***Various proximities***

SPSS macros by Kirill Orlov

kior@akado.ru, ttnphns@gmail.com

<https://www.spsstools.net/en/KO-spssmacros>

All rights reserved

*Разные меры близости.* Вычисление большого числа мер близости или связи (сходства, расстояния, корреляции), многие из которых отсутствуют в SPSS. Среди них сходство Гауэра для сравнения респондентов по количественным и качественным признакам сразу.

*Прочтите «*[*О SPSS макросах*](https://www.spsstools.net/ru/KO-aboutmacros)*» что они такое и как их запускать.*

*Ошибка “Protected directory”.* Некоторые из макросов, описанных в текущем документе, пишут временные файлы на жесткий диск. Если вы не обладаете полными правами Администратора вашего компьютера, это может вызвать ошибку, сообщающую среди прочего: *“SPSS Statistics cannot access a file... specifies a protected directory...”* и значащую, что дефолтная директория, какую макрос хочет использовать, защищена на вашем ПК. Чтобы решить эту проблему, в окне синтаксиса скомандуйте: CD 'myfolder'., где 'myfolder' есть путь/имя некоторой папки, куда вам разрешено сохранять файлы.

* [!KO\_EDPROXMX](#_МАКРОС_!EDPROXMX:_РЕДАКТИРОВАНИЕ) это подсобный макрос, позволяющий сделать некоторую редакторскую работу над сущестующей квадратной матрицей близости, например вычленить из нее подматрицу, сделать матрицу симметричной, заменить диагональ и т.д.
* Сходство Гауэра [!KO\_GOWER](#_МАКРОС_!GOWER:_СХОДСТВО_1) для калькуляции подобия между наблюдениями по количественным и качественным признакам сразу.
* Разные меры близости – расстояния, сходства – для количественных данных [!KO\_PROXQNT](#_МАКРОС_!PROXQNT:_РАЗНЫЕ).
* Разные меры близости – расстояния, сходства – для двоичных данных [!KO\_PROXBIN](#_МАКРОС_!PROXBIN:_РАЗНЫЕ_1).
* Тетрахорический коэффициент корреляции [!KO\_TETRACH](#_МАКРОС_!TETRACH:_ТЕТРАХОРИЧЕСКИЙ).
* Бисериальный коэффициент корреляции [!KO\_BISER](#_МАКРОС_!BISER:_БИСЕРИАЛЬНЫЙ_1).
* Автокоррелятивное расстояние [!KO\_ACORRD](#_МАКРОС_!ACORRD:_АВТОКОРРЕЛЯТИВНОЕ_1).
* Перешкалированный коэффициент корреляции Пирсона [!KO\_RESCR](#_МАКРОС_!RESCR:_ПЕРЕШКАЛИРОВАННЫЙ).

*Внимание*, для всех этих макросов рекомендуются имена входящих переменных не длиннее 8 байтов, иначе на выходе они будут обрезаны.

# МАКРОС !KO\_EDPROXMX: РЕДАКТИРОВАНИЕ МАТРИЦЫ БЛИЗОСТЕЙ

Version 2, Sep 2023 (Version 1, Mar 2020). Tested on SPSS Statistics 22, 26, 28.

!KO\_edproxmx matrix= var1 to var250 /\*Столбцы, образующие тело матрицы; можно через to

/seq= ROW /\*Последовательность рядов/столбцов в выводимой матрице: (COL, тж п/у)

/\*ROW, GRAD1, GRAD2, IND 'файл'

/oper= AVER /\*Операция внутри тела: NONE (тж п/у), AVER, GMEAN, HMEAN, UBYL, LBYU,

/\*SWITHG, GWITHS, FLIP, DIAG v, ABS, RANK, URANK, LRANK,

/\*UADD v, LADD v, UMULT v, LMULT v, UEXP v, LEXP v

/numvars= height bmi /\*Опционально: прихватить эти числовые переменные

/strvars= v3 to v5 /\*Опционально: прихватить эти короткие текстовые переменные

/dgelem= /\*При seq=GRAD1/GRAD2: диагональные элементы участвуют (YES, тж п/у) или нет (NO)

/savedg= /\*Опционально: сохранить диагональ как внешний файл (путь/имя файла или массив)

/savetr= /\*Опционально: развернув, сохранить треугольники как внешний файл

/\*(путь/имя файла или массив)

/plot= GREY /\*Выдать тепловую карту: NONE (тж п/у), COLOR или GREY или RGREY;

/\*после можно добавить слово LABEL

/bounds= OBS /\*При заданном PLOT: пределы для тепловой шкалы на карте: AUTO (тж п/у),

/\*OBS или min max.

