Федеральное государственное бюджетное образовательное учреждение высшего образования «Кемеровский государственный медицинский университет» Министерства здравоохранения Российской Федерации

Кафедра общественного здоровья, здравоохранения и медицинской информатики

Штернис Т. А.

БИОСТАТИСТИКА

Учебно-методическое пособие для обучающихся в аспирантуре по направлениям подготовки: 31.06.01 «Клиническая медицина», 32.06.01 «Медико-профилактическое дело», 06.06.01 «Биологические науки»

Кемерово 2020

УДК [614.2:311](075.9)
ББК 51.1(2)я73
Ш 904

Штернис, Т. А. Биостатистика : учебно-методическое пособие для обучающихся в аспирантуре по направлениям подготовки: 31.06.01 «Клиническая медицина», 32.06.01 «Медико-профилактическое дело», 06.06.01 «Биологические науки» / Т. А. Штернис. – Кемерово, 2020. - с 183.

Учебно-методическое пособие содержат информацию о целях и задачах раздела биостатистика в целом и отдельных его тем, вопросы для подготовки к занятиям, ситуационные задачи, тесты, критерии оценки приложения, рекомендуемый список литературы. Предназначено для обучающихся в аспирантуре по направлениям подготовки 31.06.01 Клиническая медицина, 32.06.01 Медико-профилактическое дело, 06.06.01 Биологические науки

Автор:

Штернис Татьяна Александровна – канд. мед. наук, доцент кафедры общественного здоровья, здравоохранения и медицинской информатики ФГБОУ ВО КемГМУ Минздрава России.

Рецензенты:

Леванова Л.А., д-р. мед. наук., доцент, заведующий кафедрой микробиологии, вирусологии, иммунологии ФГБОУ ВО «Кемеровский государственный медицинский университет» Минздрава России.

Пивовар О.И., канд. мед. наук., доцент, заведующий кафедрой инфекционных болезней ФГБОУ ВО «Кемеровский государственный медицинский университет» Минздрава России.

Рекомендовано Центральным методическим советом Кемеровского государственного медицинского университета в качестве учебно-методического пособия для обучающих в аспирантуре по направлениям подготовки 31.06.01 Клиническая медицина, 32.06.01 Медико-профилактическое дело, 06.06.01 Биологические науки

СОДЕРЖАНИЕ

Цели и задачи освоения биостатистики

Целями освоения биостатистики является овладение базисными теоретическими знаниями и практическими умениями по планированию, проведению научных исследований, статистической обработки информации и ее интерпретации.

Задачами раздела являются:

- обучение ориентированию в базовых теоретических положениях статистической науки
- обучение планированию научных исследований
- обучение выбору и применению статистических методов для обработки результатов собственных исследования;
- обучение самостоятельной интерпретации результатов статистической обработки информации собственных исследований, материалов научных публикаций, отчетов и др. статистической информации;
- привитие навыков самостоятельности, в том числе в сфере проведения научных исследований.

Тема 1.Основные понятия биостатистики. Статистические оценки и их свойства

Цель занятия: освоить базовые понятия биостатистики.

Учебно-целевые задачи:

- изучить основные понятия биостатистики.
- научиться интерпретировать базовые понятия биостатистики и применять их в практической деятельности врача-эпидемиолога

В результате освоения темы обучающиеся **должны знать:** основные термины и понятия используемые в биостатистике.

В результате освоения темы обучающиеся **должны уметь:** применять получение знания для решения профессиональных задач; определять

характер распределения признака в совокупности; проводить описание признаков в статистической совокупности и интерпретировать результаты.

В результате освоения темы обучающиеся **должны владеть:** методиками расчета относительных показателей и средних величин и их доверительных границ и мер разброса средних величин

ИНФОРМАЦИОННЫЙ МАТЕРИАЛ
ОСНОВНЫЕ ПОНЯТИЯ БИОСТАТИСТИКИ

Для изучения любого явления, в том числе заболеваемости населения необходимо выбрать ***объект*** исследования, *единицу наблюдения* и ее *учетные признаки,* определить *статистическую совокупность* и *необходимый объем наблюдений*. Объем выборочного исследования должен быть достаточным для получения статистически значимых результатов исследования.

Теоретическое обоснование выборочного метода строится на основе теории вероятностей и закона больших чисел. Суть теории вероятности заключается в том, что появление любого признака рассматривается как случайное событие.

Например, несмотря на случайность заболеваний гриппом, эпидемии возникают с определенной закономерностью и, как правило, в холодное время года, так что эпидемиологи могут с большой вероятностью прогнозировать распространенность этого вида заболевания среди населения.

В последнее время особый интерес медицинской общественности вызывают результаты исследований, полученные на принципах доказательной медицины, основой которой является теория вероятности и закон больших чисел, а также современные методы обработки информации.

Вероятность – это мера возможности возникновения случайных событий в конкретных условиях.

В биологии и медицине мы чаще всего имеем дело с выборочными совокупностями.

Вероятность наступления какого-либо события (P) определяется отношением наступивших событий (m) к числу всех возможных случаев (n):

$$P = \frac{m}{n};$$

Альтернативной противоположностью вероятности наступившего события является вероятность его отсутствия (q).

$$q = \frac{n-m}{n} = 1 - \frac{m}{n} = 1 - p; \qquad q = 1 - p;$$

$$p + q = 1;$$

Вероятность наступления события находится в пределах от 0 до 1. Соответственно, чем P ближе к 1, тем выше вероятность наступления того или иного события и чем больше P приближается к 0, тем меньше возможность его наступления.

Теория вероятности является постулатом закона больших чисел.

Закон больших чисел позволяет утверждать:

- *с увеличением числа наблюдений результаты выборочного исследования стремятся воспроизвести закономерности генеральной совокупности;*
- *достаточный объем наблюдений позволяет установить закономерности, которые не удается обнаружить при небольшом числе наблюдений.*

Для иллюстрации сказанного в математике применяют упрощенные схемы моделирования этих процессов с подбрасыванием монет или использованием ящиков с цветными шарами.

Доля, полученная на выборочной совокупности (P_1), при большой выборке весьма близка к доле, которую составляет явление в генеральной

совокупности (P). При большом числе наблюдений вероятность этого несовпадения настолько мала, что практически с ней можно не считаться.

Теорией статистики установлено, что при большой выборке (n>30) с вероятностью, равной 95%, можно утверждать, что разность долей, полученных из этой выборки (P_1), и генеральной совокупности (P), будет составлять 2 m; с вероятностью, равной 99,7%, можно утверждать, что разность этих долей (P_1-P) не превысит 3 m.

Числа 1, 2, 3 и др., на которые умножают ошибку репрезентативности (m), носят название доверительных коэффициентов и обозначают их буквой t. С увеличением t возрастает степень вероятности, с которой можно утверждать, что разность долей, полученных из выборки и генеральной совокупности будет находиться в пределах: $\Delta = tm$; где Δ – предельная ошибка, допустимая для данного исследования. Предельная ошибка (Δ) может быть с положительным и отрицательным знаком ($\pm\Delta$). Следовательно: $P = P_1 \pm \Delta$.

Пользуясь законом больших чисел, увеличивая объем выборки, можно регулировать размер предельной ошибки, доводя ее до минимальных размеров. При планировании исследования используют формулы, основывающиеся на законе больших чисел, по которому рассчитывают необходимую численность (n) выборки. Для этого надо знать, с какой точностью в зависимости от задач исследования, необходимо получить результаты, т.е. иметь представление о допустимой для данного исследования ошибке (Δ).

Основные и наиболее общие положения теории вероятностей и закона больших чисел разработаны отечественными учеными - математиками П.Л. Чебышевым, А.М. Ляпуновым, А.А. Марковым. Дальнейшая разработка теории вероятностей произведена советским математиком А.Н. Колмогоровым. Теорема Чебышева формулируется следующим образом: с

вероятностью, сколь угодно близкой к единице, можно утверждать, что при достаточно большом числе независимых наблюдений средняя величина изучаемого признака, полученная на основе выборки, будет сколь угодно мало отличаться от средней величины изучаемого признака во всей генеральной совокупности.

РАСПРЕДЕЛЕНИЕ ПРИЗНАКА В СТАТИСТИЧЕСКОЙ СОВОКУПНОСТИ

Статистическая совокупность состоит из элементов, имеющих различные значения изучаемого признака. Количество элементов (единиц наблюдения) в совокупности варьирует. *Величина, свидетельствующая о количестве единиц наблюдения с одинаковой величиной признака, называется частотой.* Наличие в совокупности элементов с определенным значением признака выражается мерой вероятности, что позволяет с помощью теории вероятностей определить закономерности распределения признака изучаемого явления.

Характер распределения, возможно, установить только на достаточном объеме наблюдений. Знание характера распределения признака в совокупности способствует рациональному выбору статистических критериев для оценки результатов исследования.

Распределения, которые наблюдаются в медицинских, в том числе и в социально-гигиенических исследованиях, довольно разнообразны по своему характеру.

Различают основные типы распределений: альтернативное, нормальное (симметричное) и ассиметричное (правостороннее, левостороннее, двугорбое – бимодальное и др.).

Часто встречаются явления, которые распределяются по типу асиммет-ричного распределения. При правостороннем распределении наибольшее число случаев наблюдения скапливается не на уровне середины ряда, а

сдвигается в сторону меньшего значения признака. В этом случае M>Me>Mo. Если наибольшее число случаев наблюдения сдвинуто в сторону большего значения признака, наблюдается левосторонняя асимметрия (M<Me<Mo). Если наибольшее число случаев наблюдения скапливается по концам ряда, отмечается двугорбое бимодальное распределение.

Правосторонняя асимметрия характерна для распределения такого признака, как число детей в семье или кратность случаев временной утраты трудоспособности. Как известно, в большинстве семей имеется 1-2 ребенка. С увеличением числа детей в семьях соответственно уменьшается число семей. Если проанализировать ряд по кратности случаев нетрудоспособности в связи с заболеванием в течение года, то он будет иметь вид правосторонней асимметрии, так как основная масса работающих имеет минимальное число случаев временной нетрудоспособности 1-2 (т.е. значительное число болеющих скапливается у наименьшей градации данного признака).

Реже встречается распределение нормальное (симметричное). Обычно нормальное распределение наблюдается при построении рядов, вариантами которых являются количественные признаки: рост, масса тела, уровень артериального давления, сроки госпитализации и др.

При нормальном типе распределения число случаев наблюдений с различной величиной признака располагается симметрично по отношению к середине ряда: от меньшего значения признака к большему его значению, наибольшее число случаев наблюдений приходится на середину ряда. При нормальном распределении признака в совокупности значения моды, медианы и средней арифметической величины равны (M=Me=Mo). Если признаки имеют только положительные значения, можно оценить соответствие эмпирического распределения нормальному с помощью среднего квадратического отклонения. Если среднее квадратическое отклонение меньше половины среднего (S<M/2), распределение можно

считать нормальным, т.к. симметричность одна из основных характеристик нормального распределения.

Двугорбое – бимодальное распределение имеет две вершины. Как правило, такой ряд нуждается в дополнительном анализе. Двугорбый тип распределения указывает, что совокупность неоднородна. Например, если включить в совокупность мальчиков и девочек и измерить их рост, то полученное распределение будет бимодальным.

При альтернативном типе распределения признака в совокупностях, специальных методов его определения не требуется. Обработка данных, полученных на альтернативных совокупностях, предполагает вычисление относительных величин, оценку статистической значимости различий между ними и установление других закономерностей.

Определить отличается ли эмпирическое распределение от нормального можно с помощью коэффициента асимметрии и эксцесса. Чем выраженнее асимметрия (As), тем больше значение коэффициента. Знак величины коэффициента асимметрии связан с направлением асимметрии, если распределение вытянуто в сторону отрицательных значений, то коэффициент асимметрии является положительным(As>0), если распределение вытянуто в сторону положительных значений, коэффициент асимметрии является отрицательным (As<0; рис. 1 а).

Количественное описание островершинности дает коэффициент эксцесса (Ex). Если эксцесс больше нуля, распределение принято считать островершинным, в противоположном случае – туповершинным рис.1 б).

Таким образом, если асимметрия и эксцесс значительно отличаются от нуля, можно считать, что распределение не подчиняется закону нормального распределения (рис.1 б).

В эпоху компьютерных технологий рассчитать эти коэффициенты не сложная задача. В доступной для любого пользователя программе Microsoft

Office Excel имеется встроенная опция «Анализ данных», которая позволяет получить значения описательной статистики.

С целью определения отличия изучаемого распределения от нормального с помощью коэффициентов асимметрии и эксцесса, применяют обычный в биометрии метод сравнения коэффициентов с их ошибками репрезентативности. Ошибки репрезентативности определяют по формулам:

$$mAs = \sqrt{\frac{6}{n}}; \quad tAs = \frac{As}{mAs} \geq 3$$

$$mEx = 2\sqrt{\frac{6}{n}}; \quad tEx = \frac{Ex}{mEx} \geq 3$$

Если показатели эксцесса и асимметрии превышают свою ошибку более чем в три раза, можно говорить об отличии эмпирических распределений от нормального.

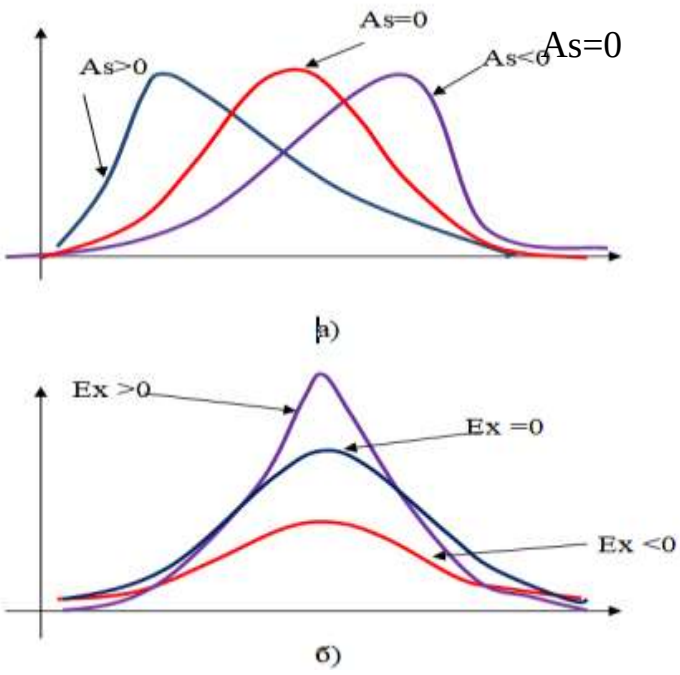


Рисунок 1. – Определение характера распределения признака с помощью коэффициентов эксцесса и асимметрии

С помощью специализированных статистических программ можно оценить характер распределения признака несколькими способами: построить гистограмму распределения; наглядно оценить, насколько диаграмма близка к графику нормального распределения; проверить нулевую гипотезу о том, что эмпирическое распределение соответствует закону нормального распределения; альтернативную гипотезу о том, что эмпирическое распределение не соответствует закону нормального распределения.

Проверить гипотезы можно с помощью критериев:

- Колмогорова-Смирнова. Условие: среднее значение и среднее квадратическое отклонение признака известны заранее, а не вычисляются по выборке;
- Лиллиефорса. Условие: среднее значение и среднее квадратическое отклонение признака вычислены по выборке;
- Шапиро-Уилка. Условие: среднее значение и среднее квадратическое отклонение признака не известны. Этот критерий является наиболее мощным и самым «строгим» из перечисленных.

Если полученное значение p для используемого критерия больше критического уровня статистической значимости (0,05), то эмпирическое распределение приближенно считают нормальным.

СТАТИСТИЧЕСКИЙ АНАЛИЗ

Программа анализа материала основана на свойствах статистической совокупности: распределение признаков, средний уровень признаков, вариабельность, репрезентативность и взаимосвязь признаков (корреляции и ассоциации).

Оценить распределение признаков можно абсолютными числами и относительными показателями.

Средний уровень признаков описывается средними величинами (мода, медиана, средняя геометрическая, средняя арифметическая). Какую именно среднюю величину применить в каждом конкретном случае исследователь решает, предварительно оценив характер распределения признаков в совокупности.

Вариабельность или разнообразие признаков характеризуют размах (амплитуда (А), интерпроцентильный размах, интерквартильный размах и среднее квадратическое отклонение. Для принятия решения о том, какие меры разнообразия использовать, оценивается соответствие распределения признака закону нормального распределения.

Репрезентативность – представительность признаков, то есть выборочная совокупность, должна обладать всеми свойствами генеральной совокупности. Мерами репрезентативности являются ошибка показателя и доверительные интервалы.

Еще одно свойство статистической совокупности – это взаимосвязь признаков. Установить наличие взаимосвязи количественных или порядковых признаков – это значит рассчитать коэффициент корреляции, который покажет, в какой степени изменение значения одного признака сопровождается изменениями значения другого признака. Метод Пирсона применяется для признаков, характер распределения которых соответствует закону нормального распределения. Для других распределений используются непараметрические методы оценки корреляции. Описать взаимосвязь качественных признаков или ассоциацию можно, используя методы Спирмена, Кендалла, Гамма.

Перечисленные методы подходят для выявления линейной взаимосвязи. Эта такая взаимосвязь, при которой увеличение (уменьшение) средних значений одного признака приводит к увеличению (уменьшению) средних значений другого. Но встречается и другой тип взаимосвязи между явлениями. Например, исследователь проверяет гипотезу о влиянии

продолжительности рабочей смены на производительность труда. Производительность труда с начала рабочей смены постепенно повышается, затем продолжительное время держится на высоком уровне, а к концу смены начинает снижаться. В данном случае можно говорить о криволинейной зависимости, которая коэффициентами линейной корреляции не улавливается. Одним из способов оценки корреляционной связи в этом случае является применение корреляционного отношения.

Таким образом, в программе анализа необходимо указать, какие методы применялись для оценки характера распределения признака в совокупности и для проверки статистических гипотез, какие параметры распределения использовались для описания признаков, явлений.

План исследования включает в себя перечень организационно-методических вопросов. В соответствии с планом исследователь определяет, что исследовать, в каком направлении, сколько единиц наблюдения взять для исследования, где проводить исследование и другие вопросы.

При выборе места исследования необходимо исходить из того, что это должно быть типичное лечебно-профилактическое учреждение, либо типичное предприятие, типичный район. Только в этом случае исследование будет представлять ценность, а именно возможность использования результатов научного труда на практике, возможность распространения полученных данных на другие подобные учреждения, территории и т.п.

Под объектом наблюдения понимается статистическая совокупность, состоящая из отдельных предметов или явлений (единиц наблюдения), взятых вместе в единых границах времени и пространства. Объект исследования должен иметь четко установленный формат, которым определяются место (территория), сроки, объем и единица наблюдения. Объектами исследований в области организации здравоохранения и

общественного здоровья могут быть медицинские организации, население областей, краев, городов и районов, ресурсы системы здравоохранения и др.

Тему научной работы определяет предмет исследования. *Предмет – это свойства, особенности, процессы объекта исследования, которые следует изучить***.** В качестве изучаемого явления можно рассматривать общественное здоровье населения, организацию медицинской помощи, ресурсное обеспечение отрасли и т.п.

В качестве объекта исследования может быть представлено население территории Н., а в качестве предмета, в данном случае, будет заболеваемость населения изучаемой территории. В одном и том же объекте может быть несколько предметов исследования. Границы между предметом и объектом условны и подвижны. Например, изучение качества жизни связанного со здоровьем, является предметом исследования для практикующего врача, а для организатора здравоохранения – объектом исследования. Точное определение предмета избавляет исследователя от безнадежных попыток «объять необъятное» и позволяет представить новую информацию об объекте исследования.

Единица наблюдения – это первичный элемент статистической совокупности, наделенный всеми признаками, подлежащими учету и регистрации. Если объектом исследования являются работающие угольных предприятий, то один работающий – это единица наблюдения. Если же объектом исследования является здравоохранение территории Н., единицей наблюдения в исследовании будет каждое конкретное лечебно-профилактическое учреждение.

Общее число единиц наблюдения в исследовании характеризует его объем. Если исследование является не сплошным, объем выборки должен быть достаточным для получения статистически значимых результатов, т.е., исследование должно иметь необходимую статистическую мощность.

ОПИСАНИЕ КАЧЕСТВЕННЫХ ПРИЗНАКОВ АБСОЛЮТНЫЕ И ОТНОСИТЕЛЬНЫЕ ВЕЛИЧИНЫ

В медицине и здравоохранении нередко пользуются абсолютными величинами. Они несут важную информацию о численности и составе населения регионов, городов, районов, количестве родившихся, заболевших, умерших, сменивших место жительства и т.д. Эти данные используются для организации медицинской помощи населению и решения других проблем.

Абсолютная величина должна быть указана, если данные получены на малых выборках (менее 20 единиц наблюдения), потому что в этом случае процентные значения оказываются значительно больше абсолютного числа единиц наблюдения.

Если количество единиц наблюдения в исследовании от 20 до 100, то проценты должны быть представлены в виде целых чисел. Например, 81% и 19%. При числе наблюдений более 100 относительная величина указывается с одним разрядом десятичной дроби.

В некоторых случаях абсолютные числа без их преобразования в относительные величины (показатели) имеют ограниченное познавательное значение. Например, можно судить об эффективности противоэпидемических мероприятий по поводу инфекционных заболеваний, если располагать данными об их распространенности в двух городах без учета численности населения городов и определения соответствующих показателей.

Различают следующие виды относительных величин: интенсивные, экстенсивные, соотношения и наглядности.

Интенсивный показатель, или показатель частоты, вычисляется на 100, 1000, 10000 в однородной среде, то есть среде, «продуцирующей» это явление.

Пример. Численность населения района К. составляет 30000. За последний год в названном районе родилось 450 детей. Требуется определить уровень рождаемости.

В данном случае необходимо найти интенсивный показатель (показатель частоты), который вычисляется на 100, 1000, 10000 в однородной среде.

Среди 30000 жителей родилось	450 детей;
на 1000 жителей	*x* детей

$$x = \frac{1000 \times 450}{30000} = 15{,}0 \text{‰}$$

Экстенсивный показатель, или показатель структуры распределения частей в целом, выражается в %. Однако общее число случаев может приниматься не только за 100, но и за 1000 и 10000. Тогда рассматриваемый показатель вычисляется в ‰ и %.

Пример. В районе С были зарегистрированы следующие случаи инфекционных заболеваний (см. табл. 1). Требуется рассчитать показатель, характеризующий структуру заболеваемости.

Для характеристики распределения частей в целом необходимо определить экстенсивный показатель (показатель структуры), который выражается в % по отношению к итоговым данным.

Таблица 1 – Распределение больных по нозологическим формам

Название заболеваний	Количество случаев	В %
Корь	8	13,3
Скарлатина	1	1,7

Эпидемический гепатит	9	15,0
Коклюш	15	25,0
Энтерит	20	33,3
Прочие	7	11,7
Всего:	60	100,0

60 случаев заболеваний – 100%
8 случаев заболеваний корью – x%

$$x = \frac{8 \times 100}{60} = 13{,}3\%$$

60 случаев заболеваний – 100%
1 случай заболевания скарлатиной – x%

$$x = \frac{1 \times 100}{60} = 1{,}7\%$$

Показатель соотношения характеризует отношение между самостоятельными совокупностями и вычисляется на 100, 1000 и 10000.

Пример. В районе Н с численностью населения 40000 развернуто 480 больничных коек. Какой вид относительных величин целесообразно вычислить для характеристики обеспеченности населения больничными койками?

В данном случае нужно рассчитывать показатель соотношения, который вычисляется на 100, 1000, 10000 в разнородной среде.

На 40000 человек развернуто 480 больничных коек, на 1000 человек развернуто x больничных коек.

$$x = \frac{1000 \times 480}{40000} = 12{,}0 \text{ ‰}$$

Показатель наглядности используется для изучения динамики изучаемого явления во времени, вычисляется в % к начальному уровню или к средней величине числового ряда, принятым за 100 %.

Пример. При изучении заболеваемости с временной утратой трудоспособности в динамике были получены следующие данные (табл. 2).

Таблица 2 – Характеристика динамики снижения заболеваемости с временной нетрудоспособностью по годам

Годы	2007	2008	2009	2010	2011
Количество случаев заболеваний	120	110	105	100	94
Показатель наглядности	100%	91,7%	87,5%	83,3%	78,3%

Требуется рассчитать показатели, позволяющие наглядно представить сведения о заболеваемости.

Для характеристики динамики изучаемого процесса необходимо определить показатель наглядности, который вычисляется в % по отношению к начальному уровню или к средней величине числового ряда, принятым за 100%.

Вычисление показателя наглядности для 2008 г.

120 случаев заболеваний – 100%
110 случаев заболеваний – x%

$$x = \frac{110 \times 100\%}{120} = 91{,}7\%$$

Вычисление показателя наглядности для 2009 г.

120 случаев заболеваний – 100%
105 случаев заболеваний – x%

$$x = \frac{105 \times 100\%}{120} = 87{,}5\%$$ и т.д.

Относительная частота (доля, пропорция, вероятность, шанс)

Описание качественных бинарных признаков начинается с подсчета абсолютных и относительных частот. Анализируются данные величины путем построения таблиц сопряженности (табл. 3).

Описывать порядковые данные можно с помощью медианы, моды и квартили.

Таблица 3 – сопряженности для сравнения групп по бинарному признаку.

Мед. учреждения	Осложнения есть (абс./отн.)		Осложнений нет (абс./отн.)		Итого
Больница А	64 (a)	16,7%	320 (в)	83,3%	384 (100%)
Больница Б	58 (c)	13,2%	383 (d)	86,8%	441 (100%)

Рассмотрим описание качественных бинарных признаков на примере.

В хирургическом отделении больницы А за год было прооперировано 384 человека, у 64 больных в послеоперационном периоде возникли осложнения (т.е в 64 случаях из 384). В хирургическом отделении больницы Б за год был прооперирован 441 человек, у 58 больных в послеоперационном периоде возникли осложнения (т.е. в 58 случаях из 441). В таблице 3 представлены абсолютные и относительные частоты наблюдаемых явлений.

Относительные величины следует представлять в отчете одновременно с указанием абсолютного значения той величины, которая принималась за 100%. Например, 16,7% из 384 единиц наблюдения. Допускается указание одновременно абсолютных и относительных значений признака: 16,7% (64 из 384 единиц наблюдения).

Описать относительную частоту (долю, пропорцию, вероятность) бинарного признака можно с помощью доверительного интервала (ДИ).

Рассчитаем 95% ДИ для относительной частоты возникновения послеоперационных осложнений, которая составила 16,7 на 100 прооперированных (точечная оценка). Необходимо вычислить интервальную оценку.

С целью определения интервальной оценки вычислим 95% доверительный интервал для частоты послеоперационных осложнений. Для этого нужно определить ошибку показателя (m). Ошибка показателя является мерой отличия выборочной совокупности от генеральной, и свидетельствует о пределе возможных колебаний коэффициента при повторном исследовании.

$$m = \pm\sqrt{\frac{p \times q}{n}}$$

m – ошибка показателя

p – вероятность наступления события

q – вероятность не наступления события

$q = 1 - p$, если показатель вычислен в десятичных долях 1;

$q = 100 - p$, если показатель вычислен на 100;

$q = 1000 - p$, если показатель вычислен на 1000;

$q = 10000 - p$, если показатель вычислен на 10000;

q = число наблюдений

$$m = \pm\sqrt{\frac{16{,}7 \times (100 - 16{,}7)}{384}} = \pm 1{,}9$$

Доверительные границы статистического показателя определяются по формуле:

$$p \pm tm,$$

где

p – показатель,

t – доверительный коэффициент,

m – ошибка показателя

В рассмотренном примере показатель равен 16,7 на 100 обследованных, его ошибка соответствует ±1,9.

95% ДИ равен 16,7±2 х 1,9 = (12,9; 20,5).

Следовательно, существует 95% уверенность, что истинная частота послеоперационных осложнений колеблется в популяции в частотном интервале от 12,9 до 20,5 на 100 прооперированных.

ОПИСАНИЕ КОЛИЧЕСТВЕННЫХ ПРИЗНАКОВ

Если количественные данные получены на малых выборках (менее 20 единиц наблюдения), следует приводить их в виде таблицы данных. В этом случае использование методов описательной статистики может дать неадекватные реальности результаты, поэтому нет необходимости вычислять параметры распределения.

Если количество единиц наблюдения более 20, следует приводить таблицу с описанием центральных тенденций и рассеяния значений признака. Таблицу можно заменить диаграммой диапазонов.

Меры центральной тенденции показывают наиболее типичное количественное значение для данной выборки, а меры рассеяния – разброс значений признака в совокупности. При описании признаков необходимо учитывать характер их распределения. Технология описания количественных признаков с учетом характера их распределения представлена в таблице 4.

Таблица 4 – Технология описания количественных признаков с учетом характера их распределения

Параметры распределения	
Меры центральной тенденции	Меры рассеяния
Средняя величина (М) сжатая числовая характеристика изучаемого явления. Медиана (Ме) величина, которая	Размах (или амплитуда (А) разница между максимальным и минимальным значением вариационного ряда. Интерпроцентильный размах

делит вариационный ряд на две равные части. Мода (Mo) величина, которая наиболее часто встречается в данном вариационном ряду.	(интервал) – чаще всего это значения 10-го и 90-го процентилей распределения. Этот интервал включает 80% значений признака в выборке. Интерквартильный размах (интервал) – это значение 25-го и 75-го квартилей. Интерквартильный размах включает 50% значений признака в выборке. Среднее квадратическое отклонение (s) абсолютная мера разброса значений признака около средней величины.
Характер распределения признака в статистической совокупности	
Распределение признака подчиняется закону нормального распределения	Характер распределения признака отличается от нормального
Описание количественного признака производится с помощью средней величины и среднего квадратического отклонения, в формате M(s). Необходимо также показать число наблюдений (n).	Количественный признак описывается медианой и интерквартильным размахом (либо интерпроцентильным размахом). При представлении данных указывается число наблюдений (n), Me (25-й; 75-й квартили или 10-й; 90-й процентили).

ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ, ДОВЕРИТЕЛЬНАЯ ВЕРОЯТНОСТЬ, УРОВЕНЬ СТАТИСТИЧЕСКОЙ ЗНАЧИМОСТИ

Выборка из генеральной совокупности позволяет получить точечную оценку интересующего параметра и вычислить стандартную ошибку для определения точности оценки. Стандартная ошибка используется для вычисления интервальной оценки – доверительного интервала для параметра популяции.

Доверительный интервал расширяет оценки в обе стороны некоторой величиной, кратной стандартной ошибке найденной величины.

Доверительные границы интервала обычно отделяют запятой и ставят в скобки.

Величина доверительного интервала задается вероятностью безошибочного прогноза (доверительная вероятность, надежность). Величина доверительной вероятности может задаваться t – коэффициентом Стьюдента.

При достаточном числе наблюдений значения коэффициента t и доверительной вероятности соотносятся следующим образом: если $t = 1$, то с доверительной вероятностью в 68,3% результаты выборочного исследования могут быть перенесены на генеральную совокупность; при $t = 2$ вероятность перенесения результатов выборочного исследования на генеральную совокупность увеличивается до 95,5% и при $t = 3$ – до 99,7%.

При интерпретации доверительных интервалов необходимо учитывать их ширину. Широкий доверительный интервал указывает на неточную оценку, узкий – на точную оценку. Ширина доверительного интервала зависит от размера стандартной ошибки, которая в свою очередь зависит от объема выборки и при рассмотрении числовой переменной от изменчивости данных. Исследования с небольшим набором данных дают более широкие доверительные интервалы, чем исследования многочисленного набора данных немногих переменных.

Доверительный интервал принято определять для средней величины, медианы, относительной частоты встречаемости признака в совокупности и др. Доверительный интервал с определенной долей уверенности (95% или 99%) показывает, в каких пределах находится изучаемый показатель в генеральной совокупности. Например, 95% доверительный интервал представляет собой область, в которую попадает истинное значение изучаемого показателя в 95% случаев. Иными словами, можно с 95% надежностью сказать, что истинное значение изучаемого показателя в

генеральной совокупности будет находиться в пределах 95% доверительного интервала.

Значению доверительной вероятности соответствует свой уровень статистической значимости (p). Уровень статистической значимости выражает вероятность нулевой гипотезы (вероятность того, что в сравниваемых совокупностях отсутствуют различия в распределении частот). Чем выше уровень статистической значимости, тем меньше можно доверять утверждению о том, что различия существуют.

Для вероятности безошибочного прогноза – 95%, уровень статистической значимости: p = 1-0,95 = 0,05. Для доверительной вероятности – 99%-p = 0,01.

Таким образом, статистическая значимость полученных на выборке данных представляет собой меру уверенности в их «истинности». Уровень статистической значимости находится в убывающей зависимости от доверительной вероятности. Высокая статистическая значимость соответствует низкому уровню доверия к найденной по выборке величине. Именно уровень статистической значимости представляет собой вероятность ошибки, связанной с распространением наблюдаемого результата на всю генеральную совокупность. Выбор того или иного уровня значимости в большинстве случаев является произвольным. Для медицинских и биологических исследований допустимой является доверительная вероятность не менее 95%, соответственно уровень значимости не более 0,05. В тех случаях, когда необходима особая уверенность в полученных результатах, в качестве критического уровня статистической значимости, принимается p=0,01 или p=0,001.

Интерпретировать уровни статистической значимости следует так:

$p \geq 0,1$ – данные согласуются с нулевой гипотезой;

$p \geq 0,05$ – есть сомнения в истинности как нулевой так и альтернативной гипотезы;

$p < 0,05$ – нулевая гипотеза может быть отвергнута;

$p \leq 0,01$ – нулевая гипотеза может быть отвергнута, сильный довод.

$p \leq 0,001$ – нулевая гипотеза не подтверждается, очень сильный довод.

Исследователю необходимо различать клиническую и статистическую значимость полученных результатов. Целесообразно, употреблять термин «клиническая важность». Клинически важное заключение – это заключение о том, что при тестируемом методе лечения есть последствия для здоровья пациента. Статистически значимое заключение основано на вероятности. Статистическая значимость отражает влияние случая на результат. Клиническая важность отражает биологическую ценность полученных в ходе исследования данных. Ошибочно делать вывод о «тенденции к значимости» при клинически важных, но статистически не значимых результатах. Результаты исследования не могут ни «стремиться к значимости», ни «приближаться к значимости». Вместо использования этих выражений следует указать отмеченную разность и 95% доверительный интервал для неё.

ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ ДЛЯ ЧАСТОТ И ДОЛЕЙ

Для описания качественных данных их интервальная оценка предпочтительнее точечной для описания частоты встречаемости изучаемой характеристики в генеральной совокупности. Действительно, поскольку исследования проводятся с использованием выборочных данных, проекция результатов на генеральную совокупность должна содержать элемент неточности выборочной оценки. Доверительный интервал представляет собой меру точности оцениваемого параметра. Интересно, что в некоторых книгах по основам статистики для медиков тема доверительных интервалов для частот полностью игнорируется. В данной статье мы рассмотрим

несколько способов расчета доверительных интервалов для частот, подразумевая такие характеристики выборки, как бесповторность и репрезентативность, а также независимость наблюдений друг от друга. Под частотой в данной статье понимается не абсолютное число, показывающее, сколько раз встречается в совокупности то или иное значение, а относительная величина, определяющая долю участников исследования, у которых встречается изучаемый признак.

В биомедицинских исследованиях чаще всего используются 95 % доверительные интервалы. Данный доверительный интервал представляет собой область, в которую попадает истинное значение доли в 95 % случаев. Другими словами, можно с 95 % надежностью сказать, что истинное значение частоты встречаемости признака в генеральной совокупности будет находиться в пределах 95 % доверительного интервала.

В большинстве пособий по статистике для исследователей от медицины сообщается [3, 6, 7–10, 16], что ошибка частоты рассчитывается с помощью формулы

$$s_p = \sqrt{\frac{p(1-p)}{N}},$$

где p – частота встречаемости признака в выборке (величина от 0 до 1). В большинстве отечественных научных статей указывается значение частоты встречаемости признака в выборке (p), а также ее ошибка (s) в виде p ± s. Целесообразнее, однако, представлять 95 % доверительный интервал для частоты встречаемости признака в генеральной совокупности, который будет включать значения от

$$p - 1.96\sqrt{\frac{p(1-p)}{N}} \quad \text{до} \quad p + 1.96\sqrt{\frac{p(1-p)}{N}}.$$

В некоторых пособиях рекомендуется при малых выборках заменять значение 1,96 на значение t для N – 1 степеней свободы, где N – количество

наблюдений в выборке. Значение t находится по таблицам для t-распределения, имеющимся практически во всех пособиях по статистике. Использование распределения t для метода Вальда не дает видимых преимуществ по сравнению с другими методами, рассмотренными ниже, и потому некоторыми авторами не приветствуется.

Представленный выше метод расчета доверительных интервалов для частот или долей носит имя Вальда в честь Авраама Вальда (Abraham Wald, 1902–1950), поскольку широкое применение его началось после публикации Вальда и Вольфовица в 1939 году. Однако сам метод был предложен Пьером Симоном Лапласом (1749–1827) еще в 1812 году.

Метод Вальда очень популярен, однако его применение связано с существенными проблемами. Метод не рекомендуется при малых объемах выборок, а также в случаях, когда частота встречаемости признака стремится к 0 или 1 (0 % или 100 %) и просто невозможно для частот 0 и 1. Кроме того, аппроксимация нормального распределения, которая используется при расчете ошибки, «не работает» в случаях, когда $n \cdot p < 5$ или $n \cdot (1 – p) < 5$. Более консервативные статистики считают, что $n \cdot p$ и $n \cdot (1 – p)$ должны быть не менее 10. Более детальное рассмотрение метода Вальда показало, что полученные с его помощью доверительные интервалы в большинстве случаев слишком узки, то есть их применение ошибочно создает слишком оптимистичную картину, особенно при удалении частоты встречаемости признака от 0,5, или 50 %. К тому же при приближении частоты к 0 или 1 доверительный интревал может принимать отрицательные значения или превышать 1, что выглядит абсурдно для частот. Многие авторы совершенно справедливо не рекомендуют применять данный метод не только в уже упомянутых случаях, но и тогда, когда частота встречаемости признака менее 25 % или более 75 %. Таким образом, несмотря на простоту расчетов, метод Вальда может применяться лишь в очень ограниченном числе случаев. Зарубежные исследователи более категоричны в своих выводах и однозначно

рекомендуют не применять этот метод для небольших выборок, а ведь именно с такими выборками часто приходится иметь дело исследователям-медикам.

При частотах, не превышающих 25 % или превышающих 75 %, отечественные авторы рекомендуют рассчитывать доверительный интервал с помощью arcsin-преобразования (оно также часто упоминается как угловое преобразование Фишера), при котором сначала рассчитывается вспомогательная переменная (φ) по формуле:

$$\varphi = 2\arcsin\sqrt{p},$$

где p – выборочное значение частоты встречаемости признака. Затем рассчитывается стандартная ошибка вспомогательной переменной по формуле:

$$s_\varphi = \frac{1}{\sqrt{N}}.$$

Поскольку новая переменная имеет нормальное распределение, нижняя и верхняя границы 95 % доверительного интервала для переменной φ будут равны φ-1,96 s_φ и φ+1,96 s_φ соответственно, а 95 % доверительный интервал для частоты встречаемости признака в генеральной совокупности будет

от $\sin^2\frac{\varphi - 1.96 s_\varphi}{2}$ до $\sin^2\frac{\varphi + 1.96 s_\varphi}{2}$.

Вместо 1,96 для малых выборок рекомендуется подставлять значение t для N – 1 степеней свободы. Данный метод не дает отрицательных значений и позволяет более точно оценить доверительные интервалы для частот, чем метод Вальда. Кроме того, он описан во многих отечественных справочниках по медицинской статистике, что, правда, не привело к его широкому использованию в медицинских исследованиях. Расчет

доверительных интервалов с использованием углового преобразования не рекомендуется при частотах, приближающихся к 0 или 1.

На этом описание способов оценки доверительных интервалов в большинстве книг по основам статистики для исследователей-медиков обычно заканчивается, причем эта проблема характерна не только для отечественной, но и для зарубежной литературы. Оба метода основаны на центральной предельной теореме, которая подразумевает наличие большой выборки.

Принимая во внимание недостатки оценки доверительных интервалов с помощью вышеупомянутых методов, Клоппер (Clopper) и Пирсон (Pearson) предложили в 1934 году способ расчета так называемого точного доверительного интервала с учетом биномиального распределения изучаемого признака. Данный метод доступен во многих онлайн-калькуляторах, однако доверительные интервалы, полученные таким образом, в большинстве случаев слишком широки. В то же время этот метод рекомендуется применять в тех случаях, когда необходима консервативная оценка. Степень консервативности метода увеличивается по мере уменьшения объема выборки, особенно при $N < 15$. А. Н. Герасимов описывает применение функции биномиального распределения для анализа качественных данных с использованием MS Excel, в том числе и для определения доверительных интервалов, однако расчет последних для частот в электронных таблицах не «затабулирован» в удобном для пользователя виде, а потому, вероятно, и не используется большинством исследователей.

По мнению многих статистиков, наиболее оптимальную оценку доверительных интервалов для частот осуществляет метод Уилсона (Wilson), предложенный еще в 1927 году, но практически не используемый в отечественных биомедицинских исследованиях. Данный метод не только позволяет оценить доверительные интервалы как для очень малых и очень

больших частот, но и применим для малого числа наблюдений. В общем виде доверительный интервал по формуле Уилсона имеет вид

от

$$\frac{p+\frac{z^2_{1-\alpha/2}}{2N}-z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{N}+\frac{z^2_{1-\alpha/2}}{4N^2}}}{1+\frac{z^2_{1-\alpha/2}}{N}}$$

до

$$\frac{p+\frac{z^2_{1-\alpha/2}}{2N}+z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{N}+\frac{z^2_{1-\alpha/2}}{4N^2}}}{1+\frac{z^2_{1-\alpha/2}}{N}},$$

где $z_{1-\alpha/2}$ принимает значение 1,96 при расчете 95 % доверительного интервала, N – количество наблюдений, а p – частота встречаемости признака в выборке. Данный метод доступен в онлайн-калькуляторах, поэтому его применение не является проблематичным. В. А. Медик и М. С. Токмачев не рекомендуют использовать этот метод при $n \cdot p < 4$ или $n \cdot (1 – p) < 4$ по причине слишком грубого приближения распределения p к нормальному в такой ситуации, однако зарубежные статистики считают метод Уилсона применимым и для малых выборок.

