МИНИСТЕРСТВО ЗДРАВООХРАНЕНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

УЧРЕЖДЕНИЕ ОБРАЗОВАНИЯ
«ГОМЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ МЕДИЦИНСКИЙ УНИВЕРСИТЕТ»

КАФЕДРА ПАТОЛОГИЧЕСКОЙ ФИЗИОЛОГИИ

Ж. А. Чубуков, Т. С. Угольник

ОПИСАТЕЛЬНАЯ СТАТИСТИКА

Учебно-методическое пособие
для студентов всех факультетов медицинских вузов, аспирантов, магистрантов, соискателей и преподавателей

Гомель
ГомГМУ
2012

УДК 31:61(072)
ББК 60.6я7
Ч 81

Рецензент:
кандидат биологических наук, доцент,
заведующий кафедрой медицинской и биологической физики
Гомельского государственного медицинского университета
В. А. Игнатенко

Чубуков, Ж. А.
Ч 81 Описательная статистика: учебно-методическое пособие для студентов всех факультетов медицинских вузов, аспирантов, магистрантов, соискателей и преподавателей / Ж. А. Чубуков, Т. С. Угольник. — Гомель: учреждение образования «Гомельский государственный медицинский университет», 2012. — 28 с.
ISBN 978-985-506-395-8

В учебно-методическом пособии представлен материал о методах и критериях описательной статистики в контексте изучения медико-биологических дисциплин. Предназначено для студентов всех факультетов медицинских вузов, аспирантов, магистрантов, соискателей и преподавателей.

Утверждено и рекомендовано к изданию Центральным учебным научно-методическим советом учреждения образования «Гомельский государственный медицинский университет» 28 июня 2011 г., протокол № 7.

УДК 31:61(072)
ББК 60.6я7

ISBN 978-985-506-395-8

Оглавление

ОСНОВЫ ДОКАЗАТЕЛЬНОЙ МЕДИЦИНЫ И МЕДИКО-БИОЛОГИЧЕСКОЙ СТАТИСТИКИ

Любой специалист медико-биологического профиля в силу специфики профессии является исследователем и обязан непрерывно обучаться и совершенствоваться. Медицина испытывает острую потребность во внедрении инновационных технологий. В данной сфере деятельности предложение значительно превышает спрос, а разработка и внедрение той или иной технологии сопряжены как с риском для пациентов, так и со значительными финансовыми затратами для учреждений.

Медицинская технология — это система взаимосвязанных необходимых и достаточных научно обоснованных профилактических, диагностических, лечебных и реабилитационных мероприятий, выполнение которых позволяет наиболее рациональным образом обеспечить достижение максимального соответствия научно прогнозируемых результатов реальным при минимизации затрат.

Принципы медицинской технологии:

- Объективность создания и совершенствования. Медицинская технология может базироваться только на знаниях, полученных научным путем.
- Целесообразность. Медицинская технология должна иметь цель и критерии ее достижения.
- Дуализм. Медицинскую технологию следует рассматривать во всей полноте взаимосвязи субъективного и объективного.
- Эмерджентность. Внедрение медицинской технологии придает новое качество всей системе медицинского учреждения.
- Эмпирический контроль. Только результаты практического использования могут служить критерием действенности медицинской технологии.
- Мультивариантность. Технологическая задача не может иметь только одно решение.
- Оптимум соотношения цена/качество.
- Композиционность. Анализ и синтез являются неотъемлемыми составляющими разработки и внедрения технологии.
- Принцип разложения на части.
- Открытость. В любую медицинскую технологию возможно внесение уточнений и дополнений.
- Адаптивность. Медицинская технология должна иметь возможность приведения ее к конкретным условиям, если это не противоречит ее сущности.

На всех уровнях принятие решений неизбежно требует объективного обоснования выбора. В ситуации, когда риску подвергается благополучие человека недопустимо рассуждать субъективно, поэтому эффективность каждой медицинской технологии, как на этапе разработки, так и при внедрении должна рассматриваться критично и на основе доказанных сведений.

Оценка технологии относится к области стратегических исследований, призванных обеспечить лиц, принимающих решения, необходимой информацией о возможных последствиях внедрения новой технологии или модификации старой. Она включает выявление как прямых, так и косвенных эффектов, как благоприятных, так и неблагоприятных.

На практике не существует идеальных технологий, но среди доступных разумно выбирать ту, которая обладает достаточной доказательной базой. Таким образом, медицина нуждается в объективных доказательствах, поэтому всякий эффект при выборе медицинской технологии следует принимать как количественный параметр, а всякое доказательство эффективности выстраивать с позиций математики.

Доказательная медицина («медицина, основанная на доказательствах» от *англ.* evidence based medicine, EBM) — это сознательное, четкое и беспристрастное усиление навыков клинициста в диагностике, лечении и профилактике, а так же других областях путем систематического формулирования вопросов и применения математических оценок вероятности и риска но основе имеющихся доказанных сведений.

Основные постулаты доказательной медицины:

- каждое решение врача должно основываться на научных данных;
- вес каждого факта тем больше, чем строже методика научного исследования, в ходе которого он был получен.

Клиническая эпидемиология — наука, разрабатывающая методы клинических исследований, которые дают возможность делать обоснованные заключения, сводя к минимуму влияние систематических и случайных ошибок на результаты исследований. Клиническая эпидемиология является методической основой доказательной медицины.

Основные положения клинической эпидемиологии:

- диагноз, прогноз и результаты лечения для конкретного пациента не могут быть определены однозначно, и поэтому следует прибегать к вероятностным методам оценки данных;
- вероятность для каждого частного случая оценивается в контексте взаимосвязи с результатами предыдущих исследований, которые были получены на репрезентативных выборках;
- клинические исследования подвержены случайным и систематическим ошибкам, что следует учитывать при экстраполяции результатов на конкретный частный случай;
- исследования должны выполняться так, чтобы максимально снизить количество систематических и случайных ошибок;
- систематическая ошибка зависит от дизайна исследования;
- случайные ошибки оцениваются количественно при анализе результатов с использованием методов медико-биологической статистики.

Таким образом, клиническая эпидемиология позволяет произвести экстраполяцию полученных в исследовании результатов на уровень повседневной практической деятельности специалистов медико-биологического профиля.

Медико-биологическая статистика является ведущим инструментом клинической эпидемиологии и представляет собой прикладную интегральную науку на стыке математической статистики, кибернетики и биологии. Необходимость существования данной науки как самостоятельного направления обусловлена как спецификой потребностей отрасли, ее кадровым составом, так и особенностями изучаемых систем, объектов и явлений. Математическая статистика обеспечивает взаимодействие и преемственность между прикладными и фундаментальными исследованиями различных направлений. Медико-биологическая статистика не требует от исследователя глубинного знания высшей математики, но обязует его к пониманию методов и критериев математической статистики, а так же условий и ограничений использования этих методов.

