УДК 519.243

Д. Е. Красильников

АЛГОРИТМ ВЫЧИСЛЕНИЯ КОЭФФИЦИЕНТА ВЫБОРОЧНОЙ ДЕТЕРМИНАЦИИ В MS-EXCEL

Нижегородский почтамт. Отделение почтовой связи №24

Рассматривается коэффициент выборочной детерминации как критерий однородности выборок в социально-экономических исследованиях. Приводится геометрическое доказательство закона разложения дисперсии, предлагается алгоритм вычисления коэффициента выборочной детерминации в MS-Excel, рассматривается случай, когда закон разложения дисперсии не выполняется, показана связь между коэффициентом выборочной детерминации и эмпирическим корреляционным отношением.

Ключевые слова: коэффициент выборочной детерминации, закон разложения дисперсии, MS-Excel, критерий однородности выборок, дисперсионный анализ, эмпирическое корреляционное отношение.

При проведении социологических, психологических, экономических и маркетинговых исследований почти всегда встает вопрос о репрезентативности исследуемой выборки. Под репрезентативностью выборки, чаще всего, понимается ее однородность. При этом в современной литературе по соответствующим дисциплинам не дается универсальный метод проверки гипотезы об однородности. Как правило, для такой проверки используют так называемый *t*-критерий, *F*-критерий или критерий “Хи-квадрат” (см., например, [1]), которые базируются на сравнении средних величин со значением функции Стьюдента, Фишера или Хи-квадрат. Однако эти критерии слабо чувствительны к социально-экономическим данным ввиду небольшого разброса значений таких данных, а применение указанных функций недостаточно обосновано, так как эти критерии были разработаны для биологических, а не социально-экономических исследований.

Другим распространенным подходом к оценке репрезентативности является обоснованность выборки с позиций той или иной задачи. Например, при изучении спроса на автомобили стоимостью от миллиона рублей выборка, сделанная из лиц с доходом 8–10 тыс. руб., будет всегда нерепрезентативной.

Тем не менее, в Советском Союзе была разработана специальная статистика (функция от выборочной совокупности), позволяющая оценить однородность любой выборки при условии ее стратификации – коэффициент выборочной детерминации (R^2выб). Его не следует путать с коэффициентом детерминации (R^2), который характеризует качество аппроксимации с помощью линейной функции и не имеет отношения к выборочному методу.

Данная статистика основана на разложении дисперсии на межгрупповую и внутригрупповую. Это разложение также используется в дисперсионном анализе. «Первоначально (1918 г.) дисперсионный анализ был разработан английским математиком-статистиком Р.А. Фишером для обработки результатов агрономических опытов по выявлению условий получения максимального урожая различных сортов сельскохозяйственных культур. Сам термин “дисперсионный анализ” Фишер употребил позднее [2, с. 392].

Чтобы понять, на чем основано разложение дисперсии, рассмотрим так называемый «прямоугольный выборочный план», используемый в однофакторном дисперсионном анализе (табл. 1).

Этот план представляет собой таблицу, в которой каждый столбец является выборкой с n элементами. Всего делается m таких выборок. В литературе эти столбцы часто называют факторами, группами или стратами, а само расположение элементов выборок – стратификацией.

В этой статье при обозначении элемента таблицы символом y_{ij} первый индекс указывает номер строки, а второй – номер столбца, в соответствии с правилом обозначения элементов матриц, принятым в Советском Союзе. Замечу, что в английской традиции принята обратная запись, то есть сначала пишут столбец, а затем строку, а в современной российской литературе встречаются оба варианта.

Очевидно, что общее число элементов в таблице (N) есть

$$N = mn. \tag{1}$$

Таблица 1

Прямоугольный выборочный план

n \ m	1	2	…	j	…	m
1	y_{11}	y_{12}	…	y_{1j}	…	y_{1m}
2	y_{21}	y_{22}	…	y_{2j}	…	y_{2m}
⋮	…	…	…	…	…	…
i	y_{i1}	y_{i2}	…	y_{ij}	…	y_{im}
⋮	…	…	…	…	…	…
n	y_{n1}	y_{n2}	…	y_{nj}	…	y_{nm}
Среднее	$\bar{y}_1$	$\bar{y}_2$	…	$\bar{y}_j$	…	$\bar{y}_m$

