Д.А. Новиков

СТАТИСТИЧЕСКИЕ МЕТОДЫ В ПЕДАГОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

Настоящая работа представляет собой сжатое изложение одноименной книги – ***Новиков Д.А. «Статистические методы в педагогических исследованиях (типовые случаи)». М.: МЗ-Пресс, 2004. – 67 с.*** *– и содержит «рецепты» применения статистических методов в типовых случаях анализа экспериментальных данных в педагогических исследованиях.*

Книгу можно заказать в интернет-магазинах, или на сайте издательства МЗ-Пресс http://mzpress.narod.ru

СОДЕРЖАНИЕ КНИГИ

Приводится алгоритм выбора статистического критерия, методики определения достоверности совпадений и различий характеристик исследуемых объектов. Изложение сопровождается примерами анализа результатов педагогических экспериментов. Работа рассчитана на педагогов-исследователей, в первую очередь, на аспирантов и соискателей.

Рассматриваемые ниже инструменты анализа данных имеют программную реализацию в виде свободно распространяемой ***компьютерной программы "Статистика в педагогике"****, которую можно загрузить с адреса* http://www.mtas.ru/uploads/stat.zip *(2.9 Мб).*

Введение. 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 и формулы для вычислений приведены ниже). Это число сравнивается с известным (табличным) числом – критическим значением критерия (критические значения[1] для всех рекомендуемых нами критериев приведены ниже). 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**".

<u>Типовые задачи анализа данных в педагогических исследованиях.</u> 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*[3] для экспериментальной группы и $y = (y_1, y_2, \ldots, y_M)$ – выборка для контрольной группы, где x_i – *элемент выборки* – значение исследуемого показателя (признака[4]) у i-го члена экспериментальной группы, $i = 1, 2, \ldots, N$, а y_j – значение исследуемого показателя у j-го члена контрольной группы, $j = 1, 2, \ldots, M$. 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*[5] или *порядковой шкале*[6] – производились измерения, получаем следующие <u>два случая</u>.

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[7]. 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 1, строки кото-

---

[1] *Ниже мы ограничимся 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.*

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

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

[5] *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. Для шкалы отношений допустимым является преобразования подобия – умножение на положительное число.*

[6] *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). Для порядковой шкалы допустимым является любое монотонное преобразование.*

[7] *Данный пример рассматривается на протяжении всей работы. Все таблицы, диаграммы и графики экспортированы из компьютерной программы Microsoft Excel для Windows.*

рой 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.

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

Таблица 1

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	–

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

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 2, в которой указаны верхние границы диапазонов.

Таблица 2

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. Вычислим на основании данных таблицы 1, 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[8] занесем в таблицу 3.

Таблица 3

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 1 по аналогии с таблицей 3 определяем распределение членов экспериментальной и контрольной групп по уровням знаний и получаем таблицу 4.

Таблица 4

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

Таблица 4 построена по таблице 1 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 4. Поэтому для задач анализа результатов измерений, произведенных в шкале отношений, будем считать, что данные эксперимента имеют вид таблицы 1, а для задач анализа результатов измерений, произведенных в шкале порядка, будем считать, что данные эксперимента имеют вид таблицы 4.

---

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

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

- *описание данных* (компактное и информативное отражение результатов измерений характеристик исследуемых объектов);
- *установление совпадения* характеристик двух групп;
- *установление различия* характеристик двух групп (например, экспериментальной и контрольной).

Два типа шкал (отношений и порядка) и три перечисленные типа задач анализа данных позволяют выделить шесть базовых (типовых) задач, приведенных в таблице 5 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 5

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

Перечисленные шесть задач являются **базовыми** по следующим причинам. Во-первых, они включают большинство (порядка 90 %) задач анализа данных, встречающихся в экспериментальных исследованиях по педагогическим наукам. Во-вторых, они сформулированы для простейшей схемы организации педагогического эксперимента – 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), то описание и сравнение экспериментальной и контрольной групп[9] 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 5).

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, ознакомиться с ними можно в публикациях, перечисленных в упомянутой во введении работе автора.

Рассмотрим методы решения типовых для педагогических исследований задач анализа данных.

