 35 – некурящих.



а)

б)

Рисунок 13 – Алгоритм (а) и рабочее окно (б) команды «Crosstabs»

Case Processing Summary						
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
пол пациента * курение пациента	100	100,0%	0	,0%	100	100,0%

пол пациента \* курение пациента Crosstabulation

		курение пациента		Total
		курит	не курит	
пол пациента	женщина	35	20	55
	мужчина	30	15	45
Total		65	35	100

Рисунок 14 – Таблицы результатов вычисления команды «Crosstabs»

### 3.4. Частоты + Выборки

Выбор или отбор данных – это выбор наблюдений по определенным условиям, например – сколько в данном файле женщин, или только мужчин, или только пациентов старше 50 лет. После конкретного отбора все вычисления в файле будут проводиться только с этими конкретно отобранными данными. В данном

случае необходимо использовать дополнительную команду «*Select Cases*», которая работает со всеми типами переменных.

Для того чтобы произвести отбор необходимо:

1. Открыть файл «Иванов\_ГБ.sav».
2. Поставить **задачу**, например – Посчитать, сколько в данном файле курящих мужчин, у которых установлен диагноз – Эссенциальная гипертензия I степени.

3. Для решения поставленной задачи необходимо сначала выполнить конкретное условие отбора, то есть создать фильтр, а потом посчитать пациентов, соответствующих данному условию отбора. Поэтому переходим во вкладку «*VARIABLE VIEW*», для того чтобы посмотреть, какие переменные содержат в себе необходимую нам информацию. Это такие переменные, как «привычки», «пол» и «диагноз». Далее заходим в меню

***Data → Select cases***

Появится диалоговое окно команды «*Select Cases*» (рисунок 15, а). Для того чтобы начать вводить условие отбора, необходимо поставить галочку в позицию «*if condition is satisfied*» – «если выполняется условие», а затем нажать кнопку «*if*» – «если» (рисунок 15, б). После нажатия кнопки «*if*» появляется окно условий «*Select cases-if*» (рисунок 15, в).

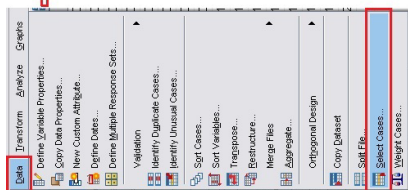
Окно условий разделено на несколько частей (см. рисунок 15, в):

1. **Список переменных:** содержит все переменные, имеющиеся в данном открытом файле.

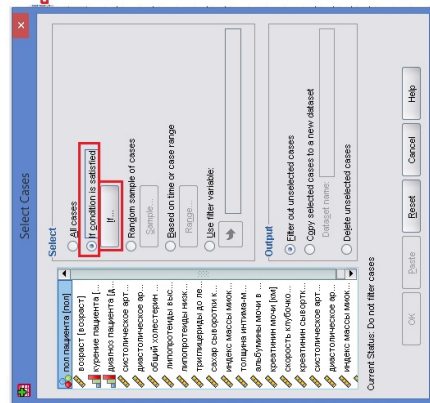
2. **Редактор условий:** в нем записывается логическое выражение или, другими словами, условие отбора.

3. **Клавиатура:** содержит цифры, а также арифметические, логические операторы и операторы отношения; с ней можно работать, как с обыкновенным калькулятором при написании логического выражения.

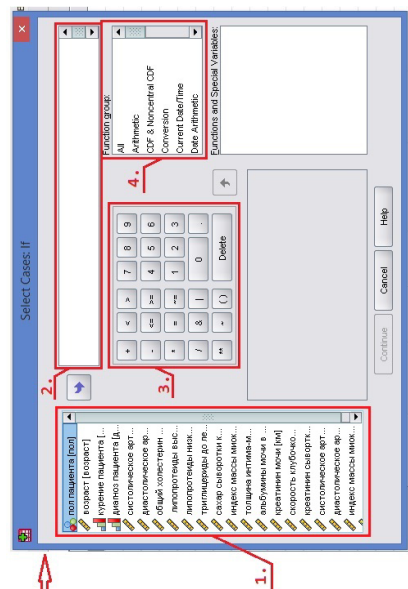
4. **Список функций:** содержит около 140 функций. Каждую из которых, можно скопировать в редактор условий двойным щелчком.



а)



б)



в)

Рисунок 15 — Алгоритм (а), рабочее окно (б) и окно условий (в) команды «Select Cases»

**Для справки:** в программе SPSS наиболее часто используются операторы, которые делятся на арифметические, логические и операторы отношения. Арифметические операторы применяются в математических формулах, которые при отборе данных имеют лишь второстепенное значение.

Логические операторы и операторы отношения применяются исключительно в логических выражениях, которые мы рассмотрим далее (рисунок 16).

### **Операторы отношения**

- < – меньше, чем...;
- > – больше, чем...;
- ≤ – меньше или равно чему-то...;
- ≥ – больше или равно чему-то...;
- = – равно чему-то... .

Например, «курение = 1»; «возраст > 50» и так далее.

### **Логические операторы**

- & – and (и) – означает «и то, и другое»;

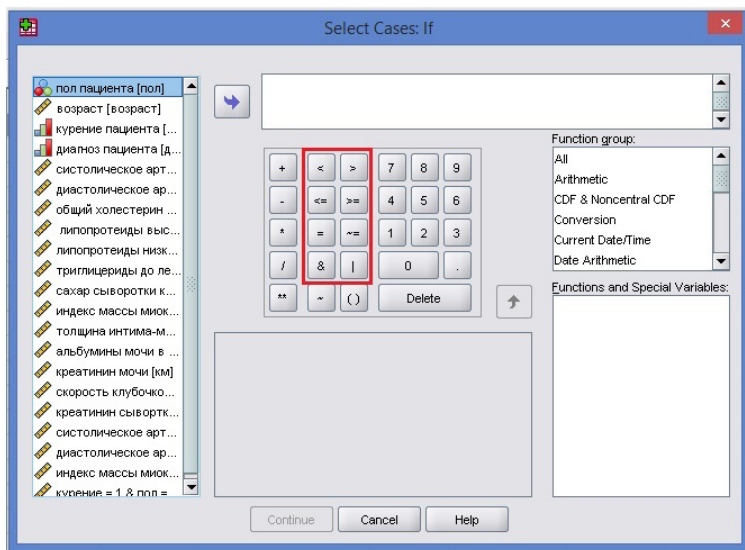


Рисунок 16 – Операторы отношения и логические операторы команды «Select Cases if»

| – or (или) – означает «или то, или другое, или оба».

Когда используются две и более переменных, их соединяют одним из двух логических операторов, в зависимости от условия.

### **Использование оператора «И»**

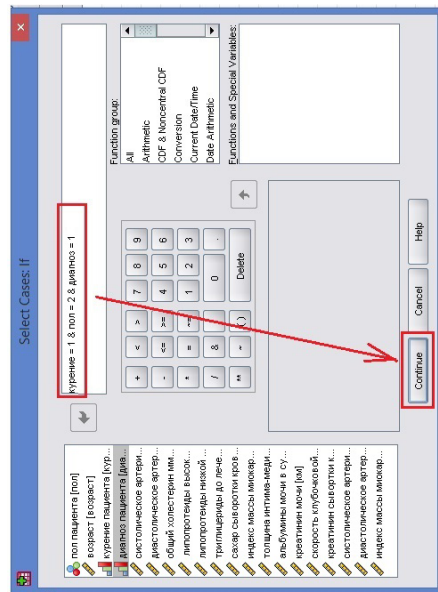
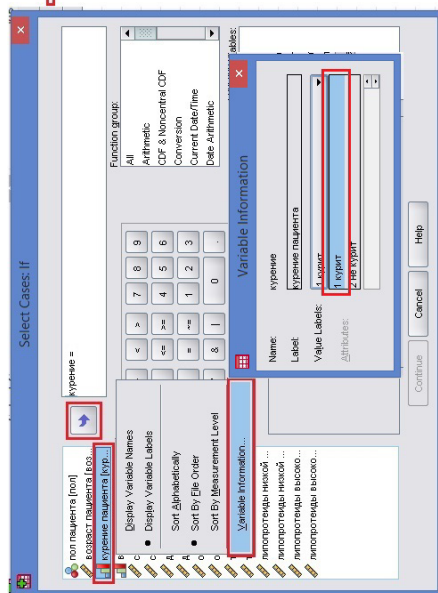
**Пример 1.** Необходимо отобрать курящих мужчин, в данном случае необходимо выполнить два условия: 1 – только курящие «И»; 2 – только мужчины, то есть условие принимает вид – «курение = 1 & пол = 2».

**Пример 2.** Необходимо отобрать курящих женщин старше 50 лет, в данном случае необходимо применить логические операторы и выполнить уже три условия: 1 – только курящее «И»; 2 – только женщины «И»; 3 – только старше 50 лет. Плюс к третьему условию необходимо добавить оператор отношения: старше 50 лет, это значит возраст больше «>» 50, то есть условие принимает вид – «курение = 1 & пол = 1 & возраст > 50».

### **Использование оператора «ИЛИ»**

**Пример.** Необходимо отобрать женщин с эссенциальной гипертензией I степени или эссенциальной гипертензией II степени. В данном случае необходимо применить оба логических оператора «И» и «ИЛИ» и выполнить уже четыре условия: 1 – только женщины «И»; 2 – только с диагнозом Эссенциальная гипертензия I степени «ИЛИ»; 3 – только женщины «И»; 4 – только с диагнозом Эссенциальная гипертензия II степени, то есть условие принимает вид – «пол = 1 & диагноз = 1 | пол = 1 & диагноз = 2».

4. Давайте разберем, как же работать с данным окном. Вспомним поставленную задачу – посчитать, сколько в данном файле курящих мужчин, у которых установлен диагноз Эссенциальная гипертензия I степени. Мы знаем, что информация о курящих пациентах зашифрована в переменной «курение». Находим в списке переменных переменную «курение» и переносим ее в центральное окно при помощи кнопки со стрелкой. Затем нам надо выяснить, под какой цифрой зашифрованы курящие пациенты, для этого щелкаем правой клавишей мыши на переменной «курение» в левой части окна в списке переменных и из вспомогательного меню выбираем пункт «*Variable Information*» (рисунок 17, а).



а)

б)

Рисунок 17 – Опция «Variable Information» (а) и готовое условие отбора (б) команды «Select Cases if»

В открывшемся окне мы видим, что курящие пациенты зашифрованы под номером 1, закрываем это вспомогательное окно (оно нужно нам только для получения информации о зашифрованных данных) и продолжаем записывать условие для отбора в центральном окне. При записи условия отбора, желательно пользоваться экранной клавиатурой. У нас получилось следующее условие отбора «курение = 1». Далее мы знаем, что по условию задачи, нас интересуют не только курящие пациенты. В данной задаче условие отбора состоит из трех частей, каждую из которых необходимо связать между собой, то есть выбрать логический оператор. В итоге должны получить следующее условие: 1 – только курящие «И»; 2 – только мужчины «И»; 3 – только с диагнозом Эссенциальная гипертензия I степени. Давайте запишем его в окне условия для отбора, которое примет следующий вид (рисунок 17, б).

Далее, когда были сформированы все необходимые части условия для отбора, нажимаем кнопки «Continue» и «ОК». После этого в файле выдаче появляется текстовое сообщение о том, что был применен фильтр и указано созданное для данной задачи условие отбора (рисунок 18).

5. И только теперь, после выполнения условия отбора, мы можем приступить к подсчету, то есть воспользуемся уже известной нам командой «*Frequencies*», только переносить в правую часть будем не переменную, а созданный нами фильтр, который находится в самом конце списка переменных (рисунок 19, а) в левой части окна.

6. После нажатия кнопки «ОК», мы получаем таблицу результатов (рисунок 19, б), из которой видно, что пациентов, соответствующих данному условию поставленной задачи, – 9.

```
USE ALL.
COMPUTE filter_$=(курение = 1 & пол = 2 & диагноз = 1).
VARIABLE LABEL filter_$ 'курение = 1 & пол = 2 & диагноз = 1 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
```

Рисунок 18 – Сообщение о применении условия отбора

7. Записываем **ответ**. В данном файле 9 курящих мужчин, у которых установлен диагноз Эссенциальная гипертензия I степени.

**Важно запомнить!** После решения каждой задачи с применением фильтра или условием отбора, его надо отключать, для этого необходимо воспользоваться командой «*Data → Select cases*» и поставить галочку в первую позицию «*All cases*» (т. е. все случаи) → «ОК» (рисунок 20).

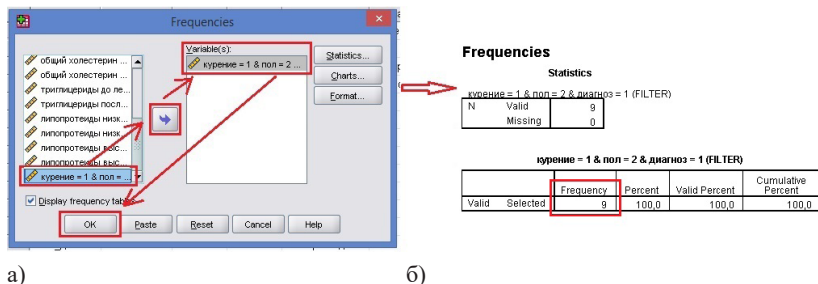


Рисунок 19 – Использование условия отбора (а) и таблицы результатов вычисления (б) команды «Frequencies»

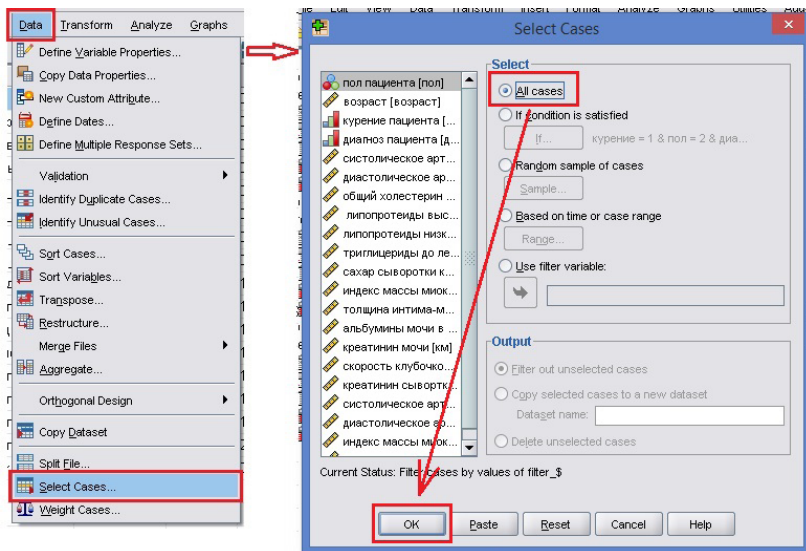


Рисунок 20 – Отключение фильтра в команде «Select Cases»

### 3.5. Задачи для закрепления материала

- Посчитайте, сколько в данном файле курящих и некурящих пациентов.
- Посчитайте, сколько в данном файле пациентов имеют различные диагнозы.
- Посчитайте, сколько в данном файле курящих и некурящих пациентов, которым выставлен различный диагноз.
- Посчитайте, сколько в файле женщин и мужчин имеют различный диагноз.
- Посчитайте, сколько в данном файле курящих женщин.
- Посчитайте, сколько в файле мужчин с диагнозом реноваскулярная гипертония, старше 70 лет.
- Посчитайте, сколько в данном файле курящих мужчин с диагнозом Реноваскулярная и ренопривная гипертония.
- Посчитайте, сколько в данном файле женщин, у которых систолическое давление до лечения выше 180.
- Посчитайте, у скольких пациентов общий холестерин до лечения выше 7,0 и сахар сыворотки крови выше 7,5.

### 3.6. Задание для самостоятельной работы

Для выполнения данного задания необходимо:

1. Поставить и решить по одной задаче на вашем созданном файле на тему «частоты» и «частоты + выборки», сохранить решение в файле выдачи. Сохраненный файл выдачи сдать преподавателю, ответить на его вопросы.

2. Продолжить заполнение листа СРС (форму СРС см. в приложении 2) по следующему образцу:

<b>Файл 2 – задачи «Сколько»</b>
<b>Чего и сколько в данном файле есть?</b>
<b>Задача 1.</b> Посчитать сколько в файле мужчин и женщин?
<b>Ответ.</b> В данном файле имеются 55 женщин (55,0 %) и 45 мужчин (45,0 %).

**Задача 2.** Посчитать, сколько в данном файле курящих мужчин, у которых установлен диагноз Эссенциальная гипертензия I степени.

**Ответ.** В данном файле 9 курящих мужчин, у которых установлен диагноз Эссенциальная гипертензия I степени (курение = 1 & пол = 2 & диагноз = 1)

### 3.7. Тестовые вопросы

**1. Укажите, с помощью каких команд можно дать ответ на вопрос «Кого и сколько в данном файле имеется?»**

- a) frequencies;
- b) crosstabs;
- c) select cases;
- d) correlate;
- e) transform.

**2. Укажите, какие типы переменных можно использовать с командами frequencies и crosstabs:**

- a) ordinal;
- b) nominal;
- c) scale;
- d) columns;
- e) decimals.

**3. Укажите, с помощью каких команд можно отобрать и посчитать данные, соответствующие какому-либо поставленному условию:**

- a) frequencies;
- b) select cases;
- c) crosstabs;
- d) correlate;
- e) transform.

**4. Укажите, с помощью какой команды, можно произвести отбор данных, соответствующих какому-либо поставленному условию:**

- a) select cases;
- b) frequencies;

- c) crosstabs;
- d) correlate;
- e) transform.

**5. Укажите, какие типы переменных можно использовать с командой select cases:**

- a) ordinal;
- b) nominal;
- c) scale;
- d) columns;
- e) decimals.

**6. Определите, какая задача является некорректной, то есть в ее условии смысла нет?**

- a) посчитайте, сколько в данном файле показателей по переменной лейкоциты;
- b) посчитайте, сколько в данном файле некурящих женщин старше 60 лет;
- c) посчитайте, сколько в данном файле мужчин с эссенциальной гипертензией;
- d) посчитайте, сколько в данном файле пациентов с легкой тяжестью заболевания;
- e) посчитайте, сколько в данном файле мужчин и женщин младше 50 лет.

**7. Определите, какая задача является некорректной, то есть в ее условии смысла нет?**

- a) посчитайте, сколько в данном файле показателей по переменной возраст;
- b) посчитайте, сколько в данном файле курящих мужчин старше 60 лет;
- c) посчитайте, сколько в данном файле женщин с реноваскулярной гипертонией;
- d) посчитайте, сколько в данном файле пациентов младше 50 лет;
- e) посчитайте, сколько в данном файле женщин с тяжелой тяжестью заболевания.

**8. Определите, какая задача является некорректной, то есть в ее условии смысла нет?**

- а) посчитайте, сколько в данном файле показателей по переменной «вес»;
- б) посчитайте, сколько в данном файле курящих и некурящих пациентов;
- в) посчитайте, сколько в данном файле женщин и мужчин с гипертензией;
- г) посчитайте, сколько в данном файле курящих младше 40 лет;
- д) посчитайте, сколько в данном файле девочек, занимающихся спортом.

**9. Определите, какая задача является некорректной, то есть в ее условии смысла нет?**

- а) посчитайте, сколько в данном файле показателей по переменной «рост»;
- б) посчитайте, сколько в данном файле мальчиков с хобби «спорт»;
- в) посчитайте, сколько в данном файле учеников в 11 Б классе;
- г) посчитайте, сколько в данном файле некурящих старше 60 лет;
- д) посчитайте, сколько в данном файле учеников, выбравших технический вуз.

**10. Определите, какая задача является некорректной, то есть в ее условии смысла нет?**

- а) посчитайте, сколько в данном файле показателей по переменной САД;
- б) посчитайте, сколько в данном файле учеников с хобби «компьютер»;
- в) посчитайте, сколько в данном файле девочек учащихся в 11 А классе;
- г) посчитайте, сколько в данном файле пациентов младше 75 лет;
- д) посчитайте, сколько в данном файле девочек, выбравших экономический вуз.

**11. Определите, какая задача является некорректной, то есть в ее условии смысла нет?**

- а) посчитайте, сколько в данном файле показателей по переменной ДАД;
- б) посчитайте, сколько в данном файле девочек с хобби «искусство»;
- с) посчитайте, сколько в данном файле учеников учащихся в 11 А и В классах;
- д) посчитайте, сколько в данном файле мальчиков и девочек в 11 В классе;
- е) посчитайте, сколько в данном файле девочек, выбравших гуманитарный вуз.

**12. В файле имеются три параметра: «возраст»; «пол» (1 – женщина; 2 – мужчина); стадия заболевания (1 – лёгкая; 2 – средняя; 3 – тяжёлая). Задача – отберите женщин с тяжёлой стадией заболевания.**

- а)  $\text{пол} = 1 \ \& \ \text{стадия заболевания} = 3$ ;
- б)  $\text{пол} = 2 \ \& \ \text{стадия заболевания} = 1$ ;
- с)  $\text{пол} = 1 \ \& \ \text{стадия заболевания} = 2$ ;
- д)  $\text{пол} = 1 \ \& \ \text{стадия заболевания} = 1$ ;
- е)  $\text{пол} = 2 \ \& \ \text{стадия заболевания} = 3$ .

**13. В файле имеются три параметра: возраст; пол (1 – женщина; 2 – мужчина); стадия заболевания (1 – лёгкая; 2 – средняя; 3 – тяжёлая). Задача – отберите мужчин в возрастной категории от 40 до 60 лет.**

- а)  $\text{пол} = 2 \ \& \ \text{возраст} \geq 40 \ \& \ \text{возраст} \leq 60$ ;
- б)  $\text{пол} = 2 \ \& \ \text{возраст} \leq 40 \ \& \ \text{возраст} \geq 60$ ;
- с)  $\text{пол} = 1 \ \& \ \text{возраст} \leq 40 \ \& \ \text{возраст} \geq 60$ ;
- д)  $\text{пол} = 1 \ \& \ \text{возраст} \geq 40 \ \& \ \text{возраст} \leq 60$ ;
- е)  $\text{пол} = 1 \ \& \ \text{возраст} \geq 40 \mid \text{возраст} \leq 60$ .

**14. В файле имеются три параметра: возраст; пол (1 – женщина; 2 – мужчина); стадия заболевания (1 – лёгкая; 2 – средняя; 3 – тяжёлая). Задача – отберите женщин со средней или тяжёлой стадией заболевания.**

а) пол = 1 & стадия заболевания = 2 | пол = 1 & стадия заболевания = 3;

б) пол = 1 | стадия заболевания = 2 & пол = 1 | стадия заболевания = 3;

с) пол = 1 & стадия заболевания = 2 | стадия заболевания = 3;

д) пол = 1 | стадия заболевания = 2 | стадия заболевания = 3;

е) пол = 1 & стадия заболевания = 2 & стадия заболевания = 3.

**15. В файле имеются два параметра: пол (1 – женщина; 2 – мужчина); ИММЛЖ (индекс массы миокарда левого желудочка). Гипертрофия левого желудочка сердца диагностируется у женщин, если ИММЛЖ больше 150 г, а у мужчин, если ИММЛЖ больше 180 г. Задача – отберите всех людей с гипертрофией левого желудочка.**

а) пол = 1 & ИММЛЖ > 150 | пол = 2 & ИММЛЖ > 180;

б) пол = 1 | ИММЛЖ > 150 | пол = 2 | ИММЛЖ > 180;

с) пол = 1 | пол = 2 | ИММЛЖ > 150 & ИММЛЖ > 180;

д) пол = 1 | ИММЛЖ > 150 & пол = 2 | ИММЛЖ > 180;

е) пол = 1 & ИММЛЖ > 180 & пол = 2 & ИММЛЖ > 150.

**16. В файле имеются три параметра: возраст; пол (1 – девочка; 2 – мальчик); хобби (1 – спорт; 2 – компьютер; 3 – искусство). Задача – отберите девочек, старше 13 лет, выбравших хобби «искусство».**

а) пол = 1 & возраст > 13 & хобби = 3;

б) пол = 1 & возраст < 13 & хобби = 1;

с) пол = 2 & возраст = 13 & хобби = 2;

д) пол = 2 & возраст > 13 & хобби = 2;

е) пол = 1 & возраст = 13 & хобби = 3.

**17. В файле имеются три параметра: возраст; пол (1 – девочка; 2 – мальчик); хобби (1 – спорт; 2 – компьютер;**

**3 – искусство). Задача – отберите мальчиков, выбравших хобби «спорт».**

- a)  $\text{пол} = 2 \ \& \ \text{хобби} = 1$ ;
- b)  $\text{пол} \ \& \ \text{хобби} = 2$ ;
- c)  $\text{пол} = 1 \ \& \ \text{хобби} = 1$ ;
- d)  $\text{пол} = 2 \ \& \ \text{хобби}$  ;
- e)  $\text{пол} = 1 \mid \text{хобби} = 3$ .