По каждому столбцу вычисляется среднее арифметическое $\bar{y}_j$ (внутригрупповая средняя), которое заносится в последнюю строку таблицы,

$$\bar{y}_j = \frac{\sum_{i=1}^{n} y_{ij}}{n} \quad . \tag{2}$$

Затем вычисляется межгрупповая средняя $\bar{y}$:

$$\bar{y} = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n} y_{ij}}{mn} = \frac{1}{m}\sum_{j=1}^{m}\left(\frac{\sum_{i=1}^{n} y_{ij}}{n}\right) = \frac{\sum_{j=1}^{m}\bar{y}_j}{m} \quad . \tag{3}$$

Величины $\left|y_{ij} - \bar{y}\right|$ и $\left|y_{ij} - \bar{y}_j\right|$ называются отклонениями элемента y_{ij} от межгрупповой средней и от внутригрупповой средних соответственно, а величина $\left|\bar{y}_j - \bar{y}\right|$ – отклонением внутригрупповой средней от межгрупповой средней. Изобразим эту ситуацию графически (рис. 1). Таким образом, получился треугольник. Рассмотрим его свойства. Для этого воспользуемся теоремой косинусов для остроугольного треугольника, выполняющейся для любого треугольника, имеющего острый угол,

$$(\mathrm{y_{ij}} - \bar{y})^2 = (y_{ij} - \bar{y}_j)^2 - 2(y_{ij} - \bar{y}_j)(\bar{y}_j - \bar{y}) + (\bar{y}_j - \bar{y})^2.$$

В качестве примера рассмотрим значение первой строки и первого столбца в табл. 3–1. Межгрупповая средняя по всей табл. 2 составила $\bar{y}$ = 2,35, групповая средняя $\bar{y}_j$=2,86. Таким образом, уравнение примет следующий вид:

$(1\text{-}2{,}35)^2=(1\text{-}2{,}86)^2\text{-}2(1\text{-}2{,}86)(2{,}35\text{-}2{,}86)+ (2{,}35\text{-}2{,}86)^2$,

1,8225=3,4596-2*1,86*0,51+0,2601,

1,8225=3,4596-1,8972+0,2601 – верно.

Теперь рассмотрим, как ведет себя слагаемое $2(y_{ij} - \bar{y}_j)(\bar{y}_j - \bar{y})$ для суммы всех элементов табл. 3. Для этого переделаем табл. 3: вместо ответов респондентов будем записывать их отклонение от внутригрупповой средней $(y_{ij} - \bar{y}_j)$. Также следует отметить, что отклонение внутригрупповой средней от межгрупповой средней $(\bar{y}_j - \bar{y})$ для каждого столбца будет величиной постоянной. В последней строке запишем сумму отклонений от внутригрупповых средних, для расчетов используем MS-Excel.

Таблица 2

Отклонение от внутригрупповой средней $(y_{ij} - \bar{y}_j)$ ответов респондентов

Респондент / Вопрос	1	2	3	4	5	6	7
1	-1,88	-1,88	-1,44	-1,56	-0,44	-0,68	-1,56
2	2,14	2,14	0,57	1,43	-0,43	1,29	1,43
3	2,14	0,14	2,57	1,43	0,57	1,29	1,43
4	0,14	0,14	-1,43	-0,57	-0,43	-0,71	0,43
5	-0,86	0,14	-1,43	-1,57	0,57	-0,71	-1,57
6	0,14	1,14	2,57	1,43	-0,43	0,29	0,43
7	-1,86	-1,86	-1,43	-0,57	0,57	-0,71	-0,57
$\sum_{j=1}^{7}(y_{ij} - y_j)$	0	0	0	0	0	0	0

Из табл. 2 очевидно, что сумма отклонений от внутригрупповой средней по каждому страту равна нулю $\sum_{j=1}^{7}(y_{ij} - y_j) = 0$, а отклонение внутригрупповой средней от межгрупповой средней $(\bar{y}_j - \bar{y})$ для каждого столбца - величина постоянная, то сумма удвоенного произведения отклонений внутригрупповой средней на межгрупповую среднюю равна нулю: $2\sum_{j=1}^{m}\sum_{i=1}^{n}(y_{ij} - \bar{y}_j)(\bar{y}_j - \bar{y}) = 0$.