<u>Алгоритм выбора статистического критерия.</u> 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** c^2**.**

Возможные модификации этого правила принятия решений (учитывающие большее число факторов) приведены на рисунке 1.

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

<u>Для шкалы отношений</u> следует решить, состоит ли решаемая задача в обнаружении различия средних значений (математических ожиданий). Если – да, то можно использовать критерий Крамера-Уэлча (описание методик применения всех упоминаемых статистических критериев приведено ниже). Если же следует обнаружить произвольные различия характеристик выборок, то следует использовать критерий Вилкоксона-Манна-Уитни или критерий χ^2.

Если число различающихся между собой значений[10] в сравниваемых выборках велико (более десяти), то целесообразно использование критерия Вилкоксона-Манна-Уитни.

---

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

[10] *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.*

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[11] ($N, M \leq 50$), 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.

Рис. 1. Алгоритм выбора статистического критерия

<u>Для порядковой шкалы</u> 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.

Использование компьютера при анализе результатов педагогических экспериментов, несомненно, целесообразно. Однако использовать статистические критерии, "зашитые" в пакеты программ следует осторожно. Все четыре предлагаемых к использованию для обработки результатов педагогического эксперимента статистических критерия (Крамера-Уэлча, Вилкоксона-Манна-Уитни, χ^2 и Фишера) корректно реализованы в профессиональных статистических пакетах, среди которых можно выделить и рекомендовать 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 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[12], поэтому можно посоветовать производить расчет

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

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

эмпирических значений критериев вручную[13] (все необходимые формулы приведены ниже), используя компьютер или калькулятор для получения описательной статистики и автоматизации расчетов. Альтернативой является использование **компьютерной программы "Статистика в педагогике"**, которую можно загрузить с адреса http://www.mtas.ru/uploads/stat.zip (2.9 Мб).

Методы обработки данных и примеры. Приведем методики анализа данных для выделенных выше шести типовых задач (см. таблицу 5): описательная статистика, анализ совпадений и различий характеристик экспериментальной и контрольной групп на основании измерений, проведенных в порядковой шкале или шкале отношений. В качестве иллюстрации рассмотрим реализацию этих методик для числового примера (см. таблицы 1 и 4).

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

Для результатов измерений в шкале отношений (задача 1.1 – см. таблицу 5) показатели описательной статистики можно разбить на несколько групп:

- *показатели положения* описывают положение экспериментальных данных на числовой оси. Примеры таких данных – *максимальный и минимальный элементы выборки*, *среднее значение*[14], *медиана*[15], *мода*[16] и др.;

- *показатели разброса* описывают степень разброса данных относительно своего центра (среднего значения). К ним относятся: *выборочная дисперсия*[17], *разность между минимальным и максимальным элементами* (*размах*, *интервал* выборки) и др.

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

- *гистограмма*[18] и др.

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

Приведем формулы расчета основных показателей. *Среднее арифметическое* $\bar{x}$ выборки $\{x_i\}_{i=1 \ldots N}$ (*выборочное среднее*) рассчитывается следующим образом[19]:

(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 1 (числа правильно решенных задач в контрольной группе до начала эксперимента) приведена в таблице 6.

---

[13] *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.*

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

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

[16] *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.*

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

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

[19] *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: $x_1 + x_2 + \ldots + x_n$.

Таблица 6

Описательная статистика числа правильно решенных задач в контрольной группе до начала эксперимента (см. первый столбец таблицы 1)

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

Целый ряд приведенных в таблице 6 показателей описательной статистики педагогу-исследователю не понадобятся (далее используются только среднее (формула (1), первая строка таблицы 6), дисперсия (формула (2), шестая строка таблицы 6) и "счет" (объем выборки) – последняя строка таблицы 6). Тем не менее, мы приводим все показатели, которые автоматически выводит "Описательная статистика" в компьютерной программе Microsoft Excel для Windows (таблица 6 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 1) представлена на рисунке 2.

