Теория

ПРИМЕНЕНИЕ СТАТИСТИЧЕСКИХ МЕТОДОВ В ПЕДАГОГИЧЕСКИХ ИЗМЕРЕНИЯХ

Вадим Аванесов
testolog@mail.ru

В статье рассматриваются актуальные вопросы применения статистических методов педагогических измерений. Даётся краткий обзор публикаций журнала по данной проблематике, приводятся доказательства эквивалентности некоторых формул, удобных для анализа ответов испытуемых по отдельным заданиям и по тесту в целом. Даны примеры применения корреляционного и регрессионного анализа.

Ключевые слова: педагогические измерения, статистические методы, вариация, корреляция, линейный парный и множественный регрессионный анализ.

Введение

Журнал «Педагогические измерения» время от времени обращался к изложению статистических методов, делая это, к сожалению, не так часто, как того требует проблема развития педагогических измерений в России. Причина проста. Трудно найти авторов, умеющих просто и понятно, для большинства недостаточно подготовленных математически читателей, писать о статистических мето-

дах, необходимых для разработки качественных педагогических тестов.

В ПИ №2 2005 г. публиковалась статья «Вычисление средних тенденций»[1], где были представлены принятые в трудах классиков зарубежной психометрики символика и формулы расчёта показателей средней тенденции: средних арифметических значений, моды, медианы исходных тестовых баллов испытуемых, а также долей правильных и неправильных ответов по каждому заданию. Там же было показан метод построения гистограмм, свидетельствующих о распределении исходных тестовых баллов, представлены таблица и матрица результатов испытуемых по всем заданиям проектируемого теста.

В ПИ №4 того же года рассматривались методы вычисления показателей вариации, асимметрии, эксцесса, проверки статистических гипотез, формулы расчёта линейного классического коэффициента корреляции Пирсона, а также излагались понятия и методы матричной алгебры, используемые в педагогических измерениях[2].

В числе четырёх основных показателей, позволяющих оценить меру вариации данных относительно средней арифметической, рассматривалась сумма квадратов отклонений от средней арифметической, обозначаемая символом *SS*x, где *SS*x означают аббревиатуру английского словосочетания sum of squares — сумму квадратов — сокращённое выражение (символ) упомянутой выше суммы квадратов отклонений тестовых результатов испытуемых от средней арифметической, по вектору *X*. Данная сумма записывалась там так:

$$SS_x = \sum (X_i - M)^2. \qquad (1)$$

В настоящей статье приведено доказательство эквивалентности её варианта,

$$SS_x = \sum X^2 - \frac{\left(\sum X\right)^2}{N}, \qquad (2)$$

широко применяемого в практике обоснования качества тестовых результатов.

В ПИ №1 2006 г. была напечатана статья по методам проведения текущего и итогового рейтинга испытуемых на основе исходных тестовых результатов. На конкретных примерах определялось место каждого испытуемого в совокупности данных. В основу итогового рейтинга был положен используемый в западных статистических пакетах метод определения процентного ранга испытуемых[3].

В ПИ №1 2008 года была опубликована статья Романа Дубинки, в которой была показана возможность успешного использования электронных таблиц Excel для расчёта стати-

1 *Аванесов В.С.* Вычисление средних тенденций// Педагогические измерения. 2005. № 2. С. 121–128.

2 *Аванесов В.С.* Введение в статистические и математические методы педагогических измерений // Педагогические измерения. 2005. № 4. С. 91–116.

3 *Аванесов В.С.* Рейтинг. // Педагогические измерения. 2006, № 1. С. 91–116.

стических показателей для тестовых заданий и теста в целом[4].

И, наконец, в ПИ № 1 2010 г. была напечатана статья о статистических методах получения структурных характеристик вариационного ряда. Эти методы позволяют определить и сравнить учебные достижения испытуемых, полученные по тестам с разным числом заданий[5].

Кромке упомянутых публикаций, были опубликованы статьи, связанные с обоснованием качества тестовых заданий посредством специализированных статистических пакетов. Эти публикации тяготеют к методике разработки тестов. А потому их полезнее рассматривать в рамках вопросов теории и методики педагогических измерений.

Процесс педагогических измерений

Применение статистических методов является существенной частью процесса педагогических измерений. Процесс измерения латентных свойств личности охватывает всех испытуемых, разработчиков тестов и тестовых заданий, а также тех, кто применяет тесты, статистические методы и пакеты, интерпретирует результаты. Для большинства педагогов самым главным свойством или качеством личности[6] является уровень подготовленности учащихся (студентов), по одной или нескольким учебным дисциплинам.

В перечисленное множество участников процесса педагогических измерений нецелесообразно включать тех исследователей, кто использует различные виды педагогического оценивания. Причина проста — они занимаются этим явно вне педагогической теории измерения латентных величин. По мнению автора, это другая, оценочная, а не метрическая деятельность, слишком часто и необоснованно подменяющая педагогические измерения. Об этом приходилось неоднократно говорить в связи с попытками ввести в России и многих других странах методы т.н. внешнего оценивания, однако вне явной связи с теорией педагогических измерений. Различия между этими видами деятельности есть, они заслуживают специального научного исследования. Пока мы будем исходить из утверждения, что это не одно и то же[7].

Есть различия между применением статистических методов в теоретических исследованиях и для обработки результатов в практике. Наиболее существенные различия между теорией и практикой педагогических измерений можно видеть в употреблении языка, символики и формул. Искусство вычис-

4 *Дубинка Р.* Проведение статистического анализа качества заданий в среде MS EXCEL// Педагогические измерения. 2008. №1. С. 111–117.

5 *Аванесов В.С.* Структурные характеристики вариационного ряда: применение квантилей для интерпретации тестовых результатов // Педагогические измерения. 2010. №1. С. 104–116.

6 Здесь не затрагиваются интересные различия этих важных понятий.

7 *Вадим Аванесов.* Проблема демаркации педагогических измерений. http://viperson.ru/wind.php?ID=592151&soch=1

ления интересующих статистических характеристик посредством удобных для практики вариантов формул процветало в психометрике до появления компьютеров. В русском языке такие формулы нередко называют рабочими. Их используют, как иногда говорят, для «ручного» подсчёта интересующих статистических характеристик тестовых результатов. Такого рода формулам учат студентов гуманитарных факультетов вузов.

