ОЦЕНКА СОГЛАСОВАННОСТИ СУБЪЕКТИВНЫХ КЛАССИФИКАЦИЙ ПРИ ЗАДАННЫХ КЛАССАХ

А. А. Заславский, Т. А. Пригарина

(Москва)

Предлагается статистический метод оценки согласованности m субъективных классификаций, когда классы, к которым относятся объекты, заданы заранее. Вводится статистика парной согласованности которая представляет собой модификацию известной каппа-статистики. Общий коэффициент согласованности для случая m> 2 определяется как среднее значение парных статистик. Наблюдаемая величина коэффицента признается значимой, если отвергается нулевая гипотеза о статистической независимости и случайности классификаций. При нулевой гипотезе исследуются точные и предельные распределения введенных статистик. Учитывается, что классы могут быть упорядоченными и неупорядоченными. Приводимые таблицы квантилей точных распределений и аппроксимационные результаты позволяют применить предложенный метод на практике.

Ключевые слова: субъективная классификация, таблица сопряженности, согласованность классификаций, метод оценки, каппа-статистика, распределение.

1. Постановка задачи

В социологических обследованиях респондентов часто просят отнести указанные объекты к одному из заданных классов. В ка -

честве примера приведем вопрос Службы общественного мнения VP (руководитель - профессор Б. А. Грушин) [1,с. 5], который предлагался по телефону в июле 1991 года: "Как сказались минувшие события на авторитете Э. А. Шеварднадзе (В. В. Жириновского, М. С. Горбачева) - он значительно повысился, несколько повысился, не изменился, понизился или значительно понизился? ". Ответ на этот вопрос представляет собой классификацию известных политических деятелей (объектов), причем классы или категории, характеризующие в данном случае изменение рейтинга, заданы заранее, и каждый объект может быть отнесен только к одной категории.

Анализ ответов такого типа обычно осуществляется визуально по таблице, содержащей данные о процентах лиц, которые поместили очередной объект в тот или иной класс. Расчет может производиться и для отдельных подгрупп респондентов - возрастных, профессиональных и т. д. Проблема согласованности полученных ответов в пределах всей выборки или внутри ее подгрупп, как правило, не рассматривается. Между тем с точки зрения выяснения групповых позиций она представляет несомненный интерес. Если классификации респондентов из какой-нибудь группы достаточно согласованны, можно утверждать, что существует общее мнение, которое с незначительными индивидуальными отклонениями разделяет большинство ее членов. Наоборот, при низком согласии единое мнение-"представитель" выделить нельзя. В этом случае иногда удается разбить множество респондентов на внутренне согласованные подмножества - коалиции и рассматривать набор точек зрения, присущих коалициям.

Когда классифицируется единственный объект, оценить степень разброса ответов нередко можно непосредственно - по данным о частотах выбора различных классов. Сложнее обстоит дело, если объектов несколько. В качестве условного примера приведем таблицу распределения гипотетических ответов ста народных депутатов, указавших тип изменения рейтинга политиков X, Y и Z (табл. 1).

Таблица 1

Частоты выбора различных категорий для X, Y, Z (%)

Тип изменения рейтинга	Политики		
	X	Y	Z
Значительно повысился	15	12	10
Несколько повысился	44	19	34
Не изменился	20	21	38
Понизился	15	38	15
Значительно понизился	6	10	3
Итого	100	100	100

Пользуясь только этой таблицей, не так просто установить, являются ли мнения депутатов похожими и можно ли говорить о более или менее единой позиции группы по отношению к всем трем политикам. Общепринятый метод оценки согласованности заключается в построении некоторого показателя близости ответов и проверке гипотезы о том, что наблюдаемое значение показателя могло бы быть превышено со значимой вероятностью хаотичными, случайными ответами. Если гипотеза отвергается, мнения респондентов признаются похожими или согласованными. По-видимому, "неслучайность" согласия является минимальным требованием к его уровню. На практике могут устанавливаться более жесткие пороговые значения показателя близости.

Описанный метод широко применяется для анализа связи численных оценок объектов (расчет коэффициентов парной и множественной корреляции) и при проверке сходства ранжировок (коэффициенты Кендалла и Спирмена, коэффициент конкордации). В данной статье он развивается применительно к классификациям с заданными классами. Отличие предлагаемых показателей от известных коэффициентов связи для номинальных и ординальных признаков (см., например, [2; 3; 4,с. 62-74; 5,с. 50-54]) состоит в том, что этими показателями измеряется связь определенного типа: степень совпадения и (или) близости между классами, к которым отнесены объекты в альтернативных классификациях. Если имеются объективные признаки с одинаковым числом категорий, с помощью предложенных показателей определяется степень взаимооднозначного соответствия между категориями разных признаков.

2. Оценка согласованности двух классификаций

Частный случай, когда имеется всего два респондента, лежит в основе разработки общего метода. Кроме того, этот случай представляет самостоятельный интерес, так как для выделения коалиций необходима информация о степени согласованности каждых двух мнений.

Пусть респонденты А и В получили задание распределить n объектов по k категориям. Результат их работы может быть представлен матрицей $T = [t_{ij}]_{k \times k}$, где t_{ij} - число объектов, которые были отнесены к категории i респондентом А и к категории j респондентом В, i, j=1,..., k; $\sum_i \sum_j t_{ij} = n.$ Матрицу Т обычно называют таблицей сопряженности. Если расхождения по всем объектам решено учитывать одинаковым образом, знания Т достаточно для исследования проблемы.

