УДК 37.012:004

А.Г. Колгатин

Информационные технологии в научно-педагогических исследованиях

Отмечено, что развитие информационных технологий создает предпосылки для пересмотра подходов к статистическому анализу данных в научно-педагогических исследованиях. Методом статистических испытаний проведен анализ условий применения критерия Пирсона для проверки статистических гипотез. Показано, что в случае неравномерного распределения частот по категориям в условиях малых выборок точность определения вероятности ошибки первого рода может снижаться. Исследовано влияние поправки на непрерывность.

It is noted that the development of the information technologies creates the prerequisites for the approaches review to the statistical analysis of data in scientific and educational research. The statistical tests method was used to analyse the conditions of Pearson criterion application for statistical hypotheses testing. It is shown that in the case of uneven distribution of frequencies by categories in small samples, the determination accuracy of the error probability of the first kind can be reduced. The effect of the correction for a continuity is investigated.

Зазначено, що розвиток інформаційних технологій створює передумови для перегляду підходів до статистичного аналізу даних в науково-педагогічних дослідженнях. Методом статистичних випробувань здійснено аналіз умов застосування критерію Пірсона для перевірки статистичних гіпотез. Показано, що у випадку нерівномірного розподілу частот за категоріями за умов малих вибірок точність визначення ймовірності похибки першого роду може знижуватися. Досліджено вплив поправки на неперервність.

Введение. Применение информационно-коммуникационных технологий коренным образом преобразует все сферы системы образования [1], и научно-педагогические исследования – не исключение. Рассмотрим один из аспектов педагогического исследования – статистическую обработку результатов педагогического эксперимента. Традиционно эта задача решается методами математической статистики на основе проверки статистических гипотез. Выдвигаются две гипотезы: нулевая, утверждающая, что различий между сравниваемыми случайными величинами по исследуемому параметру нет, а наблюдаемые в эксперименте различия или изменения возникли как результат действия случайных факторов, и альтернативная, утверждающая, что наблюдаемые различия вызваны воздействием, исследование которого есть целью эксперимента. Затем выбирается критерий, интегрирующий наблюдаемые различия в виде конкретного числа, и вычисляется вероятность получения такого или большего различия. В педагогических исследованиях число участников эксперимента обычно невелико и выборки результатов измерений малы, поэтому обычно принимают альтернативную гипотезу, если вероятность того, что наблюдаемые различия могли быть вызваны случайными факторами (ошибка первого рода), не превышает пяти процентов. Применение современных табличных процессоров (*SPSS*, *STATISTIKA*, *Microsoft Excel*, *Open Office* и др.) делает несущественными вопросы сложности вычислений, связанных с оценкой означенной вероятности, позволяет избавиться от работы с громоздкими таблицами критических значений. Однако традиции статистического анализа по-прежнему ориентированы на упрощенные методы, которые избавляют от громоздких вычислений с помощью карандаша и бумаги, но зачастую менее мощны, т.е. обеспечивают меньшую вероятность принятия альтернативной гипотезы в тех случаях, когда различия действительно есть. Применение информационно-коммуникационных технологий обеспечивает возможность по-новому взглянуть на систему методов индуктивной статистики, выделить наиболее мощные методы, определить границы их применимости, что особенно важно в психолого-педагогических исследованиях, где выборки малы и нет возможности доказать, что распределение исследуемого признака подчиняется одному из известных законов распределения, например нормальному распределению. Более того, информационно-коммуникационные технологии открывают возможность разработки и эффективного использования новых подходов, таких как

метод статистических испытаний. Остановимся на одном из популярных классических критериев проверки статистических гипотез, критерии Пирсона χ-квадрат:

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{\left(E_{i,j} - T_{i,j}\right)^2}{T_{i,j}}, \qquad (1)$$

где $E_{i,j}, T_{i,j}$ – эмпирические и теоретические частоты соответственно; i, j – номера строк и столбцов таблиц частот соответственно; m, k – число строк и столбцов в таблицах частот соответственно.

В свете сказанного, информационно-коммуникационные технологии открывают новые перспективы для анализа границ применения критерия Пирсона, исследования мощности критерия, разработки подходов к применению критерия на малых выборках, что, по мнению автора, актуально для совершенствования методов статистической обработки данных в педагогических исследованиях.