**18. В файле имеются три параметра: класс (1 – 11 А; 2 – 11 Б; 3 – 11 В); пол (1 – девочка; 2 – мальчик); хобби (1 – спорт; 2 – компьютер; 3 – искусство). Задача – отберите мальчиков, учащихся в 11 А и 11 В классах.**

- a)  $\text{пол} = 2 \ \& \ \text{класс} = 1 \mid \text{пол} = 2 \ \& \ \text{класс} = 3$ ;
- b)  $\text{пол} = 2 \mid \text{класс} = 1 \ \& \ \text{пол} = 2 \mid \text{класс} = 3$ ;
- c)  $\text{пол} = 2 \ \& \ \text{класс} = 2 \mid \text{класс} = 3$ ;
- d)  $\text{пол} = 2 \mid \text{класс} = 2 \mid \text{класс} = 3$ ;
- e)  $\text{пол} = 2 \ \& \ \text{класс} = 2 \ \& \ \text{класс} = 3$ .

**19. В файле имеются три параметра: класс (1 – 11 А; 2 – 11 Б; 3 – 11 В); пол (1 – девочка; 2 – мальчик); хобби (1 – спорт; 2 – компьютер; 3 – искусство). Задача – отберите девочек, учащихся в 11 Б классе и увлекающихся искусством.**

- a)  $\text{пол} = 1 \ \& \ \text{класс} = 2 \ \& \ \text{хобби} = 3$ ;
- b)  $\text{пол} = 1 \mid \text{класс} = 2 \mid \text{хобби} = 3$ ;
- c)  $\text{пол} = 2 \ \& \ \text{класс} = 1 \mid \text{хобби} = 3$ ;
- d)  $\text{пол} = 2 \mid \text{класс} = 2 \mid \text{хобби} = 3$ ;
- e)  $\text{пол} = 2 \ \& \ \text{класс} = 2 \ \& \ \text{хобби} = 2$ .

## **Глава 4. ОПИСАТЕЛЬНЫЕ СТАТИСТИКИ**

### **ЧАСТЬ 2**

- 4.1. Что такое описательные статистики?**
- 4.2. Виды статистических величин.**
- 4.3. Закон распределения случайной величины.**
- 4.4. Описательные статистики, нормальное распределение.**
- 4.5. Описательные статистики, распределение, отличное от нормального.**
- 4.6. Задачи для закрепления материала.**
- 4.7. Задания для самостоятельной работы.**
- 4.8. Тестовые вопросы.**

#### **4.1. Что такое описательные статистики?**

Те, кто впервые в своей работе сталкивается с обработкой и описанием данных, не всегда четко представляют, в какой форме их корректно отображать и обрабатывать для того, чтобы в дальнейшем подвергнуть статистическому выводу.

Исследование и обобщение полученных данных является первым и одним из наиболее важных этапов статистического анализа. Его целью является получение адекватного описания исследуемых переменных для лучшего понимания их особенностей. Этот этап играет ключевую роль, поскольку выбранная модель определяет дальнейшие подходы к тестированию статистических гипотез. Таким образом, основными задачами описательной статистики являются исследование данных, определение типа их распределения, выявление ошибок, а также нетипичных значений (выбросов). Описательная статистика позволяет определить долю пропущенных значений для каждой переменной и помогает оценить риск систематической ошибки [12].

В целом адекватное и понятное представление переменных обычно свидетельствует о качественном подходе к сбору, анализу и интерпретации данных.

Вот что входит в описательную статистику в качестве основных показателей [13]:

- экстремумы, или максимумы и минимумы значений самой переменной;
- среднее арифметическое (*Mean*) – сумма всех значений, отнесенная к общему числу наблюдений;
- стандартная ошибка среднего (*S.E.Mean*) – определяется как стандартное отклонение, деленное на квадратный корень из объема выборки.
- стандартное отклонение (*Standard Deviation*) – равно квадратному корню из дисперсии.
- мода (*mode*) представляет собой наиболее часто встречающееся значение в выборке;
- медиана (*Median* – аналог среднего) – это значение, делящее распределение пополам. Другими словами, это значение, ниже которого находятся 50 % значений, и выше также 50 % всех значений в распределении;
- дисперсия (*Dispersion*) – равна сумме квадратов отклонений каждого значения от среднего, деленной на N, где N – число значений в распределении.

Помимо этого, в понятие описательные статистики входят еще такие показатели, как квантили.

Квантиль – общее понятие, а проценти́ли, деци́ли и кварта́ли – три его примера.

Кварти́ли (*quartiles*) – это 3 точки – значения признака (P25, P50, P75), которые делят упорядоченное (по возрастанию) множество наблюдений на 4 равные по численности части. Первый квартиль соответствует 25-му процентилю, второй – 50-му процентилю, или медиане, третий квартиль соответствует 75-му процентилю.

Каждый из них играет существенную роль для корректного отображения получаемых данных [13].

Популяционное значение параметра (среднее значение, медиану, долю и т. д.) получить невозможно (исключение составляют случаи, когда исследование проводится на группе, которая включает всех членов популяции). Однако популяционное значение параметра можно оценить по выборке. Точность такой оценки зависит от метода измерения (ошибки измерения), объема и репрезентативности выборки (ошибка выборки) и биологической вариации [14].

Показатели описательной статистики можно разбить на несколько групп:

- показатели положения, описывающие положение экспериментальных данных на числовой оси;
- показатели разброса, описывающие степень разброса данных относительно центральной тенденции;
- показатели асимметрии;
- графические представления результатов [14].

## 4.2. Виды статистических величин

При проведении медико-социального исследования статистическая сводка материала завершается получением ряда таблиц, числа в которых отражают абсолютные размеры изучаемого явления и составляющих его частей (**абсолютные величины** – количественное выражение любого явления).

Абсолютные статистические показатели характеризуются определенной размерностью – единицей измерения. Примером абсолютных показателей являются данные о численности населения, числе работающих врачей, числе функционирующих амбулаторно-поликлинических и стационарных учреждений. Так, например, абсолютная численность населения в Китае, Индии показывает, что эти две страны являются странами с наибольшим числом населения в мире.

Переход к анализу статистического материала включает его обработку, то есть вычисление на основе этих абсолютных величин системы обобщающих показателей [9, 14].

Однако неправильно было бы рассматривать абсолютные величины как промежуточный этап и лишать их самостоятельного значения и ценности.

Абсолютные величины имеют в медицинской статистике определенное значение.

В большинстве случаев они интересны сами по себе, характеризуя, например, численность населения стран и населенных пунктов, число врачей, больничных коек, количество рождений и смертей, редкие случаи некоторых заболеваний (малярия, полиомиелит, столбняк, сыпной тиф) или наоборот массовое распространение других (туберкулез, ВИЧ инфекция) [9, 14].

Таким образом абсолютные цифры применяются:

- для характеристики абсолютных размеров явления в целом (показывают массовость явления);
- для характеристики редко встречающихся явлений (показывают единичность явлений).

Сравнивая размеры двух явлений (рождаемость, смертность, заболеваемость и т. п.), либо изучая изменения этих явлений во времени, необходимо абсолютные числа, выражающие эти размеры преобразовать в **производные величины** – относительные и средние [9, 14].

**Относительные величины** представляют собой соотношение двух абсолютных величин в виде кратных величин или в виде процентов, получающихся при сравнении двух значений одного признака или сравнении части с целым. Эталоном для сравнения служат базисные данные, принимаемые за 100, 1000, 10 000, 100 000, то есть за единицу с нулями [9, 14].

По своему содержанию относительные величины (показатели, коэффициенты) делятся на 4 основные группы (вида):

- экстенсивные;
- интенсивные;
- наглядности;
- соотношения.

К особым видам относительных статистических показателей относятся:

- относительные показатели, характеризующие динамику процесса, изменение во времени (темп роста);
- относительные показатели – отношения фактически наблюдаемых величин признака к его нормативным, плановым, оптимальным или максимально возможным величинам;
- относительные показатели, характеризующие взаимосвязи между разными признаками объекта, объектом и средой и т. п. К ним принадлежат коэффициенты регрессии, эластичности, детерминации, корреляции, а также аналитические индексы [9, 14].

**Экстенсивные показатели**, или показатели структуры (распределения), характеризуют распределение явлений (структуру объекта) внутри одной совокупности (целого) на составляющие его части по их удельному весу или отношение части к целому.

Обычно экстенсивные показатели выражаются в процентах (%) к итогу.

Примеры экстенсивного показателя: лейкоцитарная формула (доля отдельных форменных элементов во всей массе лейкоцитов), структура населения по возрасту, полу, структура заболеваний и причин смерти и т. д. [9, 14].

Экстенсивные показатели (коэффициенты) нужны для определения структуры статистической совокупности и сравнительной оценки соотношения составляющих ее частей и применяются при анализе:

- структуры заболеваемости;
- причин смерти;
- структуры коечного фонда;
- структуры финансовых расходов на здравоохранение и т. д.

**Интенсивные показатели** (показатели частоты, распространенности) характеризуют частоту явления в той среде, где мы это явление наблюдаем. Эти показатели указывают о том, как часто встречается данное явление в той или иной среде.

Для вычисления показателя интенсивности недостаточно знать лишь величину интересующего нас явления, необходимо

знать еще величину той среды, в которой данное явление наблюдается.

В медико-социальных исследованиях за «среду», как правило, принимается численность населения в целом или отдельных его групп (по возрасту, полу, профессии, месту жительства и т. д.) При этом численность населения берут на середину изучаемого периода времени (чаще всего года), либо, что правильнее, рассчитывают среднюю численность населения (полусумму численности населения на начало и конец периода наблюдения) [9, 14].

### **4.3. Нормальное распределение случайной величины**

Нормальное распределение, также называемое распределением Гаусса, – распределение вероятностей, играет одну из важнейших ролей в применении и выборе статистических методов.

Каждому биологическому и психологическому свойству соответствует свое распределение в генеральной совокупности [9, 10]. Чаще всего оно нормальное и характеризуется своими параметрами – средним и сигмой. Графически нормальное распределение можно изобразить кривой представленной на рисунке 21. Так, среднее значение задает положение кривой на числовой оси, а сигма – задает ширину этой кривой и выступает как масштаб измерения.

Наиболее важным свойством кривых нормального распределения является одинаковая доля площади под кривой между одними и теми же значениями признака, выраженными в единицах стандартного отклонения [9, 12, 14].

Все многообразие нормальных распределений может быть сведено к одной кривой, если применить  $z$ -преобразование ( $z$ -преобразование состоит в том, что от каждого измерения отнимается среднее и полученный результат делится на сигму). Тогда каждое свойство будет иметь среднее 0 и сигму 1 – это называется единичным нормальным распределением, которое используется как эталон (см. рисунок 21).

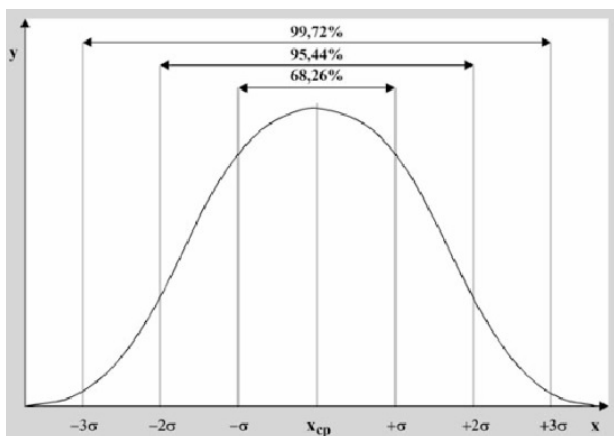


Рисунок 21 – Графическое представление нормального распределения

Формула, описывающая нормальный закон распределения случайной величины, имеет следующий вид:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2},$$

где

$\mu$  – генеральное среднее арифметическое;

$\sigma$  – генеральное стандартное отклонение;

$e$  – основание натуральных логарифмов, приблизительно равное 2,719;

$\pi$  – число, приблизительно равное 3,142;

$x_i$  – конкретное значение признака.

Площадь под кривой интерпретируется как вероятность или относительная частота.

Полезно знать, что если распределение является нормальным, то:

- 90 % всех случаев располагается в диапазоне значений  $M$  (среднее)  $\pm 1,64 \sigma$  (сигма);

- 95 % всех случаев располагается в диапазоне значений  $M (\text{среднее}) \pm 1,96 \sigma$  (сигма);
- 99 % всех случаев располагается в диапазоне значений  $M (\text{среднее}) \pm 2,58 \sigma$  (сигма).

Соответствие или несоответствие нормальности распределения влияет на выбор статистических методов обработки данных. Так, данные, соответствующие нормальному распределению, обрабатываются с помощью параметрических методов, а те данные, которые не соответствуют нормальности распределения, обрабатываются с помощью непараметрических методов статистики. К тому же, в каждом последующем сложном методе обработки данных существуют условия использования того или иного метода, в которые часто входит и нормальность распределения [14].

Может сложиться впечатление, что если величина случайная, то и предсказать результат испытания (опыта) невозможно. Если говорить о точном значении результата опыта, то это действительно так.

Но все-таки у нас есть понятие, которое позволяет очень плодотворно работать со случайными величинами. Это понятие – **вероятность!**

Напомним только, что вероятность обычно обозначается буквой  $P$  и имеет значения  $0 < P < 1$ .  $P = 0$  – невероятное событие,  $P = 1$  – событие, которое обязательно произойдет.

Таким образом, если мы знаем Закон распределения случайной величины, то мы знаем о данной случайной величине все!

При помощи теста Колмогорова – Смирнова можно проверить, соответствует ли реальное распределение переменной нормальному (Гаусса), равномерному, экспоненциальному распределению или распределению Пуассона. Разумеется, самым распространённым видом проверки является проверка наличия нормального распределения [9, 14].

#### 4.4. Описательные статистики, нормальное распределение

Для того чтобы начать рассчитывать описательные статистики, соответствующие нормальному распределению случайной величины, необходимо проделать следующие действия:

1. Открыть файл «Иванов\_ГБ.sav».
2. Поставить соответствующую **задачу**, например – Посчитать, описательные статистики по переменной «*систолическое давление до лечения*» у женщин с диагнозом Эссенциальная гипертензия I степени.

3. Для решения поставленной задачи вспомним простое, но очень важное правило – сначала отбор, потом счет? но есть сначала необходимо отобрать женщин с диагнозом Эссенциальная гипертензия, а только потом посчитать описательные статистики по переменной «*систолическое давление до лечения*».

4. Применим фильтр, то есть воспользуемся командой «*Select Cases*», в результате применения этой команды получим следующую надпись – «*COMPUTE filter\_\$=(nol = 1 & диагноз = 1)*».

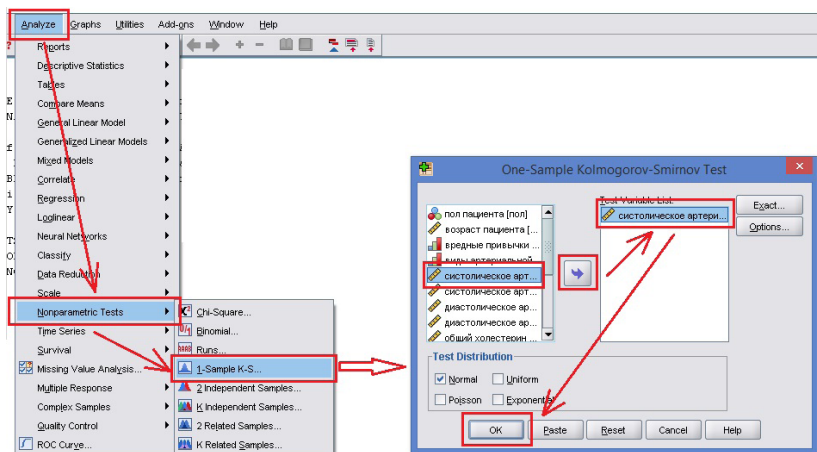
5. Далее нам необходимо проверить переменную «систолическое давление до лечения» на распределение по закону Гаусса, то есть применим тест Колмогорова – Смирнова – «*1-Sample K-S*». Для этого заходим в меню:

<i>Analyze → Nonparametric Test → 1-Sample K-S</i>
----------------------------------------------------

появится диалоговое окно «1-Sample K-S» (рисунок 22, а). Переносим в правую часть переменную «систолическое давление до лечения». Нажать кнопку «ОК» (рисунок 22, б).

Также необходимо помнить, что на тест Колмогорова – Смирнова проверяют только переменные «Scale».

6. После применения данного теста, мы видим, что  $Sig = 0,271$  (рисунок 23). Давайте проанализируем данный результат, то есть сравним с 0,05. В нашем случае  $Sig > 0,05$ , это говорит о том,



а)

б)

Рисунок 22 – Алгоритм (а) и рабочее окно (б) команды «1-Sample K-S»

## NPar Tests

One-Sample Kolmogorov-Smirnov Test		
		систолическое артериальное давление до лечения
N		14
Normal Parameters <sup>a</sup>	Mean	155,36
	Std. Deviation	10,463
Most Extreme Differences	Absolute	,267
	Positive	,267
	Negative	-,161
Kolmogorov-Smirnov Z		,999
Asymp. Sig. (2-tailed)		,271

a. Test distribution is Normal.

Рисунок 23 – Таблица результатов вычисления команды «1-Sample K-S»

переменная «систолическое давление до лечения» подчиняется закону Гаусса, то есть ее значения распределены по нормальному закону.

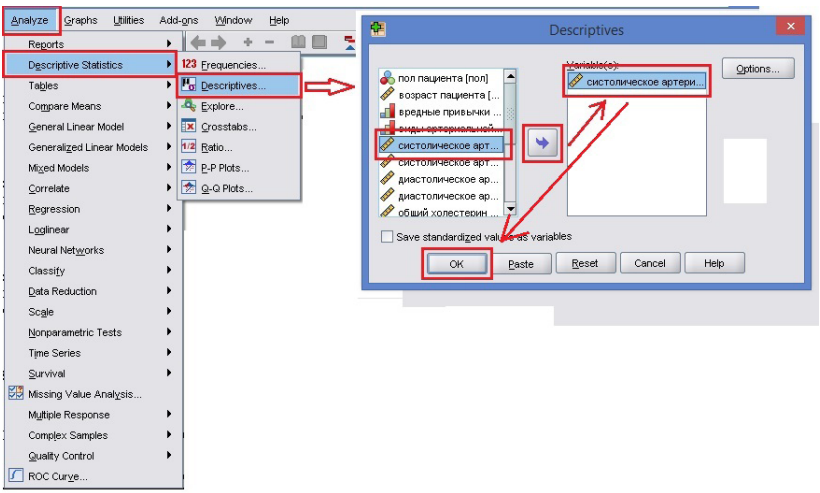
7. Раз переменная распределена по нормальному закону, значит, нам необходимо найти следующие описательные статистики: минимальное и максимальные значения, среднее и отклонение

от среднего. Для этого воспользуемся командой «*Descriptives*», то есть заходим в меню:

*Analyze* → *Descriptive Statistics* → *Descriptives*

появится диалоговое окно «*Descriptives*» (рисунок 24, а). Перенесим в правую часть ту же самую переменную, которую проверяли на закон Гаусса, то есть переменную «систолическое давление до лечения», нажимаем кнопку «ОК» (рисунок 24, б).

8. Результаты наших вычислений появились в виде таблицы (рисунок 25).



а) б)  
Рисунок 24 – Алгоритм (а) и рабочее окно (б) команды «*Descriptives*»

**Descriptives**

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
систолическое артериальное давление до лечения	14	140	180	155,36	10,463
Valid N (listwise)	14				

Рисунок 25 – Таблица результатов вычисления команды «*Descriptives*»

Давайте прочитаем ее. Мы видим, что в нашем случае описательные статистики подсчитаны по значениям 14 пациентов, отвечающих условию отбора; минимальное значение «*Minimum*» систолического давления до лечения – 140, максимальное «*Maximum*» – 180; среднее значение «*Mean Statistic*» – 155,36 и стандартное отклонение «*Std. Deviation*» – 10,463 (см. рисунок 25). Среднее значение и стандартное отклонение имеют свое устойчивое написание  $M \pm \sigma = 155,36 \pm 10,463$ .

9. Мы разобрали полученную таблицу, осталось написать ответ:  $155,36 \pm 10,463$ . В данном файле у женщин с диагнозом Эссенциальная гипертензия I степени по переменной «систолическое давление до лечения» были рассчитаны среднее арифметическое и стандартное отклонение, так как распределение показателя, по критерию Колмогорова – Смирнова, соответствует закону Гаусса, поскольку  $\text{Sig} = 0,271 > 0,05$ .

Не забудьте, что после решения каждой задачи надо отключить фильтр, если вы забыли, как это делать, посмотрите параграф 3.4.

Надеемся, вам было понятно, как проводить расчет описательных статистик для переменных, значения которых распределены по нормальному закону.

#### **4.5. Описательные статистики, распределение, отличное от нормального**

Для того чтобы начать рассчитывать описательные статистики, распределение которых отлично от нормального, необходимо проделать следующие действия:

1. Открыть файл «Иванов\_ГБ.sav».
2. Поставить соответствующую задачу, например – Посчитать, описательные статистики по переменной «креатинин мочи» у курящих пациентов старше 60 лет.
3. Для решения поставленной задачи вспомним простое, но очень важное правило – сначала отбор, потом счет, то есть сначала необходимо отобрать курящих пациентов старше 60 лет,

а только потом посчитать описательные статистики по переменной «креатинин мочи».

4. Создадим условие для отбора, при помощи команды «Select Cases», в результате применения фильтра получим следующую надпись – «COMPUTE filter\_\$(курение = 1 & возраст > 60)».

5. Далее нам необходимо проверить переменную «креатинин мочи» на распределение по закону Гаусса, то есть применим тест Колмогорова – Смирнова – «1-Sample K-S», то есть заходим в меню

*Analyze → Nonparametric Test → 1-Sample K-S*

Появится диалоговое «1-Sample K-S» (рисунок 26, а). Переносим в правую часть переменную «креатинин мочи», то есть нужно нажать стрелочку между двумя частями окна. Нажать кнопку «OK» (рисунок 26, б).

Также необходимо помнить, что на тест Колмогорова – Смирнова проверяют только переменные «Scale».

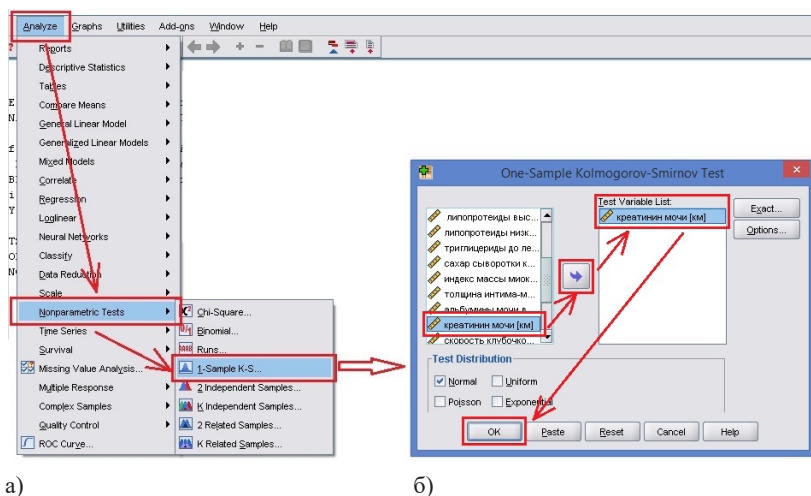


Рисунок 26 – Алгоритм (а) и рабочее окно (б) команды «1-Sample K-S»

6. После применения данного теста, мы видим, что  $Sig = 0,026$  (рисунок 27). Давайте проанализируем данный результат, то есть сравним с  $0,05$ . В нашем случае  $Sig < 0,05$ , это говорит о том, что переменная «креатинин мочи» не подчиняется закону Гаусса, то есть ее значения распределены по закону отличному от нормального.

## NPar Tests

One-Sample Kolmogorov-Smirnov Test		
		креатинин мочи
N		22
Normal Parameters <sup>a</sup>	Mean	4,932
	Std. Deviation	3,2201
Most Extreme Differences	Absolute	,314
	Positive	,314
	Negative	-,190
Kolmogorov-Smirnov Z		1,472
Asymp. Sig. (2-tailed)		,026

a. Test distribution is Normal.