Показатель согласованности классификаций может быть введен, исходя из следующих простых соображений. Предположим, имеется множество чисел $W = \{ w_{11}, w_{12} \ldots w_{k, k-1}, w_{kk}\}$, где w_{ij} - оценка близости, которая приписывается решениям респондентов А и В, если они помещают очередной объект в классы i и j соответственно. Чем выше близость, тем больше приписываемое число. В качестве общей исходной характеристики близости P_0 возьмем сумму оценок, набранных респондентами

$$P_0 = P_0 (W) = \sum_i \sum_j w_{ij} t_{ij} . \quad (1)$$

Когда объекты классифицируются наугад, характеристика P_0 может рассматриваться как случайная величина с ожидаемым значением

$$P_e = E (P_0). \quad (2)$$

Показателем согласованности классификаций æ будем считать нормированное отклонение P_0 от этого ожидаемого значения, т. е.

$$æ = æ (W) = \frac{P_0 - Pe}{(\max P_0) - Pe}. \quad (3)$$

В литературе æ получил название каппа-статистики, так как уже сам его расчет предполагает выдвижение некоторой статистической гипотезы о случайности ответов. Впервые он был введен Дж. Коэном в [6,7] и затем независимо - Г. Раушенбахом и А. Заславским [8, с. 126-141]. В важных частных случаях характеристики типа P_0 предлагались авторами работ [9,10]. Относительно полная сводка зарубежных результатов на 1985 г. содержится в [11].

При максимальном сходстве классификаций значение æ равно единице, а в ситуации, когда оба респондента отвечают наугад, колеблется в окрестности нуля. Минимум этой статистики зависит от конкретной системы оценок w, которая задается исследователем из содержательных соображений. В дальнейшем, следуя установившейся традиции, оценки w_{ij} мы будем называть весами. Отметим одно полезное свойство æ, которое, по-видимому, не было замечено ранее.

Утверждение 1. *Каппа-статистика инвариантна относительно линейного преобразования весов, т. е.*

$$æ\,(W') = æ\,(W). \qquad (4)$$

где $W' = \{ W'_{ij} \mid W'_{ij} = aw_{ij} + b;\ i, j=1,..., k;\ a> 0 \}$.

Используя линейное преобразование, минимальный вес любой системы можно сделать нулевым, а максимальный - равным единице. Далее везде будем придерживаться этого соглашения, заметив, что ему соответствует определенное значение максимума P_0: $max\ P_0 = (max\ w_{ij}) \cdot n = n$.

Модели случайных ответов могут быть различными. Мы будем рассматривать *гипотезу С*, *согласно которой каждый респондент независимо от другого помещает очередной объект в любой из классов с вероятностью* 1/k:

$$\Pr \{ \text{s-й } \textit{объект отнесен к категории } \text{i} \} = 1/k, \qquad (5)$$

для всех s, i; s=1,2... n; i=1,..., k.

Если респонденты действуют в соответствии с (5), вероятность попадания произвольного объекта в i-ю строку и j-й столбец матрицы T равна $1/k^2$ и, следовательно, число таких объектов t_{ij} - биномиально распределенная случайная величина с параметрами распределения $(1/k^2,n)$. Отсюда находим значение P_e в случае нашей гипотезы

$$P_e = E(\sum_{i,j} w_{ij} t_{ij}) = \sum_{i,j} w_{ij} E(t_{ij}) = \sum_{i,j} w_{ij} (n/k^2) = (n/k^2) \sum_{i,j} w_{ij}. \qquad (6)$$

Подставляя его в (3), получаем вариант каппа-статистики

$$æ^c = \frac{k^2 \sum_{i,j} w_{ij} t_{ij} - n_i \sum_{i,j} w_{ij}}{k^2 n - n_i \sum_{i,j} w_{ij}}. \qquad (7)$$

где $æ^c$ - показатель согласованности для данного набора объектов.

В большинстве зарубежных работ имеющиеся объекты рассматриваются как выборка из генеральной совокупности, и уровень согласия между субъектами оценивается для совокупности в целом. Этой постановке соответствует другой вариант каппа-статистики. Подробнее мы остановимся на этом в разд. 4.

Фактическое значение статистики (7), рассчитанное по полученным данным, признается существенным, если вероятность получить такое же или большее значение в рамках гипотезы *C* пренебрежимо мала. Для проверки существенности необходимо знать соответствующее распределение $æ^c$ при заданных n, k и W. Прежде всего укажем, что оно асимптотически нормально.

Утверждение 2. *Если выполняется гипотеза C, при* $n \to \infty$ *распределение величины* $æ^c / \sigma_{æ}$, *где* $\sigma_{æ}$ - *дисперсия* $æ^c$, *сходится к нормальному распределению, где*

$$\sigma_{æ} = \mathrm{Var}(æ^c) = \frac{k^2 \sum w_{ij}^{\ 2} - (\sum w_{ij})^2}{n(k^2 - \sum w_{ij})^2}. \qquad (8)$$

Если число объектов достаточно велико, порядка двухсот и выше, данный результат позволяет оценивать согласованность классификаций при любой системе W. Чтобы определить точность нормальной аппроксимации для умеренных значений и уметь проверять согласованность при малых значениях n, нужно знать точные распределения $æ^c$. Осуществить их аналитический расчет в общем случае произвольных весов затруднительно. К счастью, в большинстве практических задач рассматриваются всего два типа близости между классификациями. Они формализуются двумя системами весов, при которых точные распределения могут быть найдены.