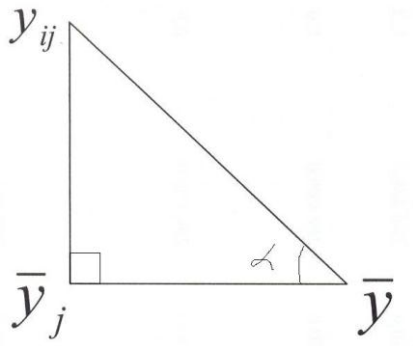

Рис. 1. Разложение вариации

Равенство нулю суммы отклонений от внутригрупповой средней по каждому страту объясняется тем, что «сумма отклонений от средней равна нулю» [3, с. 73]. Докажем это утверждение. Пусть дана выборка из n элементов. Обозначим каждый отдельный элемент выборки как y_i. Найдем среднюю по выборке как внутригрупповую по формуле (2). Тогда сумма отклонений от средней примет вид $\sum_{i=1}^{n}(y_i - \bar{y}) = \sum_{i=1}^{n} y_i - n\bar{y} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} y_i = 0$.

Таким образом, для суммы всех элементов прямоугольного выборочного плана теорема разложения вариации примет вид

$$\sum_{j=1}^{m}\sum_{i=1}^{n}(y_{ij} - \bar{y})^2 = \sum_{j=1}^{m}\sum_{i=1}^{n}(y_{ij} - \bar{y}_j)^2 - 2\sum_{j=1}^{m}\sum_{i=1}^{n}(y_{ij} - \bar{y}_j)(\bar{y}_j - \bar{y}) + \sum_{j=1}^{m}\sum_{i=1}^{n}(\bar{y}_j - \bar{y})^2$$

$$\sum_{j=1}^{m}\sum_{i=1}^{n}(y_{ij} - \bar{y})^2 = \sum_{j=1}^{m}\sum_{i=1}^{n}(y_{ij} - \bar{y}_j)^2 + \sum_{j=1}^{m}\sum_{i=1}^{n}(\bar{y}_j - \bar{y})^2 \quad (4)$$

Другими словами, из теоремы косинусов для остроугольного треугольника, выполняющейся для любого треугольника, имеющего острый угол, получится теорема Пифагора, выполняющаяся лишь для прямоугольных треугольников. Отсюда следует, что треугольник, изображенный на рис. 1, является прямоугольным.

Легко заметить, что число элементов в каждой группе в прямоугольном выборочном плане одинаковое (n) или что последнее слагаемое не зависит от i. Таким образом, формулу разложения суммы квадратов отклонений можно упростить:

$$\sum_{j=1}^{m}\sum_{i=1}^{n}(y_{ij} - \bar{y})^2 = \sum_{j=1}^{m}\sum_{i=1}^{n}(y_{ij} - \bar{y}_j)^2 + n\sum_{j=1}^{m}(\bar{y}_j - \bar{y})^2 \quad (5)$$

Это тождество и характеризует разложение дисперсии на групповую и межгрупповую.

Таким образом, словесная формулировка закона разложения дисперсии будет иметь вид: «сумма квадратов отклонений от межгрупповой средней равна сумме квадратов отклонений от внутригрупповых средних и сумме квадратов отклонений межгрупповой средней от внутригрупповых средних» [3, с. 114–118].

В литературе приняты следующие обозначения для каждого слагаемого в формуле (4):

$D_{\text{общ}} = \sum_{j=1}^{m}\sum_{i=1}^{n}(y_{ij} - \bar{y})^2$ - общая сумма квадратов отклонений от межгрупповой средней во всем прямоугольном выборочном плане; $D_{\text{внутр}} = \sum_{j=1}^{m}\sum_{i=1}^{n}(y_{ij} - \bar{y}_j)^2$ - сумма квадратов отклонений от внутригрупповых средних. $D_{\text{меж}} = \sum_{j=1}^{m}\sum_{i=1}^{n}(\bar{y}_j - \bar{y})^2$ - сумма квадратов отклонений внутригрупповых средних от межгрупповой средней.