Минимум надо задать MATRIX.

Данный макрос не вычисляет близостей или статистик, а служит средством редактирования существующей матрицы близостей между предметами одного набора – т.е. «квадратной матрицы» (соответствует однодольному графу). Макрос может сделать следующие вещи:

* Собрать матрицу в любом нужным вам, но правильном порядке – т.е. чтобы предметы, задающие ряды/столбцы, шли в рядах и столбцах в одной и той же последовательности – *соупорядоченно* (следовательно, элементы *i,i* будут лежать на диагонали).
* Вычленить подматрицу.
* Сделать кое-какие операции в теле матрицы (например, сделать асимметричную матрицу симметричной, прибавить константу, транспонировать, поставить диагональ).
* Сохранить во вне диагональ или развернутые треугольники матрицы.
* Нарисовать тепловую карту (heatmap) матрицы.

Макрос принимает матрицу, делает свою работу и выводит получившуюся матрицу в новый безымянный массив. Макрос не создает переменной *ROWTYPE\_* при матрице. Вы можете создать ее вручную или взять из исходного массива (см. п/к STRVARS).

ПРИМЕР 1. Вычленение и переупорядочивание.

proximities v1 to v100 /view= case /measure= euclid /matrix= out(\*) /print= none.

dataset name prox.

!KO\_edproxmx matrix= var15 var14 var10 var20 to var24 /numvars= caseno\_ /strvars= rowtype\_.

value label rowtype\_ 'PROX' 'DISSIMILARITY EUCLID'.

* Команда PROXIMITIES создала квадратную матрицу расстояний между наблюдениями данных и вывела ее как новый массив.
* Макрос !KO\_EDPROXMX взял из нее только ряды/столбцы следующие и в таком порядке: 15 14 10 20 to 24. И сохранил эту матрицу как новый безымянный массив.
* Макрос также прихватил и вывел переменные исходного массива CASENO\_ и ROWTYPE\_ (они были вспомогательными при матрице). Инструктивный ярлык значения в ROWTYPE\_ восстановлен командой VALUE LABELS.

ПРИМЕР 2. Просто нарисовать теплокарту для корреляционной матрицы.

dataset declare corr.

correlations /variables= v1 to v10 /matrix= out(corr).

dataset activate corrmat.

select if rowtype\_='CORR'.

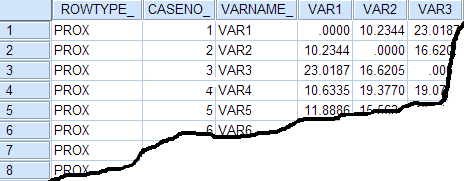
!KO\_edproxmx matrix= v1 to v10 /plot= COLOR.

* Команда CORRELATIONS создала квадратную матрицу корреляций между переменными данных и вывела ее как массив *CORR*.
* Этот массив активирован и в матрице оставлены только ряды, соответствующие корреляциям.
* !KO\_EDPROXMX взял целиком тело матрицы и построил тепловую карту.

***Входящая матрица***

Массив данных должен содержать «матрицу попарных близостей» между предметами одного набора. Вообще, в матрице могут быть любые числовые значения, кроме пропусков. Матрица может быть симметричной или асимметричной. Имена переменных – столбцов матрицы – до 8 байтов. Обязательно присутствие короткой текстовой переменной *VARNAME\_*, именующей ряды в соответствие столбцам. Имена, являющиеся значениями этой переменной, должны быть написаны в том же регистре, как тождественные им имена среди имен переменных. Макрос *не* требует, чтобы ряды и столбцы шли в одинаковом порядке или чтобы их число и состав были полностью одинаковы: как раз задача макроса состоит в том, чтобы выбрать из входящей матрицы одинаковые своими именами ряды и столбцы и соупорядочить их, и вернуть вам матрицу квадратную и соупорядоченного (диагонализованного) строения. Переменная *ROWTYPE\_* и прочие вспомогательные – не обязательны при входящей матрице.

Кроме переменных, задающих тело матрицы и переменной *VARNAME\_*, массив может содержать любые другие переменные.



***Подкоманды***

**MATRIX**

Укажите переменные рабочего массива, являющиеся столбцами собственно матрицы близостей. Вы можете указать все или только нужные столбцы и в произвольном порядке. Можно использовать “to” для задания диапазоном.