Важным свойством любой классификации является наличие или отсутствие упорядоченности классов. Классы (категории) могут быть упорядочены по любому признаку. Пример из разд. 1 демонстрирует классификацию с упорядоченными категориями. Такие классификации мы будем называть *сортировками*. Если категории не соответствуют уровням интенсивности какого-либо признака или такое соответствие не существенно, будем говорить о *разбиении* объектов. Примером разбиения может служить определение респондентами авторства литературных отрывков, когда список возможных авторов дается заранее.

Понятно, что близость между сортировками и между разбиениями должна измеряться по-разному. Рассмотрим эти случаи по отдельности.

1. Оценка согласованности разбиений. Если нет каких-то дополнительных содержательных соображений, при назначении весов естественно учитывать только две возможности: совпадение и несовпадение категорий, выбранных для объекта. Когда решения респондентов совпадают, им приписывается единичный вес, а при несовпадении - нулевой:

$$\begin{cases} w_{ii} = 1, \\ w_{ij} = 0, i \neq j. i, j = 1,..., k. \end{cases}$$

Из (1) видно, что характеристика P_0 окажется в этом случае общим числом одинаковых решений: $P_0 = \sum_i t_{ii}$. Каппа-статис -

тика (7), которую в качестве показателя согласованности разбиений мы обозначим $æ^c$, примет следующий вид

$$\overline{æ}^c = \frac{\sum_i t_{ii} - n}{kn - n}. \qquad (9)$$

Минимум этого показателя, как это легко заметить, равен -1/ (k-1), а выражение для дисперсии находится из (8): Var ($æ^c$) = 1/ (n· (k-1)). С точки зрения проверки гипотезы *C* статистики P_0 и $æ^c$ эквивалентны. При введенных весах P_0 представляет собой число "попаданий" объектов на главную диагональ матрицы T, причем каждое "попадание" не зависит от других и происходит с вероятностью 1/к. Таким образом, эта величина имеет биномиальное распределение с параметрами (1/к, n). Если ее фактическое значение превышает верхнюю α%-ю точку данного распределения, где α - допустимая вероятностм ошибки (уровень значимости), то соответствующее значение $æ^c$ признается существенным, а разбиение - согласованными. Публикации подробных таблиц биноминального распределения и его процентных точек указаны в [12].

2. Оценка согласованности сортировок.

Будем считать, что номера классов соответствуют их местам в имеющемся упорядочении. Чем "дальше" друг от друга места тех категорий, к которым отнесен объект, тем меньшее сходство решений демонстрируют респонденты. Поэтому кажется разумным, что веса w_{ij} должны расходиться в обратной зависимости к величине |i-j|. Авторами работы [10] предложена линейная форма зависимости

$$w_{ij} = 1 - \frac{|i-j|}{k-1}, \quad \text{i, j=1,2…k.} \qquad (10)$$

В силу утверждения 1 любая система "линейных" весов вида $a_1|i-j|+ a_2$, где $a_1>0$, может быть сведена к (10). Эту форму мы и будем использовать в дальнейшем.

Обозначим $t^{(r)}$ число объектов, для которых модуль разности номеров их категорий равен r, а именно

$$t^{(r)} = \sum_{i,j} t_{ij}, \text{ где } |i-j| = r, 0,1,\ldots,k-1.$$

Подставляя веса (10) в (7), получим показатель согласованности сортировок

$$\vee$$

$$æ^c = 1 - \frac{3k}{n(k^2-1)} P_0', \text{ где } P_0' = P_0'(n) = \sum_{r=0}^{k-1} r t^{(r)}. \qquad (11)$$

$$\vee \qquad —$$

При к=2 показатели $æ^c$ и $æ^c$ совпадают. Минимум $æ^c$ достигается, когда по всем объектам мнения респондентов полярны ($t^{(k-1)} = n$), и равен -(2k-1) / (k+1). С ростом k он уменьшается, приближаясь к-2. Вычислив сумму весов (10) и сумму их квадратов, по формуле (8) рассчитываем дисперсию введенной статистики

$$\vee$$

$$\mathrm{Var}(æ^c) = \frac{k^2+2}{2n(k^2-1)}.$$

$$\vee$$

Чем меньше величина P_0', тем больше $æ^c$. Для проверки того, что положительное фактическое значение этого показателя неслучайно, нужно знать верхние процентные точки его распреде -

ления при гипотезе *C*. Им соответствуют нижние процентные точки статистики P_0', таблицы которых можно составить, основываясь на следующих соображениях.

Пусть респонденты, действуя согласно (5), отнесли некоторый объект к категориям i и j. Обозначим вероятность события |i-j|=r через π_r. Легко проверить, что

$$\pi_r = \begin{cases} 1/k, r = 0, \\ 2(k-r)/k^2, r = 1,....,k-1. \end{cases} \qquad (12)$$

Когда классифицируется единственный объект (n=1), P_0' - это модуль разности номеров категорий, выбранных для данного объекта, и следовательно

$$\Pr\left\{P_0'(1) = r\right\} = \pi_r, r = 0,1,....,k-1. \qquad (13)$$

Таким образом, при n=1 распределение P' известно.