Анализ публикаций относительно условий применения таблиц критических значений критерия Пирсона, построенных на статистике χ-квадрат, показывает, что рекомендации разных авторов несколько отличаются, но, в целом, отражают тот факт, что замена реального распределения значения критерия распределением χ-квадрат есть приближение, точность которого зависит от объема выборки:

- при сравнении двух случайных величин, имеющих два возможных значения (или значения сгруппированы по двум категориям), критерий не рекомендуется применять, если «сумма объемов двух выборок меньше 20» [2] или «хотя бы одна из абсолютных частот … в таблице 2×2, составленной на основе экспериментальных данных, меньше пяти» [2], некоторые исследователи «второе условие заменяют следующим: хотя бы одна из ожидаемых частот ... меньше пяти» [2];
- «объем выборки должен быть достаточно большим: $n \geq 30$» [3], «теоретическая частота для каждой ячейки таблицы не должна быть меньше пяти» [3];
- «... если число степеней свободы больше единицы, то критерий χ-квадрат нельзя применять, когда в 20 и более процентах случаев теоретические частоты меньше пяти … (*Siegel*, 1956)» [4];
- в работе [5] применительно к четырехпольной таблице (2×2) не рекомендуется применять критерий χ-квадрат, «... если число опытов в каждом из сравниваемых распределений меньше 10» [5], взамен предлагается использовать точный критерий Фишера.

Часто в популярной литературе по применению математической статистики в психолого-педагогических исследованиях (например, [2, 3]) предлагается коррекция критерия Пирсона (поправка на непрерывность) для случая малых выборок, при этом авторы отмечают дискуссионный характер такой поправки:

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{\left(\left|E_{i,j} - T_{i,j}\right| - 0{,}5\right)^2}{T_{i,j}} \quad . \qquad (2)$$

Рекомендации в психолого-педагогической литературе, во-первых, отдают предпочтение методам с простым алгоритмом вычислений, редко упоминая наиболее точные методы, такие как точный критерий Фишера, во-вторых, конкретные оценки погрешностей упоминаются только в примерах. С другой стороны, развитие ресурсов Интернет делает сложные алгоритмы анализа данных все более доступными. Так, например, упомянутый точный критерий Фишера можно применять не только при использовании дорогостоящих профессиональных табличных процессоров, но и свободно, воспользовавшись услугами сайтов, разрабатываемых энтузиастами [6]. Оценка вероятности ошибки первого рода по критерию χ-квадрат Пирсона включена как стандартная функция в табличные процессоры общего назначения [7].

Нерешенные проблемы. Наблюдается дисбаланс между современным развитием методов математической статистики, доступностью информационных технологий и готовностью широкой научно-педагогической общественности к эффективному применению новых информационных технологий в экспериментальь-

ной педагогике. Отсутствует системное рассмотрение как погрешности определения вероятности ошибки первого рода при применении того или иного критерия, так и ознакомление педагогов и психологов с оценками мощности критериев. Отдельного внимания требует проблема коррекции критерия Пирсона (поправка на непрерывность).

Цель данной статьи состоит в демонстрации возможностей информационных технологий на примере анализа условий применения критерия Пирсона.

Модель и процедура статистических испытаний

В каждой из рассматриваемых ситуаций задача может быть решена точными методами на основе комбинаторики. Однако выберем для анализа метод статистических испытаний ввиду полной независимости его алгоритма от условий и вида критерия. Для проведения статистических испытаний подготовлена простая модель: на основе одного и того же генератора случайных чисел создаются две выборки чисел объемов n_a и n_b соответственно. Полученные значения распределяются по m категориям, когда можно управлять законом распределения, обеспечивая равномерное распределение или преобладание частот в определенных категориях. Следует отметить, что порядок категорий для критерия Пирсона безразличен; для полученных таблиц частот вычисляется критерий Пирсона, сравнивается с критическим значением этого критерия на заданном уровне значимости (в педагогике принято пять процентов) и принимается решение о возможности отклонения нулевой гипотезы. Поскольку обе выборки генерируются одним генератором случайных чисел, известно, что в действительности справедлива нулевая гипотеза, однако в части испытаний будет принята альтернативная как результат наложения случайных факторов; относительная частота таких ошибочных решений дает оценку вероятности ошибки первого рода и должна соответствовать тому уровню значимости, на котором было выбрано критическое значение критерия Пирсона. Все исследования в данной статье проведены для уровня значимости пять процентов, критические значения критерия Пирсона по данным [2] приняты соответственно числу степеней свободы: для двух категорий – 3,841; для трех – 5,991.