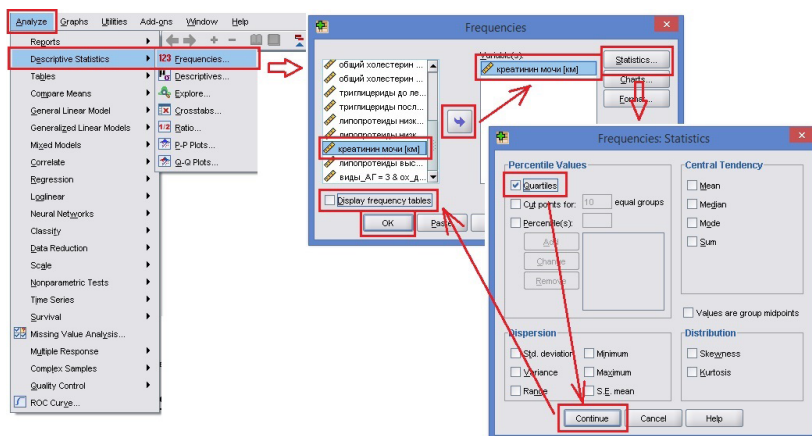
Рисунок 27 – Таблица результатов вычисления команды «1-Sample K-S»

7. Раз переменная распределена по закону, отличному от нормального, значит, нам необходимо найти следующие описательные статистики: квартили и медиану. Для этого воспользуемся командой «Frequencies», то есть заходим в меню

**Analyze → Descriptive Statistics → Frequencies**

Появится диалоговое окно «Frequencies» (рисунок 28, а). Переносим в правую часть ту же самую переменную, которую проверяли на закон Гаусса, то есть переменную «креатинин мочи», нажимаем кнопку «Statistics», галочку в позицию «Quartiles», нажимаем кнопки «Continue», убираем галочку с позиции «Display frequency tables» → «OK» (рисунок 28, б).

8. Результаты наших вычислений появились в виде таблицы (рисунок 29).



а)

б)

Рисунок 28 – Алгоритм (а) и рабочее окно (б) команды «Frequencies»

## Frequencies

Statistics		
кратинин мочи		
N	Valid	22
	Missing	0
Percentiles	25	3,075
	50	4,200
	75	4,975

Рисунок 29 – Таблица результатов вычисления команды «Frequencies»

Давайте прочитаем ее. Мы видим, что в нашем случае описательные статистики подсчитаны по значениям 22-х пациентов, отвечающих условию отбора. Далее отображены первый, второй и третий квартили. Первый квартиль – это точка на шкале измеренных значений (3,075), ниже (левее) которой располагаются 25 % измеренных значений. Второй квартиль (4,200) – это точка, ниже которой располагаются 50 % измеренных значений. Второй квартиль также называется медианой. Третий квартиль (4,975) – это точка на шкале измеренных значений, ниже которой

располагаются 75 % значений. Они также имеют свое стандартное написание 4,200 (3,075; 4,975).

9. Мы разобрали полученную таблицу, осталось написать ответ: 4,200 (3,075; 4,975). Для переменной «креатинин мочи» у курящих пациентов старше 60 лет были рассчитаны медиана и квартильный размах, так как распределение показателя, по критерию Колмогорова – Смирнова, не соответствует нормальному распределению Гаусса, поскольку  $\text{Sig} = 0,026 < 0,05$ .

Не забудьте, что после решения каждой задачи надо отключить фильтр, если вы забыли, как это делать, посмотрите параграф 3.4.

Надеемся, вам было понятно, как проводить расчет описательных статистик для переменных, значения которых распределены по закону отличному от нормального.

#### **4.6. Задачи для закрепления материала**

- Посчитайте описательные статистики по переменной «сахар сыворотки крови» до лечения.
- Посчитайте описательные статистики по переменной «систолическое давление» до лечения у курящих женщин старше 30 лет.
- Посчитайте описательные статистики по переменной «креатинин сыворотки крови» у курящих пациентов.
- Посчитайте описательные статистики по переменной «возраст» у мужчин с систолическим давлением до лечения выше 160.
- Посчитайте описательные статистики по переменной «общий холестерин» у некурящих с липопротеидами высокой плотности до лечения ниже 1.
- Посчитайте описательные статистики по переменной «систолическое давление» до лечения у пациентов младше 50 лет с диагнозом «Реноваскулярная гипертензия».

- Посчитайте описательные статистики по переменной «триглицериды» до лечения у женщин с индексом массы миокарда левого желудочка до лечения выше 150.
- Посчитайте описательные статистики по переменной «креатинин мочи» у женщин старше 50 лет с общим холестерином до лечения ниже 5,1.
- Посчитайте описательные статистики по переменной «систолическое давление» после лечения.
- Посчитайте описательные статистики по переменной «альбумины мочи в сутки» у курящих с диагнозом эссенциальная гипертензия I степени.

#### 4.7. Задание для самостоятельной работы

Для выполнения данного задания необходимо:

1. Поставить и решить по одной задаче на вашем созданном файле на тему «Описательные статистики» для переменных, подчиняющихся и не подчиняющихся закону Гаусса, сохранить решение в файле выдачи. Сохраненный файл-выдачи сдать преподавателю, ответить на его вопросы.

2. Продолжить заполнение листа СРС (форму СРС см. в приложении 2) по следующему образцу:

<b>Файл 2</b> – задачи «Описательные статистики»
<b>Описательные статистики 1</b>
<b>Задача.</b> Посчитать, описательные статистики по переменной «систолическое давление до лечения» у женщин с диагнозом Эссенциальная гипертензия I степени
<b>Ответ.</b> $155,36 \pm 10,463$ . В данном файле у женщин с диагнозом Эссенциальная гипертензия I степени по переменной «систолическое давление» до лечения были рассчитаны среднее арифметическое и стандартное отклонение, так как распределение показателя по критерию Колмогорова – Смирнова соответствует закону Гаусса, поскольку $\text{Sig} = 0,271 > 0,05$

Описательные статистики 2
<b>Задача.</b> Посчитать, описательные статистики по переменной «креатинин мочи» у курящих пациентов старше 60 лет
<b>Ответ.</b> 4,200 (3,075; 4,975). Для переменной «креатинин мочи» у курящих пациентов старше 60 лет были рассчитаны медиана и квартильный размах, так как распределение показателя по критерию Колмогорова – Смирнова не соответствует нормальному распределению Гаусса, поскольку $\text{Sig} = 0,026 < 0,05$

## 4.8. Тестовые вопросы

**1. Определите, как называются различные вычисляемые показатели, характеризующие распределение значений переменной?**

- a) описательные статистики;
- b) описательные таблицы;
- c) сравнительные статистики;
- d) корреляционные статистики;
- e) регрессионные статистики.

**2. Определите, как называется сумма всех значений в распределении, деленная на их количество?**

- a) среднее;
- b) медиана;
- c) минимум;
- d) максимум;
- e) дисперсия.

**3. Определите, как называются величины, наиболее часто встречающиеся в публикациях значений выборки?**

- a) медиана;
- b) среднее;
- c) минимум;
- d) максимум;
- e) дисперсия.

**4. Определите, как называется сумма квадратов отклонений значений от среднего арифметического?**

- a) дисперсия;
- b) медиана;
- c) среднее;
- d) минимум;
- e) максимум.

**5. Определите, как называется переменная, которая в результате испытания в зависимости от случая принимает одно из возможного множества своих значений?**

- a) случайная величина;
- b) качественная величина;
- c) количественная величина;
- d) относительная величина;
- e) абсолютная величина.

**6. Определите, как называется всякое соотношение, устанавливающее связь между возможными значениями случайной величины и соответствующими им вероятностями?**

- a) закон распределения;
- b) ряд распределения;
- c) таблица распределения;
- d) полигон распределения;
- e) многоугольник распределения.

**7. Определите, какая задача является некорректной, то есть в ее решении смысла нет:**

- a) найти среднее арифметическое и стандартное отклонение (или квартили и медиану) по показателю «диагноз»;
- b) найти среднее арифметическое и стандартное отклонение (или квартили и медиану) по показателю «возраст»;
- c) найти среднее арифметическое и стандартное отклонение (или квартили и медиану) по показателю «сахар крови»;
- d) найти среднее арифметическое и стандартное отклонение (или квартили и медиану) по показателю «рост»;
- e) найти среднее арифметическое и стандартное отклонение (или квартили и медиану) по показателю «сахар мочи».

**8. Определите, какая задача является некорректной, то есть в ее решении смысла нет:**

- а) найти среднее арифметическое и стандартное отклонение (или квартили и медиану) по показателю «пол»;
- б) найти среднее арифметическое и стандартное отклонение (или квартили и медиану) по показателю «вес»;
- с) найти среднее арифметическое и стандартное отклонение (или квартили и медиану) по показателю «ИММЛЖ»;
- д) найти среднее арифметическое и стандартное отклонение (или квартили и медиану) по показателю «Hb»;
- е) найти среднее арифметическое и стандартное отклонение (или квартили и медиану) по показателю «креатинин».

**9. Определите, какая задача является некорректной, то есть в ее решении смысла нет:**

- а) найти среднее арифметическое и стандартное отклонение (или квартили и медиану) по показателю «курение»;
- б) найти среднее арифметическое и стандартное отклонение (или квартили и медиану) по показателю «лейкоциты»;
- с) найти среднее арифметическое и стандартное отклонение (или квартили и медиану) по показателю «ИМТ»;
- д) найти среднее арифметическое и стандартное отклонение (или квартили и медиану) по показателю «САД»;
- е) найти среднее арифметическое и стандартное отклонение (или квартили и медиану) по показателю «моноциты».

**10. Определите, какая задача является некорректной, то есть в ее решении смысла нет:**

- а) найти среднее арифметическое и стандартное отклонение (или квартили и медиану) по показателю «тяжесть заболевания»;
- б) найти среднее арифметическое и стандартное отклонение (или квартили и медиану) по показателю «СОЭ»;
- с) найти среднее арифметическое и стандартное отклонение (или квартили и медиану) по показателю «тромбоциты»;
- д) найти среднее арифметическое и стандартное отклонение (или квартили и медиану) по показателю «рост»;
- е) найти среднее арифметическое и стандартное отклонение (или квартили и медиану) по показателю «триглицериды».

**11. Определите, какая задача является некорректной, то есть в ее решении смысла нет:**

- a) найти среднее арифметическое и стандартное отклонение (или квартили и медиану) по показателю «хобби»;
- b) найти среднее арифметическое и стандартное отклонение (или квартили и медиану) по показателю «успеваемость»;
- c) найти среднее арифметическое и стандартное отклонение (или квартили и медиану) по показателю «счет в уме»;
- d) найти среднее арифметическое и стандартное отклонение (или квартили и медиану) по показателю «осведомленность»;
- e) найти среднее арифметическое и стандартное отклонение (или квартили и медиану) по показателю «время помощи».

**12. Определите, какая задача является некорректной, то есть в ее решении смысла нет:**

- a) найти среднее арифметическое и стандартное отклонение (или квартили и медиану) по показателю «вуз»;
- b) найти среднее арифметическое и стандартное отклонение (или квартили и медиану) по показателю «числовой ряд»;
- c) найти среднее арифметическое и стандартное отклонение (или квартили и медиану) по показателю «словарь»;
- d) найти среднее арифметическое и стандартное отклонение (или квартили и медиану) по показателю «память»;
- e) найти среднее арифметическое и стандартное отклонение (или квартили и медиану) по показателю «счет в уме».

**13. Выберите, каким тестом необходимо воспользоваться, чтобы проверить случайную величину на закон распределения:**

- a) 1-Sample K-S;
- b) Independent-Samples T Test;
- c) Paired-Samples T Test;
- d) 2 Independent Samples;
- e) 2 Related Samples.

**14. Выберите, какой командой необходимо воспользоваться, чтобы посчитать описательные статистики для переменной с  $\text{sig} < 0,05$  в тесте Колмогорова – Смирнова.**

- a) Frequencies;
- b) 1-Sample K-S;
- c) Descriptives;
- d) Crosstabs;
- e) Compare Means.

**15. Выберите, какой командой необходимо воспользоваться, чтобы посчитать описательные статистики для переменной с  $\text{sig} > 0,05$  в тесте Колмогорова – Смирнова?**

- a) Descriptives;
- b) Frequencies;
- c) 1-Sample K-S;
- d) Crosstabs;
- e) Compare Means.

**16. Выберите, общий алгоритм, с помощью которого будет решена любая задача по нахождению описательных статистик.**

a) при необходимости выполняем условие для отбора, затем проверяем переменную Scale на соответствие закону Гаусса, если  $\text{sig} > 0,05$ , находим min, max, среднее и стандартное отклонение, если  $\text{sig} < 0,05$ , то находим квартили и медиану;

b) при необходимости выполняем условие для отбора, затем проверяем переменную Nominal на соответствие закону Гаусса, если  $\text{sig} > 0,05$ , находим min, max, среднее и стандартное отклонение, если  $\text{sig} < 0,05$ , то находим квартили и медиану;

c) при необходимости выполняем условие для отбора, затем проверяем переменную Scale на соответствие закону Гаусса, если  $\text{sig} > 0,05$ , находим квартили и медиану, если  $\text{sig} < 0,05$ , то находим min, max, среднее и стандартное отклонение;

d) при необходимости выполняем условие для отбора, затем проверяем переменную Ordinal на соответствие закону Гаусса, если  $\text{sig} > 0,05$ , находим квартили и медиану, если  $\text{sig} < 0,05$ , то находим min, max, среднее и стандартное отклонение;

e) при необходимости выполняем условие для отбора, затем проверяем переменную Ordinal на соответствие закону Гаусса, если  $\text{sig} > 0,05$ , находим min, max, среднее и стандартное отклонение, если  $\text{sig} < 0,05$ , то находим квартили и медиану.

**17. Укажите правильную последовательность при решении задачи: посчитать описательные статистики по показателю «сахар в крови» для людей старше 60 лет:**

а) отобрать людей по показателю возраст → проверить на соответствие закону Гаусса показатель «сахар крови», → если закону Гаусса соответствует → посчитать стандартные описательные статистики по показателю «сахар в крови»;

б) проверить на соответствие закону Гаусса показатель «сахар крови» → отобрать людей по показателю возраст, → если закону Гаусса соответствует → посчитать стандартные описательные статистики по показателю «сахар в крови»;

с) отобрать людей по показателю возраст → проверить на соответствие закону Гаусса показатель «возраст», → если закону Гаусса соответствует → посчитать стандартные описательные статистики по показателю «возраст»;

д) отобрать людей по показателю «возраст» → проверить на соответствие закону Гаусса показателю «фильтр» → если закону Гаусса соответствует → посчитать стандартные описательные статистики по показателю «фильтр»;

е) отобрать людей по показателю «возраст» → проверить на соответствие закону Гаусса показатель «возраст», → если закону Гаусса не соответствует → посчитать стандартные описательные статистики по показателю «сахар в крови».

**18. Выберите алгоритм, с помощью которого будет решена следующая задача: посчитать описательные статистики по переменной «возраст» у курящих мужчин:**

а) отбираем курящих мужчин, проверяем «возраст» на соответствие закону Гаусса, если закону соответствует находим min, max, среднее и стандартное отклонение, если не соответствует, то находим квартили и медиану;

б) отбираем курящих мужчин, проверяем «возраст» на соответствие закону Гаусса, если закону соответствует, находим квартили и медиану, если не соответствует, то находим min, max, среднее и стандартное отклонение;

с) отбираем курящих, проверяем «пол» на соответствие закону Гаусса, если закону соответствует, находим квартили и медиану, если не соответствует, то находим min, max, среднее и стандартное отклонение;

д) отбираем мужчин, проверяем «курение» на соответствие закону Гаусса, если закону соответствует, находим квартили и медиану, если не соответствует, то находим min, max, среднее и стандартное отклонение;

е) отбираем мужчин, проверяем «возраст» на соответствие закону Гаусса, если закону соответствует, находим квартили и медиану, если не соответствует, то находим min, max, среднее и стандартное отклонение.

**19. Определите, соответствует ли закону Гаусса показатель «систолическое артериальное давление» после лечения:**

		систолическое артериальное давление после лечения
N		100
Normal Parameters <sup>a</sup>	Mean	151,61
	Std. Deviation	16,862
Most Extreme Differences	Absolute	,124
	Positive	,124
	Negative	-,097
Kolmogorov-Smirnov Z		1,244
Asymp. Sig. (2-tailed)		,090

а) закону Гаусса соответствует, так как Asymp. Sig. (2-tailed) > 0,05;

б) закону Гаусса не соответствует, так как Negative < 0,05;

с) закону Гаусса не соответствует, так как Asymp. Sig. (2-tailed) < 0,05;

д) закону Гаусса соответствует, так как Mean = 151,61 > 0,05;

е) закону Гаусса соответствует, так как Std. Deviation = 16,862 > 0,05.

**20. Определите, соответствует ли закону Гаусса показатель «липопротеиды высокой плотности до лечения».**

One-Sample Kolmogorov-Smirnov Test		
		липопротеиды высокой плотности до лечения
N		100
Normal Parameters <sup>a</sup>	Mean	,957
	Std. Deviation	,7364
Most Extreme Differences	Absolute	,187
	Positive	,187
	Negative	-,142
Kolmogorov-Smirnov Z		1,867
Asymp. Sig. (2-tailed)		,002

- a) закону Гаусса не соответствует, так как Asymp. Sig. (2-tailed) < 0,05;
- b) закону Гаусса соответствует, так как Asymp. Sig. (2-tailed) > 0,05;
- c) закону Гаусса не соответствует, так как Negative < 0,05;
- d) закону Гаусса соответствует, так как Mean = 0,957 > 0,05;
- e) закону Гаусса соответствует, так как Std. Deviation = 0,736 > 0,05.

**21. Выберите, какие описательные статистики были получены для показателя «креатинин мочи»:**

креатинин мочи		
N	Valid	100
	Missing	2
Percentiles	25	3,100
	50	4,200
	75	6,175

- a) 4,2 (3,1; 6,175);
- b) 3,1 (4,2; 6,175);
- c) 4,2 (6,175; 3,1);
- d) 4,2; 6,175;
- e) 4,2; 3,100.

**22. Выберите, чему равняется среднее арифметическое значение по показателю «креатинин сыворотки крови»:**

	N	Minimum	Maximum	Mean		Std. Deviation
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic
креатинин сыворотки крови	100	50	150	108,71	2,475	24,751
Valid N (listwise)	100					

- a) 108,71;
- b)  $(50 + 150)/2 = 100$ ;
- c) 2,475;
- d) 24,751;
- e) 100.

**23. Выберите, какие описательные статистики были получены для показателя «сахар сыворотки до»:**

	N	Minimum	Maximum	Mean		Std. Deviation
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic
сахар сыворотки крови до лечения	100	3,3	9,0	5,966	,1294	1,2938
Valid N (listwise)	100					

- a)  $\min = 3,3$ ;  $\max = 9,0$ , среднее и стандартное отклонение =  $5,966 \pm 1,2938$ ;
- b)  $\min = 3,3$ ;  $\max = 9,0$ , среднее и стандартное отклонение =  $0,1294 \pm 1,2938$ ;
- c)  $\min = 3,3$ ;  $\max = 9,0$ , среднее и стандартное отклонение =  $5,966 \pm 1,2938$ ;
- d)  $\min = 3,3$ ;  $\max = 9,0$ , среднее и стандартное отклонение =  $100 \pm 0,1294$ ;
- e)  $\min = 3,3$ ;  $\max = 9,0$ , среднее и стандартное отклонение =  $100 \pm 1,2938$ .

**24. Выберите, чему равняется стандартное отклонение по показателю «возраст»:**

	N	Minimum	Maximum	Mean		Std. Deviation
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic
возраст	100	33	87	56,59	1,085	10,846
Valid N (listwise)	100					

- a) 1,085;
- b) 56,59;
- c) 10,846;
- d) 33;
- e) 87.

## **Глава 5. СРАВНИТЕЛЬНЫЕ СТАТИСТИКИ**

### **5.1. Что такое сравнительные статистики?**

**5.2. Сравнение средних с применением парного t-критерия Student'a.**

**5.3. Сравнение средних с применением непарного t-критерия.**

**5.4. Сравнение средних с применением парного непараметрического критерия Вилкоксона.**

**5.5. Сравнение средних с применением непараметрического критерия Манна – Уитни.**

**5.6. Задачи для закрепления материала.**

**5.7. Задания для самостоятельной работы.**

**5.8. Тестовые вопросы.**

### **5.1. Что такое сравнительные статистики?**

Сравнение средних значений различных выборок относится к наиболее часто применяемым методам статистического анализа. При этом всегда должен быть выяснен вопрос, можно ли объяснить имеющееся различие средних значений статистическими колебаниями или нет. Если нет, говорят о статистически значимом различии между сравниваемыми группами [11].

Выбор критерия зависит от поставленной задачи, типа данных и количества измерений. При сравнении средних значений выборок предполагается, что, если обе выборки подчиняются нормальному распределению, используют параметрические методы, основанные на таких показателях, как среднее значение и стандартное отклонение. Если это не так, то вычисляются медианы и для сравнения выборок используется непараметрические тесты. В первом случае, для сравнения двух независимых

выборки используется непарный t-критерий, для двух зависимых выборок используется парный t-критерий.

Исходя из вышеизложенного, введем следующие понятия:

**Независимые выборки.** Допустим, мы изучаем ударный объем сердца у жителей Ошской и Иссык-Кульской областей. Для этого мы делаем репрезентативные выборки в этих двух регионах, проводим измерения ударного объема, находим средние по каждой выборке и хотим определить, достоверно ли эти средние различаются друг от друга.

**Зависимые выборки.** Допустим, нашей задачей является испытание препарата, влияющего на работу сердечной мышцы. Для решения задачи мы измеряем ударный объем сердца у группы людей, после этого мы даем препарат и опять измеряем ударный объем у тех же людей. В данном случае сравнение средних должно идти по модели зависимых выборок.

Другими словами, выборки могут быть **независимыми**, если идёт сравнение контрольной и опытной групп, или **зависимыми**, если обе выборки представлены одними и теми же пациентами, но до и после вмешательства.

Первый шаг в статистическом исследовании – формулировка подходящей гипотезы об исследуемой переменной. Обычно такая гипотеза называется нулевой, обозначается  $H_0$ :

➤  $H_0$  гипотеза об отсутствии различий (*нулевая гипотеза*), то есть предположение о том, что на самом деле между параметрами генеральных совокупностей различий нет, а наблюдаемые различия между выборочными показателями носят чисто случайный характер;

➤  $H_1$  гипотеза о значимости различий (*альтернативная гипотеза*), то есть наблюдаемые различия выборочных показателей носят систематический, а не случайный характер и отражают действительное различие неизвестных генеральных показателей [14].

Правило, по которому принимается решение – принять или отклонить нулевую гипотезу, называется критерием. Таким образом, целью проверки гипотезы является принятие решения

о значимости различия средних значений двух генеральных совокупностей по ограниченному числу наблюдений.

При сравнении групп мы изначально исходим из того, что они не отличаются (это –  $H_0$ ). Если вероятность того, что выявленные различия являются случайным результатом весьма мала, тогда правомочным будет отвергнуть нулевую гипотезу и заключить, что различие действительно имеется (верна  $H_1$ ). Показатель различий обозначается «р» (probability, в англоязычной литературе встречается обозначение Р или р). Величиной «р» (или «пи-величина», *англ.* – «P-value») для конкретной выборки называют вероятность получения по крайней мере таких же или ещё больших отличий наблюдаемого от ожидаемого, чем в данной конкретной выборке, при условии, что выдвинутая гипотеза верна. Величина «р» меняется от выборки к выборке, то есть является случайной величиной на множестве выборок (причём, с равномерным распределением на интервале 0 – 1) [11, 14].

Исследователю необходимо решить вопрос о случайности выявленных различий, от этого зависит принятие решения о том, являются ли выявленные различия свидетельством различного состояния и/или свидетельством эффекта от вмешательства. Количественную характеристику случайности представляет теория вероятностей в виде «р»-значения. Чем это значение больше, тем больше вероятность отсутствия различий в пользу нулевой гипотезы, и чем оно меньше, тем больше вероятность наличия различий в пользу альтернативной гипотезы [14].

**ВАЖНО!!!** – «р»-значение является количественной характеристикой только лишь статистической, **НО** не клинической значимости. При наличии статистической значимости необходимо принять решение о клинической важности выявленных различий. Особенно это касается вторичного анализа данных, незапланированного. При первичном запланированном анализе данных обычно проверяется статистическая значимость клинически важных различий [11, 14].

Только принятие (в конкретном случае) правильной или отклонение неправильной гипотезы можно считать верным

решением. Если нулевая гипотеза отвергается, а на самом деле она верна, то возникает ошибка, называемая ошибкой первого рода. Наоборот, если ошибочная гипотеза принимается, то совершается ошибка второго рода [15].

**Ошибка первого рода** – допустимость ошибочного признания различий, то есть альтернативной гипотезы. В медико-биологических исследованиях в качестве критического порога значимости традиционно выбирается уровень 0,05, что допускает наличие ошибки первого рода 5 раз в 100 сравнениях. При  $p \leq 0,05$  различия принимаются статистически значимыми. И чем меньше «р»-значение, тем меньше подобных ошибок: например, при  $p = 0,01$  считается, что ошибка первого рода возможна 1 раз в 100 сравнениях, при  $p = 0,001$  – 1 раз в 1000 сравнениях [14].

**Ошибка второго рода** заключается в принятии неверной нулевой гипотезы.

Выбор статистического теста является чрезвычайно важной задачей. От его правильности будет зависеть качество анализа и, в конечном итоге, надежность выводов. Выбор теста является задачей нетривиальной, но разбираясь в статистических характеристиках данных и используя пошаговый алгоритм, исследователь в состоянии осуществить его корректно. Успешное продвижение по алгоритму выбора подходящего статистического метода анализа предполагает знание ответов на следующие вопросы:

- а) тип данных;
- б) данные зависимые или независимые;
- в) распределение параметрическое (нормальное) или непараметрическое (отличное от нормального);
- г) количество сравниваемых групп.