Утверждение 3. *Если* n≥2, *и* x - *произвольное целое число*, *то*

$$\Pr\left\{P_0'(n) = x\right\} = \sum_{r=0}^{k-1} \pi_r \Pr\left\{P_0'(n-1) = x-r\right\}. \qquad (14)$$

Нетрудно показать, что статистика P_0' (n) принимает с положительной вероятностью любое целочисленное значение в диапазоне от нуля до своего максимума, равного n (к-1). Соотношение (14) позволяет рассчитывать распределение P_0' (n) в этом диапазоне по распределению P_0' (n-1). Нами были проведены соответствующие расчеты, в результате которых составлены таблицы процентных точек P_0' для n=1 (1) 200, к=2 (1) 10 и шести уровней значимости: 0,001; 0,005; 0,01; 0,025; 0,05; 0,10.

Если ξ_ϕ - наблюдаемое значение статистики P_0', а ξ_α - его нижняя процентная точка, соответствующая принятому уровню значимости α, то при $\xi_\phi \le \xi_\alpha$ гипотеза *C* отвергается и сортировки признаются согласованными. При $\xi_\phi > \xi_\alpha$ констатируется рассогласованность.

Так как распределение P_0' асимптотически нормально, то начиная с некоторого n точная проверка существенности ξ_ϕ может быть заменена приближенной. В этом случае рассчитывается стандартизированное значение

$$Z_\phi = \frac{\xi_\phi - E(P_0') + 1/2}{\sqrt{Var(P_o')}}, \qquad (15)$$

где

$$E(P_0') = \frac{n(k^2-1)}{3k}, \qquad (16)$$

$$\mathrm{D}(P_0') = \frac{n(k^2-1)(k^2-1)}{18k^2}, \qquad (17)$$

а 1/2 - величина так называемой поправки на непрерывность.

Заметим, что (16) и (17) непосредственно получаются с учетом (11) из известных выражений для математического ожидания и дисперсии æ^c. Затем значение Z_ϕ сопоставляется с нижней α%-й точкой Z_ϕ стандартного нормального распределения. При $Z_\phi \le Z_\alpha$ констатируем согласованность и наоборот.

Приближенная проверка проще, так как не требует использования специальных таблиц. Чтобы оценить величину n, начиная с которой аппроксимацию можно считать удовлетворительной, при ni10 для всех табулированных процентных точек ξ_α точного распределения P_0' рассчитывалась разность

$$\Delta\alpha = \Delta\alpha\ (n.\ k) = \xi_\alpha - [\xi_\alpha{}^*], \qquad (18)$$

где $\xi_\alpha^* = Z_\alpha\sqrt{\mathrm{Var}(P_0')} + E(P_0')$ - значение статистики P_0', соответствующее (без учета дисперсии) нормальной точке Z_α; $[\xi]$ - целая часть числа ξ.

Модуль $\Delta\alpha$ - это количество значений P_0', относительно которых в случае приближенной проверки будут сделаны ошибочные выводы. Если $\Delta\alpha<0$, то при $\xi_\alpha<\xi_\phi\le[\xi_\alpha^*]$ произойдет ложное отвержение гипотезы *C*, а при $\xi_\alpha\ge\xi_\phi>[\xi_\alpha^*]$, когда $\Delta\alpha>0$, приближенный тест окажется более консервативным, и значение ξ_ϕ будет ошибочно объявлено несущественным.

Приведенные вычисления показали следующее. Во-первых, во всех рассмотренных случаях разность (18) оказалась неотрицательной; таким образом ложная констатация согласованности из-за ошибки приближения исключается. Во-вторых, в подавляющем большинстве случаев $\Delta\alpha\le2$, а при $\alpha\ge0{,}01$ (т. е. для уровней значимости, равных 0,01; 0,025; 0,05 и 0,10) величина $\Delta\alpha$ за несколькими исключениями не превышает единицы. При $\alpha=0{,}05$ число единичных значений по всем сочетаниям параметров n и k меньше 1/3, а при $\alpha=0{,}10$ - примерно равно 15%. Для остальных сочетаний $\Delta\alpha=0$. Как видим, в практических расчетах, когда n>10, с полным основанием могут использоваться формулы (15) - (17) и таблицы нормального распределения. Если $n\le10$, процентные точки ξ_α для α, равного 0,01 и 0,05 можно определить по таблице П1, которая приведена в Приложении.

В заключение отметим, что статистика $æ^c$ имеет еще одну область применения. Предположим, отвечая на некоторый вопрос, респонденты должны были выбирать единственный из k вариантов ответ, например, один из трех: "да", "нет", "не знаю". Предположим также, что они были заранее разбиты на три группы (в общем случае на k каких-то групп): "мужчины", "женщины", "дети". Если исследователь считает, что каждая из выделенных групп должна склоняться к определенному варианту ответа, скажем, мужчины - "нет", женщины - "да", дети - "не знаю", то оценить степень такого соответствия можно при помощи предложенной статистики. Объектами при этом окажутся респонденты, а вместо согласованности альтернативных классификаций будет оцениваться согласованность классификаций, осуществленных по двум разным признакам. Подчеркнем, что речь идет о согласованности определенного типа.

3. Случай m классификаций

Рассмотрим теперь случай, когда имеется m альтернативных классификаций (m> 2). В качестве показателя согласованности естественно использовать величину

$$V = \frac{1}{m(m-1)} \sum_{\alpha \neq \beta} æ^{c}_{\alpha\beta}(W), \qquad (19)$$

где $æ^{c}_{\alpha\beta}$ (W) рассчитывается по формуле (7) для классификаций с номерами α и β.