Для получения удовлетворительной точности анализа необходимо значительное число испытаний. Проведено по 1 млн испытаний для каждого случая, в 95 процентах доверительный интервал оценивался как $1{,}96 \cdot s$, где s – среднее квадратическое отклонение полученных значений вероятности ошибки первого рода в последовательных идентичных испытаниях. Оценка погрешности составляет порядка 0,0005, и можно считать правильными две значащие цифры в получаемых на основе модели оценках вероятности ошибки первого рода при принятии альтернативной статистической гипотезы. В некоторых тестах были получены нулевые значения теоретических частот, что не позволяло вычислить значения критерия Пирсона (возникает ситуация деления на ноль в формуле (1)), такие результаты изымались из анализа, и, если их доля в общем числе тестов превышала один процент, то исследование при соответствующих условиях не проводилось, поскольку наблюдается выход за пределы применимости критерия Пирсона.

Анализ вероятности ошибки первого рода при равномерном распределении частот по категориям

На основе предложенной модели проведены серии статистических испытаний для случаев равномерного распределения значений сравниваемых случайных величин по двум, трем и десяти категориям. Для трех категорий выполнен полный перебор всех возможных комбинаций объемов выборок в диапазоне от девяти до 120. Минимальный проанализированный объем выборки равен девяти, т.е. в среднем по три значения в каждой категории. Такое минимальное значение выбрано для удобства, поскольку оно кратно трем, а при шести вариантах в каждой из выборок доля случайно сгенерированных наборов данных, в которых хотя

бы одной из категорий не соответствует ни одного значения в обеих выборках, составляет 2,3 процента от общего числа испытаний и гарантировать точность анализа невозможно. Анализ результатов статистических испытаний для случая распределения частот по трем категориям (табл. 1 и 2) дает основания для таких выводов:

- уменьшение объема выборки в целом снижает точность определения вероятности ошибки первого рода, причем решающее значение имеет не сумма объемов двух выборок, а объем *каждой* выборки;

Таблица 1. Относительная частота ошибочного отклонения нулевой гипотезы в статистических испытаниях при распределении сравниваемых случайных величин по трем категориям для некоторых объемов выборок

n_a	n_b						
	9	12	15	20	30	60	120
9	5,3						
12	5,2	5,6					
15	4,7	5,8	5,2				
20	5,2	5,1	5,5	5,0			
30	4,7	5,0	5,0	5,1	5,1		
60	4,5	4,9	5,1	5,0	5,0	5,1	
120	4,4	4,7	4,8	5,0	5,0	5,0	5,1

Таблица 2. Относительная частота ошибочного отклонения нулевой гипотезы в статистических испытаниях при распределении сравниваемых случайных величин по трем категориям для объемов выборок от 30 до 40

n_a	n_b										
	30	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**
30	5,1	5,2	5,1	5,2	5,3	5,0	5,0	5,3	5,2	5,1	5,1
31	5,1	5,1	5,2	5,1	5,1	5,1	5,1	5,0	5,1	5,3	5,1
32	5,1	5,1	5,1	5,1	5,2	5,0	5,1	5,3	5,0	5,0	5,3
33	5,1	5,1	5,1	5,1	5,1	5,1	5,0	5,0	5,2	5,1	5,0
34	5,3	5,1	5,1	5,1	5,1	5,0	5,1	5,1	5,0	5,0	5,2
35	5,0	5,1	5,0	5,1	5,1	5,1	5,0	5,0	5,2	5,1	5,0
36	5,0	5,1	5,1	5,1	5,1	5,0	5,2	5,1	5,0	5,2	5,1
37	5,3	5,0	5,3	5,0	5,1	5,0	5,1	5,0	5,1	5,0	5,1
38	5,2	5,1	5,0	5,2	5,1	5,2	5,1	5,1	5,1	5,2	5,0
39	5,1	5,3	5,1	5,1	5,0	5,1	5,1	5,1	5,2	5,1	5,1
40	5,2	5,1	5,3	5,0	5,2	5,0	5,1	5,1	5,1	5,2	5,1

Примечание: значения симметричные относительно главной диагонали рассчитаны в отдельных статистических испытаниях, их сравнение дает информацию о погрешности модели и округления.

