РОССИЙСКАЯ АКАДЕМИЯ ОБРАЗОВАНИЯ
Институт управления образованием

Д.А. Новиков

СТАТИСТИЧЕСКИЕ МЕТОДЫ В ПЕДАГОГИЧЕСКИХ ИССЛЕДОВАНИЯХ (ТИПОВЫЕ СЛУЧАИ)

Москва
МЗ-Пресс
2004

УДК 519.6
ББК 65, 74
Н 73

Новиков Д.А. Статистические методы в педагогических исследованиях (типовые случаи). М.: МЗ-Пресс, 2004. – 67 с.

ISBN 5-94073-073-6

Серия «Статистические методы»

Редакционный совет серии: Богданов Ю.И., Вощинин А.П., Горбачев О.Г., Горский В.Г., Кудлаев Э.М., Натан А.А., Новиков Д.А., Орлов А.И. (*председатель*), *Татарова Г.Г., Толстова Ю.Н., Фалько С.Г., Шведовский В.А.*

Работа содержит "рецепты" применения статистических методов в типовых случаях анализа экспериментальных данных в педагогических исследованиях. Приводится алгоритм выбора статистического критерия, методики определения достоверности совпадений и различий характеристик исследуемых объектов. Анализируются наиболее распространенные ошибки. Изложение сопровождается примерами анализа результатов педагогических экспериментов.

Работа рассчитана на педагогов-исследователей, в первую очередь, на аспирантов и соискателей.

Рецензенты: А.М. Новиков – д.п.н., проф., академик Российской академии образования
А.И. Орлов – д.т.н., проф., президент Российской ассоциации статистических методов

УДК 519.6
ББК 65, 74

ISBN 5-94073-073-6

СОДЕРЖАНИЕ

ПРЕДИСЛОВИЕ

С большим удовольствием представляю читателю замечательную книгу, которая может осчастливить начинающего исследователя. В ней всё рассказано о статистических методах, всё то, что надо знать для успешного самостоятельного применения этих методов в педагогических исследованиях. А дальше – выход в море более продвинутых методов. Конечно, если такой выход нужен.

Статистические методы – это набор инструментов научного работника. Одни инструменты предназначены для первичной обработки, другие – для более тонкой отделки. Одни используются чаще, другие – реже. Одни – современные, другие устарели. Но есть базовый набор, которым должен владеть каждый научный работник. Этот набор и представлен в книге профессора Д.А. Новикова.

В настоящее время теория измерений – это базовая общенаучная теория, с которой должен быть знаком каждый научный работник. В книге рассмотрены основные шкалы измерения. Из них в педагогических исследованиях, да и в любых иных, наиболее часто применяются шкалы порядка и отношений. На основе теории измерений дается обоснованная критика распространенной практике использования «среднего балла».

Изложение построено на основе выделенной автором структуры педагогического эксперимента. Эта структура такова. Создаются экспериментальная и контрольная группы. Проверяется отсутствие различий между ними. Затем в экспериментальной группе применяется исследуемая методика. А в контрольной – традиционная. Если в конечном состоянии группы различаются, то налицо эффект (превосходство) исследуемой методики.

В книге рассмотрены методы решения шести базовых задач. Для каждой из двух наиболее часто применяемых шкал измерения (порядковой и отношений) разобраны методы описания данных, проверки совпадения характеристик двух групп и установления различия двух групп. Приведены все необходимые формулы и алгоритмы расчетов. Нет необходимости обращаться к иной литературе – все есть в этой книге!

Однако статистические методы отнюдь не исчерпываются базовыми задачами. «Продвинутым» исследователям целесообразно обратиться к существенно более толстым сочинениям, многие из которых указаны в списке литературы. В частности, при различии групп в начальном состоянии может помочь технология стандартизации выборки. Более того, контрольная группа не всегда нужна, например, при изучении взаимосвязи признаков.

Наконец – самое важное. Настоящая книга полезна не только при проведении педагогических исследований. Столь же хорошо она может быть использована и в научных медицинских исследованиях. А также и в любых иных областях науки, отраслях народного хозяйства.

Книга выходит в серии «Статистические методы» издательства МЗ-Пресс. Прочитаете ее – переходите к другим книгам серии.

Президент Российской ассоциации
статистических методов

А.И. Орлов

1. ВВЕДЕНИЕ

Экспериментальные исследования играют существенную роль во всех науках. Можно утверждать, что, чем менее строгой является наука, тем более значимую роль в ней играет *эксперимент*[1]. Действительно, в науках сильной версии (см. [14]), использующих математический аппарат, многие результаты могут быть получены и обоснованы теоретически, на базе существующего эмпирического материала. В науках же слабой версии, к которым на сегодняшний день принадлежит и *педагогика*, эксперимент зачастую является единственным способом подтверждения справедливости гипотезы и результатов теоретического исследования, так как отсутствие общепринятой аксиоматики и адекватного формального аппарата не позволяет привести должного обоснования, не прибегая к эксперименту. Например, можно ли априори сказать, что та или иная новая методика обучения или воспитания более эффективна, чем известные и применяемые до нее? Вряд ли – пока эта методика не будет апробирована, и результаты ее применения не будут сопоставлены с результатами применения традиционных методик, никаких выводов сделать нельзя.

При планировании и подведении результатов эксперимента существенную роль играют *статистические методы*, которые дают, в том числе, возможность устанавливать степень достоверности сходства и различия исследуемых объектов на основании результатов измерений их показателей.

Анализ диссертационных исследований по педагогическим наукам (см. четвертый раздел настоящей работы) позволяет констатировать, что на сегодняшний день складывается следующая картина. С одной стороны, большинство исследователей четко представляет, что использование статистических методов необходимо (хотя бы потому, что это является общепринятым требованием в науке), и существует обширная литература по теоретической и прикладной статистике. С другой стороны, статистические мето-

---

[1] *Эксперимент – общий эмпирический метод исследования, суть которого заключается в том, что явления и процессы изучаются в строго контролируемых и управляемых условиях. Основной принцип любого эксперимента – изменение только одного фактора при неизменности и контролируемости всех остальных факторов.*

ды в педагогике либо не используются вообще, либо часто используются некорректно.

Объяснений этому несколько. Во-первых, необходимо признать, что существующая литература в большинстве своем ориентирована на людей, имеющих математическое или техническое образование, и практически недоступна гуманитариям (немногочисленные книги по математической статистике для гуманитариев [4, 5, 8, 9, 10, 12, 23, 26, 30] подавляют своим объемом и, все таки, наверное, слишком сложны). Во-вторых, класс *типовых* (наиболее распространенных, массовых) *задач* (случаев) анализа данных, возникающих в педагогических исследованиях, достаточно узок, и для эффективного решения этих задач вовсе не требуется знакомства со всем богатейшим арсеналом статистических методов. Все это приводит к тому, что педагоги-исследователи боятся использовать статистические методы, а если и используют, то на уровне "шаманских заклинаний", особо не понимая, что и как надо делать, что они делают и какие результаты получают.

Поэтому основной целью настоящей работы[1] является изложение "рецептов" применения статистических методов для решения типовых задач анализа данных в педагогических исследованиях. Желающим же получить более полное представление о том, как и в каких ситуациях, какие методы можно и нужно использовать, порекомендуем ознакомиться с перечисленными в списке литературы многочисленными учебниками и книгами, содержащими методики и опыт применения статистических методов в различных областях научного знания.

Дальнейшее изложение имеет следующую структуру. Во втором разделе описана модель педагогического эксперимента и алгоритм действий исследователя при организации эксперимента и обработке его результатов. Третий раздел содержит минимально необходимые сведения из теории измерений относительно того, какого рода данные существуют, и какие операции к ним применимы. В четвертом разделе проводится анализ использования статистических методов в диссертационных исследованиях по педагогике, что позволяет перечислить наиболее распространен-

[1] *Следует признать, что иногда мы были вынуждены немного жертвовать корректностью изложения в пользу его доступности.*

ные ошибки, и сформулировать в пятом разделе типовые задачи анализа данных в педагогических экспериментальных исследованиях. Шестой раздел включает описание методов решения этих задач и примеры, а также алгоритм выбора статистического критерия – принятия решения относительно того, какой метод следует использовать в той или иной конкретной ситуации.

2. СТРУКТУРА ПЕДАГОГИЧЕСКОГО ЭКСПЕРИМЕНТА

Целью эксперимента, в том числе в диссертационном исследовании по педагогическим наукам, является эмпирическое подтверждение или опровержение гипотезы исследования и/или справедливости теоретических результатов.

Рассмотрим следующую *модель педагогического эксперимента*. Пусть имеется некоторый педагогический *объект*, изменение *состояния* которого исследуется в ходе эксперимента. В качестве объекта может выступать отдельный индивид, группа, коллектив и т.д., например, множество учащихся, обучаемых по новой (предлагаемой в диссертации) методике. Состояние объекта измеряется[1] теми или иными показателями[2] (*характеристиками*) по *критериям*[3], отражающим его существенные характеристики. Примерами критериев являются: успеваемость, уровень знаний и т.д., примерами характеристик – время выполнения заданий, число сделанных учащимися ошибок, число правильно решенных задач и т.д.

Эксперимент заключается в целенаправленном *воздействии* на объект, призванном изменить его определенным образом. Собственно, это воздействие – его состав, структура, свойства и т.д. – и есть результат теоретического (теоретической части) исследова-

---

[1] *Измерение – "процесс определения какой-либо мерой величины чего-либо". Величина – "то* (*предмет, явление и т.д.*)*, что можно измерить, исчислить". Другими словами, величина – мера некоторого множества, относительно элементов которого имеют смысл утверждения – больше, меньше, равно. Мера – "единица измерения". Все определения здесь и далее взяты, если не оговорено особо, из словаря русского языка С.И. Ожегова.*

[2] *Показатель – "то, по чему можно судить о развитии и ходе чего-либо".*

[3] *Критерий – "1*) *средство для вынесения суждения; стандарт для сравнения; правило для оценки; 2*) *мера степени близости к цели".*

ния. Примерами воздействия являются новые содержание и формы, методы, средства обучения и т.д.

Следовательно, при проведении педагогического эксперимента необходимо обосновать, что состояние объекта изменилось, причем в требуемую сторону. Но этого оказывается недостаточно. Ведь нужно обосновать, что изменения произошли именно в результате произведенного воздействия.

Действительно, на утверждение о том, что успеваемость повысилась в результате использования новой методики, можно всегда возразить, – а, может быть, она сама повысилась бы, без каких-либо нововведений, или в результате каких-либо других воздействий? Аналогично, на утверждение о том, что успеваемость учащихся, прошедших обучение по новой методике, выше успеваемости тех, кто обучался по традиционной методике, можно возразить, – а, может быть, успеваемость первых до начала применения новой методики была выше, и, если бы новая методика не применялась, то она в результате оказалась бы выше наблюдаемой?

Таким образом, для того, чтобы выделить в явном виде результат целенаправленного воздействия на исследуемый объект, необходимо взять аналогичный объект и посмотреть, что происходит с ним в отсутствии воздействий.

Традиционно эти два объекта в экспериментальных исследованиях называют соответственно *экспериментальной группой* (например, обучаемой по предложенной методике) и *контрольной группой* (например, обучаемой по традиционной методике).

На рисунке 1 представлена в общем виде структура любого педагогического эксперимента (двойными пунктирными стрелками отмечены процедуры сравнения[1] характеристик объектов).

[1] *При этом мы по умолчанию подразумеваем, что методы* (*методики, тесты и т.д.*) *измерения характеристик объектов одинаковы. Например, сравнивать уровни знаний членов экспериментальной и контрольной группы, предлагая им различные наборы задач, нельзя.*

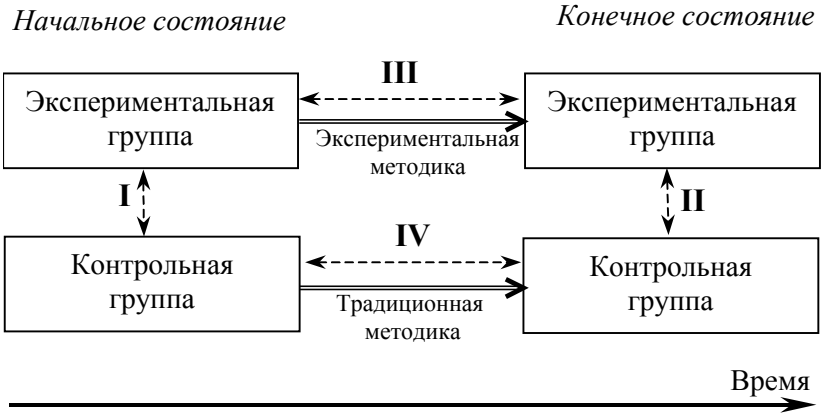

Рис. 1. Структура педагогического эксперимента

Констатации (в результате сравнения III – см. рисунок 1) различий начального и конечного состояний (динамики) экспериментальной группы недостаточно – быть может, аналогичные изменения происходят и с контрольной группой, что может быть установлено сравнением IV. Поэтому **алгоритм** действий исследователя заключается в следующем:

1) На основании сравнения I установить совпадение[1] начальных состояний экспериментальной и контрольной группы;

2) Реализовать воздействие на экспериментальную группу[2];

3) На основании сравнения II установить различие конечных состояний экспериментальной и контрольной группы.

Легко видеть, что, выполняя перечисленные шаги[3], мы, фактически, косвенным образом реализуем процедуру сравнения III,

---

[1] Если говорить корректно, то с точки зрения математической статистики совпадение установить невозможно – можно установить различие или отсутствие статистически значимого различия.

[2] При выполнении данного шага необходимо быть уверенным, что и экспериментальная, и контрольная группы находятся в одинаковых условиях, за исключением целенаправленно изменяемых исследователем.

[3] Эксперимент может следовать и более сложной, но укладывающейся в рамки описанной идеологии, схеме – например, характеристики контрольных и экспериментальных групп могут измеряться и сравниваться неоднократно, в различные моменты времени.

исключая влияние общих для экспериментальной и контрольной группы условий и воздействий.

Спрашивается, а где же место статистических методов? Роль их заключается в том, чтобы корректно и достоверно обосновать совпадение или различие состояний контрольной и экспериментальной группы. Однако, прежде чем описывать эти методы, давайте рассмотрим, что понимается под "состоянием объекта" и как это состояние измерять. Проблемами измерений занимается теория измерений, поэтому приведем минимально необходимые сведения из этой теории.