Чтобы получить удобные формулы вычисления V для разбиений и сортировок, отметим прежде всего следующее полезное свойство статистики $æ^{c}$.

Утверждение 4. *Пусть от двух респондентов получены классификации двух непересекающихся совокупностей объектов объема* n_1 *и* n_2. *Показатели согласованности* $æ^{c}_1$ *и* $æ^{c}_2$ *рассчитаны для каждой из совокупностей по формуле (7). Тогда показатель согласованности объединенной совокупности равен*

$$æ^{c} = (n_1 æ^{c}_1 + n_2 æ^{c}_2) / (n_1 + n_2).$$

Перейдем теперь к вычислению V.

Утверждение 5. *Для разбиений*

$$V = 1 - \frac{nm^2 - \sum_{s=1}^{n} \sum_{j=1}^{k} m^2_{sj}}{\qquad}, \qquad (20)$$

где m_{sj} - *число респондентов, отнесших объект* s *в категорию* j.

Как видно из (20), для оценки согласованности можно пользоваться линейно связанной с V суммой $\Sigma\Sigma m^2_{sj}$, которую мы обоз -

начим s. Отметим, что к s как статистике для проверки гипотезы *C* можно прийти и исходя из схемы дисперсионного анализа. Действительно, если n=1, ожидаемое значение m_{sj} равно m/k, поэтому величина S $(m_{sj}-m/k)^2/(m/k)$ в пределе при $m \to \infty$ имеет распределение X^2 (Хи-квадрат), т. е.

$$\sum_{j=1}^{k} \frac{(m_{sj} - m/k)^2}{(m/k)} = (k/m)S - m \sim X^2_{k-1}.$$

Когда n>1, $S=S_1+\ldots + S_n$, где $S_1 = \Sigma m^2_{lj}$, $1 \le l \le n$. Как известно, свертка n распределений X^2_{k-1} дает распределение X^2_{nk-1}. Таким образом доказано следующее утверждение.

Утверждение 6. *При* $m \to \infty$ (k/m) S - nm ~ X^2_{nk-1}. Отсюда следует, что при больших m для проверки гипотезы *C* можно пользоваться распределением X^2. Однако при m <20 точность аппроксимации невысока, поэтому следует обращаться к точному распределению статистики S. Таблицы критических точек S приводятся в [8].

Перейдем теперь к анализу сортировок. **Утверждение 7.** *Для* m *сортировок единственного объекта* s ($1 \le s \le n$)

$$V_s = 1 - \frac{6k}{m(m-1)(k^2-1)} \sum_{i=1}^{k-1} M_{si}(m - M_{si}), \qquad (21)$$

где $M_{si} = \sum_{j=1}^{i} m_{si}$.

Отметим следующие свойства V_s, непосредственно вытекающие из (21):

1) максимальное значение V_s, равное 1, достигается при совпадении всех сортировок;

2) при справедливости гипотезы о случайности V_s принимает значения, близкие к нулю;

3) при наличии двух групп сортировок, далеких друг от друга, V_s принимает отрицательные значения. Минимальное значение V_s достигается при $m_{s1} = m_{sk} = m/2$, $m_{s2} =... = m_{s,k-1} = 0$. Эти свойства позволяют не только проверять гипотезу С, но и делать выводы о распределении оценок респондентов. Так, если значение V_s меньше нижнего критического, можно сделать вывод о наличии двух далеких друг от друга групп респондентов, если же оно больше верхнего критического, мнения респондентов можно считать достаточно согласованными.

Пользуясь (19) и (21), можно оценить моменты распределения V_s и показать, что если k> 2, величина $\sqrt{mV_s}$ при $m \to \infty$ имеет асимптотически нормальное распределение с нулевым средним и дисперсией $(k^2-4) /[5(k^2-1)]$. Однако при m<25 точность аппроксимации невысока, поэтому следует пользоваться точным распределением, критические точки которого приведены в Приложении в таблице П2. При k=2 асимптотика аналогична неупорядоченному случаю.

В общем случае, когда n> 1, коэффициент V для сортировок равен среднему арифметическому коэффициентов, рассчитанных для каждого объекта

$$V = 1 - \frac{6k}{nm(m-1)(k^2-1)} \sum_{s=1}^{n} \sum_{i=1}^{k-1} M_{si}(m - M_{si}), \qquad (22)$$

Поэтому указанные выше свойства сохраняются за исключением того, что отрицательные значения V позволяют сделать вывод не о наличии двух групп респондентов с разными токами зрения, а лишь о поляризации мнений по каждому объекту, так как разбиения респондентов на две группы могут быть различными для разных объектов.

Поскольку распределение V_s является асимптотически нормальным, распределение V также стремится к нормальному с дисперсией $(k^2-4) /[5nm (k^2-1)]$.

В заключение воспользуемся статистикой V для оценки согласованности респондентов в примере из разд. 1. В этом приме -

ре категории упорядочены, их число равно пяти. Таким образом, n=3, k=5, m=100.