*Задание открытым диапазоном с помощью “?”*. В некоторых случаях возникает нужда указать диапазон, заключенный между парой переменных, которые сами не входят в диапазон. Используйте для этого “?” по краям. Например, *?VARNAME\_ to ENDVAR?* означает все переменные, находящиеся в массиве между переменными *VARNAME\_* и *ENDVAR*, не включая их самих. Для обозначения диапазона, открытого с одного конца, используйте “?” только у этого конца. Например: *?VARNAME\_ to VAR100* или *VAR1 to ENDVAR?*.

Когда вы желаете, чтобы макрос вычленил для вас квадратную подматрицу из большей матрицы, вы можете пойти двумя путями: (1) указать только нужные вам столбцы – это делается в п/к MATRIX; и/или (2) выбрать только нужные вам ряды – это делается селекцией или фильтрацией кейсов массива. В любом случае макрос соберет квадратную соупорядоченную (под)матрицу, потому что существует переменная *VARNAME\_*, где записаны имена предметов, и есть имена столбцов матрицы – всё те же имена предметов.

**SEQ**

Эта подкоманда задает, к какой последовательности должны идти предметы (т.е. ряды/столбцы) в выходящей матрице:

COL - (тж по умолчанию/незаданию) ряды и столбцы будут идти в порядке списка имен столбцов в подкоманде MATRIX.

ROW - ряды и столбцы будут идти в порядке расположения рядов (значений VARNAME\_) во входящей матрице. Если матрица, введенная подкомандой MATRIX, уже квадратна и с соупорядоченными рядами и столбцами, то это даст тот же результат, что SEQ=COL.

GRAD1 - сделать сортировку рядов/столбцов по «градиенту мультипликативному». Для каждой пары *одноименных* ряда *R* и столбца *C* матрицы размером *I×I*, готовящейся на выход, вычисляется величина , т.е. кросспроизведение. Потом ряды/столбцы сортируются по убыванию этой величины.

GRAD2 - сделать сортировку рядов/столбцов по «градиенту аддитивному». Для каждой пары *одноименных* ряда *R* и столбца *C* матрицы размером *I×I*, готовящейся на выход, вычисляется величина , т.е. объединенная сумма. Потом ряды/столбцы сортируются по убыванию этой величины.

IND *файл* - сортировать по индексам, хранящимся как внешний .SAV-файл. Укажите после кл. слова путь/имя этого файла в кавычках или апострофах. Файл должен представлять собой единственную переменную (с любым именем), значения которой – индексы (позиции) предметов в выходящей матрице. Например, если 8-е наблюдение в *файле* есть значение 17, то 8-й ряд/столбец входящей матрицы станет 17-м в выходящей матрице. Индексы должны быть положительными целыми (нецелые значения макрос обрежет, округлит в меньшую сторону) не выше, чем число рядов/столбцов в матрице.

Для GRAD1/GRAD2 имеет значение подкоманда DGELEM. Также отметить, переупорядочиваине GRAD1/GRAD2 зависит от значений элементов тела матрицы и делается оно *после* опциональной операции с элементами (п/к OPER).

*Замечания к* IND.

1. Эта опция требует, чтобы входящая матрица (матрица, как она берется подкомандой MATRIX) была уже квадратной и с соупорядоченными рядами/столбцами. В противном случае макрос сообщит об ошибке.
2. Индексы в *файле* могут повторяться; число наблюдений в нем может превышать или недостигать числа предметов во взятой подкомандой MATRIX матрице. Таким образом, SEQ=IND позволяет не только задать порядок рядов/столбцов, но и брать подматрицу из матрицы или же, напротив, распложать некоторые ряды/столбцы сколько угодно раз.

ПРИМЕР 3. Изменение порядка предметов в матрице на иной, случайный.

compute sortvar= uniform(1).

sort cases by sortvar.

!KO\_edproxmx matrix= var1 to var100 /seq= ROW.

* Строки входящей матрицы сортируются по значениям случайной переменной SORTVAR.
* SEQ=ROW макроса составляет матрицу с порядком рядов/столбцов, следующим порядку рядов во входящей матрице.

**OPER**

Опционально вы можете заказать одну операцию над элементами тела матрицы. Выберите:

NONE - (тж по умолчанию/незаданию) ничего не делать с элементами.

AVER - сделать матрицу симметричной путем усреднения (обычная средняя) симметричных по положению элементов.

GMEAN - сделать матрицу симметричной путем усреднения (геометрическая средняя) симметричных по положению элементов; требует неотрицательных значений.