3. ЭЛЕМЕНТЫ ТЕОРИИ ИЗМЕРЕНИЙ

Информация, имеющаяся о начальных и конечных состояниях экспериментальной и контрольной группы, определяется проведенными измерениями. Любое измерение производится в той или иной *шкале*, и выбранная шкала определяет тип получающихся данных и множество операций, которые можно с этими данными осуществлять. Поэтому в настоящем разделе дается краткий обзор свойств основных шкал измерений, а затем описываются наиболее распространенные в педагогических исследованиях типы экспериментальных данных и методы их первоначальной обработки (до применения статистических методов).

3.1. ШКАЛЫ ИЗМЕРЕНИЙ

Состояние объекта оценивается по тем или иным критериям. В качестве критериев могут выступать: успеваемость учащихся, эффективность управления образовательным учреждением и т.д.

Оценки измеряются в той или иной шкале. *Шкала* (условно говоря, шкала – это множество возможных значений оценок по критериям) – числовая система, в которой отношения между различными свойствами изучаемых явлений, процессов переведены в свойства того или иного множества, как правило – множества чисел.

Различают несколько *типов шкал*. Во-первых, можно выделить *дискретные шкалы* (в которых множество возможных значе-

ний оцениваемой величины конечно – например, школьная оценка в баллах – "1", "2", "3", "4", "5") и *непрерывные шкалы* (например, время, затрачиваемое учащимися на выполнение задания, в минутах). Во-вторых, выделяют *шкалы отношений*, *интервальные шкалы, порядковые* (*ранговые*) *шкалы* и *номинальные шкалы* (шкалы наименований) – см. рисунок 2, на котором отражена также мощность шкал – то есть их "разрешающая способность".

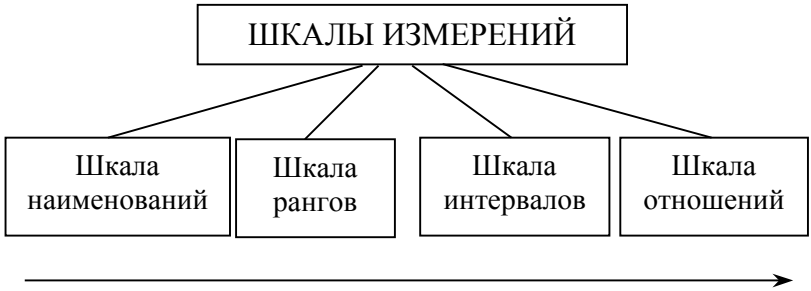

Рис. 2. Классификация шкал измерений

Рассмотрим, следуя, в основном [15, 22], свойства четырех основных типов шкал, перечисляя их в порядке убывания мощности.

Шкала отношений – самая мощная шкала. Она позволяет оценивать, во сколько раз один измеряемый объект больше (меньше) другого объекта, принимаемого за эталон, единицу. Для шкал отношений существует естественное начало отсчета (нуль), но нет естественной единицы измерений.

Шкалами отношений измеряются почти все физические величины – время, линейные размеры, площади, объемы, сила тока, мощность и т.д. В педагогических измерениях шкала отношений будет иметь место, например, когда измеряется время выполнения того или иного задания (в секундах, минутах, часах и т.п.), количество ошибок или число правильно решенных задач. В отдельных случаях, в том числе в исследованиях по трудовому и профессиональному обучению, применяются оценки и в мерах физических величин – величина допускаемых ошибок в миллиметрах при,

допустим, токарной обработке деталей, величина силы нажатия учащимся на слесарный инструмент в ньютонах (килограммах), величина электрической активности мышц в милливольтах и т.п.

Шкала интервалов применяется достаточно редко и характеризуется тем, что для нее не существует ни естественного начала отсчета, ни естественной единицы измерения. Примером шкалы интервалов является шкала температур по Цельсию, Реомюру или Фаренгейту. Шкала Цельсия, как известно, была установлена следующим образом: за ноль была принята точка замерзания воды, за 100 градусов – точка ее кипения, и, соответственно, интервал температур между замерзанием и кипением воды поделен на 100 равных частей. Здесь уже утверждение, что температура 30^0C в три раза больше, чем 10^0C, будет неверным. Справедливо говорить лишь об интервалах температур – температура 30^0C на 20^0C больше, чем температура 10^0C.

Порядковая шкала (шкала рангов) – шкала, относительно значений которой уже нельзя говорить ни о том, во сколько раз измеряемая величина больше (меньше) другой, ни на сколько она больше (меньше). Такая шкала только упорядочивает объекты, приписывая им те или иные ранги (результатом измерений является нестрогое упорядочение объектов).

Например, так построена шкала твердости минералов Мооса: взят набор 10 эталонных минералов для определения относительной твердости методом царапанья. За 1 принят тальк, за 2 – гипс, за 3 – кальцит и так далее до 10 – алмаз. Любому минералу соответственно однозначно может быть приписана определенная твердость. Если исследуемый минерал, допустим, царапает кварц (7), но не царапает топаз (8) – соответственно его твердость будет равна 7. Аналогично построены шкалы силы ветра Бофорта и землетрясений Рихтера.

Шкалы порядка широко используются в педагогике, психологии, медицине и других науках, не столь точных, как, скажем, физика и химия. В частности, повсеместно распространенная шкала школьных отметок в баллах (пятибалльная, двенадцатибалльная и т.д.) может быть отнесена к шкале порядка. В школах некоторых стран применяется и другая оценка успеваемости учащихся (как итоговая): порядковое место, которое данный ученик занимает в данном классе (выпуске). Это тоже шкала порядка.

Частным случаем порядковой шкалы является *дихотомическая шкала*, в которой имеются всего две упорядоченные *градации* – например, "справился с заданием", "не справился с заданием".

Шкала наименований (номинальная шкала), фактически, уже не связана с понятием "величина" и используется только с целью отличить один объект от другого: фамилии учеников, номера автомобилей, телефонов и т.п.

3.2. ДОПУСТИМЫЕ ПРЕОБРАЗОВАНИЯ

Результаты измерений необходимо анализировать, а для этого нередко приходится строить на их основании производные показатели, то есть, применять к экспериментальным данным то или иное преобразование. Используемая шкала определяет множество преобразований, которые допустимы для результатов измерений в этой шкале (подробнее см. публикации [13, 21, 22, 25, 29] по теории измерений).

Начнем с наиболее слабой шкалы – *шкалы наименований*, которая выделяет попарно различимые классы объектов. Например, в шкале наименований измеряются значения признака "пол": "девочки" и "мальчики". Эти классы будут различимы независимо от того, какие различные термины или знаки для их обозначений будут использованы: "лица женского пола" и "лица мужского пола", или "girls" и "boys", или "А" и "Б", или "1" и "2", или "2" и "3" и т.д. Следовательно, для шкалы наименований применимы любые взаимно-однозначные преобразования, то есть сохраняющие четкую различимость объектов (таким образом, самая слабая шкала – шкала наименований – допускает самый широкий диапазон преобразований).

Отличие *порядковой шкалы* (шкалы рангов) от шкалы наименований заключается в том, что в шкале рангов классы (группы) объектов упорядочены. Поэтому произвольным образом изменять значения признаков нельзя – должна сохраняться упорядоченность объектов (порядок следования одних объектов за другими). Следовательно для порядковой шкалы допустимым является любое монотонное преобразование. Например, если ученик Иванов набрал 5 балов, а ученик Сидоров – 10, то их упорядочение не изменится, если мы число баллов умножим на одинаковое для всех

учеников положительное число, или сложим с некоторым одинаковым для всех числом, или возведем в квадрат и т.д. (например, вместо "1", "2", "3", "4", "5" используем соответственно "3", "5", "9", "17", "102"). При этом изменятся разности и отношения "баллов", но упорядочение сохранится. В некоторых школах, используются ранговые нечисловые шкалы, например, пятерка соответствует букве А или, например, пятиугольнику, четверка – букве В или четырехугольнику, и т.д., и учащиеся знают, что А лучше В, В лучше С и т.д.

Для *шкалы интервалов* допустимо уже не любое монотонное преобразование, а только такое, которое сохраняет отношение разностей оценок, то есть линейное преобразование – умножение на положительное число и добавление постоянного числа. Например, если к значению температуры в градусах Цельсия добавить минус 273^0С, то получим температуру по Кельвину, причем разности любых двух температур в обоих шкалах буду одинаковы.

И, наконец, в наиболее мощной шкале – *шкале отношений* – возможны лишь преобразования подобия – умножение на положительное число. Содержательно это означает, что, например, отношение масс двух предметов не зависит от того, в каких единицах измерены массы – граммах, килограммах и т.д.

Суммируем сказанное в таблице 1, которая отражает соответствие между шкалами и допустимыми преобразованиями.

Таблица 1

Шкалы и допустимые преобразования

Шкала	**Допустимое преобразование**
Наименований	Взаимно-однозначное
Порядковая	Строго монотонное
Интервальная	Линейное
Отношений	Подобия

Как отмечалось выше, результаты любых измерений относятся, как правило[1], к одному из основных (перечисленных выше)

[1] *Результатами измерений могут быть и более сложные данные – ранжировки, последовательности и т.д., встречающиеся в педагогических исследованиях чрезвычайно редко, поэтому их рассмотрение выходит за рамки настоящей работы – см., например,* [*1-4, 13, 22, 33*].

типов шкал. Однако получение результатов измерений не является самоцелью – эти результаты необходимо анализировать, а для этого нередко приходится строить на их основании *производные показатели*. Эти производные показатели могут измеряться в других шкалах, нежели чем исходные. Например, можно для оценки знаний учащихся применять 100-балльную шкалу. Но она слишком детальна, и ее можно перестроить в пятибалльную ("1" – от "1" до"10"; "2" – от "10" до "30" и т.д.), или двухбалльную (например, положительная оценка – все, что выше 50 баллов, отрицательная – 50 и меньше). Следовательно, возникает проблема – какие преобразования можно применять к тем или типам исходных данных. Другими словами, переход от какой шкалы к какой является корректным. Эта проблема в теории измерений получила название *проблемы адекватности*.

Для решения проблемы адекватности можно воспользоваться свойствами взаимосвязи шкал и допустимых для них преобразований, так как отнюдь не любая операция при обработке исходных данных является допустимой. Так, например, такая распространенная операция, как взятие среднего арифметического, не может быть использована, если измерения получены в порядковой шкале [13, 22]. Общий вывод таков – всегда возможен переход от более мощной шкалы к менее мощной, но не наоборот (например, на основании оценок, полученных в шкале отношений, можно строить балльные оценки в порядковой шкале, но не наоборот).

Завершая обсуждение шкал измерений, в качестве отступления отметим, что мы рассматриваем процесс обработки результатов измерений, но вовсе не затрагиваем проблемы, связанные, во-первых, с процедурой измерений (то есть с тем, каким образом получается информация об объекте), во-вторых, с тем, какого рода информация представляет интерес с точки зрения проводимого педагогического исследования, и, наконец, в-третьих, с тем, что понимать под "улучшением" или "ухудшением" состояния исследуемого объекта, то есть, каковы критерии эффективности [12, 15-17] (подобные содержательные аспекты находятся вне компетенции математики – статистические методы позволяют лишь установить и обосновать сходство или различие объектов, а как их интерпретировать – вопрос педагогики).

Не останавливаясь на том очевидном требовании, что для сравнения результатов измерений ко всем объектам должны применяться одинаковые процедуры измерений (например, нельзя сравнивать результаты выполнения двумя различными учениками двух различных тестов), а также не перечисляя методы измерений, используемые в педагогике (с ними можно ознакомиться в [12, 16, 17]), отметим, что отдельной и чрезвычайно интересной областью исследований является выбор показателей, наиболее адекватно, и, в то же время, емко отражающих изучаемые свойства объекта. К этой содержательной области относятся задачи построения тестов, выбора методик оценки знаний и умений и т.д. Кроме того, необходимо подчеркнуть, что проблема адекватности возникает не только при переходе от одной шкалы к другой, но и при выборе шкалы для получения первоначальных оценок – непосредственной информации об объекте. И здесь опять справедлив вывод о том, что шкала должна быть адекватна – если она слишком мощная, то возможен большой произвол (например, при измерении качественных характеристик в шкале отношений), если слишком слабая, то происходят потери информации (например, при измерении количественных показателей в номинальной шкале). Например, наверное, нецелесообразно, с одной стороны, оценивать результаты решения одной задачи в 100-балльной шкале, а с другой стороны, результаты решения 100 задач в двухбалльной шкале.

Теперь, когда мы совершили небольшой экскурс в теорию измерений, рассмотрим вопрос о применении шкал измерений в педагогических исследованиях.

3.3. ПРИМЕНЕНИЕ ШКАЛ ИЗМЕРЕНИЙ В ПЕДАГОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

Наиболее распространенная мера педагогических оценок – шкала оценки знаний и умений учащихся в баллах. Школьные оценки (отметки) – удобный аппарат для практики обучения, который выполняет не только оценивающие, но и определенные воспитательные функции (стимулирования одних учащихся, "наказания" других и т.д.).

В педагогических исследованиях используются также и другие шкалы балльных оценок (порядковые шкалы). Например, выделив какие-либо уровни сформированности у учащихся определенных качеств личности или овладения той или иной деятельностью, исследователь приписывает этим уровням соответствующие значения баллов: "1", "2", "3" и т.д., или "0", "10", "100", что принципиально безразлично. Но использование порядковой шкалы как критерия оценки для педагогических исследований нежелательно, хотя и не исключено. И дело здесь не только в известной необъективности отметок, о чем уже говорилось, но и в свойствах самой шкалы порядка. В этой шкале ничего нельзя сказать о равномерности или неравномерности интервалов между соседними значениями оценок. Мы не вправе, к примеру, сказать о том, что знания учащегося, оцененные на "5", настолько же отличаются от знаний, оцененных на "4", как знания, оцененные на "4", отличаются от знаний, оцененных на "3". С тем же успехом можно было бы приписывать баллам значения не "1", "2", "3", "4", "5", а, допустим "1", "10", "100", "1000", "10000". И поэтому совершенно некорректно использование так широко применяемой в диссертациях по педагогике величины среднего балла (по классу, группе учащихся и т.д.), поскольку усреднение предполагает сложение значений величины, а операция суммы для порядковых шкал не может быть корректно определена. Соответственно не могут быть определены и все остальные арифметические и алгебраические действия.

Поэтому, например, утверждение о том, что знания учащихся в экспериментальных классах в среднем на 0,5 балла выше, чем в контрольных, будет неправомочным, некорректным. Тем более при использовании балльных оценок некорректны (даже абсурдны) утверждения типа: "эффективность экспериментальной методики в 2,6 раза выше контрольной".

