Широбоков И.Г.

[1)] *МАЭ РАН, Отдел антропологии, Университетская наб., 3, Санкт-Петербург, 199034, Россия*

О ПОПРАВКЕ НА ЧИСЛО НАБЛЮДЕНИЙ ПРИ РАСЧЕТЕ РАССТОЯНИЙ МАХАЛАНОБИСА

Введение. *Расстояния Махаланобиса (D^2) применяются в краниологических исследованиях для обобщенной оценки различий между выборками с учётом дисперсии признаков и корреляций между ними. При этом выборочные расстояния, особенно в случае небольшого размера выборок, в среднем демонстрируют смещение в сторону завышенных значений по сравнению с истинной величиной D^2. Внесение поправки на число наблюдений, предложенная Д. Райтмайром, является одним из простых способов компенсации этого смещения. Однако условия ее применения и даже способы расчета переменных могут различаться в зависимости от особенностей выборки. Цель данного исследования состоит в оценке эффективного влияния поправки на величину расстояний Махаланобиса и поиске подходов к снижению смещения выборочных оценок.*

Материалы и методы. *В анализе использованы три обобщенные серий мужских черепов башкир, чувашей и латышей. Для расчета D^2 применялась усредненная ковариационная матрица. Были рассмотрены три метода вычисления расстояний Махаланобиса: без поправки на число наблюдений, с поправкой Д. Райтмайра, примененной ко всем расстояниям, и с поправкой, учитывающей только статистически значимые расстояния. Кроме того, протестированы альтернативные подходы к внесению поправки при резких различиях в числе наблюдений отдельных признаков: использование среднего гармонического числа наблюдений и раздельное вычисление расстояний D^2 для линейных и угловых признаков с последующим суммированием.*

Результаты. *Поправка Д. Райтмайра, применяемая ко всем расстояниям, в среднем позволяет получать достаточно точные несмещенные оценки D^2. При внесении поправок величина D^2 может оказаться близкой к нулю или даже отрицательной, в т.ч. при наличии значимых различий между популяциями. Поскольку медианы выборочных D^2 в наибольшей степени сближаются с истинными значениями D^2, все отрицательные значения D^2 могут без ущерба для расчетов преобразованы в нули. Выборочное $D^2=0$ необязательно означает, что между выборками отсутствуют морфологические различия. Для приближения к истинному значению расстояния между выборками можно воспользоваться вычислением доверительных интервалов, например, при помощи процедуры бутстрэппинга.*

Заключение. *Поправка Д. Райтмайра позволяет получать несмещенные оценки расстояний Махаланобиса при использовании усредненной ковариационной матрицы и небольших размерах выборок. Однако выбор конкретного метода коррекции должен учитывать размер выборок и вариативность числа наблюдений по разным признакам. При работе с сериями черепов плохой сохранности в поправке целесообразно использовать среднее гармоническое число наблюдений или раздельный расчет расстояний для угловых и линейных признаков.*

Ключевые слова: расстояния Махаланобиса; размер выборки; краниометрия

DOI: 10.55959/MSU2074-8132-25-2-9

Введение

Расстояния Махаланобиса (здесь и далее для краткости расстояниями или просто D^2 именуются квадраты расстояний Махаланобиса) широко применяются в многомерном статистическом анализе для обобщенной оценки различий между выборками с учётом ковариационной структуры исследуемых признаков. Изначально (почти столетие назад) метод был предложен П.Ч. Махаланобисом именно для анализа краниологический серий, однако затем получил широкое признание за пределами краниометрии, преимущественно в задачах классификации объектов. В среде отечественных краниологов метод получил распространение в первую очередь благодаря работам А.Г. Козинцева [Козинцев, 2007; 2009; и др.] и программе Б.А. Козинцева CANON. Программа предназначена для проведения канонического дискриминантного анализа по средневыборочным значениям 14 признаков (№№ по Мартину и др.: 1, 8, 17, 9, 45, 48, 55, 54, 51, 52, 77, zm, SS:SC, 75(1))[1] и основана на использовании усредненной ковариационной матрицы. Программа также рассчитывает матрицу квадратов расстояний Махаланобиса с поправкой Д. Райтмайра [Rightmire, 1969]. Формула скорректированной величины расстояния имеет вид:

$$D^2_c = D^2 - p*(1/n_1 + 1/n_2),$$

где p – число признаков, а n_1 и n_2 – число наблюдений в сравниваемой паре выборок. Число наблюдений, как правило, оказывается несколько меньше числа черепов, составляющих выборку, поскольку сохранность скелетов, особенно происходящих из археологических раскопок, редко позволяет измерить каждый череп по полной программе. В программе CANON число наблюдений рассчитывается как среднее число наблюдений всех признаков.

Описанный выше вариант алгоритма был по умолчанию принят большинством российских краниологов, использующих расстояния Махаланобиса для оценки различий между выборками. Другие статистические программы позволяют проводить расчеты только по индивидуальным данным (которые чаще всего отсутствуют в распоряжении исследователей для большинства сравнительных серий)[2]. Программа CANON была написана около тридцати лет назад, и, к сожалению, она не позволяет пользоваться данными из таблиц Excel без предварительной обработки, а главное – не запускается на современных компьютерах без специальных программ-эмуляторов. Не так давно московскими коллегами была выпущена программа MultiCan [Гончаров, Гончарова, 2016]. Программа также выполняет канонический дискриминантный анализ на основе усредненной ковариационной матрицы и вычисляет матрицу расстояний Махаланобиса. Программа удобна в использовании и основана на том же алгоритме, что и CANON (описанным в работах В.Е. Дерябина), за единственным исключением. MultiCan вычисляет матрицу именно расстояний Махаланобиса (D), а не квадратов расстояний (D^2), а поправка на число наблюдений к ним не применяется.