- зависимость погрешности определения вероятности ошибки первого рода от объема выборок не монотонна, что объясняется дискретной природой сравниваемых частот; при определенном сочетании малых объемов выборок обеспечивается высокая точность аппроксимации реальной вероятности ошибки первого рода таблицами критических значений критерия Пирсона (например, при $n_a = 10$ и $n_b = 10$ значение $\alpha = 0{,}048$ и погрешность не превышает 5 процентов, при $n_a = 11$ и $n_b = 11$ имеем $\alpha = 0{,}051$, при $n_a = 10$ и $n_b = 12$ получаем $\alpha = 0{,}050$; в этих случаях погрешность в пределах возможности используемой для статистических испытаний модели при выбранном числе тестов, погрешность которой оценена как 2 процента); с другой стороны, при определенном сочетании достаточно больших объемов выборок возможно расхождение более 5 процентов между истинной вероятностью ошибки первого рода и ее оценкой путем распределения χ-квадрат (например, при $n_a = 44$, $n_b = 37$ значение $\alpha = 0{,}053$ вместо 0,05 и погрешность превышает 5 процентов);
- в исследованном диапазоне объемов выборок (от девяти до 120) максимальное значение вероятности ошибки первого рода составило $\alpha = 0{,}058$ (например, $n_a = 15$ и $n_b = 12$), минимальное – $\alpha = 0{,}044$ (например, $n_a = 9$ и $n_b = 120$), таким образом, при очень малых объемах выборки (больше девяти) оценки с применением распределения χ-квадрат обеспечивают погрешность не хуже 16 процентов с отклонениями как в большую, так и в меньшую сторону.

Ввиду сложного поведения зависимости погрешности традиционного применения критерия Пирсона от объема выборок, предоставим читателю возможность определить допустимые пределы применения этого критерия для сравнения двух случайных величин, распределенных по трем категориям, руководствуясь диаграммой (рис. 1), которая показывает для каждого объема выборки то предельное значение объема второй выборки, погрешность ко-

торого не превышает 5 процентов (т.е. α в пределах от 0,048 до 0,052). В реальной практике педагогического эксперимента исследователь редко принимает решение при значениях вероятности ошибки первого рода, столь близких к критическому, чтобы указанная погрешность оказала заметное влияние. Обычно стараются продолжить эксперимент, увеличив объем выборки.

Статистические испытания по оценке погрешности определения вероятности ошибки второго рода при сравнении двух случайных величин, распределенных по двум категориям (четырехпольные таблицы), показали, что точность критерия Пирсона в этом случае ниже, чем для трех категорий. В диапазоне объемов выборок от 21 до 50 (в среднем не менее 10 вариантов в каждой категории) среднее квадратическое отклонение оценки вероятности ошибки первого рода составило 0,0035; минимальное – 0,042; максимальное – 0,0659. Полученные результаты существенно отклоняются от уровня значимости 0,05, который должен быть обеспечен.

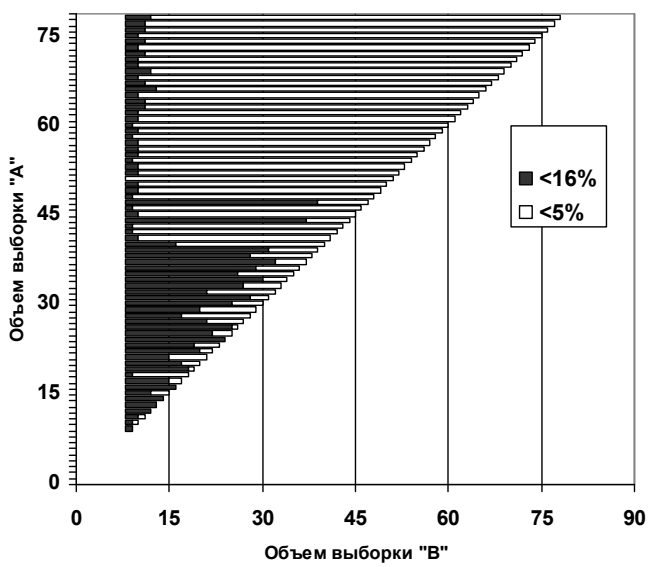

Рис. 1. Зоны погрешности приближения вероятности ошибки первого рода на основе критерия Пирсона

Примечание: зона погрешности < 16 процентов может содержать и более точные данные в отдельных точках

Исходя из полученных результатов, следует усомниться в целесообразности применения критерия Пирсона для четырехпольных таблиц, поскольку существует альтернатива – точный тест (метод) Фишера, – разработаны соответствующие таблицы [5] и доступные сайты [6], предоставляющие программное обеспечение для таких расчетов.