Важно понимать, что в самом факте подсчёта полученных баллов испытуемых ещё нет достаточных признаков педагогического измерения. Последние возникают при появлении так называемых метрических шкал, обладающих, как минимум, свойствами интервальности[8]. Измерение появляется после перевода результатов счета в специальную трансформационную шкалу. Например, в шкалу логитов, предложенную Г. Рашем[9]. Иначе говоря, самым важным научным признаком отличия исходных тестовых баллов от настоящих тестовых баллов заключается в факте трансформации данных исходной шкалы в результаты шкалы логитов.

Величина

Выделяются два вида переменных величин. Первые — это наблюдаемые и непосредственно измеряемые. Например, скорость движущихся автомобилей наблюдаема и измеряема с помощью спидометра. Так же наблюдаема и измеряема, с помощью весов, масса чемоданов. Второй вид переменных величин проявляет себя при научном подходе к измерению таких, например, свойств личности, как интеллект и знания. При этом наблюдаемо не само интересующее свойство, а признаки его проявления, такие, например, как умение правильно ответить на задания теста.

В процессе научно организуемого педагогического измерения каждое интересующее свойство личности становится величиной. Значения исходных баллов испытуемых по переменной величине могут принимать различные значения. Сущность измеряемого свойства, подтверждаемая подходящей концепцией и термином, а также наблюдаемые средние и дисперсия значений испытуемых является самым важными признаками переменной величины.

Переменная величина начинается с общей идеи измеряемого свойства — что надо измерить. Затем готовятся заданий для выявления признаков интересующего свойства. В основе отбора содержания заданий положена та же идея. Таким образом тестовые задания становятся операциональным определением измеряемого свойства. Но

8

Wright B.D. Raw Scores are NOT measures. In: Measurement for Social Science and Education. A history of social science measurement. И мн. др., на том же сайте: http://www.rasch.org/memo62.htm . Wright B.D., Linacre J.M. Observations are Always Ordinal; Measurements, however, Must be Interval. Archives of Physical Medicine and Rehabilitation 70 (12) pp. 857-860, November 1989. http://www.rasch.org/memo44.htm .

9

Аванесов В.С. Метрическая система Георга Раша — RASCH MEASUREMENT (RM) // Педагогические измерения № 2, 2010. С. 3–36.

этого недостаточно. Нужны основания, чтобы считать, что переменная величина реализуется данным набором заданий. Мы должны предъявить задания подходящим испытуемым и проверить — соответствую ли задания реальным профилям ответов испытуемых.

В процессе измерения переменной величины испытуемые часто имеют различающиеся значения. Это и есть определение понятия вариации. При отсутствии вариации данных говорят о постоянной величине.

Само измеряемое свойство личности рассматривается как латентное. Проявления интересующего латентного элементарного свойства посредством задания называются эмпирическим индикатором. Упорядоченная система эмпирических индикаторов и шкалированных ответов на систему образует величину, или иначе, метрический показатель. Есть ещё отдельные счётные индикаторы, агрегированные показатели и др.

Педагогическое измерение позволяет локализовать каждого испытуемого на латентной шкале. Графический образ локализации испытуемого среднего уровня подготовленности на латентной переменной величине представлен на рис. 1[10].

При формировании показателя важно понимать текущие и отдалённые последствия от его введения в практику. Последствия могут быть положительными и отрицательными, положительными в одном отношении и отрицательными — в других отношениях. Можно привести пример с ЕГЭ, который стал разрушать российское образование, потому что школа стала теперь оцениваться по результатам сдачи экзаменов по небольшому числу учебных предметов, включённых в ЕГЭ. А это побуждает учителей, детей и родителей изучать преимущественно эти предметы. От идеалов всесторонне развитой личности теперь остаются одни только воспоминания.

Вместо подлинной образовательной деятельности в школах теперь преимущественно занимаются целенаправленной

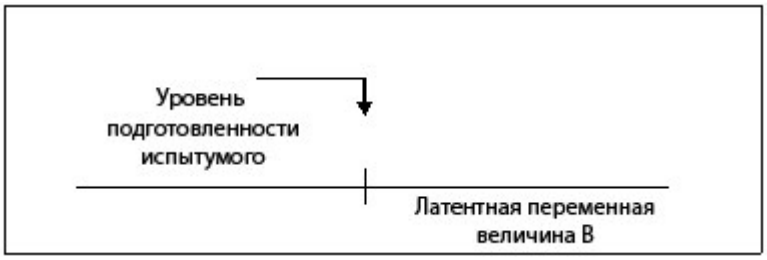

Рис. 1. Графический образ латентной переменной величины

подготовкой к ЕГЭ. Изменился — или, скорее, подменился — предмет деятельности. Знания остальных учебных дисциплин теперь стали заметно выводиться из фокуса внимания участников образовательного процесса. К тому сам ЕГЭ был изначально задуман келейно, некачественно, без научного проекта и обсуждения, а от этого ситуация в образовании только усугубилась. Похоже, что власть потеряла — а, может быть, и не имела вовсе — отзывчивость к научной критике[11].

История науки полна губительными примерами иррационального поведения, функционирования не ради достижения сущностных целей и задач, а ради выполнения плана по несовершенным показателям. В итоге возник известный в науке эффект *реификации (овеществления) показателя*, следствием чего становится работа не на суть, а на негодный показатель сути. Раньше это были, например, процент учащихся, не имеющих двойки, процент отличников боевой и политической подготовки, госплановские количественные показатели производства обуви, одежды, автомобилей. Качество во внимание не принималось. Сейчас нечто похожее сложилось в системе образования, где некачественные оценки ЕГЭ заменяют чиновникам суть и смысл образовательной деятельности.

В результате произошло резкое падение общего уровня образованности молодёжи. Прав был известный философ Альбер Камю: «Чувство абсурдности поджидаёт нас на каждом шагу»[12].