## **5.2. Сравнение средних с применением парного t-критерия Стьюдента**

Параметрическая статистика используется для анализа непрерывных (численных) переменных, значения которых распределены нормально.

Если данные являются зависимыми (например, получены в процессе повторных наблюдений за одним и тем же пациентом или используются показатели пациентов, подобранных в пары (по возрасту или полу), рекомендуется парный (paired) t-тест [9, 14].

Парный t-критерий Стьюдента – одна из модификаций метода Стьюдента, используемая для определения статистической значимости различий повторных измерений. Он был разработан Уильямом Госсетом для оценки качества пива в компании Гиннесс. В связи с обязательствами перед компанией по неразглашению коммерческой тайны, статья Госсета вышла в 1908 году в журнале «Биометрика» под псевдонимом «Student» (Студент) [14].

Парный t-критерий Стьюдента используется для сравнения двух зависимых (парных) выборок. Зависимыми являются измерения, выполненные у одних и тех же пациентов, но в разное время, например, артериальное давление у больных гипертонической болезнью до и после приема гипотензивного препарата. Нулевая гипотеза гласит об отсутствии различий между сравниваемыми выборками, альтернативная – о наличии статистически значимых различий [14].

Для применения парного t-критерия необходимо, чтобы исходные данные имели нормальное распределение.

Давайте рассмотрим это на практике.

1. Открыть файл «Иванов\_ГБ.sav».
2. Поставить соответствующую **задачу**, например – Определить, есть ли различия между показателями «систолического давления до» и «систолического давления после лечения» у пациентов старше 60 лет.

3. Выясняем, связанные выборки или нет. В данном случае выборки связанные, так как параметры измерены у одних и тех же пациентов, но в разное время – до и после лечения.

4. Далее вспомним простое, но очень важное правило – сначала отбор, потом счет. Сначала необходимо отобрать пациентов старше 60 лет, а только потом сравнить средние значения по переменной «систолическое давление до и после лечения».

5. Применим фильтр, то есть воспользуемся командой «*Select Cases*», в результате применения фильтра получим следующую надпись – «*COMPUTE filter\_\$=(возраст > 60)*».

6. Далее нам необходимо проверить переменные «*систолическое давление до лечения*» и «*систолическое давление после лечения*» на соответствие распределения закону Гаусса, то есть применим тест Колмогорова – Смирнова – «*1-Sample K-S*». Также необходимо помнить, что на тест Колмогорова – Смирнова проверяют только переменные «*Scale*». После применения данного теста, мы видим, что  $Sig_{до} = 0,229$ ,  $Sig_{после} = 0,710$  (рисунок 30). В нашем случае оба значения  $Sig > 0,05$ , это говорит о том, что переменные «*систолическое давление до лечения*» и «*систолическое давление после лечения*» подчиняются закону Гаусса, то есть их значения распределены по нормальному закону.

## NPar Tests

### One-Sample Kolmogorov-Smirnov Test

		систолическое артериальное давление до лечения	систолическое артериальное давление после лечения
N		34	34
Normal Parameters <sup>a</sup>	Mean	177,35	152,47
	Std. Deviation	17,367	17,090
Most Extreme Differences	Absolute	,179	,120
	Positive	,106	,120
	Negative	-,179	-,111
Kolmogorov-Smirnov Z		1,041	,701
Asymp. Sig. (2-tailed)		,229	,710

a. Test distribution is Normal.

Рисунок 30 – Таблица результатов вычисления команды «1-Sample K-S»

7. Раз переменные распределены по нормальному закону и выборки связанные, значит, нам необходимо применить t-критерий Стьюдента для парных выборок. Для этого воспользуемся командой «*Paired-Samples T test*», то есть заходим в меню

**Analyze → Compare Means → Paired-Samples T test**

Появится диалоговое окно «Paired-Samples T test» (рисунок 31, а). Переносим в правую часть переменные «систолическое давление до лечения» и «систолическое давление после лечения». При переносе переменных необходимо в обязательном случае пользоваться кнопкой со стрелочкой, расположенной между двумя частями окна, нажимаем «ОК» (рисунок 31, б).

8. Результаты наших вычислений появились в виде трех таблиц (рисунок 32).

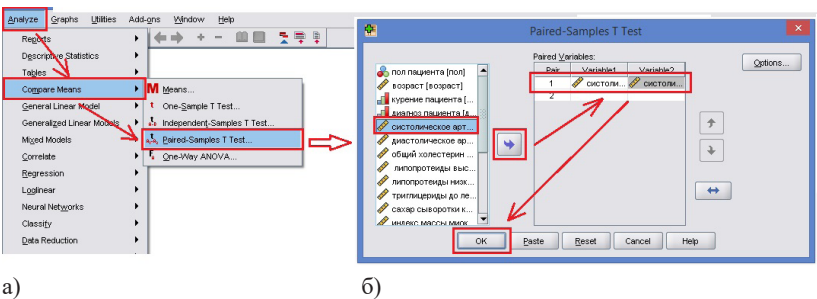


Рисунок 31 – Алгоритм (а) и рабочее окно (б) команды «Paired-Samples T test»

**T-Test**

Paired Samples Statistics				
	Mean	N	Std. Deviation	Std. Error Mean
Pair 1 систолическое артериальное давление до лечения	177,35	34	17,367	2,978
систолическое артериальное давление после лечения	152,47	34	17,090	2,931

Paired Samples Correlations			
	N	Correlation	Sig.
Pair 1 систолическое артериальное давление до лечения & систолическое артериальное давление после лечения	34	,633	,000

Paired Samples Test							
		Paired Differences		95% Confidence Interval of the Difference			
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	
Pair 1 систолическое артериальное давление до лечения - систолическое артериальное давление после лечения		24,882	14,768	2,533	19,729	30,035	9,824 33
							Sig. (2-tailed)
							,000

Рисунок 32 – Таблицы результатов вычисления команды «Paired-Samples T test»

Давайте разберем их. Нам необходимо обратить внимание на значение *Sig (2-tailed)* в третьей таблице,  $Sig = 0,000$ , сравниваем его со значением 0,05 и делаем следующий вывод – в нашем случае полученное значение  $Sig < 0,05$ , поэтому существуют **статистически значимые различия**. Далее нам необходимо найти клинические различия, для этого обращаем внимание на среднее значение «Mean» в первой таблице:  $Mean_{до} = 177,35$ ,  $Mean_{после} = 152,47$ . На основании этого можно сделать следующий вывод: в данном случае существуют клинические значимые различия, так как в ходе лечения у данных пациентов старше 60 лет значение систолического давления после лечения существенно ниже, чем до начала лечения, то есть применяемая тактика лечения данных пациентов эффективна. Различие составляет 24,88 мм рт. ст., то есть  $177,35 - 152,47 = 24,88$ .

9. Мы разобрали полученные таблицы, осталось сформулировать ответ. В данном файле у пациентов старше 60 лет значение систолического давления после лечения статистически значимо ниже, чем значение систолического давления до лечения, так как  $Sig = 0,000$ ,  $Mean_{до} = 177,35$  и  $Mean_{после} = 152,47$ .

**Пояснение 1.** Для сравнения выбран t-критерий Стьюдента для связанных выборок, так как оба показателя соответствуют нормальному распределению Гаусса и относятся к одной группе пациентов.

**Пояснение 2.** Существуют статистически значимые различия, так как  $Sig = 0,000 < 0,05$ .

**Пояснение 3.** По моему мнению, 24,88 мм рт. ст. – это клинически значимое различие ( $177,35 - 152,47 = 24,88$ ).

Надеемся, вам было понятно, как проводить расчет сравнения средних для связанных выборок и переменных, значения которых распределены по нормальному закону.

*P.S.* Зачастую бывает так, что при решении задач данного рода *Sig* в последней таблице больше 0,05. Это говорит о том, что **нет статистически значимых различий**, а имеющиеся клинические различия случайны.

### 5.3. Сравнение средних с применением непарного t-критерия

Для сравнения средних по второму варианту мы воспользуемся так называемым непарным t-тестом, с помощью которого возможно провести проверку гипотезы ( $H_0$ ) об отсутствии различия средних значений переменной в двух независимых выборках.

Непарный t-критерий используется для сравнения двух независимых (непарных) выборок. Независимыми являются измерения, выполненные у разных групп пациентов, в независимости от времени, например, артериальное давление у больных с гипертонической болезнью I и II степени до приема гипотензивного препарата. Нулевая гипотеза гласит об отсутствии различий между сравниваемыми выборками, альтернативная – о наличии статистически значимых различий [14].

Для применения непарного t-критерия необходимо, чтобы исходные данные имели нормальное распределение [14].

Давайте рассмотрим это на практике.

1. Открыть файл «Иванов\_ГБ.sav».
2. Поставить соответствующую **задачу**, например – Определить, есть ли различия между показателями «альбумины мочи в сутки» у курящих и некурящих женщин.
3. Выясняем, связанные выборки в данном случае или нет? Выборки несвязанные, так как параметры измерены у двух разных групп пациентов: группы курящих женщин и группы некурящих женщин, данные измерения также не зависят от времени.
4. Так как выборки несвязанные, необходимо проверить переменную «альбумины мочи в сутки» на закон Гаусса отдельно для группы курящих женщин и группы некурящих женщин.
5. Отбираем группу курящих женщин (*COMPUTE filter\_\$ = (курение = 1 & пол = 1)*) и проверяем переменную «альбумины мочи в сутки» на закон Гаусса. Мы видим, что *Sig = 0,052*, оно больше 0,05, то есть будем считать, что переменная для этой группы подчиняется закону Гаусса (рисунок 33, а).

COMPUTE filter\_\$(курение = 1 & пол = 1)

**NPar Tests**

One-Sample Kolmogorov-Smirnov Test

		альбумины мочи в сутки
N		35
Normal Parameters <sup>a</sup>	Mean	199,57
	Std. Deviation	150,718
Most Extreme Differences	Absolute	,228
	Positive	,228
	Negative	-,177
Kolmogorov-Smirnov Z		1,349
Asymp. Sig. (2-tailed)		,052

a. Test distribution is Normal.

а)

COMPUTE filter\_\$(курение = 2 & пол = 1)

**NPar Tests**

One-Sample Kolmogorov-Smirnov Test

		альбумины мочи в сутки
N		20
Normal Parameters <sup>a</sup>	Mean	179,40
	Std. Deviation	160,421
Most Extreme Differences	Absolute	,262
	Positive	,262
	Negative	-,174
Kolmogorov-Smirnov Z		1,172
Asymp. Sig. (2-tailed)		,128

a. Test distribution is Normal.

б)

Рисунок 33 – Таблицы результатов вычисления команды «1-Sample K-S» для группы курящих женщин (а) и группы некурящих женщин (б)

6. Эти же действия повторим и для группы некурящих женщин (*COMPUTE filter\_\$(курение = 2 & пол = 1)*). Мы видим, что *Sig* = 0,128, оно также больше 0,05, данная переменная для второй группы также подчиняется закону Гаусса (рисунок 33, б).

7. Перед сравнением средних значений необходимо удалить из фильтра группирующую переменную. В нашем случае группирующей переменной является переменная «курение», так как именно данная переменная дала возможность сформировать группы курящих и некурящих. В итоге осталось следующее условие отбора (*COMPUTE filter\_\$(пол = 1)*).

8. Обе переменные распределены по нормальному закону, выборки несвязанные, значит, нам необходимо применить t-критерий для непарных выборок. Для этого воспользуемся командой «Independent-Samples T test», то есть заходим в меню

*Analyze → Compare Means → Independent-Samples T test*

Появится диалоговое окно «Independent-Samples T test» (рисунок 34, а). Переносим в правую часть в поле «Test variable(s)» переменную «альбумины мочи в сутки», а в поле «Grouping Variable» группирующую переменную «курение». Далее необходимо ввести номера групп, которые использовали в фильтре, для этого нажимаем кнопку «Define Groups» и указываем

цифры 1 – для первой группы и 2 – для второй группы, которые мы использовали в фильтре, для формирования групп курящих и не курящих женщин. Нажимаем кнопки «Continue» → «OK» (рисунок 34, б).

9. Результаты наших вычислений появились в виде двух таблиц (рисунок 35).

Давайте разберем их. Во-первых, необходимо отметить, что у наших респондентов наблюдается микроальбуминурия,

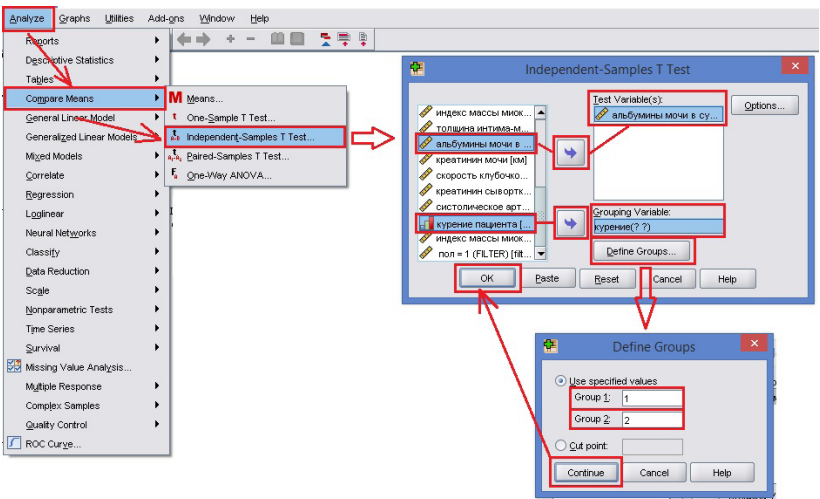


Рисунок 34 – Алгоритм (а) и рабочее окно (б) команды «Independent-Samples T test»

T-Test

Group Statistics

курение пацие	N	Mean	Std. Deviation	Std. Error Mean
альбумины мочи в сутки	35	199,57	160,718	25,476
курит	20	179,40	160,421	35,871

Independent Samples Test

Levene's Test for Equality of Variances				t-test for Equality of Means					
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
альбумины мочи в сутки	,353	,556	-.409	53	,643	20,171	43,242	-66,561	106,904
Equal variances assumed									
Equal variances not assumed			,458	37,648	,649	20,171	43,997	-68,924	109,267

Рисунок 35 – Таблицы результатов вычисления команды «Independent-Samples T test»

поскольку значения альбумина мочи в сутки очень высокие. Во-вторых, нам необходимо обратить внимание на значение *Sig Ливиня* во второй таблице третий столбец,  $Sig = 0,555$ , сравниваем его со значением 0,05, полученное значение  $Sig > 0,05$ , поэтому обращаем внимание на верхнее значение *Sig* во второй таблице в шестом столбце. Во-вторых, итоговое  $Sig = 0,643$ , это больше, чем 0,05, поэтому делаем следующий вывод – **статистически значимых различий не существует!** Мы так же можем обратить внимание на клинические различия  $Mean_{\text{курящие}} = 199,57$ ,  $Mean_{\text{не курящие}} = 179,40$ , по нашему мнению, 20,17 мг/сут., при микроальбуминурии – это клинически незначимое различие, которое можно считать случайным.

10. Мы разобрали полученные таблицы, осталось сформулировать ответ: в данном файле нет статистически значимых различий, по среднему значению альбуминов мочи в сутки, у курящих и не курящих женщин.

**Пояснение 1.** Для сравнения средних выбран непарный t-критерий для несвязанных выборок, так как показатель в обеих выборках соответствует нормальному распределению Гаусса, и сравниваются две разные группы пациентов.

**Пояснение 2.** Статистическую значимость определяем по верхнему *Sig*, так  $Sig_{\text{Ливиня}} = 0,555 > 0,05$ . Нет статистически значимых различий, так как верхний  $Sig = 0,643 > 0,05$ .

**Пояснение 3.** По моему мнению, 20,17 мг/сут. – это клинически незначимое различие, поскольку у пациентов фиксируется микроальбуминурия.

Надеемся, вам было понятно, как проводить расчет сравнения средних для несвязанных выборок и переменных, значения которых распределены по нормальному закону.

## 5.4. Сравнение средних с применением парного непараметрического критерия Вилкоксона

Критерий Вилкоксона для связанных выборок (также используются названия t-критерий Уилкоксона, критерий знаковых

рангов Уилкоксона, критерий суммы рангов Уилкоксона) – непараметрический статистический критерий, используемый для сравнения двух связанных (парных) выборок по уровню какого-либо количественного признака, измеренного в непрерывной или в порядковой шкале [9, 14].

Суть метода состоит в том, что сопоставляются абсолютные величины выраженности сдвигов в том или ином направлении. Для этого сначала все абсолютные величины сдвигов ранжируются, а потом суммируются ранги. Если сдвиги в ту или иную сторону происходят случайно, то и суммы их рангов окажутся примерно равны. Если же интенсивность сдвигов в одну сторону больше, то сумма рангов абсолютных значений сдвигов в противоположную сторону будет значительно ниже, чем это могло бы быть при случайных изменениях [14].

Тест был впервые предложен в 1945 году американским статистиком и химиком Фрэнком Уилкоксоном (1892–1965). В той же научной работе автором был описан еще один критерий, применяемый в случае сравнения независимых выборок.

Т-критерий Вилкоксона используется для оценки различий между двумя рядами измерений, выполненных для одной и той же совокупности исследуемых, но в разных условиях или в разное время. Данный тест способен выявить направленность и выраженность изменений, то есть являются ли показатели больше сдвинутыми в одном направлении, чем в другом [9, 14].

Классическим примером ситуации, в которой может применяться t-критерий Вилкоксона для связанных совокупностей, является исследование «до – после», когда сравниваются показатели до и после лечения. Например, при изучении эффективности гипотензивного средства сравнивается артериальное давление до приема препарата и после приема.

Критерий Вилкоксона является непараметрическим критерием, поэтому, в отличие от парного t-критерия Стьюдента, не требует наличия нормального распределения сравниваемых совокупностей [14].

Число исследуемых при использовании t-критерия Вилкоксона должно быть не менее 5.

Исследуемый признак может быть измерен как в количественной непрерывной (артериальное давление, ЧСС, содержание лейкоцитов в 1 мл крови), так и в порядковой шкале (число баллов, степень тяжести заболевания, степень обсемененности микроорганизмами).

Данный критерий используется только в случае сравнения двух рядов измерений. Аналогом t-критерия Вилкоксона для сравнения трех и более связанных совокупностей является критерий Фридмана.

Давайте рассмотрим это на практике.

1. Открыть файл «Иванов\_ГБ.sav».
2. Поставить соответствующую **задачу**, например – Определить, есть ли различия между показателями «*диастолического давления до*» и «*диастолического давления после лечения*» у женщин с диагнозом Эссенциальная гипертензия II степени.

3. Выясняем, связанные выборки в данном случае или нет. Выборки связанные, так как параметры измерены у одних и тех же пациентов, но в разное время – до и после лечения.

4. Далее вспомним простое, но очень важное правило – сначала отбор, потом счет, то есть сначала необходимо отобрать женщин с диагнозом «Эссенциальная гипертензия II степени», а только потом сравнить среднее значение по переменной «*диастолическое давление до и после лечения*».

5. Применим фильтр, то есть воспользуемся командой «*Select Cases*», в результате применения фильтра получим следующую надпись – «*COMPUTE filter\_\$=(пол = 1 & диагноз = 2)*».

6. Далее нам необходимо проверить переменные «диастолическое давление до лечения» и «диастолическое давление после лечения» на распределение по закону Гаусса, то есть применим тест Колмогорова – Смирнова – «*1-Sample K-S*». Также необходимо помнить, что на тест Колмогорова – Смирнова проверяют только переменные «*Scale*». После применения данного теста, мы видим, что  $Sig_{до} = 0,020$ ,  $Sig_{после} = 0,114$  (рисунок 36). В нашем

случае одно значение  $Sig < 0,05$ , а другое  $Sig > 0,05$ , это говорит о том, переменная «диастолическое давление до лечения» не подчиняется закону Гаусса, а переменная «диастолическое давление после лечения» подчиняется закону Гаусса, то есть в сумме их значения распределены по закону отличному от нормального.

### NPar Tests

One-Sample Kolmogorov-Smirnov Test

		диастолическое артериальное давление до лечения	диастолическое артериальное давление после лечения
N		12	12
Normal Parameters <sup>a</sup>	Mean	99,17	87,50
	Std. Deviation	15,885	10,335
Most Extreme Differences	Absolute	,438	,346
	Positive	,268	,154
	Negative	-,438	-,346
Kolmogorov-Smirnov Z		1,516	1,197
Asymp. Sig. (2-tailed)		,020	,114

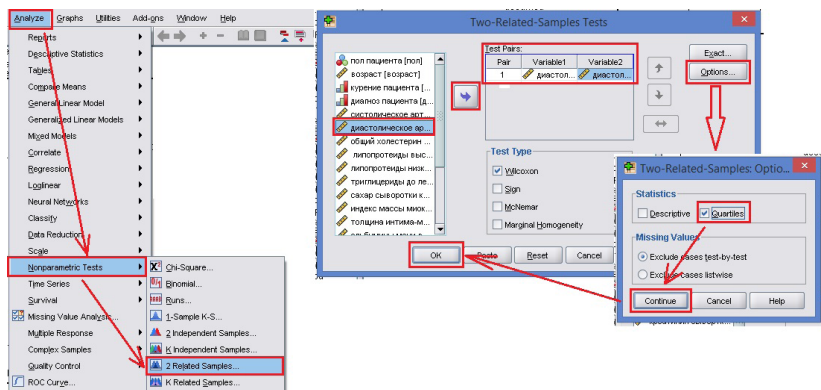
a. Test distribution is Normal.

Рисунок 36 – Таблица результатов вычисления команды «1-Sample K-S»

7. Поскольку переменные распределены по закону, отличному от нормального, и выборки связанные, значит, нам необходимо применить непараметрический t-критерий Вилкоксона для парных выборок. Для этого воспользуемся командой «2 Related Samples Tests», то есть заходим в меню:

*Analyze → Nonparametric Test → 2 Related Samples Tests*

Появится диалоговое окно «2 Related Samples Tests» (рисунок 37, а). Переносим в правую часть переменные «диастолическое давление до лечения» и «диастолическое давление после лечения». При переносе переменных необходимо в обязательном случае пользоваться кнопкой со стрелочкой, расположенной между двумя частями окна. Далее нажимаем кнопку «Options» добавляем галочку в позицию «Quartiles», нажимаем кнопки «Continue» и «ОК» (рисунок 37, б).



а)

б)

Рисунок 37 – Алгоритм (а) и рабочее окно (б) команды «2 Related Samples Tests»

8. Результаты наших вычислений появились в виде трех таблиц (рисунок 38).

Давайте разберем их. Нам необходимо обратить внимание на значение *Sig (2-tailed)* в третьей таблице,  $Sig = 0,040$ , сравниваем его со значением  $0,05$  и делаем следующий вывод – в нашем случае полученное значение  $Sig < 0,05$ , поэтому **существуют статистически значимые различия**. Далее нам необходимо найти клинические различия, для этого обращаем внимание на медиану или  $50^{th}$  (*Median*):  $50^{th}_{до} = 102,50$ ,  $50^{th}_{после} = 90,00$ . На основании этого можно сделать следующий вывод: в данном случае существуют клинически значимые различия, так как в ходе лечения у женщин с диагнозом Эссенциальная гипертензия II степени значение диастолического давления после лечения ниже, чем до начала лечения, то есть применяемая тактика лечения данных пациентов эффективна, но возможно нуждается в корректировке. Различие составляет  $12,5$  мм рт. ст., то есть  $102,5 - 90,00 = 12,5$ .

9. Мы разобрали полученные таблицы, осталось сформулировать ответ: в данном файле у женщин с диагнозом Эссенциальная гипертензия II степени среднее значение диастолического

## NPar Tests

### Descriptive Statistics

	N	Percentiles		
		25th	50th (Median)	75th
диастолическое артериальное давление до лечения	12	100,00	102,50	107,25
диастолическое артериальное давление после лечения	12	82,50	90,00	93,75

### Wilcoxon Signed Ranks

#### Ranks

		N	Mean Rank	Sum of Ranks
диастолическое артериальное давление после лечения - диастолическое артериальное давление до лечения	Negative Ranks	10 <sup>a</sup>	5,60	56,00
	Positive Ranks	1 <sup>a</sup>	10,00	10,00
	Ties	1 <sup>c</sup>		
	Total	12		

a. диастолическое артериальное давление после лечения < диастолическое артериальное давление до лечения

b. диастолическое артериальное давление после лечения > диастолическое артериальное давление до лечения

c. диастолическое артериальное давление после лечения = диастолическое артериальное давление до лечения

#### Test Statistics<sup>a</sup>

	диастолическое артериальное давление после лечения - диастолическое артериальное давление до лечения
Z	-2,055 <sup>a</sup>
Asymp. Sig. (2-tailed)	,040

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

Рисунок 38 – Таблицы результатов вычисления команды «2 Related Samples Tests»

давления после лечения статистически значимо ниже, чем значение диастолического давления до лечения.