Чтобы продемонстрировать, что может получиться с использованием "среднего" балла, приведем такой гипотетический пример [15]. Пусть исследовалась сравнительная эффективность двух каких-либо методик обучения, А и В. В обеих группах учащихся – контрольной и экспериментальной – было по 80 человек. Оценки производились по двум шкалам – пятибалльной и десятибалльной. Предположим, что оценки по десятибалльной шкале могут быть

пересчитаны в оценки по шкале пятибалльной: оценки "10" и "9" будут отнесены к "5", "8" и "7" – к "4" и так далее. Пусть оценки по десятибалльной шкале распределились следующим образом (в числителе указано количество учащихся, получивших соответствующую оценку в группе, обучавшейся по методике А, в знаменателе – по методике Б): "10" $\frac{20}{0}$, "9" $\frac{0}{30}$, "8" $\frac{30}{0}$, "7" $\frac{0}{30}$, "6" $\frac{20}{0}$, "5" $\frac{0}{20}$, "4" $\frac{10}{0}$, оценки "3", "2", "1" не получил никто.

Соответственно "средний балл" составит 7,50 (методика А) и 7,25 (методика В). Казалось бы, можно сделать вывод, что методика А лучше методики В. Вычислим оценки по пятибалльной шкале, в том же порядке: "5" $\frac{20}{30}$, "4" $\frac{30}{30}$, "3" $\frac{20}{20}$, "2" $\frac{10}{0}$, "1" $\frac{0}{0}$. "Средний балл" в этом случае составит 3,750 в группе, обучавшейся по методике А, и 4,125 в группе, обучавшейся по методике В. Таким образом мы получили как бы противоположный "результат": методика В лучше методики А.

Заметим, что этот "парадокс" никак не связан со статистической достоверностью различий – он будет иметь место и при очень больших выборках данных (числе учащихся). Просто это свойство слабой шкалы измерений. Сказанное будет относиться и к любым другим критериям оценки, использующим шкалу порядка.

Внимательный читатель может сказать (и будет прав), что использованное в приведенном выше примере преобразование (из десятибалльной в пятибалльную шкалу) некорректно, так как не является взаимно-однозначным. Поэтому рассмотрим еще один пример [22], в котором "парадокс" имеет место при взаимно-однозначном преобразовании. Предположим для простоты, что и экспериментальная, и контрольная группы состоят из двух учеников. Ученики в первой группе получили следующие баллы: $x_1 = 2$, $x_2 = 5$, во второй – $y_1 = 3$, $y_2 = 4$. "Средний балл" экспериментальной группы: $3{,}5 = (2 + 5) / 2$ равен "среднему баллу" контрольной группы: $3{,}5 = (3 + 4) / 2$. Применим строго монотонное (возрастающее) преобразование: “2” $\rightarrow$ “6”, “3” $\rightarrow$ “8”, “4” $\rightarrow$ “12”, “5” $\rightarrow$ “15”. Средний балл экспериментальной группы $(10{,}5 = (6 + 15) / 2)$ стал строго больше среднего балла контрольной

группы (10 = (8 + 12) / 2). Таким образом, несмотря на то, что строго монотонное преобразование является допустимым для порядковой шкалы (см. выше), соотношение между «средними» изменилось. Обусловлено это тем, что **операция вычисления среднего арифметического не является корректной в порядковой шкале**.

В принципе, шкалу балльных оценок, также как и другие шкалы порядка, можно использовать в педагогических исследованиях, но в этом случае необходимо применять адекватные методы обработки данных, не вычисляя "среднего балла". Корректной характеристикой набора балльных оценок является *медиана* (такое значение оценки, справа и слева от которого расположено одинаковое число оценок в их упорядоченной совокупности). Однако, при порядковых шкалах, имеющих малое число "разрядов" – "баллов", медиана малоинформативна (более подробно методы обработки результатов измерений в порядковой шкале рассмотрены ниже – в шестом разделе).

По приведенным выше соображениям целесообразно использовать такие способы оценки, которые позволяют применить шкалу отношений или шкалу интервалов, а не шкалу порядка (шкалы наименований в педагогических исследованиях практически не используются). Например, использовать тесты – серии коротко и точно сформулированных вопросов, заданий и т.д., на которые учащийся должен дать краткие и однозначные ответы, в правильности (или неправильности) которых нельзя сомневаться. Результатом измерений будет число правильных ответов, которое уже может измеряться в шкале отношений. Точно так же могут быть построены письменные контрольные работы, результаты обработки анкет (процент учащихся, давших положительные ответы на тот или иной вопрос) и т.д.

В общем же случае можно выделить следующие характеристики, измеряемые в шкале отношений [18]:

- временн*ы*е (время выполнения действия, операции, время реакции, время, затрачиваемое на исправление ошибки, и т.д.);
- скоростные (производительность труда, скорость реакции, движения и т.д. – величины, обратные времени);

- точностные (величина ошибки в мерах физических величин (миллиметрах, углах и т.п.), количество ошибок, вероятность ошибки, вероятность точной реакции, действия и т.д.);

- информационные (объем заучиваемого материала, перерабатываемой информации, объем восприятия и т.д.).

Методы обработки величин, измеренных в шкале отношений, рассмотрены ниже – в шестом разделе.

В заключение настоящего подраздела приведем некоторые типичные (то есть, наиболее часто встречающиеся в диссертационных исследованиях по педагогике, анализ которых приведен ниже в четвертом разделе) характеристики: уровень (степень) знаний, усвоения, обучаемости, компетентности, подготовки, адаптируемости, отношения, сформированности, удовлетворенности, профессионализма, самостоятельности, становления, развития и т.д.; качество обучения; эффективность деятельности (учебной, преподавательской, воспитательной, управленческой).

Данные характеристики в диссертационных исследованиях в большинстве случаев измерялись в порядковой шкале (чаще всего, в двух-, трех- или пятибалльной), реже – в шкале отношений (количество учащихся, успешно выполнивших задание или набравших тот или иной балл; объем усвоенного материала; время, затрачиваемое на изучение установленного объема учебного материала и т.д.).

3.4. АГРЕГИРОВАННЫЕ ОЦЕНКИ

Как правило, в любом педагогическом эксперименте имеется значительное число (десятки, сотни, а иногда и тысячи) участников – учеников, учителей, образовательных учреждений и т.д. В результате измерения показателей этих участников получается набор их *индивидуальных оценок*. Понятно, что сравнивать между собой и анализировать одновременно все индивидуальные оценки невозможно, да и нецелесообразно, так как всегда существует их разброс, обусловленный неконтролируемым различием участников эксперимента (каждый человек неповторим).

Поэтому для того, чтобы, во-первых, получить обозримое число характеристик и, во-вторых, для того, чтобы сгладить индивидуальные колебания, используют так называемые *агрегированные*

(коллективные, групповые, производные) *оценки*. Например, если имелись индивидуальные оценки успеваемости учеников, то агрегированной оценкой будет успеваемость группы.

Получение агрегированных оценок на основании индивидуальных является их преобразованием, и преобразование это следует выполнять корректно (см. обсуждение проблемы адекватности выше). Приведем некоторые корректные процедуры агрегирования для наиболее распространенных в педагогических исследованиях показателей (см. также раздел 6.1 "Описательная статистика").

Для абсолютных величин, измеренных в шкале отношений (см. их перечисление выше), наиболее типичным является вычисление среднего арифметического по группе. Эта процедура вполне корректна, и обычно ее реализация не вызывает затруднений.

Наибольшее число ошибок в педагогических исследованиях возникает при агрегировании показателей, измеренных в порядковых шкалах – пресловутый "средний балл" неискореним! Еще раз повторим – **не следует складывать, вычитать, умножать или делить баллы друг на друга, да и на чтобы то ни было – все это абсолютно бессмысленные операции**.

Если имеется набор индивидуальных баллов, то единственной адекватной характеристикой группы будет число ее членов, получивших тот или иной балл[1] (например, 20 человек получили балл "4"). Аналогичным образом агрегируется и информация о выделении уровней – если введены три уровня (например, уровни знаний: низкий, средний и высокий) и имеется информация о распределении всех членов нескольких групп (контрольных или экспериментальных) по этим уровням, то агрегированной информацией об объединенной группе будет число ее членов, обладающих тем или иным уровнем знаний (вычисляемое как сумма по всем группам числа их членов, обладающих данным уровнем знаний) – соответствующие примеры приводятся ниже.

Если в настоящем разделе речь шла об агрегировании индивидуальных оценок по группе с целью получения характеристик группы в целом, то в следующем разделе рассматривается пробле-

[1] *Отметим, что такая агрегированная характеристика группы как число ее членов* (*учащихся*)*, получивших данный балл, является величиной, измеренной в шкале отношений.*

ма агрегирования показателей, характеризующих один и тот же объект.

3.5. КОМПЛЕКСНЫЕ ОЦЕНКИ

Нередко встречаются случаи, когда какое-либо изучаемое явление, процесс характеризуется несколькими показателями – *вектором показателей*. При этом часто возникает вопрос о возможности однозначной оценки этого явления, процесса или изучаемых их свойств одной величиной – *комплексной оценкой*. Так, во многих спортивных состязаниях победитель выявляется по сумме очков, баллов, набранных на отдельных этапах состязания или в отдельных играх, в многоборье – в отдельных видах спорта. Или же другой пример из образовательной практики – аккредитация учебного заведения производится на основании оценки результатов его деятельности по фиксированному и утвержденному Министерством образования РФ набору показателей (квалификация преподавателей, обеспеченность учащихся методическими материалами и т.д.).

На практике комплексные оценки встречаются довольно часто и, очевидно, без них не обойтись, хотя способы их определения нередко и вызывают множество недоуменных вопросов. Но в любом случае такие комплексные оценки, применяемые в повседневной жизни, являются либо результатом определенных общественных соглашений, которые признаются всеми участниками, либо установлены каким-либо нормативным актом определенного директивного органа – министерства, ведомства и т.д. и в силу этого также признаются всеми заинтересованными лицами.

Другое дело – применение комплексных оценок в научном исследовании. Здесь сразу на первое место встает вопрос о научной, в том числе математической, строгости применяемой оценки. В частности, не вызывает сомнений возможность использования такой векторной оценки, как суммарные затраты времени на выполнение учащимся отдельных заданий, или суммарное количество ошибок, допущенных учащимся при выполнении отдельных заданий. Здесь суммируются однородные величины, заданные шкалами отношений.

Но, как только начинают суммироваться баллы, выставляемые одному и тому же учащемуся за выполнение, допустим, разных заданий – исследование сразу выходит за рамки научной строгости. Как уже говорилось, операция суммы для порядковой шкалы не определена. Если 5 + 2 = 4 + 3, то "5" и "2" балла – это не одно и то же, что "4" и "3" балла!

Между тем суммирование баллов довольно часто встречается в диссертациях по педагогике. Так, в одной работе диссертант для оценки деятельности учителей использовал большое количество показателей, оцениваемых по пятибалльной шкале [15]:

- структура знаний учителя (общенаучные, специальные);
- педагогические умения (проективные, конструктивные, организаторские, коммуникативные, гностические);
- нравственно-психологическая направленность педагога (внимательность к людям, справедливость, гуманизм, увлеченность делом, ответственность, самоорганизованность);
- общая одаренность (качества ума, качества речи, качества воли, характера, эмоциональные и другие качества личности) и так далее.

Общая же оценка учителю в этой работе давалась по сумме набранных баллов. Но в данном случае диссертант должен был бы задаться большой серией вопросов. Во-первых, любой учитель – личность, он осуществляет сложнейшую деятельность, и насколько правомерно оценивать его однозначно каким-то числом баллов и утверждать, что учитель Иванов, допустим, хуже учителя Петрова на 11 баллов?! Во-вторых, насколько выделенные качества равнозначны, что, к примеру, специальные знания "стоят" сколько же, сколько гуманизм?! Кроме того, вычисление суммы подразумевает взаимозаменяемость критериев[1], то есть, снижение общей одаренности на один балл может быть компенсировано таким же увеличением оценки педагогических умений?! И так далее, эту череду недоуменных вопросов можно было бы продолжать долго. И если бы диссертант над ними задумался, вряд ли бы он так легко вводил подобные "оценки".

В педагогических диссертациях, к сожалению, встречаются и другие, самые разнообразные неудачные попытки введения ком-

[1] *Данное замечание справедливо и для величин, измеряемых в шкалах отношений.*

плексных оценок, вплоть до полных курьезов. Так, для оценки эффективности деловой игры в одной из диссертационных работ была использована следующая "формула": $P = 50 – K – (B – 40)$, где P – "комплексная" оценка в баллах, 50 – максимально возможное количество баллов, K – количество замечаний, сделанных ведущим, B – время в минутах. Как видим, здесь уж, что называется, "смешались в кучу кони, люди...". Под знак суммы (разности) поставлены совершенно разнородные величины: баллы, количество замечаний, время. Кроме того, в некоторых работах предметом "исследования" является построение подобных комплексных оценок, и на полном серьезе приводятся "обоснования", чем предлагаемая автором оценка лучше других ей подобных.

В некоторое оправдание подобным неверным построениям комплексных оценок следует отметить, что проблема агрегирования векторных оценок на сегодняшний день исследована не полностью, а существующие результаты, даже для их применения на практике, зачастую требуют хорошего знания высшей математики. Достаточно простым и интуитивно понятным (но, в то же время, корректным) методом агрегирования балльных оценок является использование так называемых матриц свертки [7, 19], элементы которых содержат значения агрегированного показателя, а агрегируемые баллы задают номер строки и столбца. Например, если с целью получения оценки знаний по естественнонаучным предметам агрегируются баллы, полученные по физике и по химии, то матрица свертки будет содержать баллы[1], соответствующие всем возможным комбинациям исходных оценок – можно условно считать, что, если по физике набраны 4 балла, а по химии – 3, то агрегированная оценка равна, допустим, трем баллам, если по физике набраны 3 балла, а по химии – 4, то агрегированная оценка равна четырем баллам (при этом приоритет явно отдается химии) и т.д.

Для тех, кто глубже заинтересуется проблемой комплексных оценок и принятия решений при многих критериях, можно реко-

---

[1] *Как отмечалось выше, общим свойством порядковых шкал является то, что сравниваемые результаты их преобразований должны быть измеримы в исходной шкале – например, если используется пятибалльная шкала (1, 2, 3, 4 и 5), то результат агрегирования набора измерений может принимать только одно из этих пяти значений.*

мендовать ознакомиться с соответствующими публикациями [20, 24, 31]. Но в любом случае при построении комплексных оценок нужно быть предельно внимательным и осторожным. Кстати, **нередко можно обойтись и без них**. Если получены количественные результаты по отдельным показателям, то можно ограничиться их качественной интерпретацией, не "загоняя под общий знаменатель", проанализировать и сравнить исследуемые объекты отдельно по каждому из показателей. И пусть по каким-то показателям результаты экспериментальных групп будут лучше контрольных, а по каким-то хуже – от этого исследование только обогатится, станет достовернее [15].