Между тем, поправка необходима для компенсации смещения полученной оценки D^2 от истинной величины расстояния между популяциями, выборки из которых анализируются исследователями [Sjøvold, 1975]. Такая необходимость возникает даже при условии использования усредненной ковариационной матрицы, когда учитываемые алгоритмом различия между выборками сводятся к различиям в средних значениях признаков. Отклонения средних от реальных значений признаков в генеральных совокупностях могут быть как положительными, так и отрицательными. Для каждого отдельно взятого признака эти отклонения компенсируют друг друга, и математическое ожидание среднего значения выборки равно истинному среднему в соответствующей популяции. Однако, чем больше признаков используется при расчете обобщенной меры различий между выборками и чем меньше при этом размер (число индивидов) последних, тем больше итоговая величина D^2 смещается в сторону завышения истинного расстояния между популяциями.

Существуют различные способы получения несмещенных оценок расстояний Махаланобиса. Несомненное преимущество поправки, предложенной Д. Райтмайром более пятидесяти лет назад, заключается в простоте ее вычисления. Ее влияние на итоговую оценку D^2 при разном числе наблюдений может заметно различаться. При работе с большими выборками величина поправки оказывается ничтожной и не имеет значения, а при небольшом числе наблюдений может, напротив, оказаться больше самого расстояния. В этом случае применение поправки Д. Райтмайра нередко приводит к тому, что расстояния Махаланобиса оказываются отрицательными, несмотря на то, что

[1] Состав и число признаков можно менять при наличии у исследователя необходимой информации о корреляциях между ними и величине стандартных отклонений.

[2] Справедливости ради, стоит отметить, что если в распоряжении исследователя есть подходящая ковариационная матрица, то для вычисления D^2 по средним значениям признаков достаточно воспользоваться встроенными функциями Excel.

квадраты расстояний не могут быть отрицательными по определению. Однако А.Г. Козинцев отмечает, что получение отрицательных значений «вопреки мнению некоторых антропологов, не только возможно, но и необходимо, т.к. речь идет не о генеральных совокупностях, а о выборках, притом очень небольших. Лишь при учете отрицательных значений средняя величина D^2 может получиться нулевой при отсутствии реальных различий между двумя группами» [Козинцев, 2007, с. 145].

С другой стороны, Д. Райтмайр и Т. Шевольд указывали, что поправка должна применяться только в случае, если величина D^2 является статистически значимой. Перед внесением поправки необходимо проверить значимость расстояний с использованием статистики хи-квадрат согласно Д. Райтмайру [Rightmire, 1969] и F-статистики согласно Шевольду [Sjøvold, 1975]. Проверка проводится путем домножения расстояний Махаланобиса на соответствующие статистикам коэффициенты и сопоставлением полученных величин с критическими значениями. Между тем, ни CANON, ни MultiCan не предполагают оценки значимости расстояний, но в первом случае поправка автоматически вводится ко всем рассчитанным D^2, а во втором не предусмотрена вовсе.

Вторая проблема заключается в том, что из-за недостаточно хорошей сохранности черепов нередко число измерений значительно (иногда в несколько раз) различается между отдельными признаками. В первую очередь это касается угловых признаков лицевого скелета, часть, которых к тому же отличается относительно высокой внутригрупповой вариабельностью. Недооценка случайных колебаний угловых признаков и указателей из-за использования усредненного числа наблюдений при расчете поправки теоретически может приводить к завышению реальных различий между популяциями. Конечно, можно попытаться оценить «вес» каждого отдельного признака в величину D^2 между каждой парой сравниваемых выборок, однако такой подход потребует большого объема вычислений. Более простой выход заключается в использовании среднего гармонического числа наблюдений – оно всегда будет меньше среднего и увеличит поправку. Более сложный подход заключается в отдельном расчете расстояний Махаланобиса для наборов линейных и угловых признаков. На внутригрупповом уровне корреляция между ними невелика или полностью отсутствует, а это означает, что D^2, полученные для разных наборов и скорректированные при помощи отдельно вносимых поправок, могут быть затем суммированы для получения общего расстояния. Рассчитанное таким образом расстояние будет преимущественно корректировано именно за счет учета случайных вариаций угловых признаков. Но есть ли смысл в таком дифференцированном подходе?

Данное исследование посвящено указанным двум проблемам. Первая задача заключается в демонстрации смещения выборочных оценок D^2 от истинного значения между популяциями и роли поправки Д. Райтмайра (с предварительной оценкой статистической значимости D^2 и без нее) в ее снижении. Вторая задача состоит в оценке обоснованности использования специальных подходов к расчету расстояния Махаланобиса при значительном разбросе в числе наблюдений между линейными и угловыми признаками.

Материалы и методы

Для анализа были использованы три обобщенные серии мужских черепов: башкир, чувашей и латышей [Алексеев, 1969; 1971; Юсупов, 1989]. Каждая серия включает в себя черепа из нескольких могильников. Средние значения краниометрических признаков в сериях представлены в таблице 1. Для генерации случайных подвыборок, расчета средних значений признаков и оценки распределения значений D^2 при разном числе наблюдений признаков использовались индивидуальные данные, опубликованные в сводке [Широбоков с соавт., 2017]. При вычислении расстояний Махаланобиса использовалась усредненная ковариационная матрица, рассчитанная А.Г. Козинцевым для 14 краниометрических признаков и по умолчанию используемая в программе CANON.