Считается, что помимо метода Уилсона метод Вальда с коррекцией по Агрести – Коуллу также дает оптимальную оценку доверительного интервала для частот. Коррекция по Агрести – Коуллу представляет собой замену в формуле Вальда частоты встречаемости признака в выборке (p) на p`, при расчете которой к числителю добавляется 2, а к знаменателю добавляется 4, то есть $p` = (X + 2) / (N + 4)$, где X – количество участников исследования, у которых имеется изучаемый признак, а N – объем выборки. Такая модификация приводит к результатам, очень похожим на результаты применения формулы Уилсона, за исключением случаев, когда частота

события приближается к 0 % или 100 %, а выборка мала. Кроме вышеупомянутых способов расчета доверительных интервалов для частот были предложены поправки на непрерывность как для метода Вальда, так и для метода Уилсона для малых выборок, однако исследования показали, что их применение нецелесообразно.

Рассмотрим применение вышеописанных способов расчета доверительных интервалов на двух примерах. В первом случае мы изучаем большую выборку, состоящую из 1 000 случайно отобранных участников исследования, из которых 450 имеют изучаемый признак (это может быть фактор риска, исход или любой другой признак), что составляет частоту 0,45, или 45 %. Во втором случае исследование проводится с использованием малой выборки, допустим, всего 20 человек, причем изучаемый признак имеется всего у 1 участника исследования (5 %). Доверительные интервалы по методу Вальда, по методу Вальда с коррекцией по Агрести – Коуллу, по методу Уилсона рассчитывались с помощью онлайн-калькулятора, разработанного Jeff Sauro (http://www.measuringusability.com/wald.htm). Доверительные интервалы по методу Уилсона с поправкой на непрерывность рассчитывались с помощью калькулятора, предложенного порталом Wassar Stats: Web Site for Statistical Computation (http://faculty.vassar.edu/lowry/prop1.html). Расчеты с помощью углового преобразования Фишера производились «вручную» с использованием критического значения t для 19 и 999 степеней свободы соответственно. Результаты расчетов представлены в таблице 5 для обоих примеров.

Как видно из таблицы, для первого примера доверительный интервал, рассчитанный по «общепринятому» методу Вальда заходит в отрицательную область, чего для частот быть не может. К сожалению, подобные казусы нередки в отечественной литературе. Традиционный способ представления данных в виде частоты и ее ошибки частично маскирует эту проблему. Например, если частота встречаемости признака (в процентах) представлена

как 2,1 ± 1,4, то это не настолько «режет глаз», как 2,1 % (95 % ДИ: –0,7; 4,9), хоть и обозначает то же самое. Метод Вальда с коррекцией по Агрести – Коуллу и расчет с помощью углового преобразования дают нижнюю границу, стремящуюся к нулю. Метод Уилсона с поправкой на непрерывность и «точный метод» дают более широкие доверительные интервалы, чем метод Уилсона. Для второго примера все методы дают приблизительно одинаковые доверительные интервалы (различия появляются только в тысячных), что неудивительно, так как частота встречаемости события в этом примере не сильно отличается от 50 %, а объем выборки достаточно велик.

Таблица 5. Доверительные интервалы, рассчитанные шестью разными способами для двух примеров, описанных в тексте

Способ расчета доверительного интервала	95% ДИ для X=1, N=20, P=0,0500, или 5%	95% ДИ для X=450, N=1000, P=0,4500, или 45%
Вальда	–0,0455–0,2541	0,4192–0,4810
Вальда с коррекцией по Агрести – Коуллу	<,0001–0,2541	0,4194–0,4810
Уилсона	0,0089–0,2361	0,4194–0,4810
Уилсона с коррекцией на непрерывность	0,0026–0,2694	0,4189–0,4815
«Точный метод» Клоппера – Пирсона	0,0013–0,2487	0,4189–0,4814
Угловое преобразование	<0,0001–0,1967	0,4193–0,4809

Для обучающихся, заинтересовавшихся данной проблемой, можно порекомендовать работы R. G. Newcombe и Brown, Cai и Dasgupta, в которых приводятся плюсы и минусы применения 7 и 10 различных методов расчета доверительных интервалов соответственно. Из отечественных пособий рекомендуется книга В. А. Медика и М. С. Токмачева, в которой помимо подробного описания теории представлены методы Вальда, Уилсона, а также

способ расчета доверительных интервалов с учетом биномиального распределения частот. Кроме бесплатных онлайн-калькуляторов (http://www.measuringusability.com/wald.htm и http://faculty.vassar.edu/lowry/prop1.html) доверительные интервалы для частот (и не только!) можно рассчитывать с помощью программы CIA (Confidence Intervals Analysis), которую можно загрузить с http://www.medschool.soton.ac.uk/cia/.

Использован материал А. М. Гржибовский. Доверительные интервалы для частот и долей. Экология человека. 2008. №5. С. 57-60.

Вопросы для подготовки к занятию

1. Определение статистики. Основные разделы и область применения медико-биологической статистики.
2. Объект исследования, единица наблюдения, учетные признаки.
3. Определение статистической, генеральной и выборочной совокупностей.
4. Теория вероятности и закон больших чисел.
5. Предельная ошибка исследования, методика ее расчета.
6. Основные типы распределения признака в статистической совокупности. Какой тип распределения признака чаще всего встречается в медицинской и биологической практике?
7. Отличие эмпирического распределения от нормального.
8. Основные требования к оформлению статистических таблиц.
9. Основные свойства статистической совокупности и способы их оценки.
10. Репрезентативность и рандомизация.
11. Доверительный интервал и доверительная вероятность.
12. Абсолютные и относительные величины.

13. Методика вычисления интенсивного, экстенсивного показателей, показателей соотношения и наглядности.
14. Характеристика качественных, бинарных и порядковых признаков.
15. Доверительный интервал для средней и относительной величин.
16. Сравнение совокупностей с использованием доверительных интервалов. Интерпретация результатов сравнения.

Тесты

1. Статистическая совокупность - это:

а) группа определенных признаков

б) группа объектов, обладающих признаками сходства и различия

в) группа относительно однородных элементов (единиц наблюдения), взятых в единых границах времени и пространства

г) группа явлений, объединенных в соответствии с целью исследования

2. Первичным элементом статистической совокупности является:

а) объект наблюдения

б) признак

в) единица наблюдения

г) группа признаков

3. Единица наблюдения в статистической совокупности - это:

а) признак

б) первичный элемент совокупности, обладающий учитываемыми признаками

в) группа признаков

г) заболевание

4. Единица совокупности – это:

а) описка по рассеянности или невнимательности

б) первичный элемент объекта статистического наблюдения, являющийся носителем признаков, подлежащих регистрации

в) разметка бланков по условным знакам

г) первичный элемент, из которого состоит вся наблюдаемая статистическая совокупность

5. Признак - это:

а) объект статистического исследования

б) первичный элемент стат. совокупности

в) свойство, проявлением которого один предмет отличается от другого

г) характеристика статистической совокупности

6. К качественным признакам относятся:

а) рост

б) пол

в) масса тела

г) жизненная емкость легких

7. К количественным признакам относятся:

а) рост

б) пол

в) исход заболевания

г) вид заболевания

8. Выборочная совокупность это:

а) группа, состоящая из относительно однородных элементов, взятых в единых границах времени и пространства

б) совокупность, состоящая из всех единиц наблюдения, которые могут быть к ней отнесены в соответствии с целью исследования

в) часть генеральной совокупности, отобранная специальными методами и предназначенная для ее характеристики

г) всех единиц наблюдения, которые могут быть отнесены к ней в соответствии с целью исследования

9. Репрезентативность - это:

а) достаточный объем генеральной совокупности

б) достаточный объем выборочной совокупности

в) непохожесть выборочной совокупности на генеральную

г) способность выборочной совокупности наиболее полно представлять генеральную

10. Репрезентативность выборочной совокупности по отношению к генеральной обеспечивает:

а) обязательное соблюдение временных границ

б) достаточный объем наблюдений

в) оценка показателей в динамике

г) обязательное соблюдение пространственных границ

11. Достоинства средней величины состоят в том, что она:

а) позволяет анализировать большое число наблюдений

б) позволяет выявить закономерности при малом числе наблюдений и большом разбросе показателей

в) позволяет с помощью одного числа получить представления о совокупности массовых явлений

г) позволяет с помощью одного числа получить представления о распространенности массовых явлений

12. Единица наблюдения определяется в зависимости от:

а) программы исследования

б) плана исследования

в) цели и задач исследования

г) количества наблюдений

13. Вариационный ряд - это:

а) ряд числовых измерений признака, расположенных в ранговом порядке и характеризующихся определенной частотой

б) ряд цифровых значений различных признаков

в) генеральная совокупность

г) ряд чисел, отражающих частоту (повторяемость) цифровых значений изучаемого признака

14. Средняя арифметическая - это:

а) варианта с наибольшей частотой

б) разность между наибольшей и наименьшей величиной

в) обобщающая величина, характеризующая размер варьирующего признака совокупности

г) варианта, находящаяся в середине ряда

15. Медиана – это:

а) варианта с наибольшей частотой

б) разность между наибольшей и наименьшей величиной

в) обобщающая величина, характеризующая размер варьирующего признака совокупности

г) варианта, находящаяся в середине ряда

16. Мода – это:

а) варианта с наибольшей частотой

б) разность между наибольшей и наименьшей величиной

в) обобщающая величина, характеризующая размер варьирующего признака совокупности

г) варианта, находящаяся в середине ряда

Ситуационные задачи

I. Для выполнения задания выберете в электронном файле «База данных инфекционной заболеваемости» лист «Тема 1. Задача 1». Определите интенсивный показатель, экстенсивный показатель, показатель соотношения и наглядности, рассчитайте 95% доверительные интервалы по методам Уилсона и Вальда для показателей заболеваемости.

II. Оцените, с использованием 95%ДИ, отличается ли уровень заболеваемости в городах и районах Кемеровской области от общероссийского показателя. Сделайте выводы. Население Кемеровской области 2 694 900, РФ –

146 780 720. Уровень заболеваемости туберкулезом в РФ 48,3 на 100 000 населения, Кузбассе – 83,6 на 100 000 населения

III. Для выполнения задания III выберете в файле «База данных инфекционной заболеваемости» лист «Тема 1. Задача 3». Рассчитайте долю лиц имеющих инфекционные, паразитарные, неинфекционные заболевания в двух выборочных совокупностях. Сравните, отличаются ли эти две совокупности по частоте встречаемости заболеваемости с помощью доверительных интервалов.

Тема 2. Проверка гипотез. Анализ мощности и оценка объема выборки

Цель занятия: получить представление о статистических гипотезах и мощности исследования.

Учебно-целевые задачи:

- Ознакомиться со статистическими гипотезами их видами, критериями для их проверки.

В результате освоения темы обучающиеся **должны знать:** понятия нулевой и альтернативной гипотезы; односторонней и двухсторонне гипотезы; условиях применения критериев для проверки статистических гипотез; мощности исследования и статистической значимости различий; ошибках первого и второго рода.

В результате освоения темы обучающиеся **должны уметь:** выбирать и применять критерии для проверки статистических гипотез, оценить мощность исследования,

В результате освоения темы обучающиеся **должны владеть:** методами планирования исследования, расчета необходимого объема наблюдений и мощности исследования.

ИНФОРМАЦИОННЫЙ МАТЕРИАЛ ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

При проверке статистических гипотез формулируется две гипотезы: нулевая (H_0) и альтернативная (H_1). Нулевая гипотеза (H_0) – это гипотеза об отсутствии различий между сравниваемыми группами, гипотеза об определенных значениях параметров, или гипотеза о соответствии распределения закону нормального распределения. Альтернативная гипотеза (H_1) – это гипотеза о существовании различий между группами, гипотеза об отличающихся от заданных значений параметров, или гипотеза о несоответствии распределения закону о нормальном распределении.

Нулевая гипотеза должна быть противоположна рабочей гипотезе, которая послужила поводом для проведения исследования. Возможны следующие ситуации при проверке статистических гипотез:

H_0 неверна и отклонена согласно статистическому критерию – истинноположительный результат;

H_0 верна, но ошибочно отклонена согласно статистическому критерию – ложноположительный результат (ошибка первого рода, или α-ошибка);

H_0 не верна, но ошибочно не отклонена согласно статистическому критерию – ложноотрицательный результат (ошибка второго рода или β-ошибка);

H_0 верна и не отклонена согласно статистическому критерию – истинно отрицательный результат.

Вероятность α-ошибки равна доверительной вероятности, значению которой соответствует свой уровень статистической значимости (p).

Вероятность β-ошибки на основании доверительной вероятности рассчитать нельзя, нужно знать, какие именно гипотезы проверяются. При уменьшении доверительной вероятности, уменьшается вероятность ошибок первого рода, но увеличивается вероятность ошибок второго рода. Поэтому доверительную вероятность следует выбирать, принимая во внимание величину ущерба от ошибок первого и второго рода.

При сравнении групп с применением методов проверки статистических гипотез вычисляется значение p. Это достигнутое значение уровня статистической значимости критерия, означающее вероятность ошибки первого рода. Проверка гипотез заключаются в сравнении достигнутого уровня статистической значимости (p) с критическим уровнем. Чаще всего в медицинских и биологических исследованиях в качестве критического уровня значимости принимается p=0,05. Если достигнутый уровень статистической значимости оказывается больше 0,05, нулевую гипотезу не отклоняют. В этом случае различия групп признаются статистически не значимыми. Если же достигнутый уровень статистической значимости менее

0,05 ($p<0,05$), различия являются статистически значимыми, либо статистически высокозначимыми при $p<0,01$.

Как было отмечено ранее, критический уровень статистической значимости следует выбирать с учетом величины ущерба от ошибок первого и второго рода.

Герасимов А.Н. считает, что в случаях, когда ущерб от ошибок первого и второго рода сопоставим, $p=0,05$ действительно разумен. Так, при постановке диагноза ошибка первого рода – поставить неправильный диагноз, ошибка второго рода – отказаться от предполагаемого диагноза и оставить пациента без диагноза. Ущерб от обеих ошибок близок, и выбор «мягкого» критерия правомочен. Если же речь идет о проверке готовности самолета к рейсу, то ошибка первого рода – выпустить в рейс самолет, который разобьется, а ошибка второго рода – не выпустить в рейс самолет, который благополучно долетит. Здесь ущерб от ошибки первого рода много больше, чем от ошибки второго рода, и нужны значительно более жесткие критерии проверки.

Если бы инженеры, так же как и врачи, работали с доверительной вероятностью в 0,05, то они бы строили самолеты, которые разбиваются в каждом двадцатом рейсе, и мосты, которые разваливались бы при прохождении каждого двадцатого поезда. В некоторых случаях выбор $p=0,05$ не просто обоснован, а даже слишком жёсток.

Если в больнице, начинается увеличение заболеваемости внутрибольничными инфекциями, то проведение противоэпидемических мероприятий обосновано с доверительной вероятностью повышения заболеваемости равной 0,1.

Таким образом, исследователь сам принимает решение, какой уровень p принять в качестве критического 0,05;0,01 или 0,001, т.е. допустить α-ошибку в 5%, 1% или 0,1%.

Поэтому рекомендуется указывать точное значение достигнутого уровня статистической значимости (p), а не использовать формат «p<0,05» при описании статистически значимых результатов, или «p>0,05» при описании незначимых различий.

Реброва О.Ю. отмечает, что результаты p=0,051 и p=0,049 следует интерпретировать практически одинаково. Указание точного значения p позволяет читателю самостоятельно интерпретировать статистическую значимость результата. Значения p принято указывать в тексте статей с точностью до трех десятичных знаков, и только в случае, если p меньше 0,001, то в формате «p<0,001», т.е. в формате указания лишь интервальной оценки.

Выбор методов сравнения зависит от многих критериев. Необходимо знать, являются ли группы зависимыми (сопряженными) или независимыми (несопряженными).

Группы считаются зависимыми, если явление изучается в динамике, или это исследование типа «случай-контроль», когда отбор в группы наблюдения осуществляется путем подбора пар.

Группы относятся к независимым, если единицы наблюдения в одну группу набраны независимо от того, какие единицы наблюдения включены в другую.

Например, если ставится задача, сравнить распространенность курения среди медицинских работников, в одну группу могут быть включены средние медицинские работники, в другую – врачи.

Необходимо учитывать количество сравниваемых групп, признаки (количественные или качественные), характер распределения количественных признаков.

На выбор метода влияет и задача, которую исследователь ставит в своем исследовании. Это может быть сравнение групп или оценка взаимосвязи признаков.

Реброва О.Ю. рекомендует к использованию следующие статистические методы проверки статистических гипотез в зависимости от задач и типа данных (табл. 6).

Все перечисленные методы могут быть реализованы в пакетах прикладных статистических программ STATISTICA, SPSS и др.

Таблица 6 – Статистические методы проверки статистических гипотез в зависимости от задач и типа данных.

Задача	Методы	
	Параметрические (для количественных нормально распределенных признаков)	Непараметрические (для количественных признаков независимо от вида распределения, а также для качественных – порядковых или номинальных – признаков)
Выполнение описательной статистики	Вычисление средних значений, средних квадратических отклонений и т.д.	Вычисление медиан и интерквартильных интервалов, пропорций
Сравнение двух независимых групп по одному признаку	t-критерий Стьюдента для независимых выборок	Критерий Манна-Уитни, Колмогорова-Смирнова, Вальда-Вольфовица, χ^2, точный критерий Фишера
Сравнение двух зависимых групп по одному признаку	t-критерий Стьюдента для зависимых выборок	Критерий Вилкоксона, критерий знаков, критерий МакНемара
Сравнение трех независимых групп и более по одному признаку	Дисперсионный анализ (ANOVA)	Дисперсионный анализ (ANOVA) по Краскелу-Уоллису, медианный критерий, критерий χ^2
Сравнение трех зависимых групп и	Критерий Кокрана	ANOVA по Фридману, критерий Кокрана

более по одному признаку		
Анализ взаимосвязи двух признаков	Корреляционный анализ по Пирсону	Критерий χ^2, корреляционный анализ по Спирмену, Кендаллу, гамма и др.
Одновременный анализ трех признаков и более	Регрессионнный анализ, дискриминантный анализ, факторный анализ, кластерный анализ	Логистический регрессионный анализ, логлинейный анализ, анализ древовидных диаграмм, анализ конъюнкций и др.

При представлении результатов исследования необходимо указывать: соответствующую описательную статистику, статистический критерий, который использовался при проверке статистических гипотез, достигнутый уровень статистической значимости, число наблюдений.

При использовании параметрических методов приводятся аргументированные доводы о выполнении условий применимости методов. Характер распределения признака должен соответствовать закону нормального распределения (указывается критерий, с использованием которого проводилась проверка нормальности распределения).

Вторым необходимым условием применения параметрических методов является равенство дисперсий. Равенство дисперсий оценивается с использованием критерия Левена. Если выявлены различия дисперсий, принимаются во внимание только значение p для критерия с раздельными оценками дисперсий.

Сравнение групп с помощью доверительных интервалов и методов проверки статистических гипотез дополняют друг друга, поэтому рекомендуется представлять результаты применения этих методов одновременно.

МОЩНОСТЬ ИССЛЕДОВАНИЯ

В соответствии с принципами доказательной медицины врач принимает решения не только на основании своего личного опыта, но и на результатах исследований, опубликованных в научной литературе. Нередко перед доктором ставится задача проведения самостоятельного исследования и представления достоверных результатов широкой медицинской аудитории. Первый вопрос, который возникает у начинающего врача-исследователя «Сколько пациентов, респондентов включить в исследование, сколько отобрать проб для анализа и др.?».

Познакомимся с методикой подготовки и планировании исследования. Предполагается, что обучающийся знаком с элементарными понятиями статистики и основными видами дизайна эпидемиологических исследований. Данная статья посвящена лишь некоторым технологиям расчета необходимого объема наблюдений и дает представление о том, что решение вопроса о необходимом количестве данных не является делом одной формулы.

Расчет необходимого объема наблюдений - это одна из существенных составляющих исследования, которая дает право говорить о том, что результаты являются достоверными. Нередко происходит подмена понятий «статистическая значимость» и «достоверность». Следует сказать, что каждый термин в статистике имеет свое значение и смысл, поэтому использоваться они должны с пониманием. Неграмотное применение терминологии в научной работе может привести к абсурдным выводам. Достоверность это то, насколько дизайн исследования соответствует его цели и задачам, а результаты являются справедливыми в отношении изучаемого явления. Поэтому оценить, достоверны или нет результаты исследования, может либо читатель, либо рецензент. Формулировка: «Результаты достоверны не менее чем на 0,05» является некорректной. В этом случае речь

идет, скорее всего, не о достоверности, а о статистической значимости различий между сравниваемыми выборками.

«Статистическая значимость различий» или «p-уровень» - еще одна важная величина, которая требуется для расчета необходимого объема наблюдений и, соответственно является важной характеристикой для обеспечения достоверности результатов исследования. Уровень статистической значимости - это расчетная величина критерия, который применяется для проверки статистических гипотез в исследовании. Другими словами, это вероятность отвергнуть нулевую гипотезу, когда она верна, или ошибка первого рода(α). Статистически значимым принимается уровень вероятности меньше чем критический уровень α, который точно задается для каждого исследования. В клинических и эпидемиологических исследованиях он соответствует 0,05. Таким образом, достигнутый в исследовании "p"-уровень, показывает вероятность найти различия там, где их нет. На ошибку первого рода влияет объем наблюдений. Увеличивая объем наблюдений, можно обнаружить статистически значимые различия. Многими исследователями такая ситуация может трактоваться неверно. Например: «Различия не были обнаружены из-за недостатка единиц наблюдения в выборке». Здесь необходимо вспомнить об ошибке второго рода (β), которую выражают через мощность (1-β). В биомедицинских исследованиях общепринятой является мощность 80-90%. Следовательно, 20-10% приходится на вероятность не обнаружения различий, если они в действительности существуют.

Прежде чем формулировать вывод об отсутствии различий следует показать, что исследование спланировано таким образом, что его мощность составила не менее 80%. На мощность влияет объем наблюдений и избыточная мощность такое же неблагоприятное явление, как и недостаточная. Поэтому исследование планируется так, чтобы необходимый

уровень мощности достигался для клинически значимого результата. Что такое клинически важный результат рассмотрим на гипотетическом примере.

Допустим, исследователь проверяет гипотезу (Но) об отсутствии различий в двух несвязанных группах пациентов по нормально распределенному количественному непрерывному признаку (уровень эритроцитов кл/л). Исследуемый признак составил 5,1 (1) кл/л и 5,2 (1) кл/л соответственно для групп №1 и №2. Данные представлены в формате M (s).

Заданные в примере условия позволяют для проверки гипотезы использовать двухсторонний критерий Стьюдента для независимых выборок. При объеме наблюдений n=10 достигнутый уровень статистической значимости критерия составит более чем 0,05. Увеличив объем наблюдений до 1500 различия признаются статистически значимыми при $p<0,05$. Данный результат хоть и является статистически значимым,однако никакой практической ценности = клинической значимости для врача не представляет. Это всего лишь констатация факта, что количество эритроцитов в двух группах пациентов различается на 0,1, и это переделы нормы (4,3-5,5 кл/л).

Именно поэтому на этапе планирования важно определиться, какая же разница в показателях будет клинически значимой. Логично будет предположить, что это такая разница, которая превышает границы нормы. Для рассматриваемого примера клинически значимая разница может быть принята за 1. Известно, что на этапе планирования исследователь не имеет собственных данных как в рассмотренном ранее примере, но он может воспользоваться данными подобных, ранее проведенных исследований, результаты которых были опубликованы в научной литературе. Так, в примере уровень эритроцитов в группе №1 составил 5,1 (1) кл/л. Следовательно, с учетом клинически значимой разницы в 1, можно предположить, что в группе №2 этот показатель составит 6,1 (1).

Для расчета необходимого объема наблюдений воспользуемся формулой:

$n=(A+B)^2 \bullet 2 \bullet S^2 / F^2$,

где n – размер выборки для каждой группы. Общий размер выборки будет в два раза больше;

S – стандартное отклонение;

F–клинически важная разница;

A – значение, зависящее от уровня статистической значимости различий, и при 0,05 составит 1,95, при 0,01 – 2,58;

B – значение, зависящее от мощности, при мощности 80% = 0,84; 90% =1,28.

Тогда, для достижения уровня мощности не менее 80% нужно включить 16 пациентов в группу №1 и 16 пациентов в группу №2. Для расчета необходимого объема наблюдений при нормальном распределении количественного непрерывного признака в двух независимых группах можно использовать также онлайн-калькуляторhttps://www.stat.ubc.ca/~rollin/stats/ssize/n2.html (рис 2).

Далее, при проверке статистической гипотезы о различии сопоставляемых групп по изучаемому признаку, в случае $p>0,05$ можно аргументированно говорить об отсутствия статистически значимых различий выявленных на достаточном объеме наблюдений, а читатель имеет возможность оценить представленные в работе данные как достоверные.

При планировании исследования важно знать и понимать его цель и задачи, какие результаты ожидаются в итоге, какую научную гипотезу проверяет исследователь. Также необходимо определиться с дизайном исследования, анализируемыми данными (качественными и количественными), критериями, которые будут использоваться для проверки

статистических гипотез. Исходя из выбранных критериев, рассчитывается необходимый объем выборки.

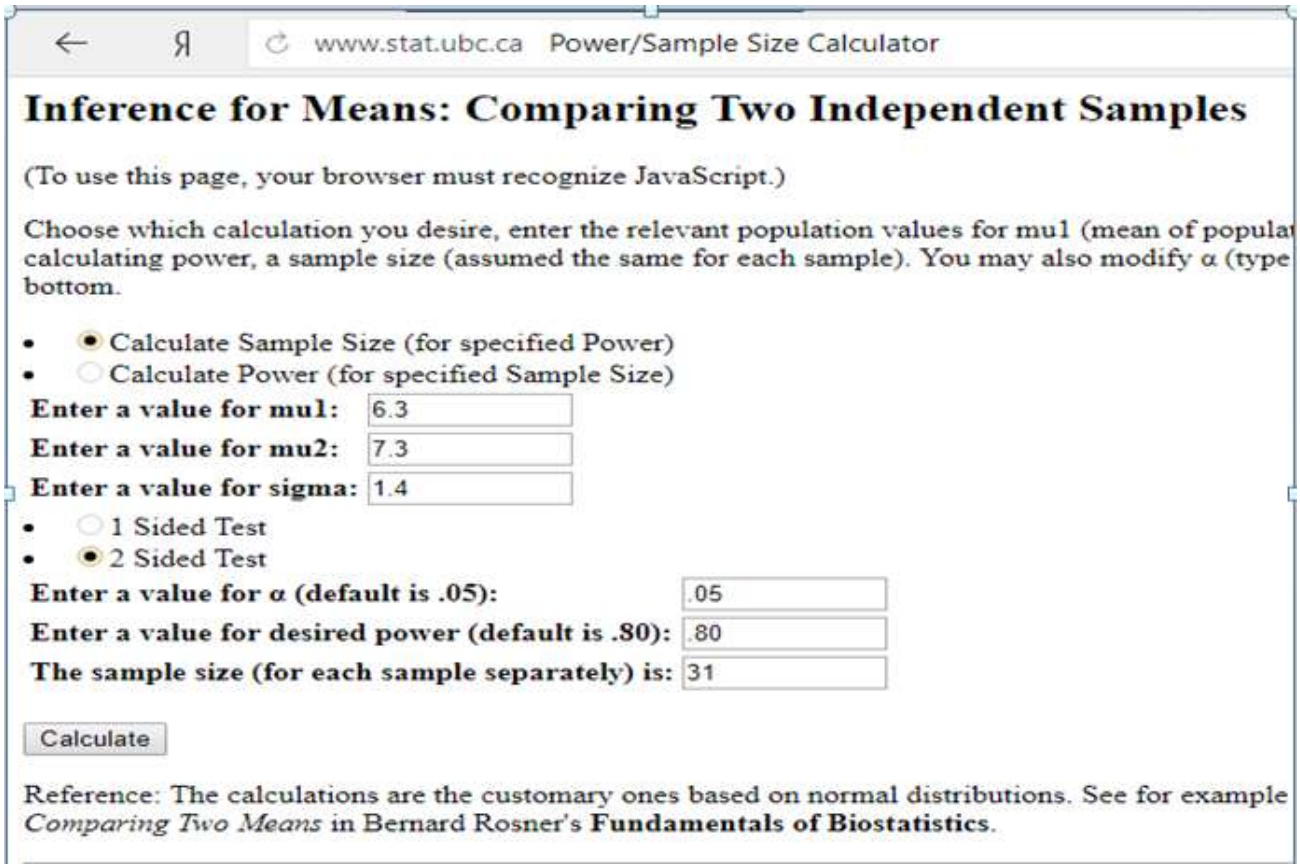


Рис. 2. Онлайн-калькулятор для расчета необходимого объема наблюдений

ОПРЕДЕЛЕНИЕ НЕОБХОДИМОГО ОБЪЕМА НАБЛЮДЕНИЙ ДЛЯ ПРОВЕДЕНИЯ ПОПЕРЕЧНОГО ИССЛЕДОВАНИЯ, ИССЛЕДОВАНИЯ ТИПА «СЛУЧАЙ-КОНТРОЛЬ»

Часто в практике врача-эпидемиолога встречаются поперечные исследования и исследования типа «случай-контроль». Рассмотрим технологию расчета необходимого объема наблюдений для поперечных исследований.

Достаточно высоким является интерес к проблеме приверженности родителей к профилактическим прививкам детей от гриппа. Гипотезу о

низкой приверженности планируется проверить в ходе поперечного исследования, проведенного в г. Кемерово (N=558 973). Важным аспектом проведения данного исследования является обеспечение репрезентативности выборки. Существуют специальные техники отбора единиц наблюдения для выборочного исследования (случайный, механический, серийный методы отбора и др.). Однако вопрос, который требуется решить в первую очередь – «Сколько родителей нужно опросить, чтобы получить достоверную информацию?».

Данные о количестве жителей города можно найти на сайте федеральной службы государственной статистики. Официальный сайт службы Росстата www.gks.ru. Ожидаемая распространенность изучаемого явления (приверженность к прививкам) на этапе планирования может быть получена из обзора литературных источников. Так, по данным исследований, 96% родителей знают о профилактических прививках от гриппа, однако ежегодно прививают своих детей только 33,7% родителей.

Для расчета необходимого объема наблюдений доступна бесплатная программа «EpiInfo™», размещенная на официальном сайте: http://www.cdc.gov/epiinfo. После установки программы необходимо совершить ряд последовательных действий: «StatCalc» → «SampleSizeandPower» → «Populationsurvey». В открывшемся окне ввести показатели численности населения г. Кемерово, частоту изучаемого явления (данные литературы), уровень точности оценки (не менее 5%; рис. 3).

Таким образом, для того чтобы ответить на вопрос с 95% доверительной вероятностью о приверженности населения к прививкам от гриппа нужно опросить не менее 343 родителей детей проживающих в г. Кемерово. Данный факт будет демонстрировать показатель частоты изучаемого явления и 95% доверительный интервал для неё, рассчитанный в ходе статистической обработки данных в последующем. В этом случае

границы доверительного интервала не превысят 10% (5% по обе стороны интервала). Размер доверительно интервала зависит от размера выборки. Чем шире доверительный интервал, тем ниже уровень доверия к результатам исследования.

Рис. 3. Расчет необходимого объема наблюдений с использованием программы «EpiInfo™» для поперечного исследования.

Как правило, исследователь не ограничивается изучением только одного явления, как в рассмотренном примере о приверженности к профилактике гриппа методом вакцинации. В этом случае для каждого изучаемого явления (признака) определяется необходимый объем наблюдений. Только после этого исследователь вправе сказать, что данные получены на достаточном объеме наблюдений. Следует отметить, что такие исследования проводятся с использованием опросника. При заполнении опросного листа респондентами могут возникнуть различные непредвиденные ситуации (утраченные или не заполненные опросные листы

и др.). Поэтому целесообразно увеличить рассчитанный объем наблюдений не менее, чем на 25%.

Взяв за основу литературные данные из предыдущего примера, проверим гипотезу о том, что шансы быть не привитым среди детей перенесших грипп, как минимум в два раза выше, чем у тех, кто не заболел гриппом. В данном случае фактором риска выступает отсутствие прививки от гриппа. В исследованиях «случай-контроль» оцениваются не шансы заболеть или не заболеть, для тех лиц, кто подвергался воздействию фактору риска, а шансы быть или не быть подвергнутым фактору риска в случае наличия заболевания. Проведем расчет необходимого объема наблюдений в программе «EpiInfo™». Для исследований «случай-контроль» выберем «Unmatchedcase-control» (рис. 4).

Для этого установим принятые для медико-биологических исследований 95% доверительный интервал (two-sidedconfidencelevel) и уровень мощности 80% (power). Соотношение «контролей» и «случаев» (ratioofcontrolstocases) может быть различным. Для данного расчета используем наиболее распространенное 1:1. Из литературных данных известно, что доля детей, находящихся под воздействием фактора риска составляет 66,3%. Будем считать, что эпидемиологически важным будет отношение шансов (oddsratio), равное 2, т.е. шансы не иметь прививки от гриппа у заболевших в два раза выше, чем у тех детей, которые не заболели. В итоговой таблице (см. рис. 2) представлены расчеты объема выборки тремя методами (KelseyJ.L., Fleiss., Fleiss с поправкой на непрерывность). Формулировка заключения может быть представлена следующим образом: при мощности 80% и уровне доверительной вероятности 95% выборка объемом 344 ребенка (172 ребенка в группе «случай» и 172 ребенка в группе «контроль») будет достаточной для обнаружения того, что шансы не иметь прививки от гриппа у заболевших детей в два раза выше, чем у тех детей,

которые не заболели. Или: вероятность заболеть гриппом у непривитых детей как минимум, в 2 раза выше, чем у привитых.

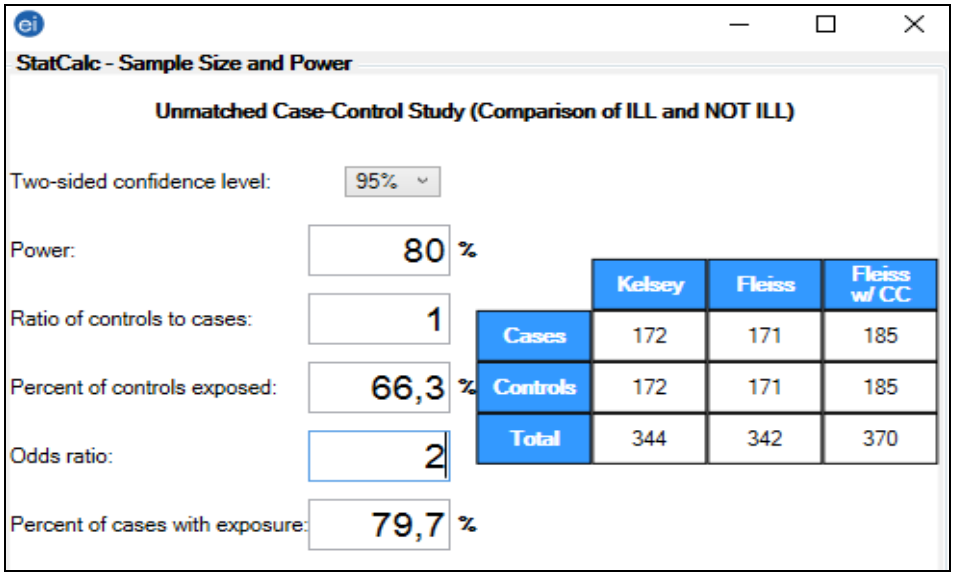


Рис. 4. Расчет необходимого объема наблюдений с использованием программы «EpiInfo™» для исследования «случай-контроль».

Использованы материалы: Гржибовский А. М., Иванов С. В. Поперечные (одномоментные) исследования в здравоохранении / / Наука и Здравоохранение. 2015. № 2. С. 5-18; Крамарь, Л. В. Роль врача педиатра в формировании приверженности родителей к вакцинации детей против гриппа / Л. В. Крамарь, А. Б. Невинский // Детские инфекции. – 2015. – №3. – С. 64-67.

РЕШЕНИЕ ПРОБЛЕМЫ МНОЖЕСТВЕННЫХ СРАВНЕНИЙ

Часто на практике возникает необходимость анализа нескольких групп, т.е. требуется провести межгрупповые множественные сравнения.

Максимально допустимое число сравнений можно определить по формуле: k (k-1)/2, где k – число групп для сравнения.

Нельзя сравнивать несколько средних значений поочередно с помощью параметрических или непараметрических критериев. Увеличение числа сравнений повышает вероятность выявления различий там, где их нет.

Если поочередно сравнивать более 10 выборок, можно с довольно высокой вероятностью найти одно ложное различие.

В случае множественных сравнений рекомендуется вводить поправку Бонферрони. В основе поправки Бонферрони лежит утверждение о том, что если исследователь при числе сравнений, равном k, желает обеспечить вероятность ошибки, равную a, то в каждом из сравнений уровень значимости при отклонении нулевой гипотезы должен быть равен или больше значения a/k.

Таким образом, если исследователь выполнил 5 сравнений, используя критерий Стьюдента, то в любом из этих 5 сравнений уровень значимости должен быть меньше 0,01, чтобы сделать вывод об имеющихся различиях сравниваемых групп с уровнем значимости 0,05. Особенностью критерия Бонферрони является то, что он плохо работает при большом числе сравниваемых групп (k>8).

Существует и такой подход к преодолению проблемы множественных сравнений, как принятие более жесткого уровня статистической значимости 0,01; 0,005; 0,001 и т.д.

Множественные сравнения можно осуществить, используя критерии Тьюки, Ньюмена-Кейлса, Шефе. В компьютерных статистических программах SPSS и Statistica эти критерии реализованы в разделе post – hoc comparisons. Критерий Тьюки, из всех перечисленных, имеет хорошую чувствительность и умеренную жёсткость.

Вопросы для подготовки к занятию

1. Статистическая мощность исследования.
2. Доверительный интервал и доверительная вероятность.
3. Уровень статистической значимости, его интерпретация: $p \geq 0,1$; $p \geq 0,05$; $p < 0,05$; $p \leq 0,01$; $p \leq 0,001$
4. Технология оценки статистической значимости различий с использованием статистических гипотез.
5. Интерпретация истинно положительного, ложноположительного, ложноотрицательного и истинно отрицательного результатов.
6. Вероятность α- и β-ошибок.
7. Параметрические и непараметрические методы оценки статистической значимости различий.
8. Проблема множественных сравнений. Основные подходы к ее решению.

Тесты

1. Выбор подходящего метода сравнения выборочных совокупностей определяется:

а) различиями в характеристиках сравниваемых рядов

б) длинами выборок и максимальным разбросом вариант

в) числом сопоставляемых групп, зависимостью или независимостью выборок, видом распределения признака

г) средними значениями и дисперсиями

2. Примером независимых выборок является:

а) группа пациентов и группа их родственников

б) группа пациентов до и после хирургического вмешательства

в) показатели сахара крови группы пациентов в разные моменты времени

г) результаты двух анкетирований группы пациентов

3. Зависимыми выборками являются:

а) совокупность мужчин и совокупность женщин

б) показатели сахара крови группы пациентов в разные моменты времени

в) больные сахарным диабетом и больные гриппом

г) группа пациентов и группа их родственников

4. Параметрические критерии основаны на:

а) оценке параметров распределения

б) типе распределения

в) выдвигаемых гипотезах

г) требуемой точности

5. Параметрические критерии применимы, если:

а) распределение отличается от нормального

б) требуются достаточно грубые оценки

в) варианты выборок различны

г) численные данные подчиняются нормальному распределению

6. При анализе данных выдвигаются следующие гипотезы:

а) нулевая гипотеза и гипотеза однородности

б) нулевая и альтернативная гипотезы

в) нулевая гипотеза и гипотеза равенства средних

г) гипотеза однородности и гипотеза отсутствия ошибок репрезентативности

7. Если вероятность нулевой гипотезы окажется выше некоторого наперед заданного уровня значимости, то:

а) нулевая гипотеза может быть отвергнута

б) альтернативная гипотеза может быть принята

в) нулевая гипотеза не может быть отвергнута

г) уровень значимости нулевой гипотезы возрастает

8. К параметрическим критериям относятся:

а) критерий Стьюдента и критерий Вилкоксона

б) критерий Вилкоксона и критерий Манна-Уитни

в) критерий Фишера и критерий Манна-Уитни

г) критерий Стьюдента и критерий Фишера

9. Критерий Стьюдента основан на сравнении:

а) частот изучаемого признака в вариационном ряду

б) средних значений выборок

в) числа наблюдений выборок

г) выборочных дисперсий

10. Критерий Фишера основан на сравнении:

а) частот изучаемого признака в вариационном ряду

б) средних значений выборок

в) числа наблюдений выборок

г) выборочных дисперсий

11. Критерий Стьюдента обозначается символом:

а) t

б) U

в) Z

г) F

12. Полученное значение критерия Стьюдента сравнивают с:

а) рассчитанным по формуле значением критерия Стьюдента

б) табличным значением критерия Стьюдента

в) стандартной ошибкой

г) выборочным средним

13. Если полученное значение t-критерия превышает табличное для выбранного уровня значимости alpha = 0,05, это означает что:

а) различие выборочных средних статистически значимо с вероятностью 95 %

б) различие выборочных средних статистически значимо с вероятностью 5%

в) различие выборочных средних статистически незначимо

г) различие выборочных средних статистически значимо с вероятностью 0.95

14. Сходство-различие форм сравниваемых распределений можно определить, пользуясь:

а) критерием Манна-Уитни

б) t-критерием

в) критерием хи-квадрат

г) критерием Вилкоксона

15. Для корректного использования критерия Пирсона объем выборочной совокупности должен быть:

а) не менее10

б) не менее 30

в) не менее 50

г) не менее 150

16. На малых выборках работают:

а) параметрические критерии

б) непараметрические критерии

в) критерии согласия

г) параметрические и непараметрические критерии

17. Степень соответствия эмпирических и теоретических распределений вероятностей, а также двух эмпирических распределений, позволяют определить:

а) непараметрические критерии

б) параметрические и непараметрические критерии

в) параметрические критерии

г) критерии согласия

18. К непараметрическим критериям относятся:

а) критерий Стьюдента и критерий Вилкоксона

б) критерий Вилкоксона и критерий Манна-Уитни

в) критерий Фишера и критерий Манна-Уитни

г) критерий Стьюдента и критерий Фишера

19. Критерий Манна-Уитни это:

а) ранговый критерий для сравнения независимых выборок

б) ранговый критерий для сравнения зависимых выборок

в) параметрический критерий для сравнения независимых выборок

г) параметрический критерий для сравнения зависимых выборок

20. Критерий Вилкоксона это:

а) ранговый критерий для сравнения независимых выборок

б) ранговый критерий для сравнения зависимых выборок

в) параметрический критерий для сравнения независимых выборок

г) параметрический критерий для сравнения зависимых выборок

21. Непараметрические критерии могут быть применены:

а) для данных, имеющих произвольное распределение

б) только для данных, имеющих нормальное распределение

в) только для данных, имеющих распределение Пирсона

г) только для параметров распределения

22. Критерий согласия Пирсона называется:

а) U-критерий

б) t-критерий

в) хи-квадрат

г) Z-критерий

23. Суммарная вероятность нулевой (Н 0)и альтернативной (Н 1) гипотезы,равна:

а) 0

б) 1

в) 5

г) 100

24. Мерой сходства/ различия формы сравниваемых распределений вероятностей, является критерий:

а) Стьюдента

б) Вилкоксона

в) Пирсона

г) Манна-Уитни

25. К ранговым критериям относится:

а) критерий Манна-Уитни

б) критерий Стьюдента

в) критерий Фишера

г) критерий Пирсона

26. Допущение об отсутствии того или иного интересующего исследователя события, явления или эффекта, это:

а) альтернативная гипотеза

б) нулевая гипотеза

в) дизайн исследования

г) погрешность

27. Под альтернативной гипотезой подразумевается:

а) наличие того или иного события, явления или эффекта

б) отсутствие события, явление или эффекта

в) возможность возникновения события

г) погрешность

28. Если вероятность нулевой гипотезы увеличивается, то вероятность альтернативной гипотезы:

а) не изменяется

б) увеличивается

в) равна 1

г) снижается

29. В случае, если максимальное значение одного из сравниваемых выборочных вариационных рядов заведомо меньше минимального значения другого вариационного ряда, то:

а) необходим расчет критерия Стьюдента

б) расчетов с применением критерия Стьюдента не требуется

в) необходим расчет критерия Манна-Уитни

г) необходим расчет критерия Вилкоксона

30. Если набор объектов исследования в каждую из групп осуществлялся независимо от того, какие объекты исследования включены в другую группу, то такие выборки называются:

а) зависимыми

б) независимыми

в) случайные

г) возможные

Ситуационные задачи

I. Проведите оценку различий между относительными величинами с использованием t критерия Стьюдента (файл «База данных инфекционной заболеваемости» лист «Тема 2. Задача 1».). Оцените, есть ли различия между уровнями инфекционной заболеваемости в 2017 и 2018 годах.