При изучении биологических объектов и явлений следует помнить, что каждый частный случай является динамически изменяющимся элементом сверхсложной системы со многими неизвестными, и в тоже время сам фактически является системой. Описать подобную систему во всей полноте взаимосвязей на текущем уровне технического развития не представляется возможным. Тем не менее, всякое исследование прямо или косвенно направлено на выявление, изучение и практическое использование особенностей и закономерностей объектов или явлений. Композиционность медицинских технологий обеспечивается единством анализа и синтеза. Таким образом, для эффективной разработки и внедрения медицинской технологии современные специалисты медико-биологического профиля неизбежно в той или иной мере обращаются к кибернетике.

Кибернетика — это фундаментальная теоретическая наука о связи и управлении в биологических и небиологических системах.

Медицинская кибернетика является прикладной дисциплиной, которая изучает связь и управление в системах в контексте потребностей медицинской отрасли.

Современная техника позволяет специалистам медико-биологического профиля не производить рутинные вычисления вручную. Но всякое исследование оригинально, поэтому не может существовать универсального программного пакета, идеально отвечающего потребностям исследователя. Специалист по медико-биологической статистике принимает участие на всех этапах разработки, внедрения и сопровождения медицинской технологии и, подобно другим сотрудникам коллектива, осуществляет выбор «инструментов» и методов для достижения поставленных задач с учетом возможностей и потребностей коллег.

ТЕОРИЯ ВЕРОЯТНОСТИ И МЕДИКО-БИОЛОГИЧЕСКАЯ СТАТИСТИКА

Медико-биологическая статистика хотя и оперирует методами и терминологией математической теории вероятности, но отличается свой узкой прикладной направленностью. Теория вероятностей как один из разделов высшей математики (как, впрочем, и любой другой ее раздел) стремится максимально абстрагироваться от параллелей с реальной жизнью. В высшей математике, по мнению многих авторов, приближенные вычисления не являются приемлемыми и допустимыми. Всякая теорема доказывается «идеально». Только такой подход позволяет обеспечить развитие математики как фундаментальной науки.

Прикладные интегральные науки, основанные на математической теории вероятности (в том числе и медико-биологическая статистика) широко используют приближенные вычисления, аппроксимацию и интерполяцию. Методы и критерии прикладных статистических наук, хотя и имеют сходный смысл, но могут значительно различаться в возможностях и ограничениях при использовании в различных отраслях. Так, экономическая статистика и прогнозирование, как правило, оперируют большими и средними выборками (тысячи и десятки тысяч случаев наблюдений) количественных данных, распределение которых поддается корректной оценке, аппроксимации и преобразованию к нормальному закону, что дает возможности для широкого применения параметрических методов. Сходная ситуация наблюдается и в технических науках.

Исследователь медико-биологического профиля, как правило, работает с малыми выборками категориальных и дискретных данных, что обязывает к использованию непараметрических методов и критериев с соответствующими поправками на размер выборки и количество проверок гипотезы. На данный момент количество специфических методов медико-биологической статистики настолько возросло, что без соответствующей переподготовки специалист-математик не сможет проводить анализ медико-биологических данных. Специалист же медицинского профиля без освоения азов математических знаний и теории вероятности будет испытывать трудности в понимании сути статистических методов и интерпретации результатов. Кроме того, из-за значительных различий в программах подготовки специалистов медицинского и технического профилей, а также широкого использования терминологии несомненно будут возникать трудности коммуникативного характера: медик не может четко поставить задачу математику, а математик — не может в полной мере предоставить информацию о ходе анализа и его результатах медику в понятной форме.

Таким образом, медико-биологическая статистика не является статистикой в строгом понимании этого слова, но является самостоятельной областью знаний, интегральной наукой со своей специфической терминологией, методологией и областью применения.

ТИПЫ ДАННЫХ

Как правило, основной задачей любого эксперимента является получение информации об изучаемых объектах и/или явлениях. Саму суть понятия «информация» достаточно сложно сформулировать, хотя бы из-за того, что любое определение данного понятия будет являться тавтологией по своей сути. Кроме того, еще А. Н. Колмогоров, стоявший у истоков создания теории информации, задавался вопросом о том, существует ли информация независимо от ее восприятия или определяется индивидуальными особенностями исследователя. Таким образом, возникает противоречие, которое можно лишь отчасти разрешить, если описывать информацию как потенциальное свойство.

Информация — это потенциальные свойства некого объекта или системы, которые доступны для хранения, передачи, преобразования и выявления при изучении разумным существом.

Данные — это информация, представленная в формализованном виде.

Формализация данных может достигаться различными методами.

Тип данных — это метод формализации, который определяется сущностью изучаемого параметра.

Переменная — это совокупность первичных сигналов, содержащая данные определенного типа, доступные для изменения.

Важность понимания различий типов данных обусловлена потребностью исследователя в их анализе и последующем осмыслении результатов. Различия в типах данных являются отправной точкой для выбора математических методов, которые должны применяться для анализа результатов эксперимента или наблюдения, а также для правомочности использования того или иного способа представления данных (рисунок 1).

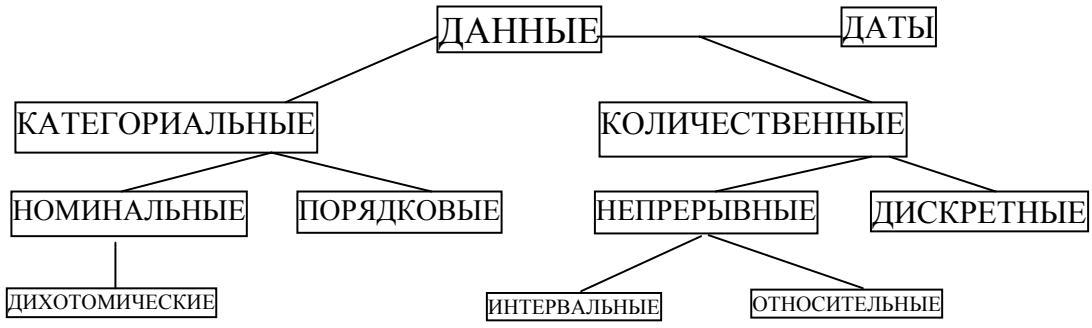


Рисунок 1 — Различные типы данных

Все переменные и результирующие показатели можно подразделить на два типа: категориальные и количественные. Дата и время представляют собой не абсолютно формализованную, но необходимую информацию, которая в зависимости от масштабов оценки может служить источником

переменных различных типов. Поэтому в структуре типов данных дата и время будут располагаться вне рассматриваемой классификации.