На первом этапе расчет квадратов расстояний Махаланобиса между подвыборками осуществлялся в трех вариантах: без поправки на число наблюдений; с поправкой Д. Райтмайра на число наблюдений, которое определялось как среднее арифметическое числа наблюдений отдельных признаков; с поправкой Д. Райтмайра для статистически значимых расстояний. Для оценки статистической значимости при $\alpha=0.05$ применялась F-статистика. Предполагалось, что величина $D^2*(n_1*n_2*(n-k-p+1)/((n-k)*p*(n_1+n_2))$ подчиняется F-распределению со степенями свободы $df_1=p$ и $df_2=n-k-p-1$, где n – общее число индивидов во всех k выборках, учитываемых при подготовке ковариационной матрицы по p признакам [Sjøvold, 1975]. Поскольку соотношение

Таблица 1. Средние значения краниометрических признаков в сериях
Table 1. Average values of craniometric features in samples

№ признаков по Мартину и др.	БАШКИРЫ		ЧУВАШИ		ЛАТЫШИ	
	n	M	n	M	n	M
1	324	182,4	126	180,2	128	183,6
8	325	147,4	126	143,1	128	144,2
17	317	133,2	111	132,8	111	133,2
9	325	96,7	128	97,0	129	98,2
45	315	138,6	111	133,5	116	133,6
48	292	74,1	118	72,3	113	70,2
55	313	54,7	121	52,1	115	51,6
54	310	26,1	115	25,3	108	25,1
51	312	44,5	119	41,6	116	42,4
52	312	34,7	121	33,1	119	32,7
77	326	142,5	131	140,5	132	140,0
zm	326	131,8	131	130,0	132	126,6
SS:SC	326	48,5	131	46,7	132	48,1
75 (1)	319	25,3	98	25,0	94	31,0

(n-k-p+1)/(n-k) при большом n стремится к 1, а в анализе использовалась усредненная ковариационная матрица, исходное выражение может быть упрощено до вида $D^{2*}(n_1*n_2)/(p*(n_1+n_2))$.

На втором этапе проводилось тестирование двух подходов к снижению смещения в оценке расстояний, вызываемого значительным разбросом в числе наблюдений между отдельными признаками. В первом варианте число наблюдений, учитываемое в поправке Д. Райтмайра, рассчитывалось как среднее гармоническое. Во втором варианте расстояния между подвыборками отдельно рассчитывались для двух наборов признаков: 10 линейных признаков (D^2_l) и 4 угловых и симотического указателя (D^2_a). Поправка Д. Райтмайра вносилась в величину каждого отдельно рассчитанного расстояния, а число наблюдений рассчитывалось как среднее числа наблюдений внутри соответствующих наборов признаков. Затем рассчитывалось обобщенное $D^2_{sum}=D^2_l+D^2_a$. Распределения вычисленных таким образом расстояний сравнивались с распределениями расстояний, рассчитанных традиционным способом. Для наглядности на втором этапе анализа число наблюдений в наборе угловых признаков и симотического указателя было искусственно снижено путем случайного исключения половины индивидуальных значений соответствующих признаков.

Для проведения всех расчетов и визуализации результатов была написана программа на Python с использованием библиотек pandas, numpy и matplotlib. Первоначально код программы был написан при помощи ChatGPT (бесплатная версия GPT-4o), а затем по необходимости редактировался автором. Чат-бот хорошо справился с написанием кода, но, как показали результаты предварительного тестирования программы, для правильного расчета числа наблюдений, а также оценки статистической значимости расстояний все же потребовалось введение некоторых корректировок.

Результаты

В таблице 2 представлены средние значения и медианы D^2 при разных способах подсчета, а также интервалы, в которых находятся 95% расстояний между подвыборками с разным число наблюдений. На рисунке 1 в качестве примера представлены варианты распределения D^2 между 10 000 парами случайных подвыборок из общей башкирской серии при разных способах подсчетов. Вполне предсказуемым образом среднее и медиана D^2 в наибольшей степени сближаются с ожидаемым $D^2\approx0$ в случае автоматического внесения поправки Д. Райтмайра во все расстояния. Вариант без поправки всегда дает завышенные значения. Вариант с предварительной оценкой статистической значимости D^2 сближается с вариантом без внесения поправки даже при 50 наблюдениях в каждой выборке, но позволяет избавиться от длинного хвоста «больших» расстояний.[3] Этот результат также предсказуем: мы заранее знаем, что нулевая гипотеза справедлива и большая часть расстояний не может иметь статистически значимую величину, но при этом не может быть и отрицательной, а значит смещение неизбежно.

На графике хорошо заметно, что поскольку распределение расстояний Махаланобиса между выборками, особенно небольшими, аппроксимируется именно F-распределением, не среднее значение D^2, а мода или медиана множества подвыборок будет лучше соответствовать истинному значению D^2 между популяциями. F-распределение асимметрично и имеет вытянутый правый

[3] Очевидно, что тот же результат мы получим, если для тестирования значимости расстояний вслед за самим Д. Райтмайром воспользуемся статистикой хи-квадрат. В этом случае для сравнения с критическим значением величина D^2 предварительно корректируется путем умножения на величину $(n_1*n_2)/(n_1+n_2)$.

Таблица 2. Средние и медианы расстояний Махаланобиса (D^2), а также 95% доверительный интервал, рассчитанный для 10 000 пар случайных подвыборок
Table 2. Mean and median Mahalanobis distances (D^2) and 95% confidence interval calculated for 10 000 pairs of random subsamples