Таким образом, символьная запись закона разложения дисперсии примет вид

$$D_{\text{общ}} = D_{\text{внутр}} + D_{\text{меж}} \quad (6)$$

Если разделить левую и правую части (5) на общее количество элементов (mn), то получим величины, называемые соответствующими дисперсиями.

$$\frac{\sum_{j=1}^{m}\sum_{i=1}^{n}(y_{ij} - \bar{y})^2}{mn} = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n}(y_{ij} - \bar{y}_j)^2}{mn} + \frac{n\sum_{j=1}^{m}(\bar{y}_j - \bar{y})^2}{mn}.$$

После упрощения выражение закона разложения дисперсии примет вид

$$\frac{\sum_{j=1}^{m}\sum_{i=1}^{n}(y_{ij}-\bar{y})^2}{mn}=\frac{\sum_{j=1}^{m}\sum_{i=1}^{n}(y_{ij}-\bar{y}_j)^2}{mn}+\frac{\sum_{j=1}^{m}(\bar{y}_j-\bar{y})^2}{m}.$$

С целью укорочения записи обычно пишут

$$S^2_{\text{общ}}=S^2_{\text{внутр}}+S^2_{\text{меж}}, \tag{7}$$

где $$S^2_{\text{общ}}=\frac{\sum_{j=1}^{m}\sum_{i=1}^{n}(y_{ij}-\bar{y})^2}{mn}$$ – общая дисперсия; (8)

$$S^2_{\text{внутр}}=\frac{\sum_{j=1}^{m}\sum_{i=1}^{n}(y_{ij}-\bar{y}_j)^2}{mn}$$ – внутригрупповая дисперсия; (9)

$$S^2_{\text{меж}}=\frac{\sum_{j=1}^{m}(\bar{y}_j-\bar{y})^2}{m}$$ – межгрупповая дисперсия. (10)

Выборочным коэффициентом детерминации ($R^2_{\text{выб}}$) называют отношение внутригрупповой дисперсии к общей:

$$R^2_{\text{выб}}=\frac{S^2_{\text{внутр}}}{S^2_{\text{общ}}}=\frac{D_{\text{внутр}}}{mn}/\frac{D_{\text{общ}}}{mn}=\frac{D_{\text{внутр}}}{D_{\text{общ}}} \tag{11}$$

Очевидно, что данная величина изменяется от 0 до 1. Чем ближе она к 1, тем более однородны группы, следовательно, сделанная выборка репрезентативней. В то же время нулю и единице данная статистика никогда не равна, поскольку в этих крайних случаях перестают выполняться предпосылки теоремы Пифагора. Треугольник на рис. 1 превратится в отрезок, и исчезнут границы между стратами.

На практике выборку считают однородной, если коэффициент выборочной детерминации больше 80% ($R^2_{\text{выб}} \geq 0{,}8$).

Следует отметить, что в статистические программные продукта не заложена стандартная функция для автоматического расчета этой статистики. По этой причине мною предлагается к рассмотрению алгоритм вычисления этой статистики с помощью программы MS-Excel. Такой выбор обоснован двумя причинами. Во-первых, MS-Excel – самый распространенный табличный процессор в России; во-вторых, использование этой статистики в большинстве случаев требует аналитической группировки, которую удобней всего проводить в табличных процессорах, а не в статистических программных продуктах.

В MS-Excel номера групп (столбцы) обозначаются латинскими буквами *A*, *B*, *C*, …, а номера элементов в группе (строки) – арабскими числами 1, 2, 3, … Все значения следует округлять не менее, третьего знака после запятой, поскольку он не влияет на точность расчета коэффициента выборочной детерминации ($R^2_{\text{выб}}$). Сам алгоритм выглядит следующим образом.

1. Заносим значения элементов исследуемых выборок согласно схеме, рассмотренной в табл 1 и 3.

2. Находим внутригрупповую среднюю (2) при помощи функции СРЗНАЧ из раздела «статистические». Ее аргументом являются *i*-е значения элементов в каждой *j*-й группе. Набрав выражение для этой функции один раз для первого столбца, для остальных столбцов набирать заново не надо, достаточно увеличить значение индекса *j* на единицу и т.д. Назовем эту строку «Среднее», как в табл. 3.

3. Находим дисперсию по каждому столбцу при помощи функции ДИСПР из раздела «статистические». Ее аргументом являются *i*-е значения элементов в каждой j-ой группе. Набрав выражение для этой функции один раз для первого столбца, для остальных столбцов набирать заново не надо, достаточно увеличить значение индекса *j* на единицу и т.д. Заносим ее значение в строку под названием «Дисперсия», которая находится под строкой «Среднее» (табл. 3).

4. После этого находим внутригрупповую дисперсию (9) и заносим ее значение в отдельную ячейку под названием «Внутригрупповая дисперсия» при помощи уже известной функции СРЗНАЧ. Ее аргументом являются все *j*-е значения в строке «Дисперсия».