II. Известно, что одним из показателей, характеризующих здоровье работающего населения является доля лиц ни разу не болевших в течение года. Рассчитайте долю лиц ни разу не болевших ОРВИ на обследуемых предприятиях «База данных инфекционной заболеваемости» лист «Тема 2. Задача 2» и оцените, достаточным ли является объем наблюдений для получения статистически значимых результатов при условии, что выборка является бесповторной. Всего на предприятиях трудятся 680 человек в рабочих профессиях. Определите также необходимый объем наблюдений.

III. Проверте гипотезу (Но) об отсутствии различий в двух несвязанных группах пациентов по нормально распределенному количественному непрерывному признаку (уровень эритроцитов кл/л). Исследуемый признак составил 8,7 (1) кл/л и 3,2 (1) кл/л соответственно для групп №1 и №2. Данные представлены в формате M (s). Рассчитайте необходимый объем

наблюдений по формуле и с использованием онлайн-калькуляторhttps://www.stat.ubc.ca

IV. Рассчитайте необходимый объем наблюдений для того чтобы ответить на вопрос с 95% доверительной вероятностью о приверженности населения к прививкам от клещевого энцефалита.

V. Рассчитайте необходимый объем наблюдений для исследования дизайна «случай-контроль». Проверяемая гипотеза: шансы встретить непривиых детей в группе заболевших в 2 раза выше, чем в группе не заболевших. Из литературных данных известно, что доля детей, находящихся под воздействием фактора риска заболевания Х составляет 70,3%. Сформулируйте вывод.

Тема 3. Корреляционный анализ. Анализ зависимостей и связей.

Цель занятия: овладеть навыками оценки связей и выявления зависимостей

Учебно-целевые задачи:

- освоить методы корреляционного анализа (Пирсона, Спирмена).
- научиться проводить корреляционно-регрессионный анализ с оценкой значимости коэффициентов и характеристик зависимости, интерпретировать результаты
- научиться прогнозировать заболеваемость в зависимости от влияния различных факторов

В результате освоения темы обучающиеся **должны знать:** методы анализа связей и зависимостей

В результате освоения темы обучающиеся **должны уметь:** интерпретировать результаты корреляционно-регрессионного анализа

В результате освоения темы обучающиеся **должны владеть:** технологией расчета коэффициентов Пирсона, Спирмена; проведения корреляционно-регрессионный анализ в Excel

ИНФОРМАЦИОННЫЙ МАТЕРИАЛ

Среди статистически взаимосвязанных признаков одни могут рассматриваться как определенные факторы, влияющие на изменение других, а вторые – как следствие, или результат изменения первых. Соответственно, первые – это факторные признаки, а вторые – результативные. Связь между двумя переменными X и Y является функциональной, если определенному значению переменной X соответствует строго определенное значение Y. Это жестко детерминированная связь. Но существует и другая взаимосвязь, при которой взаимно действуют многие факторы, неравномерно влияющие на изменение результативного признака. Такие связи являются стохастическими (вероятностными).

Корреляционная связь является частным случаем стохастической связи. Это соотношение, соответствие между средним значением результативного признака и признаками-факторами. При этом если рассматривается связь средней величины результативного показателя Y с одним признаком-фактором X, корреляционная связь называется «парной», а если факторных признаков два и более множественной.

По характеру изменений Y, X в парной корреляции различают прямую (положительная) и обратную (отрицательная) взаимосвязи (рис. 5 а). При прямой связи – с увеличением X возрастает и Y, при обратной – уменьшается. По форме связи она делится на прямолинейные (линейные) и криволинейные (нелинейные); (рис.5). Если направление изменения одной переменной не меняется с возрастанием (или убыванием) другой переменной

то такая взаимосвязь называется монотонной, в противоположном случае – немонотонной (рис. 5.).

Область допустимых значений линейного коэффициента парной корреляции от –1 до +1. Знак коэффициента корреляции указывает направление связи. Если $r_{xy}>0$, то связь прямая; если $r_{xy}<0$, то связь обратная.

Связи между признаками могут быть слабыми и сильными (тесными). Их критерии оцениваются по шкале Чеддока:

$0,1 < r_{xy} < 0,3$ – слабая;

$0,3 < r_{xy} < 0,5$ – умеренная;

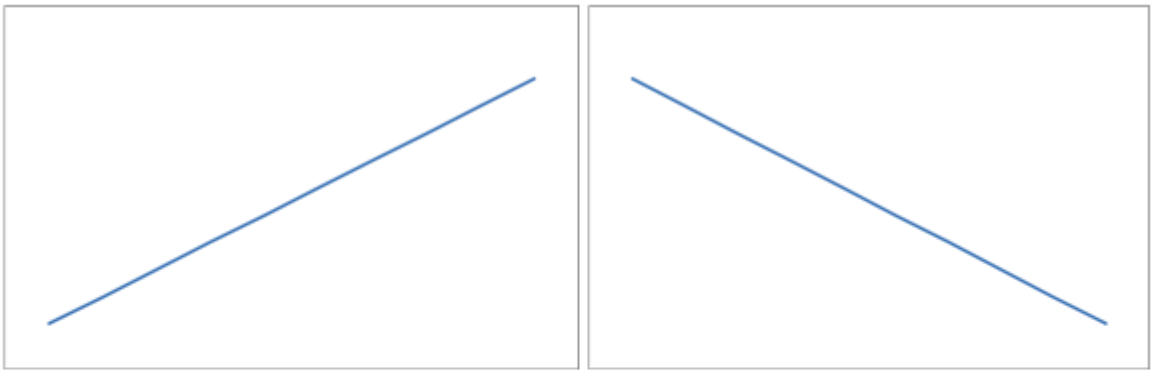
$0,5 < r_{xy} < 0,7$ – заметная;

$0,7 < r_{xy} < 0,9$ – высокая;

$0,9 < r_{xy} < 1$ – весьма высокая;

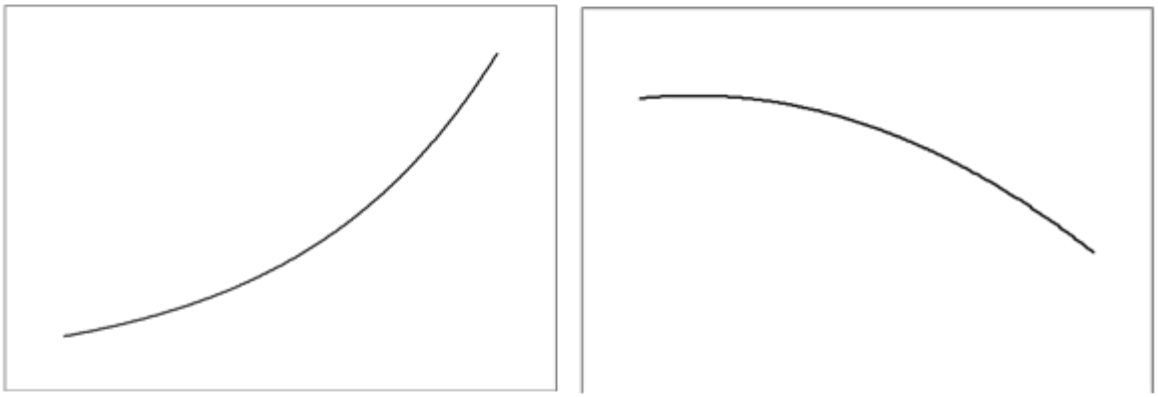
Если данный коэффициент по модулю близок к единице, то связь между признаками может быть интерпретирована как довольно тесная линейная. Если его модуль равен единице, то связь между признаками функциональная линейная. Если признаки x и y линейно независимы, то r_{xy} близок к 0.

В медицине и биологии чаще всего для определения направления и силы связи между явлениями используются коэффициенты Пирсона (r_{xy}) ранговой корреляции Спирмена(ρ_{xy}). Параметрический корреляционный анализ Пирсона используется для установления взаимосвязи признаков, имеющих нормальное распределение.



а. б.

Прямолинейная связь: а) положительная; б) отрицательная



а. б.

Нелинейная связь: а) положительная; б) отрицательная

а. б.

Нелинейная немонотонная связь: а) положительная; б) отрицательная

Рис. 5. Виды связи.

Наглядное представление о характере связи дает диаграмма рассеивания (рис. 6).

Рис. 6. Диаграммы рассеивания: а) r_{xy}=+0,85; б) r_{xy}=0

Диаграмма рассеивания представляет собой график, оси которого соответствуют значениям двух переменных. Точка на графике это каждая единица наблюдения.

Непараметрические коэффициенты (Спирмена, Кендалла, гамма) применяются с целью исследования взаимосвязи количественных признаков независимо от вида их распределения, количественного и качественного признака, двух порядковых признаков. Нередко исследуется взаимосвязь качественных (порядковых) признаков. В этом случае корректнее использовать термин «ассоциация» вместо термина «корреляция». Измерить ассоциацию бинарных данных можно при помощи «φ-коэффициента сопряженности» (Phi).

Изучение корреляционных связей сводится к решению следующих задач:

1) выявление наличия или отсутствия корреляционной связи между изучаемыми признаками, эта задача может быть решена на основе параллельного сопоставления (сравнения) значений X и Y у n единиц совокупности, а также с помощью группировок и путем построения и анализа специальных корреляционных таблиц;

2) измерение тесноты связи между двумя и более признаками с помощью специальных коэффициентов (коэффициентов корреляции, $r_{x,y}$), и эта часть исследований называется «корреляционным анализом»;

3) определение уравнения регрессии – математической модели, в которой среднее значение результативного признака Y рассматривается как функция одной или нескольких переменных факторных признаков X и эта часть исследования носит название «регрессионный анализ». Он включает следующие этапы: выбор формы связи (вида аналитического уравнения регрессии); оценку параметров уравнения; оценку качества аналитического уравнения регрессии.

Также метод регрессионного анализа используется для статистического прогноза – вычисление значения результативного показателя Y для любых значений факторов X и восполнения пропусков в данных.

Общий термин «корреляционно-регрессионный анализ» подразумевает всестороннее исследование корреляционных связей, в том числе и определение уравнений регрессии, измерение тесноты связей, а также определение возможных ошибок как параметров уравнений регрессии, так и показателей тесноты связей.

Наиболее часто для описания статистической связи признаков используется линейная форма. Внимание к линейной связи объясняется четкой интерпретацией ее параметров, ограниченной вариацией переменных и тем, что в большинстве случаев нелинейные формы связи для выполнения расчетов преобразуют (путем логарифмирования или замены переменных) в линейную форму.

Корреляционный анализ по Пирсону

На основе данных таблицы 7 требуется определить зависимость цветного показателя крови (ряд x) от величины эритроцитов (ряд y).

В данном случае числовые значения коррелируемых рядов соответствуют нормальному распределению. Поэтому для определения наличия связи между изучаемыми явлениями целесообразно вычислять коэффициент линейной корреляции Пирсона.

При вычислении коэффициента линейной корреляции Пирсона необходимо помнить, что названный коэффициент является более мощным по сравнению с коэффициентом ранговой корреляции Спирмена, однако область его применения ограничивается нормальным распределением.

Таблица 7. Алгоритм вычисления коэффициента линейной корреляции Пирсона

Условие		Решение (схема)				
x (в ед.)	y (во фл.)	d_x	d_y	d_x^2	d_y^2	$d_x d_y$
0,91	78	- 0,05	- 4	0,0025	16	0,20
0,92	79	- 0,04	- 3	0,0016	9	0,12
0,95	80	- 0,01	- 2	0,0001	4	0,02
0,93	81	- 0,03	- 1	0,0009	1	0,03
0,91	82	- 0,05	0	0,0025	0	0,00
0,98	83	0,02	1	0,0004	1	0,02
0,99	84	0,03	2	0,0009	4	0,06
0,95	82	- 0,01	0	0,0001	0	0,00
0,98	85	0,02	3	0,0004	9	0,06
1,05	86	0,09	4	0,0081	16	0,36
Mx=0,96	My=82,0	$\Sigma d_x=0$	$\Sigma d_y=0$	$\Sigma d_x^2=0,0175$	$\Sigma d_y^2=60,0$	$\Sigma d_x d_y=0,87$

Определение коэффициента линейной корреляции по Пирсону проводят по формуле:

$$r_{xy} = \frac{\Sigma dx d_y}{\sqrt{\Sigma d_x{}^2 \Sigma d_y^2}}$$

гд r_{xy} – коэффициент линейной корреляции;

; $d_x d_y$ – отклонение каждого числового значения от средней величины по ряду x и d в ряду y

Начальным моментом вычисления коэффициента линейной корреляции Пирсона является нахождение средней величины по ряду x (M_x) и ряду y (M_y).

$$M_x = \frac{\Sigma V_x}{n} = \frac{0{,}91 + 0{,}92 + 0{,}96 + \ldots\ldots + 1{,}05}{10} = 0{,}96$$

$$M_y = \frac{\Sigma V—}{n} = \frac{78 + 79 + 80 + \ldots\ldots + 86}{10} = 82$$

$d_x = V_x - M_x = 0{,}91 - 0{,}96 = -0{,}05;\ 0{,}92 - 0{,}96 = -0{,}04$

и т.д.

$d_y = V_y - M_y = 78 - 82 = -4;$ $79 - 82 = -3$ и т.д.

$$r_{xy} = \frac{0{,}87}{\sqrt{60 \times 0{,}0175}} = 0{,}85$$

Проверка значимости коэффициента корреляции Пирсона

Для определения статистической значимости полученного коэффициента используются два метода:

1. Найденный результат r_{xy} сравниваем с критическим значением оценочных таблиц (см. приложение). Критическое значение таблицы при n = 10 соответствует 0,632 - -0,765, следовательно, полученный показатель

является статистически значимым: цветной показатель крови зависит от величины эритроцитов (p<0,001).

2. Определение статистической значимости коэффициента корреляции проводят по формуле:

$$t = \frac{r_{xy}}{m}$$

где: t - коэффициент Стьюдента;

r_{xy} – коэффициент корреляции;

m - ошибка коэффициента корреляции;

n - число коррелируемых пар

$$m = \pm\sqrt{\frac{1 - r_{xy}^{\ 2}}{n-2}}; \quad m = \pm\sqrt{\frac{1-0{,}87^2}{10-2}} = 0{,}17; \quad t = \frac{0{,}87}{0{,}17} = 5{,}1$$

Для оценки статистической значимости критерия t нужно определить число степеней свободы по формуле $f = n - 2 = 10 - 2 = 8$. Критическое значение t = 2,31 – 3,36 – 5,04 (см. табл. 1 приложения). Таким образом, и второй метод оценки позволяет сделать аналогичные выводы, т.е. полученный коэффициент корреляции является статистически значимым; цветной показатель крови зависит от величины эритроцитов (p<0,001).

Алгоритм компьютерной обработки в программе IBM SPSS Statistics:

«Анализ» → «Корреляции» → «Парные» (рис. 7).

В отрывшемся диалоговом окне «Парные корреляции» перенесите в окно «Переменные» количественные переменные ЦП и величина эритроцитов. Выберите коэффициент корреляции Пирсона, критерий значимости двухсторонний и поставьте флажок «Метить значимые корреляции» → «ОК» (рис. 7).

В окне результатов анализа (рис. 8) будет представлена информация о том, что корреляция значима на уровне 0,01. Соответственно, как и при

расчетах «вручную», установлено, что цветной показатель крови зависит от величины эритроцитов (p=0,01).

Если переменные, между которыми анализируется связь, являются качественными, либо хотя бы одна из анализируемых переменных не подчиняется закону нормального распределения, применяются ранговые коэффициенты корреляции (Спирмена или Кенделла).

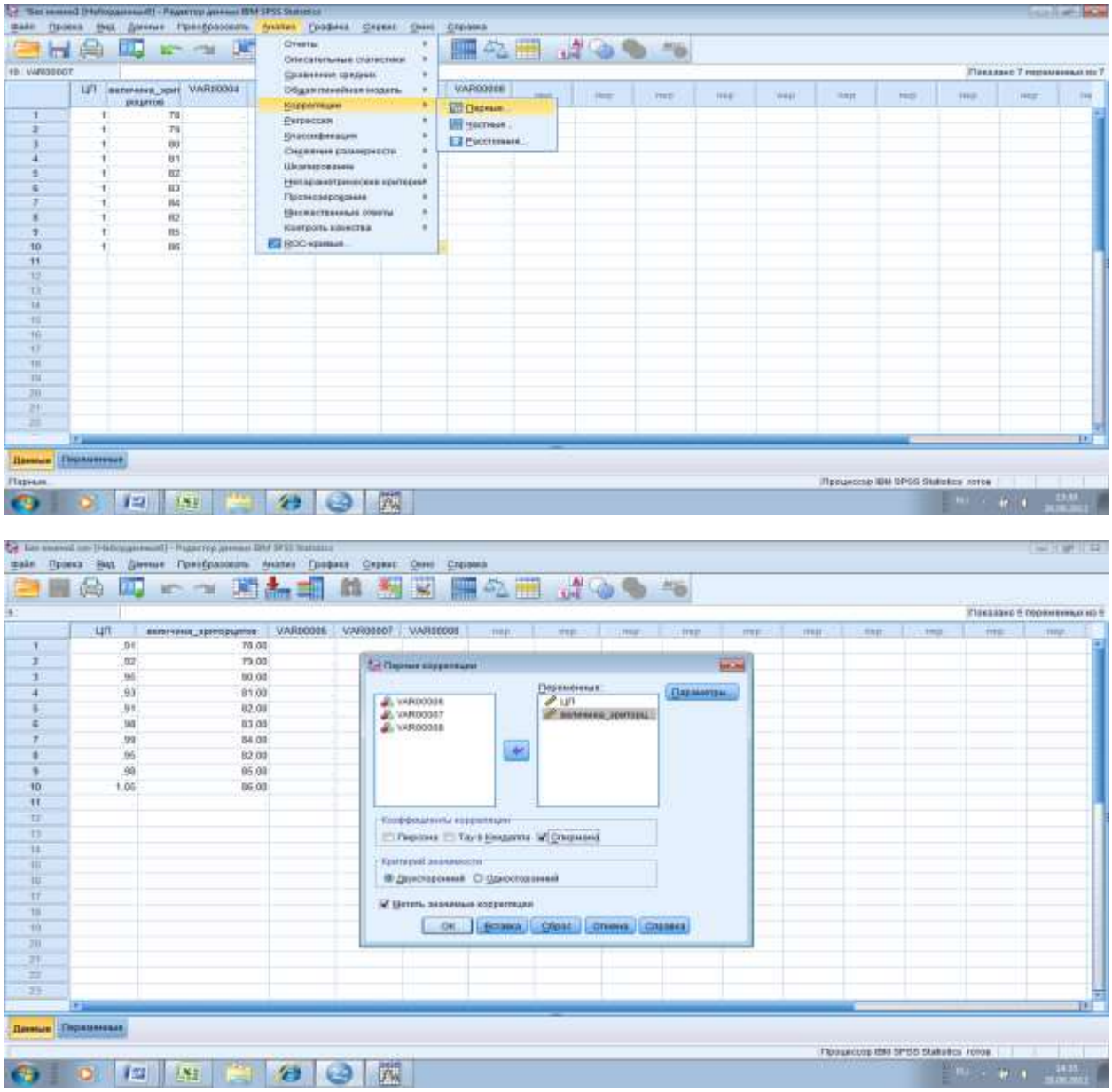


Рис. 7. Настройка корреляционного анализа по Пирсону в программе IBM SPSS Statistics

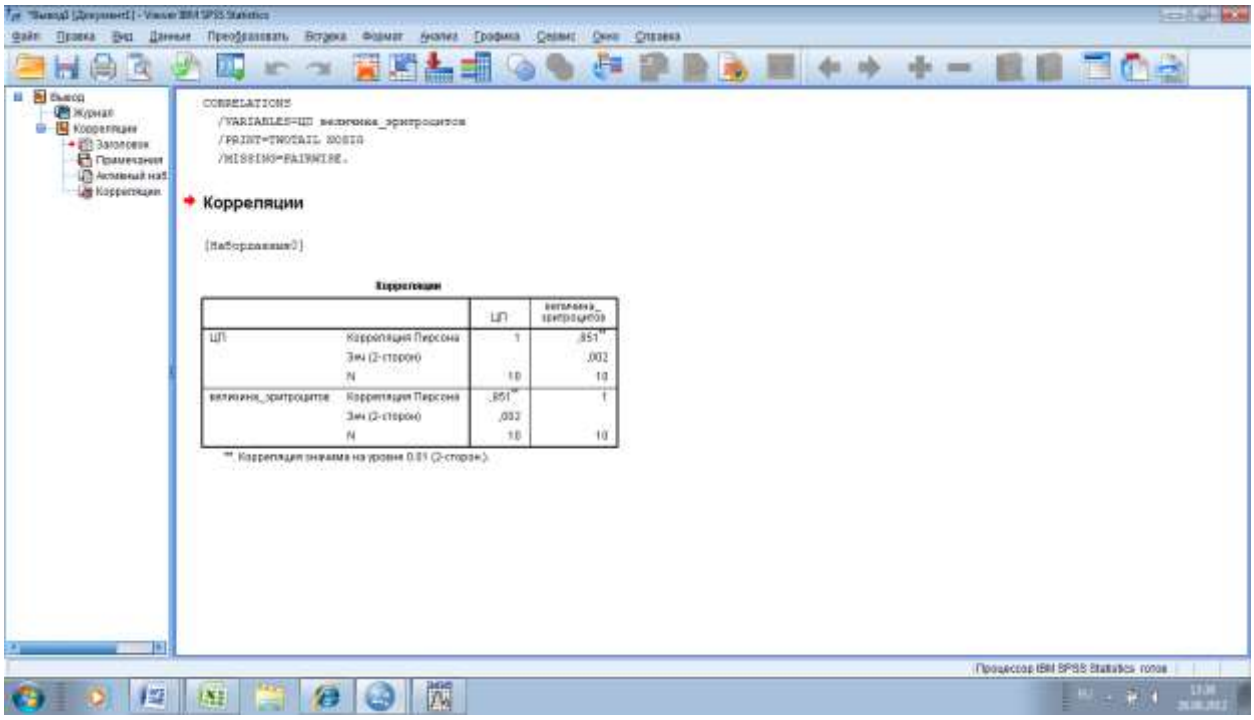


Рис. 8. Результаты корреляционного анализа по Пирсону в программе IBM SPSS Statistics

Корреляционный анализ по Спирмену

На основании данных таблицы 8 требуется определить зависимость цветного показателя (ряд x) от уровня насыщения крови кислородом (ряд y).

В данном случае один из рядов (ряд y) представлен показателями, выраженными в %, что соответствует качественному характеру распределения. Поэтому для определения зависимости между величиной цветного показателя и уровнем насыщения крови кислородом нужно использовать коэффициент ранговой корреляции Спирмена.

При вычислении коэффициента ранговой корреляции Спирмена необходимо учесть, что этот коэффициент является менее мощным критерием по сравнению с коэффициентом линейной корреляции, однако имеет более широкую область применения. Названный коэффициент корреляции используется в тех случаях, когда коррелируемые данные

соответствуют количественному, качественному или порядковому распределениям. Коэффициент корреляции Спирмена рассчитывается по формуле:

$$\rho_{xy} = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$$ где p_{xy}- коэффициент ранговой корреляции Спирмена ;

d - разница между рангами;

n - число коррелируемых пар

Начальным этапом определения коэффициента ранговой корреляции является проведение ранжировки коррелируемых рядов (построение числовых значений рядов x и y по возрастающей или убывающей величине) путем присвоения числовым значениям каждого ряда порядковых номеров. Если числовые значения в ряду одинаковые, то их ранги будут соответствовать среднему значению порядковых мест, которые они занимают. Так в ряду x самая малая величина 0,9 (см. табл. 20) встречаются дважды, следовательно, для определения ее ранга необходимо сложить порядковые места и найти их среднее значение (1+2) : 2 = 1,5; следующим по величине является число 1,0, которое встречается 5 раз. Это число должно занять места с 3 по 7. Для нахождения рангов необходимо: (3 + 4 + 5 + 6 + 7) : 5 = 5 и т.д. Аналогично определяются ранги и для ряда y.

$$\rho_{xy} = 1 - \frac{6 \times 163{,}75}{10 \times (100 - 1)} = 0{,}01$$

Таблица 8. Алгоритм вычисления коэффициента ранговой корреляции Спирмена.

Условие		Решение (схема)			
x (в ед.)	y (в %)	Ранги по ряду *x*	Ранги по ряду *y*	*d*	d^2
1,2	94,0	9,5	9,0	0,5	0,25
1,0	94,2	5,0	10,0	- 5,0	25,00

0,9	93,1	1,5	6,5	- 5,0	25,00
1,0	93,3	5,0	8,0	- 3,0	9,00
0,9	92,5	1,5	5,0	- 4,0	16,00
1,0	92,2	5,0	4,0	1,0	1,0
1,1	93,1	8,0	6,5	1,5	2,25
1,0	91,1	5,0	2,0	3,0	9,0
1,0	91,2	5,0	3,0	2,0	4,0
1,2	90,1	9,5	1,0	8,5	72,25
					Σd^2=163,75

Проверка значимости коэффициента корреляции Спирмена

Найденный результат сравниваем с критическим значением оценочных таблиц. Критическое значение таблицы при n = 10 соответствует 0,64 – 0,79, следовательно, полученный показатель является статистически незначимым, поэтому влияние уровня насыщения крови кислородом на величину цветного показателя не доказано (p>0,05).

Оценку статистической значимости полученного коэффициента корреляции Спирмена можно проводить и с использованием критерия *t*.

$$t = \frac{\delta_{xy}}{m_e};$$ где t - коэффициент Стьюдента;

p_{xy} - коэффициент ранговой корреляции;

m - ошибка коэффициента корреляции;

n - число коррелируемых пар

$$n = \pm\sqrt{\frac{1-\rho_{xy}^{\ 2}}{n-2}};$$

$$m = \pm\sqrt{\frac{1-0{,}0001}{10-2}} = 0{,}35; \qquad t = \frac{0{,}01}{0{,}35} = 0{,}03;$$

Для оценки значимрсти критерия t нужно определить число степеней свободы по формуле: $f = n - 2 = 10 - 2 = 8$. Критическое значение t = 2,31 – 3,36 – 5,04. Таким образом, второй метод оценки позволяет сделать аналогичные выводы. Полученный коэффициент ранговой корреляции по Спирмену является статистически незначимым. Уровень насыщения крови кислородом не влияет на величину цветного показателя (p>0,05).

Алгоритм компьютерной обработки в программе IBM SPSS Statistics:

«Анализ» → «Корреляции» → «Парные». В отрывшемся диалоговом окне «Парные корреляции» перенесите в окно «Переменные» количественные переменные ЦП и насыщение кислородом. Выберите коэффициент корреляции Спирмена, критерий значимости двухсторонний и поставьте флажок «Метить значимые корреляции» → «ОК» (рис. 9). Результаты анализа представлены на рисунке 10.

Рис. 9. Настройка корреляционного анализа по Спирмену в программе IBM SPSS Statistics

Соответственно, как и при расчетах «вручную», установлено, что уровень насыщения крови кислородом не влияет на величину цветного показателя (p=0,907).

Если одна из переменных является качественной, а другая количественной связь между этими переменными изучается путем сравнения групп по уровню выраженности количественной переменной. Предположим, изучается взаимосвязь профессиональной заболеваемости (качественная переменная) и стажа (количественная переменная). Связь между этими переменными может быть изучена путем сравнения со средними значениям стажа группы лиц имеющих проф. заболевание со стажем группы лиц, не имеющих проф. заболевания. При обнаружении статистически значимых различий связь между стажем и формированием профессиональной заболеваемости будет доказана.

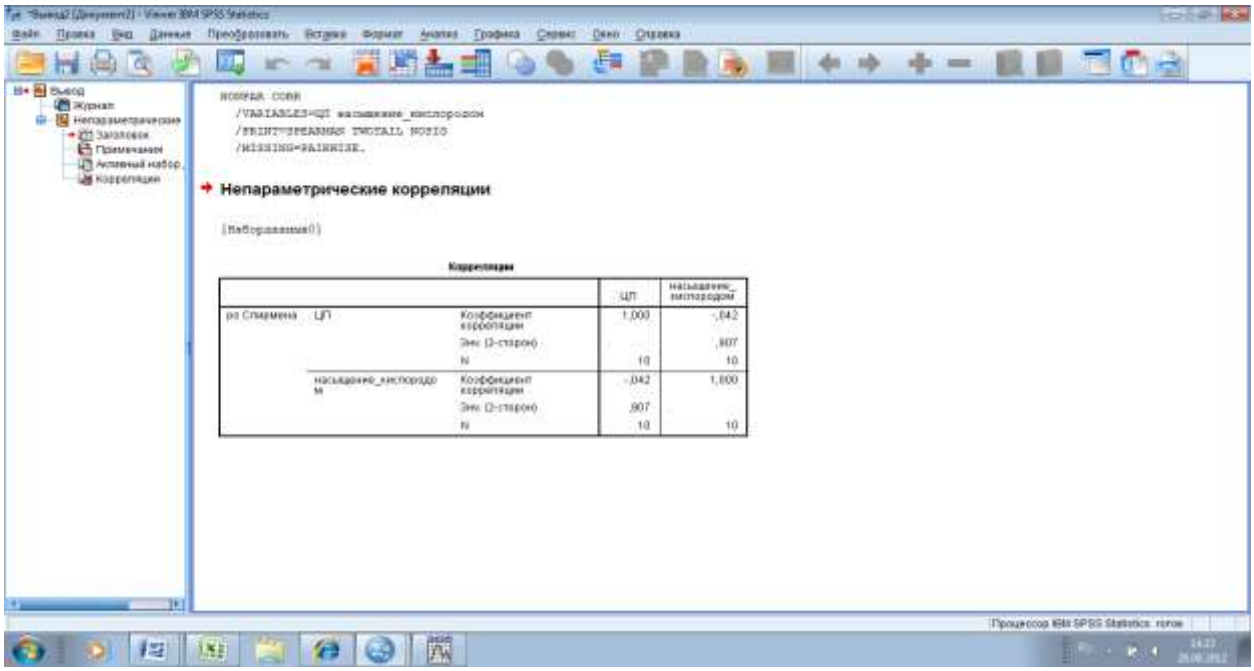


Рис. 10. Результаты корреляционного анализа по Спирмену в программе IBM SPSS Statistics

Коэффициент детерминации.

Коэффициент детерминации R^2 характеризует долю вариации (дисперсии) результативного признака, объясняемую фактором, в общей вариации (дисперсии). Коэффициент детерминации R^2 принимает значения от 0 до 1.

При парной линейной корреляции коэффициент детерминации равен квадрату коэффициента корреляции $R^2=r_{xy}^2$.

Тесноту связи удобно выражать в процентах ($R^2 \times 100\%$).

Вернемся к рассмотренному ранее примеру. При изучении зависимости цветового показателя крови от величины эритроцитов был $r_{xy}=0,85$; $R^2=r_{xy}^2=0,7225$ ($R^2=72,25\%$).

Проверка значимости коэффициента детерминации

Таблицы для критических значений R^2 отсутствуют, поэтому для проверки значимости используется F-критерий Фишера:

$$F = \frac{R^2}{1-R^2} \times (n-2) = \frac{0,7225}{0,2775} \times 8 = 20,8$$

Критический уровень $F_{крит}$ находится по таблице Фишера-Снедекора.

Если выполняется неравенство $F>F_{крит}$, то нулевая гипотеза (отсутствие связи между y и x) отвергается. Число степеней свободы определяется так: $k_1= n_1-1$ и $k_2= n_2-1$, где n_1 – объем выборки с большей исправленной дисперсий, а n_2 – с меньшей. При $k_1=9$ и $k_2=9$, $F_{крит}=3,18$ ($F=20,8 > F_{крит}=3,18$; см. приложение). Следовательно, цветовой показатель крови на 72,25% зависит от величины эритроцитов ($p<0,05$).

КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ В EXCEL

Рассчитаем коэффициент корреляции на конкретном примере. Имеем таблицу, в которой помесячно расписаны в отдельных колонках затраты на

здравоохранение и количество лиц ни разу не болевших. Нам предстоит выяснить степень зависимости состояния здоровья от суммы затрат на здравоохранение.

Способ 1: определение корреляции через Мастер функций

Одним из способов, с помощью которого можно провести корреляционный анализ, является использование функции КОРРЕЛ. Сама функция имеет общий вид **КОРРЕЛ (массив1;массив2)**.

1. Выделяем ячейку, в которой должен выводиться результат расчета. Кликаем по кнопке **«Вставить функцию»**, которая размещается слева от строки формул.

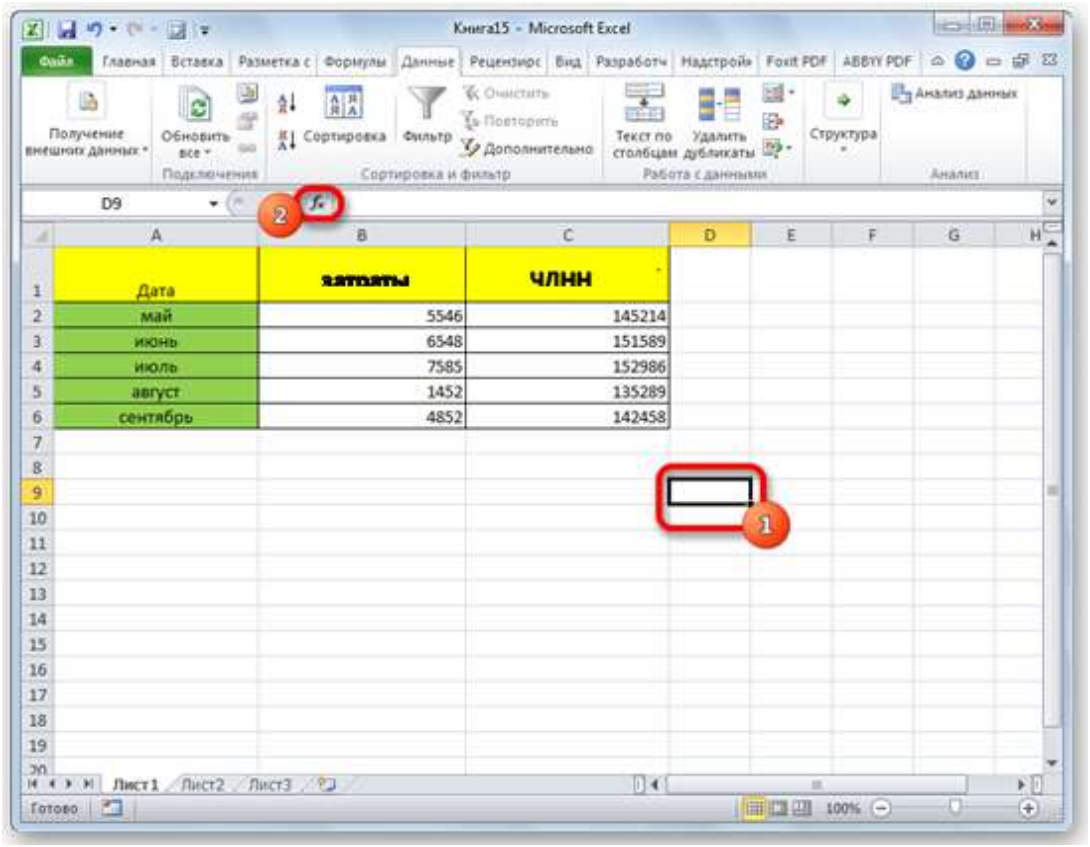


Рис. 11. Опция «Вставить функцию» в Excel

2. В списке, который представлен в окне Мастера функций, ищем и выделяем функцию **КОРРЕЛ**. Жмем на кнопку **«ОК»**.

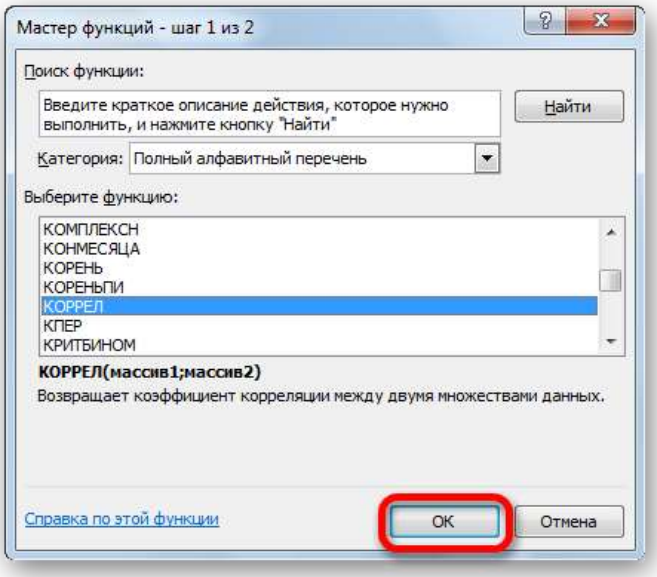


Рис. 12. Выбор функции «Корреляция» в Excel

3. Открывается окно аргументов функции. В поле **«Массив1»** вводим координаты диапазона ячеек одного из значений, зависимость которого следует определить. В нашем случае это будут значения в колонке «ЧЛНН» - частота лиц, ни разу не болевших. Для того, чтобы внести адрес массива в поле, просто выделяем все ячейки с данными в вышеуказанном столбце.

В поле **«Массив2»** нужно внести координаты второго столбца. У нас это затраты на здравоохранение. Точно так же, как и в предыдущем случае, заносим данные в поле.

Жмем на кнопку **«OK»**

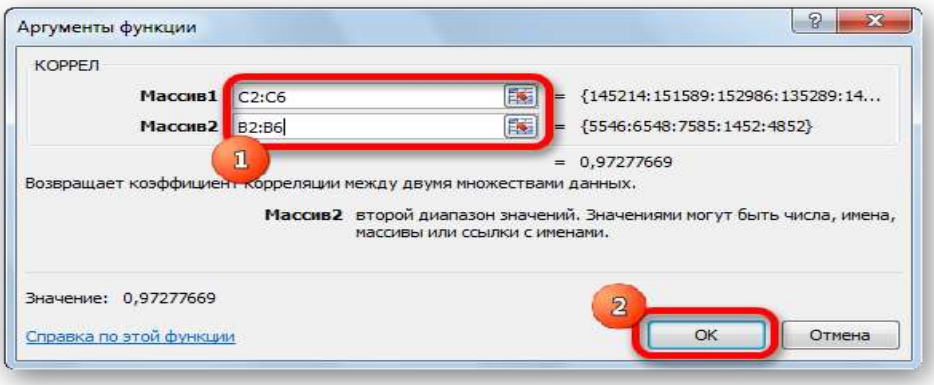


Рис. 13. Применение функции «Корреляция» в Excel

Как видим, коэффициент корреляции в виде числа появляется в заранее выбранной нами ячейке. В данном случае он равен 0,97, что является очень высоким признаком зависимости одной величины от другой.