Сравниваемые подвыборки (по 10 черепов в каждой)	**Среднее**	**Медиана**	**95%-ый интервал**
Башкиры-чуваши	6,1 (8,7)	5,4 (8,3)	1,3 – 11,9
Башкиры-латыши	10,3 (11,8)	9,5 (11,4)	2,9 – 19,5
Чуваши-латыши	4,6 (8,1)	3,8 (7,7)	-0,3 – 12,8
Башкиры-башкиры	1,0 (3,8)	0,8 (3,5)	-1,4 – 4,6
Сравниваемые подвыборки (по 20 черепов в каждой)	**Среднее**	**Медиана**	**95%-ый интервал**
Башкиры-чуваши	5,6 (6,7)	5,3 (6,5)	2,2 – 9,6
Башкиры-латыши	9,8 (11,0)	9,1 (10,7)	4,3 – 15,3
Чуваши-латыши	4,8 (6,0)	4,2 (5,8)	1,1 – 9,1
Башкиры-башкиры	0,4 (1,9)	0,3 (1,7)	-0,7 – 2,1
Сравниваемые подвыборки (по 30 черепов в каждой)	**Среднее**	**Медиана**	**95%-ый интервал**
Башкиры-чуваши	5,2 (6,3)	5,3 (6,2)	2,7 – 8,7
Башкиры-латыши	8,8 (10,1)	9,1 (10,1)	5,3 – 14,1
Чуваши-латыши	4,6 (5,8)	4,1 (5,6)	1,4 – 8,0
Башкиры-башкиры	0,4 (1,4)	0,2 (1,3)	-0,5 – 1,7
Сравниваемые подвыборки (по 50 черепов в каждой)	**Среднее**	**Медиана**	**95%-ый интервал**
Башкиры-чуваши	5,4 (6,1)	5,4 (6,0)	3,5 – 7,8
Башкиры-латыши	9,0 (9,7)	8,9 (9,5)	5,5 – 13,1
Чуваши-латыши	4,3 (4,9)	4,2 (4,8)	2,1 – 7,2
Башкиры-башкиры	0,2 (0,8)	0,2 (0,7)	-0,3 – 1,1
Сравниваемые подвыборки (по 10 и 50 черепов)	**Среднее**	**Медиана**	**95%-ый интервал**
Башкиры-чуваши	5,8 (9,2)	5,5 (8,8)	1,0 – 12,4
Башкиры-латыши	9,3 (12,8)	8,8 (12,2)	2,3 – 19,2
Чуваши-латыши	5,2 (7,7)	4,7 (7,3)	0,1 – 12,8
Башкиры-башкиры	1,2 (3,8)	0,9 (3,5)	-1,3 – 5,2

Примечания. Расстояния Махаланобиса указаны с поправкой Райтмайра [1969], примененной ко всем D^2, в скобках приведены расстояния без поправки.

Notes. Mahalanobis distances are given with the Rightmire correction [1969] applied to all D^2, distances in brackets are given without correction.

хвост. Среднее значение будет давать несколько завышенные оценки расстояния, даже если применять поправку Д. Райтмайра ко всем расстояниям без учета их статистической значимости.

Таким образом, если мы исходим из того, что исследуемые выборки происходят из одной генеральной совокупности, поправка Д. Райтмайра, автоматически применяемая ко всем расстояниям, в среднем демонстрирует наилучший результат. Означает ли что такой подход оптимален во всех случаях?

На следующем графике (рис. 2) представлены варианты распределения D^2 для 10 000 пар случайных подвыборок из серий башкир и чувашей при разных способах подсчетов. Истинное значение D^2 между сериями составляет 5.1, т.е. популяции довольно заметно различаются между собой. Некорректированные расстояния между подвыборками, включающими по 5 черепов каждая, как и следовало ожидать, в среднем заметно превышают расстояние между полными сериями. Все расстояния между подвыборками, включающими по 20 черепов и больше, статистически значимы, поскольку в этом случае величина истинного значения между чувашами и башкирами в два с лишним раза превышает критическую величину. Поэтому

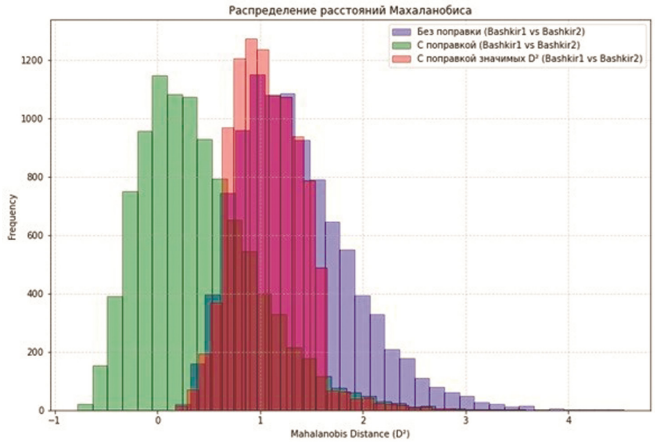

Рисунок 1. Распределение расстояний Махаланобиса между 10 000 парами случайных подвыборок башкир при разных способах подсчета (по 30 индивидов в каждой подвыборке)
Figure 1. Distribution of Mahalanobis distances between 10,000 pairs of random subsamples of Bashkirs using different counting methods (30 individuals in each subsample)

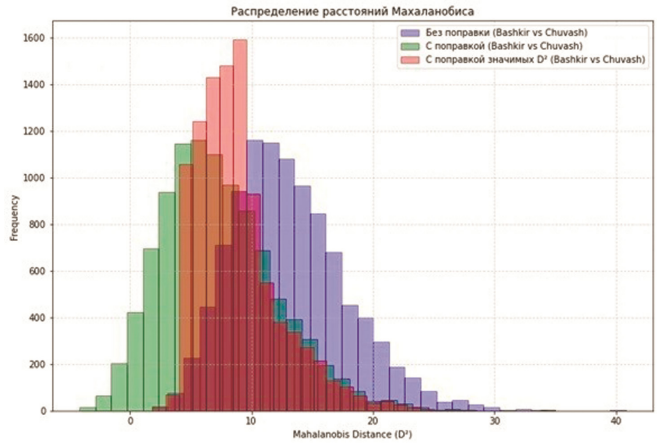

Рисунок 2. Распределение расстояний Махаланобиса между 10 000 парами случайных подвыборок башкир и чувашей при разных способах подсчета
Figure 2. Distribution of Mahalanobis distances between 10,000 pairs of random subsamples of Bashkirs and Chuvash using different counting methods (5 individuals in each subsample)

распределения соответствующим образом рассчитанных расстояний полностью совпадают.