В соответствии с утверждением 5 общий коэффициент согласованности равен среднему арифметическому частных коэффициентов: $V = (V_x+V_y+V_z)/3$. Вычислим их по (22):

$$V_x = 1 - \frac{6\times 5}{(5^2-1)\times 100\times 99}(15\times 85 + 59\times 41 + 79\times 21 + 94\times 6) = 0{,}253;$$

$$V_y = 1 - \frac{6\times 5}{(5^2-1)\times 100\times 99}(12\times 88 + 31\times 69 + 52\times 48 + 90\times 10) = 0{,}168;$$

$$V_z = 1 - \frac{6\times 5}{(5^2-1)\times 100\times 99}(10\times 90 + 44\times 56 + 82\times 18 + 97\times 3) = 0{,}352.$$

Таким образом, V = 0,258. Поскольку число респондентов достаточно велико, можно считать, что каждый частный коэффициент имеет нормальное распределение с нулевым средним и дисперсией

$$\sigma^2 = \frac{1}{100}\cdot\frac{5^2-4}{5(5^2-1)} = 0{,}00175.$$

Тогда распределение общего коэффициента V также нормально с нулевым средним и дисперсией $\sigma^2/3$. Для оценки согласованности достаточно сравнить значение величины $V/\sqrt{\sigma^2/3}$ с выбранной процентной точкой стандартного нормального распределения. Нетрудно убедиться, что это значение, равное 10,7, достаточно велико, чтобы считать ответы респондентов согласованными при любом разумном уровне значимости.

4. Обсуждение метода

Статистика парной согласованности æ определяется в общей форме соотношениями (1) - (3). Ее конкретный вид зависит не только от задаваемых весов, но и от выбора определенной модели случайных ответов. В рамках этой модели рассчитывается величина P_e и производится проверка существенности наблюдаемого значения статистики. Можно указать по меньшей мере три механизма случайной классификации. Обозначим t_i^a и t_i^b количество объектов, которые были отнесены к категории i респондентами A и B соответственно; $\sum_{i=1}^{k} t_i^a = \sum_{i=1}^{k} t_i^b = n$.

F-модель. Предположим, число объектов, попадающих в каждый класс, фиксированно. Респонденту A выделено в некотором классе i t_i^a "мест", а респонденту B - t_i^b. Отвечающие наугад распределяют n объектов по n "местам", так что вероятность занять определенное "место" в индивидуальной классификации для любого объекта равна 1/n.

P-модель. Принимая решение, респондент A помещает очередной объект в класс i с вероятностью π_i^a, а B - с вероятностью π_i^b. Отвечающие не различают объекты, но в силу субъективных причин могут одни категории указывать чаще, чем другие.

C-модель (частный случай *P*-модели). Респонденты не различают ни объекты, ни категории: $\pi_i^a = \pi_i^b = 1/k$, i=1,..., k.

Именно третья, наиболее простая гипотеза о случайности использовалась нами при введении статистики $æ^c$ и разработке изложенной процедуры оценки согласованности. В известных нам зарубежных работах (см., например, [11]) исследовалась задача измерения согласия между экспертами по всей генеральной совокупности объектов, а выделенные объекты рассматривались как выборка. Каппа-статистика была предложена и затем исследовалась большинством авторов в варианте

$$æ^c = \frac{\sum_{j,j} w_{ij} P_{ij} - \sum_{j,j} w_{ij} P_{ij}^{\ a} P_{ij}^{\ b}}{1 - \sum_{j,j} w_{ij} P_{ij}^{\ a} P_{ij}^{\ b}}. \qquad (23)$$

где P_{ij}, P_{ij}^a и P_{ij}^b - доли (пропорции) объектов, попавших в определенные классы: $P_{ij} = t_{ij}/n$, $P_{ij}^a = t_i^a/n$ и $P_{ij}^b = t_i^b/n$.

В (23) величина P_e ($P_e = \Sigma w_{ij} P_{ij}^a P_{ij}^b$) имеет смысл оценки ожидаемого уровня согласия при независимости (но, вообще говоря, неслучайности) субъективных классификаций. Если вернуться к измерению согласованности для заданного набора объектов, не являющегося выборкой, æ^p будет соответствовать либо *F*-, либо *P*-модели, причем во втором случае теоретические вероятности π_i^a и π_i^b в выражении для P_e заменяются эмпирическими пропорциями.

F-модель с психологической точки зрении выглядит довольно искусственной, Так как предполагает фиксированными частоты выбора различных категорий. В явном виде она использовалась в [13], автору которой удалось получить формулы для расчета дисперсии æ^p. Исключительная трудоемкость вычислений по этим формулам свидетельствует о том, что и в практическом отношении использование *F*-модели вряд ли целесообразно.

Рассмотрим *P*-модель. Теоретические вероятности π_i^a и π_i^b могут быть установлены только в специальных экспериментах и в реальных опросах не известны. Считать эмпирические пропорции P_{ij}^a и P_{ij}^b оценками этих вероятностей, что обычно делается в рамках выборочной модели, в данном случае недопустимо, так как нет никакой гарантии, что респонденты действительно отвечают наугад и поэтому с ростом числа объектов эмпирические доли будут близки к теоретическим. Чтобы показать, к чему приводит такой подход, рассмотрим два примера.

Пример 1. С помощью показателя æ^p оценивается согласованность двух пар респондентов. Респонденты распределяли восемь объектов по трем неупорядоченным категориям: n=8, k=3. Таблицы сопряженности T_1 и T_2, построенные для каждой пары, представлены на рис. 1. Пустые клетки в них соответствуют нулям.

T_1:

4	4	

T_2:

4		
	4	

Рис. 1.

Так как речь идет о разбиениях, существенно, что в обоих случаях половина объектов была отнесена респондентами к одной и той же (первой) категории, а половина - к разным. Естественно ожидать, что значения $æ^p$ для приведенных таблиц совпадут. Однако, подставляя в (23) полученные пропорции и нужные веса ($w_{ii} = = 1$, $w_{ii} = 0$ при $i \neq j$), находим, что $æ^p(T_1) = 0$, но $æ^p(T_2) = 1/3$.