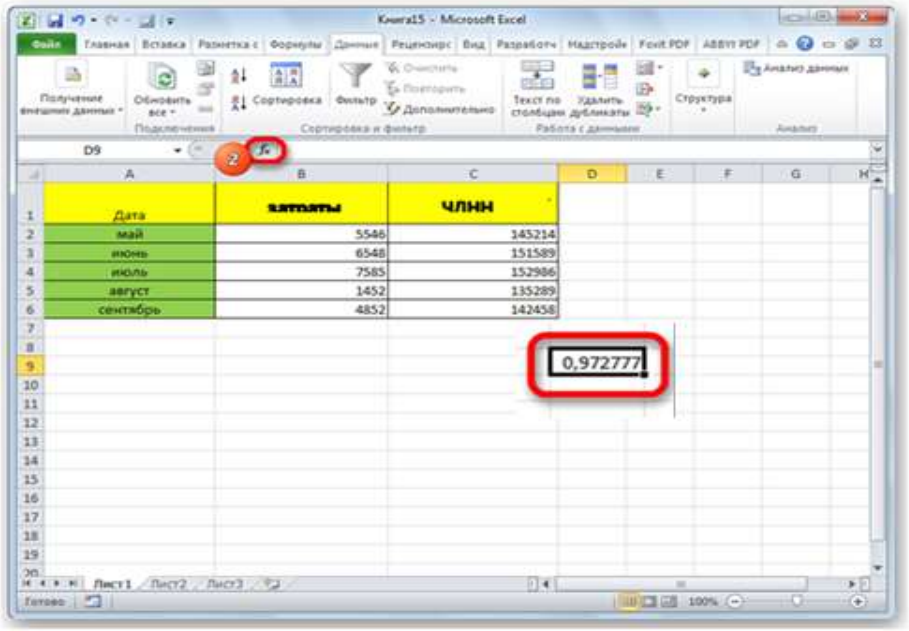


Рис. 14. Вычисление коэффициента корреляции в Excel

Способ 2: вычисление корреляции с помощью пакета анализа

Кроме того, корреляцию можно вычислить с помощью одного из инструментов, который представлен в пакете анализа.

1. Переходим во вкладку **«Данные»** – **«Анализ»**.

2. Открывается список с различными вариантами анализа данных. Выбираем пункт **«Корреляция»**. Кликаем по кнопке **«ОК»**.

3. Открывается окно с параметрами корреляционного анализа. В отличие от предыдущего способа, в поле **«Входной интервал»** мы вводим интервал не каждого столбца отдельно, а всех столбцов, которые участвуют в анализе. В нашем случае это данные в столбцах «Затраты» и «ЧЛНН».

Параметр **«Группирование»** оставляем без изменений – **«По столбцам»**, так как у нас группы данных разбиты именно на два столбца. Если бы они были разбиты построчно, то тогда следовало бы переставить переключатель в позицию **«По строкам»**.

В параметрах вывода по умолчанию установлен пункт **«Новый рабочий лист»**, то есть, данные будут выводиться на другом листе. Можно изменить место, переставив переключатель. Это может быть текущий лист (тогда вы должны будете указать координаты ячеек вывода информации) или новая рабочая книга (файл).

Когда все настройки установлены, жмем на кнопку **«ОК»**.

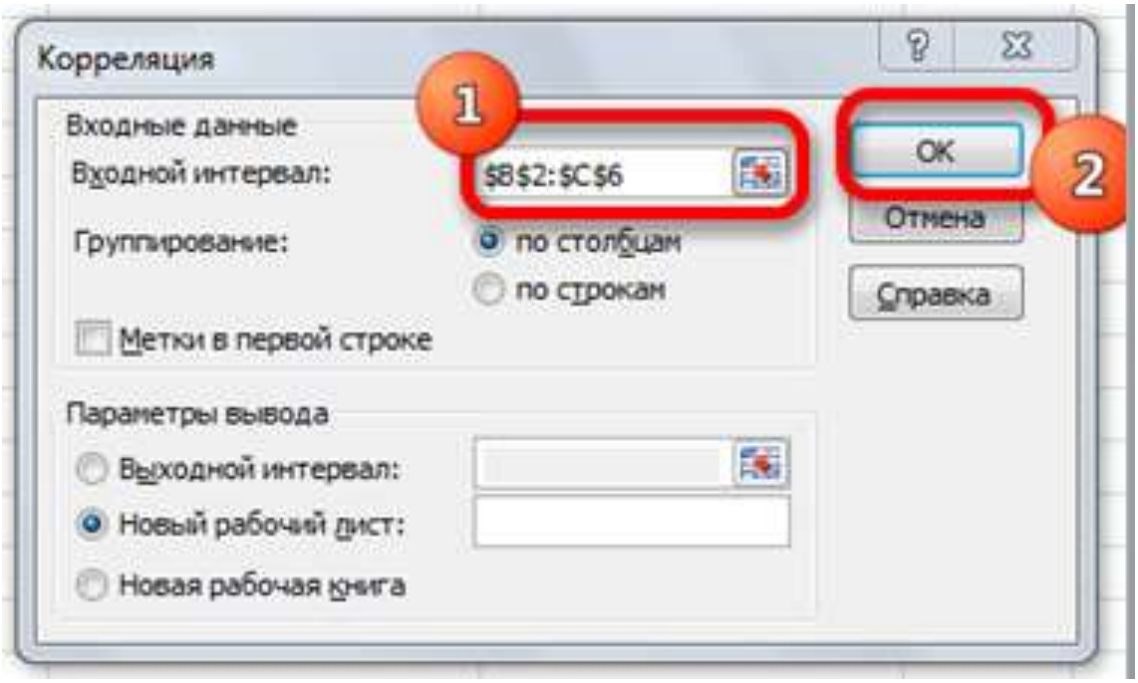


Рис. 15. Настройки функции «Корреляция» в Excel

Так как место вывода результатов анализа было оставлено по умолчанию, мы перемещаемся на новый лист. Как видим, тут указан коэффициент корреляции. Естественно, он тот же, что и при использовании первого способа – 0,97. Это объясняется тем, что оба варианта выполняют одни и те же вычисления, просто произвести их можно разными способами.

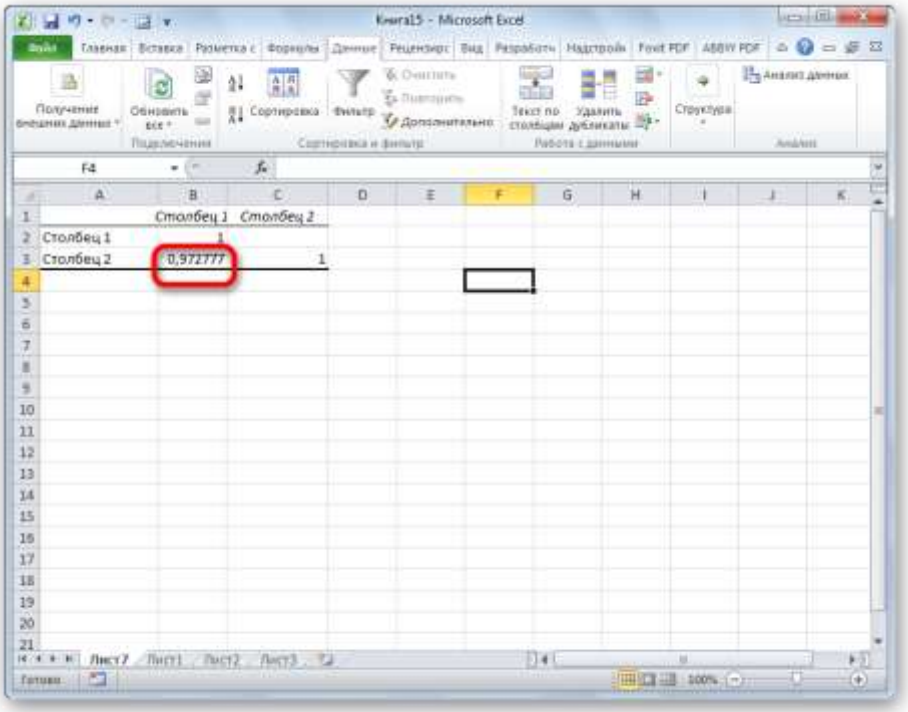


Рис. 16. Вычисление коэффициента корреляции в Excel

Как видим, приложение Excel предлагает сразу два способа корреляционного анализа. Результат вычислений, если вы все сделаете правильно, будет полностью идентичным. Но, каждый пользователь может выбрать более удобный для него вариант осуществления расчета.

РЕГРЕССИОННЫЙ АНАЛИЗ

Регрессионный анализ является одним из самых востребованных методов статистического исследования. С его помощью можно установить степень влияния независимых величин на зависимую переменную. В функционале Microsoft Excel имеются инструменты, предназначенные для проведения подобного вида анализа.

Подключение пакета анализа

Для того, чтобы использовать функцию, позволяющую провести регрессионный анализ, прежде всего, нужно активировать Пакет анализа.

Только тогда необходимые для этой процедуры инструменты появятся на ленте Excel. Перемещаемся во вкладку «Файл».

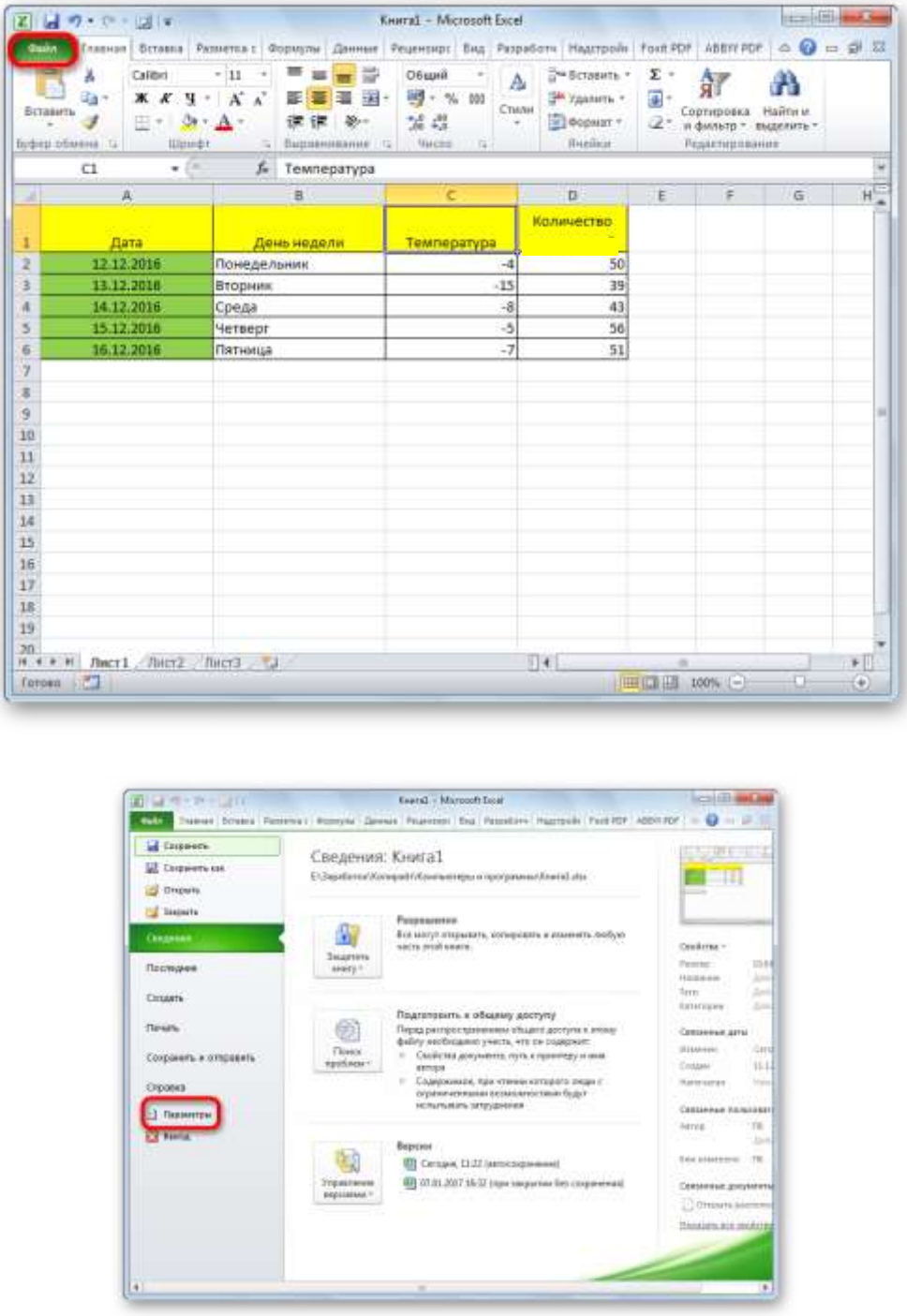


Рис. 17а. Настройка регрессионного анализа в Excel

Переходим в раздел «Параметры».

Открывается окно параметров Excel. Переходим в подраздел «Надстройки».

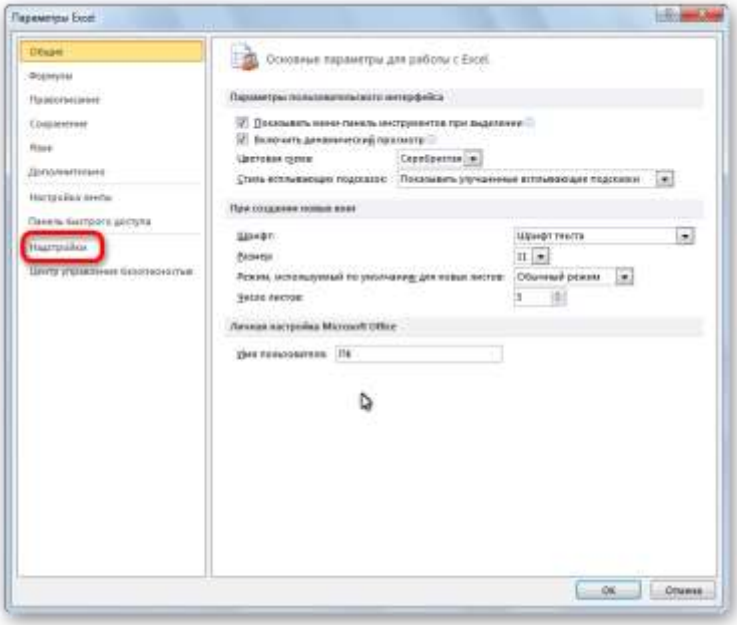


Рис. 17 б. Настройка регрессионного анализа в Excel

В самой нижней части открывшегося окна переставляем переключатель в блоке «Управление» в позицию «Надстройки Excel», если он находится в другом положении. Жмем на кнопку «Перейти».

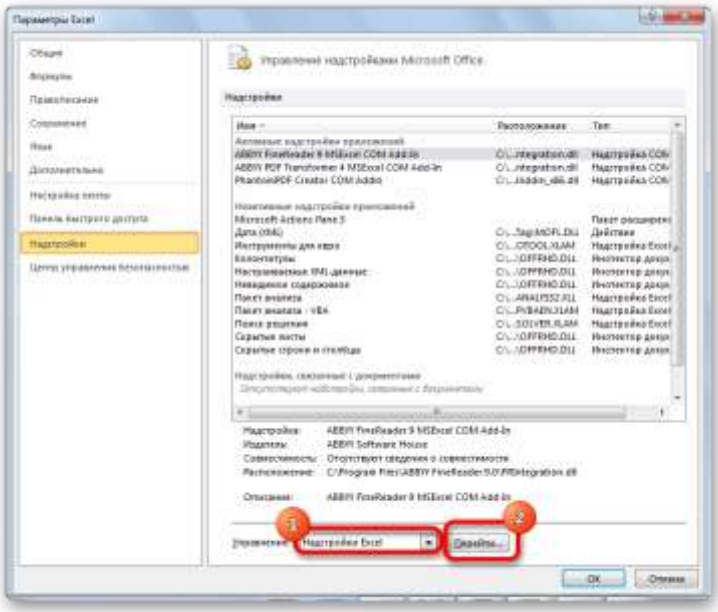


Рис. 17 в. Настройка регрессионного анализа в Excel

Открывается окно доступных надстроек Excel. Ставим галочку около пункта «Пакет анализа». Жмем на кнопку «ОК».

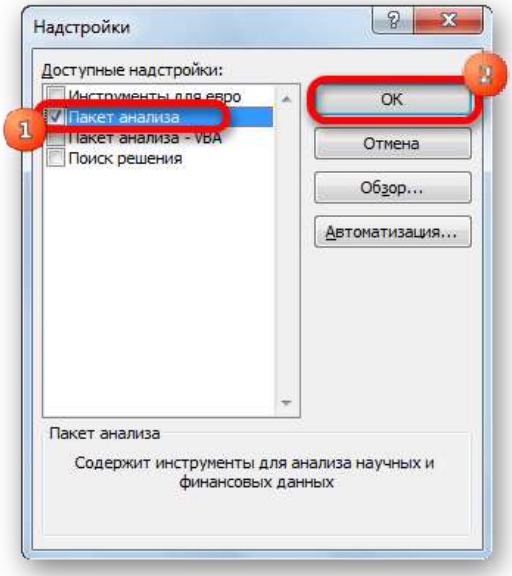


Рис. 17 г. Настройка регрессионного анализа в Excel

Теперь, когда мы перейдем во вкладку «Данные», на ленте в блоке инструментов «Анализ» мы увидим новую кнопку – «Анализ данных».

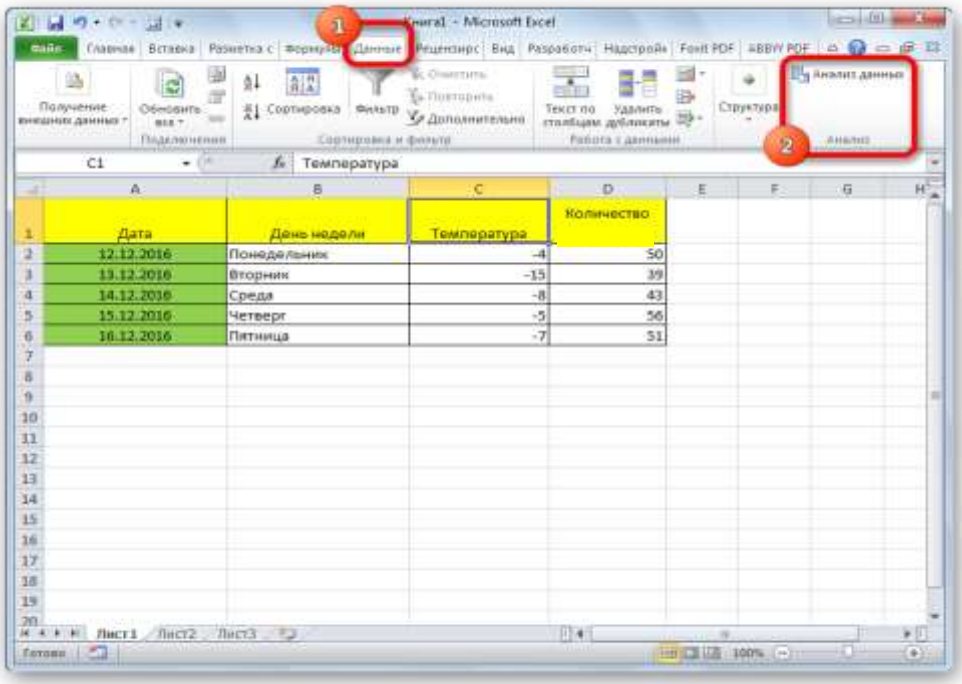


Рис. 17 д. Настройка регрессионного анализа в Excel

Виды регрессионного анализа

Существует несколько видов регрессий:

- параболическая;
- степенная;
- логарифмическая;
- экспоненциальная;
- показательная;
- гиперболическая;
- линейная регрессия.

О выполнении линейного регрессионного анализа в Excel подробнее поговорим далее.

Линейная регрессия в программе Excel

Внизу, в качестве примера, представлена таблица, в которой указана среднесуточная температура воздуха на улице, и количество обращений за медицинской помощью. Давайте выясним при помощи регрессионного анализа, как именно погодные условия в виде температуры воздуха могут повлиять на обращаемость в медицинские учреждения.

Общее уравнение регрессии линейного вида выглядит следующим образом: У = a0 + a1x1 +…+акхк. В этой формуле Y означает переменную, влияние факторов на которую мы пытаемся изучить. В нашем случае, это количество обращений за медицинской помощью. Значение x – это различные факторы, влияющие на переменную. Параметры a являются коэффициентами регрессии. То есть, именно они определяют значимость того или иного фактора. Индекс k обозначает общее количество этих самых факторов.

Кликаем по кнопке «Анализ данных». Она размещена во вкладке «Главная» в блоке инструментов «Анализ».

Открывается окно настроек регрессии. В нём обязательными для заполнения полями являются «Входной интервал Y» и «Входной интервал X». Все остальные настройки можно оставить по умолчанию.

В поле «Входной интервал Y» указываем адрес диапазона ячеек, где расположены переменные данные, влияние факторов на которые мы пытаемся установить. В нашем случае это будут ячейки столбца «Количество обращений за медицинской помощью». Адрес можно вписать вручную с клавиатуры, а можно, просто выделить требуемый столбец. Последний вариант намного проще и удобнее.

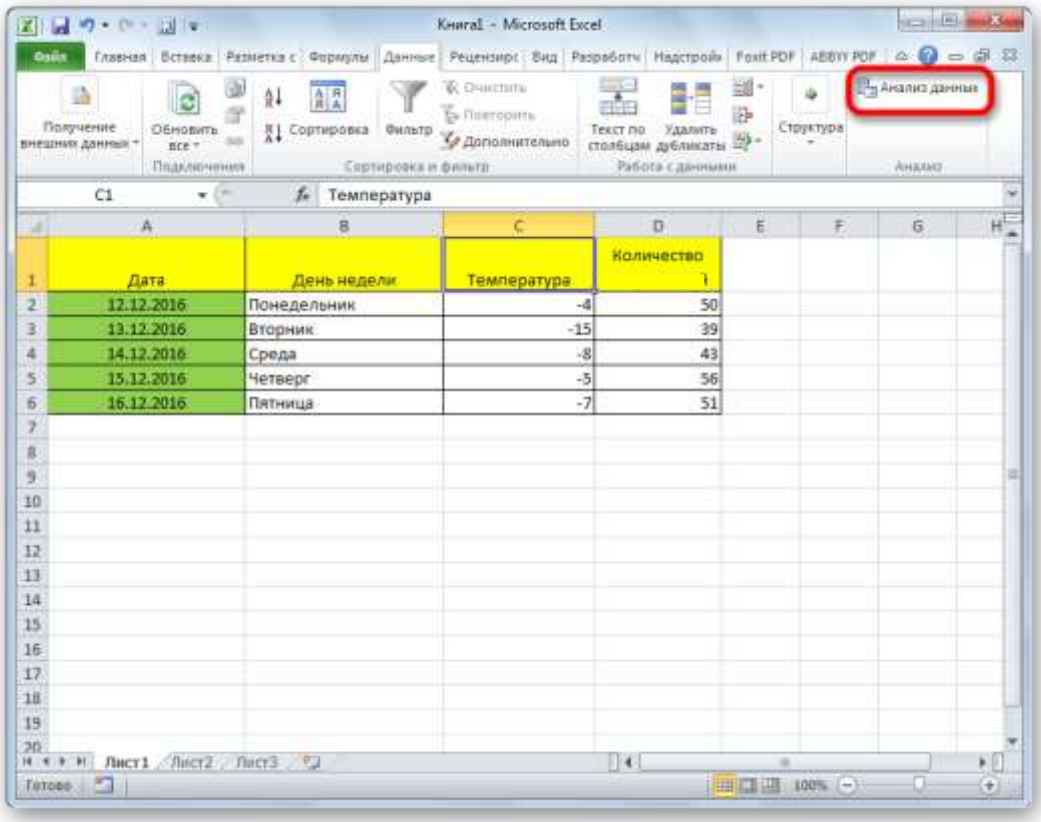


Рис. 18 а. Настройка регрессионного анализа в Excel. Анализ данных

Открывается небольшое окошко. В нём выбираем пункт «Регрессия». Жмем на кнопку «ОК».

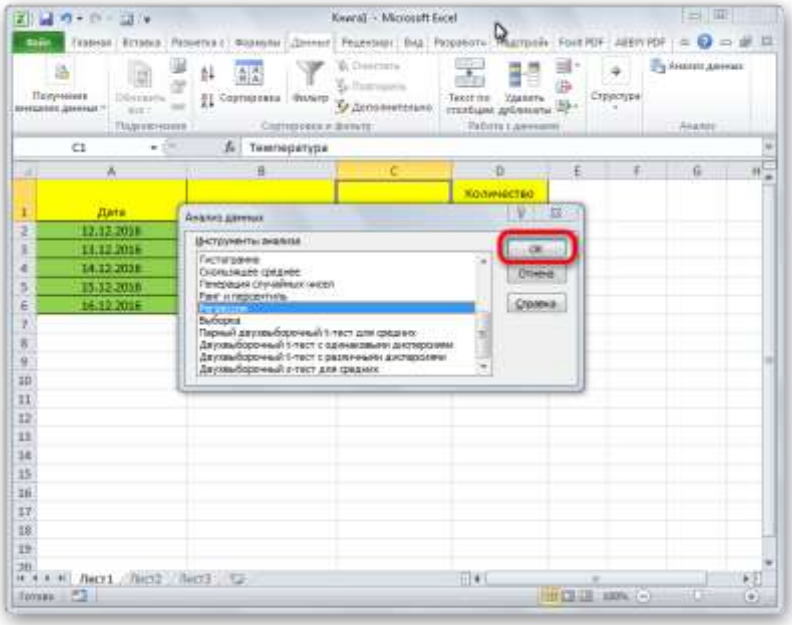


Рис. 18 б. Настройка регрессионного анализа в Excel. Анализ данных

В поле «Входной интервал X» вводим адрес диапазона ячеек, где находятся данные того фактора, влияние которого на переменную мы хотим установить. Как говорилось выше, нам нужно установить влияние температуры на количество обращений за медицинской помощью, а поэтому вводим адрес ячеек в столбце «Температура». Это можно сделать теми же способами, что и в поле «Количество».

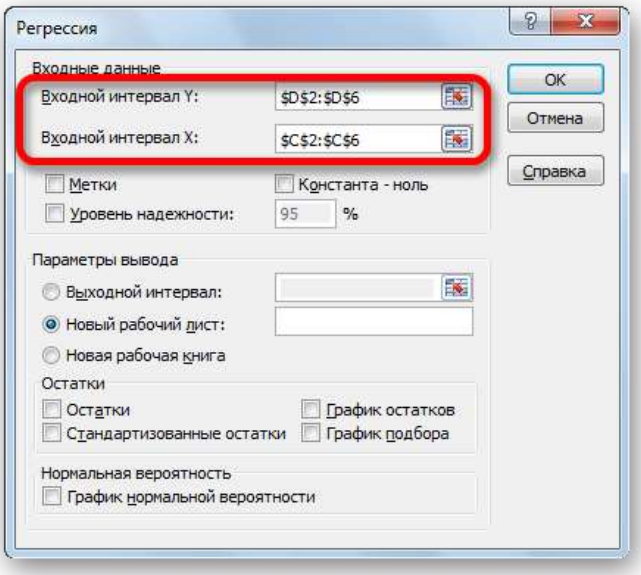


Рис. 18 в. Настройка регрессионного анализа в Excel. Анализ данных

С помощью других настроек можно установить метки, уровень надёжности, константу-ноль, отобразить график нормальной вероятности, и выполнить другие действия. Но, в большинстве случаев, эти настройки изменять не нужно. Единственное на что следует обратить внимание, так это на параметры вывода. По умолчанию вывод результатов анализа осуществляется на другом листе, но переставив переключатель, вы можете установить вывод в указанном диапазоне на том же листе, где расположена таблица с исходными данными, или в отдельной книге, то есть в новом файле.

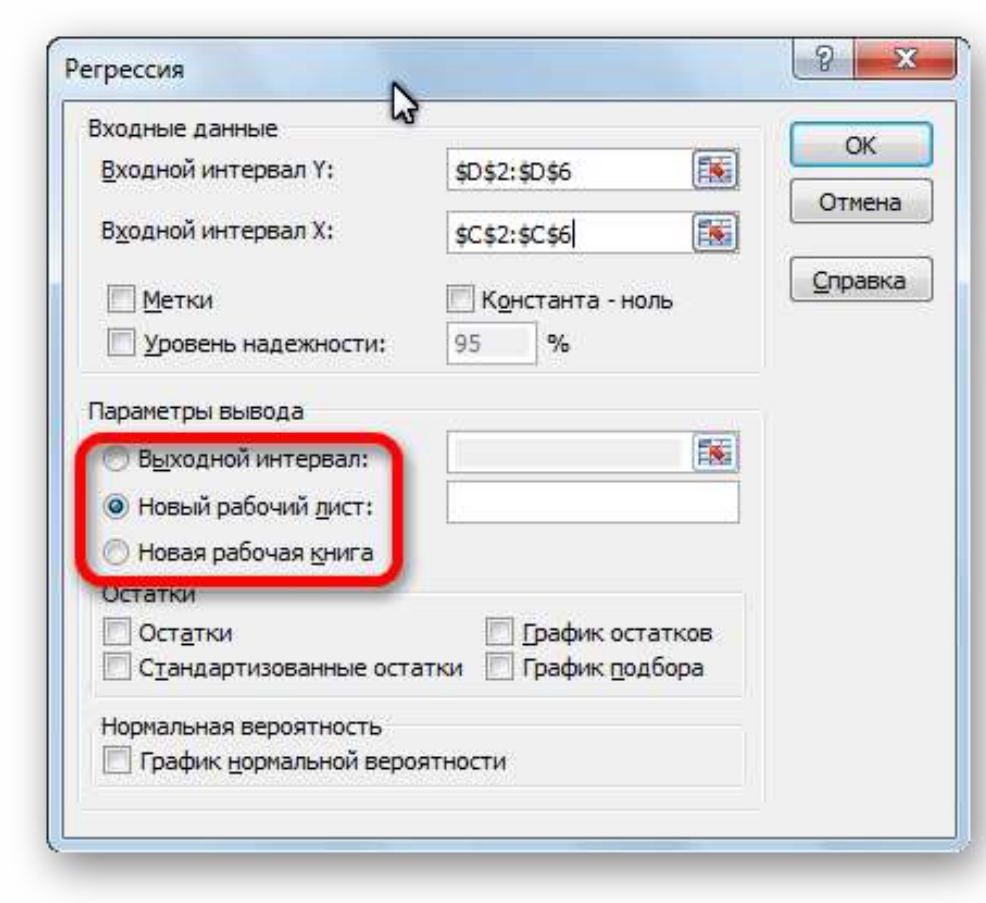


Рис. 18 г. Настройка регрессионного анализа в Excel. Анализ данных

После того, как все настройки установлены, жмем на кнопку «ОК».

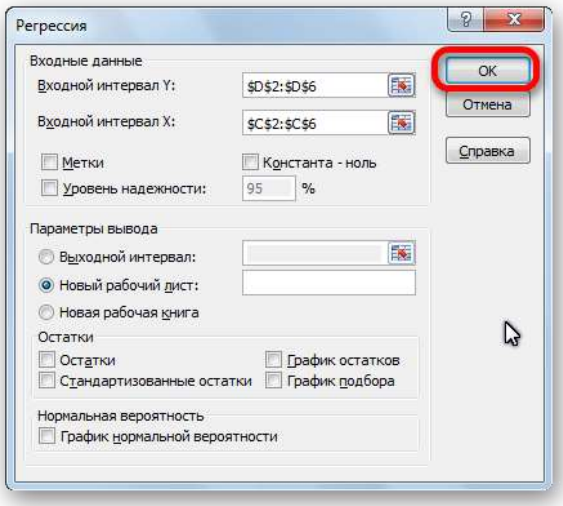


Рис. 18 д. Настройка регрессионного анализа в Excel. Анализ данных

Разбор результатов анализа

Результаты регрессионного анализа выводятся в виде таблицы в том месте, которое указано в настройках.

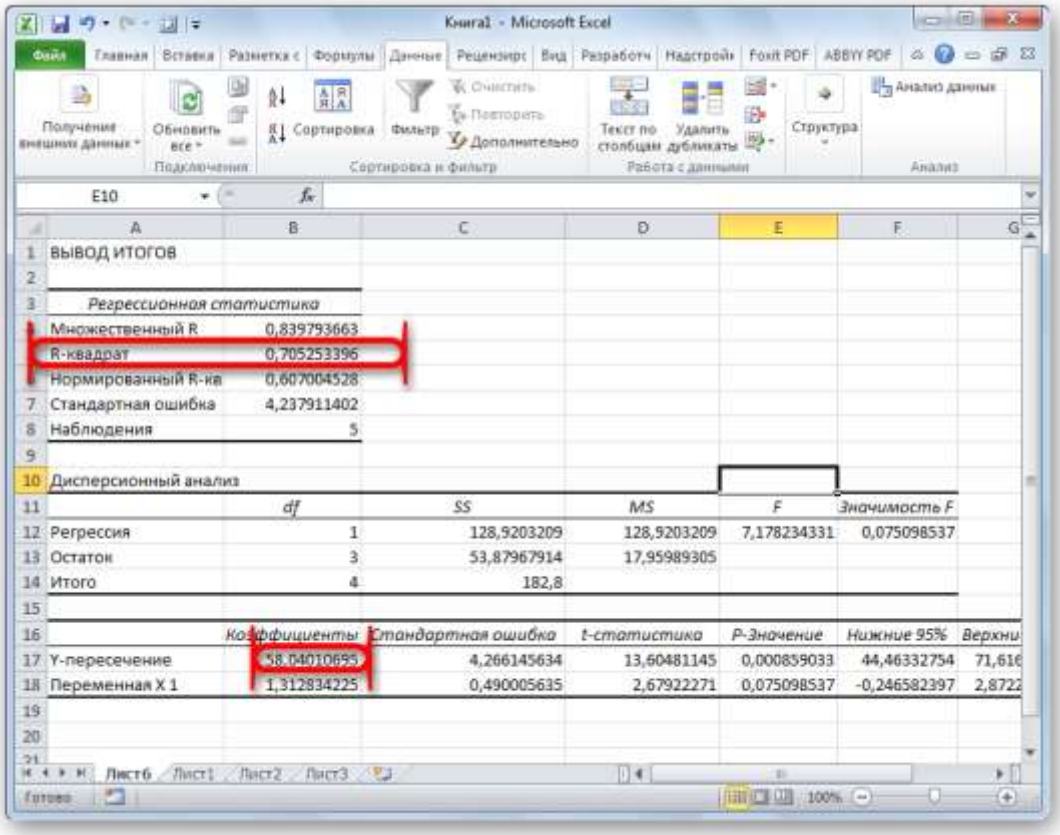


Рис. 19. Результаты регрессионного анализа в Excel.

Одним из основных показателей является R-квадрат. В нем указывается качество модели. В нашем случае данный коэффициент равен 0,705 или около 70,5%. Это приемлемый уровень качества. Зависимость менее 0,5 является плохой.

Ещё один важный показатель расположен в ячейке на пересечении строки «Y-пересечение» и столбца «Коэффициенты». Тут указывается какое значение будет у Y, а в нашем случае, это количество обращений за медицинской помощью, при всех остальных факторах равных нулю. В этой таблице данное значение равно 58,04.

Значение на пересечении граф «Переменная X1» и «Коэффициенты» показывает уровень зависимости Y от X. В нашем случае — это уровень зависимости количества обращений за медицинской помощью от температуры. Коэффициент 1,31 считается довольно высоким показателем влияния.

Итак, с помощью программы Microsoft Excel довольно просто составить таблицу регрессионного анализа. Но, для того чтобы работать с полученными на выходе данными, и понимать их суть, необходимо иметь хорошую теоретическую подготовку по теме «Корреляционно-регрессионный анализ».

Вопросы для подготовки к занятию

1. Виды связи между явлениями.
2. Область применения и методика вычисления коэффициентов корреляции. Коэффициентов корреляции.
3. Коэффициент детерминации (R^2). Методика его расчета и сфера применения.
4. Регрессия. Методы регрессионного анализа. Область применения уравнения регрессии.

5. Регрессионный анализ: понятие, задачи, основные цели

Тесты

1. Термин «корреляция» в статистике понимают как:

а) связь, зависимость

б) отношение, соотношение

в) функцию, уравнение

г) коэффициент

2. Связь между признаками можно считать средней при значении коэффициента корреляции:

а) r=0,13

б) r=0,45

в) r=0,71

г) r=1,0

3. Коэффициент корреляции r = - 0,82 говорит о том, что корреляционная связь:

а) прямая, средней силы

б) обратная, слабая

в) прямая, сильная

г) обратная, сильная

4. При значении коэффициента корреляции в диапазоне от 0 до 0,3 сила связи оценивается, как:

а) слабая

б) средняя

в) сильная

г) полная

5. Связь между признаками можно считать сильной при значении коэффициента корреляции:

а) r= - 0,25

б) r= 0,62

в) r= - 0,95

г) r= 0,55

6. Зависимость, при которой увеличение или уменьшение значения одного признака ведет к увеличению или уменьшению – второго, характеризует следующий вид связи:

а) прямая

б) обратная

в) полная

г) неполная

7. Зависимость, при которой увеличение одного признака дает уменьшение второго характеризует следующий вид корреляционной связи:

а) прямая

б) обратная

в) полная

г) неполная

8. Коэффициент корреляции Пирсона определяет:

а) статистическую значимость различий между переменными

б) степень разнообразия признака в совокупности

в) силу и направление связи между зависимой и независимой переменными

г) долю дисперсии результативного признака объясняемую влиянием независимых переменных

9. Условием для расчета коэффициента корреляции Пирсона является:

а) распределение переменных неизвестно

б) нормальное распределение по крайней мере, одной из двух переменных

в) по крайней мере, одна из двух переменных измеряется в ранговой шкале

г) отсутствует нормальное распределение переменных

10. Ранговый коэффициент корреляции спирмэна рассчитывается, когда:

а) присутствует нормальное распределение переменных

б) необходимо оценить связь между качественными и количественными признаками

в) необходимо определить статистическую значимость различий между переменными

г) необходимо оценить степень разнообразия признака в совокупности

11. Зависимость, когда каждому значению одного признака соответствует точное значение другого, называется:

а) прямой

б) обратной

в) корреляционной

г) функциональной

12. Зависимость, когда при изменении величины одного признака изменяется тенденция (характер) распределения значений другого признака, называется:

а) прямой

б) обратной

в) корреляционной

г) функциональной

13. Для изображения корреляционной зависимости используется график:

а) линейный

б) график рассеяния точек

в) радиальный

г) динамический

14. Если коэффициент корреляции равен 1, то связь является:

а) сильной, прямой

б) сильной обратной

в) средней, прямой

г) полной (функциональной), прямой

15. Связь между Y и X можно признать более существенной при следующем значении линейного коэффициента корреляции:

а) r= 0,35

б) r= 0,15

в) r= -0,57

г) r= 0,46

16. Корреляционный анализ используется для изучения:

а) взаимосвязи явлений

б) развития явления во времени

в) структуры явлений

г) статистической значимости различий между явлениями

17. Коэффициент корреляции может принимать значения:

а) от 0 до 1

б) от -1 до 0

в) от -1 до 1

г) любые положительные

18. Коэффициент детерминации может принимать значения:

а) от 0 до 1

б) от -1 до 0

в) от -1 до 1

г) любые положительные

19. В результате проведения регрессионного анализа получают уравнение, описывающее ... показателей:

а) взаимосвязь

б) соотношение

в) структуру

г) темпы роста

20. Линейная связь между факторами исследуется с помощью уравнения регрессии:

а) y=a+bx

б) y=a+b/x

в) y=a+b1x1+b2x2

г) y=ax

21. Параметр b (b = 0,016) линейного уравнения регрессии показывает, что:

а) с увеличением признака "x" на 1 признак "y" увеличивается на 0,678

б) с увеличением признака "х" на 1 признак "у" увеличивается на 0,016

в) с увеличением признака "х" на 1 признак "у" уменьшается на 0,678

г) с увеличением признака "х" на 1 признак "у" уменьшается на 0,016

22. Независимая переменная в уравнении регрессии называется:

а) вариантой

б) уровнем

в) предиктором

г) переменной отклика

23. Зависимая переменная в уравнении регрессии называется:

а) вариантой

б) уровнем

в) предиктором

г) переменной отклика

24. Для прогнозирования ожидаемого систолического давления ребенка используется:

а) квадратное уравнение

б) отношение правдоподобия

в) коэффициент вариации

г) уравнение регрессии

25. Для оценки корреляционной связи между качественными признаками применяется коэффициент корреляции:

а) Пирсона

б) Спирмена

в) Кендела

г) Чупрова

26. О сильной обратной связи можно говорить при коэффициенте корреляции равном:

а) 0,9

б) -0,59

в) -0,9

г) 0,2

27. Для изучения связи, в которой присутствует более одной независимой переменной используется:

а) линейная регрессия

б) множественная регрессия

в) ранговая корреляция Спирмэна

г) расчет темпа прироста

28. Для расчета коэффициента корреляции Спирмэна необходимо:

а) рассчитать сумму данных

б) расположить переменные в порядке чередования

в) возвести переменные в квадрат

г) присвоить переменным в порядке возрастания последовательные ранги (номера 1, 2, 3, .., n)

29. Зависимость веса от роста человека (росто-весовой индекс) описывается при помощи:

а) логистической регрессии

б) множественной регрессии

в) экспоненциальной регрессии

г) линейной регрессии

30. Зависимость положительного или отрицательного результата лечения от ряда факторов описывается при помощи:

а) логистической регрессии

б) множественной регрессии

в) экспоненциальной регрессии

г) линейной регрессии

31. Коэффициент корреляции измеряется в:

а) процентах

б) тех же единицах, что и изучаемый признак

в) промилле

г) не имеет единиц измерения

32. Из нижеперечисленных величин для определения размера одного признака при изменении другого на единицу измерения применяется:

а) среднеквадратическое отклонение

б) коэффициент корреляции

в) коэффициент регрессии

г) коэффициент вариации

Ситуационные задачи

I. Проведите корреляционно-регрессионный анализ заболеваемости. В качестве экспериментальных данных представлена еженедельная заболеваемость ОРВИ на территории населённого пункта К. в зимний период (с декабря по февраль).

Неделя	Заболевания	Температура воздуха, - t °С
1	25	19
2	41	21
3	22	22
4	34	24

5	35	20
6	35	21
7	37	24
8	31	21
9	28	18
10	26	20
11	38	30
12	44	32

Требуется подобрать наиболее качественную модель прогнозирования заболеваемости с учетом её адекватности и статистической значимости параметров модели, надежности и точности.

Найдите параметры уравнения линейной регрессии, дайте интерпретацию коэффициента регрессии.

Осуществите проверку значимости параметров уравнения регрессии с помощью *t*-критерия Стьюдента (p=0,05).

Вычислите коэффициент детерминации, проверьте значимость уравнения регрессии с помощью *F*-критерия Фишера (p=0,05).

Сделайте вывод о качестве модели.