Категориальные (качественные) данные встречаются, когда объект изучения может принадлежать лишь к одной из взаимоисключающих (альтернативных) категорий.

• *Порядковые данные* — это категориальные данные, поддающиеся логическому упорядочению.

• *Номинальные данные* — это категориальные данные, не поддающиеся логическому упорядочению.

а) Бинарные данные — это номинальные данные, которые можно описать с использованием одной из двух альтернативных категорий.

Количественные данные — это данные, которые можно описать с использованием числового значения.

• *Дискретные данные* — это количественные данные, которые можно описать с идеальной точностью.

• *Непрерывные данные* — это количественные данные, которые можно описать с точностью, которая была достигнута при измерении на непрерывной шкале.

а) Интервальные данные — это непрерывные данные о величинах, имеющих физический смысл.

б) Относительные данные — это непрерывные данные о безразмерных величинах.

Следует отметить, что определение типов данных для переменных следует проводить до начала сбора информации об изучаемых системах, объектах или явлениях. Это позволяет улучшить дизайн исследования, что значительно снижает вероятность возникновения систематических ошибок. Кроме того, при разработке и внедрении масштабных медицинских технологий, работа осуществляется с гигантскими массивами данных, хранение и анализ которых требуют значительных вычислительных ресурсов.

Корректный выбор типов данных позволяет значительно сократить потребление вычислительных ресурсов, как на этапе разработки баз данных, так и на этапе их анализа. ЭВМ все расчеты, со всеми типами данных выполняет на уровне элементарных логических операций с бинарными значениями. На всех этапах работы с переменной происходит преобразование содержащейся в ней информации в бинарные данные и обратно. С точки зрения машины, логическая «1» («истина»), число «1» и символ «1» являются совершенно различной информацией, на работу с которой требуются различные вычислительные ресурсы. Правильный выбор типа данных позволяет оптимизировать производительность ЭВМ и сократить потребление дискового пространства для хранения данных. В несение информации в базу данных должно производиться с использованием программного обеспечения, которое исключает саму возможность присвоения переменным значений, не относящихся к заранее определенному типу данных.

ВАРИАЦИОННЫЙ РЯД И РАСПРЕДЕЛЕНИЕ

Математическое понятие *множества* традиционно принимается в качестве интуитивного. Множество состоит из элементов. Соответствие каждому элементу (аргументу) из одного множества, некоторого элемента (образа) другого множества называется **функцией**.

Вариационный ряд — это формальный способ представления информации о функции, где аргументом является значение (или диапазон значений) переменной, а образом — частота либо количество элементов множества, принимающих данное значение (либо находящихся в пределах рассматриваемого диапазона).

Следует понимать, что в независимости от типа данных, переменные в вариационном ряду всегда будут описываться как конечные множества, состоящие из дискретных данных. Такая ситуация обусловлена тем, что максимальное количество аргументов функции не может превышать количества наблюдений, а всякий способ измерения ограничен в разрешающей способности. Вариационный ряд, содержит значения переменных, на основе которых устанавливается тип распределения.

Распределением — называется функция, которая описывает как выполняется соответствие аргументов и образов вариационного ряда.

На практике всякое распределение – не более чем идеализованное теоретическое предположение, которое с определенной степенью уверенности отражает реальность. В прикладных статистических дисциплинах исследователь всегда аппроксимирует функцию реального распределения к функции «идеального» распределения.

Упрощенную классификацию распределений можно представить в следующем виде:

В зависимости от типа данных аргумента:

1. Дискретные
2. Непрерывные

В зависимости от количества переменных функции:

1. Одномерные
2. Многомерные

От места распределения в приведенной упрощенной классификации зависит выбор способа описания и подхода к анализу данных. Дискретные распределения анализируются в медицинских исследованиях преимущественно с использованием методов и критериев Байесовой статистики, описание данных приводится в процентах, долях, либо абсолютных числах. Непрерывные распределения описывают неким набором входных аргументов, которые называют параметрами распределения: количество степеней свободы, параметры формы, меры положения, меры рассеяния и т. д. Рассмотрение распределения как одномерного или многомерного формирует выбор подхода к анализу и интегральной оценке функционально взаимосвязанных величин.

ПОНЯТИЕ ОБ ОПИСАТЕЛЬНОЙ СТАТИСТИКЕ

Ключевым моментом всякого статистического анализа является количественное представление данных. При проведении статистического анализа, исследователь, как упоминалось ранее, работает с «идеализованным» образом распределения. Для того чтобы охарактеризовать количественно особенности распределения изучаемых величин применяется описательная статистика.

Основные задачи описательной статистики:

- Описание групп объектов исследования.
- Статистическая оценка параметров распределения.
- Компактное визуальное представление данных о показателях.

Следует упомянуть о том, что описание данных должно производиться способом, который дает возможность максимально составить представление о распределении, т. е. значения показателей описательной статистики должны находиться в непосредственной взаимосвязи с параметрами функции распределения. Чем более тесной является взаимосвязь показателя описательной статистики с параметрами распределения, тем более подходящим для задач описательной статистики является данный показатель.

В каком виде представлять данные — в текстовом или графическом? Наша рекомендация может выглядеть следующим образом: информация должна предоставляться в виде наиболее соответствующим задачам исследователя и виду изложения информации.

Без всякого сомнения, если исследование представляет собой диссертацию на соискание ученой степени или журнальную статью, то лучше использовать текстовый формат представления данных, графику использовать по минимуму. Это даст возможность коллегам и экспертам получить интересующую информацию в наиболее полном объеме: количественные данные, полученные в ходе исследования, возможность экспертной проверки и т. д.

Если же данные приводятся в устном выступлении с презентацией, а выступление ограничено по времени, то приоритет должен быть у визуальных методов представления данных, которые позволяют донести информацию быстрее за короткий промежуток времени и акцентировать внимание на ключевых моментах работы.

МЕРЫ ПОЛОЖЕНИЯ: СРЕДНЕЕ АРИФМЕТИЧЕСКОЕ, МОДА, МЕДИАНА, ПРОЦЕНТИЛИ

Меры положения определяют, как будет располагаться функция распределения преимущественно относительно оси абсцисс. В высшей математике для интегрального обозначения меры положения используется термин «математическое ожидание», традиционно обозначаемый греческим символом ξ.

При выборе параметра математического ожидания для целей описания данных следует исходить из принципа: максимум информации при минимуме расхода вычислительных ресурсов, но не менее, чем это необходимо для того, чтобы создать корректное представление о данных.