В случае, когда обе или одна сравниваемых выборок невелики (например, каждая включает всего по 5 черепов), расстояния, рассчитанные с поправкой Д. Райтмайра, в среднем одинаково близки к истинному значению в независимости от предварительной оценки статистической значимости. Однако предварительная оценка статистической значимости D^2 позволяет избежать излишне оптимистичных заключений о морфологическом сходстве популяций. При автоматическом внесении поправки расстояние между подвыборками в ряде случаев оказывается близким к нулю или даже отрицательным.

На втором этапе сравнивались расстояния Махаланобиса между подвыборками башкир, чувашей и латышей, рассчитанные на основе одной и двух ковариационных матриц (табл. 3). Расстояния, вычисленные по полному набору признаков, в целом несколько превышают расстояния, рассчитанные как сумма D^2 для линейных и угловых признаков. В последнем случае перед получением суммарных D^2_{sum} поправка Райтмайра вносилась независимо в величину каждого расстояния[4]. В обоих вариантах средние величины примерно в равной степени близки к истинным значениям D^2, чем при стандартном способе расчета (при условии, что в обоих случаях использована поправка Д. Райтмайра).

[4] Как уже указывалось выше, вероятно, наибольшей точности можно достигнуть при использовании медиан.

Таблица 3. Реальные и выборочные расстояния Махаланобиса, рассчитанные с поправкой Райтмайра [1969] для общего набора признаков и двух независимых наборов (10 000 пар случайных подвыборок). В скобках приведены расстояния с поправкой, где в качестве числа наблюдений использовано среднее гармоническое числа наблюдений отдельных признаков

Table 3. Real and sample Mahalanobis distances between the studied samples, calculated with the Rightmire correction [1969] for the common set of features and two independent sets (10,000 pairs of random subsamples). The corrected Mahalanobis distances are given in brackets, where the harmonic mean of the number of observations of individual features is used as the number of observations

Сравниваемые группы (D^2 между полными сериями)	Среднее расстояние Махаланобиса для 10000 пар случайных выборок					
	D^2			$D^2_{sum}=D^2_l+D^2_a$		
	по 10 черепов	по 20 черепов	по 30 черепов	по 10 черепов	по 20 черепов	по 30 черепов
Башкиры-чуваши (5,1)	6,1	5,6	5,2	5,4	5,4	5,3
Башкиры-латыши (8,8)	10,3	9,8	8,8	9,1	8,6	8,4
Чуваши-латыши (4,2)	4,6	4,8	4,6	4,4	4,0	3,8
Башкиры-башкиры (0)	1,1	0,5	0,4	0,7	0,4	0,3
	Число наблюдений угловых признаков снижено в 2 раза					
	D^2			$D^2_{sum}=D^2_l+D^2_a$		
Башкиры-чуваши (5,1)	6,8 (5,5)	6,5 (5,0)	5,7 (4,7)	6,6	5,6	5,3
Башкиры-латыши (8,8)	10,2 (7,6)	9,9 (7,3)	9,5 (6,9)	9,9	8,8	8,7
Чуваши-латыши (4,2)	6,4 (4,8)	5,5 (4,1)	5,1 (3,9)	5,4	4,4	4,3
Башкиры-башкиры (0)	2,6 (1,4)	1,3 (0,7)	1,0 (0,5)	1,0	0,7	0,5

Несмотря на использование усредненной ковариационной матрицы и автоматически применяемой к каждому расстоянию поправки Райтмайра, средние суммарные квадраты расстояний Махаланобиса, рассчитанные для двух наборов признаков, оказались приблизительно или незначительно меньше расстояний, рассчитанных для общего набора $D^2_{sum}=D^2_l+D^2_a \leq D^2$

При случайном исключении части значений угловых признаков и симотического указателя (что соответствует гипотетической ситуации плохой сохранности черепов), расстояния D^2_{sum} в среднем оказались ближе к истинным значениям D^2. По величине угловых признаков серии башкир и чувашей демонстрируют большее сходство, нежели каждая из них имеет с серией латышей. Однако в рассматриваемом случае это не повлияло значимым образом на разницу между соответствующими расстояниями Махаланобиса, вычисленными двумя способами. Использование в поправке Д. Райтмайра вместо среднего арифметического среднего гармонического числа наблюдений позволило приблизить средние D^2 к истинным значениям, хотя и с чуть меньшей эффективностью по сравнению с D^2_{sum}.

Обсуждение

Среднее гармоническое оказывается наиболее простым и достаточно точным способом учесть число наблюдений в поправке при расчете D^2, когда между числом измерений отдельных признаков наблюдается заметный разброс. Если же сравниваемые выборки включают всего по несколько черепов, то, вероятно, для получения правдоподобной оценки D^2 стоит сократить число рассматриваемых признаков до группы показателей с наиболее высокой ожидаемой (исходя из исторического контекста) дифференцирующей способностью и наибольшим число наблюдений. Естественно такие оценки D^2 также требуют внесения поправок и могут сравниваться только с расстояниями, рассчитанными для того же сокращенного набора признаков.

Сама по себе необходимость корректировки выборочных расстояний Махаланобиса не вызывает сомнений. Использованный в программе CANON алгоритм автоматического применения поправки Д. Райтмайра ко всем расстояниям оказывается эффективным, несмотря на рекомендации Т. Шевольда и Д. Райтмайра вводить поправки только в значения статистически значимых расстояний.

Т. Шевольд указывает, что модификации выборочных расстояний Махаланобиса (включая предложенную Д. Райтмайром), могут быть приняты как «асимптотически последовательный результат из нецентрального F-распределения» и именно поэтому должны применяться лишь к статистически значимым расстояниям [Sjøvold, 1975, p. 555]. Нецентральное F-распределение отличается от центрального (обычного) F-распределения наличием параметра нецентральности λ. При λ=0 оба распределения идентичны. Когда мы тестируем значимость расстояния Махаланобиса, изначально мы исходим из того, что его распределение подчиняется именно центральному F-распределению, а нулевая гипотеза состоит в том, что истинное D^2 равно 0. Если мы обнаруживаем, что в случае справедливости этого условия вероятность получить некоторое выборочное D^2 или еще более экстремальные значения достаточно невелика (например, составляет менее 5%), мы отвергаем нулевую гипотезу и принимаем решение о модификации расстояния, необходимой в случае справедливости гипотезы о подчинении D^2 нецентральному F-распределению.