Найдите коэффициент эластичности и среднюю относительную ошибку аппроксимации линейной регрессии.

Составьте уравнения нелинейной регрессии (гиперболическая; степенная; показательная).

Найдите коэффициенты детерминации, коэффициенты эластичности и средние относительные ошибки аппроксимации.

Сравните модели по всем характеристикам и сделайте вывод.

Тема 4. Таблицы сопряженности

Цель занятия: овладеть навыками формирования таблиц сопряженности и установления наличия или отсутствия связи между изучаемыми признаками.

Учебно-целевые задачи:

- научиться формировать таблицы сопряженности
- научиться рассчитывать показатели на основе таблиц сопряженности, анализировать и интерпретировать их
- научиться оценивать статистическую значимость выявленной связи

В результате освоения темы обучающиеся **должны знать:** методы выявления связи между категориальными переменными; оценки статистической значимости связи

В результате освоения темы обучающиеся **должны уметь:** определять вид переменных (категориальные, зависимые и независимые переменные); рассчитывать показатели: отношение шансов, относительный риск и доверительные интервалы для них

В результате освоения темы обучающиеся **должны владеть:** технологией изучения связи между категориальными переменными; методами оценки статистической значимости выявленной связи

ИНФОРМАЦИОННЫЙ МАТЕРИАЛ

В случае изучения взаимосвязи между двумя качественными переменными лучше изучать путем сравнения градаций одной переменной по распределению другой переменной. Например, предположим, что изучается связь формирования невротических реакций и количества лет совмещения работы с учебой. Результаты изучения связей количественных переменных принято представлять в виде таблицы сопряженности:

Таблица 9. Характеристика частоты возникновений невротических реакций в зависимости от количества лет совмещения работы с учебой

Название реакций	Количество лет совмещения работы с учебой		
	До 3 лет	4 и более	всего
	Количество лиц с невротическими реакциями		
Неврастенические	120	81	201
Обсессивно-фобические	75	64	139
Истерические	52	41	93
ИТОГО:	247	186	433

Для анализа связи бинарных переменных строят четырехпольные таблицы сопряженности (табл. 10). Изучить взаимосвязь двух бинарных переменных можно методом Пирсона. При этом рассчитывается φ-коэффициент сопряженности (аналог коэффициента Пирсона). А так же можно использовать метод χ^2. Использование φ-коэффициента значительно ограничено. Чем больше асимметрия распределения «да» и «нет» по каждой переменной, тем менее точно φ-коэффициент отражает связь между переменными, т.е. количество «да» и «нет» по каждой переменной должно быть примерно одинаковым.

Ассоциация бинарных данных

Рассмотрим пример. Требуется определить зависимость между патогенностью микроорганизмов и их устойчивостью к сульфаниламидным препаратам (см. табл. 10).

Таблица 10. Характеристика устойчивости патогенных микроорганизмов к сульфаниламидным препаратам

Штаммы микроорганизмов	Число штаммов		Всего
	устойчивых к сульфаниламидам	не устойчивых к сульфаниламидам	
Патогенные	a 60	b 39	a+b 99

Не патогенные	c 64	d 106	c+d 170
ИТОГО:	(a+c) 124	(b+d)145	N 269

Рассчитаем **φ-коэффициент** сопряженности (*a, b, c, d* – числовые значения частот сопоставляемых совокупностей.

$$\varphi = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} = \frac{60 \times 106 - 39 \times 64}{\sqrt{(60+39)(60+64)(39+106)(64+106)}} =$$

$$= \frac{6360 - 2496}{\sqrt{99 \times 124 \times 145 \times 170}} = \frac{3864}{17395{,}5} = 0{,}22$$

Выявлена слабая ($0{,}1 < r_{xy} < 0{,}3$ – слабая) положительная взаимосвязь между патогенностью микроаргонизмов и устойчивостью к сульфаниламидам. Т.к. φ-коэффициент сопряженности есть коэффициент Пирсона вычисленный на бинарных данных интерпретация и оценка статистической значимости проводятся аналогичным образом.

Вычисление χ^2 проводим по формуле:

$$\chi^2 = \frac{(ad - bc)^2 \times (a+b+c+d)}{(a+b)(a+c)(b+d)(c+d)},$$

$$\chi^2 = \frac{(60 \times 106 - 39 \times 64)^2 \times (60+39+64+106)}{(60+39) \times (60+64) \times (39+106) \times (64+106)} = 13{,}3.$$

Для оценки статистической значимости полученного результата необходимо определить число степеней свободы по формуле: $K = (R - 1)$ x $(S - 1) = (2 - 1)$ x $(2 - 1) = 1$ (R и S – количество строк и граф без итоговых данных). При $K = 1$ критическое значение χ^2 соответствует 3,8 – 6,6 – 9,5. Следовательно, H_0 отвергается. ($p < 0{,}001$).

Вывод. Между патогенностью микроорганизмов и их устойчивостью к сульфаниламидным препаратам существует взаимная зависимость ($p < 0{,}001$).

Вопросы для подготовки к занятию

1. Для каких целей строятся таблицы сопряженности?
2. Понятие зависимые и независимые переменные (факторные и результативны)
3. Интерпретация показателя отношения шансов
4. Интерпретация доверительного интервала для показателя отношения шансов.
5. Интерпретация показателя относительного риска
6. Интерпретация доверительного интервала для показателя относительного риска
7. Оценка статистической значимости показателей относительного риска и отношения шансов

Тесты

1. Признак: «наличие или отсутствие болезни» является:

а) количественным

б) непрерывным

в) дискретным

г) дихотомическим

2. Зависимый признак, изменяющий свое значение под влиянием другого:

а) факторный

б) результативный

в) дискретный

г) непрерывный

3. Какая шкала отображает степень тяжести заболевания:

а) номинальная

б) интервальная

в) порядковая

г) логарифмическая

4. В медицинских исследованиях при установлении доверительных границ любого показателя принята вероятность безошибочного прогноза:

а) 80%

б) 68%

в) 95% и более

г) 50%

5. Для вероятности безошибочного прогноза 95,0% величина критерия t составляет:

а) 3

б) 2

в) 1

г) 10

6. Для вероятности безошибочного прогноза 99,0% величина критерия t составляет:

а) 3

б) 2

в) 1

г) 5

7.По роли в статистической совокупности учетные признаки можно подразделить на:

а) достоверные и невозможные

б) первичные и вторичные

в) качественные и вероятные

г) факторные и результативные

8. Шанс события выражается формулой:

а) вероятность события/ 1 + вероятность события

б) вероятность события/ 1 - вероятность события

в) вероятность события/ вероятность события + 1

г) вероятность события/ вероятность события - 1

15. Вероятность события выражается формулой:

а) шансы события / шансы события – 1

б) шансы события / 1 – шансы события

в) шансы события /1 + шансы события

г) шансы события / шансы события + 1

Ситуационные задачи

I. В исследовании была поставлена цель - оценить действие БЦЖ, направленное на предупреждение развития менингита туберкулезной этиологии. В исследование было включено 60 человек с диагнозом туберкулезного менингита. Такое же количество участников отобрано в контрольную группу с учетом возраста, пола и места проживания. наличие вакцинации против туберкулеза вакциной БЦЖ исследователи определяли с помощью опроса участников. В результате установлено, что 25 участников из основной группы и 50% из контрольной сообщили о прививке вакциной БЦЖ.

Задание

1. Определите дизайн представленного исследования.

2. Укажите фактор риска и исход в данном исследовании.

3. Заполните четырехпольную таблицу и рассчитайте необходимые показатели.

4. Обозначьте возможные систематические ошибки в исследовании.

II. В 1981 году Б. МакМахон и коллеги сообщили о проведенном исследовании случай-контроль причин развития рака поджелудочной железы. Случаями была группа пациентов с гистологически подтвержденным панкреатического рака в 11 бостонских и рок-айлендских больницах с 1974 по 1979 годы. Контрольные группы отбирались из пациентов, которые были госпитализированы в тоже самое время, что и случаи с другими диагнозами. пациенты для контрольной группы были отобраны из числа госпитализированных теми же врачами, которые направляли на госпитализацию больных, ставших случаями. Одной из находок в этом исследовании была очевидная дозозависимая корреляция между потреблением кофе и раком поджелудочной железы, особенно среди женщин (таб. 1.)

Задание

1. Рассчитайте показатели отношения шансов, отражающие выявленную зависимость.

2. При наличии вычислительных средств сделайте расчеты доверительных интервалов отношений шансов.

3. Было ли заключение по результатам этого исследования верным?

4. Предположите, какие систематические ошибки могли быть в данном исследовании.

5. Как можно иначе организовать данное исследование?

Распределение пациентов из групп случаев и контролей в зависимости от употребления кофе.

пол		Потребление кофе (чашек/день)				Всего
		0	1-2	3-4	≥5	
м	Число случаев	9	94	53	60	216
	Число контролей	32	119	74	82	307
ж	Число случаев	11	59	53	28	151
	Число контролей	56	152	80	48	336

III. В таблице представлено число новых случаев артериальной гипертонии женщин 20-30 лет города А в зависимости от гиперхолистеринемии:

Группы	**Новые случаи АГ**		**Всего**
	Есть	Нет	
Основная группа F+	64	79	143
Контрольная группа F-	219	815	1034
Всего	283	894	1177

Задание: рассчитайте показатели, заполните таблицу

	Показатели	95%ДИ
Инцидентность в группе F+		
Инцидентность в группе F-		
Атрибутивный риск		
Относительный риск		
Этиологическая доля		
Относительный риск		

IV. В таблице представлено число случаев сердечно-сосудистой патологии, выявленной у мужчин 40-50 лет города С в зависимости от курительного статуса:

Группы	**сердечно-сосудистая патология**		**Всего**
	Нет	Да	
Основная группа F+ (Курение +)	51	790	841
Контрольная группа F- (Курение -)	152	80	232

Всего	203	870	1073

Задание: рассчитайте показатели, заполните таблицу

	Показатели	95%ДИ
Инцидентность в группе F+		
Инцидентность в группе F-		
Атрибутивный риск		
Относительный риск		
Этиологическая доля		
Отношение шансов		

Тема 5. Дисперсионный анализ.

Цель занятия: раскрыть содержание основных понятий и процедуры проведения дисперсионного анализа

Учебно-целевые задачи:

- знакомство с основными понятиями и критериями дисперсионного анализа
- анализ особенностей применения и процедуры дисперсионного анализа
- освоение пошаговой процедуры проведения дисперсионного анализа

В результате освоения темы обучающиеся **должны знать:** цели, область и особенности применения, методики дисперсионного анализа

В результате освоения темы обучающиеся **должны уметь:** выбирать и применять разные методики дисперсионного анализа; интерпретировать полученные данные

В результате освоения темы обучающиеся **должны владеть**: технологией проведения дисперсионного анализа

ИНФОРМАЦИОННЫЙ МАТЕРИАЛ

Для сравнения двух, трех и более групп по количественному признаку применяется дисперсионный анализ для связанных (когда исследуется влияние разных градаций фактора или разных условий на одну и ту же

выборку испытуемых, а также, если группы сравниваются в динамике) и несвязанных совокупностей. Выборки по численности могут быть как равными, так и неравными. Если выборки не равны по численности, обязательным условием для проведения дисперсионного анализа является равенство дисперсий. Если различие дисперсий выборок статистически значимо, следует воспользоваться непараметрическим аналогом дисперсионного анализа (метод Краскела-Уоллиса (Н). Численность выборок не должна быть менее 2 объектов (лучше не менее 5). Дисперсионный анализ относится к группе параметрических методов и поэтому его следует применять только тогда, когда доказано, что распределение является нормальным (Шеффе Г., 1980).

Метод дисперсионного анализа позволяет сравнивать выборки, изучить влияние независимой переменной (одной или нескольких) на зависимую. Независимая переменная это признак, имеющий две или более градации. Каждой градации независимой переменной соответствует выборка объектов наблюдения, для которых установлены значения зависимой переменной. Независима переменная (или фактор) также имеет несколько градаций.

По количеству изучаемых факторов дисперсионный анализ может быть однофакторным (при этом изучается влияние одного фактора на результативный признак), двухфакторным (при изучении влияния двух факторов) и многофакторным (позволяет оценить не только влияние каждого из факторов в отдельности, но и их взаимодействие).

Математическая идея дисперсионного анализа основана на соотнесении межгрупповой и внутригрупповой дисперсий зависимой переменной. При объединении выборок с одинаковой дисперсией, но разными средними величинами дисперсия увеличивается пропорционально различиям средних величин этих выборок. Это объясняется тем, что к внутригрупповой дисперсии добавляется дисперсия, обусловленная различиями между группами. В дисперсионном анализе внутригрупповая

дисперсия рассматривается как обусловленная случайными причинами, а межгрупповая – действием изучаемого фактора на зависимую переменную. В общей дисперсии зависимой переменной выделяют внутригрупповую (случайную) и межгрупповую (факторную) дисперсию. Чем больше отношение межгрупповой дисперсии к внутригрупповой, тем выше факторный эффект, соответственно, тем больше различаются средние величины разных градаций фактора.

Проведение дисперсионного анализа вручную довольно трудоемкая задача. Однако, для того чтобы понять суть этого анализа, уметь правильно интерпретировать его результаты необходимо иметь общее представление о вычислительных операциях. Рассмотрим упрощенный способ проведения однофакторного дисперсионного анализа вручную.

ПАРАМЕТРИЧЕСКИЙ МЕТОД СРАВНЕНИЯ ТРЕХ НЕЗАВИСИМЫХ ГРУПП И БОЛЕЕ ПО ОДНОМУ КОЛИЧЕСТВЕННОМУ ПРИЗНАКУ ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ ДЛЯ НЕЗАВИСИМЫХ ГРУПП

Разберем основные этапы однофакторного дисперсионного анализ для независимых групп на примере: на территории области изучена онкологическая заболеваемость. Необходимо установить, влияет ли урбанизация на уровень онкологической заболеваемости.

1. Построим дисперсионный комплекс. Количество градаций факторного признака 3, наблюдений 5, результативный признак онкологическая заболеваемость (первичная) на 100 тыс. населения (табл. 11)

Таблица 11. Онкологическая заболеваемость (первичная) на 100 тыс. населения.

Наблюдения	Города с населением до 50 тыс. чел.	Города с населением от 50 тыс. до 150 тыс. чел.	Города с населением от 150 тыс. чел. и более

(n)	(v)	(v)	(v)
1	256,8	248,6	332,0
2	292,0	289,8	348,6
3	270,1	322,8	352,5
4	299,3	322,1	385,1
5	291,5	313,9	381,6
M (общее среднее)		313,78	
Mj (групповое среднее)	281,94	299,44	360,0

Алгоритм вычислений:

Рассчитываем показатель общей изменчивости зависимой переменной (SS_{total}):

$$SS_{total} = \sum_{i=1}^{N} (v - M)^2$$

SS_{tota} =(256,8-313,8)^2+(292,0-313,8)^2+(270,1-313,8)^2+(299,3-313,8)^2+(291,5-313,8)^2+…+(381,6-313,8)^2=24038,904

Рассчитываем показатель изменчивости между 3 группами каждая численностью 5 объектов:

$$SS_{bg} \sum_{j-1}^{k} n \times (Mj - M)^2$$

SS_{bg}=5x[(281,94-313,78)^2+(299,44-313,78)^2+(360,0-313,78)^2]=16760,068

NB!

При проведении дисперсионного анализа с выборками разной численности формула для вычисления межгрупповой суммы квадратов приобретает вид:

$$SS_{bg} \sum_{j-1}^{k} n \times (Mj - M)^2$$

n_j – численность по каждой группе

Кроме того, необходима проверка равенства дисперсий.

Рассчитываем показатель случайной изменчивости (внутри групп)

$$SS_{wg} = SS_{total} - SS_{bg} = 24038{,}904 - 16760{,}068 = 7278{,}836$$

Определим число степеней свободы:
df_{bg}=k-1=3-1=2;
df_{wg}= N-k=15-3=12

Рассчитаем средние квадраты:

$$MS_{bg} = \frac{ss_{bg}}{df_{bg}} = \frac{16760{,}07}{2} = 8380{,}03$$

$$MS_{wg} = \frac{ss_{wg}}{df_{wg}} = \frac{7278{,}84}{12} = 606{,}6$$

Рассчитаем F-отношение:

$$F = \frac{MS_{bg}}{MS_{wg}} = \frac{8380{,}03}{606{,}6} = 13{,}8$$

По таблице критических значений для критерия F при p=0,05; $df_{числ.}$=2; $df_{знам.}$=12 критическое значение F=3,9; для p=0,01; $df_{числ.}$=2; $df_{знам.}$=12 критическое значение F=6,972. Следовательно, p=0,01.

Рассчитаем коэффициент детерминации:

$$R^2 = \frac{SS_{bg}}{SS_{total}} = \frac{16760{,}07}{24038{,}904} = 0{,}697 \approx 0{,}7$$

Вывод: Для всех территорий можно с вероятностью 99,7%; (Fфакт. > Fкрит.; 13,8>3,9) утверждать, что урбанизация влияет на уровень онкологической патологии. Доля влияния урбанизации на уровень онкологической патологии составляет 70% в общем числе других факторов.

Можно предположить, что высокий уровень онкологической заболеваемости в крупных городах связан с лучшей выявляемостью данной патологии.

Возьмем данные уже рассмотренного примера (рис. 20) и проведем обработку на компьютере с помощью Microsoft Excel. Для этого запустите в меню «Сервис» выберите команду «Анализ данных». Выберите необходимую строку в списке “Инструменты анализа” «Однофакторный дисперсионный анализ» (рис. 20).

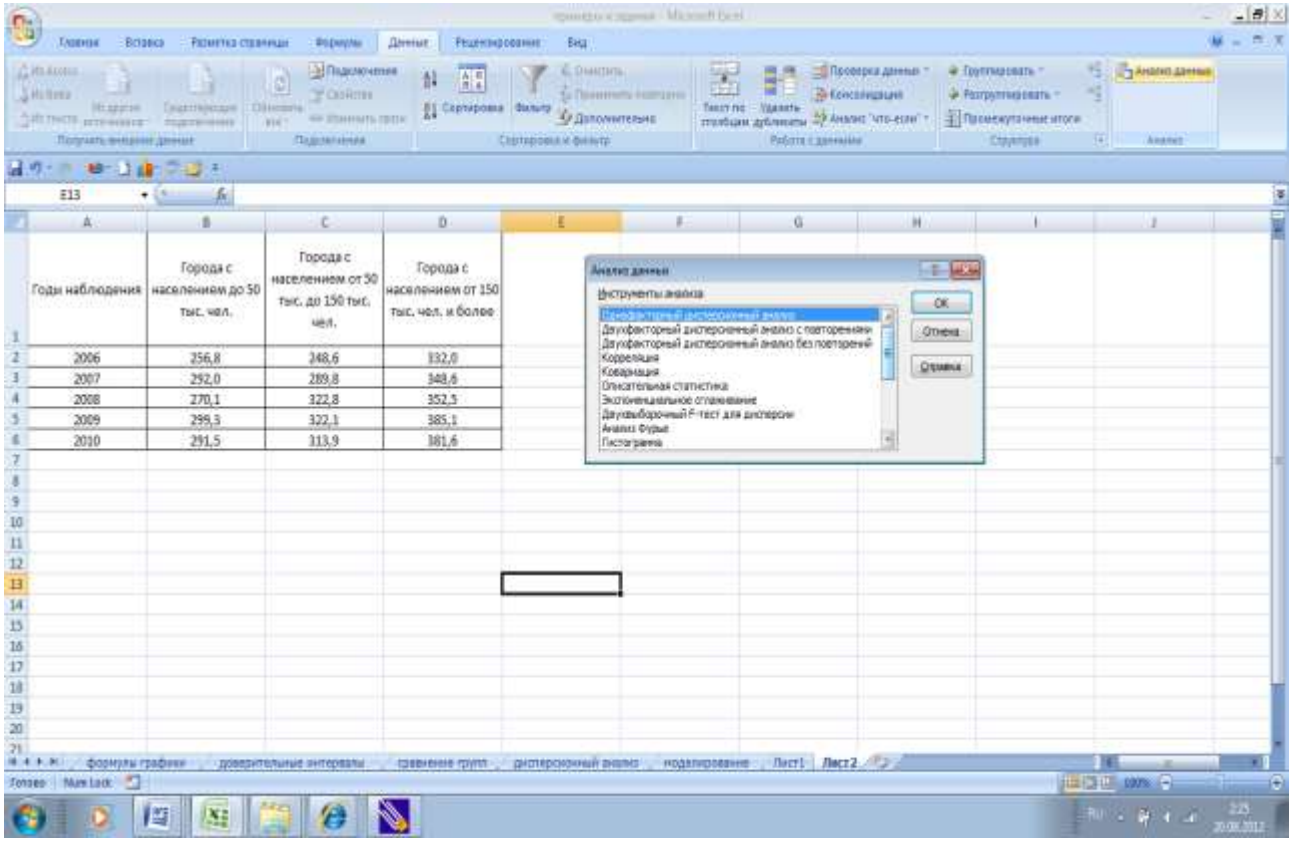


Рис. 20. Выбор инструмента анализа

В диалоговом окне однофакторного дисперсионного анализа указываем входной интервал ($B$2:$D$6), группирование по столбцам, выходной интервал ($A$8), уровень значимости 0,05; рис. 21).

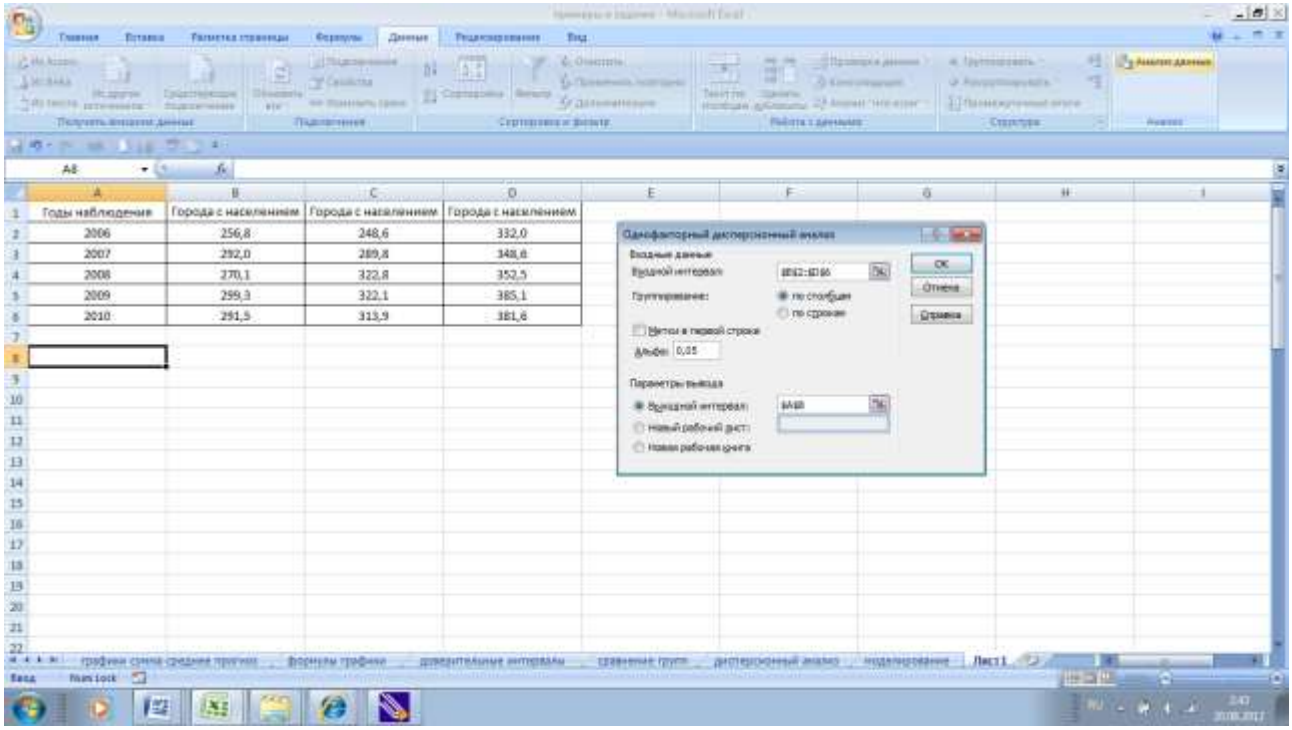


Рис. 21. Диалоговое окно однофакторного дисперсионного анализа

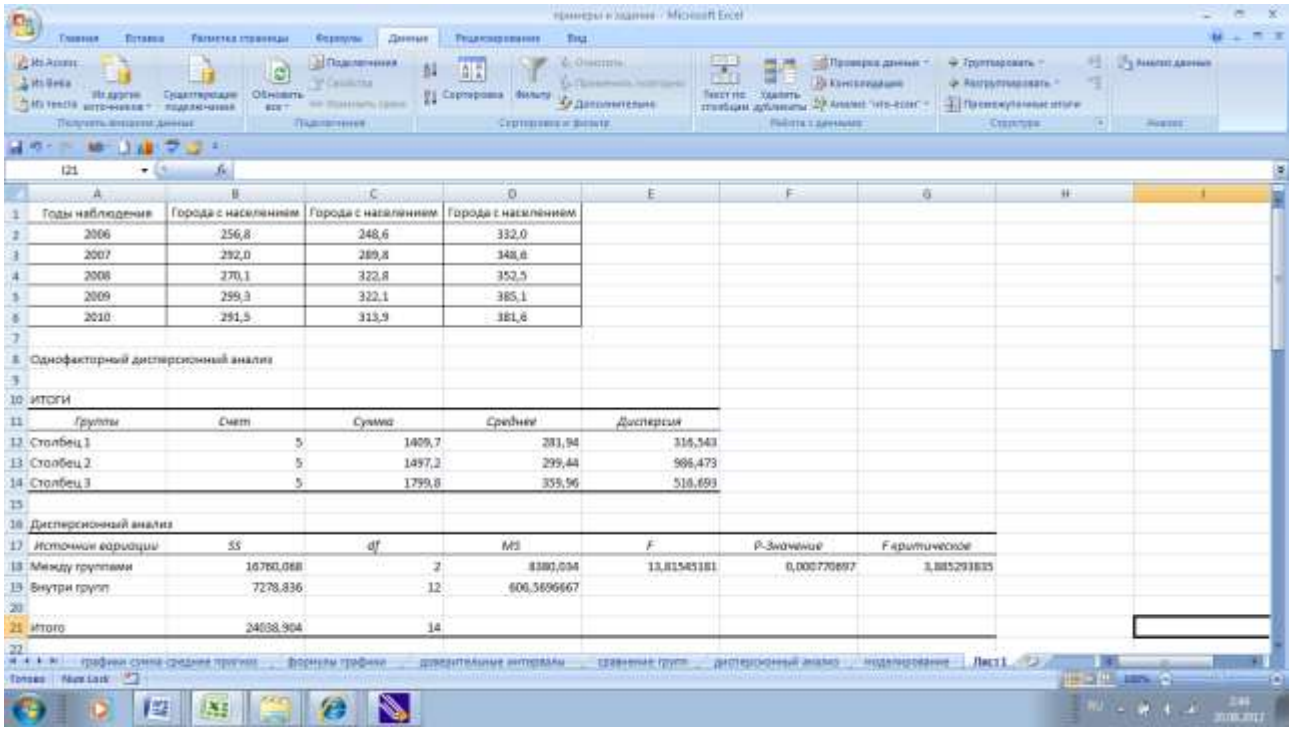
В таблице ИТОГИ (рис. 22) представлены “Счет” – число повторностей, “Сумма” – сумма значений показателя по строкам, “Дисперсия” – частная дисперсия показателя.

В таблице ANOVA представлены результаты дисперсионного анализа однофакторного комплекса, в котором первая колонка “Источник вариации” содержит наименование дисперсий, графа “SS” - это сумма квадратов отклонений, “df” - степень свободы, графа “MS” - средний квадрат, “F” - критерий фактического F – распределения..“P - значение” - вероятность того, что дисперсия, воспроизводимая уравнением, равна дисперсии остатков. Определяет вероятность того, что полученная количественная определенность взаимосвязи между факторами и результатом может считаться случайной. “F - критическое” - это значение F – теоретического, которое впоследствии сравнивается с F – фактическим.

Вручную остается рассчитать коэффициент детерминации и сформулировать выводы.

$$R^2 = \frac{SS_{bg}}{SS_{total}} = \frac{16760,07}{24038,904} = 0,697 \approx 0,7$$

Результаты расчетов вручную и обработанные с помощь компьютерной программы совпадают.



Годы наблюдения	Города с населением	Города с населением	Города с населением
2006	256,8	248,6	332,0
2007	292,0	289,8	348,8
2008	270,1	322,8	352,5
2009	299,3	322,1	385,1
2010	291,5	313,9	381,6

Однофакторный дисперсионный анализ

ИТОГИ

Группы	Счет	Сумма	Среднее	Дисперсия
Столбец 1	5	1409,7	281,94	316,543
Столбец 2	5	1497,2	299,44	986,473
Столбец 3	5	1799,8	359,96	516,693

Дисперсионный анализ

Источник вариации	SS	df	MS	F	P-Значение	F критическое
Между группами	16760,068	2	8380,034	13,81545181	0,000770697	3,885293835
Внутри групп	7278,836	12	606,5696667			
Итого	24038,904	14				

Рис. 22. Результаты однофакторного дисперсионного анализа

Вывод: Для всех территорий можно с вероятностью 99,7%; (Fфакт. > Fкрит.; 13,8>3,9) утверждать, что урбанизация влияет на уровень онкологической патологии. Доля влияния урбанизации на уровень онкологической патологии составляет 70% в общем числе других факторов. Можно предположить, что высокий уровень онкологической заболеваемости в крупных городах связан с лучшей выявляемостью данной патологии.

Рассмотрим технологию проведения однофакторного дисперсионного анализа, для независимых групп в программе **SPSS Statistics**.

Построим дисперсионный комплекс непосредственно в программе **SPSS Statistics** (рис. 23).

Выбираем «Анализ» → «Сравнение средних» → «Однофакторный дисперсионный анализ» (рис. 24). В открывшемся окне выделяем и переносим из

левого окна переменные при помощи соответствующей кнопки: зависимую переменную «заболеваемость по городам» в «список зависимых переменных», переменную «города» переносим в нижнее окно «Фактор» (рис. 25). Открываем вкладку «Параметры». В открывшемся окне диалога отмечаем «Описательная» и «Проверка однородности дисперсий» → «Продолжить» (рис. 25).

О необходимости равенства дисперсий при проведении параметрического дисперсионного анализа иже упоминалось ранее. Если в выводе не подтвердиться равенство дисперсий следует воспользоваться процедурой непараметрического дисперсионного аналога для несвязанных групп – методом Краскела-Уоллиса (H).

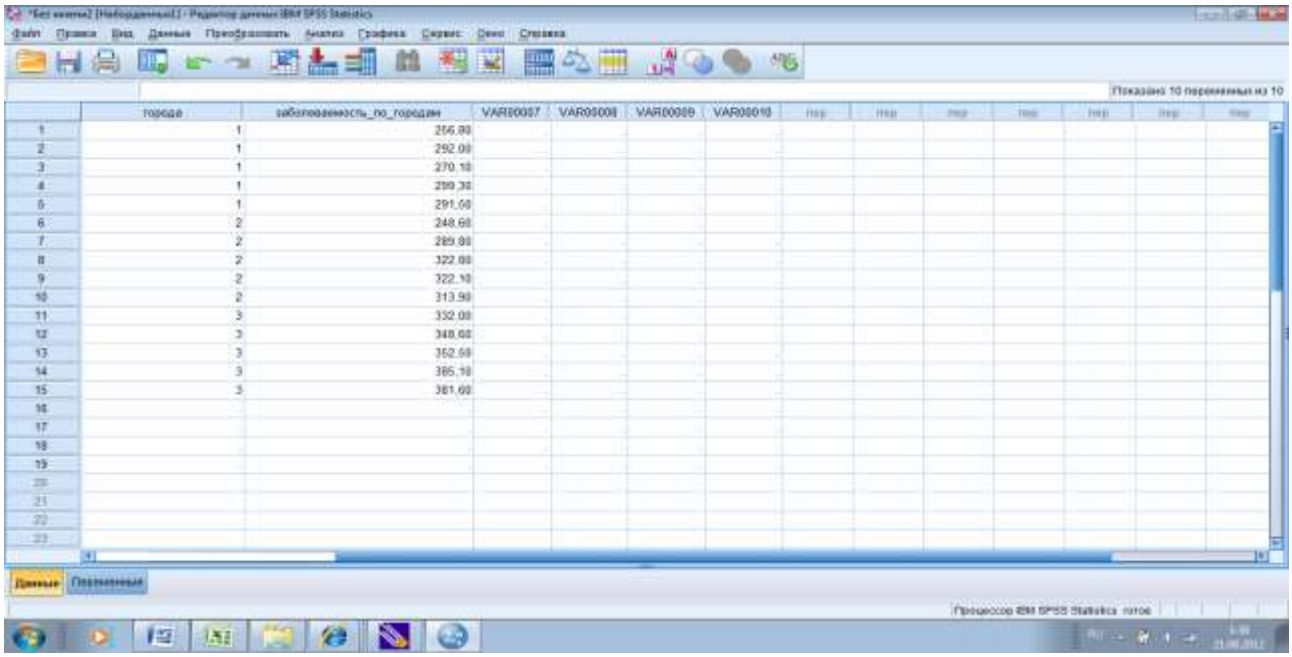


Рис. 23. Ввод данных для сравнения трех несвязанных совокупностей в программе SPSS Statistics методом однофакторного дисперсионного анализа.

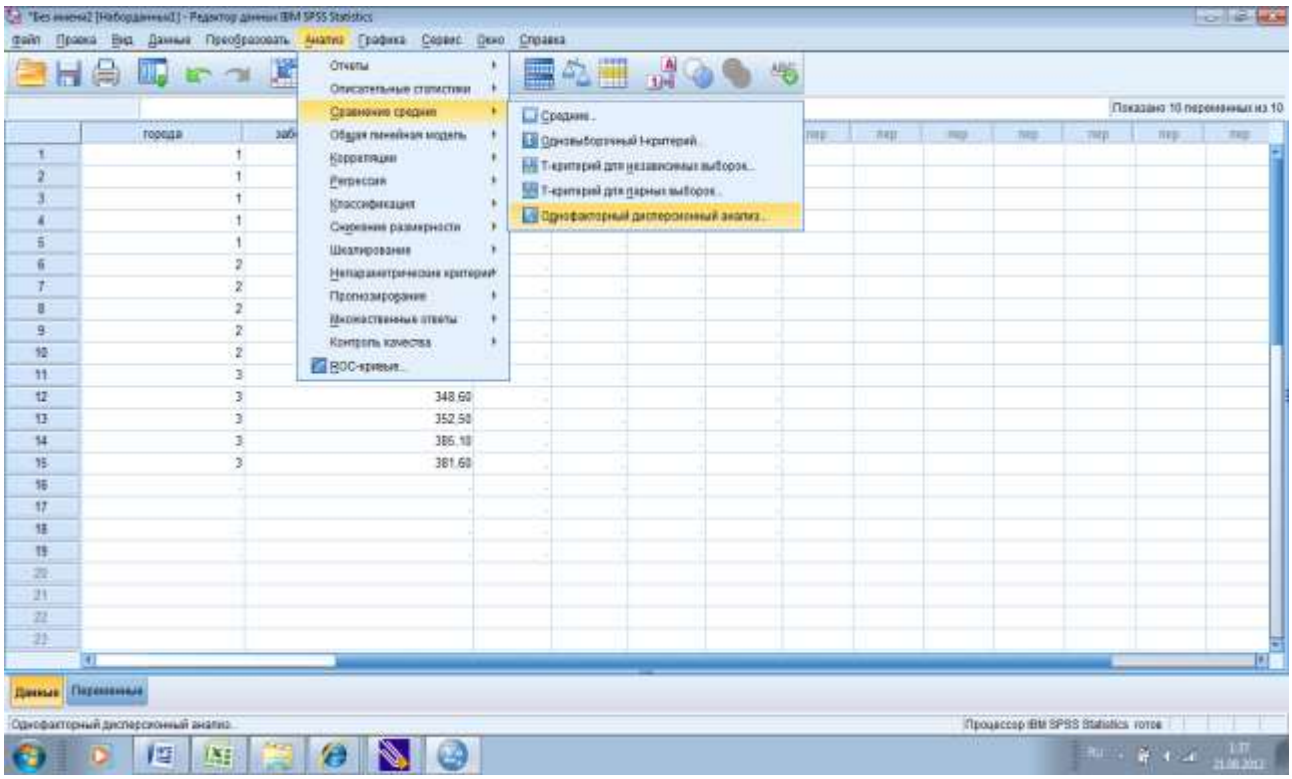


Рис. 24. Настройка однофакторного дисперсионного анализа для независимых групп в программе SPSS Statistics

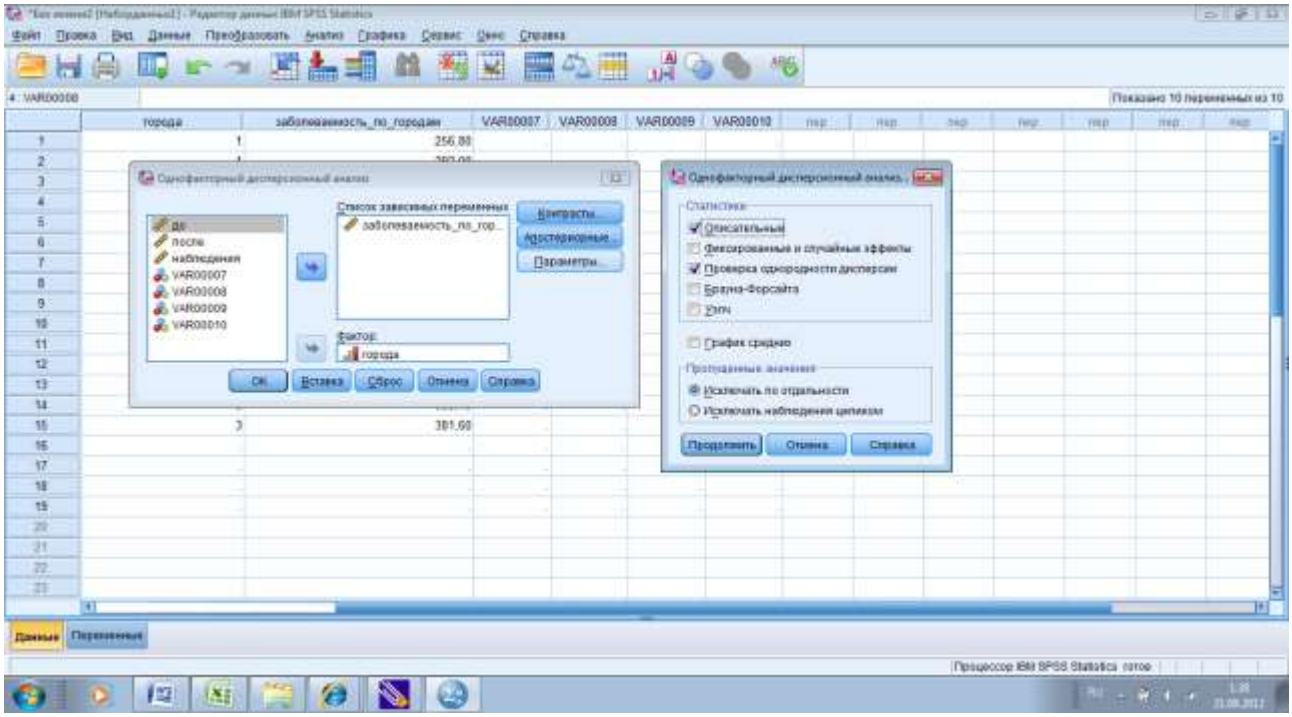


Рис. 25. Настройка однофакторного дисперсионного анализа для независимых групп в программе SPSS Statistics

Для более конкретного вывода о том, какие именно совокупности различаются, в однофакторном дисперсионном анализе предусмотрена процедура множественных сравнений. Для этого откройте вкладку «Апостериорные» , выберете критерий Шеффе → «Продолжить» → «ОК» (рис. 26).

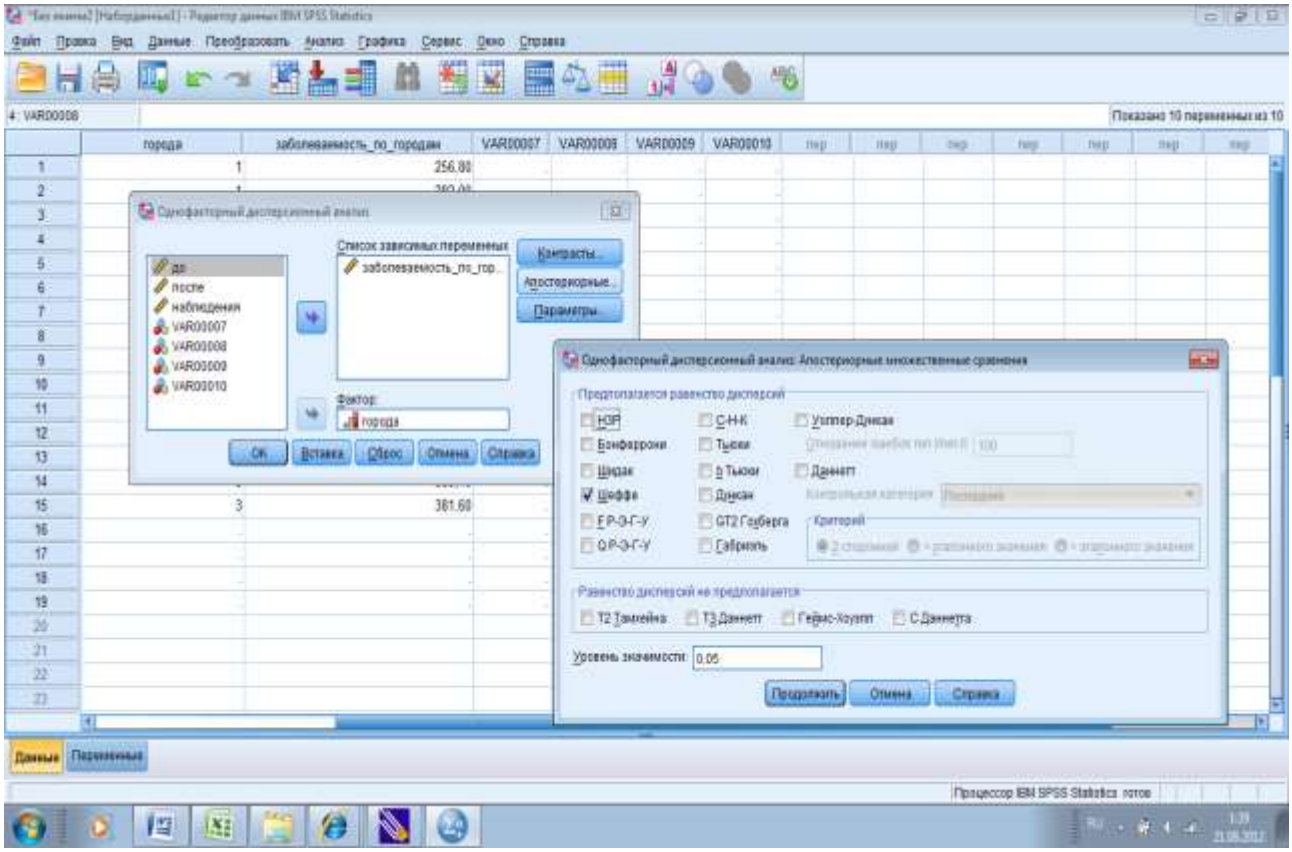


Рис. 26. Настройка однофакторного дисперсионного анализа для независимых групп в программе SPSS Statistics

В статистическом отчете выведены результаты описательной статистики, результаты проверки однородности дисперсий и результаты дисперсионного анализа (рис. 27).