В роли такого параметра может быть использован соответствующий аргумент из функции распределения. Не всякий аргумент, определяющий положение функции распределения, является достаточно простым для вычисления и, соответственно, понимания: если понимание смысла среднего арифметического не вызывает затруднений у большинства исследователей медико-биологического профиля, то описание данных с использованием, например, среднего гармонического уже создает определенные трудности в интерпретации результатов.

Таким образом, параметр описательной статистики, применяемый в качестве математического ожидания должен не только нести в себе информацию о положении функции распределения, быть доступным для вычисления, но и находиться в пределах способностей и навыков восприятия данной информации большинством специалистов медико-биологического профиля.

Наиболее простым для расчета и понимания исследователями показателем описательной статистики является среднее арифметическое (формула 1):

$$\overline{X} = n^{-1} \sum_{i=1} x_i \,, \qquad (1)$$

где n — количество случаев;

x_i — значение i-го случая.

Среднее арифметическое (Mean) является параметром выбора для описания математического ожидания для симметричных одномерных распределений. У среднего арифметического есть одно замечательное свойство: сумма квадратов отклонений значений признака от значения данной меры центральной тенденции минимальна относительно аналогичного показателя для других мер центральной тенденции (мода, медиана, среднее геометрическое и др.). Тем не менее, значимость его для описания данных снижается пропорционально нарастанию асимметрии распределения изучаемого параметра. Среднее арифметическое идеально подходит для описания математического ожидания нормального распределения (и его разновидностей), хорошо — логистических распределений.

Вопрос целесообразности использования данного показателя для других распределений — весьма неоднозначный: не смотря на то, что среднее арифметическое находится в непосредственной взаимосвязи с аргументами функции, оно зачастую не является параметром определяющим положение функции распределения относительно оси абсцисс.

Мода (Mode) — это точка на оси абсцисс, в которой плотность вероятности распределение имеет локальный максимум. Если описать более простым языком, мода — это наиболее часто встречающееся значение в выборке. Как мера центральной тенденции применяется крайне редко, а как самостоятельный параметр — практически никогда.

Распределения в зависимости от количества мод подразделяются на мономодальные (один локальный максимум) и мультимодальные (несколько локальных максимумов).

Квантиль (Quantile) — это такое число, что заданная случайная величина не превышает его с определенной вероятностью. Квантиль есть понятие более сложное для понимания, чем его прикладные аналоги с фиксированной вероятностью — процентили, децили, квартили и медиана.

Медиана (Median) — это возможное значение признака, которое делит ранжированную совокупность на две равные части: половина значений лежит выше медианы, половина — ниже. Если говорить более простым языком, медиана — средняя позиция в упорядоченном ряду значений. Она хорошо подходит как мера центральной тенденции для одномерных распределений даже в условиях выраженной ассиметрии распределения или при наличии выраженных выбросов значений.

Если медиана делит ранжированную совокупность на две равные части, то **процентили** (Percentile) — это такие значения признака, которые делят ее на 100 равных частей.

Децили — делят ранжированную совокупность на десять равных частей.

Квартили (Quartile) — делят ранжированную совокупность на четыре равных части.

Меры центральной тенденции при их корректном использовании дают представление о положении распределения и помогают установить его вид для дальнейших расчетов.

МЕРА РАССЕЯНИЯ: ДИСПЕРСИЯ, СТАНДАРТНОЕ ОТКЛОНЕНИЕ, РАЗМАХ

Распределение случайной величины помимо меры центральной тенденции принято описывать мерами рассеяния. Меры рассеяния позволяют получить информацию об изменчивости, вариабельности значений показателя в выборке.

Дисперсия (Variance) является простейшим расчетным показателем для описания рассеяния в выборке. Она отражает степень отклонения наблюдений от среднего арифметического в выборке (формула 2):

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

где x_i — значение i-го показателя;
X — среднее арифметическое;
n — количество случаев.

Дисперсия рассчитывается на основе среднего арифметического и, следовательно, наследует все его преимущества и недостатки: простоту арифметического расчета, идеальную пригодность для описания одномерного нормального распределения, восприимчивость к выбросам и непригодность для использования в качестве параметра, описывающего рассеяние в случае асимметричных распределений. Еще одним минусом дисперсии являются единицы измерения, которые представляют собой квадрат исходных данных.

Если извлечь квадратный корень из дисперсии получим меру рассеяния, которая называется **среднеквадратичным отклонением** или **стандартным отклонением среднего** (Standart Deviation of Mean). Этот показатель более удобен для практического использования, так как имеет одинаковые со средним арифметическим единицы измерения (формула 3):

$$s = \sqrt{\frac{\sum (x_i - \overline{X})^2}{n-1}} \quad (3)$$

Следует помнить, что именно среднее и стандартное отклонение являются двумя переменными, на основе которых формируется функция нормального распределения. Поэтому для одномерного нормального распределения принята форма описания данных в виде: $X \pm s$.

В случае же если распределение отличается от нормального, количественные показатели следует описывать с использованием медианы и процентилей: Me ($Q_{25\%}$; $Q_{75\%}$).

Общепринятыми мерами оценки рассеяния в случае негауссовых непрерывных распределений являются размах и интерквартильный размах.

Размах (Range) — это разница между минимальным и максимальным значениями количественного показателя. Является довольного грубой мерой рассеяния и применяется достаточно редко. В настоящее время в некоторых программных пакетах статистической обработки используется труднопереводимое на русский язык понятие «**Non-Outlier Range**», которое фактически представляет собой не разницу между максимумом и минимумом, а разницу между 99 и 1-м процентилями. Использование данного параметра позволяет снизить влияние единичных выбросов на представление о рассеянии.

Интерквартильный размах (Interquartile Range) — это разница между значениями верхнего и нижнего квартилей.

Собственно существуют и иные варианты квантильных «размахов»: децильный, интерперсентильный и т. д. Все они имеют аналогичный смысл.

В графическом виде меры центральной тенденции и меры рассеяния в одной или нескольких группах удобно представлять в виде графиков «ящик с усами» (box-whisker plot) (рисунок 2). Отдельные элементы графика («центр», «ящик», «усы») различаются для разных видов распределения (таблица 1).