Баллы и шкалы

Различия по языку, используемому в теории и практике педагогических измерений, уже подвергались анализу[13]. Теперь обратимся к различиям формул и символики. В теории (науке) всегда вводятся необходимые уточнения и коррекции.

Начнём с понятия т.н. тестового балла. В практике результат испытуемого нередко называют именно так. Например, по сей день баллы КИМов ЕГЭ, где исходные, а где трансформированные, тоже называют «тестовыми». Хотя много раз говорилось, что настоящих тестов там нет. Естественно, нет там и тестовых баллов.

Если испытуемый отвечает на задания качественно подготовленного теста, то такой процесс есть основания называть тестированием. При этом испытуемым и их родителям сообщается о цели и задачах тестирования, профессиональной общественности сообщается об уровне надёжности и валидности тестовых результатов. Без такой информации

10 *Wright B.D., Stone, M.H.* Best Test Design. MESA Press, Chicago, 1979. P. 1.

11 *Аванесов В.С.* Проблема модернизации образования. http://viperson.ru/wind.php?ID=635807&soch=1

12 *Камю Альбер.* Миф о Сизифе. Эссе об абсурде//Сумерки богов. Политиздат. 1990. http://bibliotekar.ru/sumerki/5.htm.

13 *Аванесов В.С.* Язык педагогических измерений // Педагогические измерения. № 2. 2009. С. 29–60. http://testolog.narod.ru/Theory65.html

качественного тестирования не бывает.

По окончании тестирования к полученному результату полезно добавлять словосочетание «исходный тестовый балл», что является сокращением более точного выражения «исходный тестовый балл, полученный испытуемым в процессе тестирования по данному тесту». Этот балл получается посредством *подсчёта* числа правильных ответов (или исходных баллов), полученных испытуемым. Затем исходные тестовые баллы переводятся в ту или иную шкалу. Получается, соответственно, шкалированный тестовый балл.

При этом надо обязательно показывать два главных свойства шкалы: средний балл и стандартное отклонение, а также приводить график распределения исходных баллов. Графики распределения некачественных баллов по многим предметам с первых дней проведения ЕГЭ стали госсекретом. Засекречивание распределений оградило, на некоторое время, ЕГЭ от профессиональной критики, но, тем самым, и погубило этот экзамен в качественном отношении.

Далее вводится символика и индексация. Исходный тестовый балл испытуемого под номером i в тесте под номером j представляется как X_{ij}. Исходные тестовые баллы представлены в шкале, отражающей результаты подсчёта баллов.

Расчёт средних арифметических

Другое различие между истолкованием формул в теории и практике даёт пример расчёта средней арифметической. В предположении, что в данный момент мы имеем дело с результатами только одного теста, средняя арифметическая исходных баллов вычисляется по известной формуле:

$$\frac{\Sigma X_i}{N}. \qquad (3)$$

Если бы было несколько тестов, то у символа M появился бы индекс j. M_j тогда означал бы среднее арифметическое исходных тестовых баллов испытуемых по тесту под номерам j. В случае нескольких тестов j принимает значение номера теста: 1, 2 и т.д.

В случае, когда исходные данные результатов тестирования представлены в т.н. дихотомической шкале, где ответы испытуемых оцениваются либо 1 — за правильное решение, либо 0 — за неправильное решение, формулу 1 можно представить в непривычном для математиков виде

$$M = \frac{\Sigma\, единичек + \Sigma нулей}{N}. \qquad (4)$$

Если далее обозначить число правильных решений по множеству испытуемых символом *m* и принять, что сумма нулей есть нуль, то формула 4 приобретает приемлемый для математики вид:

$$M_j = \frac{m_j}{N}. \qquad (5)$$

Отношение $\frac{m_j}{N}$ есть определение средней арифметической, для данных, представленных в дихотомической шкале. В педагогических измерениях это отношение называется *долей* правильных ответов испытуемых (p_j) на задание теста, под номером *j*. Таким образом, средняя арифметическая (p_j) для данных этой шкалы вычисляется по формуле 6.

$$p_{.j} = \frac{m_j}{N}. \qquad (6)$$

Получается, что средний арифметический балл для данных, полученных в дихотомической шкале, выражается формулой расчёта доли правильных ответов.

Показатели вариации

В формуле 1 уже приводилось выражение для расчёта суммы квадратов отклонений исходных баллов испытуемых от средней арифметической по тесту *X*:

$$SS_x = \sum (X_i - M)^2. \qquad (1, \text{повторно})$$

В учебниках обычно сообщается, что эта сумма — наименьшая среди сумм квадратов отклонений от других значений вариационного ряда. Это свойство средней арифметической основано на равенстве

$$\sum (X_i - M) = 0. \qquad (7)$$

Доказательство вытекает из операции раскрытия скобок и замены одних членов равенства другими, эквивалентными. После раскрытия скобок получается ΣX_i $\Sigma M = 0$. Для данных, полученных в дихотомической шкале (1/0), знак S означает не что иное, как единица, взятая *N* раз. В итоге получается *N*. Поскольку средний арифметический балл *M* является константой, ΣM можно заменить произведением *NM*. Одно из доказываемых свойств средней арифметической таково: сумма баллов остаётся неизменной, если исходный тестовый балл каждого испытуемого заменить средней арифметической. Тогда получится, что $\Sigma X_i - NM = 0$.

Известную всем формулу расчёта средней арифметической $M = \frac{\Sigma X_i}{N}$ можно представить в виде произведения слева и справа на *N*. Получается $NM = \Sigma X_i$. В эту формулу вмес-

то NM поставим ΣX_i. В итоге получаем $\Sigma X_i - \Sigma X_i = 0$.

Теперь вернёмся к формуле 1 и докажем её эквивалентность формуле 2.

$$SS_x = \sum X^2 - \frac{\left(\sum X\right)^2}{N}.$$

(2, повторно)

Доказательство эквивалентности формул 1 и 2 вытекает из допустимых алгебраических операций в формуле 1.