4. АНАЛИЗ ИСПОЛЬЗОВАНИЯ СТАТИСТИЧЕСКИХ МЕТОДОВ В ДИССЕРТАЦИОННЫХ ИССЛЕДОВАНИЯХ ПО ПЕДАГОГИКЕ

Для анализа применяемых в педагогических исследованиях статистических методов были использованы 118 успешно защищенных в различных диссертационных советах и утвержденных ВАК кандидатских и докторских диссертаций.

Корректность применения статистических методов. К сожалению, в 65 диссертациях (55% от общего числа!) нет никаких упоминаний об измерении и обработке экспериментальных данных (за редким исключением описания констатирующих экспериментов), поэтому анализировать их мы не будем.

В 16 из 53 оставшихся (53 = 118 – 65) диссертационных работ[1] отсутствует сравнение начальных состояний контрольной и экспериментальной групп (при этом в 12 из упомянутых 16 работ контрольные группы отсутствовали вообще, то есть рассматривалась только динамика состояния "экспериментальной" группы – см. второй раздел выше).

[1] *Эти 53 работы распределились по специальностям следующим образом: 13.00.01 – "Общая педагогика, история педагогики и образования" – 29 диссертаций, 13.00.08 – "Теория и методика профессионального образования" – 17 диссертаций, 13.00.02 – "Теория и методика обучения и воспитания" – 7 диссертаций.*

Если продолжить последовательно вычленять группу диссертационных исследований, в которых корректно использовались статистические методы, то получится следующая картина (см. также рисунок 3).

Еще в 7 из 37 оставшихся (37 = 53 – 16) использовался "средний балл" (см. выше).

В 14 из 30 оставшихся (30 = 37 – 7) работах упоминались методы, используемые при обработке данных (надо признать, что в большинстве (39 = 53 – 14) работ о методах нет ни слова, а, если и есть, то стандартные ни о чем не говорящие выражения, например: "выявлены статистические значимые различия исследуемых параметров в пользу экспериментальных групп"). И, в большинстве случаев, упоминались они "зря", так как в 8 работах (из 14!) использовались неадекватные методы.

Остаются 22 работы (22 = 30 – 8). Мы не задавались целью самостоятельного определения уровня значимости (он упоминается только в 5 работах) всех полученных результатов, но сам по себе тот факт, что в такой считающейся экспериментальной науке, как педагогика, из 118 диссертаций лишь в 22, то есть менее чем в каждой пятой, применялись адекватные статистические методы (при этом мы не утверждаем, что они применялись правильно и сделанные выводы были ими, действительно, обоснованы), более чем огорчителен.

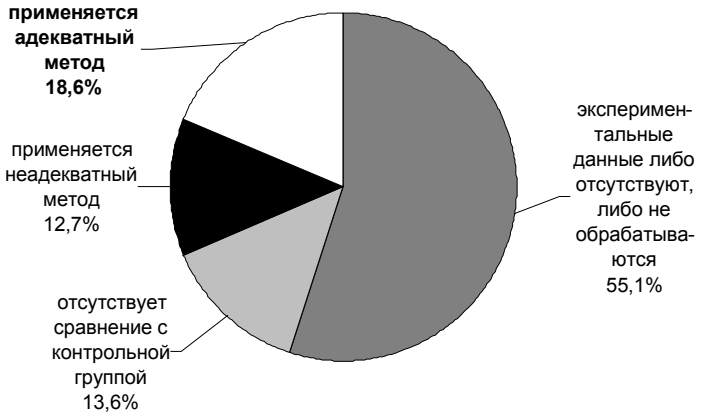

Рис. 3. Корректность и адекватность применения статистических методов в диссертационных исследованиях по педагогике

Типовые задачи (случаи). Помимо неутешительных выводов (см. рисунок 3), анализ диссертационных работ по педагогике позволил выделить *типовые задачи анализа данных*.

1. Описание данных. В части работ, не использующих статистические методы в смысле структуры педагогического эксперимента, описанной выше во втором разделе, в иллюстративных целях применялись лишь некоторые производные показатели – среднее, медиана и т.д. Краткое рассмотрение *описательной статистики*, то есть описание результатов эксперимента с помощью различных агрегированных показателей и графиков, приведено в разделе 6.1 ниже.

2. Величины, измеренные в шкале отношений. В данном классе задач результатом "измерений" являлись значения физических величин – время, затрачиваемое на выполнение упражнения; объем материала, усваиваемый в единицу времени; число и процент правильно выполненных заданий. Такие ситуации встречались в 5 диссертационных работах, что составляет около 10% от общего числа (53) работ, использующих статистические методы. Описание соответствующих статистических методов и примеры приведены ниже в разделе 6.3.

3. Величины, измеренные в порядковой шкале. В данном классе задач результатом "измерений" являлись значения таких величин, как успеваемость, удовлетворенность, заинтересованность, качество и т.д. Измерялись они в баллах, уровнях[1], рейтингах, вербальных шкалах и других порядковых величинах (см. выше).

Такие ситуации встречались в 43 диссертационных работах, что составляет более 80% от общего числа (53) работ, использующих статистические методы (см. рисунок 4).

Описание соответствующих статистических методов и примеры приведены ниже в разделе 6.4.

[1] *Простейший случай – два уровня* (*справились с заданием, не справились*) – *дихотомическая шкала.*

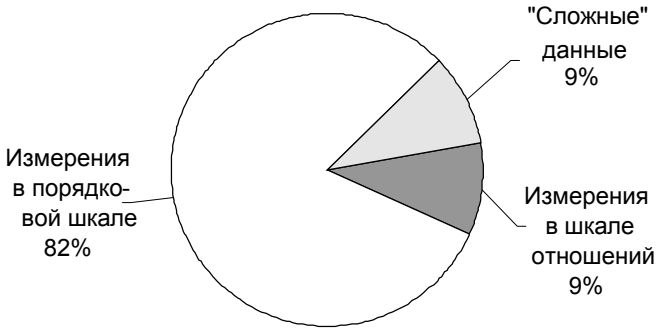

Рис. 4. Статистические методы в педагогических исследованиях

4. Задачи, требующие использования "продвинутых" статистических методов. К задачам данного класса (обработки "сложных" данных) можно отнести:

- задачи обработки результатов опросов с закрытыми вопросами, в которых респонденты могли отмечать одновременно два или более ответа на один и тот же вопрос (одна диссертационная работа) – например: "Какие факторы влияют на эффективность обучения: содержание, методы, средства, подготовленность учащихся?";

- задачи анализа ранжировок, в которых исходными данными являются упорядочения объектов (две диссертационные работы) – например, упорядочение факторов, определяющих эффективность обучения, в порядке убывания их важности (с точки зрения участников проведенного опроса);

- задачи, требующие использования факторного и регрессионного анализа (две диссертационные работы).

Так как данный класс задач (5 диссертационных работ) составляет менее 10% (5/53) от объема нашей выборки, то описывать соответствующие методы мы не будем, отослав заинтересованного читателя к многочисленным публикациям по теории и практике применения статистических методов в различных областях [1-5, 8-13, 21-23, 26-28, 30-33].

Динамика и многокритериальность. Отдельно следует отметить, что данные, отражающие:

- динамику учебно-воспитательного процесса (динамика учитывается последовательностью измерений[1]), встречались в 10 случаях из 53 (19 %);

- многокритериальность (каждый объект оценивается одновременно по нескольким критериям – см. раздел "Комплексные оценки" выше), встречались в 23 случаях из 53 (43 %);

- и динамику, и многокритериальность, встречались в 6 случаях из 53 (11 %)

Следовательно, помимо попарных однократных сравнений экспериментальной и контрольной группы по одному из критериев, возникают задачи их сравнения в динамике, а также по совокупности критериев. Обсуждение методов решения этого класса задач содержится в следующем (пятом) разделе.

Завершив анализ использования статистических методов в диссертационных исследованиях по педагогике, и выделив типовые задачи, перейдем к формализации последних.

5. ТИПОВЫЕ ЗАДАЧИ АНАЛИЗА ДАННЫХ В ПЕДАГОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

Предположим, что имеется экспериментальная группа, состоящая из N человек, и контрольная группа, состоящая из M человек (где N и M – целые положительные числа, например, $N = 25$, $M = 30$). Допустим, что в результате измерения одного и того же показателя с помощью одной и той же процедуры измерений были получены следующие данные:

$x = (x_1, x_2, \ldots, x_N)$ – *выборка*[2] для экспериментальной группы

и

$y = (y_1, y_2, \ldots, y_M)$ – выборка для контрольной группы,

---

[1] *Под "динамическими" данными подразумеваются данные, содержащие более двух измерений состояний экспериментальной и контрольной группы* (*см. рисунок 1*)*, то есть, помимо начального и конечного моментов времени, рассматривались и промежуточные.*

[2] *Выборка – совокупность значений одного и того же признака у наблюдаемых объектов. В рассматриваемом примере выборка представляет собой набор чисел, соответствующих количеству решенных учащимися задач.*

где x_i – *элемент выборки* – значение исследуемого показателя (признака[1]) у *i*-го члена экспериментальной группы, $i = 1, 2, ..., N$, а y_j – значение исследуемого показателя у *j*-го члена контрольной группы, $j = 1, 2, ..., M$. Число элементов выборки называется ее *объемом* – например, объем выборки *x* равен *N*, а объем выборки *y* равен *M*.

В зависимости от того, в какой шкале – шкале отношений или порядковой шкале – производились измерения, получаем следующие два случая.

Шкала отношений. Если измерения производились в шкале отношений (время, число и т.д.), то $\{x_i\}$ и $\{y_j\}$ – положительные, в том числе – натуральные, числа, для которых имеют смысл все арифметические операции.

Рассмотрим пример[2]. Пусть имеется экспериментальная группа, состоящая из 25 человек ($N = 25$), и контрольная группа, состоящая из 30 человек ($M = 30$), и измерение заключается в определении уровня знаний путем проведения теста, включающего 20 задач. Примем, что характеристикой учащегося (признаком) является число правильно решенных им задач. Результаты измерений уровня знаний в контрольной и экспериментальной группах до и после эксперимента приведены в таблице 2, строки которой соответствуют членам групп (отдельным учащимся). Например, первый учащийся контрольной группы до начала эксперимента правильно решил 15 задач, а третий участник экспериментальной группы после окончания эксперимента правильно решил 12 задач, и т.д.

---

[1] *Признак – свойство* (*характеристика*) *наблюдаемого объекта. В рассматриваемом примере признаком являются решенные задачи.*

[2] *Данный пример рассматривается на протяжении всего настоящего и последующего разделов. Все таблицы, диаграммы и графики экспортированы из компьютерной программы Microsoft Excel для Windows*.

Таблица 2

Результаты измерений уровня знаний в контрольной и экспериментальной группах до и после эксперимента

Контрольная группа (число правильно решенных задач до начала эксперимента)	**Экспериментальная группа (число правильно решенных задач до начала эксперимента)**	**Контрольная группа (число правильно решенных задач после окончания эксперимента)**	**Экспериментальная группа (число правильно решенных задач после окончания эксперимента)**
15	12	16	15
13	11	12	18
11	15	14	12
18	17	17	20
10	18	11	16
8	6	9	11
20	8	15	13
7	10	8	7
8	16	6	14
12	12	13	17
15	15	17	19
16	14	19	16
13	19	15	12
14	13	11	15
14	19	9	19
19	12	19	18
7	11	8	14
8	16	6	13
11	12	9	18
12	8	12	13
15	13	11	13
16	7	17	15
13	15	10	18
5	8	8	9
11	9	8	14
19	–	20	–
18	–	19	–
9	–	6	–
6	–	14	–
15	–	10	–

Результаты эксперимента могут быть получены и в порядковой шкале (или переведены из шкалы отношений в порядковую), поэтому рассмотрим представление данных в порядковой шкале.

Порядковая шкала. Если использовалась порядковая шкала (шкала рангов) с L градациям (например, в пятибалльной школьной шкале $L = 5$), то будем считать, что $\{x_i\}$ и $\{y_j\}$ – натуральные числа, принимающие одно из L значений. Для простоты можно считать, что множество значений (баллов) есть множество чисел от единицы до L. Тогда характеристикой группы будет число ее членов, набравших заданный балл (см. раздел "Агрегированные оценки" выше). То есть, для экспериментальной группы вектор баллов есть

$$n = (n_1, n_2, \ldots, n_L),$$

где n_k – число членов экспериментальной группы, получивших k-ый балл, $k = 1, 2, \ldots, L$. Для контрольной группы вектор баллов есть

$$m = (m_1, m_2, \ldots, m_L),$$

где m_k – число членов контрольной группы, получивших k-ый балл, $k = 1, 2, \ldots, L$. Очевидно, что

$$n_1 + n_2 + \ldots + n_L = N, \quad m_1 + m_2 + \ldots + m_L = M.$$

Пусть в рассматриваемом примере (в котором ($N = 25$, $M = 30$) выделены три уровня знаний ($L = 3$): низкий (число решенных задач меньше либо равно 10), средний (число решенных задач строго больше 10, но меньше либо равно 15) и высокий (число решенных задач строго больше 15). Сформируем в компьютерной программе Microsoft Excel для Windows таблицу 3, в которой указаны верхние границы диапазонов.

Таблица 3

Переход от шкалы отношений к порядковой шкале

Уровень знаний	**Максимальное число правильно решенных задач**
Низкий	10
Средний	15
Высокий	20

Поставим в соответствие уровням знаний (низкому, среднему и высокому) баллы – 1, 2 и 3 (эта операция является корректной для порядковой шкалы – см. раздел "Допустимые преобразования" выше). Вычислим на основании данных таблицы 2, например, сначала для контрольной группы до начала эксперимента число ее членов, получивших балл, принадлежащий тому или иному диапазону: $m_1 = 9$ (то есть, 9 членов контрольной группы до начала эксперимента продемонстрировали низкий уровень знаний), $m_2 = 14$, $m_3 = 7$. Результаты[1] занесем в таблицу 4.

Таблица 4

Уровни знаний членов контрольной группы до эксперимента

Уровень знаний	**Частота (число человек)**
Низкий (1 балл)	9
Средний (2 балла)	14
Высокий (3 балла)	7

Для каждого из столбцов таблицы 2 по аналогии с таблицей 4 определяем распределение членов экспериментальной и контрольной групп по уровням знаний и получаем таблицу 5.