**Пояснение 1.** Для сравнения выбран непараметрический t-критерий Вилкоксона, так как один из показателей не соответствует нормальному распределению Гаусса и относится к одной группе пациентов.

**Пояснение 2.** Существуют статистически значимые различия, так как  $\text{Sig} = 0,040 < 0,05$ .

**Пояснение 3.** По моему мнению, 12,5 мм рт. ст. – это клинически значимое различие для диастолического давления ( $102,5 - 90,00 = 12,5$ ).

Надеемся, вам было понятно, как проводить расчет сравнения средних для связанных выборок и переменных, значения которых распределены по закону отличному от нормального.

### **5.5. Сравнение средних с применением непараметрического критерия Манна – Уитни**

U-критерий Манна – Уитни – непараметрический статистический критерий, используемый для сравнения двух независимых выборок по уровню какого-либо признака, измеренного количественно. Метод основан на определении того, достаточно ли мала зона перекрещивающихся значений между двумя вариационными рядами (ранжированным рядом значений параметра в первой выборке и таким же – во второй выборке). Чем меньше значение критерия, тем вероятнее, что различия между значениями параметра в выборках статистически значимы [14].

Данный метод выявления различий между выборками был предложен в 1945 году американским химиком и статистиком Фрэнком Уилкоксоном.

В 1947 году он был существенно переработан и расширен математиками Х.Б. Манном (H.B. Mann) и Д.Р. Уитни (D.R. Whitney), по именам которых сегодня обычно и называется [14].

U-критерий Манна – Уитни является непараметрическим критерием, поэтому, в отличие от t-критерия Стьюдента, не требует наличия нормального распределения сравниваемых совокупностей.

U-критерий подходит для сравнения малых выборок: в каждой из выборок должно быть не менее 3-х значений признака. Допускается, чтобы в одной выборке было 2 значения, но во второй тогда должно быть не менее пяти [14].

Условием для применения U-критерия Манна – Уитни является отсутствие в сравниваемых группах совпадающих значений признака (все числа – разные) или очень малое число таких совпадений [14].

Давайте рассмотрим это на практике.

1. Открыть файл «Иванов\_ГБ.sav».

2. Поставить соответствующую **задачу**, например – Определить, есть ли различия между показателем «*скорость клубочковой фильтрации*» у курящих пациентов с диагнозами Эссенциальная гипертензия I степени и Ренопривная гипертензия.

3. Выясняем, связанные выборки в данном случае или нет? Выборки несвязанные, так как параметры измерены у двух разных групп пациентов: группа курящие пациенты с диагнозом Эссенциальная гипертензия I степени и группа курящих с диагнозом Ренопривная гипертензия.

4. Так как выборки несвязанные, необходимо проверить переменную «*скорость клубочковой фильтрации*» на закон Гаусса отдельно для группы курящих пациентов с диагнозом Эссенциальная гипертензия I степени и группы курящих с диагнозом Ренопривная гипертензия.

5. Отбираем группу курящих пациентов с диагнозом Эссенциальная гипертензия I степени ( $COMPUTE\ filter\_ \$ = (курение = 1 \ \& \ диагноз = 1)$ ) и проверяем переменную «*скорость клубочковой фильтрации*» на закон Гаусса. Мы видим, что  $Sig = 0,014$ , оно меньше 0,05, то есть переменная для этой группы не подчиняется закону Гаусса (рисунок 39, а).

6. Эти же действия повторим и для группы курящих пациентов с диагнозом Ренопривная гипертензия ( $COMPUTE\ filter\_ \$ = (курение = 1 \ \& \ диагноз = 4)$ ). Мы видим, что  $Sig = 0,904$ , оно больше 0,05, данная переменная для второй группы подчиняется закону Гаусса (рисунок 39, б).

Если в первом случае переменная не подчиняется закону Гаусса, а во втором подчиняется, то говорят о том, что обе переменные распределены по закону отличному от нормального.

7. Перед сравнением средних значений необходимо удалить из фильтра группирующую переменную, это такая переменная, по которой в фильтре формируются две разные группы. В нашем случае, группирующей переменной является переменная «*диагноз*», так как именно данная переменная дала возможность сформировать группы пациентов с разными диагнозами. В итоге осталось следующее условие отбора ( $COMPUTE\ filter\_ \$ = (курение = 1)$ ).

COMPUTE filter\_\$(курение = 1 & диагноз = 1)

#### NPar Tests

##### One-Sample Kolmogorov-Smirnov Test

		скорость клубочковой фильтрации
N		18
Normal Parameters <sup>a</sup>	Mean	62,11
	Std. Deviation	3,909
Most Extreme Differences	Absolute	,372
	Positive	,372
	Negative	-,295
Kolmogorov-Smirnov Z		1,579
Asymp. Sig. (2-tailed)		,014

a. Test distribution is Normal.

а)

COMPUTE filter\_\$(курение = 1 & диагноз = 4)

#### NPar Tests

##### One-Sample Kolmogorov-Smirnov Test

		скорость клубочковой фильтрации
N		15
Normal Parameters <sup>a</sup>	Mean	44,60
	Std. Deviation	7,980
Most Extreme Differences	Absolute	,147
	Positive	,147
	Negative	-,137
Kolmogorov-Smirnov Z		,568
Asymp. Sig. (2-tailed)		,904

a. Test distribution is Normal.

б)

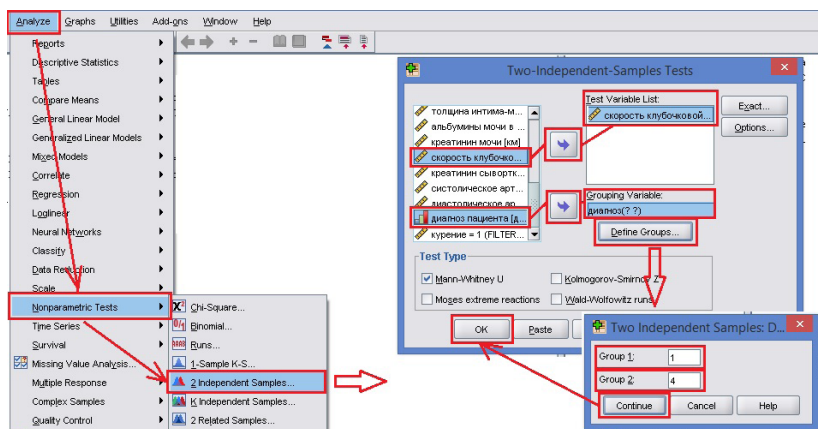
Рисунок 39 – Таблицы результатов вычисления команды «1-Sample K-S» для группы курящих с Эссенциальной гипертензией I степени (а) и группы курящих с Ренопривной гипертензией (б)

8. Обе переменные распределены по закону отличному от нормального, выборки несвязанные, значит, нам необходимо применить непараметрический U-критерий Манна – Уитни для непарных выборок. Для этого воспользуемся командой «2 Independent-Samples Test», то есть заходим в меню

Analyze → Nonparametric Test → 2 Independent-Samples Test

Появится диалоговое окно «2 Independent-Samples Test» (рисунок 40, а). Переносим в правую часть в поле «Test variable(s)» переменную «скорость клубочковой фильтрации», а в поле «Grouping Variable» группирующую переменную «диагноз». Далее необходимо ввести номера групп, которые использовали в фильтре, для этого нажимаем кнопку «Define Groups» и указываем цифры 1 – для первой группы и 4 – для второй группы, которые мы использовали в фильтре, для формирования групп курящих с Эссенциальной гипертензией I степени и курящих с Ренопривной гипертензией. Нажимаем кнопки «Continue» → «OK» (рисунок 40, б).

9. Результаты наших вычислений появились в виде двух таблиц (рисунок 41).



а)

б)

Рисунок 40 – Алгоритм (а) и рабочее окно (б) команды «2 Independent-Samples Test»

## NPar Tests

### Mann-Whitney

#### Ranks

	диагноз пациента	N	Mean Rank	Sum of Ranks
скорость клубочковой фильтрации	эссенциальная гипертензия 1-й степени	18	24,50	441,00
	ренопривная гипертензия	15	8,00	120,00
	Total	33		

#### Test Statistics<sup>a</sup>

	скорость клубочковой фильтрации
Mann-Whitney U	,000
Wilcoxon W	120,000
Z	-.5008
Asymp. Sig. (2-tailed)	,000
Exact Sig. [2*(1-tailed Sig.)]	,000 <sup>a</sup>

a. Not corrected for ties.

b. Grouping Variable: диагноз пациента

Рисунок 41 – Таблицы результатов вычисления команды «2 Independent-Samples Test»

Давайте разберем их. Нам необходимо обратить внимание на значение *Asymp. Sig (2-tailed)* во второй таблице,  $Sig = 0,000$ , сравниваем его со значением 0,05 и делаем следующий вывод – в нашем случае полученное значение  $Sig < 0,05$ , поэтому существуют **статистически значимые различия** и нам необходимо найти клинические различия.

10. Для нахождения клинических различий, нам необходимо найти квартили отдельно для каждой группы курящих пациентов и различным диагнозом, то есть необходимо воспользоваться командой «*Frequencies*» и найти квартили для каждой группы в отдельности (рисунок 42). Если вы не помните, как работает данная команда, проработайте еще раз параграф 4.5. «Описательные статистики, распределение отличное от нормального».

COMPUTE filter_\$( <b>курение = 1 &amp; диагноз = 1</b> )		COMPUTE filter_\$( <b>курение = 1 &amp; диагноз = 4</b> )	
<b>Frequencies</b>		<b>Frequencies</b>	
Statistics		Statistics	
скорость клубочковой фильтрации		скорость клубочковой фильтрации	
N	Valid 18	N	Valid 15
	Missing 1		Missing 0
Percentiles	25 60,00	Percentiles	25 38,00
	50 60,00		50 45,00
	75 63,25		75 50,00

Рисунок 42 – Таблицы результатов вычисления команды «*Frequencies*» для двух групп

Итак, найдя квартили для каждой из групп мы видим, что  $50^{th}_1 = 60,00$ ,  $50^{th}_2 = 45,00$ , по нашему мнению, 15,0 мл/мин – это клинически значимое различие.

11. Мы разобрали полученные таблицы, осталось сформулировать ответ: в данном файле среднее значение «*скорости клубочковой фильтрации*» у курящих пациентов с диагнозом Эссенциальная гипертензия I степени **статистически значимо выше**, чем «*скорость клубочковой фильтрации*» у курящих пациентов с диагнозом Ренопривная гипертония.

**Пояснение 1.** Для сравнения выбран непарный критерий Манна – Уитни, так как в одной из выборок показатель не

соответствует нормальному распределению Гаусса и сравниваются две разные группы пациентов.

**Пояснение 2.** Существуют статистически значимые различия, так как  $\text{Sig} = 0,000 < 0,05$ .

**Пояснение 3.** По моему мнению, 15,0 мл/мин – это клинически значимое различие ( $60,00 - 45,00 = 15,00$ ).

Надеемся, вам было понятно, как проводить расчет сравнения средних для несвязанных выборок и переменных, значения которых распределены по закону отличному от нормального.

## 5.6. Задачи для закрепления материала

- Определить, есть ли различия между показателями *«альбумины мочи в сутки»* у женщин и мужчин с диагнозом Эссенциальная гипертензия I степени.
- Определить, есть ли различия между показателями *«общий холестерин до лечения»* у курящих и некурящих пациентов старше 55 лет.
- Определить, есть ли различия между показателями *«толщина интима- медиа»* у женщин с диагнозом Реноваскулярная гипертония и Ренопривная гипертония.
- Определить, есть ли различия между показателями *«индекс массы миокарда левого желудочка до лечения»* у женщин и мужчин с диастолическим давлением до лечения ниже 100.
- Определить, есть ли различия между показателями *«диастолическое давление после лечения»* у курящих и некурящих.
- Определить, есть ли различия между показателями *«диастолическое давление после лечения»* у пациентов с диагнозом Эссенциальная гипертензия II степени и Реноваскулярная гипертония?
- Определить, есть ли различия между показателями *«диастолическое давление до лечения»* и *«диастолического давления после лечения»* у некурящих мужчин.
- Определить, есть ли различия между показателями *«индекс массы миокарда левого желудочка до лечения»* и *«индекс*

*массы миокарда левого желудочка после лечения» у курящих мужчин с диагнозом Эссенциальная гипертензия I степени.*

- *Определить, есть ли различия между показателями «систолическое давление до лечения» и «систолическое давления после лечения» у женщин с диагнозом Ренопривная гипертензия.*
- *Определить, есть ли различия между показателями «диастолическое давление до лечения» и «диастолического давления после лечения» у курящих пациентов с общим холестерином до лечения выше 5,5.*
- *Определить, есть ли различия между показателями «креатинин мочи» у курящих пациентов с диагнозами Эссенциальная гипертензия I степени и Ренопривная гипертензия.*
- *Определить, есть ли различия между показателями «индекс массы миокарда левого желудочка до лечения» у курящих и некурящих пациентов с диагнозом Эссенциальная гипертензия II степени.*
- *Определить, есть ли различия между показателями «триглицериды до лечения» у курящих и некурящих пациентов старше 50 лет.*
- *Определить, есть ли различия между показателями «диастолическое давление до лечения» у пациентов с Реноваскулярной и Ренопривной гипертензией.*
- *Определить, есть ли различия между показателями «креатинин сыворотки крови» у мужчин и женщин.*
- *Определить, есть ли различия между показателями «индекс массы миокарда левого желудочка до лечения» и «индекс массы миокарда левого желудочка после лечения» у курящих женщин.*
- *Определить, есть ли различия между показателями «диастолическое давление до лечения» и «диастолического давления после лечения» у женщин с диагнозом Эссенциальная гипертензия II степени.*
- *Определить, есть ли различия между показателями «систолическое давление до лечения» и «систолическое давления после*

лечения» у курящих с диагнозом Эссенциальная гипертензия II степени.

- Определить, есть ли различия между показателями «индекс массы миокарда левого желудочка до лечения» и «индекс массы миокарда левого желудочка после лечения» у некурящих мужчин.
- Определить, есть ли различия между показателями «диастолическое давление до лечения» и «диастолического давления после лечения» у курящих с диагнозом Реноваскулярная гипертония.

## 5.7. Задание для самостоятельной работы

Для выполнения данного задания необходимо:

1. Поставить и решить по одной задаче на вашем созданном файле на тему «Сравнительные статистики» для всех четырех вариантов (то есть с применением всех четырех критериев), сохранить решение в файле выдачи. Сохраненный файл выдачи сдать преподавателю, ответить на его вопросы.

2. Продолжить заполнение листа СРС (форму СРС см. в приложении 2) по следующему образцу:

<b>Файл 3 – «Сравнение средних» со связанными выборками</b>			
<b>Сравнение средних 1</b>			
<b>Задача.</b> Определить, есть ли различия между показателями «систолического давления до» и «систолического давления после лечения» у пациентов старше 60 лет			
<b>Ответ:</b>			
САДдо	САДпосле	Критерий	Значимость
177,35	152,47	t-критерий Стьюдента	$P < 0,05$
<b>Пояснение 1.</b> Для сравнения выбран t-критерий Стьюдента для связанных выборок, так как оба показателя соответствуют нормальному распределению Гаусса и относятся к одной группе пациентов			

**Пояснение 2.** Существуют статистически значимые различия, так как  $\text{Sig} = 0,000 < 0,05$ .

**Пояснение 3.** По моему мнению, 24,88 мм рт. ст. – это клинически значимое различие

### Сравнение средних 2

**Задача.** Определить, есть ли различия между показателями «диастолического давления до» и «диастолического давления после лечения» у женщин с диагнозом Эссенциальная гипертензия II степени

**Ответ:**

ДАДдо	ДАДпосле	Критерий	Значимость
102,50	90,00	t-критерий Вилкоксона	$P < 0,05$

**Пояснение 1.** Для сравнения выбран непараметрический t-критерий Вилкоксона, так как один из показателей не соответствует нормальному распределению Гаусса и относятся к одной группе пациентов

**Пояснение 2.** Существуют статистически значимые различия, так как  $\text{Sig} = 0,040 < 0,05$

**Пояснение 3.** По моему мнению, 12,5 мм рт. ст. – это клинически значимое различие

### Файл 4 – «Сравнение средних» с несвязанными выборками

### Сравнение средних 1

**Задача.** Определить, есть ли различия между показателями «альбумины мочи в сутки» (АМ) у курящих и некурящих женщин

**Ответ:**

АМ у 1-й выборки	АМ у 2-й выборки	Критерий	Значимость
199,57	179,40	непарный t-критерий	$P > 0,05$

**Пояснение 1.** Для сравнения средних выбран непарный t-критерий для несвязанных выборок, так как показатель в обеих выборках соответствует нормальному распределению Гаусса, и сравниваются две разные группы пациентов

**Пояснение 2.** Статистическую значимость определяем по верхнему  $\text{Sig}$ , так  $\text{Sig}_{\text{Левия}} = 0,555 > 0,05$ . Нет статистически значимых различий, так как верхний  $\text{Sig} = 0,643 > 0,05$

<b>Пояснение 3.</b> По моему мнению, 20,17 мг/сут. – это клинически незначимое различие, поскольку у пациентов наблюдается микроальбуминурия. В этом случае различие в 20,17 мг/сут. нельзя считать клинически значимым			
<b>Сравнение средних 2</b>			
<b>Задача.</b> Определить, есть ли различия между показателем «скорость клубочковой фильтрации» у курящих пациентов с диагнозами Эссенциальная гипертензия I степени и Ренопривная гипертония			
<b>Ответ:</b>			
СКФ у 1-й выборки	СКФ у 2-й выборки	Критерий	Значимость
60,00	45,00	непарный критерий Манна – Уитни	$P < 0,05$
<b>Пояснение 1.</b> Для сравнения выбран непарный критерий Манна – Уитни, так как в одной из выборок показатель не соответствует нормальному распределению Гаусса, и сравниваются две разные группы пациентов			
<b>Пояснение 2.</b> Существуют статистически значимые различия, так как $\text{Sig} = 0,000 < 0,05$			
<b>Пояснение 3.</b> По моему мнению, 15,0 мл/мин – это клинически значимое различие ( $60,00 - 45,00 = 15,00$ )			

## 5.9. Тестовые вопросы

1. Определите, в каком случае при применении теста Колмогорова – Смирнова говорят о том, что переменная величина распределена по нормальному закону?

- a)  $\text{sig} > 0,05$ ;
- b)  $\text{sig} < 0,05$ ;
- c)  $\text{sig} \leq 0,05$ ;
- d)  $\text{sig} = 0,05$ .

**2. Определите, в каком случае при применении теста Колмогорова – Смирнова говорят о том, что переменная величина распределена по закону, отличному от нормального?**

- a)  $\text{sig} < 0,05$ ;
- b)  $\text{sig} > 0,05$ ;
- c)  $\text{sig} \geq 0,05$ ;
- d)  $\text{sig} = 0,05$ .

**3. Определите, с помощью какого теста будет решена следующая задача: есть ли различия между средними значениями показателя сахара крови до лечения у женщин и мужчин, если переменная «сахар крови» для обеих групп соответствует нормальному распределению.**

- a) Independent-Samples T Test;
- b) Paired-Samples T Test;
- c) 2 Independent Samples;
- d) 2 Related Samples;
- e) 1-Sample K-S.

**4. Определите, с помощью какого теста будет решена следующая задача: есть ли различия между средними значениями показателя диастолического давления после лечения у женщин и мужчин при условии, что один показатель соответствует, а другой не соответствует нормальному распределению:**

- a) 2 Independent Samples;
- b) Independent-Samples T Test;
- c) Paired-Samples T Test;
- d) 2 Related Samples;
- e) 1-Sample K-S.

**5. Определите, с помощью какого теста будет решена следующая задача: есть ли различия между средними значениями показателя глюкоза плазмы крови до и после лечения при условии, что оба показателя соответствуют нормальному распределению:**

- a) Paired-Samples T Test;
- b) 2 Independent Samples;
- c) Independent-Samples T Test;

- d) 2 Related Samples;
- e) 1-Sample K-S.

**6. Определите, с помощью какого теста будет решена следующая задача: есть ли различия между средними значениями показателя индекса массы тела до и после лечения у женщин старше 60 лет при условии, что один показатель не соответствует, а второй соответствует нормальному распределению:**

- a) 2 Related Samples;
- b) Paired-Samples T Test;
- c) 2 Independent Samples;
- d) Independent-Samples T Test;
- e) 1-Sample K-S.

**7. Определите, какая задача является некорректной, то есть в ее решении смысла нет:**

- a) есть ли различия между средними значениями показателей систолического давления до лечения и диастолического давления после лечения;
- b) есть ли различия между средними значениями показателей сахара крови до и после лечения у женщин старше 40 лет;
- c) есть ли различия между средними значениями показателей гемоглобина до и после лечения у некурящих пациентов;
- d) есть ли различия между средними значениями показателей систолического давления до и после лечения у женщин;
- e) есть ли различия между средними значениями показателей сахара мочи до и после лечения у курящих мужчин младше 35 лет.

**8. Определите, какая задача является некорректной, то есть в ее решении смысла нет?**

- a) есть ли различия между средними значениями показателей «пол» и «систолическое артериальное давление до лечения»;
- b) есть ли различия между средними значениями показателей диастолического давления до лечения у мужчин и женщин;
- c) есть ли различия между средними значениями показателей общего холестерина до лечения у курящих и некурящих;

- d) есть ли различия между средними значениями показателей триглицеридов до лечения у курящих и некурящих женщин;
- e) есть ли различия между средними значениями показателей альбуминов мочи до лечения у курящих мужчин и женщин.

**9. Определите, какая задача является некорректной, то есть в ее решении смысла нет:**

- a) есть ли различия между средними значениями показателей сахара крови до лечения у мужчин и сахара крови после лечения у женщин;
- b) есть ли различия между средними значениями показателей креатинина до лечения и после лечения у мужчин;
- c) есть ли различия между средними значениями показателей СОЭ до лечения у курящих и некурящих;
- d) есть ли различия между средними значениями показателей лейкоцитов до и после лечения у некурящих женщин;
- e) есть ли различия между средними значениями показателей тромбоцитов до лечения у курящих мужчин и женщин.

**10. Определите, какая задача является некорректной, то есть в ее решении смысла нет:**

- a) есть ли различия между средними значениями показателей «рост у курящих мужчин» и «вес у некурящих женщин»;
- b) есть ли различия между средними значениями показателей «вес у курящих и некурящих женщин»;
- c) есть ли различия между средними значениями показателей общего холестерина у курящих и некурящих;
- d) есть ли различия между средними значениями показателей систолического давления у некурящих мужчин и женщин;
- e) есть ли различия между средними значениями показателей креатинина мочи у курящих мужчин и женщин.

**11. Определите, какая задача является некорректной, то есть в ее решении смысла нет:**

- a) есть ли различия между средними значениями показателей «диагноз» у пациентов с легкой степенью заболевания;

б) есть ли различия между средними значениями показателей гемоглобина у пациентов с легкой и тяжелой степенью заболевания;

с) есть ли различия между средними значениями показателей СОЭ у пациентов с легкой и средней степенью заболевания;

д) есть ли различия между средними значениями показателей возраст у женщин с легкой и тяжелой степенью заболевания;

е) есть ли различия между средними значениями показателей сахара крови у пациентов со средней и тяжелой степенью заболевания.

**12. Определите, какая задача является некорректной, то есть в ее решении смысла нет:**

а) есть ли различия между средними значениями показателей успеваемости у мальчиков 11 А и девочек 11 В;

б) есть ли различия между средними значениями показателей «кратковременная память» у мальчиков и девочек 11 А;

с) есть ли различия между средними значениями показателей «счет в уме» у мальчиков 11 А и 11 Б;

д) есть ли различия между средними значениями показателей «осведомленность» у мальчиков и девочек 11 В;

е) есть ли различия между средними значениями показателей «успеваемость» у мальчиков и девочек 11 А.

**13. Укажите правильную последовательность при решении задачи: есть ли различия между средними значениями показателя индекса массы тела до и после лечения у людей в возрасте от 18 до 35 лет.**

а) отобрать людей по показателю «возраст» → проверить на соответствие закону Гаусса индекс массы тела до и после лечения → если оба параметра соответствуют закону Гаусса, то запустим тест «Paired-Samples T Test»; если хотя бы один не соответствует закону Гаусса, то выбираем «2 Related Samples»;

б) проверить на соответствие закону Гаусса индекс массы тела до и после лечения → если оба параметра соответствуют закону Гаусса, то запустим тест «Paired-Samples T Test»;

если хотя бы один не соответствует закону Гаусса, то выбираем «2 Related Samples»;

с) отобрать людей по показателю возраст → проверить на соответствие закону Гаусса индекс массы тела до и после лечения → если оба параметра соответствуют закону Гаусса, то запустим тест «Independent-Samples T Test»; если хотя бы один не соответствует закону Гаусса, то выбираем «2 Independent Samples»;

д) отобрать людей по показателю «возраст» → проверить на соответствие закону Гаусса индекс массы тела до и после лечения → если оба параметра соответствуют закону Гаусса, то запустим тест «2 Related Samples»; если хотя бы один не соответствует закону Гаусса, то выбираем «Paired-Samples T Test»;

е) отобрать людей по показателю индекс массы тела → проверить на соответствие закону Гаусса возраст → если оба параметра соответствуют закону Гаусса, то запустим тест «2 Related Samples»; если хотя бы один не соответствует закону Гаусса, то выбираем «Paired-Samples T Test».