5. «Межгрупповая дисперсия» (10) находится при помощи уже известной функции ДИСПР и заносится в специальную ячейку. Аргументом функции ДИСПР являются все *j*-е значения по строке «Средняя».

6. Далее находим «Общую дисперсию» (8) и заносим в отдельную ячейку. Она вычисляется при помощи функции ДИСПР, ее аргументом являются все значения элементов исследуемых выборок.

7. Следует проверить, что сумма межгрупповой и внутригрупповой дисперсий равна общей дисперсии. Ошибка не должна превышать 2-3%.

8. Считаем коэффициент выборочный детерминации (11), как отношение внутригрупповой дисперсии к общей. Его значение заносим в соответствующую ячейку.

9. Делаем вывод: если полученное значение коэффициента выборочной детерминации больше 0,8 ($R^2_{\text{выб}} \geq 0{,}8$), то выборка считается однородной, в противном случае – неоднородной.

В качестве примера рассмотрим ответы семи респондентов на вопросы об их потребности в обучении (табл. 3). Подобные опросы регулярно проводятся образовательными учреждениями Нижнего Новгорода для выяснения предпочтений абитуриентов в области экономического образования. Чаще всего, это группы второго высшего образования, Master of Business Administration (MBA) или желающие прослушать определенный набор дисциплин с целью получения сертификата. Обучение в них ведется либо за счет самих слушателей, либо организаций, в которых они работают.

Такие группы всегда немногочисленны: как правило, группа считается сформированной, если в нее входят 6 чел. Чаще всего, группа состоит из представителей среднего бизнеса, поскольку крупные организации, например Газпром, РЖД и т.п., имеют собственную образовательную инфраструктуру и проводят курсы повышения квалификации по тем дисциплинам, которые необходимы их сотрудникам. По этой причине слушатели подобных программ крайне не однородны и между собой никак не связаны: например, директор магазина по продаже домашних животных сидит за одной партой с заместителем директора сервисной службы по ремонту холодильного оборудования. В силу указанных причин образовательным учреждениям сложно планировать свою деятельность и необходимо регулярно проводить опросы потенциальных слушателей.

Анкета, по которой опрашивались респонденты (табл. 3), представляла собой семь утверждений, в зависимости от степени согласия с которыми необходимо поставить балл от 1 до 5: 1 – согласен; 2 – скорее согласен; 3 – и да и нет; 4 – скорее не согласен; 5 – не согласен.

Проводя расчеты по указанному алгоритму, получили $S^2_{\text{внутр}} = 1{,}63$, $S^2_{\text{меж}} = 0{,}27$, $S^2_{\text{общ}} = 1{,}9$. Очевидно, что формула (7) выполняется. Таким образом, коэффициент выбороч-

ной детерминации $R^2_{\text{выб}} = \frac{1{,}63}{1{,}9} \approx 0{,}86$. Отсюда следует, что группа является однородной, и на ее основе можно формировать учебную программу.

Таблица 3

Ответы респондентов

Респондент / Вопрос	1	2	3	4	5	6	7
1. Постоянное повышение образовательного уровня сегодня является обязательным условием профессионализма менеджеров	1	1	1	1	1	1	1
2. Современное российское образование мало что может дать российскому бизнесу	5	5	3	4	1	3	4
3. Основной массе обучающихся менеджеров нужны документы об окончании учебного заведения	5	3	5	4	2	3	4
4. Руководство вашего предприятия заинтересовано в обучении своих сотрудников	3	3	1	2	1	1	3
5. Получение образования способствует карьерному росту	2	3	1	1	2	1	1
6. Желание менеджера получить дополнительное образование вызывает конфликты в коллективе и ревность коллег	3	4	5	4	1	2	3
7. Предприятие для повышения конкурентоспособности должно стимулировать свой персонал к получению дополнительных знаний	1	1	1	2	2	1	2
Среднее ($\bar{y}_j$)	2,86	2,86	2,43	2,57	1,43	1,71	2,57
Дисперсия	2,41	1,84	3,10	1,67	0,24	0,78	1,39

Таблица 4

Неодинаковое число элементов в группах

Респондент / Номер элемента выборки	1	2	3	4	5	6	7
1	1	1	1	1	1	1	1
2		5	3	4	1	3	4
3		3	5		2	3	4
4		3			1		
5		3			2		
6		4					
7		1					
Среднее ($\bar{y}_j$)	1	2,86	3	2,5	1,4	2,33	3
Квадрат отклонения от внутригрупповой средней	0	12,86	8	4,5	1,2	2,67	6