Таблица 1 — Значение графических элементов в графике «ящик с усами»

№	Графический элемент	Значение графического элемента в зависимости от вида распределения	
		Нормальное	Отличное от нормального
1	«Усы»	Стандартная ошибка / 95 % ДИ	Размах / Non-Outlier Range
2	«Ящик»	Стандартное отклонение	Интерквартильный размах
3	«Центр»	Среднее арифметическое	Медиана

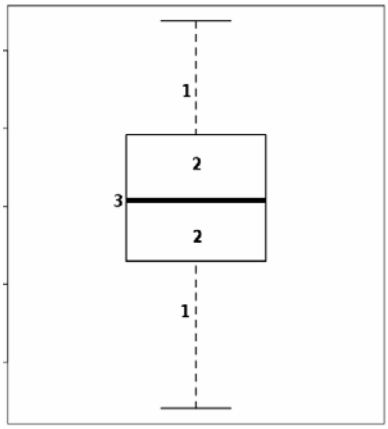


Рисунок 2 — Форма представления в виде графика «ящик с усами»

НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

Распределение вероятностей показывает вероятности всех возможных значений переменной. Каждое распределение обусловлено параметрами, обобщающими величинами, характеризующими данное распределение. При помощи соответствующих статистических методов можно производить оценку этих параметров в выборке. Нормальное распределение — одно из самых важных в статистике. Случайная величина называется нормально распределенной, если ее плотность вероятности имеет вид (формула 4):

$$\frac{1}{\sqrt{2ps}}\exp\left[-\frac{1}{2}\left(\frac{x_i-\bar{X}}{s}\right)\right], \quad (4)$$

где x_i — значение i-го показателя;

X — среднее арифметическое;

s — стандартное отклонение.

Плотность распределения случайной величины $U = (x_i — X) / s$ равна (формула 5):

$$p(U)=\left(\sqrt{2p}\right)^{-1}\exp\left(-\frac{1}{2}U^2\right), \quad (5)$$

Основные показатели, которыми можно охарактеризовать нормальное распределение:

1. Распределение описывается всего двумя параметрами: средним арифметическим и стандартным отклонением.
2. Распределение непрерывное, куполообразное, унимодальное.
3. Распределение симметрично относительно среднего.
4. Мода, медиана, среднее арифметическое совпадают.
5. Если среднее арифметическое увеличивается — распределение сдвигается вправо, если уменьшается — влево.
6. Если стандартное отклонение увеличивается — распределение уплощается, если уменьшается — становится более остроконечным.

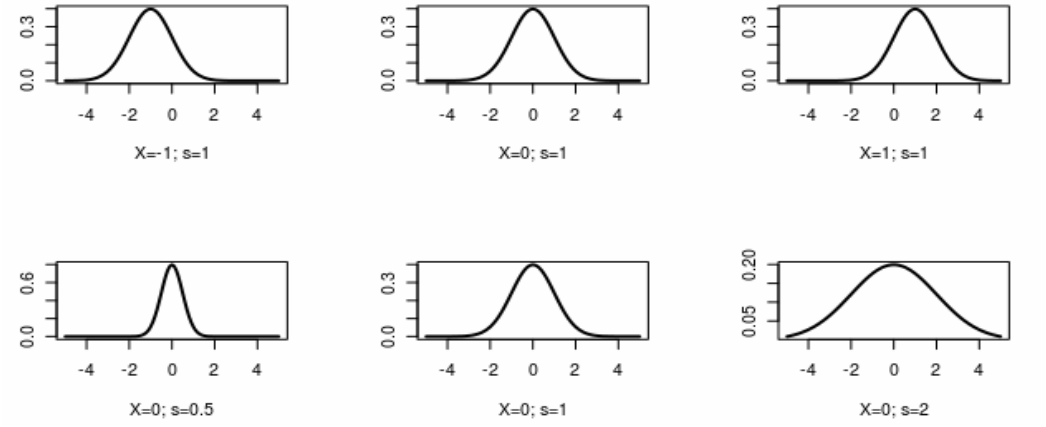


Рисунок 3 — Изменения графика нормального распределения в зависимости от значений среднего арифметического и стандартного отклонения

Нормальное распределение со средним равным 0 и стандартным отклонением равным 1 называют стандартным нормальным распределением или Z-распределением.

Собственно сама по себе информация о нормальности или «ненормальности» распределения показателя не имеет большой ценности для исследователя, так как большинство показателей в живом организме имеют распределение отличное от нормального. То есть само по себе соответствие (либо не соответствие) распределения случайной величины (особенно если это медико-биологический показатель) нормальному закону нужно воспринимать как факт, а не как характеристику качества проделанной работы.

Важность нормального распределения для статистики заключается несколько в другом. Нормальность распределения определяет дальнейшую тактику исследователя при обработке данных.

Если распределение нормальное, то существует вероятность того, что будут применяться параметрические методы, основанные на работе с основными параметрами нормального распределения: средним и стандартным отклонением. Параметрические методы обладают высокой чувствительностью, но имеют достаточно широкий ряд условий применимости, «индивидуальных» для каждого метода, которые ограничивают возможности их использования.

Если распределение отличается от нормального, следует применять непараметрические методы, критерии и показатели. Причем не только для проверки гипотез, но и для корректного описания данных. В большинстве случаев это подразумевает, что для корректного описания данных использовать среднее арифметическое и стандартное отклонение нельзя. При проверке гипотез непараметрические методы менее чувствительны (~ 90 % от чувствительности параметрических методов). Чувствительность снижается вследствие ранжирования и группировки данных, а так же преобразований. Практически все преобразования при использовании непараметрических методов направлены на то, чтобы максимально аппроксимировать распределение изучаемого показателя к стандартному нормальному Z-распределению. И именно в этом заключается ценность нормального распределения для медико-биологической статистики — в большом количестве критериев оно является «конечным этапом» расчета плотности вероятности и значений статистических погрешностей.

КРИТЕРИИ СООТВЕТСТВИЯ РАСПРЕДЕЛЕНИЯ НОРМАЛЬНОМУ ЗАКОНУ

Выяснение соответствия распределения изучаемого признака нормальному закону является одной из важнейших задач при статистической обработке данных. Нормальность наблюдаемых данных является необходимой предпосылкой для корректного применения большинства классических методов математической статистики, поэтому проверка на нормальность является обязательной процедурой в ходе проведения измерений, контроля и испытаний. Под «классическими» подразумеваются параметрические методы и критерии, вся логика применения которых, основана на предположении о том, что распределение изучаемого признака является нормальным и поддается описанию двумя параметрами, являющимися аргументами функции: средним арифметическим и стандартным отклонением (а также иными показателями описательной статистики, которые могут быть рассчитаны на основе информации об этих параметрах).

Для статистического анализа величин, распределение которых отличается от нормального применяются непараметрические методы и критерии. Не смотря на большую универсальность непараметрических методов и критериев, их использование не всегда целесообразно, так как в ходе расчетов данных параметров происходят потери информации при выполнении ранжирования и группировки данных, что снижает их мощность и чувствительность.

В Российской Федерации для оценки нормальности распределения применяется стандарт ГОСТ Р ИСО 5479-2002 «Статистические методы. Проверка отклонения распределения вероятностей от нормального распределения», введенный в действие в 2002 г. Этот стандарт представляет собой аутентичный текст международного стандарта ISO 5479-97. В упомянутом документе регламентируются предпочтительные методы и критерии, а так уже условия и ограничения их использования. Тем не менее, данный стандарт не является идеальным алгоритмом и не подразумевает использование ряда мощных критериев, а также сочетаний критериев (например z_2 критерий, предложенный D'Agostino).