**14. Укажите правильную последовательность при решении задачи: есть ли различия по среднему значению показателя индекса массы тела до лечения у людей, страдающих инсулинозависимым и инсулиннезависимым типами диабета в возрасте до 50 лет.**

а) отобрать людей с инсулинозависимым типом диабета в возрасте до 50 лет → проверить на соответствие закону Гаусса индекс массы тела до лечения → отобрать людей с инсулиннезависимым типом диабета в возрасте до 50 лет → проверить на соответствие закону Гаусса индекс массы тела до лечения → отобрать всех людей в возрасте до 50 → если оба параметра соответствуют закону Гаусса, то запустим тест «Independent-Samples T Test»; если хотя бы один не соответствует закону Гаусса, то выбираем «2 Independent-Samples»;

б) отобрать людей с инсулинозависимым типом диабета в возрасте до 50 лет → проверить на соответствие закону Гаусса индекс массы тела до лечения → отобрать людей с инсулиннезависимым типом диабета в возрасте до 50 лет → проверить на

соответствие закону Гаусса индекс массы тела до лечения → если оба параметра соответствуют закону Гаусса, то запустим тест «Independent-Samples T Test»; если хотя бы один не соответствует закону Гаусса, то выбираем «2 Independent-Samples»;

с) отобрать людей с инсулинозависимым типом диабета в возрасте до 50 лет → проверить на соответствие закону Гаусса индекс массы тела до лечения → отобрать людей с инсулиннезависимым типом диабета в возрасте до 50 лет → проверить на соответствие закону Гаусса индекс массы тела до лечения → удалить фильтр → если оба параметра соответствуют закону Гаусса, то запустим тест «Independent-Samples T Test»; если хотя бы один не соответствует закону Гаусса, то выбираем «2 Independent Samples»;

д) отобрать людей с инсулинозависимым типом диабета в возрасте 50 лет → проверить на соответствие закону Гаусса индекс массы тела до лечения → отобрать людей с инсулин независимым типом диабета в возрасте 50 лет → проверить на соответствие закону Гаусса индекс массы тела до лечения → удалить фильтр → если оба параметра соответствуют закону Гаусса, то запустим тест «Independent-Samples T Test»; если хотя бы один не соответствует закону Гаусса, то выбираем «2 Independent-Samples»;

е) отобрать людей с инсулинозависимым типом диабета в возрасте от 50 лет → проверить на соответствие закону Гаусса индекс массы тела до лечения → отобрать людей с инсулин независимым типом диабета в возрасте от 50 лет → проверить на соответствие закону Гаусса индекс массы тела до лечения → отобрать всех людей в возрасте от 50 → если оба параметра соответствуют закону Гаусса, то запустим тест «Independent-Samples T Test»; если хотя бы один не соответствует закону Гаусса, то выбираем «2 Independent-Samples».

**15. Есть ли различия по среднему значению систолического давления после лечения у людей, болеющих эссенциальной гипертензией первой и второй степени:**

Group Statistics					
	diagnosis	N	Mean	Std. Deviation	Std. Error Mean
систолическое артериальное давление после лечения	эссенциальная гипертензия I-й степени	25	136,72	15,252	3,050
	эссенциальная гипертензия 2-й степени	23	150,17	10,061	2,265

Independent Samples Test										
		Levene's Test for Equality of Variances			t-test for Equality of Means					
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
систолическое артериальное давление после лечения	Equal variances assumed	1,709	,198	-3,492	46	,034	-13,454	3,852	-21,209	-5,699
	Equal variances not assumed			-3,541	43,374	,056	-13,454	3,789	-21,114	-5,794

а) у пациентов с диагнозом эссенциальная гипертензия I степени среднее значение систолического давления после лечения статистически значимо ниже, чем у пациентов с диагнозом «Эссенциальная гипертензия II степени, так как Sig = 0,034, это меньше, чем 0,05;

б) у пациентов с диагнозом «Эссенциальная гипертензия II степени среднее значение систолического давления после лечения статистически значимо ниже, чем у пациентов с диагнозом «Эссенциальная гипертензия I степени», так как Sig = 0,034, это меньше чем 0,05;

с) у пациентов с диагнозом «Эссенциальная гипертензия II степени» среднее значение систолического давления после лечения статистически значимо ниже, чем у пациентов с диагнозом «Эссенциальная гипертензия I степени», так как Sig = 0,056, это больше чем 0,05;

д) у пациентов с диагнозом «Эссенциальная гипертензия I степени» среднее значение систолического давления после лечения статистически значимо выше, чем у пациентов с диагнозом «Эссенциальная гипертензия II степени», так как Sig = 0,198, это больше чем 0,05;

е) у пациентов с диагнозом «Эссенциальная гипертензия II степени» среднее значение систолического давления после лечения статистически значимо ниже, чем у пациентов с диагнозом «Эссенциальная гипертензия I степени», так как Sig = 0,198, это больше, чем 0,05.

## 16. Выберите правильное суждение:

Descriptive Statistics				
	N	Percentiles		
		25th	50th (Median)	75th
систолическое артериальное давление до лечения	100	160,00	170,00	190,00
систолическое артериальное давление после лечения	100	165,00	180,00	190,00

Test Statistics <sup>a</sup>	
	систолическое артериальное давление после лечения - систолическое артериальное давление до лечения
Z	-8,700 <sup>a</sup>
Asymp. Sig. (2-tailed)	,032

- а) статистически значимые различия по показателю «систолическое давление» до и после лечения есть, так как  $\text{Sig} = 0,032$ , это меньше, чем  $0,05$ , но лечение было неэффективным;
- б) статистически значимые различия по показателю «систолическое давление» до и после лечения есть, так как  $\text{Sig} = 0,032$ , это меньше, чем  $0,05$ , лечение было эффективным;
- в) статистически значимых различий по показателю «систолическое давление» до и после лечения нет, так как  $\text{Sig} = 0,032$ , это меньше, чем  $0,05$ , лечение было неэффективным;
- г) статистически значимые различия по показателю «систолическое давление» до и после лечения есть, так как медиана равна 170 и 180, лечение было эффективным;
- д) статистически значимых различий по показателю «систолическое давление» до и после лечения нет, так как медиана равна 170 и 180, лечение было не эффективным.

## 17. Выберите правильное суждение:

	Mean	N	Std. Deviation	Std. Error Mean
Pair 1 систолическое артериальное давление до лечения	173,29	100	17,490	1,749
систолическое артериальное давление после лечения	151,61	100	16,862	1,686

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Pair 1 систолическое артериальное давление до лечения - систолическое артериальное давление после лечения	21,680	11,152	1,115	19,467	23,893	19,441	99	,000

а) статистически значимые различия по показателю «систолическое давление» до и после лечения есть, так как Sig = 0,000, это меньше, чем 0,05;

б) статистически значимые различия по показателю «систолическое давление» до и после лечения есть, так как Mean = 173,29 и Mean = 151,61;

с) статистически значимые различия по показателю «систолическое давление» до и после лечения есть, так как Std Error Mean (до) = 1,749 и Std Error Mean (после) = 1,686;

д) статистически значимых различий по показателю «систолическое давление» до и после лечения нет, так как Sig = 0,000, это меньше, чем 0,05;

е) статистически значимых различий по показателю «систолическое давление» до и после лечения нет, так как Mean = 21,680.

## 18. Выберите правильное суждение:

	курение	N	Mean Rank	Sum of Ranks
общий холестерин ммоль/л до лечения	курит	65	50,35	3272,50
	не курит	35	50,79	1777,50
	Total	100		

Test Statistics<sup>a</sup>

	общий холестерин ммоль/л до лечения
Mann-Whitney U	1127,500
Wilcoxon W	3272,500
Z	-,072
Asymp. Sig. (2-tailed)	,942

а) статистически значимых различий по показателю «общий холестерин» до лечения у курящих и некурящих нет, так как Sig = 0,942, это больше, чем 0,05;

б) статистически значимых различий по показателю «общий холестерин» до лечения у курящих и некурящих нет, так как Sig = 0,072, это больше, чем 0,05;

с) статистически значимые различия по показателю «общий холестерин» до лечения у курящих и некурящих есть, так как Sig = 0,942, это больше, чем 0,05;

д) статистически значимые различия по показателю «общий холестерин» до лечения у курящих и некурящих есть, так как Sig = 0,072, это больше чем 0,05;

е) статистически значимые различия по показателю «общий холестерин» до лечения у курящих и некурящих есть, так как Sig = 0,942 и 0,072, это больше, чем 0,05.

### 19. Выберите правильное суждение:

	N	Percentiles		
		25th	50th (Median)	75th
диастолическое артериальное давление до лечения	100	95,00	100,00	108,00
диастолическое артериальное давление после лечения	100	80,00	90,00	93,75

		N	Mean Rank	Sum of Ranks
диастолическое артериальное давление после лечения - диастолическое артериальное давление до лечения	Negative Ranks	77 <sup>a</sup>	40,66	3131,00
	Positive Ranks	3 <sup>a</sup>	36,33	109,00
	Ties	20 <sup>c</sup>		
	Total	100		

	диастолическое артериальное давление после лечения - диастолическое артериальное давление до лечения
Z	-7,302 <sup>a</sup>
Asymp. Sig. (2-tailed)	,000

а) диастолическое давление до лечения статистически значимо выше, чем после, так как Sig = 0,000, это меньше, чем 0,05, лечение было эффективным;

б) диастолическое давление до лечения статистически значимо ниже, чем после, так как Sig = 0,000, это меньше, чем 0,05, лечение было эффективным;

с) диастолическое давление до лечения статистически значимо выше, чем после, так как  $\text{Sig} = 0,000$ , это меньше, чем  $0,05$ , лечение было не эффективным;

д) диастолическое давление до лечения статистически значимо ниже, чем после, так как  $\text{Sig} = 0,000$ , это меньше, чем  $0,05$ , лечение было не эффективным;

е) диастолическое давление до лечения статистически значимо выше, чем после, так как  $\text{Sig} = 0,000$ , это больше, чем  $0,05$ , лечение было не эффективным.

## **Глава 6. КОРРЕЛЯЦИОННЫЕ СТАТИСТИКИ**

**6.1. Что такое корреляция?**

**6.2. Критерии корреляции.**

**6.3. Виды корреляции.**

**6.4. Корреляция по типу Scale – Scale.**

**6.5. Корреляция по типу Nominal – Nominal, Ordinal – Ordinal, Nominal – Ordinal или Ordinal – Nominal.**

**6.6. Корреляция по типу Scale – Nominal или Scale – Ordinal.**

**6.7. Задачи для закрепления материала.**

**6.8. Задания для самостоятельной работы.**

**6.9. Тестовые вопросы.**

### **6.1. Что такое корреляция?**

Изучая историю, нетрудно заметить, что существует взаимосвязь явлений и процессов, происходящих в природе и обществе, внутри общества, во времени и пространстве.

Например, оценка исторического факта предполагает выявление факторов, способствовавших его появлению, а их оценка в историческом исследовании чаще всего бывает расплывчатой [16].

До сих пор мы выясняли лишь сам факт существования статистической зависимости различий между двумя признаками.

Корреляционный анализ – статистический метод изучения взаимосвязи между двумя и более случайными величинами.

Корреляционный анализ является очень популярным методом аналитической статистики, используемым в медико-биологических исследованиях. По мнению А. М. Гржибовского, в российских научных публикациях корреляционный анализ находится на втором месте среди используемых в научных статьях методов, после критерия Стьюдента [16].

**Корреляция** (от *лат.* correlatio), корреляционная зависимость – взаимозависимость двух или нескольких случайных величин. Данные термин употребляется в науке с конца XVIII века, его ввел французский палеонтолог Жорж Кювье, основавший «закон корреляции», согласно которому череп с рогами обязательно принадлежал травоядному животному, обладавшему копытными конечностями; ежели лапа имела когти, то животное было хищным. Об этом «законе» сохранился рассказ о неудачной шутке студентов, пытавшихся во время университетского карнавала напугать Кювье. Ряженный в шкуре и маске с рогами крикнул профессору: «Я тебя съем!» На что получил спокойный ответ от профессора, что рогатых хищников не бывает, а за незнание закона корреляции можно получить плохую оценку [16].

Однако в статистике термин «корреляция» (применительно к корреляции Спирмена) впервые использовал английский биолог и статистик Фрэнсис Гальтон (1822–1911) в конце XIX в. [5]. В 1892 году он первым предложил принципы того, как можно вычислить коэффициент корреляции. На работу Фрэнсиса Гальтона огромное влияние оказали труды Чарльза Дарвина, который приходился ему двоюродным братом. В 1888 году Гальтон сделал доклад «Корреляции и их измерение, преимущественно по антропометрическим данным» на заседании Королевского общества, который был посвящен корреляции между длиной рук и ног у пропорционально сложенного человека [16].

Также огромный вклад в развитие корреляционного анализа внес Карл Пирсон (1857–1936), английский математик и биолог, чьим научным руководителем в Университетском колледже Лондона был Фрэнсис Гальтон. Карл Пирсон является основоположником математической статистики, в частности теории корреляции.

Кроме того, в развитие корреляционного анализа внесли существенный вклад следующие ученые: Чарльз Эдвард Спирмен (1863–1945), Морис Джордж Кендалл (1907–1983), Александр Александрович Чупров (1874–1926), Джордж Юл (1871–1951) и многие другие [17].

Суть корреляции заключается в том, что при изменении значения одной переменной происходит закономерное изменение (уменьшение или увеличение) другой переменной. Корреляция – взаимозависимость случайных факторов, которая отражает только приближенную взаимосвязь, то есть не говорит о причинно-следственных связях [16].

Объясним на примере. Если бы в исследуемой выборке между ростом и весом человека существовала корреляционная зависимость, то это не значило бы, что вес является причиной, влияющей на значение величины роста человека, иначе при сбросе лишних килограммов рост человека также уменьшался. Корреляционная связь лишь говорит о взаимосвязанности данных параметров, причем в данной конкретной выборке, в другой выборке мы можем не наблюдать полученные корреляции [14, 16].

Причинная зависимость предполагает, что один из пары рассматриваемых признаков выступает как фактор, второй – как результат. Например, высокогорье может рассматриваться фактором повышения систолического и диастолического давления у людей.

Корреляционный анализ встречается в отечественной биомедицинской литературе чаще, чем в зарубежной, вероятно, из-за его кажущейся простоты. Однако грамотный корреляционный анализ имеет много тонкостей и «подводных камней». Часто при представлении результатов корреляционного анализа авторы допускают ошибки и неточности. Одними из основных ошибок являются выбор неверного коэффициента корреляции и некорректная интерпретация результатов корреляционного анализа. Кроме того, авторы допускают следующую ошибку: по величине коэффициента корреляции судят о наличии статистически значимых различий, хотя данный коэффициент указывает только на силу связи.

Для корректного описания и грамотной интерпретации корреляционной связи должны использоваться следующие характеристики (к сожалению, обычно авторы уделяют внимание только одной или двум из них):

### *1. Есть ли связь между переменными?*

В качестве критического в медико-биологических исследованиях обычно принимается  $p \leq 0,05$  ( $\text{sig} \leq 0,05$ ). Наряду с величиной коэффициента корреляции и его знаком указывать уровень значимости является обязательным. Так как, например, при исследовании малых выборок может быть выявлена средняя или сильная корреляционная связь, при этом она может оказаться статистически незначимой ( $p \geq 0,05$ ) [14, 17].

### *2. Какова сила связи?*

Коэффициент корреляции обозначается  $r$ , который изменяется от -1 до 1 и принимает следующие значения:

$r = 0$  – связь отсутствует;

$r \leq 0,29$  – связь слабая;

$r = 0,3 - 0,49$  – связь умеренная;

$r = 0,5 - 0,69$  – связь значительная;

$r = 0,7 - 0,99$  – связь сильная;

$r = 1,0$  связь функциональная.

### *3. Какой характер связи?*

Если коэффициент корреляции положительный, то связь **прямая**, то есть увеличение показателей одной переменной ведет к увеличению показателей другой переменной, и наоборот. Если коэффициент корреляции отрицательный, то связь **обратная**, то есть увеличение показателей одной переменной, ведет к уменьшению показателей другой переменной, и наоборот.

От чего же зависит выбор конкретного коэффициента корреляции? Во-первых, от типа шкалы переменной. Во-вторых, если обе переменные количественные, то выбор коэффициента корреляции объясняется формой распределения и некоторыми другими характеристиками.

## **6.2. Критерии корреляции**

### **Критерий корреляции Пирсона**

Критерий корреляции Пирсона – это метод параметрической статистики, позволяющий определить наличие или отсутствие

линейной связи между двумя количественными показателями, а также оценить ее тесноту и статистическую значимость. Другими словами, критерий корреляции Пирсона позволяет определить, изменяется ли (возрастает или уменьшается) один показатель в ответ на изменения другого? В статистических расчетах и выводах коэффициент корреляции обычно обозначается как  $r(r_{xy})$  [17].

Критерий корреляции Пирсона был разработан командой британских ученых во главе с Карлом Пирсоном (1857–1936) в 90-х годах XIX века для упрощения анализа ковариации двух случайных величин. Помимо Карла Пирсона над критерием корреляции Пирсона работали также Фрэнсис Эджуорт и Рафаэль Уэлдон [17].

Критерий корреляции Пирсона позволяет определить, какова теснота (или сила) корреляционной связи между двумя показателями, измеренными в количественной шкале. При помощи дополнительных расчетов можно также определить, насколько статистически значима выявленная связь.

Например, при помощи критерия корреляции Пирсона можно ответить на вопрос о наличии связи между температурой тела и содержанием лейкоцитов в крови при острых респираторных инфекциях, между ростом и весом пациента, между содержанием в питьевой воде фтора и заболеваемостью населения флюорозом [17].

Условия и ограничения применения коэффициента корреляции Пирсона:

1. Сопоставляемые показатели должны быть измерены в количественной шкале (например, частота сердечных сокращений, температура тела, содержание лейкоцитов в 1 мл крови, систолическое артериальное давление).

2. Посредством критерия корреляции Пирсона можно определить лишь наличие и силу взаимосвязи между величинами, а также направление (прямая или обратная).

3. Количество сопоставляемых величин должно быть равно двум.

4. Критерий корреляции Пирсона является параметрическим, в связи с чем условием его применения служит нормальное распределение каждой из сопоставляемых переменных. В случае необходимости корреляционного анализа показателей, распределение которых отличается от нормального, следует использовать коэффициент ранговой корреляции Спирмена.

### **Критерий корреляции Спирмена**

Коэффициент ранговой корреляции Спирмена – это непараметрический метод, который используется с целью статистического изучения связи между явлениями. В этом случае дается оценка тесноты установленной связи с помощью количественно выраженного коэффициента [17].

Данный критерий был разработан и предложен для проведения корреляционного анализа в 1904 году Чарльзом Эдвардом Спирменом, английским психологом, профессором Лондонского и Честерфилдского университетов.

Коэффициент ранговой корреляции Спирмена используется для выявления и оценки тесноты связи между двумя рядами сопоставляемых количественных показателей. В том случае, если ранги показателей, упорядоченных по степени возрастания или убывания, в большинстве случаев совпадают (большему значению одного показателя соответствует большее значение другого показателя. Например, при сопоставлении роста пациента и его массы тела) делается вывод о наличии прямой корреляционной связи. Если ранги показателей имеют противоположную направленность (большему значению одного показателя соответствует меньшее значение другого, например, при сопоставлении возраста и частоты сердечных сокращений), то говорят об обратной связи между показателями [17].

Коэффициент корреляции Спирмена обладает следующими свойствами:

1. Коэффициент корреляции может принимать значения от минус единицы до единицы, причем при  $r = 1$  имеет место строго прямая связь, а при  $r = -1$  – строго обратная связь.

2. Если коэффициент корреляции отрицательный, то имеет место обратная связь, если положительный, то прямая связь.

3. Если коэффициент корреляции равен нулю, то связь между величинами практически отсутствует.

4. Чем ближе модуль коэффициента корреляции к единице, тем более сильной является связь между измеряемыми величинами.

В связи с тем, что коэффициент Спирмена рассчитывается с помощью непараметрического метода, проверка на нормальность распределения не требуется.

Эффективность и качество оценки методом Спирмена снижается, если разница между различными значениями какой-либо из измеряемых величин достаточно велика. Не рекомендуется использовать коэффициент Спирмена, если имеет место неравномерное распределение значений измеряемой величины [17].

### **Критерий корреляции Фи и Ви-Крамера**

К наиболее часто используемым инструментам изучения взаимосвязи двух переменных относятся методы анализа таблицы сопряженности. Анализ таблицы является весьма простым и наглядным, и вместе с тем эффективным инструментом изучения одновременно двух переменных. Таблица сопряженности является наиболее универсальным средством изучения статистических связей, в ней могут быть представлены переменные, измеренные в любой шкале. Чтобы высказать предположение о зависимости признаков, следует анализировать не абсолютные частоты в таблице, а относительные, отнесенные либо к маргинальным частотам строк, либо к маргинальным частотам столбцов. Значительное расхождение в распределениях данных частот свидетельствует о наличии связи между признаками [17].

Основным критерием для проверки гипотезы о наличии связи между признаками на основе таблицы сопряженности является критерий Хи-квадрат Пирсона.

Заметим, что статистика критерия с точностью до обозначений совпадает со статистикой критерия Пирсона для проверки гипотезы однородности, то есть, по сути, это один и тот же критерий. Действительно, любая задача однородности нескольких

совокупностей может быть переформулирована в терминах независимости тех или иных признаков. Допустим, мы проверяем гипотезу о равенстве (различии) средних доходов двух групп респондентов, которые различаются, скажем, по половому признаку. Данная гипотеза равенства (различия) средних доходов респондентов равносильна гипотезе о независимости (зависимости) средних доходов респондентов от пола респондента [17].

Помимо критерия Хи-квадрат Пирсона при анализе таблиц сопряженности используется также критерий Хи-квадрат, основанный на методе максимального правдоподобия. На практике статистика максимума правдоподобия Хи-квадрат очень близка по величине к рассмотренной статистике Пирсона Хи-квадрат.

Для таблиц сопряженности  $2 \times 2$  используют критерий Хи-квадрат с поправкой Йетса на непрерывность, а также точный критерий Фишера. В случае зависимых выборок для таблиц сопряженности  $2 \times 2$  используют критерий Макнемара.

Для характеристики степени связи между признаками в таблице сопряженности наиболее часто используют коэффициент сопряженности Пирсона и, статистику Ви-Крамера [17].

### 6.3. Виды корреляции

*Существует три типа корреляции:*

1. Scale – Scale.
2. Nominal – Nominal, Ordinal – Ordinal или Nominal – Ordinal.
3. Scale – Ordinal или Scale – Nominal.

Рассмотрим каждый из них в отдельности.

#### **Корреляция Scale – Scale**

Этот вид корреляции является самым распространенным.

Давайте рассмотрим этот вид на практике.

1. Открыть файл «Иванов\_ГБ.sav».
2. Поставить соответствующую **задачу**, например – Определить, есть ли связь между показателями «*общий холестерин до лечения*» и «*липопротеидами низкой плотности до лечения*» у курящих пациентов.

3. Вспомним простое, но очень важное правило – сначала отбор, потом счет, то есть сначала необходимо отобрать курящих пациентов, а только потом определить связь между данными показателями.

4. Применим фильтр, то есть воспользуемся командой «*Select Cases*», в результате применения фильтра получим следующую надпись – «*COMPUTE filter\_\$=(курение = 1)*».

5. Далее нам необходимо проверить переменные «общий холестерин до лечения» и «липопротеидами низкой плотности до лечения» на закон Гаусса, то есть применим тест Колмогорова – Смирнова – «*1-Sample K-S*». Также необходимо помнить, что на тест Колмогорова – Смирнова проверяют только переменные «Scale». После применения данного теста, мы видим, что  $Sig_{хол} = 0,049$ ,  $Sig_{лппп} = 0,018$  (рисунок 43). В нашем случае оба значения  $Sig < 0,05$ , это говорит о том, переменные «общий холестерин до лечения» и «липопротеидами низкой плотности до лечения» не подчиняются закону Гаусса, то есть их значения распределены по закону отличному от нормального.