Необходимо отметить, что при неодинаковом количестве элементов в каждой выборке закон разложения дисперсии (4) не выполняется. Убедимся в этом на примере, полученном из табл. 3. Перегруппируем ее таким образом, чтобы количество элементов в каждом

столбце было неодинаковым, при помощи таблицы случайных чисел [4, с. 366]. Это позволит избежать смещения при перегруппировке. Здесь необходимо отметить, что различное число элементов в столбцах можно получить и другими способами, но этот считается наиболее научно обоснованным. Возьмем из таблицы случайных чисел семь целых чисел в интервале от 1 до 7: 1; 7; 3; 2; 5; 3; 3. Таким образом, табл. 3 примет вид (табл. 4).

Общее число элементов в табл. 4 равно 24 ($N = 24$). Поскольку число элементов в группах неодинаково, то формула (1) не будет выполняться, как и формула (7). По этой причине найти коэффициент выборочной детерминации по рассмотренному алгоритму невозможно. Следовательно, все вычисления следует проводить непосредственно по формуле (6). Для этого в последнюю строку табл. 4 вместо дисперсии были внесены значения квадратов отклонений от внутригрупповых средних.

По данным табл. 4 получили $D_{\text{внутр}} = 35{,}22$; $D_{\text{меж}} = 3{,}83$; $D_{\text{общ}} = 45{,}83$. Таким образом, формула (6) не выполняется, а следовательно, не выполняется и закон разложения дисперсии. Ошибка составила $\dfrac{D_{\text{общ}}}{D_{\text{меж}} + D_{\text{внутр}}} - 1 = \dfrac{45{,}83}{35{,}22 + 3{,}83} - 1 \approx 17\%$.

В некоторых случаях эту ошибку можно снизить, если считать межгрупповую среднюю не по формуле (3), а как сумму средних по столбцам, взятую с весами. В качестве весов выступает количество элементов в каждой группе (n_j), деленное на общее число элементов (N):

$$\bar{y} = \sum_{j=1}^{m} w_j \bar{y}_j \text{, где } w_j = \frac{n_j}{N} \qquad (12)$$

Очевидно, что $\sum_{j=1}^{m} w_j = 1$.

Для расчета межгрупповой средней по формуле (12) добавим в табл. 3 еще одну строку – веса (w_j).

Таблица 5

Неодинаковое число элементов в группах с учетом весов

Респондент / Номер элемента выборки	1	2	3	4	5	6	7
1	1	1	1	1	1	1	1
2		5	3	4	1	3	4
3		3	5		2	3	4
4		3			1		
5		3			2		
6		4					
7		1					
Среднее ($\bar{y}_j$)	1	2,86	3	2,5	1,4	2,33	3
веса (w_j)	$\frac{1}{24}$	$\frac{7}{24}$	$\frac{3}{24}$	$\frac{2}{24}$	$\frac{5}{24}$	$\frac{3}{24}$	$\frac{3}{24}$
Квадрат отклонения от внутригрупповой средней	0	12,86	8	4,5	1,2	2,67	6

По данным табл. 4 межгрупповая средняя составит $\bar{y} = 2{,}42$ (12), внутригрупповая дисперсия не изменится: $D_{\text{внутр}} = 35{,}22$; $D_{\text{меж}} = 3{,}93$; $D_{\text{общ}} = 44$.

Таким образом, ошибка составит

$$\frac{D_{\text{общ}}}{D_{\text{меж}} + D_{\text{внутр}}} - 1 = \frac{44}{35{,}22 + 3{,}93} - 1 \approx 12\%.$$

В то же время всю вариацию в данном случае можно расписать непосредственно через теорему косинусов для остроугольного треугольника, исходя из свойств треугольника, изображенного на рис. 1, который при неодинаковом количестве элементов в группах не будет прямоугольным. Для этого найдем межгрупповую среднюю для табл. 4 по формуле (3) $\bar{y} = 2{,}33$. Составим табл. 6, в которую занесем значения квадратов внутригрупповых отклонений, межгрупповых отклонений и отрицательное удвоенное произведение внутригруппового отклонения на межгрупповое для каждого из 24 элементов выборки.