Таблица 5

Результаты измерений уровня знаний в контрольной и экспериментальной группах до и после эксперимента

Уровень знаний	**Контрольная группа до начала эксперимента (чел.)**	**Экспериментальная группа до начала эксперимента (чел.)**	**Контрольная группа после окончания эксперимента (чел.)**	**Экспериментальная группа после окончания эксперимента (чел.)**
Низкий	9	7	12	2
Средний	14	12	10	13
Высокий	7	6	8	10

[1] *В компьютерной программе Microsoft Excel для Windows таблица 4 получается из таблиц 2 и 3 применением инструмента анализа данных "Гистограмма"* (*Меню/Сервис/Анализ данных/Гистограмма*).

Таблица 5 построена по таблице 2 введением диапазонов значений числа правильно решенных задач, попадание в которые считалось соответствующим уровням знаний. Отметим, что при подобном переходе от шкалы отношений к порядковой шкале часть информации теряется – в рассматриваемом примере одному и тому же уровню знаний соответствуют несколько различных чисел правильно решенных задач. Следовательно, труднее становится устанавливать совпадения и различия характеристик исследуемых объектов. Поэтому, рекомендуется использовать всю имеющуюся информацию, то есть, если при измерениях использовалась шкала отношений, то и обрабатывать данные следует в этой шкале.

Однако, во многих случаях на практике измерения производят в порядковой шкале (например, оценивают знания в баллах), и результаты эксперимента сразу имеют вид таблицы типа таблицы 5. Поэтому для задач анализа результатов измерений, произведенных в шкале отношений, будем считать, что данные эксперимента имеют вид таблицы 2, а для задач анализа результатов измерений, произведенных в шкале порядка, будем считать, что данные эксперимента имеют вид таблицы 5.

Типовые задачи анализа данных. Завершив описание используемых в качестве примера исходных данных, отметим, что с точки зрения их анализа можно выделить три типа задач:

- *описание данных* (компактное и информативное отражение результатов измерений характеристик исследуемых объектов);

- *установление совпадения* характеристик двух групп (например, экспериментальной и контрольной – см. сравнение I на рисунке 1));

- *установление различия* характеристик двух групп (например, экспериментальной и контрольной – см., сравнение II на рисунке 1, или экспериментальной группы в различные моменты времени – см., сравнение III на рисунке 1 и т.д.).

Два типа шкал (отношений и порядка) и три перечисленные типа задач анализа данных позволяют выделить шесть базовых (типовых) задач, приведенных в таблице 6 и условно обозначенных "задача 1.1" – "задача 2.3". Например, задача 1.1 заключается в описании данных, измеренных в шкале отношений и т.д.

Таблица 6

Типовые задачи анализа данных

	1. Шкала отношений	**2. Шкала порядка**
1. Описание данных	Задача 1.1	Задача 2.1
2. Установление совпадения характеристик двух групп	Задача 1.2	Задача 2.2
3. Установление различия двух групп	Задача 1.3	Задача 2.3

Введенная классификация типовых задач анализа данных в педагогических исследованиях определяет структуру дальнейшего изложения:

- описание данных заключается в создании *описательной статистики*, которая рассмотрена далее для обоих типов шкал (задачи 1.1 и 2.1) в разделе 6.1;

- задачи установления совпадений и/или различий характеристик двух групп для данных, измеренных в шкале отношений (задачи 1.2 и 1.3), рассматриваются в разделе 6.3;

- задачи установления совпадений и/или различий характеристик двух групп для данных, измеренных в шкале порядка[1] (задачи 2.2 и 2.3), рассматриваются в разделе 6.4.

Перечисленные шесть задач являются **базовыми** по следующим причинам. Во-первых, они включают большинство (90 % – см. четвертый раздел) задач анализа данных, встречающихся в экспериментальных исследованиях по педагогическим наукам. Во-вторых, они сформулированы для простейшей схемы организации педагогического эксперимента (см. второй раздел) – когда состояние исследуемых объектов описывается одним показателем и измеряется два раза – до начала и после завершения воздействия. Сделаем пояснение для других случаев.

Если возникает многокритериальность (объекты описываются одновременно по нескольким критериям – см. раздел "Комплексные оценки" выше), то описание и сравнение экспериментальной и

[1] *Отдельно рассматриваются методы обработки измерений, произведенных в дихотомической шкале – см. раздел 6.5.*

контрольной групп[1] по каждому из критериев может производиться независимо в рамках одной из базовых задач.

Аналогично, если возникает динамика (то есть, состояния объектов измеряются более, чем два раза), то описание и сравнение групп может производиться несколько раз независимо (в каждый момент времени) в рамках одной из базовых задач 1.1-2.3 (см. таблицу 6).

Если же у исследователя имеется желание сразу анализировать одновременно несколько групп (в динамике) и/или несколько показателей, то необходимо применение статистических методов многомерного анализа. Их описание выходит за рамки настоящей работы, ознакомится с ними можно в публикациях [2, 22, 28, 32].

Рассмотрим методы решения типовых для педагогических исследований задач анализа данных.

6. МЕТОДЫ ОБРАБОТКИ ДАННЫХ И ПРИМЕРЫ

Настоящий раздел содержит методики анализа данных для выделенных выше шести типовых задач (см. таблицу 6): описательная статистика, анализ совпадений и различий характеристик экспериментальной и контрольной групп на основании измерений, проведенных в порядковой шкале или шкале отношений. В качестве иллюстрации рассматривается реализация этих методик для числового примера (см. таблицы 2 и 5).

6.1. ОПИСАТЕЛЬНАЯ СТАТИСТИКА

В практических задачах обычно имеется совокупность наблюдений (десятки, сотни, а иногда – тысячи результатов измерений индивидуальных характеристик), поэтому возникает задача компактного описания имеющихся данных. Для этого используют методы *описательной статистики* – описания результатов с помощью различных агрегированных показателей и графиков.

[1] *Встречаются случаи, когда имеется несколько экспериментальных или несколько контрольных групп. При этом попарное их сравнение все равно является одной из базовых задач.*

Кроме того, некоторые показатели описательной статистики используются в статистических критериях (см. разделы 6.3 и 6.4) при определении достоверности совпадений и/или различий характеристик экспериментальной и контрольной группы.

Для результатов измерений в шкале отношений (задача 1.1 – см. таблицу 6) показатели описательной статистики можно разбить на несколько групп [32]:

- *показатели положения* описывают положение экспериментальных данных на числовой оси. Примеры таких данных – *максимальный и минимальный элементы выборки*, *среднее значение*[1], *медиана*[2], *мода*[3] и др.;

- *показатели разброса* описывают степень разброса данных относительно своего центра (среднего значения). К ним относятся: *выборочная дисперсия*[4], *разность между минимальным и максимальным элементами* (*размах*, *интервал* выборки) и др.

- *показатели асимметрии*: положение медианы относительно среднего и др.

- *гистограмма*[5] и др.

Данные показатели используются для наглядного представления и первичного ("визуального") анализа результатов измерений характеристик экспериментальной и контрольной группы.

---

[1] *Имеется в виду среднее арифметическое значение.*

[2] *Медианой называется значение исследуемого признака, справа и слева от которого находится одинаковое число элементов выборки.*

[3] *Модой называется такое значение измеренного признака, которым обладает максимальное число элементов выборки, то есть значение, которое встречается в выборке наиболее часто. Например, если исследовалось число правильно решенных учащимися задач, то модой будет такое число задач, для которого число учащихся, правильно решивших именно это число задач, максимально.*

[4] *Выборочная дисперсия рассчитывается как средняя сумма квадратов разностей между элементами выборки и средним значением. Дисперсия характеризует разброс элементов выборки вокруг среднего значения.*

[5] *Гистограммой называется графическое изображение зависимости частоты попадания элементов выборки от соответствующего интервала группировки* (*диапазона значений показателя*).

Приведем формулы расчета основных показателей. *Среднее арифметическое* $\bar{x}$ выборки $\{x_i\}_{i=1...N}$ (*выборочное среднее*) рассчитывается следующим образом[1]:

$$(1)\ \bar{x} = \frac{1}{N}(x_1 + x_2 + x_3 + \ldots + x_{n-1} + x_n) = \frac{1}{N}\sum_{i=1}^{N} x_i ,$$

а *выборочная дисперсия* D_x:

$$(2)\ D_x = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 .$$

В компьютерной программе Microsoft Excel для Windows описательная статистика получается применением инструмента анализа данных "Описательная статистика" (Сервис/Анализ данных/Описательная статистика). Описательная статистика для первого столбца таблицы 2 (числа правильно решенных задач в контрольной группе до начала эксперимента) приведена в таблице 7.

Таблица 7

Описательная статистика числа правильно решенных задач в контрольной группе до начала эксперимента (см. первый столбец таблицы 2)

Среднее	12,6
Стандартная ошибка	0,76
Медиана	13
Мода	15
Стандартное отклонение	4,16
Дисперсия выборки	17,28
Эксцесс	-0,89
Асимметричность	-0,03
Интервал (размах)	15
Минимум	5
Максимум	20
Сумма	378
Счет (объем выборки)	30

[1] *Символ* $\sum_{i=1}^{n} x_i$ *здесь и далее обозначает сумму элементов* $\{x_i\}$ *по индексу i, пробегающему последовательно все значения от единицы до n*: $x_1 + x_2 + \ldots + x_n$.

Целый ряд приведенных в таблице 7 показателей описательной статистики педагогу-исследователю не понадобятся (далее используются только среднее (формула (1), первая строка таблицы 7), дисперсия (формула (2), шестая строка таблицы 7) и "счет" – последняя строка таблицы 7). Тем не менее, мы приводим все показатели, которые автоматически выводит "Описательная статистика" в компьютерной программе Microsoft Excel для Windows (таблица 7 экспортирована из Excel), чтобы уважаемый читатель не терялся перед экраном компьютера.

Гистограмма в Excel получается применением инструмента анализа данных "Гистограмма" (Сервис/Анализ данных/Гистограмма). Гистограмма числа правильно решенных задач в контрольной группе до начала эксперимента (первый столбец таблицы 2) представлена на рисунке 5.

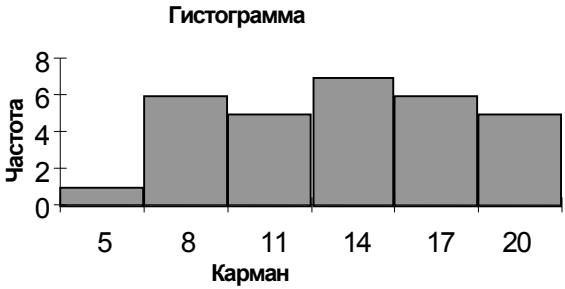

Рис. 5. Гистограмма числа правильно решенных задач в контрольной группе до начала эксперимента ("частота" *– число элементов выборки, попавших в заданный диапазон, называемый в Excel* "карманом"*)*

Рассмотрим теперь показатели описательной статистики для данных, измеренных в порядковой шкале.

Для результатов измерений в порядковой шкале (задача 2.1 – см. таблицу 6) при небольшом числе градаций единственным информативным показателем описательной статистики является гистограмма[1].

[1] *Если число градаций (различных значений) велико, то информативными также являются мода и медиана.*

Для визуального (качественного) сравнения экспериментальной и контрольной групп удобно строить для них совместные гистограммы. Например, по результатам таблицы 5 (см. выше) можно построить несколько парных гистограмм, на которых отложены одновременно частоты для двух групп (например, контрольной и экспериментальной). На рисунках 7 и 8 приведены две из них – позволяющие сравнивать контрольную и экспериментальную группу до начала и после окончания эксперимента (на самом деле визуальный анализ не дает возможности сказать, значимо ли различаются данные выборки – для этого необходимо использовать статистические методы – см. ниже раздел 6). Для их построения сначала перейдем от таблицы 5 к таблице 8, отличающейся от первой тем, что в ее ячейках стоят не абсолютное число членов той или иной группы, набравших соответствующий балл, а доля[1] (в процентах) членов группы, получивших данный балл, так как подобное преобразование (деление на одно и то же число – количество членов в данной группе) позволяет качественно сравнивать группы разных размеров (например, разного количества учащихся). Затем строим гистограммы в компьютерной программе Microsoft Excel для Windows (Меню/Вставка/Диаграмма) – см. рисунки 6 и 7, на которых по вертикали отложен процент членов той или иной группы, набравших соответствующий балл.

Таблица 8

Результаты измерений уровня знаний в контрольной и экспериментальной группах до и после эксперимента

Уровень знаний	**Контрольная группа до начала эксперимента (%)**	**Экспериментальная группа до начала эксперимента (%)**	**Контрольная группа после окончания эксперимента (%)**	**Экспериментальная группа после окончания эксперимента (%)**
Низкий	30,00	28,00	40,00	8,00
Средний	46,67	48,00	33,33	52,00
Высокий	23,33	24,00	26,67	40,00

[1] *Доля принимает значения от нуля до единицы. Для перехода к процентам следует долю умножить на 100%.*

Рис. 6. Гистограммы контрольной и экспериментальной групп до начала эксперимента

Рис. 7. Гистограммы контрольной и экспериментальной групп после окончания эксперимента

Таким образом, описательная статистика, во-первых, позволяет представить результаты педагогического эксперимента в компактном и информативном виде, что дает возможность проводить качественный анализ исследуемых объектов[1]. Во-вторых, ряд показателей описательной статистики используется в количественном анализе (при применении статистических критериев – см. разделы 6.3 и 6.4).

Завершив рассмотрение показателей описательной статистики, перейдем к общей методике определения степени достоверности совпадений и различий (следующий раздел), а затем опишем ее применение сначала для данных, измеренных в шкале отношений (раздел 6.3), а затем – для данных, измеренных в порядковой шкале (раздел 6.4).

6.2. ОБЩИЕ ПОДХОДЫ К ОПРЕДЕЛЕНИЮ ДОСТОВЕРНОСТИ СОВПАДЕНИЙ И РАЗЛИЧИЙ

В настоящем разделе рассмотрены общие подходы к определению достоверности совпадений и различий характеристик исследуемых объектов. Правило принятия решений относительно того, какой конкретный статистический критерий (метод обработки экспериментальных данных) следует использовать в том или ином случае, описано ниже в разделе 6.5 "Алгоритм выбора статистического критерия".

Как отмечалось выше, типовой задачей анализа данных в педагогических исследованиях является установление совпадений или различий характеристик экспериментальной и контрольной группы. Для этого формулируются *статистические гипотезы*:

- гипотеза об отсутствии различий (так называемая *нулевая гипотеза*);
- гипотеза о значимости различий (так называемая *альтернативная гипотеза*).