HMEAN - сделать матрицу симметричной путем усреднения (гармоническая средняя) симметричных по положению элементов; требует неотрицательных значений.

UBYL - сделать матрицу симметричной путем замены элемента верхнего треугольника симметричным по положению элементом нижнего треугольника.

LBYU - сделать матрицу симметричной путем замены элемента нижнего треугольника симметричным по положению элементом верхнего треугольника.

SWITHG - сделать матрицу симметричной путем замены меньшего элемента большим элементом в паре симметричных по положению элементов.

GWITHS - сделать матрицу симметричной путем замены большего элемента меньшим элементом в паре симметричных по положению элементов.

FLIP - транспонировать, т.е. поменять верхний и нижний треугольники местами.

ABS - отменить отрицательный знак у внедиагональных элементов.

RANK - ранжировать по возрастанию внедиагональные элементы.

URANK - ранжировать по возрастанию элементы верхнего треугольника.

LRANK - ранжировать по возрастанию элементы нижнего треугольника.

DIAG *val* или *var* - поставить новую диагональ в матрицу. После кл. слова укажите или скаляр (число), или имя числовой переменной массива; значения из этой переменной будут поставлены на диагональ; пропуски в переменной не разрешены.

UADD *val* - прибавить константу к элементам верхнего треугольника. Укажите число после кл. слова.

LADD *val* - прибавить константу к элементам нижнего треугольника. Укажите число после кл. слова.

UMULT *val* - умножить элементы верхнего треугольника на константу. Укажите число после кл. слова.

LMULT *val* - умножить элементы нижнего треугольника на константу. Укажите число после кл. слова.

UEXP *val* - экспоненциировать элементы верхнего треугольника: возвести в степень-константу. Укажите ненулевое число (степень) после кл. слова.

LEXP *val* - экспоненциировать элементы нижнего треугольника: возвести в степень-константу. Укажите ненулевое число (степень) после кл. слова.

*Замечания к* UEXP/LEXP.

1. Если в треугольнике матрицы есть отрицательные значения, экспонента (степень) должна быть целым.
2. Если в треугольнике матрицы есть нулевые значения, экспонента (степень) должна быть положительным.

ПРИМЕР 4.

!KO\_edproxmx matrix= var1 to var100 /seq= GRAD1 /DGELEM= YES /OPER= DIAG newdiag /plot= COLOR.

* Берутся столбцы VAR1 до VAR100 матрицы близостей (это может быть вся матрица, а может, только первые 100 столбцов).
* SEQ=GRAD1 заказывает переупорядочить ряды/столбцы по величине элементов, «мультипликативно», причем вклад диагональных элементов учитывается: DGELEM=YES.
* OPER=DIAG ставит новую диагональ, значения которой почерпнуты из переменной *NEWDIAG* входящего массива. SEQ=GRAD1 переупорядочивает после OPER, поэтому учитываются новые, не старые диагональные элементы.

**NUMVARS, STRVARS**

Вы можете прихватить некоторые посторонние телу матрицы переменные массива, чтобы их наблюдения (ряды) подверглись переупорядочиванию или отсеиванию вослед пересборке матрицы близостей. Укажите числовые переменные в подкоманде NUMVARS и текстовые переменные в подкоманде STRVARS. Можно использовать “to” для задания диапазоном. Имена – до 8 байтов длиной. Текстовые переменные должны быть короткими (ширина переменной до 8 байтов). Значение: -999 не разрешено в числовых переменных. *Не* указывайте переменную *VARNAME\_*: макрос сам создает такую переменную на выход.

Пропуски в переменных разрешены, однако пользовательские пропущенные значения (user-missing) на выходе в новый массив потеряют свой missing статус. Его, а также ярлыки переменных и их значений – их макрос не сохранит – вам придется восстанавливать самим (это легко сделать в меню Data – Copy Data Properties).

**DGELEM**

Эта подкоманда нужна только для SEQ= DIAG1 или DIAG2. Задайте DGELEM=YES (тж. по умолчанию), если надо, чтобы диагональные элементы матрицы (т.е. значения на пересечении одноименных рядов и столбцов) участвовали в вычислении «градианта», по падению которого будут сортированы ряды/столбцы. Задайте DGELEM=NO, если при вычислении «градиента» диагональные элементы надо игнорировать.

**SAVEDG**

Этой подкомандой можно сохранить диагональ матрицы, готовой на выход, как отдельный внешний .SAV-файл или массив. Укажите путь/имя файла в кавычках или апострофах или имя заявленного массива. Переменная *VARNAME\_* сохраняется тоже.