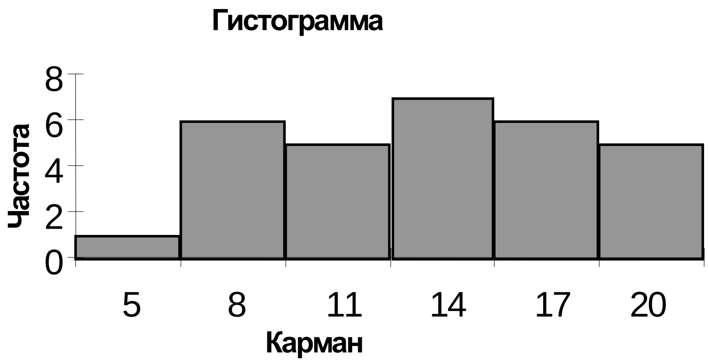

Рис. 2. Гистограмма числа правильно решенных задач в контрольной группе до начала эксперимента ("частота" – число элементов выборки, попавших в заданный диапазон, называемый в Excel "карманом")

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

Для результатов измерений в порядковой шкале (задача 2.1 – см. таблицу 5) при небольшом числе градаций единственным информативным показателем описательной статистики является гистограмма[20].

Для визуального (качественного) сравнения экспериментальной и контрольной групп удобно строить для них совместные гистограммы. Например, по результатам таблицы 4 (см. выше) можно построить несколько парных гистограмм, на которых отложены одновременно частоты для двух групп (например, контрольной и экспериментальной). На рисунках 3 и 4 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). Для их построения сначала перейдем от таблицы 4 к таблице 7, отличающейся от первой тем, что в ее ячейках стоят не абсолютное

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

число членов той или иной группы, набравших соответствующий балл, а доля[21] (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 3 и 4, на которых по вертикали отложен процент членов той или иной группы, набравших соответствующий балл.

Таблица 7

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

Таким образом, описательная статистика, во-первых, позволяет представить результаты педагогического эксперимента в компактном и информативном виде, что дает возможность проводить качественный анализ исследуемых объектов[22]. Во-вторых, ряд показателей описательной статистики используется в количественном анализе (при применении статистических критериев – см. ниже).

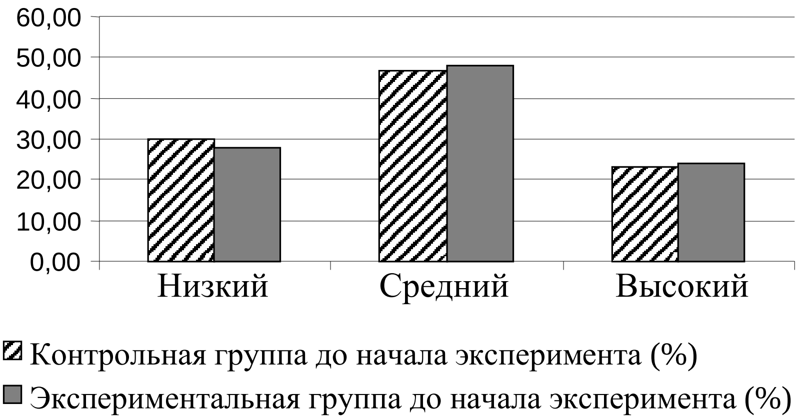

Рис. 3. Гистограммы контрольной и экспериментальной групп до начала эксперимента

Рис. 4. Гистограммы контрольной и экспериментальной групп после окончания эксперимента

[21] *Доля принимает значения от нуля до единицы. Для перехода к процентам следует долю умножить на 100%.*
[22] *Показатели описательной статистики (объем выборки, среднее, гистограммы и т.д.) обычно приводятся в тексте диссертационных работ и авторефератов по педагогике.*

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

Общие подходы к определению достоверности совпадений и различий. 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 – *статистические критерии*[23]. 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 a), 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 a, 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 $1 - a$. Например, если $a = 0{,}05$ 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.

В дальнейшем мы ограничимся уровнем значимости $a = 0{,}05$, 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 и порядковой шкале.

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: $x = (x_1, x_2, \ldots, x_N)$ – выборка для экспериментальной группы и $y = (y_1, y_2, \ldots, y_M)$ – выборка для контрольной группы, где x_i – элемент выборки – значение исследуемого показателя у i-го члена экспериментальной группы, $i = 1, 2, \ldots, N$, а y_j – значение исследуемого показателя у j-го члена контрольной группы, $j = 1, 2, \ldots, M$. 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 1) – количество правильно решенных задач.