6. Раз переменные распределены по закону, отличному от нормального, значит, нам необходимо применить критерий Спирмена (если переменные распределены по нормальному закону,

## NPAr Tests

### One-Sample Kolmogorov-Smirnov Test

		общий холестерин ммоль/л до лечения	липопротеиды низкой плотности до лечения
N		65	65
Normal Parameters <sup>a</sup>	Mean	5,528	3,786
	Std. Deviation	1,2832	1,3216
Most Extreme Differences	Absolute	,169	,190
	Positive	,169	,190
	Negative	-,069	-,116
Kolmogorov-Smirnov Z		1,363	1,536
Asymp. Sig. (2-tailed)		,049	,018

a. Test distribution is Normal.

Рисунок 43 – Таблица результатов вычисления команды «1-Sample K-S»

используется критерий Пирсона). Для этого воспользуемся командой «Bivariate», то есть заходим в меню

*Analyze → Correlate → Bivariate*

Появится диалоговое окно «Bivariate» (рисунок 44, а). Переносим в правую часть переменные «общий холестерин до лечения» и «липопротеидами низкой плотности до лечения». Убираем галочку с позиции «Pearson» и добавляем галочку в позицию «Spearman», затем нажимаем кнопку «OK» (рисунок 44, б).

7. Получаем одну таблицу, у которой два последних столбца и две последних строки идентичны по диагонали. Нам нужно смотреть или на последнюю строку, предпоследний столбец или на предпоследнюю строку последний столбец (рисунок 45).

Давайте разберем ее. Первая строка обозначает коэффициент корреляции, вторая – значение *sig*, третья – количество пациентов. Нам необходимо обратить внимание на значение *Sig* = 0,000, сравниваем его со значением 0,05 и делаем следующий вывод – существует 0. Далее необходимо дать ответ еще на два следующих вопроса.

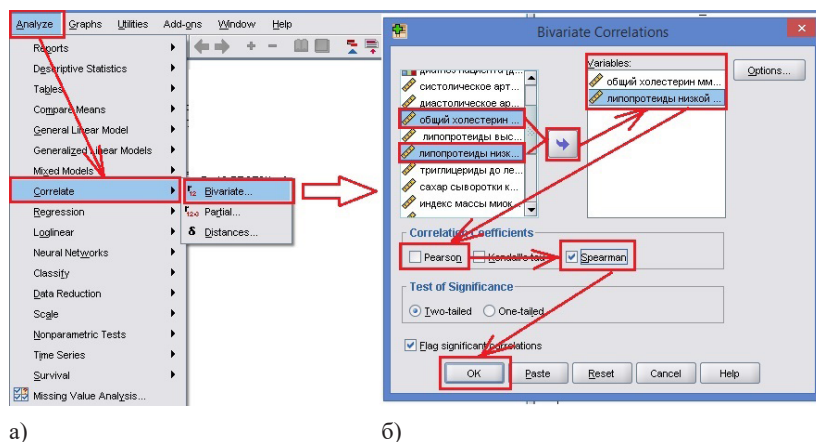


Рисунок 44 – Алгоритм (а) и рабочее окно (б) команды «Bivariate»

## Nonparametric Correlations

Correlations			общий холестерин ммоль/л до лечения	липопротеиды низкой плотности до лечения
Spearman's rho	общий холестерин ммоль/л до лечения	Correlation Coefficient	1,000	,720**
		Sig. (2-tailed)		,000
		N	65	65
	липопротеиды низкой плотности до лечения	Correlation Coefficient	,720**	1,000
		Sig. (2-tailed)	,000	
		N	65	65

\*\* . Correlation is significant at the 0.01 level (2-tailed).

Рисунок 45 – Таблица результатов вычисления команды «Bivariate»

Определим силу связи –  $r = 0,720$ , находится в диапазоне от 0,7 до 0,9, то есть связь сильная. Стоит обратить внимание на две звездочки около коэффициента корреляции – они говорят о том, что связь высокозначимая (1 звездочка – значимая).

Определим характер связи, так как коэффициент корреляции положительный, поэтому связь – прямая, то есть с увеличением значения одной переменной вторая тоже увеличивается и наоборот, но за счет чего происходит это увеличение или снижение, это история из другой области.

8. Мы разобрали полученную таблицу, осталось сформулировать ответ: в данном файле между показателями «общий холестерин до лечения» и «липопротеидами низкой плотности до лечения» у курящих пациентов существует статистически значимая связь.

**Пояснение 1.** Для расчета коэффициента корреляции был выбран критерий Спирмена, так как оба показателя не соответствуют нормальному распределению Гаусса.

**Пояснение 2.** Существует статистически значимая связь, так как  $\text{sig} = 0,000 < 0,05$ .

**Пояснение 3.** Связь сильная, потому что коэффициент корреляции ( $r$ ) находится в интервале от 0,7 до 0,9 и высокозначимая.

**Пояснение 4.** Связь прямая, так как коэффициент корреляции ( $r$ ) положительный. Это означает, что при увеличении одного показателя будем наблюдать увеличение другого.

Надеемся, вам было понятно, как проводить расчет корреляции, для переменных Scale – Scale.

**Корреляция по типу Nominal – Nominal, Ordinal – Ordinal, Nominal – Ordinal или Ordinal – Nominal**

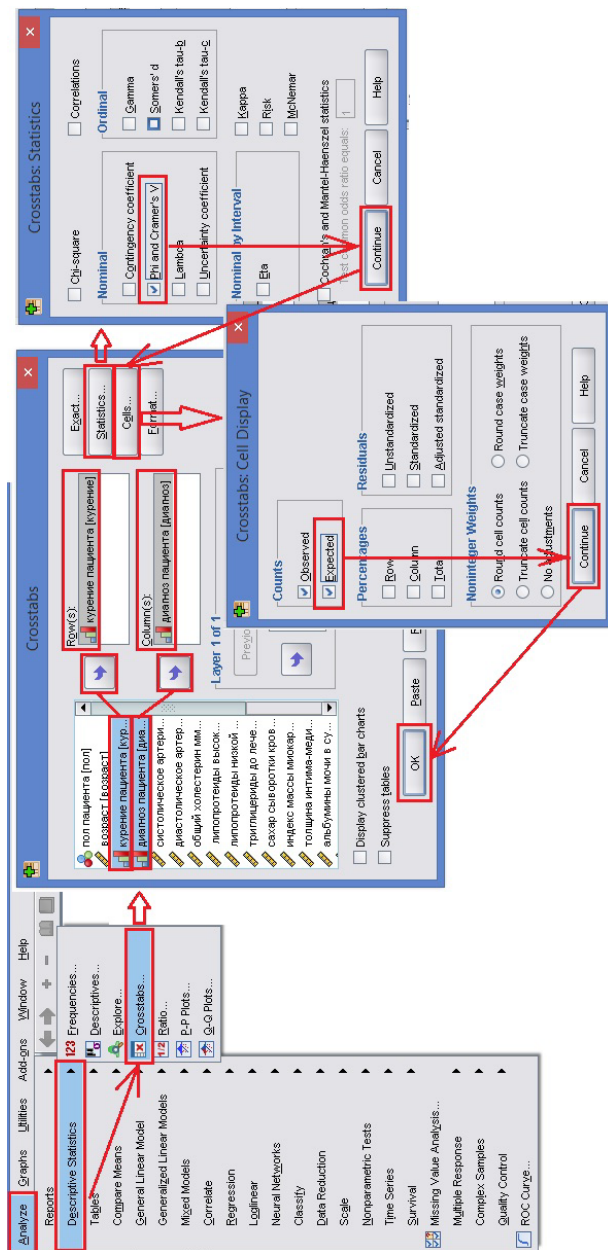
Давайте рассмотрим это на практике.

1. Открыть файл «Иванов\_ГБ.sav».
2. Поставить соответствующую **задачу**, например – Определить, есть ли связь между показателями «диагноз» и «курение».
3. В данном случае нам нет необходимости использовать фильтр, поэтому сразу перейдем к расчету корреляции.
4. Так как, наши переменные являются переменными *Ordinal* и *Nominal*, мы можем воспользоваться командой «Crosstabs», то есть заходим в меню

*Analyze → Descriptive Statistics → Crosstabs*

Появится диалоговое окно «Crosstabs» (рисунок 46, а). Переносим в правую часть в поле «Rows» переменную «курение», в поле «Columns» – переменную «диагноз». Далее нажимаем кнопку «Statistics» и добавляем галочку в позицию «Phi and Cramer's V» → «Continue», затем нажимаем кнопку «Cells», добавляем галочку в позицию «Expected», нажимаем кнопку «OK» (рисунок 46, б).

5. Получили три таблицы, обращаем внимание на размерность второй таблицы, в нашем случае переменная «курение» содержит два варианта ответа – курит и не курит, переменная «диагноз», содержит четыре варианта ответа – эссенциальная гипертензия I и II степени, реноваскулярная гипертензия и ренопривная гипертензия. Отсюда делается вывод, что размерность второй таблицы  $2 \times 4$ . Так как размерность  $2 \times 4$  смотрим строчку *Cramer's V* в третьей таблице (строчку *Phi* мы смотрим в одном единственном случае, если размерность второй таблицы  $2 \times 2$ ) (рисунок 47).



а)

б)

Рисунок 46 – Алгоритм (а) и рабочее окно (б) команды «Crosstabs»

## Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
диагноз пациента * курение пациента	100	100,0%	0	,0%	100	100,0%

диагноз пациента \* курение пациента Crosstabulation

			курение пациента		Total
			курит	не курит	
диагноз пациента	эссенциальная гипертензия 1-й степени	Count	19	6	25
		Expected Count	16,2	8,8	25,0
	эссенциальная гипертензия 2-й степени	Count	11	12	23
		Expected Count	15,0	8,0	23,0
	реноваскулярная гипертония	Count	20	8	28
		Expected Count	18,2	9,8	28,0
	ренопривная гипертония	Count	15	9	24
		Expected Count	15,6	8,4	24,0
Total	Count	65	35	100	
	Expected Count	65,0	35,0	100,0	

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,221	,180
	Cramer's V	,221	,180
N of Valid Cases		100	

Рисунок 47 – Таблицы результатов вычисления команды «Crosstabs»

Давайте разберем их. Во-первых, обращаем внимание на значение  $\text{sig} = 0,180 > 0,05$  – это означает, что не существует **статистически значимой связи**. Далее необходимо дать ответ еще на один вопрос.

6. Мы разобрали полученные таблицы, осталось сформулировать ответ: в данном файле между показателями «диагноз» и «курение» не существует статистически значимой связи.

**Пояснение 1.** Для расчета коэффициента корреляции был выбран критерий Ви-Крамера, потому что таблица  $2 \times 4$ .

**Пояснение 2.** Не существует статистически значимой связи, так как  $\text{sig} = 0,180 > 0,05$ .

**P.S.** Для чего мы ставили галочку в позицию *Expected*? Давайте посмотрим еще раз на вторую таблицу: *Count* – это реальное количество пациентов, *Expected Count* – это ожидаемое

количество пациентов при отсутствии связи. Чем больше различия между этими числами, тем связь сильнее.

Надеемся, вам было понятно, как проводить расчет корреляции для переменных Nominal – Nominal, Ordinal – Ordinal, Nominal – Ordinal или Ordinal – Nominal.

### **Корреляция по типу Scale – Nominal или Scale – Ordinal**

Корреляция по данному типу решается в два основных этапа:

1. Производим перекодировку переменной Scale в переменную Ordinal:

- a) находим min и max переменной scale;
- b) разбиваем полученный диапазон на ранги (чаще применяем диапазоны, принятые в медицине);
- c) проводим перекодировку;
- d) заполняем все необходимые поля для вновь созданной переменной.

2. Рассчитываем корреляцию по типу Ordinal – Nominal или Ordinal – Ordinal.

Давайте рассмотрим это на практике.

1. Открыть файл «Иванов\_ГБ.sav».
2. Поставить соответствующую **задачу**, например – Определить, есть ли связь между показателями «систолическое давление до лечения» и «диагноз».
3. Так как нам никого не надо отбирать в данной задаче, то сразу перейдем к I этапу.

**I этап.** Перекодируем переменную Scale – «систолического давления до лечения» в переменную Ordinal:

a) для начала необходимо найти min и max переменной «систолическое давление до лечения», то есть воспользуемся командой описательных статистик. Для тех, кто не помнит, это «Analyze – Descriptive Statistics – Descriptives». Переносим в правую часть переменную «систолическое давление до лечения» → «ОК». Получили  $min = 140$ ,  $max = 200$  (рисунок 48).

b) разбиваем полученный диапазон на ранги, то есть части. Для этого воспользуемся Всемирной паутиной – Internet и найдем классификацию по систолическому давлению (рисунок 49):

## Descriptives

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
систолическое артериальное давление до лечения	100	140	200	173,29	17,490
Valid N (listwise)	100				

Рисунок 48 – Таблица результатов вычисления команды «Descriptives»

Нормы артериального давления по классификации ВОЗ		
Артериальное давление (категория)	Верхнее артериальное давление (мм. рт. ст.)	Нижнее артериальное давление (мм. рт. ст.)
Гипотония (пониженное)	ниже 100	ниже 60
Оптимальное давление	100–119	60–79
Нормальное давление	120–129	80–84
Высокое нормальное давление	130–139	85–89
Умеренная гипертензия	140–159	90–99
Гипертензия средней тяжести	160–179	100–109
Тяжелая гипертензия	более 180	более 110

Рисунок 49 – Нормы артериального давления по классификации ВОЗ

Исходя из полученного нами min и max, мы имеем:

- 140–160 – умеренная гипертензия;
- 160–180 – гипертензия средней тяжести;
- 180 – 200 – тяжелая гипертензия. Тут подчеркивание важно, есть пояснения в тексте

При разбивке на диапазоны необходимо помнить следующее правило при использовании функции ***Recode into Different Variables*** – на какую цифру заканчивается первый диапазон, с той же и должен начинаться следующий. Так же стоит отметить, что верхнее значение не входит в интервал (кроме последнего);

с) проводим перекодировку, для этого воспользуемся командой «*Recode into Different Variables*», то есть заходим в меню

***Transform → Recode into Different Variables***

Появится диалоговое окно «Recode into Different Variables» (рисунок 50, а). Переносим в правую часть переменную «систолическое давление до лечения» и видим, что появилась надпись «сатдо →?»», это говорит о том, что программа не знает какое имя будет присвоено новой переменной. Это надо будет сделать самостоятельно, для этого в поле «Name» зададим имя для новой переменной – «кат\_сатдо» → «Change». И вот, знак вопроса изменился на «кат\_сатдо». Далее нажимаем кнопку «Old and New Values». Открывшееся окно визуально разделено на две части – левая часть «Old Value», то есть старые значения, правая часть – «New Value», то есть новые значения (рисунок 50, б).

В левой части окна ставим галочку в позицию *Range*, после этого стали активны два нижних поля под словом *Range*. В эти поля мы и будем вводить наши ранги из предыдущего пункта. В верхнее поле вводим цифру обозначающее начало первого диапазона – 140, в нижнее поле цифру, обозначающую конец первого диапазона – 160, этот ранг мы называли цифрой 1. Поэтому в правой части окна в поле *Value* ставим цифру 1 и далее нажимаем кнопку *Add* (добавить), то же самое надо проделать и с другим диапазоном. Как только введены все необходимые диапазоны, нажимаем кнопку «Continue», затем нажимаем кнопку «OK» (см. рисунок 50, б).

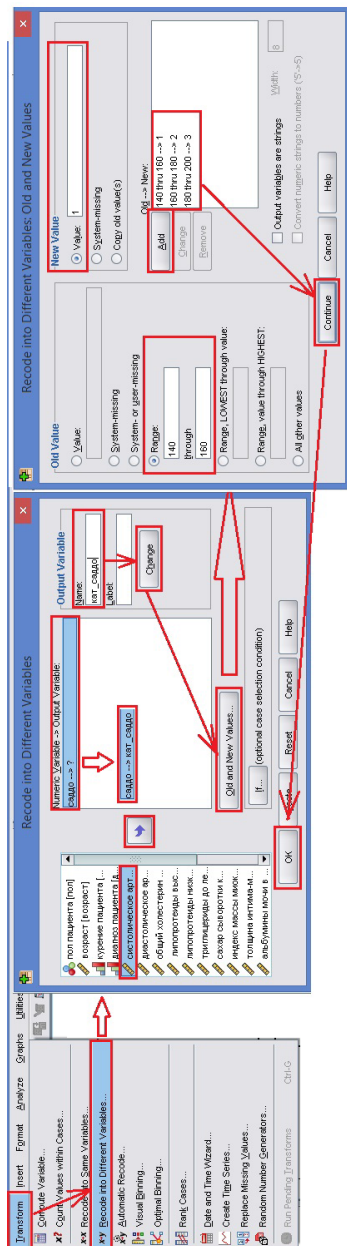
Мы получили следующую надпись (рисунок 51):

а) Теперь осталось заполнить для нашей новой созданной переменной все необходимые поля в базовом файле во вкладке «Variable View». Наша новая переменная находится в самом низу списка переменных, давайте ее заполним: Decimals = 0; Width = 1; Label – категория систолического давления до; Values – 1 – умеренная гипертензия; 2 – гипертензия средней тяжести; 3 – тяжелая гипертензия; Measure – Ordinal (рисунок 52);

б) На этом перекодировка считается завершенной.

**II этап.** Рассчитаем корреляцию по типу Ordinal – Ordinal.

Для этого воспользуемся командой «Crosstabs». В поле «Rows» переносим переменную «диагноз», в поле «Columns» – нашу новую переменную «кат\_сатдо». Нажимаем кнопку «Statistics»



a)

б)

Рисунок 50 – Алгоритм (а) и рабочее окно (б) команды «Recode into Different Variables»

RECODE систдо (140 thru 160=1) (160 thru 180=2) (180 thru 200=3) INTO кат\_систо.  
EXECUTE.

Рисунок 51 – Надпись, подтверждающая завершение команды перекодировки

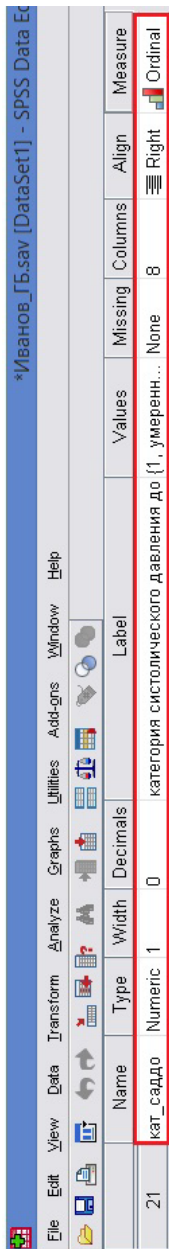


Рисунок 52 – Новая переменная кат\_систо, со всеми заполненными полями

и добавляем галочку в позицию «Phi and Cramer's V» → «Continue», затем нажимаем кнопку «Cells», добавляем галочку в позицию «Expected», нажимаем кнопку «OK».

4. Получили три таблицы, обращаем внимание на 3-ю, поэтому смотрим строчку «Cramer's V» в третьей таблице (рисунок 53).

#### Crosstabs

	Case Processing Summary					
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
диагноз пациента * категория систолического давления до	100	100,0%	0	,0%	100	100,0%

диагноз пациента * категория систолического давления до Crosstabulation				категория систолического давления до			Total
				умеренная гипертония	гипертония средней тяжести	тяжелая гипертония	
диагноз пациента	эссенциальная гипертония 1-й степени	Count		21	3	1	25
		Expected Count		9,2	7,8	8,0	25,0
	эссенциальная гипертония 2-й степени	Count		12	10	1	23
		Expected Count		8,5	7,1	7,4	23,0
	реноваскулярная гипертония	Count		3	8	17	28
		Expected Count		10,4	8,7	9,0	28,0
	реноприивная гипертония	Count		1	10	13	24
		Expected Count		8,9	7,4	7,7	24,0
Total		Count		37	31	32	100
		Expected Count		37,0	31,0	32,0	100,0

Symmetric Measures			
		Value	Approx. Sig.
Nominal by Nominal	Phi	,749	,000
	Cramer's V	,530	,000
N of Valid Cases		100	

Рисунок 53 – Таблицы результатов вычисления команды «Crosstabs»

Во-первых, обращаем внимание на значение  $\text{sig} = 0,000 < 0,05$  – это означает, что существует **статистически значимая связь**. Далее необходимо дать ответ еще на один вопрос.

Определим силу связи, в данном случае коэффициент корреляции находится в столбце *Value* и равен  $r = 0,530$ , находится в диапазоне от 0,5 до 0,7, то есть связь значительная.

5. Мы разобрали полученные таблицы, осталось сформулировать ответ: в данном файле между показателями «систолическое

*давление до лечения» и «диагноз» существует статистически значимая связь.*

**Пояснение 1.** Для расчета коэффициента корреляции был выбран критерий Ви-Крамера, потому что таблица не  $2 \times 2$ .

**Пояснение 2.** Существует статистически значимая связь, так как  $\text{sig} = 0,000 < 0,05$ .

**Пояснение 3.** Связь значительная, потому что коэффициент корреляции ( $r$ ) находится в интервале от 0,5 до 0,7.

**P.S.** Давайте обратим внимание на вторую таблицу. Если бы связи не было, то пациентов с эссенциальной гипертензией I степени и умеренной гипертонией было бы не 21, а 9; пациентов с ренопривной гипертонией и тяжелой гипертонией было бы не 13, а 8 и так далее, то есть чем больше различие, тем сильнее связь.

Надеемся, вам было понятно, как проводить расчет корреляции, для переменных Scale – Nominal, Scale – Ordinal.

## 6.4. Задачи для закрепления материала

- Определить, есть ли связь между показателями *«креатинин мочи»* и *«толщина интима-медиа»*.
- Определить, есть ли связь между показателями *«индекс массы миокарда левого желудочка до лечения»* и *«индекс массы миокарда левого желудочка после лечения»*.
- Определить, есть ли связь между показателями *«сахар сыворотки крови»* и *«креатинин сыворотки крови»* у курящих пациентов с диагнозом Реноваскулярная гипертония.
- Определить, если ли связь между показателями *«систолическое давление до лечения»* и *«общий холестерин до лечения»* у женщин старше 60 лет.
- Определить, есть ли связь между показателями *«индекс массы миокарда левого желудочка до лечения»* и *«толщина интима-медиа»* у мужчин с систолическим давлением до лечения выше 139.
- Определить, есть ли связь между показателями *«триглицериды до лечения»* и *«скорость клубочковой фильтрации»*.

- Определить, есть ли связь между показателями *«систолическое давление до лечения»* и *«триглицериды до лечения»*.
- Определить, есть ли связь между показателями *«креатинин мочи»* и *«возраст»* у некурящих мужчин.
- Определить, есть ли связь между показателями *«альбумины мочи в сутки»* и *«креатинин мочи»* у пациентов с диагнозом Эссенциальная гипертензия I степени.
- Определить, есть ли связь между показателями *«сахар сыворотки крови»* и *«общий холестерин до лечения»* у пациентов младше 50 лет.
- Определить, есть ли связь между показателями *«пол»* и *«курение»* у пациентов с диагнозом Эссенциальная гипертензия I степени.
- Определить, есть ли связь между показателями *«курение»* и *«диагноз»* у пациентов старше 55 лет.
- Определить, есть ли связь между показателями *«пол»* и *«курение»*.
- Определить, есть ли связь между показателями *«диагноз»* и *«пол»*, у пациентов с индексом массы миокарда левого желудочка до лечения  $> 200$ .
- Определить, есть ли связь между показателями *«курение»* и *«диагноз»*.
- Есть ли связь между показателями *«общий холестерин»* и *«курение»* у женщин.
- Есть ли связь между показателями *«альбумины мочи в сутки»* и *«пол»*.
- Есть ли связь между показателями *«индекс массы миокарда левого желудочка до лечения»* и *«диагноз»*.
- Есть ли связь между показателями *«триглицериды до лечения»* и *«диагноз»* у курящих.
- Есть ли связь между показателями *«креатинин сыворотки крови»* и *«пол»* у пациентов с диагнозом Эссенциальная гипертензия I степени.

## 6.5. Задание для самостоятельной работы

Для выполнения данного задания необходимо:

1. Поставить и решить по одной задаче на вашем созданном файле на тему «Корреляционные статистики» для всех трех вариантов (то есть с применением всех трех критериев), сохранить решение в файле выдачи. Сохраненный файл выдачи сдать преподавателю, ответить на его вопросы.