Вывод: Уровень заболеваемости в городах с численностью населения до 50 тыс. чел. (1) 281,9‰ (95%ДИ: 259,8-304,0), в городах с населением от 50 тыс. до 150 тыс. чел. (2) – 299,4‰ (95%ДИ: 260,4-338,4), и в города с

населением от 150 тыс. чел. и более (3) – 359,96‰ (95%ДИ: 331,7-388,2); (рис. 27).

Различие дисперсий не значимо (p=0,437), следовательно результатам анализа можно доверять (рис. 27).

Урбанизация влияет на уровень онкологической заболеваемости (p=0,001). Доля влияния урбанизации на уровень онкологической патологии составляет 70% в общем числе других факторов:

$$R^2 = \frac{SS_{bg}}{SS_{total}} = \frac{16760,07}{24038,904} = 0,697 \approx 0,7$$

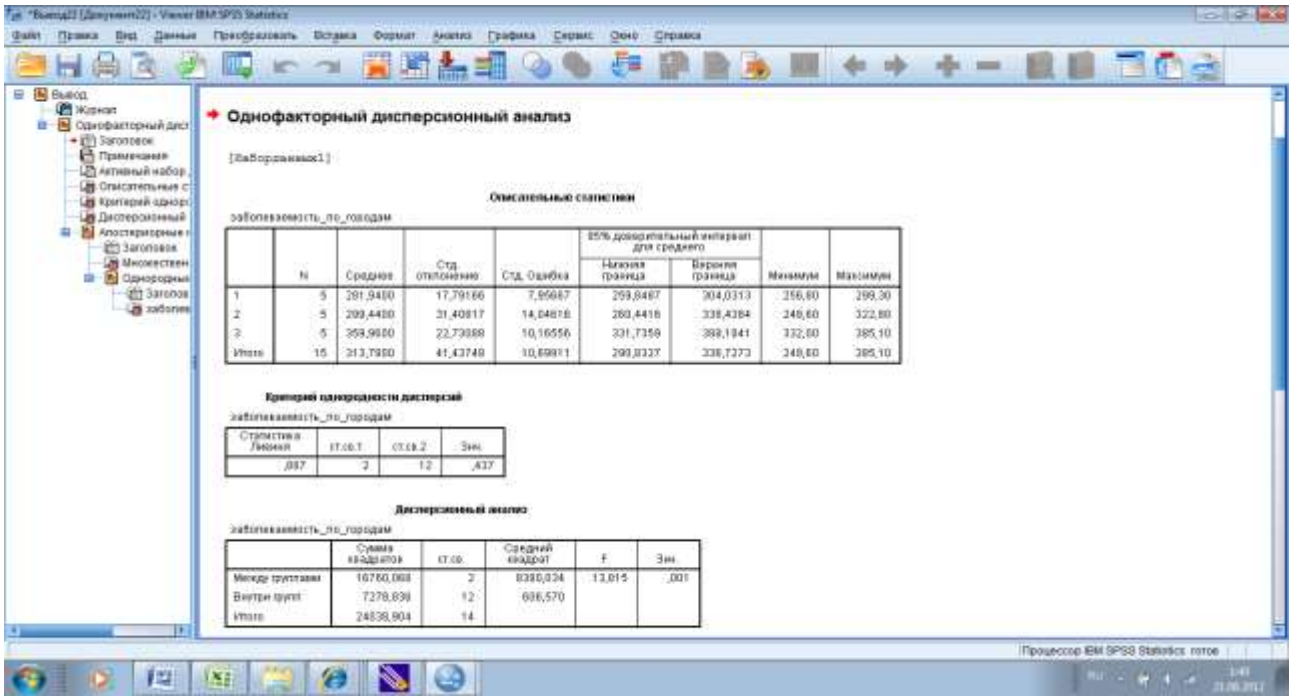


Рис. 27. Статистический отчет однофакторного дисперсионного анализа для независимых групп в программе SPSS Statistics 19

Результаты множественных сравнений показывают, что статистически значимо уровни онкологической заболеваемости отличаются в городах с численностью населения до 50 тыс. чел и в города с населением от 150 тыс. чел. годах (p=0,001), а также в городах с населением от 50 тыс. до 150 тыс. чел. и в города с населением от 150 тыс. чел. и более (p=0,008); рис. 28).

Результаты демонстрируют отсутствие статистически значимых различий дисперсий выборок 1 и 2 (Знч. 0,549), 2 и 3 (Знч. 1,0), что убеждает в корректности парных сравнений средних значений.

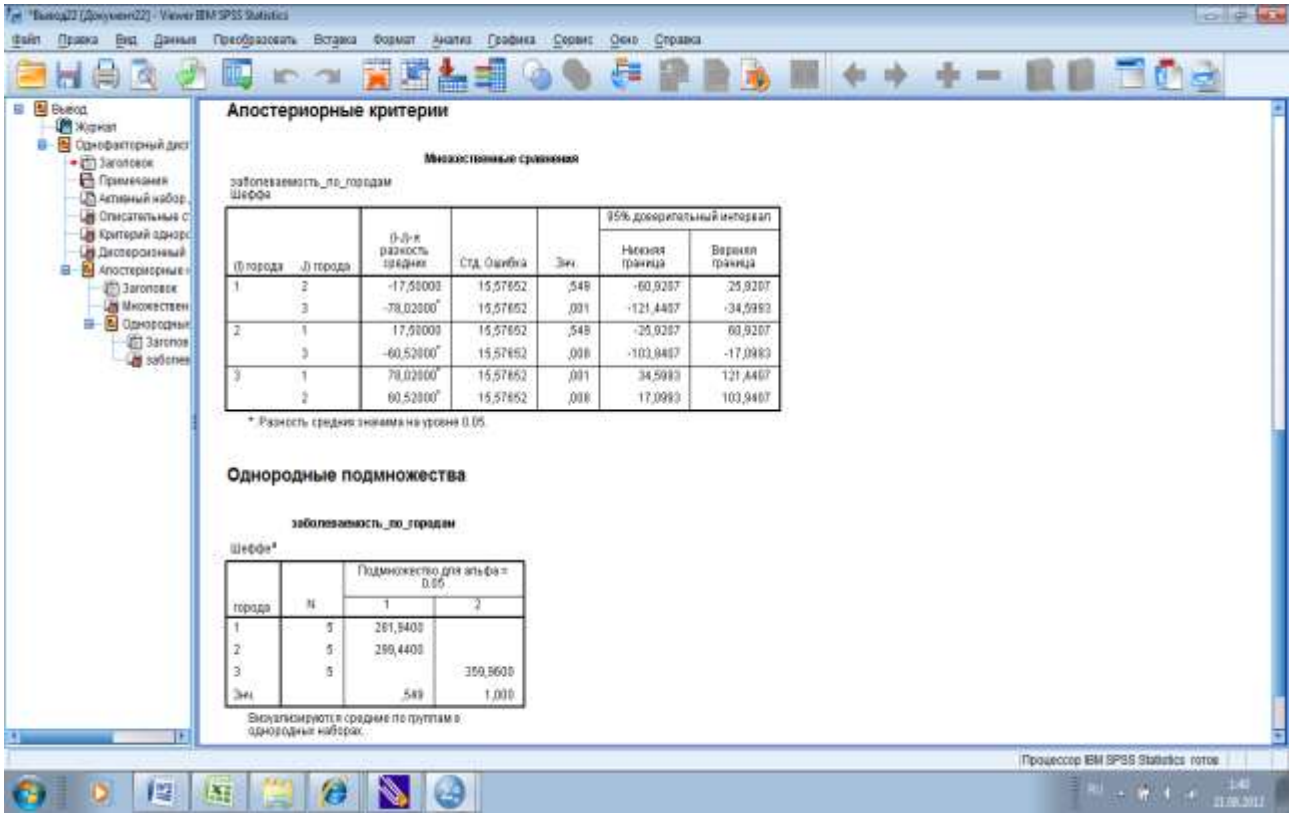


Рис. 28. Статистический отчет однофакторного дисперсионного анализа для независимых групп в программе SPSS Statistics 19

ПАРАМЕТРИЧЕСКИЙ МЕТОД СРАВНЕНИЯ ТРЕХ ЗАВИСИМЫХ ГРУПП И БОЛЕЕ ПО ОДНОМУ ПРИЗНАКУ ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ ДЛЯ ЗАВИСИМЫХ ГРУПП

При анализе зависимых выборок выделяют межгрупповые и внутригрупповые факторы. Разным градациям межгруппового фактора соответствуют разные группы объектов, а разным градациям внутригруппового фактора соответствует одна и та же группа объектов. Для анализа таких выборок применяется ***дисперсионный анализ с повторными***

измерениями. Эти выборки можно принять за независимые и сравнить их при помощи обычного дисперсионного анализа, но дисперсионный анализ с повторными измерениями является более чувствительным к влиянию изучаемых факторов в случае, когда выборки являются зависимыми. Дисперсионный анализ с повторными измерениями позволяет с большей надежностью обнаруживать факторные эффекты. Специфика этого метода заключается в том, что из остаточной изменчивости (внутригрупповой) вычитается изменчивость между средними для каждого объекта наблюдения (межиндивидуальная).

Как и для всех параметрических методов, необходимым условием для применения этого метода является допущение о том, что множество измерений зависимой переменной для каждого объекта (обследуемого) является выборкой из многомерного нормального распределения.

Если используется одномерная модель, предполагается наличие коррелированности измерений зависимой переменной и равенство дисперсий зависимой переменной для разных уровней внутригруппового фактора. Для проверки этого предположения в компьютерных программах используется тест сферичности ковариационно-дисперсионной матрицы Моучли. Если этот тест показывает статистически значимый результат, то предположение о сферичности считается ошибочным и одномерный подход не приемлем. Этот тест имеет малую чувствительность и при нарушении допущения о сферичности компьютерные программы предлагают ввести специальную поправку – эпсилон-коррекцию, но более правильным будет провести многомерный тест.

Многомерный подход не предполагает сферичности. При многомерном анализе проводятся тесты «След Пиллая и «λ-Вилкса».

При использовании межгрупповых факторов проводится проверка допущения об идентичности ковариационно-дисперсионных матриц, соответствующих разным уровням межгрупповых факторов. Для проверки

этого допущения используется М-тест Бокса. Если он показывает статистически значимый результат, то ковариационно-дисперсионные матрицы не идентичны, и применение многомерного метода не корректно.

Рассмотрим пример двухфакторного дисперсионного анализа с повторными измерениями по одному из факторов. Пациентам диспансерной группы была назначена гипохолестериновая диета. В группу обследования вошли пациенты с высокой и низкой физической активностью (фактор А межгрупповой). Забор крови на содержание холестерина осуществлялся до того как пациентам была назначена диета (Хол 1), через 1 месяц (Хол 2) и через два месяца (Хол 3); (фактор В внутригрупповой). Результаты обследования представлены в таблице 14. Врачу необходимо оценить а) способствует ли назначенная диета снижению содержания холестерина в крови; б) способствует ли физическая активность снижению содержания холестерина в крови и в) способствует ли диета и физическая активность снижению содержания холестерина в крови.

Рассчитаем общую изменчивость:

SS_{total} =(N-1) x σ^2=29 x 1,86=53,9

Рассчитаем изменчивость между уровнями межгруппового фактора (SS_A):

$$SS_A = n \times l \times \sum_{i=1}^{k} (M_i - M\ )^2 = 5 \times 3 \times [(5{,}5 - 5{,}8)^2 + (6{,}2 - 5{,}8)^2] = 5 \times 3 \times 0{,}25 = 4{,}03$$

Рассчитываем изменчивость между испытуемыми внутри уровней межгруппового фактора (SS_{IWG}).

$$SS_{IWG} = l \times \sum_{j=1}^{k} \sum_{i=1}^{n} (M_i - M_k)^2 = 3 \times [(5{,}3 - 5{,}5)^2 + (5{,}3 - 5{,}5)^2 + (5{,}7 - 5{,}5)^2 +$$
$$+ (5 - 5{,}5)^2 + (6 - 5{,}5)^2 + (6{,}3 - 6{,}2)^2 + (6 - 6{,}2)^2 + (6{,}3 - 6{,}2)^2 +$$
$$+ (5{,}3 - 6{,}2)^2 + (7 - 6{,}2)^2] = 3 \times 2{,}044 = 6{,}13$$

SS_{lWG} это мера ошибки межгрупповой факторной модели или фактора B.

Рассчитываем межиндивидуальную изменчивость (SS_{bs}):

$$SS_{bs} = l \times \sum_{i=1}^{nk} (M_i - M)^2 = 3 \times [(5{,}3 - 5{,}8)^2 + (5{,}3 - 5{,}8)^2 + (5{,}3 - 5{,}8)^2 + \ldots$$

$$\ldots + (7 - 5{,}8)^2] = 10{,}2$$

Рассчитаем внутригрупповую изменчивость (SS_{wg}):

$SS_{wg} = SS_{total} - SS_{bs} = 53{,}9 - 10{,}2 = 43{,}7$

Таблица 12. Результаты исследования крови на содержание холестерина.

Пациенты	***фактор B три уровня***			фактор А два уровня	M_{1-10}
	хол 1	***хол 2***	***хол 3***	физическая активность	
1	***8***	***5***	***3***	высокая (А1)	5,3
2	***8***	***4***	***4***	высокая (А1)	5,3
3	***8***	***4***	***5***	высокая (А1)	5,7
4	***6***	***5***	***4***	высокая (А1)	5,0
5	***7***	***5***	***6***	высокая (А1)	6,0
M	***7,4***	***4,6***	***4,4***		
6	***7***	***6***	***6***	низкая (А 2)	6,3
7	***7***	***5***	***6***	низкая (А 2)	6,0
8	***7***	***6***	***6***	низкая (А 2)	6,3
9	***6***	***4***	***6***	низкая (А 2)	5,3
10	***8***	***7***	***6***	низкая (А 2)	7,0
M	***7***	***5,6***	***6***		
$M\ B_{1-3}$	***7,22***	***5,05***	***5,13***	M	**5,8**
σ^2	1,8678				
$M\ A_1$	5,5				
$M\ A_2$	6,20				

Рассчитываем изменчивость обусловленную влиянием внутригруппового фактора B (SS_B):

$$SS_B = N \times \sum_{i=1}^{l} (M_i - M)^2 = 10 \times [(7{,}2 - 5{,}8)^2 + (5{,}1 - 5{,}8)^2 + (5{,}2 - 5{,}8)^2 = 28{,}06$$

Рассчитываем изменчивость обусловленную влиянием взаимодействия межгруппового и внутригруппового фактора (SS_{AB}):

$$SS_{AB} = n \times \sum_{i=1}^{k} \sum_{j=1}^{l} (M_{i_j} - M)^2 - SS_A - SS_B = 5 \times [(7{,}4 - 5{,}8)^2 + (4{,}6 - 5{,}8)^2 + \ldots$$

$$\ldots + (4{,}4 - 5{,}8)^2 + (7 - 5{,}8)^2 + (5{,}6 - 5{,}8)^2 + (6 - 5{,}8)^2] - 3{,}75 - 28{,}06 = 5 \times 7{,}48 - 4{,}03 -$$

$$- 28{,}06 = 5{,}26$$

Вычисляет остаточную сумму квадратов (SS_{erB}):

$SS_{erB} = SS_{wg} - SS_B - SS_{AB} = 43{,}7 - 28{,}06 - 5{,}05 = 10{,}6$

Определяем числа степеней свободы:
$df_A = k - 1 = 1$
$df_B = l - 1 = 2$
$df_{IWG} = N - k = 8$
$df_{AB} = (k-1)(l-1) = 2$
$df_{erB} = (N-k)(l-1) = 16$

Вычисляем средние квадраты:

$$MS_A = \frac{SS_A}{df_A} = \frac{4{,}03}{1} = 4{,}03$$

$$MS_{IWG} = \frac{SS_{IWG}}{df_{IWG}} = \frac{6{,}13}{8} = 0{,}76$$

$$MS_B = \frac{SS_B}{df_B} = \frac{28{,}06}{2} = 14{,}03$$

$$MS_{erB} = \frac{SS_{erB}}{df_{erB}} = \frac{10{,}6}{16} = 0{,}66$$

$$MS_{AB} = \frac{SS_{AB}}{df_{AB}} = \frac{5{,}26}{2} = 2{,}6$$

Вычисляем F-отношения:

$$F_A = \frac{MS_A}{MS_{IWG}} = \frac{4{,}03}{0{,}76} = 5{,}3$$

$$F_B = \frac{MS_B}{MS_{erB}} = \frac{14{,}03}{0{,}66} = 21{,}2$$

$$F_{AB} = \frac{MS_{AB}}{MS_{erB}} = \frac{2{,}6}{0{,}66} = 3{,}9$$

Вывод. H_o отклоняется только в отношении фактора А (физическая активность; Fфакт. < Fкрит.). Установлено, что гипохолестериновая диета способствует снижению холестерина в крови (Fфакт. > Fкрит). Выявлено, на статистически значимом уровне, что гипохолестеиновая диета и физическая нагрузка также способствуют снижению холестерина в крови (Fфакт. > Fкрит).

Введем данные рассмотренного примера в программу для статистической обработки данных. Из-за отсутствия в базовой версии IBM SPSS Statistics дисперсионного анализа с повторными измерениями проведем расчеты в программе Statistica 6.1. Это система для статистического анализа данных, включающая широкий набор аналитических процедур и методов занимает. На сегодняшний день программа Statistica занимает лидирующее положение среди специализированных статистических программ.

Откройте программу и введите цифровые данные из примера в электронную таблицу. Выберете в меню «Анализ» Дисперсионный анализ (29).

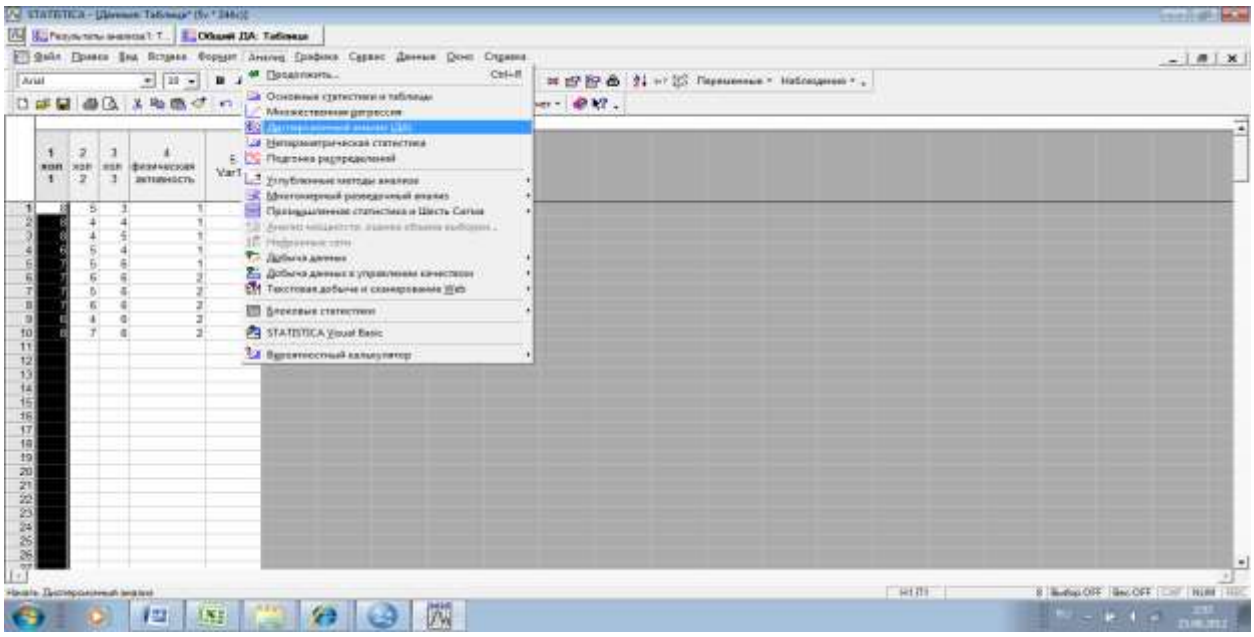


Рис. 29. Настройка дисперсионного анализа с повторными измерениями в программе Statistica

В открывшемся диалоговом окне «Вид анализа» повторные измерения ДА, в окне «Задание анализа» - диалог → «ОК» (рис. 30).

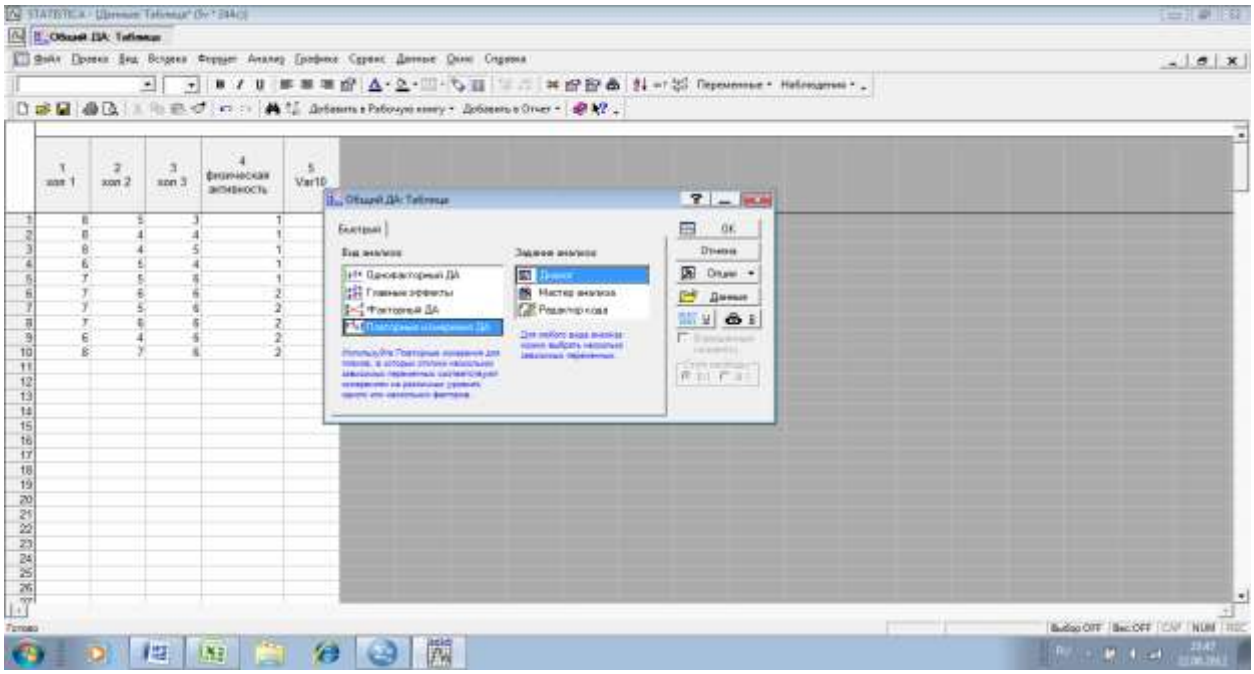


Рис. 30. Дисперсионный анализ с повторными измерениями в программе Statistica 6.1

В новом диалоговом окне выберете зависимые переменные (хол 1, хол 2, хол 3) и категориальные предикторы (физическая активность) → «ОК» (рис 31).

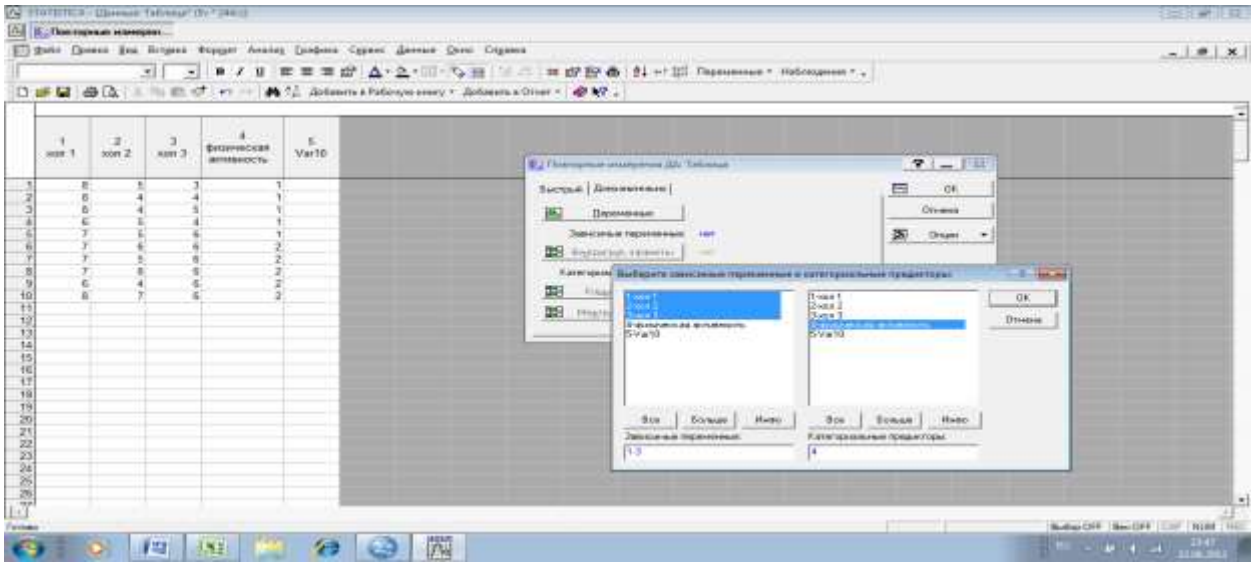


Рис. 31. Настройка дисперсионного анализа с повторными измерениями в программе Statistica

Задайте фактор с повторными измерениями и количество уровней (диета (3 уровня) → «ОК» (рис. 32).

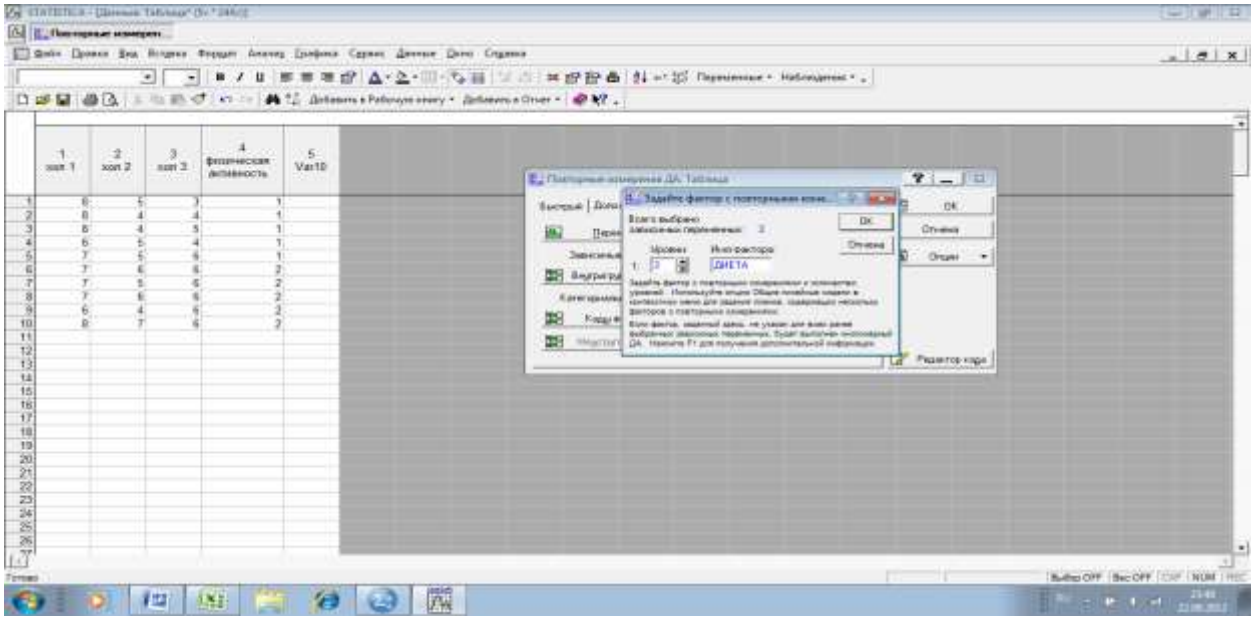


Рис. 32. Настройка дисперсионного анализа с повторными измерениями в программе Statistica

Выбираем коды факторов для независимых переменных нажатием на кнопку «Все» → «ОК» (рис. 33).

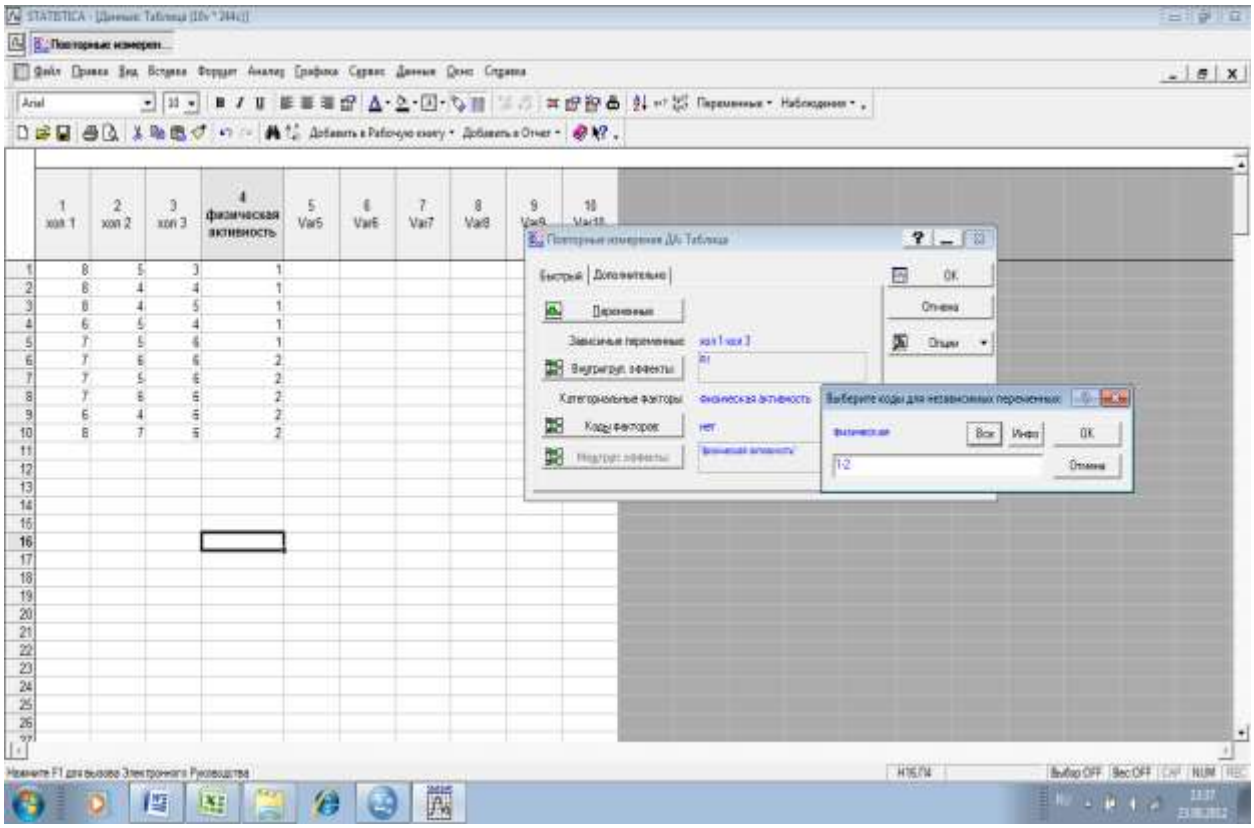


Рис. 33. Настройка дисперсионного анализа с повторными измерениями в программе Statistica

В окне «Результаты анализа 1: Таблица». Выберете кнопку «Итоги» и кликните по ней (рис. 34). Выберете критерии для многомерного анализа – «Пиллая» и «Хотеллинга» отметив их флажком.

Если используется одномерная модель, как было отмечено ранее, предполагается наличие коррелированности измерений зависимой переменной и равенство дисперсий зависимой переменной для разных уровней внутригруппового фактора. Для проверки этого предположения используется тест сферичности ковариационно-дисперсионной матрицы Моучли (рис. 67). Для проведения этого теста во внутригрупповых эффектах выберете «Сферичность».

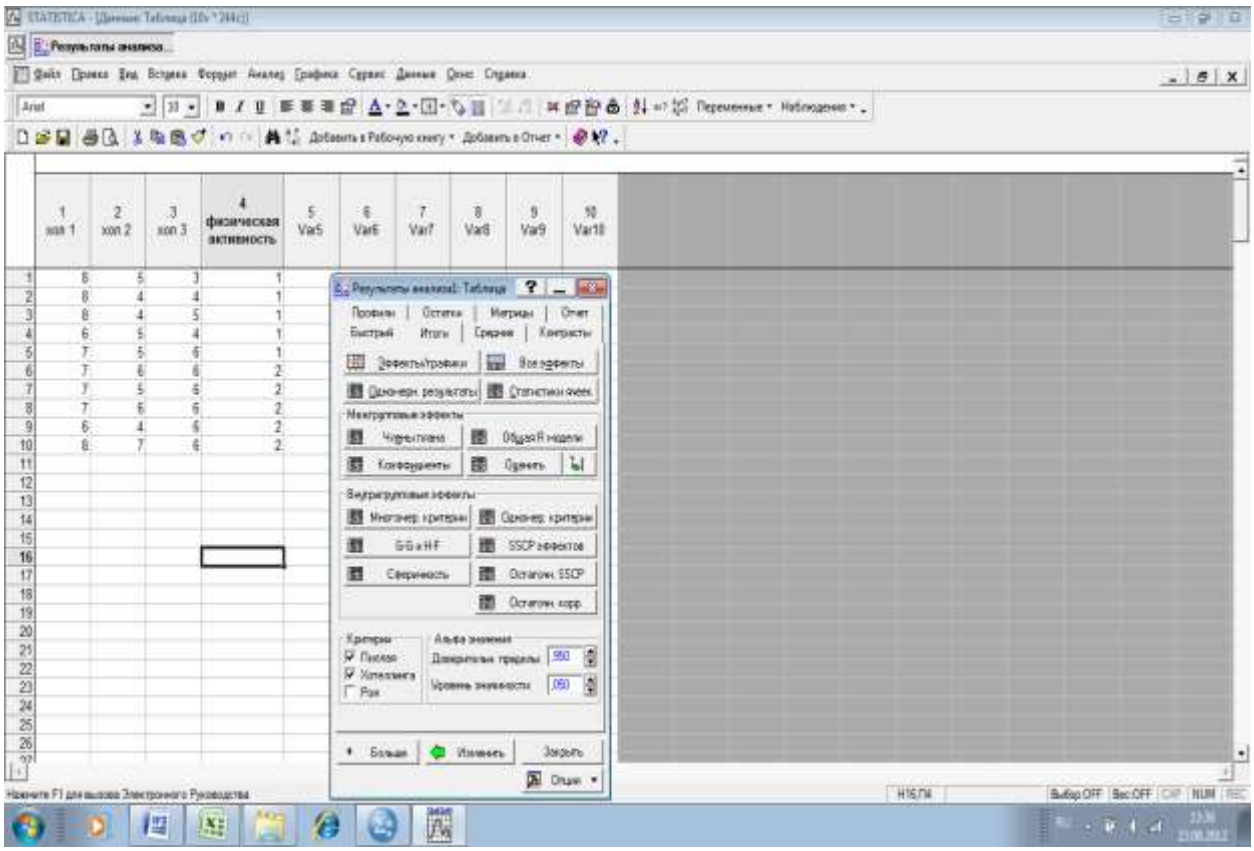


Рис. 34. Настройка дисперсионного анализа с повторными измерениями в программе Statistica

Тест Моучли показывает, что предположение о сферичности подтверждается (p=0,79), следовательно одномерный подход корректен (рис. 35).

Вернитесь к окну «Результаты Анализа» и выберете «Все эффекты» (рис. 36). Результаты одномерного анализа в появившейся таблице (рис.37) не противоречат расчетам вручную. Вывод. H_o отклоняется только в отношении фактора А (физическая активность; p=0,0509.). Установлено, что гипохолестериновая диета способствует снижению холестерина в крови (p=0,00003).

Если исследователя интересует, оказывает ли влияние совместное воздействие факторов (диета+физическая активность) на содержание холестерина в крови проводится многомерный анализ, с использованием

тестов «След Пиллая и «λ-Вилкса», «Хотеллинга» которые были отмечены ранее (рис. 34).