Для принятия решений о том, какую из гипотез (нулевую или альтернативную) следует принять, используют решающие правила

---

[1] *Показатели описательной статистики* (*объем выборки, среднее, гистограммы и т.д.*) *обычно приводятся в тексте диссертационных работ и авторефератов по педагогике.*

– *статистические критерии*[1]. То есть, на основании информации о результатах наблюдений (характеристиках членов экспериментальной и контрольной группы) вычисляется число, называемое *эмпирическим значением* критерия. Это число сравнивается с известным (например, заданным таблично) эталонным числом, называемым *критическим значением* критерия.

Критические значения приводятся, как правило, для нескольких *уровней значимости*. Уровнем значимости называется вероятность ошибки, заключающейся в отклонении (не принятии) нулевой гипотезы, то есть вероятность того, что различия сочтены существенными, а они на самом деле случайны. Обычно используют уровни значимости (обозначаемые α), равные 0,05, 0,01 и 0,001. В педагогических исследованиях обычно ограничиваются значением 0,05, то есть, грубо говоря, допускается не более чем 5% возможность ошибки.

Если полученное исследователем эмпирическое значение критерия оказывается меньше или равно критическому, то принимается нулевая гипотеза – считается, что на заданном уровне значимости (то есть при том значении α, для которого рассчитано критическое значение критерия) характеристики экспериментальной и контрольной групп совпадают. В противном случае, если эмпирическое значение критерия оказывается строго больше критического, то нулевая гипотеза отвергается и принимается альтернативная гипотеза – характеристики экспериментальной и контрольной группы считаются различными с достоверностью различий $1 - \alpha$. Например, если $\alpha = 0{,}05$ и принята альтернативная гипотеза, то *достоверность различий* равна 0,95 или 95%.

Другими словами, чем меньше эмпирическое значение критерия (чем левее оно находится от критического значения), тем больше степень совпадения характеристик сравниваемых объектов. И наоборот, чем больше эмпирическое значение критерия (чем правее оно находится от критического значения), тем сильнее различаются характеристики сравниваемых объектов.

---

[1] *Заметим, что в математической статистике исторически сложилось называть статистическими критериями не только решающие правила, но и методы расчета определенного числа* (*используемого в решающих правилах*)*, а также само это число.*

В дальнейшем мы ограничимся уровнем значимости $\alpha = 0{,}05$, поэтому, **если эмпирическое значение критерия оказывается меньше или равно критическому, то можно сделать вывод, что "характеристики экспериментальной и контрольной групп совпадают с уровнем значимости 0,05". Если эмпирическое значение критерия оказывается строго больше критического, то можно сделать вывод, что "достоверность различий характеристик экспериментальной и контрольной групп равна 95%".**

Опишем методики расчета эмпирических значений критериев для двух типовых задач анализа данных – сравнения выборок, содержащих данные, измеренные в шкале отношений (раздел 6.3) и порядковой шкале (раздел 6.4).

6.3. МЕТОДИКА ОПРЕДЕЛЕНИЯ ДОСТОВЕРНОСТИ СОВПАДЕНИЙ И РАЗЛИЧИЙ ДЛЯ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ, ИЗМЕРЕННЫХ В ШКАЛЕ ОТНОШЕНИЙ

Рассмотрим случай (см. описание исходных данных выше в пятом разделе), когда для измерений используется шкала отношений. Предположим, что имеется экспериментальная группа, состоящая из N человек, и контрольная группа, состоящая из M человек. Допустим, что в результате измерения одного и того же показателя с помощью одной и той же процедуры измерений были получены следующие данные: $x = (x_1, x_2, ..., x_N)$ – выборка для экспериментальной группы и $y = (y_1, y_2, ..., y_M)$ – выборка для контрольной группы, где x_i – элемент выборки – значение исследуемого показателя у i-го члена экспериментальной группы, $i = 1, 2, ..., N$, а y_j – значение исследуемого показателя у j-го члена контрольной группы, $j = 1, 2, ..., M$. Так как измерения производились в шкале отношений, то $\{x_i\}$ и $\{y_j\}$ – положительные, в том числе, возможно – целые, числа, для которых имеют смысл все арифметические операции. В качестве примера будем рассматривать результаты измерений уровня знаний в контрольной и экспериментальной группах до и после эксперимента (см. таблицу 2) – количество правильно решенных задач.

Для данных, измеренных в шкале отношений, для проверки гипотезы о совпадении характеристик двух групп целесообразно[1] использование либо *критерия*[2] *Крамера-Уэлча* [11, 22], либо *критерия Вилкоксона-Манна-Уитни* [2, 22, 32]. Критерий Крамера-Уэлча предназначен для проверки гипотезы о равенстве средних (строго говоря – математических ожиданий) двух выборок, критерий Вилкоксона-Манна-Уитни[3] является более "тонким" (но и более трудоемким) – он позволяет проверять гипотезу о том, что две выборки "одинаковы" (в том числе, что совпадают их средние, дисперсии и все другие показатели[4]).

Критерий Крамера-Уэлча. Эмпирическое значение данного критерия рассчитывается на основании информации об объемах N и M выборок x и y, выборочных средних $\bar{x}$ и $\bar{y}$ и выборочных дисперсиях D_x и D_y сравниваемых выборок (эти значения могут быть вычислены вручную по формулам (1)-(2) или с помощью инструмента "Описательная статистика" в компьютерной программе Microsoft Excel для Windows – см. раздел 6.1) по следующей формуле:

$$(3)\ T_{эмп} = \frac{\sqrt{M \cdot N}\ |\bar{x} - \bar{y}|}{\sqrt{M \cdot D_x + N \cdot D_y}}.$$

Алгоритм определения достоверности совпадений и различий характеристик сравниваемых выборок для экспериментальных данных, измеренных в шкале отношений, с помощью критерия Крамера-Уэлча заключается в следующем:

---

[1] Выбор критериев достаточно широк, в чем можно убедиться, ознакомившись с приведенными в списке литературы публикациями. Однако, нашей целью является описание статистических критериев, адекватных типовым для педагогических исследований задачам анализа данных.

[2] Критерий Крамера-Уэлча является более эффективным "заменителем" такого известного в физике и технике критерия как t-критерий (*критерий Стьюдента*) [22].

[3] Критерий Вилкоксона-Манна-Уитни плохо применим в условиях, когда число отличающихся друг от друга значений в выборках мало – см. ниже раздел 6.5.

[4] Две выборки могут иметь одинаковые средние (*то есть, критерий Крамера-Уэлча установит совпадение средних*), *но различаться, например, разбросом. Те различия, которые не выявит критерий Крамера-Уэлча, могут быть выявлены критерием Вилкоксона-Манна-Уитни.*

1. Вычислить для сравниваемых выборок $T_{эмп}$ – эмпирическое значение критерия Крамера-Уэлча по формуле (3).
2. Сравнить это значение с критическим значением $T_{0.05} = 1{,}96$: если $T_{эмп} \leq 1{,}96$, то сделать вывод: "характеристики сравниваемых выборок совпадают на уровне значимости 0,05"; если $T_{эмп} > 1{,}96$, то сделать вывод "достоверность различий характеристик сравниваемых выборок составляет[1] 95%".

В качестве примера применим алгоритм для данных из таблицы 2.

Для этого сравним сначала числа правильно решенных задач в контрольной и экспериментальной группе до начала эксперимента. Вычисляем[2] по формуле (3) значение $T_{эмп} = 0{,}04 \leq 1{,}96$. Следовательно гипотеза о совпадении характеристик контрольной и экспериментальной групп до начала эксперимента принимается на уровне значимости 0,05.

Теперь сравним характеристики контрольной и экспериментальной групп после окончания эксперимента. Вычисляем по формуле (3) значение $T_{эмп} = 2{,}42 > 1{,}96$. Следовательно, достоверность различий характеристик контрольной и экспериментальной групп после окончания эксперимента составляет 95%.

Итак, начальные (до начала эксперимента) состояния экспериментальной и контрольной групп совпадают, а конечные (после окончания эксперимента) – различаются. Следовательно, можно сделать вывод, что эффект изменений обусловлен именно применением экспериментальной методики обучения.

Отметим, что **мы не рассматриваем вопрос о том, "в какую сторону" экспериментальная группа отличается от контрольной, то есть, улучшились или ухудшились (с содержательной точки зрения, не имеющей отношения к статистическим методам и являющейся прерогативой педагогики) исследуемые характеристики**.

---

[1] Корректнее говорить, что достоверность различий составляет не менее 95%, однако, так мы условились считать достаточной 95%-ую достоверность различий, то будем говорить, что достоверность различий составляет 95%.

[2] Для сокращения ручных расчетов средние и дисперсии могут быть вычислены в рамках описательной статистики в компьютерной программе Microsoft Excel для Windows – см. выше таблицу 7.

Критерий Вилкоксона-Манна-Уитни[1]**.** Данный критерий оперирует не с абсолютными значениями элементов двух выборок, а с результатами их парных сравнений. Например, существенно, что учащийся Петров решил больше задач, чем учащийся Иванов, а на сколько больше – не важно.

Возьмем две выборки[2]: $\{x_i\}_{i=1...N}$ и $\{y_j\}_{j=1...M}$ и для каждого элемента первой[3] выборки x_i, $i = 1...N$, определим число a_i элементов второй выборки, которые превосходят его по своему значению (то есть число таких y_j, что $y_j > x_i$). Сумма $a_1 + a_2 + ... + a_N = \sum_{i=1}^{N} a_i$ этих чисел по всем N членам первой выборки называется *эмпирическим значением критерия Манна-Уитни* и обозначается $U = \sum_{i=1}^{N} a_i$.

Определим *эмпирическое значение критерия Вилкоксона*:

$$(4)\ W_{эмп} = \frac{\left|\frac{N \cdot M}{2} - U\right|}{\sqrt{\frac{N \cdot M \cdot (N + M + 1)}{12}}}.$$

Алгоритм определения достоверности совпадений и различий для экспериментальных данных, измеренных в шкале отношений, с помощью критерия Вилкоксона-Манна-Уитни заключается в следующем:

1. Вычислить для сравниваемых выборок $W_{эмп}$ – эмпирическое значение критерия Вилкоксона по формуле (4).
2. Сравнить это значение с критическим значением $W_{0.05} = 1{,}96$: если $W_{эмп} \le 1{,}96$, то сделать вывод: "характеристики сравниваемых выборок совпадают с уровнем значимости 0,05"; если $W_{эмп} > 1{,}96$, то сделать вывод "достоверность различий характеристик сравниваемых выборок составляет 95%".

---

[1] *Существуют два критерия – Вилкоксона и Манна-Уитни, однако, так как они однозначно связаны между собой, будем говорить об одном критерии Вилкоксона-Манна-Уитни* [*22*].

[2] *Ограничение на использование критерия Вилкоксона-Манна-Уитни следующее: каждая выборка должна содержать не менее трех элементов, если же в одной из выборок всего два элемента, то во второй их должно быть не менее пяти.*

[3] *Какую выборку считать первой, а какую второй, не имеет значения, хотя при вычислениях удобнее первой считать ту выборку, в которой меньше членов.*

В качестве примера применим алгоритм для данных из таблицы 2.

Для этого сравним сначала числа правильно решенных задач в контрольной и экспериментальной группе до начала эксперимента. В таблице 9 приведены результаты экспериментальной группы (второй столбец), и контрольной группы (пятый столбец), а также для каждого члена экспериментальной группы подсчитано число членов контрольной группы, решивших строго большее (чем он) число задач (третий столбец). Например, в таблице 9 серым цветом в пятом столбце помечены члены контрольной группы, правильно решившие строго большее число задач, чем первый член (то есть $i = 1$) экспериментальной группы, который правильно решил 12 задач. Значит $x_1 = 12$ и число таких y_j, что $y_j > x_1$ (то есть число затененных ячеек) равно 16. Следовательно, $a_1 = 16$. Аналогично заполняются остальные строки третьего столбца.

Таблица 9

Пример вычисления эмпирического значения критерия Манна-Уитни

Номер члена экспериментальной группы i	**Число задач, правильно решенных *i*-ым членом экспериментальной группы до начала эксперимента** x_i	**Число членов контрольной группы, правильно решивших строго большее число задач, чем *i*-ый член экспериментальной группы** a_i	**Номер члена контрольной группы** j	**Число задач, правильно решенных *j*-ым членом контрольной группы до начала эксперимента** y_j
1	12	16	1	15
2	11	18	2	13
3	15	7	3	11
4	17	5	4	18
5	18	3	5	10
6	6	30	6	8
7	8	25	7	20
8	10	21	8	7

Номер члена экспериментальной группы i	Число задач, правильно решенных i-ым членом экспериментальной группы до начала эксперимента x_i	Число членов контрольной группы, правильно решивших строго большее число задач, чем i-ый член экспериментальной группы a_i	Номер члена контрольной группы j	Число задач, правильно решенных j-ым членом контрольной группы до начала эксперимента y_j
9	16	5	9	8
10	12	16	10	12
11	15	8	11	15
12	14	11	12	16
13	19	1	13	13
14	13	13	14	14
15	19	1	15	14
16	12	16	16	19
17	11	18	17	7
18	16	5	18	8
19	12	16	19	11
20	8	25	20	12
21	13	13	21	15
22	7	26	22	16
23	15	7	23	13
24	8	23	24	5
25	9	22	25	11
–	–		26	19
–	–		27	18
–	–		28	9
–	–		29	6
–	–		30	15

Сумма всех 25 чисел в третьем столбце дает эмпирическое значение критерия Манна-Уитни $U = 351$. Вычисляем по формуле

(4) значение $W_{эмп} = 0{,}41 \le 1{,}96$. Следовательно, гипотеза о том, что сравниваемые выборки совпадают, принимается на уровне значимости 0,05.

Теперь аналогичным образом (построив таблицу, аналогичную таблице 9, и вычислив эмпирическое значение критерия Вилкоксона) сравним числа правильно решенных задач в контрольной и экспериментальной группе после окончания эксперимента. Эмпирическое значение критерия Манна-Уитни в этом случае равно 223. Вычисляем по формуле (4) значение $W_{эмп} = 2{,}57 > 1{,}96$. Следовательно, достоверность различий сравниваемых выборок составляет 95%.

Итак, начальные (до начала эксперимента) состояния экспериментальной и контрольной групп совпадают, а конечные (после окончания эксперимента) – различаются. Следовательно, можно сделать вывод, что эффект изменений обусловлен именно применением экспериментальной методики обучения.