Вначале возводится в квадрат разность в скобках формулы 1. Получается

$$SS_x = \Sigma(X_i^2 - 2X_i M + M^2). \quad (8)$$

Раскрытие скобок в формуле 3 даёт выражение:

$$SS_x = \Sigma X_i^2 - 2\Sigma X_i M + \Sigma M^2. \quad (9)$$

Далее в формуле 9 можно сделать замены. Известная формула расчёта средней арифметической $M = \frac{\Sigma X_i}{N}$ уже представлялась в виде произведения слева и справа на N. Получается $NM = \Sigma X_i$. Отсюда вытекает правомерность замены ΣX_i произведением NM в формуле 9. Знак Σ, используемый для данных в дихотомической шкале (1/0) означает не что иное, как единица, взятая N раз. В итоге получается N. Тогда последний член формулы 9, выражаемый символом ΣM^2 можно представить как NM^2.

Получаем для формулы 9

$$SS_x = \Sigma X_i^2 - 2NMM + NM^2. \quad (10)$$

Или, иначе:

$$SS_x = \Sigma X_i^2 - 2NM^2 + NM^2. \quad (11)$$

После приведения подобных членов в формуле (11) остаётся

$$SS_x = \Sigma X_i^2 - NM^2. \quad (12)$$

Деление левой и правой частей на N приводит к одной из самых удобных формул для расчёта стандартного показателя вариации, называемого дисперсия.

$$\frac{SS_x}{N} = \frac{\sum X^2}{N} - M^2. \quad (13)$$

Важно заметить, что в литературе на русском языке при вычислении уточнённых характеристик выборочной совокупности вместо N обычно используется делитель $N - 1$. А потому сумму квадратов отклонений от средней арифметической повсеместно рекомендуется делить не на N, а на $N - 1$. Статистическое деление на $N - 1$ членов формулы 12 даёт

$$\frac{SS_x}{N} = \frac{\sum X^2}{N-1} - M^2. \quad (14)$$

Однако в западной теории психометрики, для упрощения преобразований в формулах, сложилась традиция деления и умножения на *N*, при условии, что читатели понимают и принимают этот нюанс, а при применении формул для расчёта статистических характеристик испытуемых выборочной совокупности не забывают делить на $N - 1$.

Вернёмся, поэтому, к более экономной записи в психометрической традиции и к формуле 12. В этой формуле произведение NM^2 полезно заменить эквивалентным выражением

$$N\left(\frac{\sum X_i}{N}\right)\left(\frac{\sum X_i}{N}\right).$$

После сокращения оно упрощается и становится равным . Подставляем полученный результат в формулу 11. Теперь она имеет вид формулы 2:

$$SS_x = \sum X^2 - \frac{\left(\sum X_i\right)^2}{N}.$$

(2, повторно)

Это и есть самая распространённая в педагогических и психологических измерениях, альтернативная формула для расчёта суммы квадратов отклонений тестовых результатов испытуемых от средней арифметической. Равенство формул 1 и 7 доказано.

Дисперсия. Вторым, по счёту, но не по важности, показателем вариации тестовых баллов является дисперсия, (s^2), или по-старому, варианса. Одна из формул её расчёта такова:

$$s^2 = \frac{SS}{N-1}. \qquad (15)$$

Для тестовых заданий, в которых используется только дихотомическая оценка (1 или 0) дисперсия определяется по сравнительно простой формуле:

$$s^2 = p_j q_j, \qquad (16)$$

где p_j и q_j — доли правильных и неправильных ответов в каждом задании (j). Значение корня квадратного из дисперсии даёт стандартное отклонение.

Корреляция

В педагогических измерениях связь и влияние интересующих свойств личности изучаются посредством тестов и методов статистики. Например, исследователи часто пытаются проверить гипотезу о наличии связи между результатами испытуемых в интеллектуальных тестах с оценками знаний по различным учебным дисциплинам и с уровнем учебной мотивации. Чем выше учебная мотивация и уровень интеллектуального развития, тем выше, в среднем, по множеству испытуемых, должны быть и тестовые баллы испытуемых по учебным дисциплинам.

Гипотезой называется предположение о связи, влиянии или закономерных различиях, достоверность которых проверяется научными методами. В числе наиболее часто применяемых методов — расчёт коэффициентов корреляции, парной, множественной и частной регрессии, а также научный эксперимент. Расчёт коэффициентов корреляции используется для проверки гипотезы о связи между результатами тестов. Гипотезы о влиянии одного признака на другой проверяются посредством регрессионного анализа.

Возьмём небольшой пример двух тестов, *X* и *Y*, на которые отвечали пять испытуемых[14]. Результаты и процесс корреляционного анализа представляется в таблице (см. табл. 1) и рассчитывается посредством четырёх формул.

В табл. 1 первый столбец представляет номера испытуемых, второй столбец — их результаты по тесту *X*, третий столбец — их результаты по тесту *Y*, четвёртый столбец — произведения значений *X* на *Y*, у каждого испытуемого. В пятом и шестом столбцах представлены квадраты значений *X* и *Y* каждого испытуемого. Для удобства читателей приведём здесь все четыре формулы расчёта коэффициентов корреляции, их названия и расчёты, сделанные по ним для данных примера, в табл. 2.

Графические образы соотношения тестовых результатов

При использовании корреляционного анализа рекомендуется проводить визуальный анализ расположения точек *X* и *Y* каждого испытуемого на плоскости. При этом важно определить принцип расположения точек.

Таблица 1

Учебный пример данных для коррелирования результатов пяти испытуемых по двум тестам

№ п/п	*X*	*Y*	*X·Y*	X^2	Y^2
1	2	1	2	4	1
2	4	2	8	16	4
3	3	3	9	9	9
4	5	4	20	25	16
5	6	5	30	36	25
$\sum$:	20	15	69	90	55
M	4	3			

[14] В реальных исследованиях для проверки правдоподобности гипотезы связи число испытуемых рекомендуется иметь не менее тридцати. Если планируется статистическая обработка результатов методами многомерного статистического анализа, то расчёт выборки желательно делать из требования иметь не менее 5–10 человек на каждый тест или иной показатель. Это необходимо для получения достоверных выводов о значениях выборочных статистик.