По мнению авторов, не существует единого критерия, позволяющего однозначно дать ответ на вопрос о соответствии изучаемого признака нормальному закону, так как каждый из критериев, как будет показано ниже, помимо ряда ограничений рассматривает лишь одну из характеристик нормального распределения. Поэтому целесообразным является применение нескольких критериев, основанных на анализе нескольких характеристик распределения. Учитывая специфику получения медико-биологических данных и тяжесть возможных последствий ошибок, при статистической обработке исследователю следует обратить внимание на корректный вы-

бор нескольких критериев, учесть их возможности и ограничения использования. Заключение о нормальности распределения изучаемого признака делается на основе оценки истинности логического выражения, где заключение о нормальности распределения по каждому критерию связаны логическим «И». Только в случае истинности результирующего выражения, и корректности применения тестов, следует делать заключение о нормальности. Логика оценки нормальности распределения представлена в таблице 2.

Таблица 2 — Логика оценки нормальности распределения

Изучаемый показатель	Распределение нормальное по результатам данного теста			Распределение нормальное
	Критерий Шапиро-Уилка	Критерий проверки на эксцесс	Критерий проверки на симметричность	
Параметр 1	Истина	Истина	Истина	Истина
Параметр 2	Истина	Ложь	Истина	Ложь

Прежде всего, исследователю следует помнить о главном ограничении всех методов, предназначенных для анализа соответствия нормальному закону: изучаемый признак должен быть количественным. Отличия количественных признаков от порядковых приведены в главе «Типы данных».

Мы настоятельно рекомендуем не использовать графический метод проверки нормальности распределения, так как другие симметричные непрерывные распределения при изменении масштаба могут быть визуально неотличимы от нормального. Кроме того, при оценке графика «на глаз» невозможно привести математическое обоснование заключению о нормальности и дать оценку погрешности. Оценка соответствия распределения изучаемого признака нормальному закону должна производиться количественно, с использованием соответствующих методов и критериев с учетом условий и ограничений их использования.

Критерий проверки на симметричность

График нормального распределения, как упоминалось ранее, симметричен. Принцип метода основывается на определении значения коэффициента асимметрии (3-го центрального момента), которое в случае нормального распределения равно 0. Стандартизованный коэффициент асимметрии (skewness) так же равен 0.

Возможности метода:

— H_0: Распределение симметрично.

— H_1: Распределение асимметрично.

Справедливость H_0 свидетельствует лишь о симметричности распределения, но не позволяет сделать вывод о его соответствии нормальному закону, так как ряд других непрерывных распределений также симметричны (например: экспоненциальное, логистическое, Лапласса), т. е. условие необходимое, но не достаточное.

Ограничения метода:

1. $8 < N < 5000$, где N — количество случаев по изучаемому показателю.
2. Критерий не может использоваться самостоятельно

Критерий проверки на эксцесс

Значение коэффициента эксцесса (kurtosis) отражает «остроту пика» графика мономодального распределения, и рассчитывается как 4-й центральный момент. В случае нормального распределения значение коэффициента эксцесса равно 3, значение стандартизованного коэффициента равно 0.

Возможности метода:

— H_0: значение коэффициента эксцесса равно 3 (0 в случае стандартизованного коэффициента).

— H_1: значение коэффициента эксцесса отличается от 3 (0 в случае стандартизованного коэффициента).

Справедливость H_0 свидетельствует о соответствии значения коэффицента эксцесса таковому у нормального распределения, но не позволяет сделать вывод о его соответствии нормальному закону, так как не учитывает положение «пика» и симметричность графика, т. е. условие необходимое, но не достаточное. Кроме того, мощность данного критерия убывает с уменьшением объема выборки.

Ограничения метода:

1. $8 < N < 5000$, где N — количество случаев по изучаемому показателю.
2. Критерий не может использоваться самостоятельно.

Критерий Шапиро-Уилка

Значение критерия рассчитывается на основе анализа линейной комбинации разностей порядковых статистик. Критерий рекомендуют применять при отсутствии априорной информации о типе возможного отклонения от нормальности.

Возможности метода:

— H_0: распределение нормальное (значение коэффициента W Шапиро-Уилка стремится к 1 при любом значении p).

— H_1: распределение отличается от нормального (значение коэффициента W Шапиро-Уилка стремится к 0 при $p < 0{,}05$).

При использовании критерия важно обращать внимание не только на значение показателя W, но и на уровень статистической значимости. Так как нулевая гипотеза сформулирована о том, что распределение нормальное, то она будет приниматься при условии, что уровень статистической значимости $p > 0{,}05$ и высоких значений ($> 0{,}9$) W. В ином случае, принимается альтернативная гипотеза.

Ограничения метода:

1. $8 < N < 2000$, где N — количество случаев по изучаемому показателю.

2. В выборках объемом более 100 объектов резко снижается чувствительность критерия: крайне плохо выявляются различия между нормальным, экспоненциальным, логистическим, Лапласовым распределениями.

ПРИМЕР РАСЧЕТА ОПИСАТЕЛЬНОЙ СТАТИСТИКИ С ИСПОЛЬЗОВАНИЕМ «STATISTICA 8.0»

Данные для обработки вносятся в табличном виде как случаи (cases) и переменные (variables). Случаи представляют собой строки заполняемой таблицы данных (spreadsheet). Таблицы данных можно импортировать, изменять, сохранять и экспортировать для работы в других программных пакетах.

Каждый случай таблицы данных характеризуется набором параметров. Номера случаев представляются в соответствующем столбце данных, значения параметров для каждого случая — в соответствующих пронумерованных столбцах таблицы, неактивная область переменных обозначена темно-серым цветом.

Для удобства рассмотрения возможностей «Statistica» по расчету параметров описательной статистики создадим таблицу данных для 100 случаев, по 2 параметра для каждого, пусть это будут, например, рост и вес. Чтобы создать таблицу данных нажмем File – New и в открывшейся форме Create New Document зададим значения соответствующих полей: Number of variables – 2 и Number of cases – 100. Жмем ОК. Получим пустую таблицу 2 × 100.

Изменим названия переменных, для чего сделаем двойной щелчок левой кнопкой на имени переменной var1. Откроется форма Variable 1, в ней в поле name введем значение «Рост», нажмем ОК. Название переменной изменится на «Рост». По аналогии переименуем переменную var2 в «Вес» и заполним таблицу 4 данными.