Анализ эффективности поправки на непрерывность

Описанные статистические испытания были повторены, формула (1) в алгоритме была заменена на формулу (2). Все остальные условия сохранены. Применение поправки по формуле (2) приводит к существенному систематическому занижению вероятности ошибки первого рода в широком диапазоне значений объемов выборок как для четырехпольных таблиц, так и для большего числа категорий. Анализ результатов (рис. 2) позволяет сделать вывод, что поправка существенно занижает вероятность ошибки первого рода.

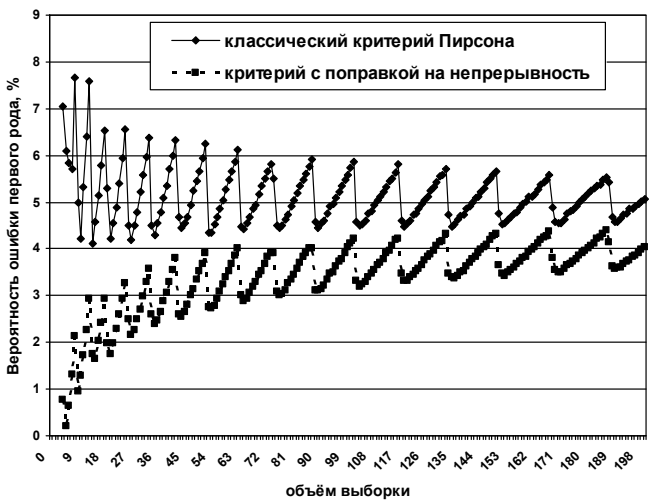

Рис. 2. Результаты статистических испытаний для четырехпольных таблиц при одинаковых объемах сравниваемых выборок

Таким образом, введение поправки на непрерывность безусловно гарантирует, что при практическом применении критерия не произойдет неоправданного отклонения нулевой гипотезы, однако существенно уменьшится и мощность критерия, которая и без того мала при малых объемах выборок. Более того, приближение результатов, полученных с поправкой к асимптотике при значительных объемах выборок происходит очень медленно (см. рис. 2).

Влияния формы распределения сравниваемых случайных величин

Одно из существенных преимуществ критерия Пирсона – его независимость от формы распределения величин в генеральных сравниваемых совокупностях. Однако в силу дискретности частот возможно нарушение этого свойства при малых объемах выборок. Для оценки возможного отрицательного эффекта была организована генерация случайных чисел с неравномерным распределением по категориям (вероятность попадания варианта в одну из категорий в два раза меньше, чем в остальные) и проведены статистические испытания. Для случайных величин, распределенных по трем категориям, пороговые погрешности и среднее квадратичекое отклонение в пределах точности проведенных статистических испытаний не изменились – попрежнему погрешность определения вероятности ошибки первого рода не превышает 16 процентов (0,008 в абсолютных величинах). Это при том, что средние частоты по одной категории в некоторых тестах (при объеме выборки от девяти до 12 составляли около двух). Однако для каждого конкретного сочетания объемов выборок различия вероятности ошибки первого рода для случаев различных распределений наблюдаются. Чтобы исследовать характер влияния формы распределения на величину вероятности ошибки первого рода проведена серия статистических испытаний (рис. 3) при фиксированном объеме выборок, равном 30 и числе категорий три.

Для значимой регистрации различий число испытаний для каждого случая было увеличено до 10 млн. Анализ результатов показывает:

- оценка вероятности ошибки первого рода на основе критерия Пирсона слабо (но статистически значимо) зависит от распределения частот при умеренном различии формы распределения (в эксперименте соотношение вероятностей отдельных категорий до 1/4);
- при существенной неравномерности распределения частот по категориям они в некоторых категориях малы (менее пяти в эксперименте) и существенно снижается точность оценки вероятности ошибки первого рода при использовании критерия Пирсона.