Таблица 2

Формулы и примеры расчёта классического коэффициента корреляции Пирсона

Название	Формула	Расчёты
Сумма квадратов отклонений от средней арифметической по вектору X	$SS_x = \sum X^2 - \frac{\left(\sum X\right)^2}{N}$	$SS_x = 90 - \frac{(20)^2}{5} = 10$
Сумма квадратов отклонений от средней арифметической по вектору Y	$SS_y = \sum Y^2 - \frac{\left(\sum Y\right)^2}{N}$	$SS_y = 55 - \frac{(15)^2}{5} = 10$
Сумма произведений X на Y, скорректированная на средние значения	$SP_{xy} = \sum XY - \frac{\sum X \sum Y}{N}$	$SP_{xy} = 69 - \frac{(20)\cdot(15)}{5}$
Классический коэффициент корреляции Пирсона	$r = \frac{SP_{xy}}{\sqrt{SS_x\, SS_y}}$	$R = \frac{9}{\sqrt{10 \cdot 10}} = 0{,}900$

В зависимости от этого принципа выбирается метод корреляции и регрессии — соответственно линейной или нелинейной. В данной статье рассматриваются наиболее простые, линейные методы.

Посмотрим примеры расположения точек на плоскости у десяти испытуемых.

На рис. 2 приведены точки, представляющие проекции исходных тестовых баллов испытуемых по двум тестам, *X* и *Y*. Общая конфигурация (расположение) таких точек может рассматриваться как близкая к линейной модели связи результатов: иначе говоря, расположение точек всех десяти испытуемых приближено к расположению на прямой линии, называемой в таких случаях линией регрессии. Это открывает возможность подобрать такую прямую линию, чтобы сумма квадратов отклонений имеющихся точек от линии стала бы минимальной по сравнению с другими аппроксимирующими линиями.

Это и есть важный признак наличия достаточно заметной положительной корреляции. Видна тенденция: по мере роста результатов испытуемых по тесту *X* наблюдается не слишком строгая, но устойчивая тенденция роста результатов по тесту *Y*. Соответственно, в таком случае для расчёта коэффициента корреляции и регрессии применяются линейные методы. Пример и формулы для расчёта

Рис. 2. Случай положительной статистической связи между результатами испытуемых по тесту *X* и *Y*

линейного коэффициента корреляции приведён в табл. 1 и 2.

Одна из интерпретаций меры связи — так называемый коэффициент детерминации, равный квадрату значения коэффициента корреляции, умноженного на сто. Коэффициент детерминации не следует толковать буквально. Это лишь традиционное, довольно устаревшее название меры связи между *X* и *Y*, выражаемое в процентах:

$$D = (0{,}9)^2 \cdot 100 = 81\%.$$

На рис. 3 видна противоположная тенденция. По мере роста результатов у испытуемых по тесту *X* заметна устойчивая статистическая тенденция снижения результатов испытуемых по тесту *Y*. В таких случаях коэффициент корреляции отрицательный.

На рис. 4 представлен пример графического представления случая отсутствия заметной корреляции между результатами испытуемых по тестам *X* и *Y*. При таком расположении точек корреляцию можно не считать: она не будет существенной.

На рис. 5 расположены точки, конфигурация которых свидетельствует о необходимости расчёта т.н. нелинейного коэффициента корреляции. Здесь надо считать т.н. корреляционное отношение. Оно будет иметь отрицательный знак.

Рис. 3. Графическое представление исходных баллов испытуемых в виде точек на плоскости для случая отрицательной корреляции

Рис. 4. Случай отсутствия заметной корреляции

Рис. 5. Случай нелинейного расположения баллов испытуемых на плоскости

Нелинейный случай связи представлен и на рис. 5. В таких случаях надо считать корреляционное отношение. Здесь корреляционное отношение будет иметь отрицательный знак.

Рис. 6 даёт ещё один пример нелинейной, но уже положительной связи между баллами испытуемых по двум тестам. Полезно обратить внимание на вертикальную линию, указывающую на границу эффективности теста *X* для прогнозирования результатов по тесту *Y*. Справа от этой линии увеличение результатов по *X* не сопровождается увеличением результатов по *Y*.

Это явление в теории профессионального отбора называется ceiling effect (эффект «потолка»). Тест *X* дифференцирует испытуемых от слабого до среднего уровня, после чего он становится бесполезным для прогнозирования результатов по *Y*.

Из двух последних примеров важно сделать полезный вывод: прежде чем считать ту или иную меру связи для полученных данных, полезно увидеть расположение точек на плоскости. И только после этого решать, какую меру связи лучше считать — линейную или нелинейную? В статистических пакетах есть опция, позволяющая вывести на печать расположение таких точек. И этим методом рекомендуется пользоваться при подготовке отчётов по разработке педагогических тестов. Применение в ситуации нелинейной связи классическо-

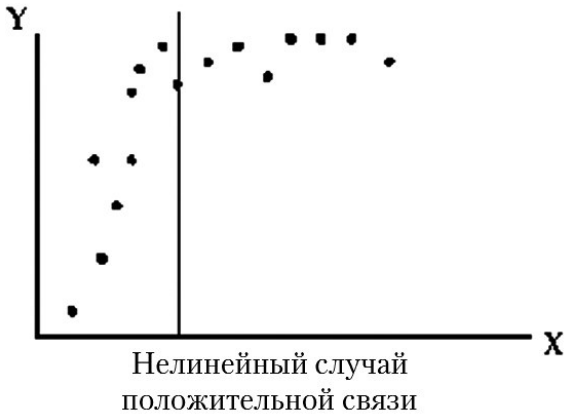

Рис. 6. Расположение результатов испытуемых по двум тестам, X и Y, на плоскости

го линейного коэффициента корреляции Пирсона заметно занижает меру реально существующей связи.

Начала регрессионного анализа

Регрессионный анализ представляет собой совокупность статистических методов, используемых, в частности, для разработки тестов и для обработки тестовых результатов испытуемых. Если в эксперименте используются один или несколько тестов, то с помощью регрессионного анализа удаётся определить меру статистического влияния результатов одного теста на результаты другого теста, а также влияние комбинации нескольких тестов (независимых переменных величин) на вариацию результатов по другому тесту (зависимой переменной величины).