---

[23] *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.*

Для данных, измеренных в шкале отношений, для проверки гипотезы о совпадении характеристик двух групп целесообразно использование либо *критерия*[24] *Крамера-Уэлча*, либо *критерия Вилкоксона-Манна-Уитни*. Критерий Крамера-Уэлча предназначен для проверки гипотезы о равенстве средних (строго говоря – математических ожиданий) двух выборок, критерий Вилкоксона-Манна-Уитни[25] 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[26]).

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. выше) по следующей формуле:

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

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. 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 $T_{эмп}$ £ 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[27] 95%".

В качестве примера применим алгоритм для данных из таблицы 1.

Для этого сравним сначала числа правильно решенных задач в контрольной и экспериментальной группе до начала эксперимента. Вычисляем[28] по формуле (3) значение $T_{эмп}$ = 0,04 ≤ 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.

Критерий Вилкоксона-Манна-Уитни[29]**.** 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[30]: $\{x_i\}_{i = 1\ldots N}$ и $\{y_j\}_{j=1\ldots M}$ и для каждого элемента первой[31] выборки x_i, $i = 1\ldots N$, определим число a_i элементов второй выборки, которые превосходят его по своему значению (то есть

---

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

[25] *Критерий Вилкоксона-Манна-Уитни плохо применим в условиях, когда число отличающихся друг от друга значений в выборках мало.*

[26] *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.*

[27] *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%.*

[28] *Для сокращения ручных расчетов средние и дисперсии могут быть вычислены в рамках описательной статистики в компьютерной программе Microsoft Excel для Windows – см. выше таблицу 6.*

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

число таких y_j, что $y_j > x_i$), а также число b_i элементов второй выборки, которые по своему значению равны ему (то есть число таких y_j, что $y_j = x_i$). Сумма

$$a_1 + a_2 + \ldots + a_N + \frac{1}{2}\ (b_1 + b_2 + \ldots + b_N\ ) = \sum_{i=1}^{N} a_i\ +\ \frac{1}{2}\ \sum_{i=1}^{N} b_i$$

по всем N членам первой выборки называется *эмпирическим значением критерия Манна-Уитни* и обозначается U.

Определим *эмпирическое значение критерия Вилкоксона*:

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

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 $W_{эмп}$ £ 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%".

В качестве примера применим алгоритм для данных из таблицы 1.

Для этого сравним сначала числа правильно решенных задач в контрольной и экспериментальной группе до начала эксперимента. В таблице 8 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, плюс полусумма числа членов контрольной группы, решивших такое же (что и он) число задач (третий столбец). Например, в таблице 8 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 $y_j > x_1$ равно 16. Следовательно, a_1 = 16. Число таких y_j, что $y_j = x_1$ равно 2. Следовательно, b_1 = 2. Итак, $a_1 + b_1 / 2$ = 17, то есть число затененных ячеек равно 17. Записываем это число во вторую строку третьего столбца. Аналогично заполняются остальные строки третьего столбца.

Таблица 8

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** $a_i + b_i / 2$	**Номер члена контрольной группы** j	**Число задач, правильно решенных *j*-ым членом контрольной группы до начала эксперимента** y_j
1	12	17	1	15
2	11	19,5	2	13
3	15	9	3	11
4	17	5	4	18
5	18	4	5	10
6	6	28,5	6	8
7	8	24,5	7	20

[30] *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.*

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

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 $a_i + b_i / 2$	Номер члена контрольной группы j	Число задач, правильно решенных j-ым членом контрольной группы до начала эксперимента y_j
8	10	21,5	8	7
9	16	6	9	8
10	12	17	10	12
11	15	9	11	15
12	14	12	12	16
13	19	2	13	13
14	13	14,5	14	14
15	19	2	15	14
16	12	17	16	19
17	11	19,5	17	7
18	16	6	18	8
19	12	17	19	11
20	8	24,5	20	12
21	13	14,5	21	15
22	7	27	22	16
23	15	9	23	13
24	8	24,5	24	5
25	9	22,5	25	11
–	–		26	19
–	–		27	18
–	–		28	9
–	–		29	6
–	–		30	15

Сумма всех 25 чисел в третьем столбце таблицы 8 дает эмпирическое значение критерия Манна-Уитни $U = 373$. Вычисляем по формуле (4) значение $W_{эмп} = 0,0338 \leq 1,96$. Следовательно, гипотеза о том, что сравниваемые выборки совпадают, принимается на уровне значимости 0,05.