Однако в действительности, нулевая гипотеза об отсутствии различий между популяциями (истинное $D^2=0$) при работе с краниологическими сериями может априори рассматриваться как неверная. Как правило, в сравниваемые выборки включаются черепа, происходящие из отдельных могильников, или отнесенные к некоторой общности (например, археологической культуре или этнической группе). Мы не формируем несколько выборок по материалам одного могильника, если у нас нет оснований полагать, что они различаются по некоторым существенным параметрам между собой. Именно в этом случае нулевая гипотеза может оказаться справедливой, но для ее проверки (точнее для принятия ее в качестве легитимной) можно воспользоваться другими инструментами, например, оценкой различий между средними значениями конкретных признаков при помощи t-критерия или его непараметрических аналогов. И это будет подход более релевантный, нежели расчет расстояний Махаланобиса по усредненной ковариационной матрице.

Между выборками, относящимися к некоторой интересующей нас общности (археологической культуре), ожидаемое расстояние D^2 будет больше ноля, поскольку мы вправе полагать то, что, хотя бы часть морфологических признаков черепа обладает определенной географической или хронологической изменчивостью. Это также означает, что расстояние D^2 между выборкой, происходящей из конкретного могильника одной культуры, и сборной серией черепов из разных могильников другой археологической культуры будет скорее всего несколько завышенным по сравнению с истинным расстоянием между центроидами данных культур даже при использовании поправки. Характеристика выборки из конкретного могильника всегда смещена не только относительно характеристики населения, его оставившего, но и имеет дополнительное смещение относительно средней характеристики населения той (мнимой или реальной как популяции) общности, к которой данный могильник относится.

Именно поэтому гипотеза о подчинении расстояний Махаланобиса нецентральному F-распределению может рассматриваться как априорная. Параметр нецентральности нам неизвестен, но для принятия решения о внесении поправок в выборочные расстояния это не имеет значения: если мы принимаем положение о том, что истинные D^2 между популяциями хотя бы на ничтожную долю отличаются от нуля, коррекция выборочных D^2 становится автоматически необходимой.

Отсюда становится очевидно, что требование оценить перед внесением поправки статистическую значимость D^2 логически несправедливо. Истинная величина D^2 нам неизвестна, но чем она меньше, тем большее число наблюдений требуется, чтобы преодолеть условный порог для отвержения нулевой гипотезы, а между тем поправка нужна именно для корректировки расстояний между небольшими выборками. Иными словами, хотя оценка статистической значимости сама по себе не имеет ничего общего с оценкой величины самого эффекта (истинного значения D^2), но избирательное внесение поправки искусственно устанавливает такую связь. Требование оценки статистической значимости D^2 перед внесением поправки снижает эффективность последней – именно это мы наблюдаем на примере полученных результатов.

С другой стороны, приходится признать, что если мы автоматически применяем поправку ко всем расстояниям, то следует отказаться от оценки $D^2\leq 0$ как надежного свидетельства морфологической общности исследуемых серий черепов, поскольку такая величина может быть следствием недостаточного числа наблюдений или неоправданно большого набора сравниваемых показателей. Все отрицательные значения выборочных D^2 без всякого ущерба для точности расчетов могут быть приравнены к нулю, поскольку не средние, а медианы расстояний Махаланобиса наиболее близко отражают истинные величины D^2 между популяциями.

Отсюда следует, что, когда нас интересует оценка сходства между двумя конкретными выборками или одной выборкой и набором сравнительных серий, лучше всего не ориентироваться на выборочные значения D^2, а попытаться установить доверительные интервалы, в которых находится реальное значение расстояния Махаланобиса. Даже если для большинства сравнительных серий доступны только средние значения признаков, исследователи, как правило, владеют данными индивидуальных измерений черепов из тех выборок, которые являются основным объектом их изучения. Таких данных достаточно для расчета доверительных интервалов путем бутстрэппинга. Для примера из каждой серии были извлечены случайные подвыборки заданного размера (10, 20, 30 и 50 черепов). Затем из первой подвыборки были сформированы 100 псевдовыборок того же размера путем случайного отбора черепов с возвращением. Для каждой из псевдовыборок и подвыборки из второй серии были рассчитаны расстояния Махаланобиса с поправкой Д. Райтмайра, вычислены средние, медианы и интервалы, в которых находились 95% всех расстояний. Результаты, приведенные в таблице 4, свидетельствуют, что таким образом можно довольно точно рассчитать интервал, в котором находится истинное значение расстояния между исходными сериями. Единичные случаи, в которых истинные значения находятся за пределами интервальных значений могут объясняться чисто статистическими причинами. В двух случаях истинные расстояния оказались ниже интервальных оценок и, вероятнее всего, это связано с проблемой общностей, описанной выше. Мы вычисляем несмещенную оценку расстояния между случайными подвыборками из серий чувашей, башкир и латышей, однако сравниваем результаты с расстояниями между полными сериями. Поскольку каждая серия включает в себя черепа из разных могильников, очевидно, что на величину отклонения характеристики случайной подвыборки от средней своей общности может повлиять компонент территориальной изменчивости.