В линейном регрессионном анализе строится прямая линия $\hat{Y} = a + bX$, где X — результаты испытуемых по тесту X, а $\hat{Y}$ — прогнозируемые по данной модели ожидаемые результаты по вектору $\hat{Y}$[15], a и b — параметры прямой, построенной таким образом, чтобы сумма квадратов отклонений всех точек от этой прямой была минимальной. Эта линия называется регрессией результатов по Y на результаты по X, что в переводе на естественный язык означает меру влияния результатов по тесту X на результаты по тесту Y. Такая вот своеобразная лексика.

При построении прямой используется известный в статистике метод наименьших квадратов. Данный метод позволяет

15

При изложении материала данного раздела автор опирался на символику, формулы и примеры из классического учебника Kerlinger F.N. & Pedhazur, Elazar. Multiple Regression in Behavioral Research. Holt..., N–Y, 1973.

ответить на вопрос: как увеличится (уменьшится) Y в зависимости от изменения значений по X?

В уравнении парной линейной регрессии $Y' = a + bX$ параметр b вычисляется по формуле: $b = \frac{SP_{xy}}{SS_x}$; a считается по формуле: $a = \overline{Y} - b \cdot M_x$.

Для определения параметра крутизны наклона прямой применяются расчёты на эмпирически получаемых данных, по формулам, использованным ранее для расчёта коэффициента корреляции (см. табл. 2).

По данным примера табл. 1 и 2 получаем:

$$b = \frac{SP_{xy}}{SS_x} = \frac{9}{10} = 0{,}900;$$

$$a = \overline{Y} - b \cdot M_x = 3 - 0{,}9 \cdot 4 =$$

$$= 3 - 3{,}6 = -0{,}6,$$

где M_x — средняя арифметическая по X.

Применение уравнения регрессии в типовой задаче профессионального отбора

Уравнение линейной регрессии выявляет закономерность, т.е. зависимость результатов по Y от баллов по X. В профотборе полученные на одной выборке параметры уравнения применяют для прогнозирования неизвестных результатов в других выборках, если там известны лишь значения испытуемых по X, но неизвестны значения по Y.

Например: в ходе тестирования у пяти испытуемых были получены такие результаты по тесту X: $X_1 = 2$, $X_2 = 4$, $X_3 = 3$, $X_4 = 5$, $X_5 = 6$. Это данные табл. 1. Необходимо определить у них или им подобным испытуемым в других сходных выборках прогнозируемые по регрессионной линейной модели результаты по вектору Y': Y_1', Y_2', Y_3', Y_4', Y_5'.

Подставляя параметры уравнения, получаем прогнозируемые баллы по критерию Y:

$$Y_1' = -0{,}2 + 0{,}9 \cdot X_1 = -0{,}2 + (0{,}9)(2) = 1{,}6;$$

$$Y_2' = -0{,}2 + 0{,}9 \cdot X_2 = -0{,}2 + (0{,}9)(4) = 3{,}4;$$

$$Y_3' = -0{,}2 + 0{,}9 \cdot X_3 = -0{,}2 + (0{,}9)(3) = 2{,}5;$$

$$Y_4' = -0{,}2 + 0{,}9 \cdot X_4 = -0{,}2 + (0{,}9)(5) = 4{,}3;$$

$$Y_5' = -0{,}2 + 0{,}9 \cdot X_5 = -0{,}2 + (0{,}9)(6) = 5{,}2.$$

Заметим, что в табл. 1 есть реальные результаты по вектору X и Y. И они несколько отличаются от прогнозируемых значений Y', основанных на уравнении линии регрессии. Этот эффект регрессии виден из табл. 3, где добавлен вектор отклонений реальных результатов

по Y от прогнозируемых по модели (Y').

Для баллов по Y ниже среднего арифметического модельные (прогнозированные) значения по Y' завышаются. Для баллов по X выше среднего арифметического прогнозируемые значения по Y' занижаются. Это и есть проявление феномена регрессии (стремления) к среднему.

Множественный регрессионный анализ

Это один из наиболее эффективных методов разработки педагогических тестов. Со статистической точки зрения исходные значения тестовых баллов педагогического теста представляют собой интегральную переменную величину, зависимую от ответов испытуемых по всем заданиям теста.

При этом важно наличие двух условий. Первое — задания должны быть действительно тестовые, отвечать требованиям, предъявляемым именно к таким заданиям[16]. Минимум требований — это известные меры трудности и значения коэффициентов корреляции ответов испытуемых на задания с суммой баллов. Второе условие — независимость заданий. Независимость понимается в статистическом смысле, совпадающем с т.н. аксиомой локальной независимости: для двух любых заданий теста вероятность правильного ответа на одно задание не должна зависеть от вероятности правильного ответа на другое задание.

Соответствие заданий теста этой аксиоме проверяется эмпирически, в форме статистической гипотезы. Если гипотеза независимости подтверждается, то аксиома считается выполненной, если не подтверждается, то имеет место факт нарушения, а потому статистически зависимые задания при разработке теста подлежат замене.

Таблица 3

Результаты регрессионного анализа

№ п/п	X	Y	$Y - \hat{Y}$	$X \cdot Y$	X^2	Y^2
1	2	1	1 – 1,6 = –0,6	2	4	1
2	4	2	2 – 3,4 = –1,4	8	16	4
3	3	3	3 – 2,5 = 0,5	9	9	9
4	5	4	4 – 4,3 = 0,2	20	25	16
5	6	5	5 – 5,2 = 1,2	30	36	25

16

Аванесов В.С. Композиция тестовых заданий. М.: Центр тестирования, 2-е изд. 2002.

Для k — числа независимых переменных (заданий теста) прогнозируемое значение исходного тестового балла каждого испытуемого считается на основе линейного уравнения множественной регрессии:

$$Y' = a + b_1 \cdot X_1 + b_2 \cdot X_2 + \ldots + b_k \cdot X_k, \quad (17)$$

где параметр a — свободный член уравнения множественной линейной регрессии, b_j — значения параметра b для каждого задания под номером j, k — число заданий теста. Это уравнение для k числа заданий теста (независимых переменных).