Теперь аналогичным образом (построив таблицу, аналогичную таблице 8, и вычислив эмпирическое значение критерия Вилкоксона) сравним числа правильно решенных задач в контрольной и экспериментальной группе после окончания эксперимента. Вычисляем по формуле (4) значение $W_{эмп} = 2,1974 > 1,96$. 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.

Методика определения достоверности совпадений и различий для экспериментальных данных, измеренных в порядковой шкале. Рассмотрим случай, когда используется порядковая шкала с L различными баллами. Характеристикой группы будет число ее членов, набравших тот или иной балл. 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, $k = 1, 2, \ldots, L$. Для рассматриваемого нами числового примера (L = 3 – "низкий", "средний" или "высокий" уровень знаний) данные приведены в таблице 4.

Для данных, измеренных в порядковой шкале (см., например, таблицу 4), целесообразно использование *критерия однородности* χ^2 ("хи" – буква греческого алфавита, название критерия читается: "*хи-*

квадрат"), эмпирическое значение $\chi^2_{эмп}$ которого вычисляется по следующей формуле[32] (пример расчета приведен ниже):

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

Критические значения $\chi^2_{0.05}$ критерия χ^2 для уровня значимости 0,05 приведены в таблице 9 (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).

Таблица 9

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 9: 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%".

Применим алгоритм для данных из таблицы 4. 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 4), 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 10. Ячейки таблицы 10 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. Например, эмпирическое значение критерия χ^2, получаемое при сравнении характеристик контрольной группы до начала эксперимента (вторая строка таблицы 10) и экспериментальной группы до начала эксперимента (третий столбец таблицы 10), равно 0,03.

В рассматриваемом примере $L = 3$ (выделены три уровня знаний – "низкий", "средний" и "высокий"). Следовательно, $L - 1 = 2$. Из таблицы 9 получаем для L – 1 = 2: $\chi^2_{0.05}$ = 5,99. Тогда из таблицы 10 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.

Следовательно "характеристики всех сравниваемых выборок, кроме экспериментальной и контрольной групп после окончания эксперимента, совпадают[33] 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%".

---

[32] *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: $n_i \ge 5$, $m_i \ge 5$, $i = 1, 2, \ldots, L$. 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.*

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

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.

Таблица 10

Эмпирические значения критерия χ^2 для данных из таблицы 4

	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

Дихотомическая шкала. Отдельно рассмотрим случай, когда используется дихотомическая шкала – порядковая шкала с всего двумя различными упорядоченными баллами – "высокий"-"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 1, 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 11 (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 11

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*[34], для которого эмпирическое значение $j_{эмп}$ вычисляется по следующей формуле (арксинус может быть вычислен в Excel):

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

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

Применим алгоритм для экспериментальных данных из таблицы 11. 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 11). Подставляя в формулу (6), получаем:

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

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 12. Ячейки таблицы 12 содержат эмпирические значения критерия Фишера для сравниваемых групп, соответствующих строке и столбцу. Жирным шрифтом выделены результаты сравнения характеристик экспериментальной и контрольной группы до начала и после окончания эксперимента.

Например, эмпирическое значение критерия Фишера, получаемое при сравнении характеристик контрольной группы до начала эксперимента (вторая строка таблицы 12) и экспериментальной группы до начала эксперимента (третий столбец таблицы 12), равно 0,16. Следовательно "состояния экспериментальной и контрольной групп до начала эксперимента совпадают с уровнем значимости 0,05".

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

Таблица 12

Эмпирические значения критерия Фишера для данных из таблицы 11

	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

[34] *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.*

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[35].

Планирование педагогического эксперимента. 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[36]. 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.

Заключение. 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.

Еще раз напомним, что рассмотренные выше инструменты анализа данных имеют программную реализацию в виде свободно распространяемой **компьютерной программы "Статистика в педагогике"**, которую можно загрузить с адреса http://www.mtas.ru/uploads/stat.zip (2.9 Мб).

---

[35] *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 c^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). 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.*

[36] *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.*