Пример 2. Восемь объектов сортируются по четырем упорядоченным категориям: n=8, k=4. Как и в предыдущем примере, должна быть оценена согласованность двух пар респондентов. Таблицы T_1 и T_2, составленные для этих случаев, показаны на рис. 2.

T_1:

2	2		
2	2		

T_2:

2			2
2			2

Рис. 2.

Как обычно, категории занумерованы в соответствии со своим упорядочением. Учитывая это, видим, что согласованность первой пары достаточно высока и заведомо выше, чем у второй. Действительно, для половины объектов в T_1 респонденты указали соседние классы, а в T_2 - противоположные. Чтобы рассчитать $æ^p$, используем предназначенную для сортировок систему весов [11]. Оказывается, $æ^p(T_1) = æ^p(T_2) = 0$.

Столь обескураживающие результаты технически связаны с неадекватными значениями величины P_e (см. [3] и [14]), которые получаются в приведенных примерах. Более глубокая причина методологическая, о ней мы уже писали. Если нет никаких сведений о

субъективной склонности респондентов к преимущественному выбору определенных категорий, для конкретизации каппа-статистики представляется естественным использовать С-модель. Эта модель была описана также в [10], но соответствующие процедуры согласованности там не развивались. Заметим, что в *примере 1*

$\overline{æ}^c(T_1) = \overline{æ}^c(T_1) = 0{,}25$, а в *примере 2* $\hat{æ}^c(T_1) = 0{,}6$ и $\hat{æ}^c(T_1) = -0{,}2$. Такие значения не противоречат здравому смыслу. Для анализа близости m разбиений (m> 2) в [11] рекомендуется коэффициент $\hat{æ}$, формула которого содержит нормирующую величину ожидаемого согласия. При этом неизвестные теоретические вероятности выбора различных категорий при расчете этой величины предлагается заменять наблюдаемыми пропорциями. Как и в случае с $æ^p$, в нашей задаче данный подход может привести к ложным выводам.

Изложенные соображения позволяют сделать вывод о предпочтительности использования статистик $æ^c$ и V для оценки согласованности субъективных классификаций данного набора объектов.

В заключение отметим, что в [8] каппа-статистика была введена на основе анализа мер близости между матрицами "объект-класс". Этот подход позволяет получать более четкую содержательную интерпретацию рассмотренных показателей. Изложим кратко его суть.

Сопоставим классификации n объектов по k классам матрицу $X = [X_{sj}]$ размера $n \times k$, где $X_{sj} = 1$ тогда и только тогда, когда объект s отнесен к категории j. Совокупность всех таких матриц обозначим R.

Определение. *Мерой близости на* R *называется функция* d (d: $R \times R \to R^1$), *удовлетворяющая условиям:*

1) $d(X, Y) \geq 0$, *причем* $d(X, X) = 0$;
2) $d(X, Y) = d(Y, X)$;
3) X, Y *принадлежат* R.

В [8] был предложен следующий показатель согласованности двух классификаций:

$$\varphi = -(d(X, Y) - Ed) / Ed, \qquad (24)$$

где Ed - математическое ожидание меры при справедливости гипотезы *C*.

В случае когда имеется единственный объект, d измеряет близость между двумя матричными строками. Такую функцию будем обозначать d'. Естественно предполагать, что она зависит только от расположения единиц в этих строках, то есть от номеров категорий i и j, к которым был отнесен объект: $d' = d'_{ij}$, i = i (X), j = j (Y). Связь между статистиками æ^c и j устанавливает следующее утверждение.

Утверждение 8. *Если*

$$d(X, Y) = \sum_{s=1} d'(X_s, Y_s), \text{ то } æ^c(W) = \varphi,$$

где $W = \{ W_{ij} \mid W_{ij} = 1 - d'_{ij} / d'_{max}, i, j=1,..., k \}$, d'_{max} - *максимальное значение* d'.

В [14,с. 5-15] для разбиений и сортировок обосновывается выбор в качестве мер близости соответственно меры

$$d_1 = \sum_s \sum_j | X_{sj} - Y_{sj} | \text{ и } d_2 = \sum_s | j(X, s) - j(Y, s) | ,$$

где j (Z, s) - номер столбца, содержащего единицу в s-й строке матрицы Z. Нетрудно убедится, что в обоих случаях получаются формулы для весов, совпадающие с приведенными выше. Таким образом оба описанных подхода (традиционный и "измерительный") оказываются эквивалентными.