6.4. МЕТОДИКА ОПРЕДЕЛЕНИЯ ДОСТОВЕРНОСТИ СОВПАДЕНИЙ И РАЗЛИЧИЙ ДЛЯ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ, ИЗМЕРЕННЫХ В ПОРЯДКОВОЙ ШКАЛЕ

Рассмотрим случай, когда используется порядковая шкала с L различным баллами. Характеристикой группы будет число ее членов, набравших тот или иной балл. Для экспериментальной группы вектор баллов есть $n = (n_1, n_2, ..., n_L)$, где n_k – число членов экспериментальной группы, получивших k-ый балл, $k = 1, 2, ..., L$. Для контрольной группы вектор баллов есть $m = (m_1, m_2, ..., m_L)$, где m_k – число членов контрольной группы, получивших k-ый балл, $k = 1, 2, ..., L$. Для рассматриваемого нами числового примера ($L = 3$ – "низкий", "средний" или "высокий" уровень знаний) данные приведены в таблице 5.

Для данных, измеренных в порядковой шкале (см., например, таблицу 5), целесообразно использование *критерия однородности* χ^2 ("хи" – буква греческого алфавита, название критерия читается:

"*хи-квадрат*") [27], эмпирическое значение $\chi^2_{эмп}$ которого вычисляется по следующей формуле[1] (пример расчета приведен ниже):

$$(5)\ \chi^2_{эмп} = N \cdot M \cdot \sum_{i=1}^{L} \frac{(\frac{n_i}{N} - \frac{m_i}{M})^2}{n_i + m_i}.$$

Критические значения $\chi^2_{0.05}$ критерия χ^2 для уровня значимости 0,05 приведены в таблице 10 (статистические таблицы критических значений статистических критериев для различных уровней значимости и различных – в том числе больших 10 – градаций шкалы отношений можно найти, практически, в любом учебнике по статистическим методам, или в специальных статистических таблицах [6]).

Таблица 10

Критические значения критерия χ^2 для уровня значимости $\alpha = 0.05$

L–1	1	2	3	4	5	6	7	8	9
$\chi^2_{0.05}$	3,84	5,99	7,82	9,49	11,07	12,59	14,07	15,52	16,92

Алгоритм определения достоверности совпадений и различий для экспериментальных данных, измеренных в порядковой шкале, заключается в следующем:

1. Вычислить для сравниваемых выборок $\chi^2_{эмп}$ – эмпирическое значение критерия χ^2 по формуле (5).
2. Сравнить это значение с критическим значением $\chi^2_{0.05}$, взятым из таблицы 10: если $\chi^2_{эмп} \le \chi^2_{0.05}$, то сделать вывод: "характеристик сравниваемых выборок совпадают с уровнем значимости 0,05"; если $\chi^2_{эмп} > \chi^2_{0.05}$, то сделать вывод "достоверность различий характеристик сравниваемых выборок составляет 95%".

---

[1] *Критерий хи-квадрат применим при условии, что для любого значения балла в любой из сравниваемых выборок не менее пяти ее членов получили данный балл, то есть*: $n_i \ge 5$, $m_i \ge 5$, $i = 1, 2, ..., L$. *Кроме того, желательно, чтобы число градаций L было не менее трех. Если L = 2, то есть используется дихотомическая шкала* (*"да" – "нет", "решил" – "не решил" и т.д.*), *то можно применять критерий Фишера – см. ниже настоящий раздел.*

Применим алгоритм для данных из таблицы 5. Сначала вычисляем по формуле (5) эмпирические значения критерия χ^2. Для примера приведем расчет. Параметры экспериментальной группы ($N = 25$) после окончания эксперимента: $n_1 = 2$, $n_2 = 13$, $n_3 = 10$ (то есть 2 учащихся продемонстрировали "низкий" уровень знаний, 13 – "средний" и 10 – "высокий" – см. выше таблицу 5), контрольной группы ($M = 30$): $m_1 = 12$, $m_2 = 10$, $m_3 = 8$. Подставляя в формулу (5), получаем:

$$\chi^2_{эмп} = 25 \cdot 30 \cdot [(\frac{2}{25} - \frac{12}{30})^2 / (2 + 12) + (\frac{13}{25} - \frac{10}{30})^2 / (13 + 10) + (\frac{10}{25} - \frac{8}{30})^2 / (10 + 8)] = 7{,}36.$$

Аналогичным образом вычисляются все оставшиеся из 16 возможных результатов парных сравнений групп (экспериментальная и контрольная группы, до начала и после окончания эксперимента). Результаты вычислений приведены в таблице 11. Ячейки таблицы 11 содержат эмпирические значения критерия χ^2 для сравниваемых групп, соответствующих строке и столбцу. Жирным шрифтом выделены результаты сравнения характеристик экспериментальной и контрольной группы до начала и после окончания эксперимента (см. сравнения I и II на рисунке 1 "Структура педагогического эксперимента"). Например, эмпирическое значение критерия χ^2, получаемое при сравнении характеристик контрольной группы до начала эксперимента (вторая строка таблицы 11) и экспериментальной группы до начала эксперимента (третий столбец таблицы 11), равно 0,03

В рассматриваемом примере $L = 3$ (выделены три уровня знаний – "низкий", "средний" и "высокий"). Следовательно, $L - 1 = 2$. Из таблицы 10 получаем для $L - 1 = 2$: $\chi^2_{0.05} = 5{,}99$. Тогда из таблицы 11 видно, что все эмпирические значения критерия χ^2, кроме результата $\chi_{эмп} = 7{,}36$ сравнения экспериментальной и контрольной групп после окончания эксперимента, меньше критического значения.

Таблица 11

Эмпирические значения критерия χ^2 для данных из таблицы 5

	Контрольная группа до начала эксперимента	**Экспериментальная группа до начала эксперимента**	**Контрольная группа после окончания эксперимента**	**Экспериментальная группа после окончания эксперимента**
Контрольная группа до начала эксперимента	0	**0,03**	1,16	4,60
Экспериментальная группа до начала эксперимента	0,03	0	1,34	3,82
Контрольная группа после окончания эксперимента	1,16	1,34	0	**7,36**
Экспериментальная группа после окончания эксперимента	4,60	3,82	7,36	0

Следовательно "характеристики всех сравниваемых выборок, кроме экспериментальной и контрольной групп после окончания эксперимента, совпадают[1] с уровнем значимости 0,05".

Так как $\chi_{эмп} = 7{,}36 > 5{,}99 = \chi^2_{0.05}$, то "достоверность различий характеристик экспериментальной и контрольной групп после окончания эксперимента составляет 95%".

Итак, начальные (до начала эксперимента) состояния экспериментальной и контрольной групп совпадают, а конечные (после окончания эксперимента) – различаются. Следовательно, можно сделать вывод, что эффект изменений обусловлен именно применением экспериментальной методики обучения.

Дихотомическая шкала. Отдельно рассмотрим случай, когда используется дихотомическая шкала – порядковая шкала с всего двумя различными упорядоченными баллами – "высокий"-

[1] *Интересно отметить, что характеристики экспериментальной группы до начала и после окончания эксперимента также совпадают с уровнем значимости 0,05.*

"низкий", "справился с заданием"-"не справился", "прошел тест"-"не прошел" и т.д. Характеристикой группы, помимо общего числа ее членов, будет число членов (или доля, процент от общего числа), набравших заданный, например – максимальный, балл (в общем случае – число членов, обладающих заданным признаком).

Для экспериментальной группы, описываемой двумя числами (n_1, n_2), где n_1 – число членов рассматриваемой группы, набравших низкий балл, n_2 – набравших высокий балл, $n_1 + n_2 = N$, доля p ее членов, набравших максимальный балл, равна: $p = n_2 / N$. Для контрольной группы, описываемой двумя числами (m_1, m_2), где $m_1 + m_2 = M$, доля q ее членов, набравших максимальный балл, равна: $q = m_2 / M$.

Рассмотрим пример: для каждого из столбцов таблицы 2, считая, что возможны два уровня знаний – "не усвоили материал" (число правильно решенных задач меньше либо равно 10) и "успешно усвоили материал" (число правильно решенных задач строго больше 10) определяем распределение членов экспериментальной и контрольной группы по двум уровням знаний и получаем таблицу 12 (для экспериментальной группы до начала эксперимента p = 0,72 (или 72%), после окончания эксперимента p = 0,92; для контрольной группы до начала эксперимента q = 0,70, после окончания эксперимента q = 0,60).

Таблица 12

Результаты дихотомических измерений уровня знаний в контрольной и экспериментальной группах до и после эксперимента

	Контрольная группа до начала эксперимента	**Экспериментальная группа до начала эксперимента**	**Контрольная группа после окончания эксперимента**	**Экспериментальная группа после окончания эксперимента**
Доля, которую составляют учащиеся, не усвоившие материал	0,30	0,28	0,40	0,08
Доля, которую составляют учащиеся, усвоившие материал	0,70	0,72	0,60	0,92

Для данных, измеренных в дихотомической шкале целесообразно использование *критерия Фишера*[1], для которого эмпирическое значение $\varphi_{эмп}$ вычисляется по следующей формуле (арксинус может быть вычислен в Excel):

(6) $\varphi_{эмп} = |2\ arcsin(\sqrt{p}\ ) - 2\ arcsin(\sqrt{q}\ )|\ \sqrt{\frac{M \cdot N}{M + N}}$.

Критическое значение $\varphi_{0.05}$ критерия Фишера для уровня значимости 0,05 равно 1,64.

Алгоритм определения достоверности совпадений и различий для экспериментальных данных, измеренных в дихотомической шкале, заключается в следующем:

1. Вычислить для сравниваемых выборок $\varphi_{эмп}$ – эмпирическое значение критерия Фишера по формуле (6).
2. Сравнить это значение с критическим значением $\varphi_{0.05} = 1{,}64$: если $\varphi_{эмп} \leq 1{,}64$, то сделать вывод: "характеристики сравниваемых выборок совпадают с уровнем значимости 0,05"; если $\varphi_{эмп} > 1{,}64$, то сделать вывод "достоверность различий характеристик сравниваемых выборок составляет 95%".

Применим алгоритм для экспериментальных данных из таблицы 12. Сначала вычисляем по формуле (2) эмпирические значения критерия Фишера. Для примера приведем расчет. Параметры экспериментальной группы ($N = 25$) после окончания эксперимента: $p = 0{,}92$, контрольной группы ($M = 30$): $q = 0{,}60$ (см. таблицу 12). Подставляя в формулу (6), получаем:

$$\varphi_{эмп} = |2\ arcsin(\sqrt{0{,}92}\ ) - 2\ arcsin(\sqrt{0{,}6}\ )|\ \sqrt{\frac{25 \cdot 30}{25 + 30}} = 2{,}94.$$

Аналогичным образом вычисляются все оставшиеся из 16 возможных результатов парных сравнений групп (экспериментальная и контрольная группы, до начала и после окончания эксперимента). Результаты вычислений приведены в таблице 13. Ячейки таблицы 13 содержат эмпирические значения критерия Фишера для сравниваемых групп, соответствующих строке и столбцу.

[1] *В математической статистике существует несколько критериев Фишера. Мы используем один из них – так называемое угловое преобразование, поэтому далее под критерием Фишера будем понимать именно угловое преобразование Фишера.*

Жирным шрифтом выделены результаты сравнения характеристик экспериментальной и контрольной группы до начала и после окончания эксперимента (см. сравнения I и II на рисунке 1 "Структура педагогического эксперимента").

Например, эмпирическое значение критерия Фишера, получаемое при сравнении характеристик контрольной группы до начала эксперимента (вторая строка таблицы 13) и экспериментальной группы до начала эксперимента (третий столбец таблицы 13), равно 0,16. Следовательно "состояния экспериментальной и контрольной групп до начала эксперимента совпадают с уровнем значимости 0,05".

Таблица 13

Эмпирические значения критерия Фишера
для данных из таблицы 12

	Контрольная группа до начала эксперимента	**Экспериментальная группа до начала эксперимента**	**Контрольная группа после окончания эксперимента**	**Экспериментальная группа после окончания эксперимента**
Контрольная группа до начала эксперимента	0	**0,16**	0,81	2,16
Экспериментальная группа до начала эксперимента	0,16	0	0,94	1,92
Контрольная группа после окончания эксперимента	0,81	0,94	0	**2,94**
Экспериментальная группа после окончания эксперимента	2,16	1,92	2,94	0

Теперь аналогичным образом сравним характеристики экспериментальной и контрольной групп после окончания эксперимента. Так как $\varphi_{эмп} = 2{,}94 > 1{,}64 = \varphi_{кр}$, то "достоверность различий состояний экспериментальной и контрольной групп после окончания эксперимента составляет 95%".

Итак, начальные (до начала эксперимента) состояния экспериментальной и контрольной групп совпадают, а конечные (после окончания эксперимента) – различаются. Следовательно, можно сделать вывод, что эффект изменений обусловлен именно применением экспериментальной методики обучения. Отметим, данный вывод (один и тот же) был получен при применении к соответствующим экспериментальным данным всех четырех критериев – Крамера-Уэлча, Вилкоксона-Манна-Уитни, χ^2 и Фишера[1].

6.5. АЛГОРИТМ ВЫБОРА СТАТИСТИЧЕСКОГО КРИТЕРИЯ

Завершив описание методик анализа данных, поясним, как следует выбирать статистические критерии, то есть приведем алгоритм выбора статистического критерия – процедуру принятия решения относительно того, какой статистический критерий использовать в той или иной ситуации.

В первом приближении этот алгоритм чрезвычайно прост: **если данные получены в результате измерений в шкале отношений, то следует использовать критерий Вилкоксона-Манна-Уитни** (ВМУ)**, если в порядковой шкале, то критерий χ^2.**

Возможные модификации этого правила принятия решений (учитывающие большее число факторов) приведены на рисунке 8.

---

[1] *Перечисленные четыре критерия обладают различной "мощностью" – возможны случаи, когда, например, применение критерия Крамера-Уэлча или критерия Вилкоксона-Манна-Уитни к данным, измеренным в шкале отношений, свидетельствует о наличии статистически значимых различий, а применение критерия χ^2 к тем же эмпирическим результатам, переведенным в порядковую шкалу, свидетельствует о совпадении характеристик* (*см. также обсуждение потерь информации при переходе от шкалы отношений к порядковой шкале выше в пятом разделе*). *Поэтому можно рекомендовать максимально использовать всю полученную в результате педагогического эксперимента информацию – если измерения проводились в шкале отношений, то и обрабатывать данные следует в этой шкале, переходя к порядковой шкале только в случае крайней необходимости* (*см. рисунок 8*).

Рис. 8. Алгоритм выбора статистического критерия

Алгоритм выбора статистического критерия.

Во-первых, необходимо определить какая шкала измерений используется – отношений или порядковая.