**SAVETR**

Этой подкомандой можно, развернув в векторы, сохранить оба треугольника матрицы, готовой на выход, как отдельный внешний .SAV-файл или массив. Укажите путь/имя файла в кавычках или апострофах или имя заявленного массива. Будут сохранены следующие столбцы. *LOWER* (нижний треугольник), *UPPER* (верхний треугольник); наблюдение – пара симметричных по положению элементов матрицы. Столбцы *COLROW* и *ELEM* показывают положение, координаты тех элементов в матрице: *COLROW* есть номер столбца (у нижнего элемента, *LOWER*) или же номер ряда (у верхнего элемента, *UPPER*), а *ELEM* это номер элемента в этом столбце/ряду. Столбцы *VARNAME1* и *VARNAME2* – это соответствующие имена из *VARNAME\_*.

**PLOT**

Строит тепловую диаграмму (heatmap) результатной, выходящей матрицы. PLOT=COLOR делает ее цветной (величина элементов передается тоном). PLOT= GREY или RGREY делает ее серо-полутоновой (при GREY чем выше значение элемента, тем он ярче, а при RGEY – наоборот, тем он темнее). После вы можете добавить второе кл. слово LABEL, для оярлычения клеток значениями.

PLOT=NONE (тж. по умолчанию) не рисует диаграмму.

**BOUNDS**

Эта подкоманда действует только при заданной PLOT. Задает границы для цветовой (или яркостной) шкалы.

AUTO - (тж по умолчанию/незаданию) позволить SPSS автоматически определить подходящие границы.

OBS - границы точно совпадают с наблюдаемыми минимальным и максимальным значениями в матрице.

*min* *max* - укажите границы вручную в виде двух чисел – минимума и максимума. Указывайте значения, в общем и целом сопоставимые со значениями матрицы.

Задание границ вручную означает, что цветовое/яркостное отображение на рисунке вами фиксировано относительно величины содержащихся в матрице значений. Становится возможным сравнивать разные матрицы между собой. Например, если вы зададите *min* *max* как *-1 1*, то корреляционная матрица, содержащая меньшие по абс. величине коэффициенты, получится более тусклой или более одноцветной, чем матрица, содержащая большие по абс. величине коэффициенты; и такие рисунки «в единой шкале» можно сравнивать.

***Особые режимы***

Макрос игнорирует взвешивание массива (впрочем, он не берет в процедуру наблюдения с пропущенными и неположительными весами) и не рассчитан на расщепленность массива. Макрос слушается выбора/фильтрации наблюдений (команды FILTER, USE, SELECT IF, N OF CASES) и временных (под TEMPORARY) преобразований.

# МАКРОС !KO\_GOWER: СХОДСТВО ГАУЭРА

Version 5, Oct 2019 (Version 1, Dec 2000). Tested on SPSS Statistics 20, 22, 25.

!KO\_gower sca= *v1 v2 v10* /\*Поименно мерные переменные, если есть

/ord= *v3 v5* /\*Поименно порядковые переменные, если есть

/bin= *v4 v8* /\*Поименно двоичные (0 и 1) переменные, если есть

/nom= *v6* /\*Поименно номинальные переменные (числовые), если есть

/cnt= *v7 v16* /\*Поименно частотные переменные, если есть

/weights= *1 1 1.5 1 1 1 1 2 1 0.5* /\*Опционально: веса для входящих переменных,

/\*в порядке SCA ORD BIN NOM CNT: любые неотрицательные числа

/missing= /\*Что делать с пропущенными значениями: LISTWISE (тж п/у) - списочно исключить

/\*наблюдения; PAIRWISE – взять все наблюдения, используя в расчетах все валидные

/\*данные, где можно. Опциональное дополнительное второе слово

/\*ASVALID – считать user-missing данные валидными

/id= /\*Опционально: текстовая переменная-идентификатор наблюдений

/flatten= /\*Опционально для мерных переменных: уплощить эксцесс у этих переменных:

/\*поименный список

/add05= /\*Опционально для частотных переменных: прибавлять 0.5 к частотам этих переменных:

/\*поименный список

/reverse= /\*Перевести сходство Гауэра в расстояние: не надо (NONE, тж п/у),

/\*вычитанием из 1 (ONEMINUS), корень после вычитания из 1 (RONEMINUS)

/\*угол в радианах (ARCCOS)

/ecorrect= /\*Проверка и коррекция расстояний к евклидовости: не делать (NONE, тж п/у),

/\*LINGOES (метод Линго), CAILLIEZ (метод Кайе).