Литература

1. А партактивы предпочитают Жириновского // Огонек. 1991. N 35.
2. *Кендалл М., Стьюарт А.* Статистические выводы и связи. М.: Наука, 1973.
3. *Аптон Г.* Анализ таблиц сопряженности. М.: Финансы и статистика, 1982.
4. *Колесов В. М.* Коэффициенты связи для совокупностей номинальных признаков // Социология: 4М. 1991. N 1.
5. *Кутенков Р. П.* , *Коростелев В. Г.* Свойства коэффициентов связи для номинальных признаков // Заводская лаборатория. 1992. N 5.
6. *Cohen J.* A coefficient of agreement for nominal scales // Educ. Psychol. Measurement. 1960. V. 20. P. 37-46.
7. *Cohen J.* Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit // Psychol. Bull. 1968. V. 70, P. 213-220.
8. *Раушенбах Г. В., Заславский А. А.* Проверка однородности в задачах классификации: статистический подход и его табличное обеспечение // Материалы I Всесоюзн. школы-семинара "Программно-алгоритмческое обеспечение анализа данных в медико-биологических исследованиях". Пущино: НЦБИ АН СССР, 1986.
9. *Goodman L. A., Kruskal W. N.* Measures of association for cross-classifications// J. of Amer. Stat. Association. 1954. V. 49. P. 732-764.
10. *Cicchetti D. V., Allison T.* A new procedure for assesingreliabilityofscoringEEGsleeprecordings//Amer. J. EEGTechnology. 1971.V. 11.P. 101-109.
11. *Флейс Дж.* Статистические методы для изучения таблиц и пропорций. М.: Финансы и статистика, 1989.
12. *Большев Л. Н., Смирнов Н. В.* Таблицы математической статистики. М.: Наука, 1983.
13. *Everitt B. S.* Moments of the statistic kappa and wheightedkappa//TheBrit. J. ofandStat. Psychology. 1968. V. 21.Part. 1. P. 97-103.
14. *Тюрин Ю. Н.* Экспертная классификация // Экспертные оценки в системных исследованиях. М.: ВНИИСИ, 1979.

Приложение

Таблица П1

Нижние процентные точки x_a статистики P_0'

для $n \le 10$, $k \le 10$: $\Pr \{ P_0' \le x_a \} \approx a$

$\alpha = 0{,}01$

n \ k	2	3	4	5	6	7	8	9	10
2	-	-	-	-	-	-	-	-	0
3	-	-	-	0	0	0	0	1	1
4	-	-	0	0	1	1	2	2	3
5	-	0	1	1	2	3	3	4	4
6	-	0	1	2	3	4	5	6	7
7	0	1	2	3	4	6	7	8	9
8	0	1	3	4	6	7	8	10	11
9	0	2	4	6	7	9	10	12	13
10	0	3	5	7	9	10	12	14	16

$\alpha = 0{,}05$

n \ k	2	3	4	5	6	7	8	9	10
2	-	-	-	0	0	0	0	0	1
3	-	0	0	1	1	1	2	2	3
4	-	0	1	2	2	3	4	4	5
5	0	1	2	3	4	5	5	6	7
6	0	1	3	4	5	6	7	9	10
7	0	2	4	5	7	8	9	11	12
8	1	3	5	6	8	10	11	13	15
9	1	3	6	8	10	12	14	15	17
10	1	4	7	9	11	13	16	18	20

Примечание. Прочерк означает отсутствие соответствующей процентной точки. При $n = 1$ и заданных α точки отсутствуют для любого $k \le 10$.

Таблица П2

Критические верхние и нижние значения $V_н$,$V_в$
для сортировок: $Pr\{V < V_н\} = Pr\{V > V_в\} \approx 0{,}05$; n=1.

n\k	2	3	4	5	6
3	- -. 334	- . 250	-. 601 . 466	-. 667 . 583	-. 715 . 657
4	- . 000	-. 500 . 367	-. 600 . 437	-. 563 . 584	-. 543 . 485
5	- . 200	-. 350 . 550	-. 440 . 520	-. 500 . 500	-. 440 . 486
6	- . 333	-. 350 . 400	-. 387 . 467	-. 375 . 417	-. 371 . 417
7	- . 048	-. 286 . 357	-. 372 . 314	-. 408 . 643	-. 372 . 608
8	- . 143	-. 245 . 358	-. 285 . 343	-. 384 . 353	-. 304 . 338
9	- . 223	-. 250 . 312	-. 289 . 289	-. 284 . 306	-. 285 . 514
10	- . 289	-. 224 . 250	-. 244 . 289	-. 250 . 292	-. 257 . 291
11	- . 119	-. 228 . 264	-. 232 . 272	-. 250 . 272	-. 253 . 270
12	- . 191	-. 193 . 216	-. 212 . 248	-. 231 . 261	-. 232 . 259
13	- . 231	-. 183 . 221	-. 204 . 240	-. 218 . 247	-. 213 . 248
14	- . 121	-. 162 . 233	-. 195 . 244	-. 202 . 236	-. 209 . 236

Таблица П2
(окончание)

n\k	2	3	4	5	6
15	- -. 162	-. 178 . 229	-. 189 . 223	-. 190 . 215	-. 195 . 226
16	- . 084	-. 162 . 184	-. 173 . 212	-. 182 . 214	-. 193 . 216
17	- . 118	-. 158 . 206	-. 176 . 200	-. 176 . 210	-. 188 . 206
18	- . 151	-. 172 . 191	-. 172 . 195	-. 172 . 195	-. 176 . 200
19	- . 088	-. 144 . 184	-. 162 . 186	-. 170 . 188	-. 173 . 188
20	- . 116	-. 131 . 176	-. 154 . 183	-. 164 . 187	-. 164 . 185
21	- . 143	-. 136 . 154	-. 151 . 177	-. 155 . 179	-. 161 . 182
22	- . 091	-. 130 . 172	-. 143 . 173	-. 152 . 174	-. 156 . 177
23	- . 115	-. 121 . 155	-. 144 . 166	-. 151 . 170	-. 154 . 171
24	- . 072	-. 121 . 160	-. 136 . 168	-. 146 . 164	-. 150 . 167
25	- . 094	-. 117 . 152	-. 136 . 157	-. 142 . 163	-. 145 . 164