Таблица 6

Разложение вариации при неодинаковом числе элементов в группах

№ п/п	Номер в таблице 4 *(i,j)*	$(y_{ij} - \bar{y}_j)^2$	$-2(y_{ij} - \bar{y}_j)(\bar{y}_j - \bar{y})$	$(\bar{y}_j - \bar{y})^2$	$(y_{ij} - \bar{y})^2$
1	1,1	0	0	1,7689	1,7689
2	1,2	3,4596	-1,9716	0,2809	1,7689
3	2,2	4,5796	2,2684	0,2809	7,1289
4	3,2	0,0196	0,1484	0,2809	0,4489
5	4,2	0,0196	0,1484	0,2809	0,4489
6	5,2	0,0196	0,1484	0,2809	0,4489
7	6,2	1,2996	1,2084	0,2809	2,7889
8	7,2	3,4596	-1,9716	0,2809	1,7689
9	1,3	4	2,68	0,4489	1,7689
10	2,3	0	0	0,4489	0,4489
11	3,3	4	2,68	0,4489	7,1289
12	1,4	2,25	-0,51	0,0289	1,7689
13	2,4	2,25	0,51	0,0289	2,7889
14	1,5	0,16	0,744	0,8649	1,7689
15	2,5	0,16	0,744	0,8649	1,7689
16	3,5	0,36	-1,116	0,8649	0,1089
17	4,5	0,16	0,744	0,8649	1,7689
18	5,5	0,36	-1,116	0,8649	0,1089
19	1,6	1,7689	0	0	1,7689
20	2,6	0,4489	0	0	0,4489
21	3,6	0,4489	0	0	0,4489
22	1,7	4	-2,68	0,4489	1,7689
23	2,7	1	1,34	0,4489	2,7889
24	3,7	1	1,34	0,4489	2,7889
	Сумма:	**35,2239**	**5,3388**	**10,8109**	**46,0136**

Таким образом, по данным табл.6 имеем: $D_{\text{внутр}} = 35{,}22$; $D_{\text{меж}} = 10{,}81$; $D_{\text{общ}} = 46$, то есть формула (6) выполняется с небольшой погрешностью, но слагаемое $-2\sum_{j=1}^{m}\sum_{i=1}^{n}(y_{ij} - \bar{y}_j)(\bar{y}_j - \bar{y})$ в нее не входит.

Следует добавить, что последние 20 лет коэффициент выборочной детерминации $R^2_{\text{выб}}$ при проведении научных исследований в отечественной литературе не используется. Причина кроется в том, что в Советском Союзе выборочными исследованиями занимались лишь органы плановой экономики (Центральное Статистическое Управление СССР), а с их

ликвидацией необходимая для этого ресурсная база была утеряна.

Разделим формулу (6) на общую сумму квадратов отклонений от межгрупповой средней ($D_{\text{общ}}$):

$$\frac{D_{\text{внутр}}}{D_{\text{общ}}} + \frac{D_{\text{меж}}}{D_{\text{общ}}} = 1 \tag{13}$$

Как уже было сказано, отношение $\frac{D_{\text{внутр}}}{D_{\text{общ}}}$ называется коэффициентом выборочной детерминации $R^2_{\text{выб}}$. Величина $\frac{D_{\text{меж}}}{D_{\text{общ}}}$ называется эмпирическим корреляционным отношением (или просто корреляционным отношением) и обозначается η^2:

$$\eta^2 = \frac{D_{\text{меж}}}{D_{\text{общ}}} \tag{14}$$

«Эмпирическое корреляционное отношение η^2 является показателем рассеяния точек корреляционного поля относительно эмпирической линии регрессии [2, с. 436]. Обычно оно используется для оценки подгонки прямой (кривой) при условии стратификации наблюдений (расчетах по корреляционной таблице). Более подробную информацию по этому вопросу можно найти в книге [2, с. 435-440].

Эта статистика также изменяется от нуля до единицы. Чем ближе она к 1, тем точнее степень подгонки прямой (кривой). В большинстве случаев эмпирическое корреляционное отношение η^2 выполняет ту же функцию, что и коэффициент детерминации R^2. Отличие состоит в том, что эмпирическое корреляционное отношение η^2 позволяет оценить степень подгонки кривой при условии стратификации данных, что нельзя сделать при помощи коэффициента детерминации R^2.