При разработке теста наибольший интерес проявляется к изучению влияния каждого отдельного задания теста на вариацию баллов испытуемых по зависимой переменной величине. Этот балл получают элементарным или взвешенным сложением баллов, полученных по всем заданиям теста. Чем больше значение b_j у какого-либо задания, тем большим может быть вклад этого задания в общую сумму баллов при условии использования стандартных шкал для каждой независимой переменной. При использовании нестандартизованных результатов, вклад в общую вариацию зависит ещё от значения дисперсии баллов по каждому заданию.

Когда значения коэффициентов b_j для всех заданий принимаются, с целью упрощения модели, равным единице, то Y_i становится элементарной суммой баллов испытуемого по всем заданиям теста. Если какие-либо задания не прошли проверку, предъявляемую к тестовым заданиям, то баллы по таким заданиям складывать для получения зависимой переменной нельзя.

Множественный линейный регрессионный анализ позволяет определить влияние независимых переменных (X_j) на зависимую (Y), построить регрессионную модель, которая показывала бы, на какое значение увеличится (уменьшится) Y в зависимости от изменения результатов по X.

Рассмотрим самый простой случай множественного линейного регрессионного анализа одной зависимой переменной (теста Y) и двух независимых переменных, тестов X_1 и X_2[17].

Вначале находим по ранее использованным формулам суммы квадратов отклонений от средних арифметических по векторам Y, X_1 и X_2:

$$SS_y = 770 - \frac{110^2}{20} = 165;$$

$$SS_{x_1} = 625 - \frac{99^2}{20} = 134{,}95;$$

$$SS_{x_2} = 600 - \frac{104^2}{20} = 59{,}20.$$

Эти суммы представлены в последней строке табл. 4.

[17] Данные приводятся по цитированной книге: Kerlinger F.N. & Pedhazur, Elazar. Multiple Regression in Behavioral Research. Holt..., N-Y, 1973. С. 33.

Таблица 4

Испы-ые п/п	Y	Y^2	X_1	X_1^2	X_2	X_2^2	$X_1 \cdot X_2$	$X_1 Y$	$X_2 Y$
1	2	4	2	4	4	16	8	4	8
2	1	1	2	4	4	16	8	2	4
3	1	1	1	1	4	16	4	1	4
4	1	1	1	1	3	9	3	1	3
5	5	25	3	9	6	36	18	15	30
6	4	16	4	16	6	36	24	16	24
7	7	49	5	25	3	9	15	35	21
8	6	36	5	25	4	16	20	30	24
9	7	49	7	49	3	9	21	49	21
10	8	64	6	36	3	9	18	48	24
11	3	9	4	16	5	25	20	12	15
12	3	9	3	9	5	25	15	9	15
13	6	36	6	36	9	81	54	36	54
14	6	36	6	36	8	64	48	36	48
15	10	100	8	64	6	36	48	80	60
16	9	81	9	81	7	49	63	81	63
17	6	36	10	100	5	25	50	60	30
18	6	36	9	81	5	25	45	54	30
19	9	81	4	16	7	49	28	36	63
20	10	100	4	16	7	49	28	40	70
$\sum$:	110	770	99	625	104	600	538	645	611
$\bar{X}$	5,50		4,95		5,20				
SS	165		134,95		59,20				

Далее находим суммы парных произведений X и Y:

$$SP_{x_1 y} = \sum X_1 Y - \frac{\sum X \cdot \sum Y}{N} =$$

$$= 645 - \frac{99 \cdot 110}{20} = 100{,}50;$$

$$SP_{x_2 y} = 611 - \frac{104 \cdot 110}{20} = 39;$$

$$SP_{x_1 x_2} = 538 - \frac{99 \cdot 104}{20} = 23{,}20.$$

Уравнение для двух независимых переменных:

$$Y' = a + b_1 \cdot X_1 + b_2 \cdot X_2.$$

При двух независимых переменных формулы для расчёта коэффициентов регрессии имеют вид:

$$b_1 = \frac{SS_{X_2} \cdot SP_{X_1Y} - SP_{X_1X_2} \cdot SP_{X_2Y}}{SS_{X_1} \cdot SS_{X_2} - SP_{X_1Y_2}};$$

$$b_2 = \frac{SS_{X_1} \cdot SP_{X_2Y} - SP_{X_1X_2} \cdot SP_{X_1Y}}{SS_{X_1} \cdot SS_{X_2} - SP_{X_1Y_2}};$$

$$b_1 = \frac{59,20 \cdot 100,50 - 23,20 \cdot 39}{134,95 \cdot 59,20 - (23,20)^2} = 0,6771;$$

$$b_2 = \frac{134,95 \cdot 39 - 23,20 \cdot 100,50}{134,95 \cdot 59,20 - (23,20)^2} = 0,3934.$$

$$a = \overline{Y} - b_1 X_1 - b_2 X_2 =$$

$$= 5,5 - 0,6771 \cdot 4,95 -$$

$$- 0,3934 \cdot 5,20 = 0,1027.$$

Таким образом, уравнение регрессии принимает следующий вид:

$$Y' = 0,1027 + 0,6771 X_1 +$$

$$+ 0,3934 X_2.$$

Данное уравнение позволяет оценить меру влияния на зависимую переменную Y результатов по независимым переменным (тестам). Заметим, что тест X_1 больше влияет на результаты теста Y, чем тест X_2. Помимо сущностных причин, есть и формальный признак большего влияния: вариация баллов по первому тесту существенно выше вариации баллов по второму тесту.

Определение достоверности регрессии

В большинстве статистических вычислений необходимо определять статистическую достоверность полученных коэффициентов по выборочным данным. Достоверность полученной регрессии определяется при помощи F-критерия Фишера. Есть два варианта формулы:

$$F = \frac{\frac{R^2}{k}}{(1 - R^2) \cdot (n - k - 1)},$$

или

$$F = \frac{SS_{reg} / d_{f_{reg}}}{SS_{res} / d_{f_{res}}},$$

где, k — число независимых переменных в модели;
$d_{f_{reg}}$ — указывает на число степеней свободы, равное $k - 1$;
R^2— коэффициент множественной детерминации, равный квадрату коэффициента множественной корреляции.