Помимо поправки Д. Райтмайра, существует еще по меньшей мере два простых варианта поправки расстояния Махаланобиса, рассмотренные в работе Т. Шевольда [Sjøvold, 1975]: поправка Г. Ван Варка, а также П. Лакенбрука и М. Микки. В варианте, предложенном Г. Ван Варком, расстояние будет всегда меньше, чем при использовании поправки Д. Райтмайра за счет коэффициента на который умножается выборочная величина D^2 перед вычитанием общей поправки: $D^2_c = D^2*(n_1+n_2-p-3)/(n_1+n_2-2)-p*(n_1+n_2)/(n_1*n_2)$. Также Г. Ван Варком был предложен второй вариант коррекции расстояния – специально для случая использования усредненной ковариационной матрицы, рассчитанной по данным k референтных групп с общим числом наблюдений n: $D^2_c=D^2*(n-k-p-1)/(n-k)-p*(n_1+n_2)/(n_1*n_2)$. Однако, как заметил Т. Шевольд, при большом размере последних этот формула становится практически идентичной предложенной Д. Райтмайром [Sjøvold, 1975, p. 550].

Формула для коррекции расстояний в соответствии с предложением П. Лакенбрука и М. Микки имеет вид: $D^2_c=D^2*(n_1+n_2-p-3)/(n_1+n_2-2)$. Ее несомненное преимущество заключается в том, что при таком способе расчета выборочное расстояние никогда не бывает отрицательным. Читатель может самостоятельно протестировать этот вариант поправки и убедиться в том, что в задаче оценки истинного значения D^2 между популяциями он, к сожалению, не имеет преимуществ перед поправкой Д. Райтмайра. Более того, в случае применения поправки только к статистически значимым расстояниям, общее распределение выборочных D^2 может приобрести даже бимодальную форму – вариант особенно нежелательный, если мы хотим оценить интервал, в котором находится истинное значение D^2.

Заключение

1. Полученные результаты демонстрируют известную необходимость использования поправки на число наблюдений при расчетах расстояний Махаланобиса даже при использовании усредненной ковариационной матрицы. Поправка Д. Райтмайра, автоматически применяемая ко всем расстояниям, в среднем позволяет получить близкие к истинным значения D^2. Именно такой алгоритм получения несмещенной оценки реализован в программе CANON. Однако для расчетов можно воспользоваться и программой MultiCan. При помощи таблиц Excel значения в матрице расстояний Махаланобиса возводятся в квадрат, а затем из каждого из них вычитается поправка Д. Райтмайра. Для удобства пользователей автор подготовил таблицу для оценки несмещенных D^2 на основе матриц выборочных квадратов расстояний Махаланобиса. Она доступна на страничке автора на Academia.edu.

2. В качестве показателей n_1 и n_2 в формуле поправки Д. Райтмайра лучше всего использовать не среднее арифметическое, а среднее гармоническое числа наблюдений, особенно

Таблица 4. Медианы и 95%-ые доверительные интервалы для расстояний Махаланобиса, рассчитанные путем бутстрэппинга (100 псевдовыборок)
Table 4. Medians and 95% confidence intervals for Mahalanobis distances calculated by bootstrapping (100 pseudo-samples)

Сравниваемые подвыборки по 10 черепов в каждой (D^2 между полными сериями)	**Среднее**	**Медиана**	**95%-ый интервал**
Башкиры-чуваши (5,1)	6,6	5,9	3,2 – 14,8
Башкиры-латыши (8,8)	12,1	11,7	8,2 – 18,1
Чуваши-латыши (4,2)	6,3	6,3	3,1 – 10,7
башкиры-башкиры (0)	2,0	1,6	-0,1 – 6,5
Сравниваемые подвыборки (по 20 черепов в каждой)	**Среднее**	**Медиана**	**95%-ый интервал**
Башкиры-чуваши (5,1)	4,4	4,3	2,5 – 6,5
Башкиры-латыши (8,8)	9,1	8,9	5,8 – 12,0
Чуваши-латыши (4,2)	6,4	6,4	3,9 – 10,0
Башкиры-башкиры (0)	0,4	0,1	-0,5 – 2,5
Сравниваемые подвыборки (по 30 черепов в каждой)	**Среднее**	**Медиана**	**95%-ый интервал**
Башкиры-чуваши (5,1)	4,5	4,5	3,0 – 6,2
Башкиры-латыши (8,8)	9,4	9,3	7,2 – 12,3
Чуваши-латыши (4,2)	6,1	6,0	4,5 – 8,2
Башкиры-башкиры (0)	1,5	1,4	0,4 – 3,3
Сравниваемые подвыборки (по 50 черепов в каждой)	**Среднее**	**Медиана**	**95%-ый интервал**
Башкиры-чуваши (5,1)	3,9	3,8	2,6 – 5,5
Башкиры-латыши (8,8)	7,3	7,1	5,5 – 9,7
Чуваши-латыши (4,2)	4,8	4,8	3,4 – 6,7
Башкиры-башкиры (0)	0,5	0,4	-0,1 – 1,5
Сравниваемые подвыборки (по 10 и 50 черепов)	**Среднее**	**Медиана**	**95%-ый интервал**
Башкиры-чуваши (5,1)	9,0	8,4	4,8 – 16,6
Башкиры-латыши (8,8)	11,0	10,8	6,0 – 18,2
Чуваши-латыши (4,2)	5,6	5,5	3,9 – 7,6
Башкиры-башкиры (0)	1,0	0,8	-0,3 – 3,5

Примечания. Расстояния указаны с поправкой Райтмайра, примененной ко всем D^2.

Notes. Distances are given with the Rightmire correction applied to all D^2.

в случае, когда число наблюдений между признаками заметно варьирует, либо поставлена задача оценки доверительного интервала, в котором находится истинное значение D^2. Более трудоемкой, но методически наиболее корректной процедурой может быть раздельный расчет расстояний Махаланобиса для линейных и угловых признаков, суммируемых после внесения отдельно оцениваемых поправок. Использование среднего арифметического позволяет получать близкие к реальным оценкам расстояния, но с тенденцией к некоторому их завышению.