Таким образом, формула (13) принимает вид

$$R^2_{\text{выб}} + \eta^2 = 1 \tag{15}$$

Перепишем формулу (15) в отклонениях от средних для одного элемента:

$$\frac{(y_{ij} - \bar{y}_j)^2}{(y_{ij} - \bar{y})^2} + \frac{(\bar{y}_j - \bar{y})^2}{(y_{ij} - \bar{y})^2} = 1 \tag{16}$$

Рассмотрим рис. 1. Из него становится очевидно, что $(y_{ij} - \bar{y})^2$ - квадрат гипотенузы прямоугольного треугольника; $(y_{ij} - \bar{y}_j)^2$ - квадрат противолежащего катета; $(\bar{y}_j - \bar{y})^2$ - квадрат прилежащего катета, но $\sin^2 \alpha = \frac{(y_{ij} - \bar{y}_j)^2}{(y_{ij} - \bar{y})^2}$, где $\alpha = \angle \bar{y}_j \bar{y} y_{ij}$ для треугольника, изображенного на рис. 1.

С другой стороны, отношение прилежащего катета к гипотенузе выражается тригонометрической функцией $\cos^2 \alpha = \frac{(\bar{y}_j - \bar{y})^2}{(y_{ij} - \bar{y})^2}$, где $\alpha = \angle \bar{y}_j \bar{y} y_{ij}$ для треугольника на рис. 1.

Таким образом, для суммы всех элементов выборок $R^2_{\text{выб}} = \sin^2 \alpha$, а $\eta^2 = \cos^2 \alpha$.
По этой причине формула (15) представляет собой основное тригонометрическое тождество угла $\alpha = \angle \bar{y}_j \bar{y} y_{ij}$ треугольника, изображенного на рис. 1: $\sin^2 \alpha + \cos^2 \alpha = 1$.

В заключение следует отметить, что в современной Росси почти не издается литера-

тура по выборочному методу, являвшемуся в XX в. «основным орудием» социально-экономического исследования. В Советском Союзе на его основе строились многоэтажные дома, магазины, детские сады... Сейчас же обосновать рентабельность строительства того или иного сооружения сложнее, чем его построить. Отсюда постоянная социальная напряженность, разговоры о том, что «власть ничего не знает о нуждах народа».

Другим важнейшим приложением выборочного метода являлось страхование, которое за последние годы стало убыточным для осуществляющих его организаций. Причина этого кроется в неправильном формировании страховых тарифов, базирующихся на эмпирических данных. В XX в. эта проблема также решалась при помощи социально-экономических исследований.

В современных образовательных стандартах выборочный метод вообще не прописывается как отдельная дисциплина, хотя является эмпирическим основанием теории вероятностей, математической статистики и эконометрики, что делает их практически бесполезными выпускникам вузов. Тем не менее, в 20-е годы XX в. этот метод входил в школьную программу.

Подводя итог всему, хотелось бы, чтобы статья не только знакомила читателя с математически обоснованным способом проверки релевантности данных, но и стимулировала его к использованию математики как прикладной науки.

Библиографический список

1. **Качанов, Ю.Л.** Алгоритм построения представительной территориальной выборки при наличии ограничений на стоимость измерений // Вопросы социологии. – 1992. – Т. 1. – №1. – С. 135–149.
2. **Кремер, Н. Ш.** Теория вероятностей и математическая статистика / Н. Ш. Кремер. – 2-е изд., исп., доп. – М.: Юнити, 2004.
3. **Елисеева, И. И.** Общая теория статистики / И. И. Елисеева, М. М. Юзбашев. – М.: Финансы и статистика, 1995.
4. **Большев, Л. Н.** Таблицы математической статистики / Л. Н. Большев, Н. В. Смирнов. – М.: Наука, 1983.

Дата поступления
в редакцию 10.05.2016

D. E. Krasilnikov

THE ALGORITHM OF SAMPLE DETERMINATION COEFFICIENT CALCULATION IN MS-EXCEL

Nizhegorodskiy pochtamt Otdeleniye pochtovoy svyazi №24

The article disposes a widely used in former Soviet Union statistic Coefficient of Sample Determination as a criterion for sample homogeneity in social and economic research and its algorithm of calculation in MS-Excel. The geometrical proof of the law of variance decomposition is proposed and the case then it fails is concerned. The liaison between the Coefficient of Sample Determination and Empirical Correlation Ratio is depicted.

Key words: coefficient of sample determination, the law of variance decomposition, MS-Excel, criterion for sample homogeneity, analysis of variance, empirical correlation ratio.