При этом полезно помнить основное равенство регрессионного анализа: $SS_{tot} = SS_{reg} + SS_{res}$, где SS_{tot} — сумма квадратов отклонений по Y; SS_{reg}— сумма квадратов отклонений, объясняемая линейной регрессией Y на результаты по X_1 и X_2; SS_{res} — остаточная сумма квадратов отклонений баллов по Y от линии регрессии.

Для расчёта значений F вычисляются R^2, SS_{reg}, SS_{res}.

По данным примера табл. 4 имеем:

$SS_{tot} = SS_y = 165$ — общая сумма квадратов отклонений по Y:

$SS_{reg} = b_1 \cdot SP_{x_1 y} + b_2 \cdot SP_{x_2 y}$ — это сумма квадратов, объясняемая регрессионной моделью;

$SS_{reg} = 0{,}6771 \cdot 100{,}50 + 0{,}3934 \cdot 39 = 83{,}3912$ — это и есть значение вариации по Y, которая объясняется вариацией баллов по X_1 и X_2.

Поскольку $SS_{tot} = SS_{reg} + SS_{res}$, то из этого равенства легко находится мера остаточной вариации $SS_{res} = SS_{tot} - SS_{reg} = 165{,}0 - 83{,}3912 = 81{,}6088$.

Это довольно большое значение, связанно с малым числом независимых переменных. Увеличение числа качественных тестов (или заданий в случае разработки теста) уменьшает число остаточной вариации до минимума, вплоть до нуля. В таких случаях можно говорить об эффективной регрессионной модели.

Из формулы $SS_{reg} = b_1 \cdot P_{x_1 y} + b_2 \cdot P_{x_2 y} SS_{reg}$ можно найти меру вариации, объясняемую регрессией: $SS_{reg} = 0{,}6771 \cdot 100{,}50 + 0{,}3934 \cdot 39 = 83{,}3912$ — это вариация, которая объясняется вариацией по двум тестам X_1 и X_2.

Рассчитаем R^2 для нашего примера:

$$R^2 = \frac{SS_{reg}}{SS_{tot}} = \frac{83.3912}{165} = 0{,}5054.$$

Это значение говорит о том, что независимые переменные X_1 и X_2 вместе, на 50,5% , влияют на вариацию результатов испытуемых по тесту Y ($p < 0{,}05$).

Расчёт F-значения по обеим формулам даёт одинаковые результаты:

$$F = \frac{0{,}5054/2}{1 - 0{,}5054/20 - 2 - 1} = \frac{0{,}2527}{0{,}0291} = 8{,}68,$$

или

$$F = \frac{83{,}3912/2}{81{,}6088/20 - 2 - 1} = \frac{41{,}6956}{4{,}8005} = 8{,}68.$$

Может возникнуть вопрос: а каков вклад каждой независимой переменной величины в отдельности в общую вариацию результатов по критерию Y? Для ответа на этот вопрос оценивается отдельно влияние каждого теста на Y. Например, посмотрим меру влияния результатов испытуемых по X_1 на Y:

$$SS_{reg} = \frac{(SP_{x_1 y})^2}{SS_{x_1}} = \frac{100{,}5^2}{134{,}95} = 74{,}84,$$

$$R^2_{yx_1x_2} = \frac{SS_{reg}}{SS_{tot}} = \frac{74{,}84}{165} = 0{,}45,$$

т.е. 45% объясняется влиянием на Y.

F-значение для этого коэффициента множественной корреляции находим по формуле:

$$F = \frac{R^2_{yx_1x_2}/k}{1-R^2/n-k-1} =$$

$$= \frac{0{,}45/1}{1-0{,}45/20-1-1} = 14{,}95,$$

$$(P < 0{,}05)$$

Полученное значение F сравнивается с табличным (имеется в учебниках по статистике). Оно оказалось больше табличного, значит влияние изучаемого фактора на Y достоверное.

Далее оценивается влияние результатов X_2 на Y:

$$SS_{reg} = \frac{(SP_{x_2y})^2}{SS_{x_2}} = \frac{39^2}{59{,}20} = 25{,}69,$$

$$SS_{res} = 165 - 25{,}69 = 139{,}307.$$

$$R^2_{yx_2} = \frac{SS_{reg}}{SS_{tot}} = \frac{25{,}69}{165} = 0{,}155,$$

иначе говоря, только 15,5% объясняется влиянием X_2 на Y. Проверяем статистическую достоверность влияния второго теста на зависимую переменную Y по той же формуле, подставляя соответствующие значения X_2:

$$F = \frac{R^2/k}{\left(1-R^2\right)\cdot\left(n-k-1\right)};$$

$$F = \frac{SS_{reg}/d_{f_{reg}}}{SS_{res}/d_{f_{res}}} =$$

$$= \frac{25{,}69/1}{139{,}307/20-1-1} = 3{,}320.$$

Полученное F сравнивают с табличным значением. Оно оказалось меньше требуемого табличного значения. Таким образом, влияние второго теста на вариацию результатов по Y есть, но оно оказалось статистически недостоверным. Это означает возможность и необходимость дальнейшей работы по улучшению качества данного показателя.

Заключение

Применение статистических методов является обязательным для разработки качественных тестов. Обязательна и публикация статистических результатов апробации тестов. Эти требования записаны в западных стандартах (требованиях) для разработчиков тестов. Без такой информации трудно сказать что-либо о качестве используемых методов и об уровне подготовленности испытуемых.

Очевидно, что Правительству РФ уже давно надо было предъявить такие же требования и российским разработчикам государственных методов оценивания выпускников школ и абитуриентов вузов. Иначе страна не сможет выбраться из топкого болота статистически непроработанных, а потому некачественных КИМов ЕГЭ и им подобных ненаучных оценочных средств.