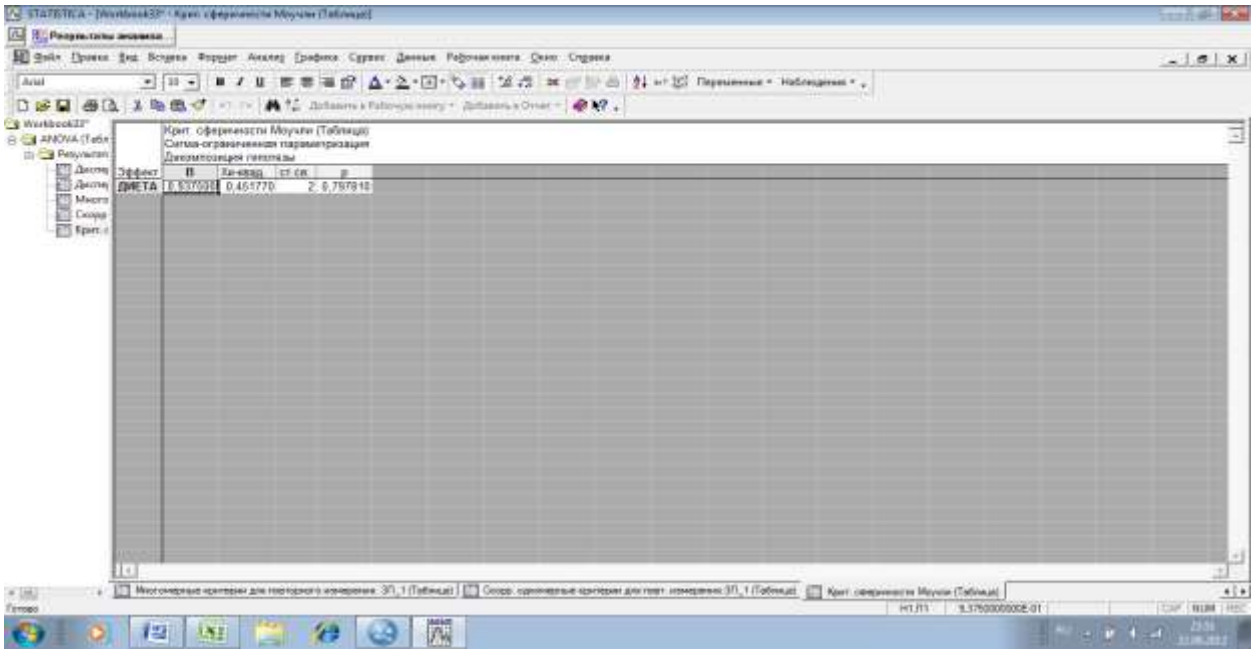


Рис. 35. Результаты дисперсионного анализа с повторными измерениями в программе Statistica

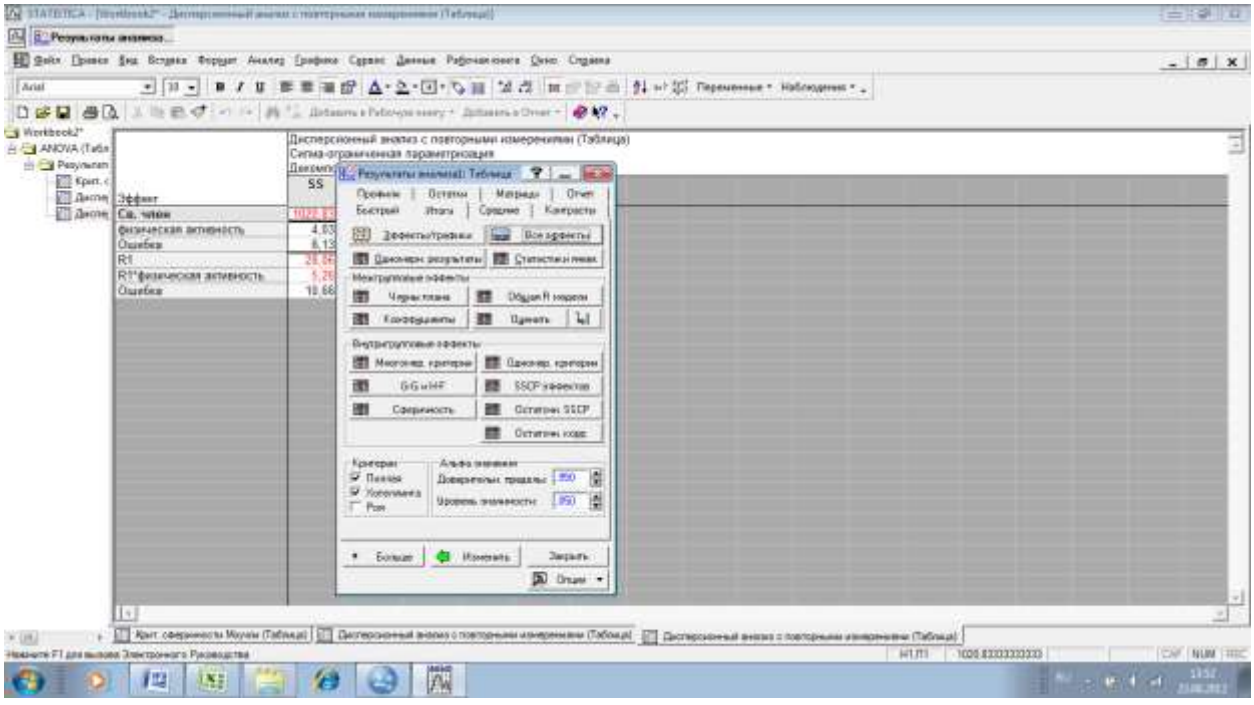


Рис. 36. Результаты дисперсионного анализа с повторными измерениями в программе Statistica

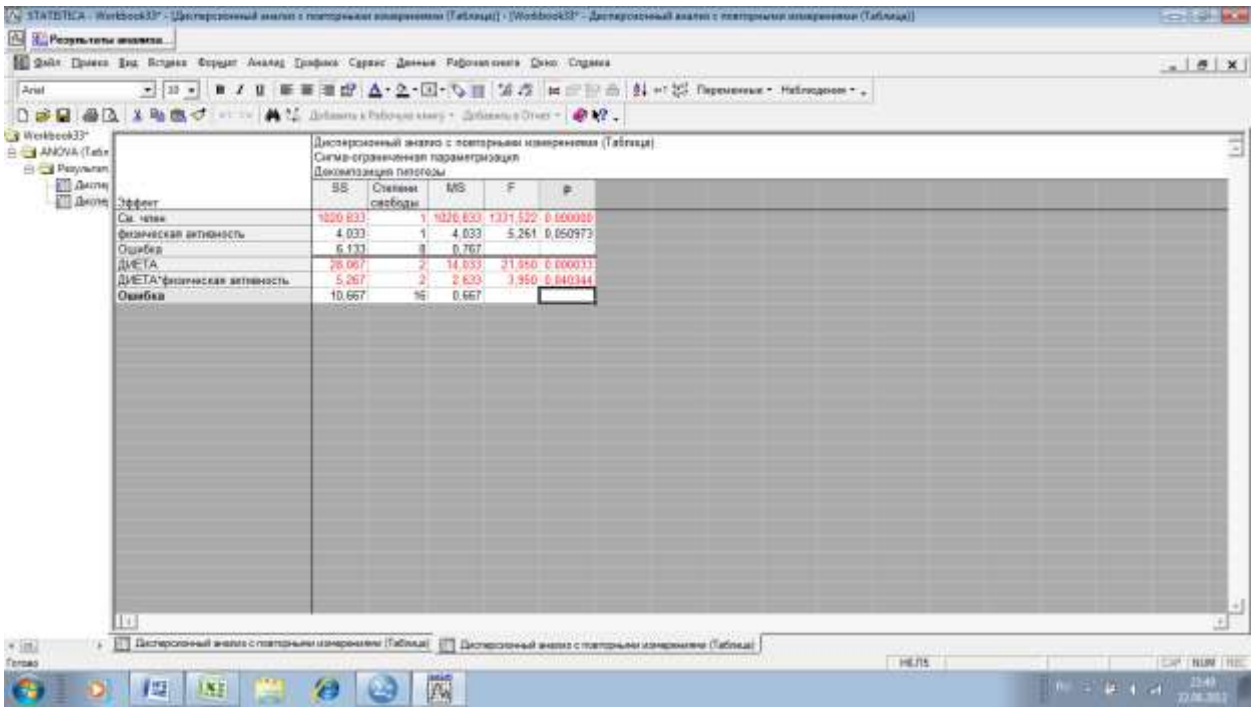


Рис. 37. Результаты одномерного дисперсионного анализа с повторными измерениями в программе Statistica

Многомерный анализ с использование критериев Пиллая, Уилкса, Хотеллинга и др. (рис. 38) показывает что совместное воздействие фактор «Диета» и «Физическая активность» не оказывают статистически значимого влияния на уровень холестерина в крови (p=0,083).

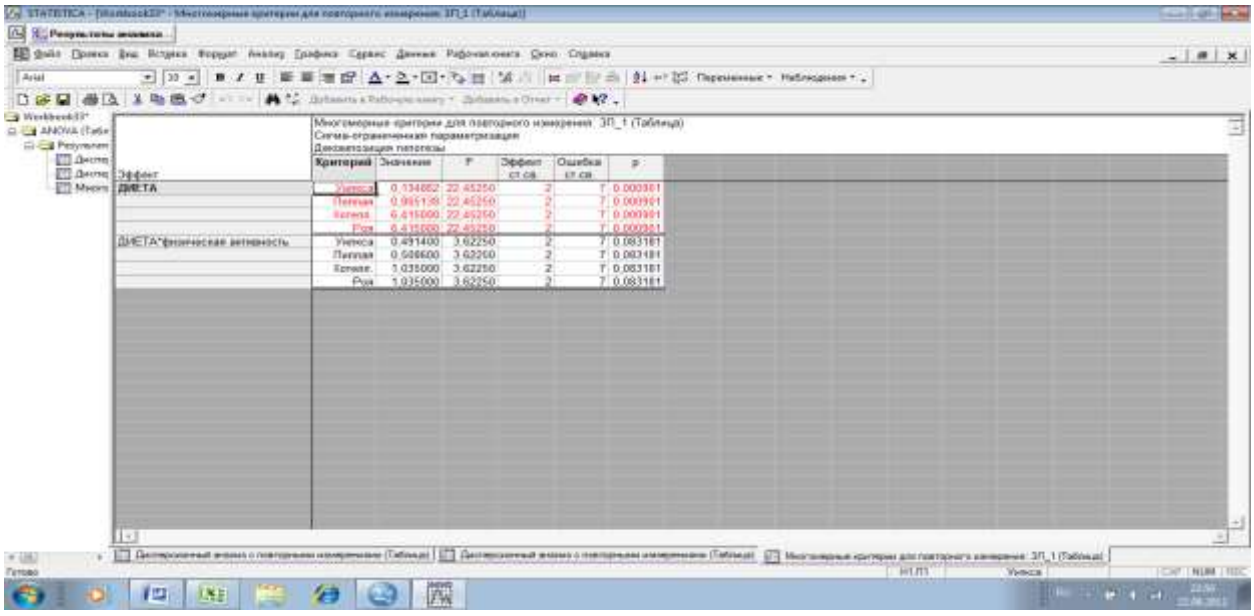


Рис. 38. Результаты многомерного дисперсионного анализа с повторными измерениями в программе Statistica

При использовании межгрупповых факторов проводится проверка допущения об идентичности ковариационно-дисперсионных матриц, соответствующих разным уровням межгрупповых факторов. Для проверки этого допущения в программе IBM SPSS Statistics используется М-тест Бокса. Если он показывает статистически значимый результат, то ковариационно-дисперсионные матрицы не идентичны, и применение многомерного метода не корректно. В программе Statistica данный тест отсутствует, но в случае не идентичности ковариационно-дисперсионных матриц программа не проводит анализ межгрупповых факторов и не выдает результатов многомерного анализа.

НЕПАРАМЕТРИЧЕСКИЕ МЕТОДЫ СРАВНЕНИЯ ТРЕХ ГРУПП И БОЛЕЕ ПО КОЛИЧЕСТВЕННОМУ ПРИЗНАКУ

В том случае если распределение признаков, хоть в одной из сравниваемых совокупностей не подчиняется закону нормального распределения, либо отсутствует равенство дисперсий, для сравнения этих совокупностей применяются непараметрические критерии.

НЕПАРАМЕТРИЧЕСКИЙ МЕТОД СРАВНЕНИЯ ТРЕХ НЕЗАВИСИМЫХ ГРУПП И БОЛЕЕ ПО ОДНОМУ ПРИЗНАКУ ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ КРАСКЕЛА-УОЛЛИСА

Метод основан на построении одного ранжированного ряда с последующим вычислением среднего ранга для каждой выборки. Значение критерия Н- Краскела-Уоллиса вычисляется по формуле:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^{\,2}}{n_i} - 3(N+1)$$

N – суммарная численность всех выборок;
k – количество сравниваемых выборок;

R_i – сумма рангов для выборки *i;*
n_i – численность выборки *i.*

Чем больше различие между выборками, тем больше вычисленное значение H и тем меньше уровень значимости (p).

Рассмотрим данный метод на примере. Изучен дорожно-транспортный травматизм на трех территориях Сибирского Федерального округа. Требуется оценить существуют ли статистически значимые различия в уровнях дорожно-транспортного травматизма на этих территориях. Данные исследования представлены в таблице 13.

Строим один упорядоченный ряд (табл. 14). Значения выборок ранжируются и вычисляются суммы рангов для каждой выборки (табл. 15). Общая сумма рангов должна быть равна N(N+1) / 2= 16 x 17/2=136. Равенство соблюдено.

Таблица 13. Количество ДТТ на 10 тыс. транспортных средств.

Территории	Количество ДТП на 10 тыс. транспортных средств							
1	52,2	53,8	47,9	33,1	40,1	35,1	44,2	30,1
2	41,4	40,7	35,0	24,6	23,1			
3	38,6	35,1	31,4					

Таблица 14. Упорядоченный ряд.

1			30,1		33,1			35,1		40,1			44,2	47,9	52,2	53,8
2	23,1	24,6				35,0					40,7	41,4				
3				31,4			35,1		38,6							

Таблица 15. Ранги

1			3		5			7,5		10			13	14	15	16
2	1	2				6					11	12				
3				4			7,5		9							

R_1=83,5
R_2=32
R_3=20,5

$$H = \frac{12}{16(16+1)}\left(\frac{83{,}5^2}{8} + \frac{32^2}{5} + \frac{20{,}5^2}{3}\right) - 3(16+1) = 2{,}7$$

Вычисленное H-значение следует сравнить с табличным значением χ^2 для числа степеней свободы 2 (df=k-1; k – число выборок 3-1=2). Эмпирическое значение H (2,7) меньше критического (6), следовательно H_o принимается. Различий в уровне дорожно-транспортного травматизма на территориях СФО не обнаружено (p>0,05).

Если установлены статистически значимые различия между выборками у исследователя может возникнуть вопрос, в какой именно выборке изучаемый признак имеет большие или меньшие значения. На основании проведенных расчетов об этом судить нельзя, следует попарно сравнить выборки по критерию Манна-Уитни. При этом необходимо помнить о проблеме множественных сравнений.

В случае множественных сравнений рекомендуется вводить поправку Бонферрони. Например, если исследователь выполнил 5 сравнений, используя критерий Манна-Уитни, то в любом из этих 5 сравнений уровень значимости должен быть меньше 0,01, чтобы сделать вывод об имеющихся различиях сравниваемых групп с уровнем значимости 0,05. Особенностью критерия Бонферрони является то, что он плохо работает при большом числе сравниваемых групп (k>8). Существует и такой подход к преодолению проблемы множественных сравнений, как принятие более жесткого уровня статистической значимости 0,01; 0,005; 0,001 и т.д

Компьютерная обработка.

Внесем данные рассмотренного примера в программу IBM SPSS Statistics (рис 39-40).

В качестве проверяемых полей зададим «Количество ДТТ», групп – «Территории» (рис. 41).

В окне «Параметры» выберем Однофакторный дисперсионный анализ Краскела-Уоллиса (для k выборок) → «Запуск»; (рис. 42).

Вывод. H_o принимается. Различий в уровне дорожно-транспортного травматизма на территориях СФО не обнаружено (p=0,263); (рис. 43).

Рис. 39. Вод данных в программу IBM SPSS Statistics для обработки методом Краскела-Уоллиса

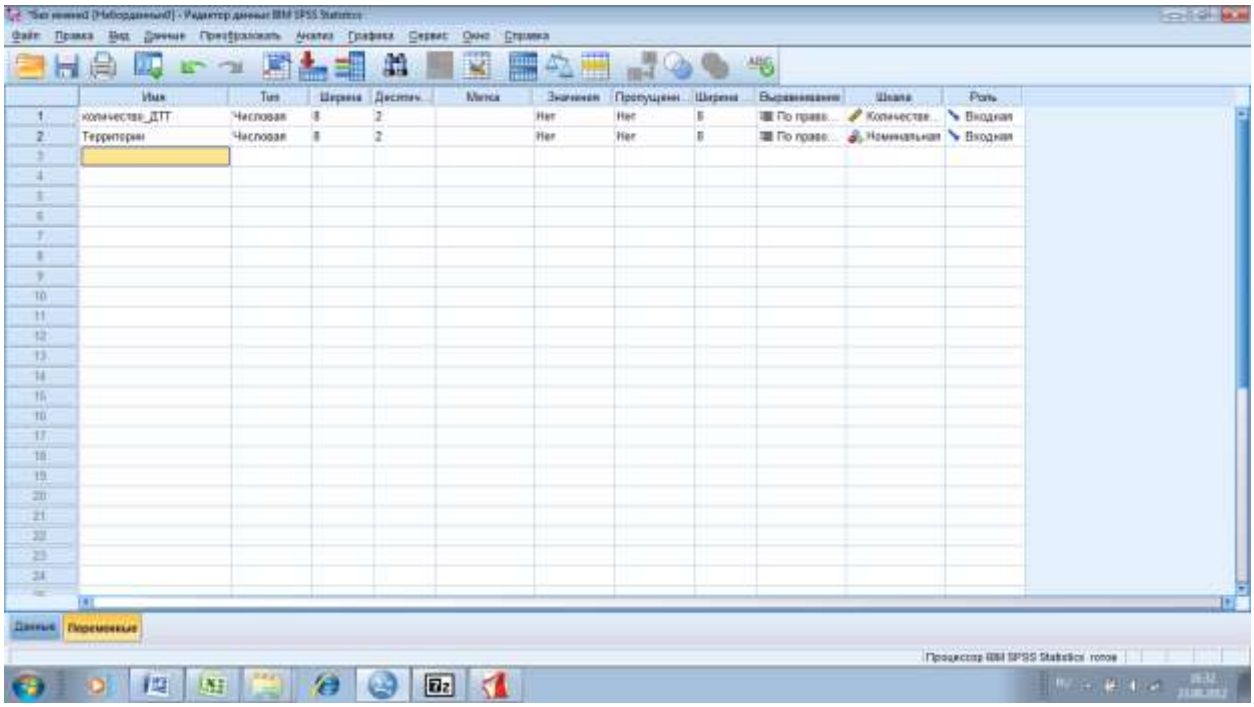


Рис. 40. Вод данных в программу IBM SPSS Statistics для обработки методом Краскела-Уоллиса

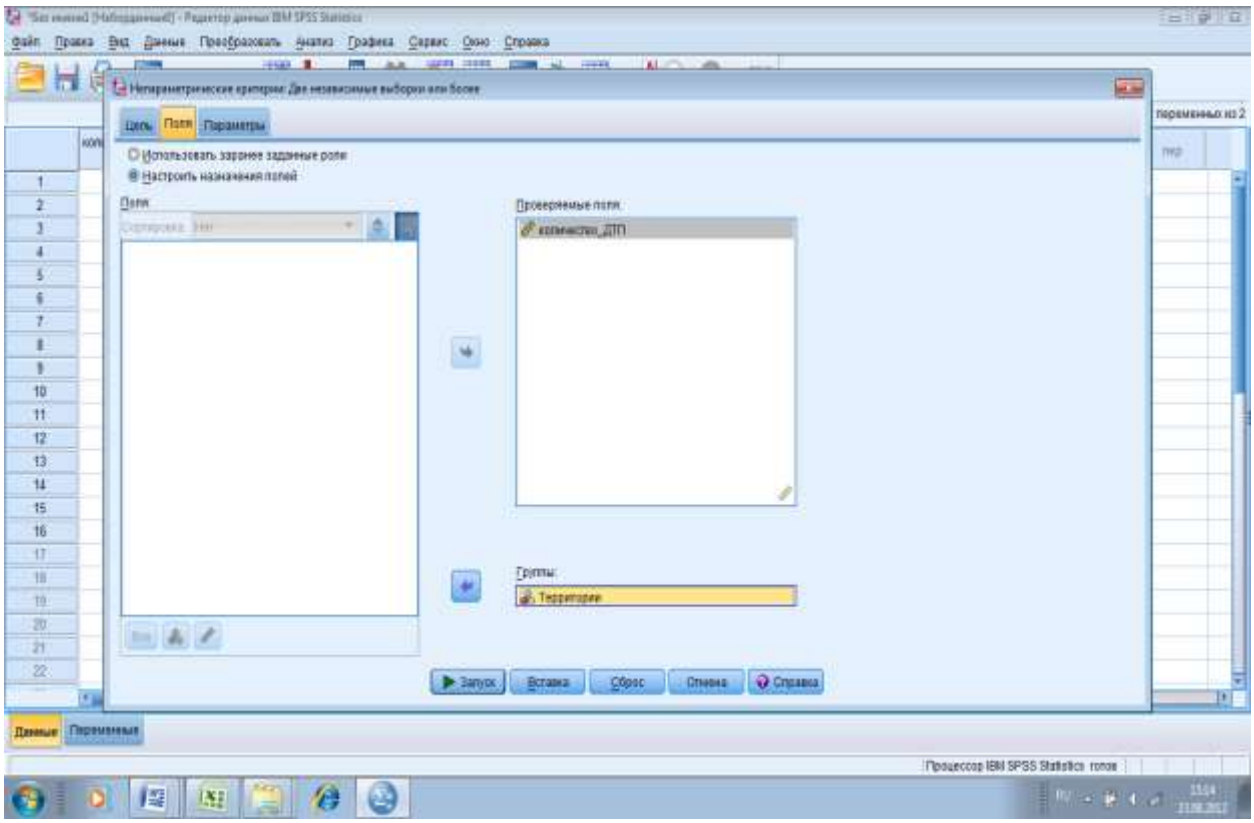


Рис. 41. Настройка анализа в программе IBM SPSS Statistics для обработки данных методом Краскела-Уоллиса

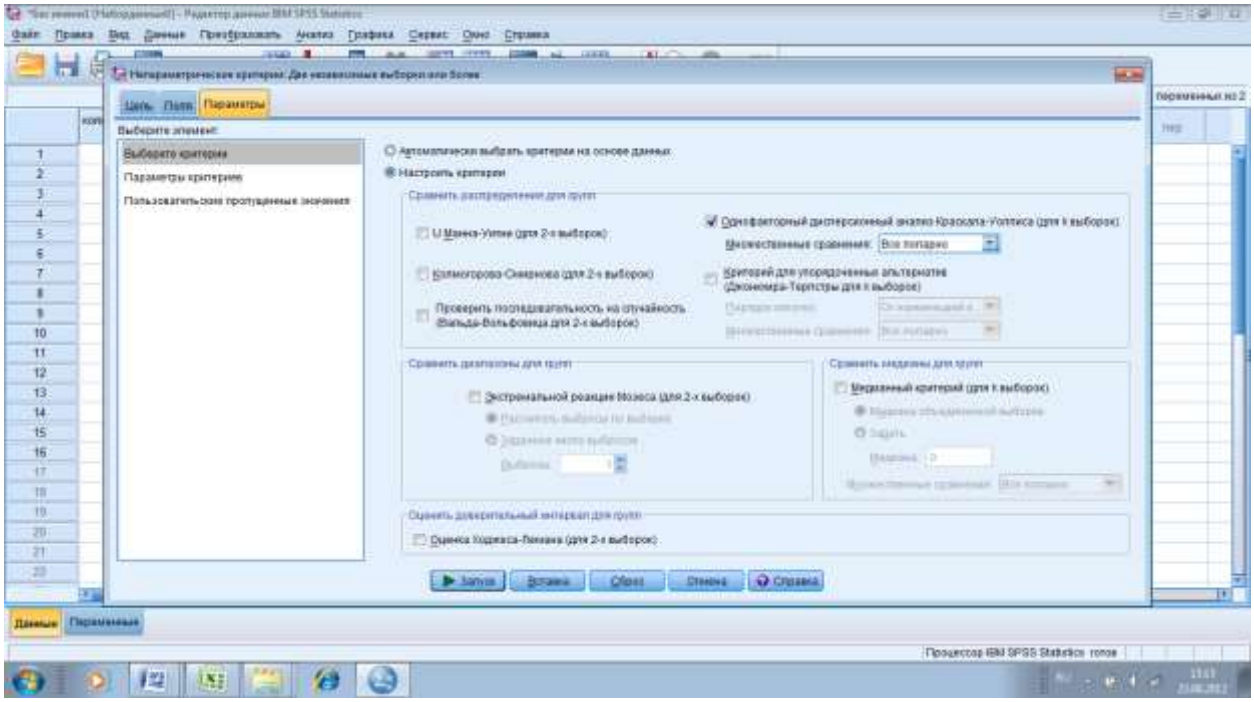


Рис. 42. Настройка анализа в программе IBM SPSS Statistics для обработки данных методом Краскела-Уоллиса

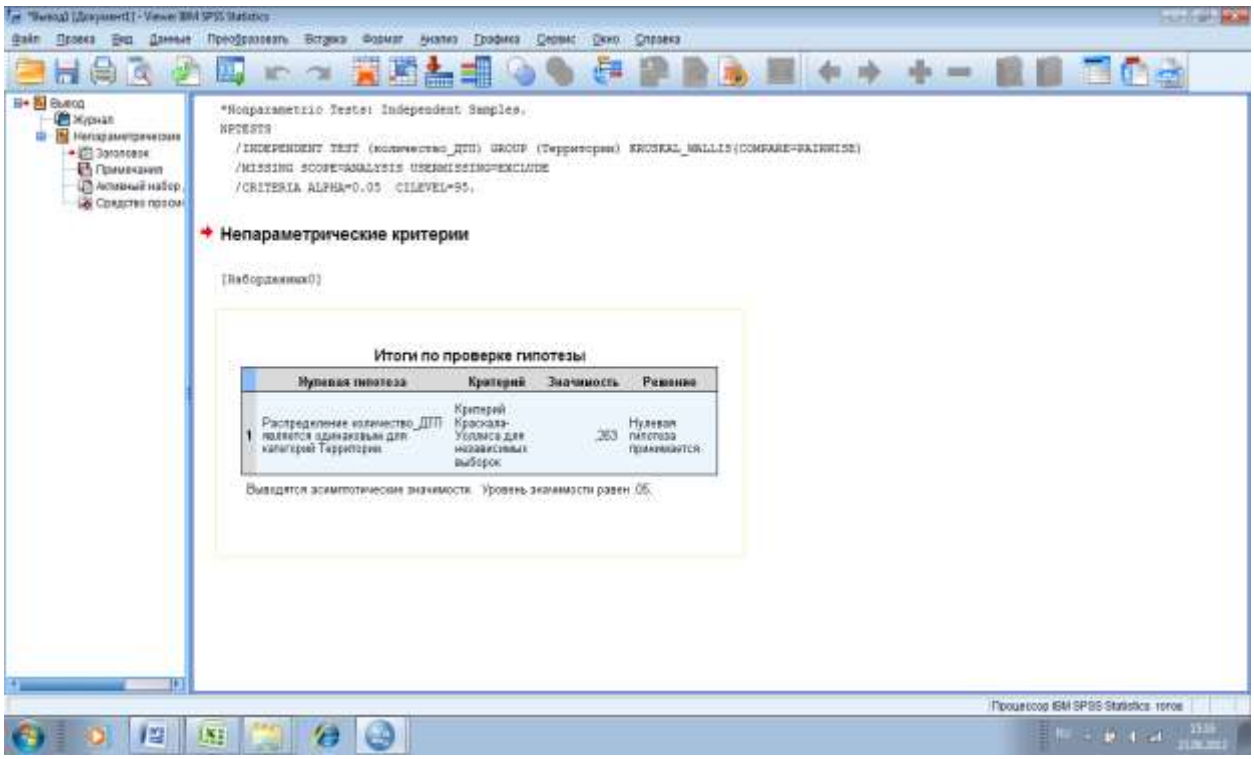


Рис. 43. Результаты анализа по методу Краскела-Уоллиса в программе IBM SPSS Statistics

НЕПАРАМЕТРИЧЕСКИЙ МЕТОД СРАВНЕНИЯ ТРЕХ ЗАВИСИМЫХ ГРУПП И БОЛЕЕ ПО ОДНОМУ ПРИЗНАКУ ДИСПЕРСИОННЫЙ АНАЛИЗ ПО ФРИДМАНУ

Критерий χ^2 Фридмана является непараметрическим аналогом однофакторного дисперсионного анализа для повторных измерений. Метод основан на ранжировании ряда повторных измерений для каждого объекта группы с последующим вычислением суммы рангов. При проведении расчетов вручную используется формула:

$$\chi^2 = \left[\frac{12}{Nk(k+1)} \times \sum_{i=1}^{k} R_i^2 \right] - 3N(k+1); df = k-1$$

N – число единиц наблюдения;
k – количество повторных измерений;
R_i – сумма рангов для повторных измерений *i;*

Чем больше различие между зависимыми выборками, тем больше рассчитанное значение χ^2 – Фридмана.

Рассмотрим пример. В группу наблюдения вошли 6 человек желающих отказаться от курения. У всех обследуемых была измерена ЖЕЛ в динамике: на момент включения в группу, через один и через два месяца после отказа от курения (табл. 16). Врачу необходимо оценить является ли статистически значимой положительная динамика увеличения показателя ЖЕЛ после отказа от курения.

Таблица 16. Результаты обследования

Группа наблюдения	На момент обследования		Через 1 мес.		Через 2 мес.	
	ЖЕЛ	ранг	ЖЕЛ	ранг	ЖЕЛ	ранг
1	2	1,5	2	1,5	3	3
2	2,1	1	3,1	2	4	3
3	2,4	2	2,1	1	3,5	3
4	2,5	1	3,5	2	4	3
5	2,3	1	3	2,5	3	2,5
6	2,6	1,5	2,6	1,5	4	3
Сумма рангов		8		10,5		17,5

Для каждого объекта ранжируются повторные измерения (по строке) и вычисляется сумма рангов для каждого повторного наблюдения. Далее вычисления проводятся по формуле:

$$\chi^2 = \left[\frac{12}{6 \times 3 \times (3+1)} \times (8^2 + 10{,}5^2 + 17{,}5^2) \right] - 3 \times 6 \times (3+1) = 79{,}76 - 72 = 7{,}8$$

Если k=3, N>9 или k>3, N>4 для оценки результатов пользуются таблицей для χ^2 ; df=k-1. Если k=3, N<9 или k=4, N<5 пользуются таблицей критических значений χ^2-Фридмена.

Рассчитанное значение χ^2 =7,8; при df =2 критическое значение χ^2 =6,0 (при p=0,05), следовательно H_o отклоняется и принимается альтернативная. При отказе от курения наблюдается увеличение ЖЕЛ (p<0,05).

Для выяснения в какой из сравниваемых групп признак выражен сильнее, необходимо попарно сравнить выборки по критерию Т-Вилкоксона. Во избежание проблемы множественных сравнений следует вводить поправку.

Для обработки данных по методу Фридмана в программе IBM SPSS Statistics внесите цифровые данные в ячейки таблицы (рис. 44).

Назовите анализируемые переменные, укажите их тип (числовая), выберете шкалу, в которой они измерены (количественная) как показано на рисунке 45.

В меню «Анализ» выберете «Непараметрические критерии» → «Устаревшие диалоговые окна» → «Для К связанных выборок» (рис. 46).

Перенесите переменные в список проверяемых переменных (рис. 47). Отметьте флажком используемый критерий (Фридмана).

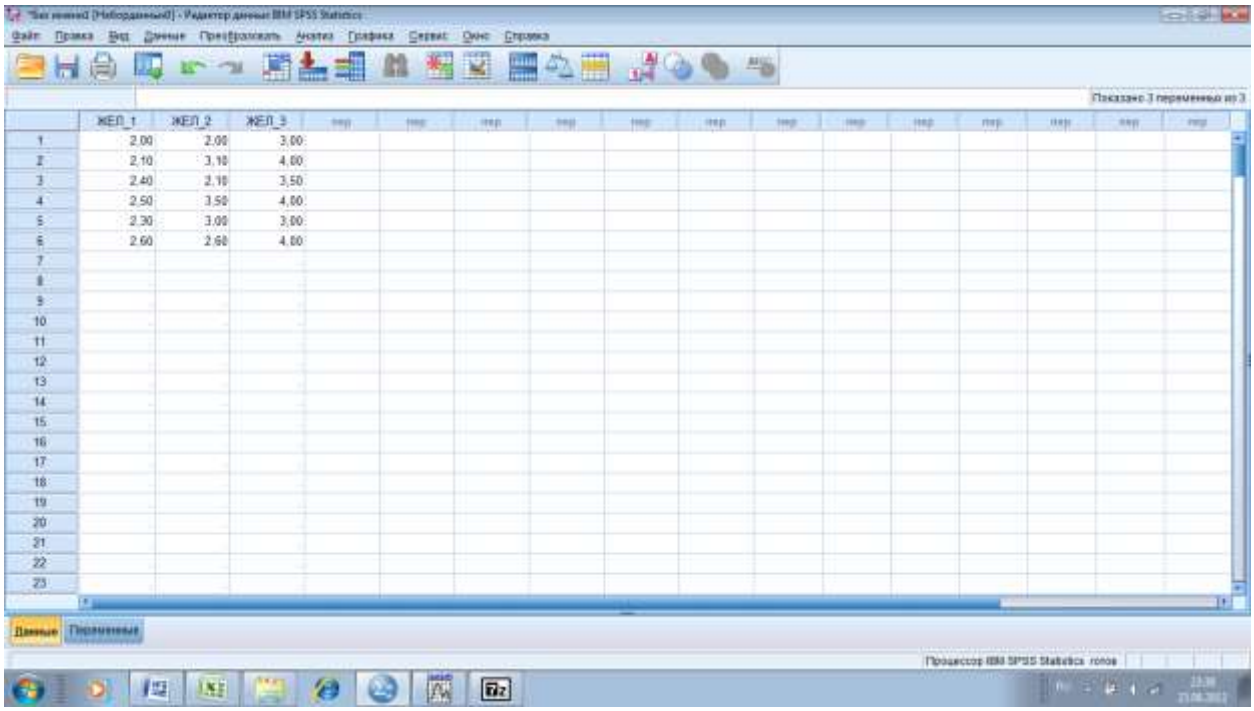


Рис. 44. Ввод данных для обработки по методу Фридмана в программе IBM SPSS Statistics

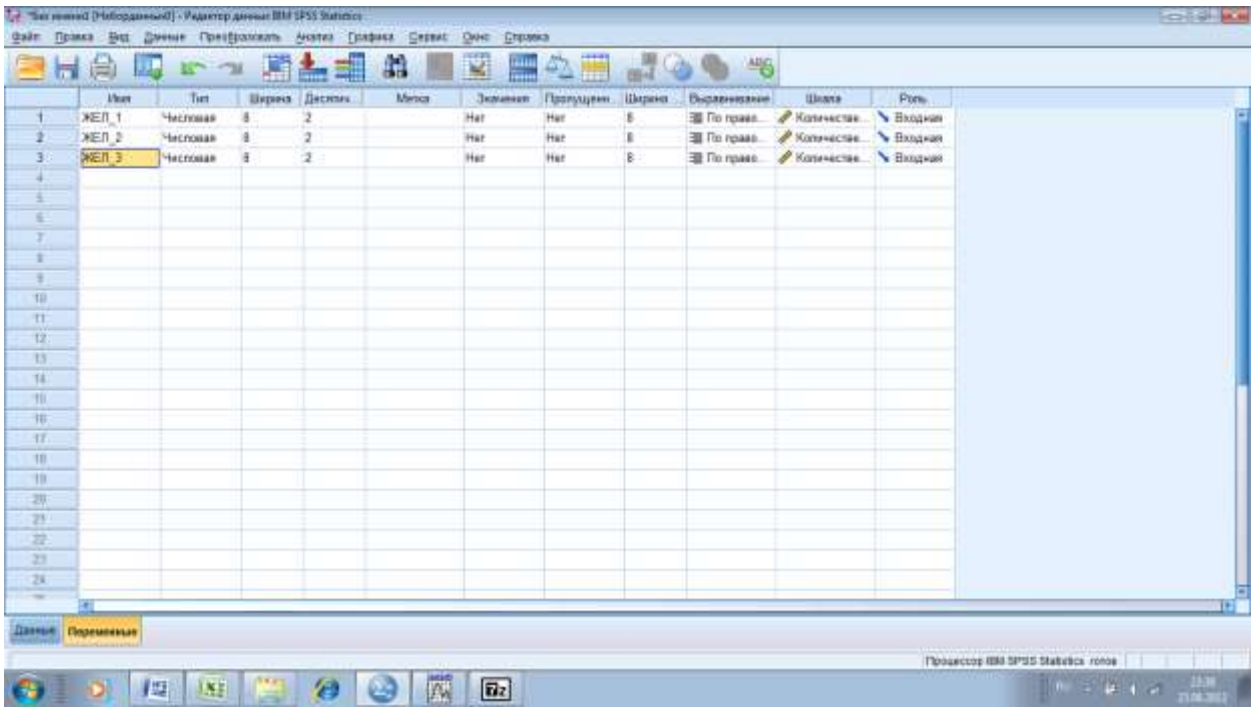


Рис. 45. Ввод данных для обработки по методу Фридмана в программе IBM SPSS Statistics

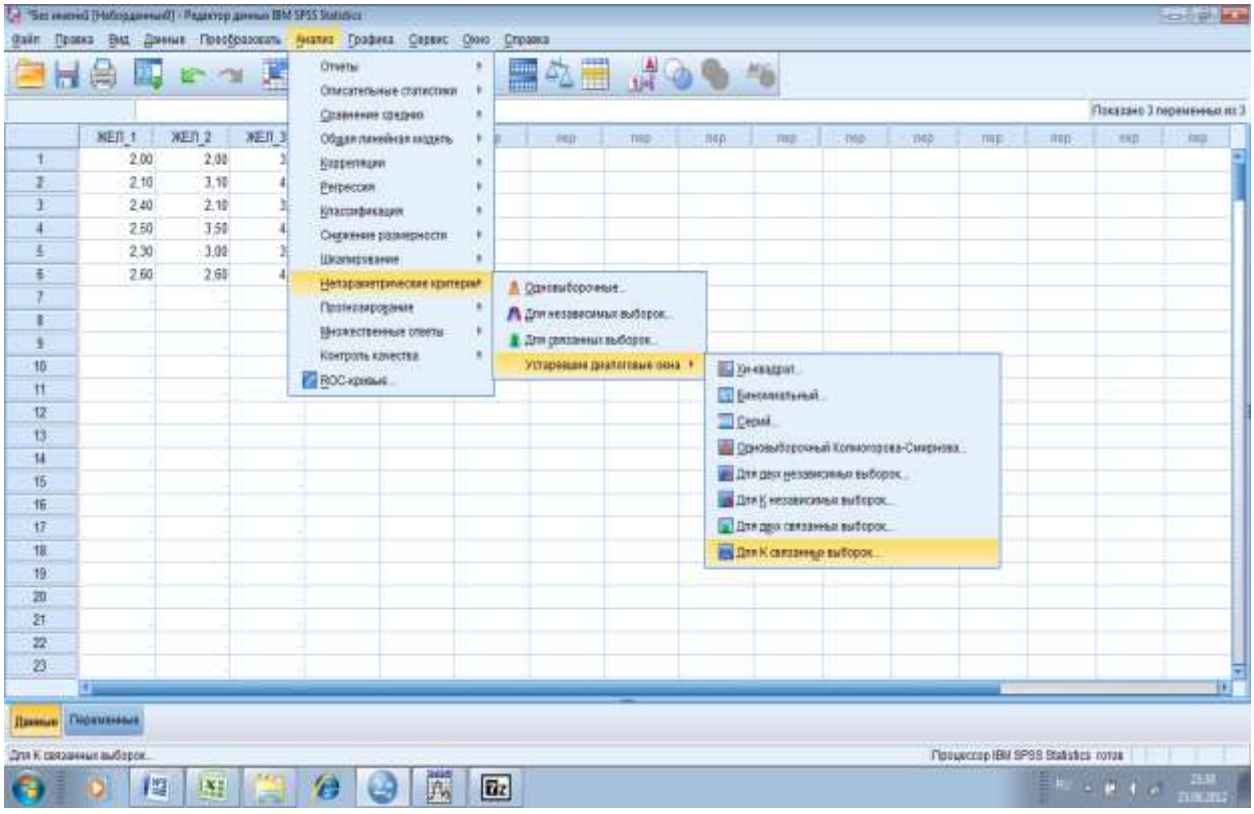


Рис. 46. Подготовка к проведению анализа по методу Фридмана в программе IBM SPSS Statistics

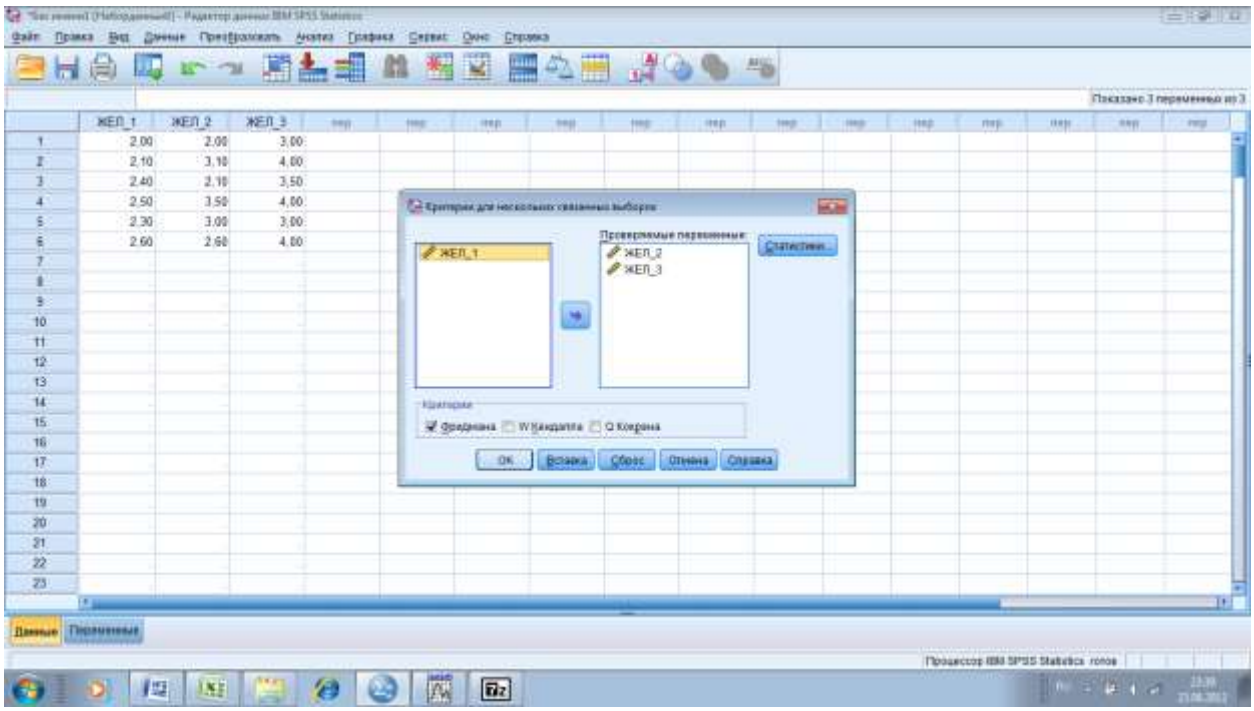


Рис. 47. Подготовка к проведению анализа по методу Фридмана в программе IBM SPSS Statistics

В окне результатов анализа указана статистика критерия Фридмана χ^2=9,238; p=0,01; следовательно, H_o отвергается (рис. 48). При отказе от курения наблюдается увеличение ЖЕЛ (p=0,01).

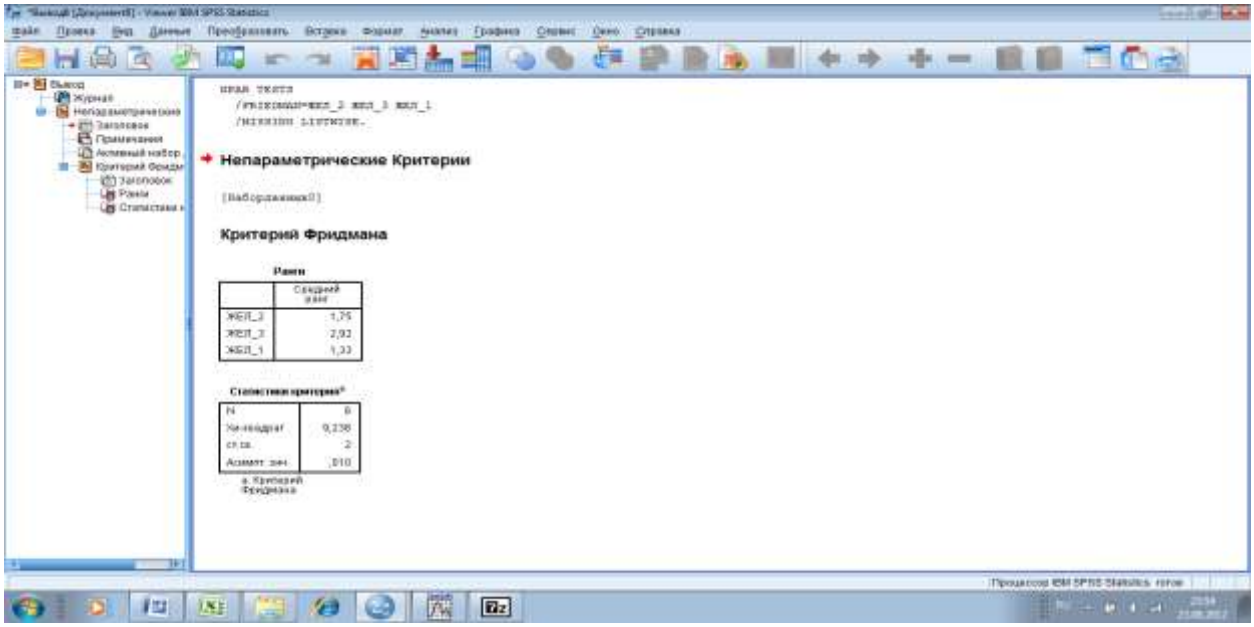


Рис. 48. Результаты анализа по методу Фридмана в программе IBM SPSS Statistics

Вопросы для подготовки к занятию

1. Что такое дисперсионный анализ?
2. Что такое корреляционный анализ?
3. Что такое регрессионный анализ?
4. Можно ли с помощью дисперсионного анализа построить математическую модель объекта?
5. Какие гипотезы проверяются в дисперсионном анализе?
6. Что такое статистика Фишера и критерий Фишера?
7. Основные предпосылки при решении задач с помощью дисперсионного анализа.
8. Основная идея однофакторного дисперсионного анализа.
9. Как проверяется гипотеза о равенстве нескольких дисперсий?
10. Основная идея двухфакторного дисперсионного анализа.
11. Что такое регрессия?
12. Как построить оценку регрессии?
13. Что такое дисперсионное отношение?