Минимум надо задать хотя бы одно из SCA, ORD, BIN, NOM, CNT.

Коэффициент сходства Гауэра (Gower, 1971) является мерой подобия между двумя объектами (именно наблюдениями, а не переменными). Его особенность состоит в том, что он способен учитывать вместе признаки, разнородные по своей измерительной природе. Одни из переменных, на базе которых устанавливается этот резюмирующий показатель подобия, могут быть количественными мерными или порядковыми, другие - двоичными (признак есть или его нет), третьи – номинальными (рядоположенные категории, к примеру, пол или профессия). Гауэр предложил коэффициент для смеси мерных, двоичных и номинальных переменных. Слагаемые для порядковых и счетных переменных добавлены позже.

Коэффициент сходства между объектами *i* и *j* является средневзвешенной сходств по каждой из *p* переменных (*p*≥1):

,

где *k* – переменная, *Sijk* – сходство по ней, *wk* – вес, придаваемый данной переменной. Веса можно задать; по умолчанию они все 1. Парциальное сходство *Sijk* варьирует в пределах [0,1] и вычисляется по-разному в зависимости от типа переменной *k*.

* Если переменная **номинальная**, то *Sijk* = 1, если *xik* =*xjk*, т.е. значения объектов *i* и *j* по переменной *k* равны; в противном случае *Sijk* = 0. Таким образом, если все *p* переменных номинальные, коэффициент Гауэра равен коэффициенту **Дайса** (Dice), который получится, если номинальные переменные перекодировать в двоичные фиктивные (dummy).
* Если переменная **двоичная** (1 vs 0), то *Sijk* = 1, если *xik* =*xjk* = 1; в противном случае *Sijk* = 0. А также, если *xik* =*xjk* = 0, то при сравнении *i* с *j* по переменной *k* вес ее зануляется: *wk* = 0. Таким образом, если все *p* переменных двоичные, то коэффициент Гауэра превращается в коэффициент **Жаккара** (Jaccard).
* Если переменная **мерная** (интервальная или отношений), то

,

где *xk\_max* и *xk\_min* есть, соответственно, максимум и минимум в переменной *k* (если они равны, то *Sijk*  принимается за 1). Таким образом, если все *p* переменных мерные, то коэффициент Гауэра есть **нормированное диапазоном манхэттенское расстояние**, переведенное в сходство.

* Если переменная **порядковая**, то, согласно нововведению Подани (Podani, 1999), переменная *k* ранжируется (*x* заменяются рангами, *r*) классическим образом (т.е. с усреднением связанных рангов, ties), и

*Sijk* = 1, если *rik* = *rjk*;

в ином случае:

,

где *rk\_max* и *rk\_min* есть, соответственно, максимальный и минимальный ранг в переменной *k*; *Tk\_max* и *Tk\_min* есть число объектов, соответственно, с максимальным и с минимальным рангом; *Tik* и *Tjk* есть число объектов с таким же рангом как у, соответственно, объектов *i* и *j* (включая сами *i* и *j*). Эта формула специально нивелирует «отдаляющее» вмешательство связанных рангов (ties). Подани предложил также более простой вариант без такой нивелировки: после ранжирования переменной использовать с ее рангами ту же формулу, что используется для мерной переменной.

* Если переменная **счетная** (частоты), то, по добавлению автора макроса, *xik* и*xjk* рассматриваются как пара взаимодополнительных частот, подлежащих сравнению хи-квадратом согласия, т.е.

,

где ожидаемая частота есть средняя двух сравниваемых: ,

так что , откуда парциальное сходство получаем таким:

; а если обе частоты нулевые, *Sijk* = 1.

Таким образом, если все *p* переменных счетные, то коэффициент Гауэра есть **канберрово расстояние**, переведенное в сходство.

Источники:

* Gower, J. C. A general coefficient of similarity and some of its properties // *Biometrics*, 1971, 27, 857-872.
* Podani, J. Extending Gower’s general coefficient of similarity to ordinal characters // *Taxon*, 1999, 48, 331-340.

Макрос выдает квадратную симметрическую матрицу коэффициентов Гауэра как новый безымянный рабочий массив. В макросе есть возможность придать входящим переменным веса, отражающие вклад каждой переменной в коэффициент. Если взяты наблюдения с пропущенными значениями (MISSING=PAIRWISE), то наблюдения сравниваются попарно по тем переменным, в которых оба наблюдения валидны (кроме того, в двоичных переменных – как сказано выше – сравнение делается лишь в случае, когда хотя бы одно из двух наблюдений имеет признак). Если окажется, что для некоторых наблюдений недостаточно данных для сравнения с другими наблюдениями, эти наблюдения макрос исключит из итоговой матрицы коэффициентов и сообщит об этом.