3. При оценке различий между популяциями, каждая из которых представлена несколькими локальными выборками, истинная величина расстояния будет ближе не к средней величине D^2, а к их медиане. Без ущерба для точности расчетов все отрицательные значения выборочных D^2 могут быть приравнены к нулю.

Библиография

Гончаров И.А., Гончарова Н.Н. Программа MultiCan для анализа многомерных массивов данных с использованием статистик выборок и параметров генеральной совокупности (MultiCan). Свидетельство о регистрации прав на ПО №2016610803, М., 2016.

Козинцев А.Г. Скифы Северного Причерноморья: межгрупповые различия, внешние связи, происхождение // Археология, этнография и антропология Евразии, 2007. №32 (4). С. 143–157.

Козинцев А.Г. О ранних миграциях европеоидов в Сибирь и Центральную Азию (в связи с индоевропейской проблемой) // Археология, этнография и антропология Евразии, 2009. №40 (4). С.125–136.

Широбоков И.Г., Моисеев В.Г., Козинцев А.Г., Хартанович В.И. Чистов Ю.К., Громов А.В. Индивидуальные краниометрические данные близких к современности групп населения Восточной и Северо-Восточной Европы. Электронное издание. СПб.: МАЭ РАН, 2017.

Информация об авторах

Широбоков Иван Григорьевич, к.и.н.; ORCID ID: 0000-0002-3555-7509; ivansmith@bk.ru;

Поступила в редакцию 31.03.2025, принята к публикации 24.04.2025

Shirobokov I.G.

Peter the Great Museum of Anthropology and Ethnography (Kunstkamera) RAS, Department of Physical Anthropology, Universitetskaya emb., 3, Saint Petersburg, 199034, Russia

ON THE CORRECTION FOR THE NUMBER OF OBSERVATIONS IN THE CALCULATION OF MAHALANOBIS DISTANCES

Introduction. *Mahalanobis distances (D^2) are used in craniological studies to provide a generalised assessment of differences between samples, taking into account trait variances and correlations. However, sample distances tend to be biased upwards compared to the true D^2 value, especially in the case of small sample sizes. An adjustment for the number of observations, proposed by D. Rightmire, is one way of compensating for this bias. However, the conditions for its application and even the methods for calculating the variables may vary according to the characteristics of the sample. The aim of this study is to evaluate the effective impact of the adjustment on Mahalanobis distances and to explore approaches to reduce bias in sample estimates.*

Materials and methods. *The analysis was based on three aggregated series of male skulls from Bashkirs, Chuvash and Latvians. D^2 was calculated using an averaged covariance matrix. Three methods of calculating Mahalanobis distances were considered: with no adjustment for the number of observations, with Rightmire's adjustment applied to all distances, and with an adjustment that considers only statistically significant distances. In addition, alternative approaches to adjustment were tested in cases where there were large differences in the number of observations for individual traits: using the harmonic mean of the sample sizes and calculating D^2 separately for linear and angular traits, followed by summation.*

Results. *Rightmire's correction, when applied to all distances, generally provides accurate and unbiased estimates of D^2. When adjustments are applied, the D^2 value may approach zero or even become negative, even in cases where there are significant differences between populations. Since the medians of the sample D^2 values are closest to the true D^2 values, all negative D^2 values can be safely transformed to zero without compromising the calculations. A D^2 value of zero does not imply that there are no morphological differences between samples. To approximate the true distance between samples, confidence intervals can be calculated, e.g. using bootstrapping procedures.*

Conclusion. *Rightmire's correction allows for unbiased Mahalanobis distance estimates when using an averaged covariance matrix and small sample sizes. However, the choice of a specific correction method should take into account the sample size and the variability in the number of observations for different traits. When working with poorly preserved cranial series, it is advisable to use the harmonic mean of the sample sizes or separate distance calculations for angular and linear traits.*

Keywords: Mahalanobis distances; sample size; craniometry

DOI: 10.55959/MSU2074-8132-25-2-9

References

Goncharov I.A., Goncharova N.N. *Programma MultiCan dlya analiza mnogomernyh massivov dannyh s ispolzovaniem statistik vyborok i parametrov generalnoj sovokupnosti (MultiCan)* [MultiCan program for analyzing multidimensional data arrays using sample statistics and population parameters (MultiCan)]. Svidetelstvo o registracii prav na PO №2016610803, Moscow, 2016. (In Russ).

Kozintsev A.G. Scythians of the North Pontic Region: Between-group cranial variation, affinities, and origins. *Archaeology, Ethnology and Anthropology of Eurasia*, 2007, 32 (4), pp. 143–157. (In Russ.). DOI:10.1134/s1563011007040135.

Kozintsev A.G. Craniometric evidence of the early Caucasoid migrations to Siberia and Eastern Central Asia, with reference to the Indo-European problem. *Archaeology, Ethnology and Anthropology of Eurasia*, 2009, 40 (4), pp. 125–136. (In Russ.). DOI: 10.1016/j.aeae.2010.02.014.

Shirobokov I.G., Moiseev V.G., Kozintsev A.G., Khartanovich V.I. Chistov Y.K., Gromov A.V. Individualnye kraniometricheskie dannye blizkih k sovremennosti grupp naseleniya Vostochnoj i Severo-Vostochnoj Evropy [Individual craniometric data of modern population groups from Eastern and North-Eastern Europe]. *Elektronnoe izdanie*. Saint Petersburg: MAE RAS Publ., 2017. (In Russ).

Rightmire G.P. On the computation of Mahalanobis' generalized distance (D^2). *American Journal of Physical Anthropology*, 1969, 30, pp. 157–160.

Sjøvold T. Some notes on the distribution and certain modifications of Mahalanobis generalized distance. *Journal of Human Evolution*, 1975, 4, pp. 549–558.

Information about the author

Shirobokov Ivan G., PhD.; ORCID ID: 0000-0002-3555-7509; ivansmith@bk.ru;