Тесты

1.Параметрические критерии основаны на:

а) оценке параметров распределения

б) типе распределения

в) выдвигаемых гипотезах

г) требуемой точности

2. Параметрические критерии применимы, если:

а) распределение отличается от нормального

б) требуются достаточно грубые оценки

в) варианты выборок различны

г) численные данные подчиняются нормальному распределению

3. Критерий Фишера основан на сравнении:

а) частот изучаемого признака в вариационном ряду

б) средних значений выборок

в) числа наблюдений выборок

г) выборочных дисперсий

4. Дисперсия альтернативного признака определяется:

а) как корень квадратный из произведения вероятностей признака, положенного в основу группировки на вероятность внешнего признака

б) как произведение вероятностей признака, положенного в основу группировки на вероятность внешнего признака

в) как произведение вероятностей наличия признака и его отсутствия

г) как произведение межгрупповой и средней из внутригрупповых дисперсий

д) как отношение межгрупповой дисперсии к средней из внутригрупповых дисперсий

5. Общая дисперсия - это…

а) произведение межгрупповой и средней из внутригрупповых дисперсий

б) отношение межгрупповой дисперсии к средней из внутригрупповых дисперсий

в) разность межгрупповой и средней из внутригрупповых дисперсий

г) сумма межгрупповой и средней из внутригрупповых дисперсий

д) корень квадратный из произведения межгрупповой и средней из внутригрупповых дисперсий

6. Межгрупповая дисперсия характеризует

а) случайную вариацию, полученную в результате действия случайных факторов

б) вариацию, полученную в результате действия внутренних факторов

в) вариацию, полученную в результате действия внешних факторов

г) вариацию, полученную в результате действия систематических и случайных факторов

д) постоянную вариацию, полученную в результате действия систематических факторов

7. Дисперсия определяется

а) как разность между максимальным и минимальным значениями признака

б) как средний коэффициент вариации ряда

в) +как средний квадрат отклонений индивидуальных значений признака от их средней величины

г) как корень квадратный из среднего квадрата отклонений индивидуальных значений признака от их средней величины

д) как среднеарифметическая из абсолютных значений отклонений отдельных вариантов от их средней

Ситуационные задачи

I. Методом дисперсионного анализа проверить нулевую гипотезу о влиянии фактора на качество объекта на основании пяти измерений для трех уровней фактора:

Номер измерения	Ф1	Ф2	Ф3
1	18	24	36
2	28	36	12
3	12	28	22
4	14	40	45

II. Используя анализ однофакторной модели, проверьте гипотезу о влиянии 3 разных подходов (1-отсутствие консультирования, 2-групповое консультирование, 3- индивидуальное консультирование) к профилактике ХНИЗ. Сформированы три группы наблюдения по 10 человек каждая. По окончании эксперимента оценивается состояние здоровья по оригинальному опроснику в баллах. Результаты представлены в таблице:

Баллы по тесту при трех разных подходах к профилактике ХНИЗ

Единицы наблюдения	1	2	3
1	28	39	41
2	33	52	49
3	42	53	56
4	47	54	62
5	48	56	63
6	50	58	64
7	50	59	65
8	51	63	72
9	60	64	77
10	71	77	87

Вопросы:

1) Влияют ли различные подходы к профилактике ХНИЗ на результат? Есть ли значимые различия между тремя выборками пациентов по результатам теста?

2) Есть ли статистически значимая тенденция возрастания показателей в порядке «отсутствие консультирования» - «групповое консультирование» - «индивидуальное консультирование»?

Тема 6. Дискриминантный анализ

Цель занятия: ознакомиться с методом дискриминантного анализа

Учебно-целевые задачи:

- усвоить область применения и цели проведения дискриминантного анализа

- научиться оценивать условия для применения дискриминантного анализа
- научиться интерпретировать результаты дискриминантного анализа

В результате освоения темы обучающиеся **должны знать:** какие вопросы решаются с помощью дискриминантного анализа; условия применения дискриминантного анализа; современные методы прогнозирования

В результате освоения темы обучающиеся **должны уметь:** интерпретировать результаты дискриминантного анализа.

ИНФОРМАЦИОННЫЙ МАТЕРИАЛ

Дискриминантный анализ представляет собой альтернативу множественного регрессионного анализа для случая, когда зависимая переменная представляет собой не количественную (номинативную) переменную. При этом дискриминантный анализ решает, по сути, те же задачи, что и множественный регрессионный анализ (МРА): предсказание значений «зависимой» переменной, в данном случае — категорий номинативного признака; определение того, какие «независимые» переменные лучше всего подходят для такого предсказания.

Дискриминантный анализ позволяет решить две группы проблем:

1. Интерпретировать различия между классами, т.е. ответить на вопросы: насколько хорошо можно отличить один класс от другого, используя данный набор переменных; какие из этих переменных наиболее существенны для различения классов. Сходную задачу решает дисперсионный анализ.

2. Классифицировать объекты, т.е. отнести каждый объект к одному из классов, исходя только из значений дискриминантных переменных. Задача классификации связана с получением по данным об «известных» объектах

дискриминантных функций «решающих правил», позволяющих по значениям дискриминантных переменных отнести с известной вероятностью каждый объект к одному из классов.

В решении задачи классификации дискриминантный анализ является не заменимым другими методами. Часто дискриминантный анализ называют еще «классификацией с обучением» или «распознаванием образов». В первом случае предполагают, что мы «учимся» классифицировать «неизвестные» объекты по дискриминантным переменным, используя данные об «известных» объектах. Во втором случае под «образом» объекта подразумевается совокупность измеренных для него значений дискриминантных переменных, И дискриминантный анализ позволяет в этом смысле распознать образ «нового» объекта путем отнесения его к известному классу объектов.

При использовании дискриминантного анализа рекомендуется:

- иметь объем выборки, в 10-20 раз превышающий число предикторов;

- количество объектов каждого класса должно превышать число предикторов (эмпирическое правило — не менее, чем в 5 раз);

- конечная модель не должна включать более 10 предикторов;

- необходимо проводить дополнительные исследования выбросов, которые могут негативно влиять на результат. Возможно проведение анализа с исключением выбросов.

- требуется принимать во внимание случаи, когда две переменные сильно коррелированы, хотя качество решающего правила при этом обычно не ухудшается.

В настоящее время на практике для прогнозирования бинарных переменных используется более совершенный статистический метод – логистическая регрессия. Этот метод позволяет не только ответить на

вопрос, какой именно исход наступит, скорее всего, но и определить вероятность, с которой наступит тот или иной исход.

Вопросы для подготовки к занятию

1. Назовите цель проведения и возможности использования результатов дискриминантного анализа.
2. Как выглядит математическое описание дискриминантной модели?
3. Какие требования предъявляются к переменным, участвующим в дискриминантном анализе, относительно типов шкал измерения переменных?
4. Какие задачи решаются в ходе проведения дискриминантного анализа?
5. Каким образом и с какой целью выявляется наличие дискриминирующих свойств у переменных, выбранных в качестве независимых (дискриминационных) переменных дискриминантной модели?

Тесты

1. Дискриминантный анализ решает ряд задач. Исключите лишнее.

а) предсказание значений независимых переменных

б) определение того, какие независимые переменные лучше всего подходят для предсказания

в) определение решающих правил, позволяющих отнести каждый объект к одному из известных классов

г) все ответы верны

2. В дискриминантном анализе для каждого из объектов имеются данные по количественным признакам, являющимся одинаковыми для этих объектов. Эти количественные признаки называются:

а) дискриминантные переменные

б) числовые значения

в) структура исходных данных

г) нет правильного варианта.

3. Канонические функции и дискриминантные переменные связывают:

а) структурные коэффициенты

б) стандартизированные коэффициенты

в) канонические коэффициенты

г) стандартизированные канонические коэффициенты.

4. Из геометрической интерпретации задач дискриминантного характера следует:

а) правило классификации объектов

б) правило систематизации объектов

в) правило интерпретации объектов

г) правило канонизации объектов.

5. Место типичных наблюдений для данных классов и их использование для описания различий между классами, называется:

а) точка отсчёта

б) исходная точка

в) центроид

г) нулевая точка.

6. Анализ канонических функций сопровождается получением важных статистических показателей качества классификации, основным из которых является:

а) λ Вилкса и χ^2 тест

б) Н – Колмогорова

в) λ Вилкса

г) χ^2 тест.

7. Мера классификации, являющаяся производной от расстояния, называется:

а) апостериорная вероятность

б) принадлежность объекта к классу

в) расстояние объекта до центроида

г) нет правильных вариантов.

8. Наиболее важным показателем в дискриминантном анализе является:

а) критерий Фишера

б) толерантность

в) статистика удаления

г) верны все варианты.

9. Для отсеивания малозначимых для дискриминатного анализа переменных применяется:

а) компьютерная программа SPSS

б) пошаговый дискриминантный анализ

в) анализ расстояний между классами

г) вычисление основных показателей качества.

10. Показателем информативности функции является:

а) собственное значение канонической функции

б) сумма всех собственных значений канонической функции

в) λ Вилкса

г) χ^2 тест.

Ситуационные задачи

1. Интерпретируйте результаты теста на равенство средних величин в группах, проводимого в ходе процедуры дискриминантного анализа, если значение *«Significance»* («Значимость») для определенной дискриминационной переменной составляет 0,637?

2. Объясните, что характеризует и с какой целью рассчитывается коэффициент корреляции между дискриминационными переменными? Как можно интерпретировать результаты таких расчетов, если значение коэффициента корреляции между двумя дискриминирующими переменными составляет 0,52?

3. Объясните, что характеризует и для чего рассчитывается коэффициент корреляции между расчетными значениями дискриминантной функции и реальной принадлежностью респондента к определенной группе? Как можно интерпретировать результаты, если значение этого коэффициента составляет 0,485?

4. В таблице 17 представлены нестандартизированные (канонические) коэффициенты дискриминантной функции, именно они используются для построения дискриминантной модели:

5. Запишите, какой вид будет иметь дискриминантная модель, построенная по результатам проведения дискриминантного анализа, в соответствии с данными, представленными в таблице. На основе построенной дискриминантной модели, сделайте прогноз наступления события, у отдельного индивида исходя из его возраста и дохода.

Таблица 17. Результаты обследования

Канонические коэффициенты дискриминантной функции

Canonical Discriminant Function Coefficients*

	Function
	1
Возраст	,076
Доход	,062
(Constant)	-4,200

* Нестандартизированные коэффициенты.

Тема 7. Факторный и кластерный анализ

Цель занятия: ознакомиться с методами факторного и кластерного анализов

Учебно-целевые задачи:

- ознакомиться с методикой факторного и кластерного анализов
- понять цель и область применения факторного и кластерного анализов
- научиться интерпретировать результаты факторного и кластерного анализов

В результате освоения темы обучающиеся **должны знать:** цель и область применения факторного анализа, его виды и методы, типы факторного анализа, условия его применения; методы детерминированного и стохастического факторного анализа; цель и область применения кластерного анализа, его виды и методы;

- В результате освоения темы обучающиеся **должны уметь:** интерпретировать результаты факторного и кластерного анализов.

ИНФОРМАЦИОННЫЙ МАТЕРИАЛ

Факторный анализ в учебной литературе трактуется как раздел многомерного статистического анализа, объединяющий методы оценки

размерности множества наблюдаемых переменных посредством исследования структуры ковариационных или корреляционных матриц.

Свою историю факторный анализ начинает в психометрике и в настоящее время широко используется не только в психологии, но и в нейрофизиологии, социологии, политологии, в экономике, статистике и других науках. Основные идеи факторного анализа были заложены английским психологом и антропологом *Ф. Гальтоном*. Разработкой и внедрением факторного анализа в психологии занимались такие ученые как: *Ч.Спирмен, Л.Терстоун и Р.Кеттел*. Математический факторный анализ разрабатывался *Хотеллингом, Харманом, Кайзером, Терстоуном, Такером* и другими учеными.

Данный вид анализа позволяет исследователю решить две основные задачи: описать предмет измерения компактно и в то же время всесторонне. С помощью факторного анализа возможно выявление факторов, отвечающих за наличие линейных статистических связей корреляций между наблюдаемыми переменными.

ЦЕЛИ ФАКТОРНОГО АНАЛИЗА

К примеру, анализируя оценки, полученные по нескольким шкалам, исследователь отмечает, что они сходны между собой и имеют высокий коэффициент корреляции, в этом случае он может предположить, что существует некоторая *латентная переменная*, с помощью которой можно объяснить наблюдаемое сходство полученных оценок. Такую латентную переменную называют фактором, который влияет на многочисленные показатели других переменных, что приводит к возможности и необходимости отметить его как наиболее общий, более высокого порядка.

Таким образом, можно выделить две **цели факторного анализа**:

- определение взаимосвязей между переменными, их классификация, т. е. «объективная R-классификация»;

- сокращение числа переменных.

Для выявления наиболее значимых факторов и, как следствие, факторной структуры, наиболее оправданно применять *метод главных компонентов*. Суть данного метода состоит в замене коррелированных компонентов некоррелированными факторами. Другой важной характеристикой метода является возможность ограничиться наиболее информативными главными компонентами и исключить остальные из анализа, что упрощает интерпретацию результатов. Достоинство данного метода также в том, что он – единственный математически обоснованный метод факторного анализа.

Факторный анализ – методика комплексного и системного изучения и измерения воздействия факторов на величину результативного показателя

ТИПЫ ФАКТОРНОГО АНАЛИЗА

Существуют следующие типы факторного анализа:

1) Детерминированный (функциональный) – результативный показатель представлен в виде произведения, частного или алгебраической суммы факторов.

2) Стохастический (корреляционный) – связь между результативным и факторными показателями является неполной или вероятностной.

3) Прямой (дедуктивный) – от общего к частному.

4) Обратный (индуктивный) – от частного к общему.

5) Одноступенчатый и многоступенчатый.

6) Статический и динамический.

7) Ретроспективный и перспективный.

Также факторный анализ может быть *разведочным* – он осуществляется при исследовании скрытой факторной структуры без предположения о числе факторов и их нагрузках и *конфирматорным*, предназначенным для проверки гипотез о числе факторов и их нагрузках.

Практическое выполнение факторного анализа начинается с проверки его условий.

Обязательные условия факторного анализа:

- Все признаки должны быть количественными;
- Число признаков должно быть в два раза больше числа переменных;
- Выборка должна быть однородна;
- Исходные переменные должны быть распределены симметрично;
- Факторный анализ осуществляется по коррелирующим переменным.

При анализе в один фактор объединяются сильно коррелирующие между собой переменные, как следствие происходит перераспределение дисперсии между компонентами и получается максимально простая и наглядная структура факторов. После объединения коррелированность компонент внутри каждого фактора между собой будет выше, чем их коррелированность с компонентами из других факторов. Эта процедура также позволяет выделить латентные переменные, что бывает особенно важно при анализе социальных представлений и ценностей.

ЭТАПЫ ФАКТОРНОГО АНАЛИЗА

Как правило, факторный анализ проводится в несколько этапов.

Этапы факторного анализа:

1 этап. Отбор факторов.

2 этап. Классификация и систематизация факторов.

3 этап. Моделирование взаимосвязей между результативным и факторными показателями.

4 этап. Расчет влияния факторов и оценка роли каждого из них в изменении величины результативного показателя.

5 этап. Практическое использование факторной модели (подсчет резервов прироста результативного показателя).

По характеру взаимосвязи между показателями различают методы детерминированного и стохастического факторного анализа.

Детерминированный факторный анализ представляет собой методику исследования влияния факторов, связь которых с результативным показателем носит функциональный характер, т. е. когда результативный показатель факторной модели представлен в виде произведения, частного или алгебраической суммы факторов.

Методы детерминированного факторного анализа: *Метод цепных подстановок; Метод абсолютных разниц; Метод относительных разниц; Интегральный метод; Метод логарифмирования.*

Данный вид факторного анализа наиболее распространен, поскольку, будучи достаточно простым в применении (по сравнению со стохастическим анализом), позволяет осознать логику действия основных факторов развития предприятия, количественно оценить их влияние, понять, какие факторы, и в какой пропорции возможно и целесообразно изменить для повышения эффективности производства.

Стохастический анализ представляет собой методику исследования факторов, связь которых с результативным показателем в отличие от функциональной является неполной, вероятностной (корреляционной). Если при функциональной (полной) зависимости с изменением аргумента всегда происходит соответствующее изменение функции, то при корреляционной связи изменение аргумента может дать несколько значений прироста функции в зависимости от сочетания других факторов, определяющих данный показатель.

Методы стохастического факторного анализа: *Способ парной корреляции; Множественный корреляционный анализ; Матричные модели;*

Математическое программирование; Метод исследования операций; Теория игр.

Необходимо также различать статический и динамический факторный анализ. Первый вид применяется при изучении влияния факторов на результативные показатели на соответствующую дату. Другой вид представляет собой методику исследования причинно-следственных связей в динамике.

И, наконец, факторный анализ может быть ретроспективным, который изучает причины прироста результативных показателей за прошлые периоды, и перспективным, который исследует поведение факторов и результативных показателей в перспективе.

КЛАСТЕРНЫЙ АНАЛИЗ

Кластерный анализ — метод *классификации объектов* по заданным признакам. Задача кластерного анализа состоит в формировании групп:

- однородных внутри (условие внутренней гомогенности);
- четко отличных друг от друга (условие внешней гетерогенности).

В современной литературе есть много определений кластерного анализа, но все они понимают под собой совокупность математических методов, предназначенные для формирования относительно «отдаленных» друг от друга групп «близких» между собой объектов по информации о расстояниях или связях (мерах близости) между ними одновременно по всем наиболее существенным признакам. Кластерный анализ применяется для решения широкого спектра задач, но чаще всего речь идет именно о задаче сегментации. Все исследования, посвященные проблеме сегментации, безотносительно того, какой используется метод, имеют целью идентифицировать устойчивы группы, каждая из которых объединяет в себя объекты похожими характеристиками. В отличие от большинства других

методов многомерного анализа, кластерный анализ параллельно развивался в нескольких дисциплинах (психология, биология, экономика...), поэтому у большинства методов существует по 2 и более названий, что существенно затрудняет взаимопонимание исследователей, в особенности, если речь идет о разных отраслях знания. Другая проблема связана с обилием вариантов при выборе метрики и метода кластеризации, а также согласования между ними. Выделяют две группы методов кластерного анализа: иерархические и неиерархические.

Основными методами иерархического кластерного анализа являются метод ближнего соседа, метод полной связи, метод средней связи и метод Варда.

Существуют также центроидные методы и методы, использующие медиану, но их применение может привести к некоторым весьма нежелательным последствиям.

Неиерархических методов больше, хотя работают они на одних и тех же принципах. По сути, они представляют собой итеративные методы дробления исходной совокупности. В процессе деления формируются новые кластеры, и так до тех пор, пока не будет выполнено правило остановки.

Между собой методы различаются выбором начальной точки, правилом формирования новых кластеров и правилом остановки. Чаще всего используется алгоритм К-средних. Он подразумевает, что аналитик заранее фиксирует количество кластеров в результирующем разбиении. Выбирая между иерархическими и неиерархическими методами, следует обратить внимание на следующие моменты. Неиерархические методы обнаруживают более высокую устойчивость по отношению к выбросам, неверному выбору метрики, включению незначимых переменных в базу для кластеризации и пр. Но платой за это является слово «априори». Исследователь должен заранее фиксировать результирующее количество кластеров, правило остановки и, если на то есть основания, начальный центр кластера. Последний момент

существенно отражается на эффективности работы алгоритма. Если нет оснований искусственно задать это условие, рекомендуется использовать иерархические методы. Нужно учесть также еще один момент, существенный для обеих групп алгоритмов: не всегда правильным решением является кластеризация всех наблюдений. Возможно, более аккуратным будет сначала очистить выборку от выбросов, а затем продолжить анализ. Можно также не задавать очень высокий критерий остановки (можно делать остановку, к примеру, когда кластеризовано более 90 % наблюдений). Из информации, приведенной выше, явно прослеживается, что от аналитика в процессе применения кластерного анализа ожидается решение ряда задач. Их можно сгруппировать следующим образом:

1. Изменение исходных данных:

• выбор метрики;

• выбор метода стандартизации;

• как работать с зависимыми выборками.

2. Принятие решений:

• сколько кластеров необходимо сформировать;

• какой метод кластеризации следует использовать;

• следует ли использовать все наблюдения или необходимо исключить некоторые под выборки.

3. Анализ полученных результатов

• насколько полученное разбиение отличается от случайного;

• является ли оно надежным и стабильным на подвыборках;

• какова взаимосвязь между результатами кластеризации и переменными, не участвовавшими в процессе кластеризации;

• можно ли проинтерпретировать полученные результаты.

Вопросы для подготовки к занятию

1. Назовите цель проведения и возможности использования результатов факторного анализа.

2. Что представляет собой результирующая факторная модель? Какие преобразования происходят с исходным массивом данных в результате проведения факторного анализа?

3. Какие задачи решаются в ходе проведения факторного анализа?

4. В чем заключается сложность факторного анализа и какие проблемы неизбежно возникают в ходе его выполнения?

5. С какой целью в ходе выполнения факторного анализа производятся тесты *«KMO»* и «*Bartlett*», как следует интерпретировать результаты, если значение теста *«KMO»* составляет 0,742, а значение величины *«Significance»* («Значимость») по результатам теста «*Bartlett*» — 0,02?

6. Какова цель проведения и возможности использования результатов кластерного анализа?

7. Какие требования предъявляются к переменным, участвующим в проведении кластерного анализа, относительно типов шкал измерения переменных?

8. Почему и в каких случаях при проведении кластерного анализа необходимо преобразование структуры исходного массива данных?

9. Чем отличается иерархический кластерный анализ от других видов кластерного анализа?

10. В чем состоит отличие между дивизионным и агломеративным алгоритмом иерархического кластерного анализа?

Тесты

1. Кластерный анализ – это

а) решает задачу построения классификации, то есть разделения исходного множества объектов на группы, класс, кластеры

б) процедура упорядочивания объектов в сравнительно однородные классы на основе попарного сравнения этих объектов по предварительно определённым и измеренным критериям

в) множество простых вычислительных процедур, используемых для классификации объектов.

2. Первые исследования с использованием кластерного анализа появились после публикации книги «Начала численной таксономии» в 1963 году, авторами которой являются:

а) Р. Сокэл и П. Снит

б) Р. Фокэл и П. Смит

в) Р. Токэл и П. Свифт

г) Р. Мокэл и П. Скитт.

3. Существует множество вариантов кластерного анализа, но наиболее широко используются методы, объединённые общим названием:

а) иерархический кластерный анализ

б) пошаговый кластерный анализ

в) последовательный кластерный анализ

г) пропорциональный кластерный анализ.

4. Результат работы кластерного анализа представляется графически в виде:

а) гистограммы

б) дендрограммы

в) центроида

г) диаграммы.

5. Непосредственными данными для применения кластерного анализа являются:

а) матрица различий между всеми парами объектов

б) дендрограмма

в) испытуемые, объекты, которые оцениваются испытуемыми

г) признаки, измеренные на выборке испытуемых.

6. В последовательности кластерного анализа последним этапом является:

а) проверка достоверности разбиения на классы

б) проверка устойчивости группировки

в) проверка значимости разбиения

г) все варианты верны.

7. К методам кластерного анализа относится:

а) метод одиночной связи

б) метод полной связи

в) метод средней связи

г) все варианты верны.

8. Для предварительного определения числа классов пользуются:

а) содержательными соображениями исследователя

б) таблицей последовательности агломерации

в) исходными данными

г) не существует формального критерия, позволяющего выделить оптимальное число классов.

9. В методе полной связи кластерного анализа наблюдается тенденция к:

а) выделению большего числа компактных кластеров

б) выделению самого далёкого элемента, который находится ближе к новому объекту

в) выделению меньшего числа компактных кластеров

г) все варианты верны.

10. Как по-другому называется метод полной связи:

а) метод дальнего соседа

б) метод ближнего соседа

в) метод межгрупповой связи

г) метод внутригрупповой связи.

11. Факторный анализ – это

а) комплекс аналитических методов, позволяющий выявить скрытые признаки, а также причины их возникновения и внутренние закономерности их взаимосвязи;

б) метод, направленный на преобразование исходного набора признаков в более простую и содержательную форму;

в) раздел многомерного статистического анализа, объединяющий методы оценки размеренности множества наблюдаемых переменных посредством исследования структуры корреляционный матриц.

12. Фактор – это

а) причина, движущая сила какого – либо психического изменения или явления;

б) связь между действием и его психическим следствием;

в) скрытая причина согласованной изменчивости наблюдаемых переменных.

13. Основателем факторного анализа является:

а) Л. Терстоун;

б) Ф. Гальтон;

в) К. Пирсон;

г) Ч. Спирмен.

14. Исходной информацией для проведения факторного анализа является:

а) анализ корреляций;

б) корреляционная матрица;

в) матрица интеркорреляционных показателей;

г) нет правильного ответа.

15. Кто в 1931 году выдвинул идею единого генерального фактора G, лежащего в основе успешности выполнения любых тестов, связанных с измерением интеллектуальных свойств:

а) Л. Терстоун;

б) Ф. Гальтон;

в) К. Пирсон;

г) Ч. Спирмен.

16. Кто в 1927 году разработал математически обоснованную методику факторного анализа, теоретической основой которого послужила однофакторная теория.

а) Л. Терстоун;

б) Ч. Спирмен.

в) Ф. Гальтон;

г) К. Пирсон;

17. Кто в 1931 году разработал мультифакторный анализ оценки многих коррелирующих между собой и относительно независимых факторов, объясняющий мультифакторную концепцию интеллекта:

а) Л. Терстоун;

б) Ч. Спирмен.

в) Ф. Гальтон;

г) К. Пирсон.

18. Основное назначение факторного анализа заключается:

а) анализ корреляций множества признаков;

б) уменьшение размерности исходных данных;

в) переход от множества исходных переменных к существенно меньшему числу новых переменных – факторов;

г) нет правильного ответа.

19. Аналоги коэффициентов корреляции в факторном анализе называются:

а) интерпретация факторов;

б) факторные нагрузки;

в) факторные выбросы;

г) абсолютная величина факторной нагрузки.

20. К основным задачам факторного анализа относятся:

а) исследование структуры взаимосвязей переменных;

б) идентификация факторов как скрытых латентных переменных;

в) вычисление значений факторов для испытуемых как новых, интегральных переменных;

г) все ответы верны.

21. Выберите правильный вариант.

Метод, который преобразует набор коррелирующих исходных переменных в другой набор – некоррелирующих переменных, называется:

а) модель главных компонент;

б) анализ главных компонент;

в) факторный анализ главных компонент;

г) метод главных компонент.

22. Каждый элемент корреляционной матрицы называется:

а) собственное значение;

б) информативность компонента;

в) компонентная нагрузка;

г) коэффициент компонента.

23. Часть дисперсии переменной, объясняемая главными компонентами (факторами) называется:

а) общность;

б) факторная структура;

в) факторные нагрузки;

г) нет правильного ответа.

24. Основной результат применения факторного анализа называется:

а) общность;

б) факторная структура;

в) факторные нагрузки;

г) нет правильного ответа.

25. Критерий, которой определяет число факторов как равное числу компонент, собственные значения которых больше 1, называется:

а) критерий Фишера;

б) критерий Манна–Уитни;

в) критерий Кайзера;

г) критерий Колмогорова–Смирнова.

26. Способ определения числа факторов, который требует построения графика собственных значений, называется:

а) критерий отсеивания Кеттелла;

б) критерий Кайзера;

в) критерий Колмогорова–Смирнова;

Ситуационные задачи

1. После запуска процедуры факторного анализа в программе SPSS, получены следующие результаты тестов *«KMO»* и «*Bartlett»: «KMO»* = 0,668. Значимость теста «*Bartlett»(Sig.)* составляет 0,0001. Исходя из результатов тестов *«KMO»* и «*Bartlett»* сделайте вывод о пригодности исходных данных для проведения факторного анализа. Назовите основные цели проведения этих тестов.

2. Объясните основное назначение критерия «*Ellbow*» при проведении кластерного анализа. Какой вывод можно сделать по результатам представленного на рисунке теста?

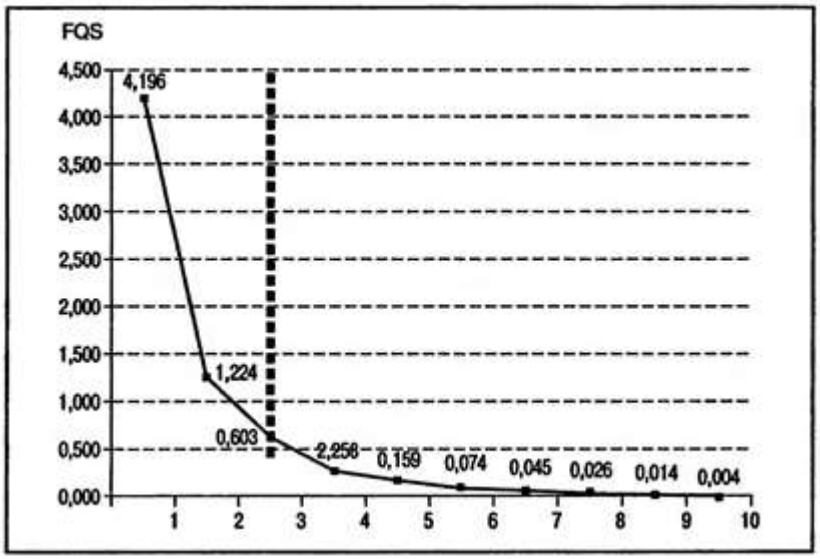


Рис. 49. Критерий «*Ellbow*»

КРИТЕРИИ ОЦЕНКИ

оценка «отлично» Дан полный, развернутый ответ на поставленный вопрос, показана совокупность осознанных знаний по заданному вопросу, доказательно раскрыты основные положения темы; в ответе прослеживается четкая структура, логическая последовательность, отражающая сущность раскрываемых понятий, теорий, явлений. Знания об объекте демонстрируются на фоне понимания его в системе данной науки и междисциплинарных связей. Ответ изложен литературным языком в терминах науки. Могут быть допущены недочеты в определении понятий, исправленные обучающимся самостоятельно в процессе ответа;

оценка «хорошо» Дан полный, развернутый ответ на поставленный вопрос, показано умение выделить существенные и несущественные признаки, причинно-следственные связи. Ответ четко структурирован, логичен, изложен литературным языком в терминах науки. Могут быть допущены недочеты или незначительные ошибки, исправленные обучающимся с помощью преподавателя;

оценка «удовлетворительно» Дан недостаточно полный и недостаточно развернутый ответ. Логика и последовательность изложения имеют нарушения. Допущены ошибки в раскрытии понятий, употреблении терминов. Студент не способен самостоятельно выделить существенные и несущественные признаки и причинно-следственные связи. Студент может конкретизировать обобщенные знания, доказав на примерах их основные положения только с помощью преподавателя. Речевое оформление требует поправок, коррекции;

оценка «неудовлетворительно» Дан неполный ответ, представляющий собой разрозненные знания по теме вопроса с существенными ошибками в определениях. Присутствуют фрагментарность, нелогичность изложения. Студент не осознает связь данного понятия,

теории, явления с другими объектами дисциплины. Отсутствуют выводы, конкретизация и доказательность изложения. Речь неграмотна. Дополнительные и уточняющие вопросы преподавателя не приводят к коррекции ответа обучающегося не только на поставленный вопрос, но и на другие вопросы дисциплины.

ПРИЛОЖЕНИЕ

Таблица критических значений коэффициентов корреляции Пирсона

Для уровня значимости α=0,05; α=0,01

Вероятность $p = \alpha$

где k – число степеней свободы

α / ***k = n - 2***	***0,05***	***0,01***	***α*** / ***k = n - 2***	***0,05***	***0,01***
5	0,75	0,87	**27**	0,37	0,47
6	0,71	0,83	**28**	0,36	0,046
7	0,67	0,80	**29**	0,36	0,046
8	0,63	0,77	**30**	0,35	0,045
9	0,60	0,74	**35**	0,33	0,42
10	0,58	0,71	**40**	0,30	0,39
11	0,55	0,68	**45**	0,29	0,37
12	0,53	0,66	**50**	0,27	0,35
13	0,51	0,64	**60**	0,25	0,33
14	0,50	0,62	**70**	0,23	0,30
15	0,48	0,61	**80**	0,22	0,28
16	0,47	0,59	**90**	0,21	0,27
17	0,46	0,58	**100**	0,20	0,25
18	0,44	0,56	**125**	0,17	0,23
19	0,43	0,55	**150**	0,16	0,21
20	0,42	0,54	**200**	0,14	0,18
21	0,41	0,53	**300**	0,11	0,15
22	0,40	0,52	**400**	0,10	0,13
23	0,40	0,51	**500**	0,09	0,12
24	0,39	0,50	**700**	0,07	0,10

25	0,38	0,49	**900**	0,06	0,09
26	0,37	0,48	**1000**	0,06	0,09

Таблица критических значений коэффициентов корреляции рангов Спирмена

Для уровня значимости α=0,05; α=0,01

Вероятность $p = \alpha$

α / *n*	**0,05**	**0,01**	*α* / *n*	**0,05**	**0,01**	*α* / *n*	**0,05**	**0,01**
5	0,94	-	**17**	0,48	0,62	**29**	0,37	0,48
6	0,85	-	**18**	0,47	0,60	**30**	0,36	0,47
7	0,78	0,94	**19**	0,46	0,58	**31**	0,36	0,46
8	0,72	0,88	**20**	0,45	0,57	**32**	0,36	0,45
9	0,68	0,83	**21**	0,44	0,56	**33**	0,34	0,45
10	0,64	0,79	**22**	0,43	0,54	**34**	0,34	0,44
11	0,61	0,76	**23**	0,42	0,53	**35**	0,33	0,43
12	0,58	0,73	**24**	0,41	0,52	**36**	0,33	0,43
13	0,56	0,70	**25**	0,39	0,51	**37**	0,33	0,43
14	0,54	0,68	**26**	0,39	0,50	**38**	0,32	0,41
15	0,52	0,66	**27**	0,38	0,49	**39**	0,32	0,41
16	0,50	0,64	**28**	0,38	0,48	**40**	0,31	0,40

Таблица значений критерия Стьюдента (*t*-критерия)

Значения критерия Стьюдента (t-критерия) для уровня значимости α=0,005; α=0,01; α=0,025; α=0,05; α=0,10

Вероятность $p = \alpha$

где k – число степеней свободы

α / k	односторонняя область				
	0,005	**0,01**	**0,025**	**0,05**	**0,10**
	двусторонняя область				
	0,01	**0,02**	**0,05**	**0,10**	**0,20**
1	63,66	31,82	12,71	6,31	3,08
2	9,93	6,97	4,30	2,92	1,89
3	5,84	4,54	3,18	2,35	1,64
4	4,60	3,75	2,78	2,13	1,53
5	4,03	3,37	2,57	2,02	1,48
6	3,71	3,14	2,45	1,94	1,44
7	3,50	3,00	2,37	1,90	1,42
8	3,36	2,90	2,31	1,86	1,40
9	3,25	2,82	2,26	1,83	1,38
10	3,17	2,76	2,23	1,81	1,37
11	3,11	2,72	2,20	1,80	1,36
12	3,06	2,68	2,18	1,78	1,36
13	3,01	2,65	2,16	1,77	1,35
14	2,98	2,62	2,15	1,76	1,35
15	2,95	2,60	2,13	1,75	1,34
16	2,92	2,58	2,12	1,75	1,34
17	2,90	2,57	2,11	1,74	1,33

18	2,88	2,55	2,10	1,73	1,33
19	2,86	2,54	2,09	1,73	1,33
20	2,85	2,53	2,09	1,73	1,33
21	2,83	2,52	2,08	1,72	1,32
22	2,82	2,51	2,07	1,72	1,32
23	2,81	2,50	2,07	1,71	1,32
24	2,80	2,49	2,06	1,71	1,32
25	2,79	2,49	2,06	1,71	1,32
26	2,78	2,48	2,06	1,71	1,32
27	2,77	2,47	2,05	1,70	1,31
28	2,76	2,47	2,05	1,70	1,31
29	2,76	2,46	2,05	1,70	1,31
30	2,75	2,46	2,04	1,70	1,31
40	2,70	2,42	2,02	1,68	1,30
60	2,66	2,39	2,00	1,67	1,30

Когда имеются основания для применения одностороннего теста, его следует предпочесть двустороннему. Односторонний критерий имеет меньшую вероятность ошибки второго рода, чем соответствующий двусторонний, при той же вероятности ошибочного отклонения нулевой гипотезы. Поэтому, предпочтительнее применение одностороннего критерия.

Таблица значений критерия Фишера (*F*-критерия)

Значения критерия Фишера (*F*-критерия) для уровня значимости $\alpha = 0{,}05$

k_1 - число степеней свободы большей дисперсии, k_2 - число степеней свободы меньшей дисперсии

k_2 \ k_1	1	2	3	4	5	6	7	8	9	10	15
1	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54	241,88	245,95
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,43
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,70
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,86
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,62
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	3,94
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,51
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,22
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,01
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,85
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,72
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,62
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,53
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,46
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,40
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,35
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,31

18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,27
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,23
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,20

http://www.exponenta.ru/educat/referat/XIkonkurs/student1/F-criteria.pdf

Таблица значений критерия χ^2 (хи-квадрат) Пирсона

Значения критерия χ^2 (хи-квадрат) для уровня значимости α=0,01; α=0,05; α=0,10

Вероятность $p = \alpha$

k - число степеней свободы

k \ α	0,01	0,05	0,10	k \ p	0,01	0,05	0,10
1	6,635	3,841	0,000	26	45,642	38,885	12,198
2	9,210	5,991	0,020	27	46,963	40,113	12,879
3	11,345	7,815	0,115	28	48,278	41,337	13,565
4	13,277	9,488	0,297	29	49,588	42,557	14,256
5	15,086	11,071	0,554	30	50,892	43,773	14,953
6	16,812	12,592	0,872	31	52,191	44,985	15,655
7	18,475	14,067	1,239	32	53,486	46,194	16,362
8	20,090	15,507	1,647	33	54,776	47,400	17,074
9	21,666	16,919	2,088	34	56,061	48,602	17,789
10	23,209	18,307	2,558	35	57,342	49,802	18,509
11	24,725	19,675	3,053	36	58,619	50,998	19,233
12	26,217	21,026	3,571	37	59,893	52,192	19,960
13	27,688	22,362	4,107	38	61,162	53,384	20,691
14	29,141	23,685	4,660	39	62,428	54,572	21,426
15	30,578	24,996	5,229	40	63,691	55,758	22,164
16	32,000	26,296	5,812	41	64,950	56,942	22,906
17	33,409	27,587	6,408	42	66,206	58,124	23,650
18	34,805	28,869	7,015	43	67,459	59,304	24,398
19	36,191	30,144	7,633	44	68,710	60,481	25,148

20	37,566	31,410	8,260	45	69,957	61,656	25,901
21	38,932	32,671	8,897	46	71,201	62,830	26,657
22	40,289	33,924	9,542	47	72,443	64,001	27,416
23	41,638	35,172	10,196	48	73,683	65,171	28,177
24	42,980	36,415	10,856	49	74,919	66,339	28,941
25	44,314	37,652	11,524	50	76,154	67,505	29,707

Рекомендуемая литература

1. КОМПЬЮТЕРНЫЕ ТЕХНОЛОГИИ ОБРАБОТКИ ИНФОРМАЦИИ В МЕДИЦИНЕ И ЗДРАВООХРАНЕНИИ. Царик Г.Н., Ивойлов В.М., Штернис Т.А., Полянская И.А., Цитко Е.А., Алешина А.А., Ткачева Е.С., Васильев Е.В., Жевняк Е.В., Мун С.А. Учебно-методическое пособие. Приложение к учебнику "Здравоохранение и общественное здоровье" / Кемерово, 2016.

2. СТАТИСТИЧЕСКИЕ МЕТОДЫ ИССЛЕДОВАНИЙ В МЕДИЦИНЕ И ЗДРАВООХРАНЕНИИ. Царик Г.Н., Ивойлов В.М., Штернис Т.А., Полянская И.А., Цитко Е.А., Алешина А.А., Ткачева Е.С., Васильев Е.В., Жевняк Е.В., Мун С.А. Учебно-методическое пособие. Приложение к учебнику "Общественное здоровье и здравоохранение" / Кемерово, 2016.

3. РЕГРЕССИОННЫЙ АНАЛИЗ В МЕДИКО-БИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ. Мун С.А., Глушков А.Н., Штернис Т.А., Ларин С.А., Максимов С.А. Методические рекомендации предназначены для врачей-специалистов, аспирантов, ординаторов, интернов, студентов медицинского вуза / Кемерово, 2012.

4. ОБЩЕСТВЕННОЕ ЗДОРОВЬЕ И ЗДРАВООХРАНЕНИЕ Царик Г.Н., Ивойлов В.М., Шпилянский Э.М., Грачева Т.Ю., Цитко Е.А., Штернис Т.А., Полянская И.А., Сергеев А.С., Цой В.К., Тё Е.А., Тё И.А., Савина Г.С., Седачева Л.А., Кирилкина Г.В., Рытенкова О.Л., Пачгин И.В., Друшляк И.А., Мурзинцева С.И., Алешина А.А., Ткачева Е.С. и др. Под редакцией Г.Н. Царик. Кемерово, 2016.