	1 Рост	2 Вес		
1	123	31		
2	184	62		
3	189	65		
4	166	53		
5	144	42		
6	122	31		
7	124	32		
8	129	35		
9	171	55		

**Рисунок 4 — Днные в программе «Statistica».
Для каждого случая проставлен рост и вес: например,
пациент 1 имеет рост 123, вес 31.**

Программа «Statistica» поставляется в виде набора модулей для анализа, который может сильно варьировать в зависимости от версии и типа лицензии. Модуль описательной статистики входит в «базовый набор» и содержится во всех версиях. Для того чтобы запустить данный модуль нажмем Statistica – Basic Statistics/Tables. Откроется форма Basics Statistics/Tables: Spreadsheet1, в которой сделаем двойной щелчок на пункте списка Descriptive

Statistics. В результате проведенных манипуляций будет запущена форма модуля описательной статистики. Перейдем на вкладку Advanced.

Проставим галочки напротив рассмотренных ранее параметров описательной статистики, которые могут быть использованы для описания количественных данных. Теперь следует выбрать переменные, по которым будут производиться расчеты: нажмем кнопку Variables. Откроется форма диалога выбора переменных Select variables for analysis. В открывшейся форме удерживая клавишу ctrl щелчком левой кнопки мыши выделим «Рост» и «Вес», после чего жмем ОК (рисунок 5).

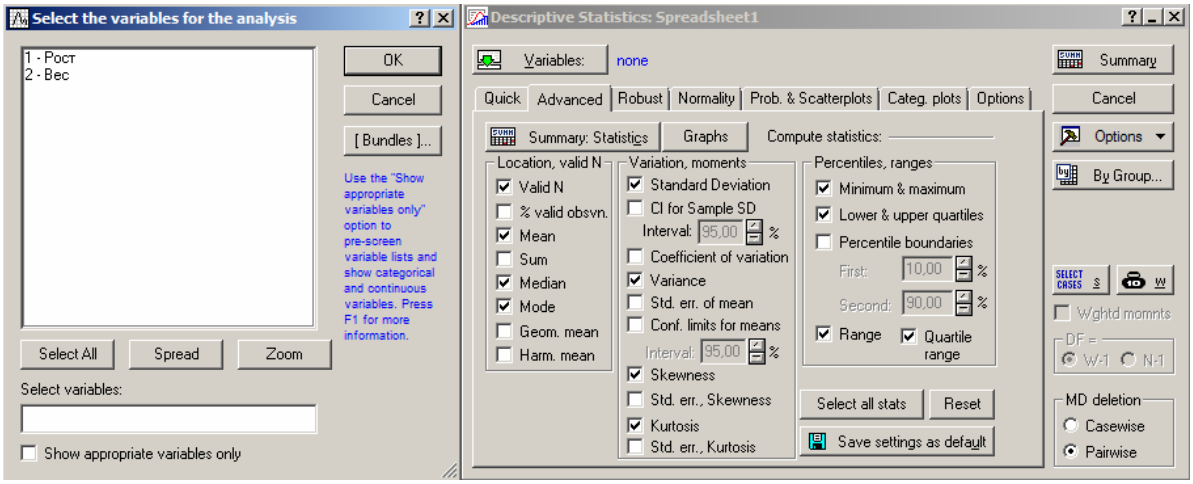
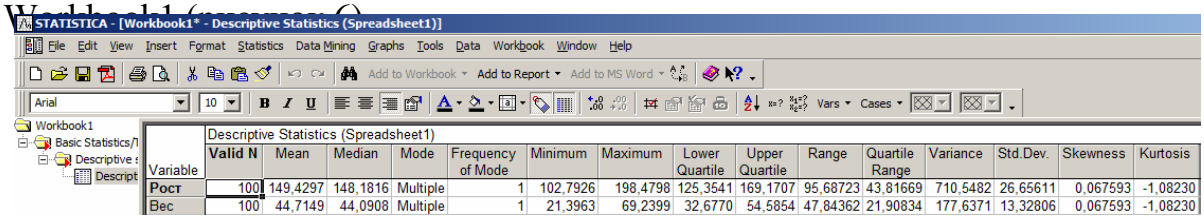


Рисунок 5 — Диалоговые формы модуля описательной статистики программы «Statistica» 8.0.

Чтобы произвести расчеты параметров описательной статистики нажмем Summary. Вывод данных в программе Statistica осуществляется в так называемые рабочие книги (workbook). В рабочих книгах приведены данные о результатах расчетов, в них же выводятся графики и результаты статистических тестов. Рабочие книги можно сохранять, изменять и экспортировать для повторного использования в других программах. В данном конкретном случае мы получим рабочую книгу с названием Workbook1 (рисунок 6).



Variable	Valid N	Mean	Median	Mode	Frequency of Mode	Minimum	Maximum	Lower Quartile	Upper Quartile	Range	Quartile Range	Variance	Std.Dev.	Skewness	Kurtosis
Рост	100	149,4297	148,1816	Multiple	1	102,7926	198,4798	125,3541	169,1707	95,68723	43,81669	710,5482	26,65611	0,067593	-1,08230
Вес	100	44,7149	44,0908	Multiple	1	21,3963	69,2399	32,6770	54,5854	47,84362	21,90834	177,6371	13,32806	0,067593	-1,08230

Рисунок 6 — Вывод результатов расчетов параметров описательной статистики в рабочую книгу программы «Statistica»

На основе расчетов попробуем составить представление о распределении переменной «Рост». Значения среднего и медианы достаточно близкие, распределение симметрично, имеет уплощенную форму, возможно наличие нескольких мод, но данный результат работы программы может

быть следствием разброса данных — стандартное отклонение и интерквартильный размах имеют достаточно большие значения.

Теперь проверим гипотезу о нормальности распределения и построим гистограммы, для чего вернемся в модуль описательной статистики (форма в свернутом виде на нижней панели). В форме модуля перейдем на вкладку Normality, поставим галочку на Shapiro-Wilk's W test и нажмем кнопку Histograms (рисунок 7).