Столбец *CASENO\_* при матрице это номер наблюдения в исходном массиве.

Входящие переменные *должны не содержать значения: -999*. В рабочем массиве должно не быть переменных с именем CASENO#$ или STRID#$.

Поскольку вычисляемые сходства нормируются у количественных переменных их диапазонами, то удаление из файла части наблюдений потенциально отражается на коэффициентах, посчитанных для оставшихся наблюдений.

Подобно большинству коэффициентов сходства, коэффициент Гауэра принимает значения от 0 до 1. Вы можете обратить сходство в расстояние п/к REVERSE.

***Подкоманды***

**SCA**

Поименно входящие мерные переменные. Это может быть интервальная или отношенческая шкала измерения. Единицы измерения и величинность в переменных могут быть разные – это не влияет на результат, поскольку сходства по каждой переменной нормируются делением на диапазон значений в ней. Однако вы должны следить за выбросами, т.к. они преувеличивают диапазон.

**ORD**

Поименно входящие числовые категориальные переменные, которые вы трактуете как порядковую шкалу измерения. Единицы измерения и величинность в переменных могут быть разные – это не влияет на результат, поскольку макрос ранжирует значения в этих переменных.

Как вариант, вы можете предпочесть не использовать подкоманду ORD, а заменить порядковые переменные ранжированными (меню Transform – Rank Cases), которые обработать как мерные, указав их в подкоманде SCA. При наличии значительного числа связанных рангов (ties) такой вариант следует однако признать менее корректным.

**BIN**

Поименно входящие числовые двоичные переменные. Имеются в виду переменные «есть признак – нет признака», а не номинальные дихотомические переменные (типа «женщина – мужчина»). Наличие признака должно быть кодировано единицей, а отсутствие нулем.

**NOM**

Поименно входящие числовые номинальные переменные. Кодировка категорий в них произвольна.

**CNT**

Поименно входящие числовые частотные счетные переменные. Значения – неотрицательные целые.

**WEIGHTS**

Необязательная подкоманда, которой можно придать веса входящим переменным – их вклад в вычисляемый коэффициент. Это могут быть любые неотрицательные числа, находящиеся в нужном вам соотношении (хотя бы некоторые должны быть положительны). Только соотношение весов играет роль. Чисел должно быть столько, сколько входящих переменных, и они должны следовать порядку SCA ORD BIN NOM CNT.

ПРИМЕР 1.

!KO\_gower sca= v1 v2 v3 /bin= /nom= w6 occup /weights= 2 2 2 3 3.

* Переменным V1 V2 V3 W6 OCCUP приданы, соответственно, веса 2 2 2 3 3.
* Порядковых входящих переменных нет: п/к ORD опущена. Двоичных входящих переменных нет, п/к BIN можно было и опустить.

**FLATTEN**

Эта опция касается только мерных переменных (SCA). Вы можете перечислить поименно те переменные из заданных в SCA, у которых вы желаете уплощить форму распределения. FLATTEN сдавливает, до вычисления близости, распределение в переменной сверху, уменьшая эксцесс и укорачивая длинные хвосты. Вы можете захотеть использовать опцию FLATTEN только для тех континуальных переменных, у которых эксцесс много больше, чем у нормального распределения. В такой островершинной переменной диапазон слишком велик по сравнению с различиями между большинством ее наблюдений, так что сходство *Sijk* в массе близится к 1, – если FLATTEN не сделать. FLATTEN уплощает распределение, преобразуя переменную по формуле , где *Z* это переменная после ее z-стандартизации.

ПРИМЕР 2.

!KO\_gower sca= v1 v2 v3 /bin= b2 b7 b8 b13 /nom= region /flatten= v2.

* Континуальная переменная V2 имеет острую вершину и длинные хвосты, поэтому решено уплощить ее распределение.