Для шкалы отношений следует решить, состоит ли решаемая задача в обнаружении различия средних значений (математических ожиданий). Если – да, то можно использовать критерий Крамера-Уэлча (раздел 6.3). Если же следует обнаружить произвольные различия характеристик выборок, то следует использовать критерий Вилкоксона-Манна-Уитни (раздел 6.3) или критерий χ^2 (раздел 6.4).

Если число различающихся между собой значений[1] в сравниваемых выборках велико (более десяти), то целесообразно использование критерия Вилкоксона-Манна-Уитни.

Если число различающихся между собой значений в сравниваемых выборках мало (менее десяти), то, произведя группировку результатов измерений (то есть, перейдя от шкалы отношений к порядковой шкале – см. выше пятый раздел), можно использовать критерий χ^2.

Далее, аналогично рассуждая, если объем выборок мал[2] ($N, M \leq 50$), то следует использовать критерий Вилкоксона-Манна-Уитни (при малом числе различающихся значений в этом случае можно использовать и критерий χ^2).

Если объем выборок велик, то, опять же с помощью группировки результатов измерений имеет смысл использовать критерий χ^2.

Для порядковой шкалы в случае, когда число градаций (различных баллов) больше либо равно трем, используется критерий χ^2, если же применялась дихотомическая шкала, то можно использовать либо критерий χ^2, либо критерий Фишера – см. раздел 6.4.

Использование компьютера при анализе результатов педагогических экспериментов, несомненно, целесообразно. Однако, использовать статистические критерии, "зашитые" в пакеты программ следует осторожно. Все четыре описанных выше статистических критерия (Крамера-Уэлча, Вилкоксона-Манна-Уитни, χ^2 и Фишера) корректно реализованы в профессиональных статистических пакетах, среди которых можно выделить и рекомендовать к использованию такие наиболее распространенные пакеты статистического анализа как: Statistica, StatGraphics и SPSS. Однако, упомянутые программы, во-первых, являются лицензионными и стоят достаточно дорого. Во-вторых, они достаточно сложны и требуют значительных временных затрат для своего освоения. Наряду с этим, существуют инструменты статистического анализа

---

[1] *Например, выборка* (*1, 2, 2, 2, 1, 1, 2, 1, 1, 1*) *содержит всего два различных значения – единицу и двойку. В то же время, например, выборка* (*2, 0, 1, 5, 8, 4, 2, 7, 3, 9*) *того же объема* (*десять элементов*) *содержит десять различных значений.*

[2] *Понятно, что приводимые границы числа различающихся между собой значений – 10, и объема выборок – 50, примерны, приблизительны.*

в электронных таблицах Microsoft Excel, входящих в стандартный комплект Microsoft Office и установленных, наверное, на любом современном компьютере. Однако, к сожалению, ни один из четырех рекомендуемых статистических критериев не реализован в Excel[1], поэтому можно посоветовать производить расчет эмпирических значений критериев вручную[2] (все необходимые формулы приведены выше), используя компьютер или калькулятор для получения описательной статистики и автоматизации расчетов.

Планирование педагогического эксперимента. В заключение настоящего раздела отметим, что, несмотря на то, что выше обсуждалось применение статистических методов к уже полученным в результате проведения педагогического эксперимента данным, знание этих методов позволяет планировать эксперимент на стадии его подготовки. Например, формулы (3)-(6), определяющие эмпирические значения критериев, совместно с фиксированными критическими их значениями, позволяют заранее (до проведения эксперимента) оценивать необходимый объем выборки и другие важные параметры[3]. Кроме того, если до начала эксперимента выявлено статистически значимое различие характеристик экспериментальной и контрольной групп по интересующему исследователя критерию (например, по успеваемости), то проводить эксперимент не имеет смысла, так как никакие результаты сравнения характеристик этих групп после окончания эксперимента, не позволят выявить вклада сравниваемого с традиционным педагогического воздействия.

---

[1] В компьютерной программе Microsoft Excel для Windows имеется критерий согласия χ^2, отличающийся от описанного выше критерия однородности χ^2, поэтому применение первого может привести к неверным результатам.

[2] Альтернативой является использование дополнительных статистических надстроек к Excel – Megastat, XLStat, которые можно найти в свободном доступе в Интернете. В этих пакетах хорошо представлены непараметрические методы – критерий Вилкоксона-Манна-Уитни и другие.

[3] Конечно, чем больше объемы выборок, тем в некотором смысле лучше, то есть тем проще будет обосновать различия, если они есть. Но, с другой стороны, привлечение к педагогическому эксперименту каждого нового участника требует от исследователя определенных усилий, поэтому целесообразно заранее примерно определить требуемый объем выборок.

7. ЗАКЛЮЧЕНИЕ

Как отмечалось выше (см. раздел 2 – "Структура педагогического эксперимента"), целью любого педагогического эксперимента является эмпирическое подтверждение или опровержение гипотезы исследования и/или справедливости теоретических результатов, то есть обоснование того, что предлагаемое педагогическое воздействие (например, новые содержание, формы, методы, средства обучения и т.д.) более эффективно (или, возможно, наоборот – менее эффективно). Для этого, как минимум, необходимо показать, что, будучи примененным к тому же объекту (например – к группе учащихся), оно дает другие результаты, чем применение традиционных педагогических воздействий.

Для этого выделяется экспериментальная группа, которая сравнивается с контрольной группой. Различие эффектов педагогических воздействий будет обосновано, если две эти группы, первоначально совпадающие по своим характеристикам, различаются после реализации педагогических воздействий. Следовательно, требуется провести два сравнения и показать, что при первом сравнении (до начала педагогического эксперимента) характеристики экспериментальной и контрольной группы совпадают, а при втором (после окончания эксперимента) – различаются.

Так как объектом педагогического эксперимента, как правило, являются люди (учащиеся, учителя, сотрудники и руководители органов управления образованием и т.д.), а каждый человек индивидуален, то говорить о совпадении или различии характеристик экспериментальной и контрольной групп можно лишь в чисто формальном, статистическом смысле. Для того, чтобы выяснить, являются ли совпадения или различия случайными, используются статистические методы, которые позволяют на основании данных, полученных в результате эксперимента, принять обоснованное решение о совпадениях или различиях.

Общий алгоритм использования статистических критериев прост: до начала и после окончания эксперимента на основании информации о результатах наблюдений (характеристиках членов экспериментальной и контрольной группы) вычисляется эмпирическое значение критерия (алгоритм выбора статистического критерия приведен выше в разделе 6.5, формулы для вычислений – в

разделах 6.3 и 6.4). Это число сравнивается с известным (табличным) числом – критическим значением критерия (критические значения[1] для всех рекомендуемых нами критериев приведены выше в разделах 6.3 и 6.4). Если эмпирическое значение критерия оказывается меньше или равно критическому, то можно утверждать, что "**характеристики экспериментальной и контрольной групп совпадают с уровнем значимости 0,05 по статистическому критерию …** (далее следует название использованного критерия: Крамера-Уэлча, Вилкоксона-Манна-Уитни, хи-квадрат, Фишера)". В противном случае (если эмпирическое значение критерия оказывается строго больше критического) можно утверждать, что "**достоверность различий характеристик экспериментальной и контрольной групп по статистическому критерию … равна 95%**".

Следовательно, если характеристики экспериментальной и контрольной групп до начала эксперимента совпадают с уровнем значимости 0,05, и, одновременно с этим, достоверность различий характеристик экспериментальной и контрольной групп после эксперимента равна 95%, то можно сделать вывод, что[2] "**применение предлагаемого педагогического воздействия** (например, новой методики обучения) **приводит к статистически значимым (на уровне 95% по критерию …) отличиям результатов**".

Итак, в настоящей работе мы попытались изложить на доступном уровне "рецепты" применения статистических методов при решении типовых задач анализа данных в педагогических исследованиях. В то же время, не следует забывать, что рассмотрены лишь несколько, хотя и наиболее распространенных, но все-таки достаточно простых ситуаций. Арсенал же современных статистических методов гораздо богаче. Быть может, освоение и применение этого арсенала подтолкнет исследователей в области педагогических наук как к расширению соответствующих предметных областей, так и к повышению уровня обоснованности научных результатов.

---

[1] Напомним, что выше мы решили ограничиться 0,05 уровнем значимости и, соответственно, 95%-ым уровнем достоверности различий.

[2] Понятно, что в каждом конкретном случае общие термины "характеристика группы", "педагогическое воздействие", "результат" заменяются на конкретные характеристики, воздействия и результаты.

ЛИТЕРАТУРА

1. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: основы моделирования и первичная обработка данных. М.: Финансы и статистика, 1983. – 472 с.

2. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. М.: ЮНИТИ, 1998. – 1022 с.

3. Айвазян С.А., Мхитарян В.С. Прикладная статистика в задачах и упражнениях. М.: ЮНИТИ, 2001. – 270 с.

4. Анализ нечисловой информации в социологических исследованиях. М.: Наука, 1985. – 220 с.

5. Артемьева Е.Ю., Мартынов Е.М. Вероятностные методы в психологии. М.: МГУ, 1975.

6. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. М.: Наука, 1983. – 416 с.

7. Бурков В.Н., Новиков Д.А. Как управлять организациями. М.: Синтег, 2004. – 404 с.

8. Грабарь М.И., Краснянская К.А. Применение математической статистики в педагогических исследованиях: Непараметрические методы. М.: Педагогика, 1977. – 136 с.

9. Гласс Д., Стенли Д. Статистические методы в педагогике и психологии. М.: Прогресс, 1976. – 495 с.

10. Ительсон Л.Б. Математические и кибернетические методы в педагогике. М.: Просвещение, 1964. – 268 с.

11. Крамер Г. Математические методы статистики. М.: Мир, 1975. – 648 с.

12. Кыверялг А.А. Методы исследований в профессиональной педагогике. Таллин: Валгус, 1980. – 334 с.

13. Литвак Б.Г. Экспертная информация: методы получения и анализа. М.: Радио и связь, 1982. – 184 с.

14. Новиков А.М. Докторская диссертация? М.: Эгвес, 2003. – 120 с.

15. Новиков А.М. Как работать над диссертацией. М.: Эгвес, 2003. – 104 с.

16. Новиков А.М. Методология образования. М.: Эгвес, 2002. – 320 с.

17. Новиков А.М. Научно-экспериментальная работа в образовательном учреждении. М.: АПО РАО, 1998. – 134 с.

18. Новиков Д.А. Закономерности итеративного научения. М.: ИПУ РАН, 1998 – 96 с.

19. Новиков Д.А. Модели и механизмы управления развитием региональных образовательных систем. М.: ИПУ РАН, 2001. – 83 с.

20. Ногин В.Д. Принятие решений в многокритериальной среде: количественный подход. М.: Физматлит, 2002. – 176 с.

21. Орлов А.И. Устойчивость в социально-экономических моделях. М.: Наука, 1986. – 294 с.

22. Орлов А.И. Эконометрика. М.: Экзамен, 2003. – 576 с.

23. Паповян С.С. Математические методы в социальной психологии. М.: Наука, 1983.

24. Подиновский В.В., Ногин В.Д. Парето-оптимальные решения многокритериальных задач. М.: Наука, 1982. – 386 с.

25. Пфанцагль И. Теория измерений. М.: Мир, 1976. – 248 с.

26. Сидоренко Е.В. Методы математической обработки в психологии. СПб.: Речь, 2000. – 350 с.

27. Смирнов Н.В., Дунин-Барковский И.В. Курс теории вероятностей и математической статистики для технических приложений. М.: Наука, 1969.

28. Справочник по прикладной статистике. М.: Финансы и статистика. Том 1, 1989. – 510 с., Том 2, 1990. – 526 с.

29. Суппес П., Зинес Д. Основы теории измерений / Психологические измерения. М.: Мир, 1967. С. 9 – 110.

30. Суходольский Г.В. Основы математической статистики для психологов. Л.: ЛГУ, 1972. – 428 с.

31. Трахтенгерц Э.А. Компьютерная поддержка принятия решений. М.: Синтег, 1998. – 376 с.

32. Тюрин Ю.Н., Макаров А.А. Статистический анализ данных на компьютере. М.: ИНФРА-М, 1998. – 528 с.

33. Тюрин Ю.Н., Литвак Б.Г., Орлов А.И., Сатаров Г.А., Шмерлинг Д.С. Анализ нечисловой информации. М.: Научный совет АН СССР по комплексной проблеме "Кибернетика", 1981. – 80 с.

***Редакционный совет серии "Статистические методы"*:**
Богданов Ю.И.
Вощинин А.П.
Горбачев О.Г.
Горский В.Г.
Кудлаев Э.М.
Натан А.А.
Новиков Д.А.
Орлов А.И. (председатель).
Татарова Г.Г.
Толстова Ю.Н.
Фалько С.Г.
Шведовский В.А.

Уважаемые читатели!

Предлагаемая книга входит в новую серию «Статистические методы» издательства «МЗ-Пресс». В этой серии будут выпускаться научные монографии по различным теоретическим и прикладным направлениям статистических методов, учебники и учебные пособия, написанные ведущими исследователями. Основная цель серии – выпуск научных монографий, являющихся одновременно учебниками и позволяющих студентам и специалистам выйти на передовой фронт современных исследований.

Книги серии посвящены прикладной статистике и другим статистическим методам обработки и анализа данных, а также применению статистических методов в технических, социально-экономических, медицинских, исторических и иных исследованиях. Они окажутся полезными для инженеров, экономистов, менеджеров, социологов, врачей, всех научных работников и специалистов, чья профессиональная деятельность связаны с обработкой и анализом данных.

Редакционный совет серии создан Правлением Российской ассоциации статистических методов (учреждена в 1990 г.). По оценке Правления, выпуск серии «Статистические методы» позволит заметно повысить научный уровень и практическую значимость отечественных научных исследований, прикладных разработок и преподавания в области статистических методов.

Надеемся, что новая серия привлечет внимание и будет полезна как студентов и преподавателям, так и профессиональным исследователям. Желаем всем потенциальным читателям найти что-то полезное для себя.

Дмитрий Александрович НОВИКОВ

СТАТИСТИЧЕСКИЕ МЕТОДЫ В ПЕДАГОГИЧЕСКИХ ИССЛЕДОВАНИЯХ (ТИПОВЫЕ СЛУЧАИ)

Подписано в печать 29.01.2004
Формат 69x90/16. Печать офсетная. Бумага офсетная.
Гарнитура "Таймс". Усл. печ. л. 4,25.
Тираж 3000 экз.

Отпечатано с готовых пленок в ДПК
г. Домодедово, Каширское шоссе, д. 4