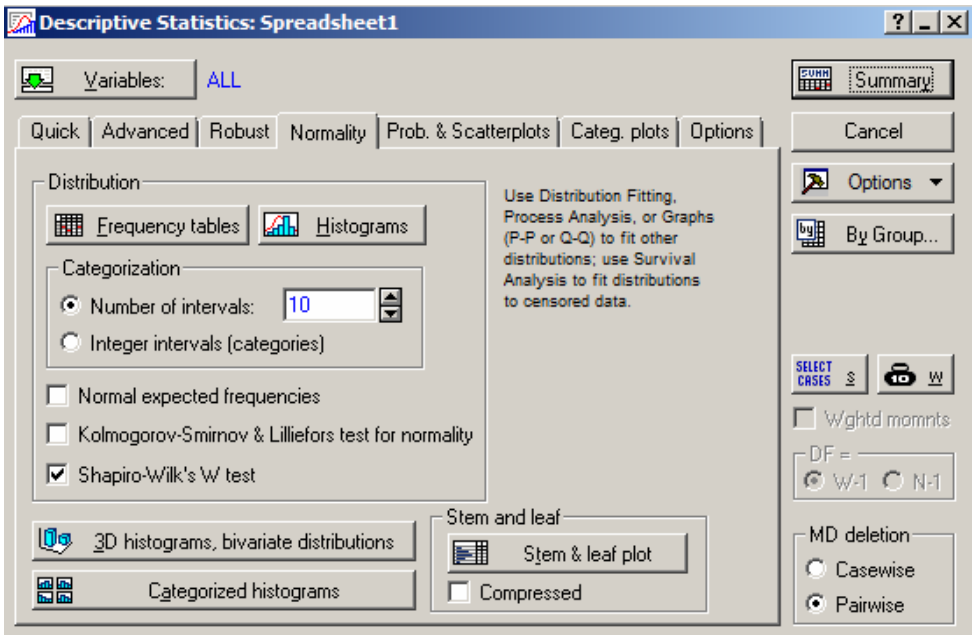


Рисунок 7 — Установка расчетных параметров для вывода гистограмм и проверки гипотезы о нормальности распределения в модуле описательной статистики программы «Statistica»

В результате в рабочую книгу будут выведены гистограммы для изучаемых параметров и значения W критерия Шапиро-Уилка для распределений (рисунок 8).

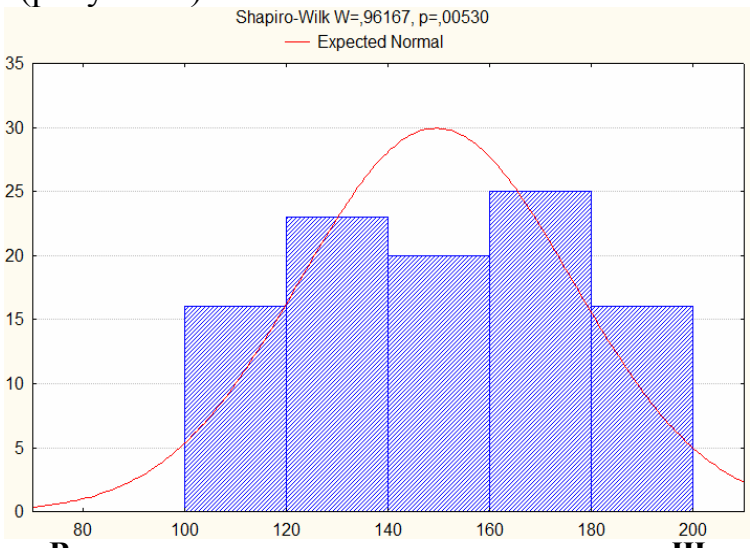


Рисунок 8 — Вывод гистограммы и значения критерия Шапиро-Уилка в модуле описательной статистике программы «Statistica»

Проанализируем полученные результаты для переменной «Рост»: распределение действительно симметрично, уплощено и имеет тенденцию к полимодальности. Критерий Шапиро-Уилка стремится к 1, но уровень статистической значимости $p < 0,05$, поэтому нулевая гипотеза о соответствии распределения изучаемой переменной нормальному отклоняется. Ожидаемая кривая нормального распределения при текущих значениях среднего и стандартного отклонения изображена красным цветом. Исходя из полученных результатов, мы сделали вывод о том, что распределение переменной «Рост» отличается от нормального, поэтому для описания данных следует использовать значения медианы и квартилей, а при проверке гипотез — методы непараметрической статистики.

Конечно, возможности программы «Statistica» не ограничиваются расчетом лишь перечисленных показателей, проверкой гипотезы о нормальности распределения и построением диаграмм, даже в модуль описательной статистики заложено гораздо больше возможностей, ознакомиться с которыми можно, воспользовавшись встроенной справочной системой.

Интерфейс программы интуитивно понятен даже начинающему пользователю. На практике произвести расчеты достаточно просто, гораздо сложнее решить вопросы интерпретации результатов и корректного выбора методов работы с данными. По этим вопросам можно получить больше информации, обратившись к книгам из списка использованной литературы.

ЛИТЕРАТУРА

1. *Петри, А.* Наглядная медицинская статистика / А. Петри, К. Сэбин; пер. с англ. под ред. В. П. Леонова. — 2-е изд., перераб. и доп. — М.: ГЭОТАР-Медиа, 2009. — 168 с.

2. *Гринхальх, Т.* Основы доказательной медицины / Т. Гринхальх; пер. с англ. под ред. К. И. Сайткулова. — М.: ГЭОТАР-Медиа, 2006. — 240 с.

3. *Лукин, Е. С.* Прикладная теория информации: учеб. пособие для студентов специальности «Информатика» / Е. С. Лукин. — Мн.:БГУИР, 2002. — 42 с.

4. *Реброва, О. Ю.* Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA / О. Ю. Реброва. — М.: МедиаСфера, 2002. — 312 с.

5. *Гланц, С.* Медико-биологическая статистика / С. Гланц; пер. англ. — М.: Практика, 1998. — 459 с.

6. *Назаренко, Г. И.* Основы медицинских технологических процессов / Г. И. Назаренко, Г. С. Осипов. — Ч. 1. — М.:ФИЗМАТЛИТ, 2005. — 144 с.

7. *Винер, Н.* Кибернетика, или управление и связь в животном и машине / Н. Винер. — 2-е изд. — М.: Наука, 1983. — 344 с.

8. *Джонсон, Н. Л.* Одномерные непрерывные распределения: в 2 ч. / Н. Л. Джонсон, С. Коц, Н. Балакришнан ; пер. 2-го англ. изд. — М.: Бином, 2010. — 703 с.

Учебное издание

Чубуков Жанн Александрович
Угольник Татьяна Станиславовна

ОПИСАТЕЛЬНАЯ СТАТИСТИКА

Учебно-методическое пособие для студентов всех факультетов медицинских вузов, аспирантов, магистрантов, соискателей и преподавателей

Редактор *О. В. Кухарева*
Компьютерная верстка *А. М. Терехова*

Подписано в печать 08.06.2012.
Формат $60\times84^{1}/_{16}$. Бумага офсетная 65 г/м^2. Гарнитура «Таймс».
Усл. печ. л. 1,62. Уч.-изд. л. 1,8. Тираж 50 экз. Заказ № 167.

Издатель и полиграфическое исполнение
Учреждение образования
«Гомельский государственный медицинский университет»
ЛИ № 02330/0549419 от 08.04.2009.
Ул. Ланге, 5, 246000, Гомель.