2. Продолжить заполнение листа СРС (форму СРС см. в приложении 2) по следующему образцу:

<b>Файл 5 – «Коэффициент корреляции»</b>			
<b>I тип Scale – Scale</b>			
<b>Задача.</b> Определить, есть ли связь между показателями « <i>общий холестерин до лечения</i> » и « <i>липопротеидами низкой плотности до лечения</i> » у курящих пациентов			
<b>Ответ:</b>			
Показатели	Критерий	r	Значимость
ОХСдо- ЛНПдо	Спирмена	0,720	p < 0,05
<b>Пояснение 1.</b> Для расчета коэффициента корреляции был выбран критерий Спирмена, потому что оба показателя не соответствуют нормальному распределению Гаусса			
<b>Пояснение 2.</b> Существует статистически значимая связь, так как sig = 0,000 < 0,05			
<b>Пояснение 3.</b> Связь сильная, потому что коэффициент корреляции (r) находится в интервале от 0,7 до 0,9 и высокосignificant			
<b>Пояснение 4.</b> Связь прямая, так как коэффициент корреляции (r) положительный. Это означает, что при повышении одного показателя будем наблюдать повышение другого			
<b>II тип Nominal – Nominal, Ordinal – Ordinal, Nominal – Ordinal или Ordinal – Nominal</b>			
<b>Задача.</b> Определить, есть ли связь между показателями «диагноз» и «курение»			
<b>Ответ:</b>			

Показатели	Критерий	г	Значимость
курение- диагноз	Ви-Крамера	0,221	p>0,05

**Пояснение 1.** Для расчета коэффициента корреляции был выбран критерий Ви-Крамера, потому что таблица  $2 \times 4$

**Пояснение 2.** Не существует статистически значимой связи, так как  $\text{sig} = 0,180 > 0,05$

### III тип Scale – Nominal, Scale – Ordinal

**Задача.** Определить, есть ли связь между показателями «систолическое давление до лечения» и «диагноз»

Категории АД	САД	Используя нормы САД, трансформируем показатель «САДдо» в показатель «кат_саддо» и найдем связь между «диагнозом» и «кат_саддо»
умеренная гипертония	140–160	
гипертония средней тяжести	160–180	
тяжелая гипертония	180–200	

**Ответ:**

Показатели	Критерий	г	Значимость
кат_саддо- диагноз	Ви-Крамера	0,530	p < 0,05

**Пояснение 1.** Для расчета коэффициента корреляции был выбран критерий Ви-Крамера, потому что таблица не  $2 \times 2$

**Пояснение 2.** Существует статистически значимая связь, так как  $\text{sig} = 0,000 < 0,05$

**Пояснение 3.** Связь значительная, потому что коэффициент корреляции (г) находится в интервале от 0,5 до 0,7

## 6.6. Тестовые вопросы

**1. Определите, что отражает корреляционная зависимость:**

- a) взаимосвязь между переменными;
- b) причинно-следственные связи;
- c) взаимосвязь между группами;
- d) причинно-следственный анализ;
- e) изменения между группами.

**2. Определите, как называется количественная мера взаимосвязи двух переменных:**

- a) коэффициент корреляции;
- b) корреляционный анализ;
- c) регрессионный анализ;
- d) коэффициент рождаемости;
- e) статистический анализ.

**3. Назовите тесты для расчета корреляционной зависимости:**

- a) тест Pearson;
- b) тест Spearman;
- c) тест Kendall's;
- d) тест Kolmagorov;
- e) тест Student.

**4. Определите, при каких условиях возможен поиск корреляционной зависимости:**

- a) только для связанных выборок;
- b) только для несвязанных выборок;
- c) вообще для любых выборок;
- d) вне зависимости от выборок;
- e) выборки тут не причем.

**5. Определите, на какие три главных вопроса дает ответ корреляционный анализ:**

- a) есть ли связь между переменными?
- b) какова сила связи между переменными?
- c) каков характер связи между переменными?

- d) какова причина связи между переменными?
- e) есть ли связь между группами переменных?

**6. Определите, в каком случае есть связь между переменными:**

- a) когда  $\text{sig} < 0,05$ ;
- b) когда  $\text{sig} > 0,05$ ;
- c) когда  $\text{sig} = 0,05$ ;
- d) когда  $\text{sig} \geq 0,05$ ;
- e) когда  $\text{sig} \leq 0,05$ .

**7. Определите, какой бывает характер связи:**

- a) прямой или обратный;
- b) положительный или отрицательный;
- c) прямой и положительный;
- d) обратный и отрицательный;
- e) сильный и слабый.

**8. Определите, какой тест необходимо применить для решения следующей задачи – Есть ли связь между возрастом и систолическим давлением, при условии, что обе переменных не подчиняются закону Гаусса:**

- a) тест Spearman;
- b) тест Kendall's;
- c) тест Pearson;
- d) тест Kolmagorov;
- e) тест Student.

**9. Определите, какой тест необходимо применить для решения следующей задачи – Есть ли связь между сахаром крови и гемоглобином, при условии, что обе переменных подчиняются закону Гаусса:**

- a) тест Pearson;
- b) тест Spearman;
- c) тест Kendall's;
- d) тест Kolmagorov;
- e) тест Student.

**10. Определите, какая задача является некорректной, то есть в ее решении смысла нет:**

- a) есть ли связь между показателями «систолическое артериальное давление» до лечения у мужчин и женщин;
- b) есть ли связь между показателями «диагноз» и «систолическое артериальное давление» до лечения;
- c) есть ли связь между индексом массы тела до лечения и полиурией до лечения, для больных с диагнозом Инсулинзависимый тип диабета;
- d) есть ли связь между лекарством и холестерином, исходное значение до лечения у пациентов старше 50 лет;
- e) есть ли связь между показателями «гемоглобин крови» до и после лечения у женщин.

**11. Определите, какая задача является некорректной, то есть в ее решении смысла нет:**

- a) есть ли связь между показателями «толщина интим-медиа» у курящих и некурящих пациентов;
- b) есть ли связь между типом диабета и глюкозой плазмы крови (натощак) до лечения;
- c) есть ли связь между показателями «липопротеиды высокой плотности» и «сахар крови» до лечения;
- d) есть ли связь между показателем «сахар крови» до лечения крови и типом диабета;
- e) есть ли связь между показателями СОЭ после лечения и весом пациентов старше 65 лет.

**12. Определите, какая задача является некорректной, то есть в ее решении смысла нет:**

- a) есть ли связь между показателями успеваемости у мальчиков и девочек, учащихся 11 А класса;
- b) есть ли связь между показателями «счет в уме» и «хобби» у девочек, учащихся в 11 А классе;
- c) есть ли связь между показателями успеваемости в 10 и 11 классах у мальчиков, выбравших хобби «спорт»;
- d) есть ли связь между показателями осведомленность и вузом у девочек, выбравших хобби «искусство»;

е) есть ли связь между показателями «кратковременная память» и «хобби» у учеников 11 А класса.

**13. Определите, какая задача является некорректной, то есть в ее решении смысла нет:**

а) есть ли связь между показателями «кратковременная память» и «пол» у учащихся 11 А и 11 Б классов;

б) есть ли связь между показателями «счет в уме» и «пол» у учеников 11 А класса;

с) есть ли связь между показателями «успеваемость» и «вуз» у девочек учащихся в 11 Б классе;

д) есть ли связь между показателями «числовые ряды» и «хобби» у мальчиков, выбравших технический вуз;

е) есть ли связь между показателями «кратковременная память» и «успеваемость» у учеников 11 А класса.

**14. Определите, какая задача является некорректной, то есть в ее решении смысла нет:**

- есть ли связь между показателями «креатинин сыворотки крови» у мужчин с легкой и тяжелой степенью заболевания;
- есть ли связь между полом и общим холестерином у пациентов со средней тяжестью заболевания старше 50 лет;
- есть ли связь между показателями «систолическое давление» до и после лечения у пациентов с тяжелой степенью заболевания;
- есть ли связь между показателями «курение» и «пол» у пациентов с легкой и тяжелой степенью заболевания младше 50 лет;
- есть ли связь между показателями «пол» и «диагноз» у женщин со средней тяжестью заболевания старше 70 лет.

**15. Определите, какая задача является некорректной, то есть в ее решении смысла нет:**

а) есть ли связь между показателями гликолизированного гемоглобина до лечения и сахаром крови после лечения у женщин;

б) есть ли связь между сахаром в крови, исходное значение до лечения и то же, через 12 месяцев, для больных принимавших бетасан;

с) есть ли связь между лекарством и систолическим артериальным давлением, исходное значение до лечения у мужчин;

д) есть ли связь между глюкозой плазмы крови и общим холестерином до лечения, для больных с диагнозом инсулиннезависимый тип диабета;

е) есть ли связь между лекарством и диастолическим давлением исходное значение до приема лекарств для больных принимавших альфасан.

**16. Укажите правильную последовательность при решении задачи: есть ли связь между систолическим артериальным давлением и уровнем холестерина:**

а) оба показателя проверить на соответствие закону Гаусса → если оба параметра соответствуют закону Гаусса, то запустим Bivariate и выберем тест «Pearson»; если хотя бы один не соответствует закону Гаусса, то выберем «Spearman»;

б) оба показателя проверить на соответствие закону Гаусса → если оба параметра соответствуют закону Гаусса, то запустим Bivariate и выберем тест «Spearman»; если хотя бы один не соответствует закону Гаусса, то выберем «Pearson»;

с) оба показателя проверить на соответствие закону Гаусса → если оба параметра соответствуют закону Гаусса, то запустим Bivariate и выберем тест «Kendall's»; если хотя бы один не соответствует закону Гаусса, то выберем «Pearson»;

д) оба показателя проверить на соответствие закону Гаусса → если оба параметра соответствуют закону Гаусса, то запустим Bivariate и выберем тест «Spearman»; если хотя бы один не соответствует закону Гаусса, то выберем «Kendall's»;

е) оба показателя проверить на соответствие закону Гаусса → если оба параметра соответствуют закону Гаусса, то запустить Crosstabs и выбрать тест «Phi and Cramer's V».

**17. Укажите правильную последовательность при решении задачи: есть ли связь между диагнозом и курением:**

а) запустить Crosstabs → в Statistics выбрать тест «Phi and Cramer's V» → в Cells поставить галочку «Expected»;

б) оба показателя проверить на соответствие закону Гаусса → запустить Crosstabs → в Statistics выбрать тест «Phi and Cramer's V»;

с) оба показателя проверить на соответствие закону Гаусса → если оба параметра соответствуют закону Гаусса, то запустим Bivariate и выберем тест «Spearman»;

д) оба показателя проверить на соответствие закону Гаусса → запустить Crosstabs → в Cells поставить галочку «Expected»;

е) оба показателя проверить на соответствие закону Гаусса → если оба параметра соответствуют закону Гаусса, то запустим Bivariate и выберем тест «Pearson».

**18. Укажите правильную последовательность при решении задачи: есть ли связь между индексом массы тела и типом диабета:**

а) с помощью Transform into different variables параметр индекс массы тела превратить из Scale в Ordinal → запустить Crosstabs → в Statistics выбрать тест «Phi and Cramer's V» → в Cells поставить галочку «Expected»;

б) с помощью Transform into different variables параметр тип диабета превратить из Nominal в Scale → запустить Crosstabs → в Statistics выбрать тест «Phi and Cramer's V» → в Cells поставить галочку «Expected»;

с) оба показателя проверить на соответствие закону Гаусса → если оба параметра соответствуют закону Гаусса, то запустим Bivariate и выберем тест «Spearman»; если хотя бы один не соответствует закону Гаусса, то выберем «Pearson»

д) оба показателя проверить на соответствие закону Гаусса → если оба параметра соответствуют закону Гаусса, то запустим Bivariate и выберем тест «Pearson»; если хотя бы один не соответствует закону Гаусса, то выберем «Spearman»;

е) оба показателя проверить на соответствие закону Гаусса → если оба параметра соответствуют закону Гаусса, то запустим Bivariate и выберем тест «Spearman»; если хотя бы один не соответствует закону Гаусса, то выберем «Kendall's».

## 19. Выберите правильное суждение:

Correlations			
		систолическое артериальное давление до лечения	общий холестерин ммоль/л до лечения
→ систолическое артериальное давление до лечения	Pearson Correlation	1	,425**
	Sig. (2-tailed)		,000
	N	100	100
общий холестерин ммоль/л до лечения	Pearson Correlation	,425**	1
	Sig. (2-tailed)	,000	
	N	100	100
**. Correlation is significant at the 0.01 level (2-tailed).			

а) связь между систолическим артериальным давлением и общим холестерином до лечения есть, она умеренная, высокодостоверная и прямая;

б) связь между систолическим артериальным давлением и общим холестерином до лечения есть, она сильная, достоверная и обратная;

в) связи между систолическим артериальным давлением и общим холестерином до лечения нет, она функциональная и прямая;

г) связь между систолическим артериальным давлением и общим холестерином до лечения есть, она значительная, высокодостоверная и обратная;

д) связь между систолическим артериальным давлением и общим холестерином до лечения нет, она значительная, высокодостоверная и обратная.

## 20. Выберите правильное суждение:

Correlations			
		общий холестерин ммоль/л до лечения	возраст
→ общий холестерин ммоль/л до лечения	Pearson Correlation	1	,137
	Sig. (2-tailed)		,174
	N	100	100
возраст	Pearson Correlation	,137	1
	Sig. (2-tailed)	,174	
	N	100	100

а) связи между общим холестерином до лечения и возрастом нет, так как  $\text{sig} = 0,174$ , это больше  $0,05$ ;

б) связь между общим холестерином до лечения и возрастом есть, она сильная, прямая,  $\text{sig} = 0,174$ , это меньше  $0,05$ ;

с) связи между общим холестерином до лечения и возрастом нет, она слабая, прямая,  $\text{sig} = 0,174$ , это меньше  $0,05$ ;

д) связь между общим холестерином до лечения и возрастом есть, умеренная, достоверная, обратная,  $\text{sig} = 0,174$ , это больше  $0,05$ ;

е) связь между общим холестерином до лечения и возрастом есть, она слабая, обратная,  $\text{sig} = 0,137$ , это больше  $0,05$ .

## 21. Выберите правильное суждение:

Correlations

			общий холестерин ммоль/л до лечения	липопротеиды высокой плотности до лечения
Spearman's rho	общий холестерин ммоль/л до лечения	Correlation Coefficient	1,000	-,484**
		Sig. (2-tailed)	.	,000
		N	100	100
	липопротеиды высокой плотности до лечения	Correlation Coefficient	-,484**	1,000
		Sig. (2-tailed)	,000	.
		N	100	100

\*\*. Correlation is significant at the 0.01 level (2-tailed).

а) связь между общим холестерином и липопротеидами высокой плотности до лечения есть, она умеренная, высокодостоверная и обратная;

б) связь между общим холестерином и липопротеидами высокой плотности до лечения есть, она сильная, достоверная, обратная;

с) связи между общим холестерином и липопротеидами высокой плотности до лечения нет, она умеренная, достоверная и прямая;

д) связи между общим холестерином и липопротеидами высокой плотности до лечения нет, она умеренная, высокодостоверная и обратная;

е) связь между общим холестерином и липопротеидами высокой плотности до лечения есть, она значительная, достоверная, обратная.

## 22. Выберите правильное суждение:

Correlations				
			систолическое артериальное давление до лечения	возраст
→ Spearman's rho	систолическое артериальное давление до лечения	Correlation Coefficient	1,000	0,430 <sup>*</sup>
		Sig. (2-tailed)	.	,043
		N	100	100
	возраст	Correlation Coefficient	0,430 <sup>*</sup>	1,000
		Sig. (2-tailed)	,043	.
		N	100	100

а) связь между систолическим артериальным давлением и возрастом есть, она умеренная, достоверная и прямая. Давление зависит от возраста;

б) связи между систолическим артериальным давлением и возрастом нет, она умеренная, достоверная и прямая. Давление не зависит от возраста;

в) связь между систолическим артериальным давлением и возрастом есть, она умеренная, высокодостоверная и обратная. Возраст зависит от давления;

г) связь между систолическим артериальным давлением и возрастом есть, она умеренная, достоверная и обратная. Давление зависит от возраста;

е) связь между систолическим артериальным давлением и возрастом есть, она умеренная, достоверная и прямая. Давление не зависит от возраста.

## СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. URL: [http://it-inform.narod.ru/index/informazija\\_i\\_ee\\_svojstva/0-2](http://it-inform.narod.ru/index/informazija_i_ee_svojstva/0-2)
2. URL: <https://forpsy.ru/works/kurosovaya/analiz-sposobov-rasprostraneniya-farmatsevticheskoy-informatsii-v-professionalnom-soobschestve/>
3. URL: <https://dic.academic.ru/dic.nsf/ruwiki/641511>
4. URL: [https://www.bsmu.by/downloads/kafedri/k\\_fiziki/2015-1/inf.pdf](https://www.bsmu.by/downloads/kafedri/k_fiziki/2015-1/inf.pdf)
5. URL: [https://ozlib.com/888929/meditsina/informatsiya\\_meditsine\\_vidy\\_osobennosti](https://ozlib.com/888929/meditsina/informatsiya_meditsine_vidy_osobennosti)
6. URL: [https://bstudy.net/1001872/meditsina/ponyatie\\_meditsinskoy\\_informatsii](https://bstudy.net/1001872/meditsina/ponyatie_meditsinskoy_informatsii)
7. URL: <http://www.dreamsmedic.com/informacionnie-stati/informatizatsiya-zdravooohraneniya.html>
8. *Чернов В.И.* Медицинская информатика: учеб. пособие / В.И. Чернов [и др.]. – Ростов н/Д: Феникс, 2007. – 320 с.
9. *Лобанова Е.Е.* Медицинская статистика: учебное пособие для студентов факультета среднего профессионального образования / Е.Е. Лобанова, А.В. Кочубей, Э.С. Антипенко [и др.]. – М.: МГМСУ, 2015. – 128 с.
10. URL: <https://www.bsmu.by/medicaljournal/242768ef4d4f96a161886c53b487037d/>
11. URL: <https://www.datuapstrade.lv/rus/spss/>
12. *Буланов Н.М.* Основные принципы применения описательной статистики в медицинских исследованиях / Н.М. Буланов, А.Ю. Суворов, О.Б. Блюсс [и др.] // Сеченовский вестник. – 2021. – Т. 12. – № 3. – С. 4–16.
13. URL: <https://sibac.info/blog/metody-opisatelnoy-statistiki>
14. *Кочетов А.Г.* Методы статистической обработки медицинских данных: методические рекомендации для ординаторов

и аспирантов медицинских учебных заведений, научных работников / сост.: А.Г. Кочетов, О.В. Лянг, В.П. Масенко [и др.]. – М.: РКНПК, 2012. – 42 с.

15. *Новикова С.* Социологические и психологические методы исследований в социальной работе: учебное пособие / С. Новикова, А. Соловьев. – М., 2016. – 496 с.
16. URL: <https://hr-portal.ru/spss/Glava11/Index6.php>
17. URL: <https://medstatistic.ru/>

### РЕФЕРАТ СРС

Артериальная гипертензия (АГ); гипертония (др.-греч. ὑπέρ – «над, выше» + τόνος «натяжение; тон») – синдром повышения систолического артериального давления (АД) (САД)  $\geq 140$  мм рт. ст. и/или диастолического АД (ДАД)  $\geq 90$  мм рт. ст. [1]. Эссенциальная гипертензия (гипертоническая болезнь) составляет 90–95 % случаев гипертонии. В остальных случаях диагностируют вторичные, симптоматические артериальные гипертензии: почечные [2] – 3–4 %, эндокринные – 0,1–0,3 %, гемодинамические, неврологические, стрессовые, обусловленные приёмом некоторых веществ (ятрогенные) и АГ беременных, при которых повышение давления крови является одним из симптомов основного заболевания.

Артериальная гипертензия – одно из самых распространённых заболеваний сердечно-сосудистой системы. Установлено, что артериальной гипертонией страдают 20–30 % взрослого населения. С возрастом распространённость болезни увеличивается и достигает 50–65 % у лиц старше 65 лет [2].

Выделяют ряд факторов риска распространения артериальной гипертонии: возраст, малоподвижный образ жизни, атеросклероз сосудов, злоупотребление алкоголем, курение, сахарный диабет, ожирение и др. [3].

Повышение АД обусловлено нарушением факторов, регулирующих деятельность сердечно-сосудистой системы. Первичным считается фактор наследственной предрасположенности. Он заключается в распространенных нарушениях ионтранспортной функции и структуры цитоплазматической мембраны клеток. В этих условиях сохранность специфической функции клеток обеспечивается механизмом клеточной адаптации, связанным с регуляцией кальциевого обмена, изменением гормонально-клеточных взаимоотношений, ростом активности нейрогуморальных

систем (гипоталамо-гипофизарно-надпочечниковой, ренин-ангиотензин-альдостероновой, инсулярной).

Кальциевая перегрузка клетки увеличивает сократительный потенциал гладких мышц сосудов и активирует клеточные факторы роста (протоонкогены). Происходящая при этом гипертрофия и гиперплазия гладких мышц сосудов и сердца ведёт к реконструкции сердца (гипертрофия) [4] и сосудов (повышенная сократимость, утолщение стенки и сужение просвета) [4], которые, являясь адаптивными, одновременно поддерживают гипертензию. Повышенное АД ведёт к возрастанию левожелудочкового систолического давления, увеличению напряжения (и гипертрофии) желудочка, возрастанию степени повреждения миокарда свободнорадикальным окислением.

Данному заболеванию присущи следующие переменные: систолическое (САД) и диастолическое (ДАД) артериальное давление, общий холестерин (ОХ), липопротеиды низкой плотности (ЛПНП), липопротеиды высокой плотности (ЛПВП), триглицериды (ТГ), сахар крови, индекс массы миокарда левого желудочка (ИММЛЖ), толщина интим-медиа (ТИМ), альбумины мочи (АМ), креатинин мочи (КМ), скорость клубочковой фильтрации (СКФ) и другие [5].

## Приложение 2

### ФОРМА СРС

КЫРГЫЗСКО-РОССИЙСКИЙ СЛАВЯНСКИЙ УНИВЕРСИТЕТ

МЕДИЦИНСКИЙ ФАКУЛЬТЕТ

Кафедра физики медицинской информатики и биологии

### ОТЧЕТ

о выполнении СРС по медицинской информатике

Тема СРС:

---

Ф.И.О. студента: \_\_\_\_\_

Группа: \_\_\_\_\_

Преподаватель: \_\_\_\_\_

Бишкек 20\_\_

ФАЙЛ 1 – база пациентов
Описание:
Источники:

Параметры				
№	Name	Value	Норма	Measure
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				

<b>Файл 2</b> – задачи «Сколько» и «Описательные статистики»
<b>Кого и сколько в данном файле имеется?</b>
<i>Задача 1:</i>
<i>Ответ:</i>
<i>Задача 2:</i>
<i>Ответ:</i>
<b>Описательные статистики 1</b>
<i>Задача:</i>
<i>Ответ:</i>
<b>Описательные статистики 2</b>
<i>Задача:</i>
<i>Ответ:</i>

<b>Файл 3</b> – сравнение средних со связанными выборками			
<b>Сравнение средних 1</b>			
<i>Задача:</i>			
<i>Ответ:</i>			
		Критерий	Значимость
<i>Пояснение 1.</i>			
<i>Пояснение 2.</i>			
<i>Пояснение 3.</i>			
<b>Сравнение средних 2</b>			
<i>Задача:</i>			
<i>Ответ:</i>			
		Критерий	Значимость
<i>Пояснение 1.</i>			
<i>Пояснение 2.</i>			
<i>Пояснение 3.</i>			

<b>Файл 4 – сравнение средних с несвязанными выборками</b>			
<b>Сравнение средних 1</b>			
<b>Задача:</b>			
<b>Ответ:</b>			
у 1-й выборки	у 2-й выборки	Критерий	Значимость
Пояснение 1.			
Пояснение 2.			
Пояснение 3.			
<b>Сравнение средних 2</b>			
<b>Задача:</b>			
<b>Ответ:</b>			
у 1-й выборки	у 2-й выборки	Критерий	Значимость
Пояснение 1.			
Пояснение 2.			
Пояснение 3.			

<b>Файл 5 – коэффициент корреляции</b>			
<b>I тип (scale – scale)</b>			
<b>Задача:</b>			
<b>Ответ:</b>			
Показатели	Критерий	r	Значимость
Пояснение 1.			
Пояснение 2.			
Пояснение 3.			
Пояснение 4.			
<b>II тип (nominal – nominal, ordinal – ordinal, nominal – ordinal)</b>			
<b>Задача:</b>			
<b>Ответ:</b>			
Показатели	Критерий	r	Значимость

<i>Пояснение 1.</i>			
<i>Пояснение 2.</i>			
<i>Пояснение 3.</i>			
<i>Пояснение 4.</i>			
<b>III тип (scale – nominal, scale – ordinal)</b>			
<b>Задача:</b>			
Категории	Показатель scale		
<b>Ответ:</b>			
Показатели	Критерий	г	Значимость
<i>Пояснение 1.</i>			
<i>Пояснение 2.</i>			
<i>Пояснение 3.</i>			
<i>Пояснение 4.</i>			

Авторы-составители:

*Елена Игоревна Кондратьева,  
Юрий Мансурович Курманбаекеев,  
Александр Анатольевич Сорокин*

МЕДИЦИНСКАЯ ИНФОРМАТИКА  
И МЕДИЦИНСКАЯ СТАТИСТИКА

Учебно-методическое пособие

Редактор *Н.В. Шумкина*  
Компьютерная верстка *М.Р. Фазлыевой*

Подписано в печать 06.02.2025.

Формат 60×84  $\frac{1}{16}$ . Офсетная печать.

Объем 11,25 п. л. Тираж 100 экз. Заказ 82.

Издательский дом КРСУ  
720000, г. Бишкек, ул. Киевская, 44