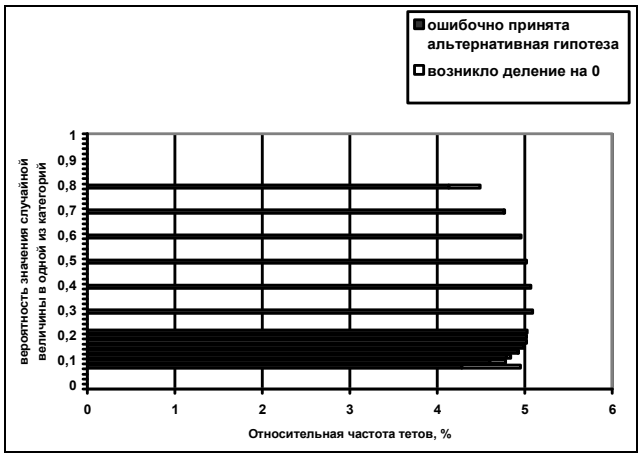

Рис. 3. Влияние формы распределения на вероятность ошибки первого рода при применении критерия Пирсона (объем выборок 30, число категорий – 3, вероятность одной из категорий варьируется, вероятности двух других категорий равны)

Анализ подтверждает, что критерий Пирсона мало чувствителен к форме распределения даже при достаточно малых объемах выборок. Однако при малых значениях частот в отдельных категориях такая зависимость становится существенной.

Заключение. Применение критерия Пирсона для четырехпольных таблиц не целесообразно, поскольку его точность в этом случае невысока даже при значительном объеме выборок; поправка на непрерывность приводит к существенному занижению мощности критерия, и, самое главное, имеется достойная альтернатива – точный метод Фишера.

При применении критического значения критерия Пирсона 5,991 для сравнения двух случайных величин, распределенных по трем категориям, вероятность ошибки первого рода оценивается с погрешностью не хуже 16 процентов, если объемы выборок не меньше девяти. Увеличение объемов выборок приводит к существенному уменьшению погрешности, которая гарантированно меньше 5 процентов для выборок объемом больше 50. Применение метода статистических испытаний вместо таблиц крити-

ческих значений, построенных на основе распределения χ-квадрат, позволяет улучшить точность оценки вероятности ошибки первого рода при применении критерия Пирсона для малых выборок.

Учитывая современный уровень развития информационных технологий, целесообразно при применении математикой статистики в педагогических исследованиях переходить от простых приближенных методов к максимально эффективным алгоритмам не зависимо от их вычислительной сложности.

Перспективным направлением совершенствования практики статистического анализа в педагогических исследованиях и системах педагогической диагностики может стать создание специализированного Интернет-ресурса, который объединит усилия математиков и педагогов-практиков в работе над созданием системы методов статистического анализа, ориентированной на максимальную эффективность решения педагогической задачи на основе применения информационно-коммуникационных технологий.

1. *Манако А.Ф.*, *Синица Е.М.* Электронные научно-образовательные пространства и перспективы их развития в контексте поддержки массовости и непрерывности // УСИМ. – 2012. – № 4. – С. 83–92.
2. *Грабарь М.И.*, *Краснянская К.А.* Применение математической статистики в педагогических исследованиях. Непараметрические методы. – М.: Педагогика, 1977. – 136 с.
3. *Сидоренко Е.В*. Методы математической обработки в психологии. – СПб.: Речь, 2002. – 350 с.
4. *Годфруа Ж.* Что такое психология: В 2 т.– М.: Мир, 1992. – Т. 2. – 496 с.
5. *Гублер Е.В*., *Генкин А.А.* Применение непараметрических критериев статистики в медико-биологических исследованиях. – Л.: Медицина, 1973. – 141 с.
6. *Fisher's* Exact Test / Øyvind Langsrud // Mode of Access. – http://www.langsrud. com/fisher.htm
7. *Білоусова Л. І.*, *Колгатін О.Г.*, *Колгатіна Л.С.* Інформаційні технології статистичної обробки даних у педагогічному університеті // Вища освіта України. – 2007. – № 2 (дод. 1), Т. 2. – Рівне: РДТУ, 2007. – С. 169–174.

Поступила 20.01.2015

E-mail: *Kolgatin@ukr.net*