**ADD05**

Эта опция касается только счетных переменных (CNT). Вы можете перечислить поименно те переменные из заданных в CNT, у которых вы желаете к значениям прибавить 0.5, чтобы избавиться от нулей. ADD05 прибавляет, до вычисления близости, всем частотам переменной величину 0.5. Сходство между частотой любой величины и нулевой частотой всегда равно 0. Например, 1 - |9-0| / (9+0) есть нулевое сходство, как и 1 - |1-0| / (1+0), несмотря на то что 9-0 куда бо̀льшая разница, чем 1-0. Если вам этот факт не нравится (хотя вообще-то он не лишен смысла для частот), то закажите ADD05 для всех переменных CNT или только для тех, где вы хотите избавиться от нулей. Тогда 1 - |9.5-0.5| / (9.5+0.5) = 0.1, что будет меньшее сходство, чем 1 - |1.5-0.5| / (1.5+0.5) = 0.5.

**MISSING**

Подкоманда для обращения с пропущенными данными. Выберите:

LISTWISE – (тж по умолчанию/незаданию) списочное исключение наблюдений с пропусками. Если у наблюдения есть хотя бы один пропуск в анализируемых макросом переменных, наблюдение исключается из анализа.

PAIRWISE – все наблюдения допускаются в анализ. Сравнение между парой наблюдений идет только по тем переменным, где оба значения валидны.

После любого из этих ключевых слов вы можете написать ASVALID, если хотите, чтобы макрос воспринял пользовательские пропущенные значения (user-missing), будь таковые в анализируемых переменных, как валидные значения. По умолчанию же макрос считает пользовательские пропуски пропусками.

**ID**

Вы можете указать текстовую переменную (не шире 8 байтов) на роль идентификатора наблюдений в выходящей матрице коэффициентов. По умолчанию/незаданию этой подкоманды макрос создаст текстовый идентификатор с именем *ID*, значения которого состоят латинского “C” и номером наблюдения, совпадающим с *CASENO\_*.

**REVERSE**

Мера Гауэра *S* является сходством. По умолчанию и при REVERSE=NONE макрос выдает матрицу сходств. Вы можете захотеть обратить сходство в «расстояние Гауэра» *D* по любой из следующих формул:

ONEMINUS *D* = 1-*S*. Это расстояние колеблется от 0 до 1, как и *S*.

RONEMINUS *D* = sqrt(1-*S*). Это расстояние колеблется от 0 до 1, как и *S*.

ARCCOS *D* = arccos(*S*). Это расстояние рассматривает *S* как косинус угла, а само является углом в радианах.

*Свойства этих расстояний*

Метричность (удовлетворение треугольному неравенству):

Все три вида расстояния, 1-*S*, sqrt(1-*S*) и arccos(*S*) являются метриками, если не задана п/к ORD. При задании ORD они могут перестать быть метриками (могут нарушать «треугольное неравенство»).

Геометрическая евклидовость (схождение расстояний в евклидовом пространстве):

sqrt(1-*S*) гарантированно евклидовы, если не задана п/к ORD. При задании ORD они могут перестать быть евклидовыми.

arccos(*S*) гарантированно евклидовы, если задана только п/к BIN.

1-*S* почти всегда не евклидовы.

Вы можете затребовать довести полученную матрицу расстояний до евклидовости подкомандой ECORRECT. Если расстояния геометрически евклидовы, то они автоматически и метричны.

**ECORRECT**

Эта опциональная подкоманда работает, если REVERSE≠NONE, иначе она игнорируется. Нередко пользователю желательно, чтобы наблюдаемые (вычисленные) расстояния сходились в евклидовом пространстве (геометрическая евклидовость): тогда с ними можно делать любой анализ, годящийся для настоящих евклидовых расстояний. Выше в п/к REVERSE написано, какие из трех видов «расстояний Гауэра» оказываются геометрически евклидовы, и когда.

Если вы получили матрицу (размером не меньше 3), расстояния в которой могут не сходится в евклидовом пространстве, а вам нужно заставить их сходиться, то воспользуйтесь подкомандой ECORRECT. Эта подкоманда прибавляет минимальную константу ко всем расстояниям матрицы, достаточную для того, чтобы сделать их евклидовыми. Не надо однако забывать, что прибавление константы есть искажение первоначальных близостей; поэтому желательно, чтобы она была малой. Макрос сообщит о величине потребовавшейся константы. Выберите способ вычисления константы:

NONE - (тж по умолчанию/незаданию) не проверять расстояния и не корректировать их к евклидовости.

LINGOES - метод Линго. Вычисляется константа, прибавляемая к квадратам расстояний, а евклидовыми становятся неквадратные расстояния. Это более быстрый метод, однако расстояния в матрице прирастают неравномерно (ведь изменяются их *квадраты*).

CAILLIEZ - метод Кайе. Вычисляется константа, прибавляемая к расстояниям как есть (неквадратным), они и становятся евклидовыми. Этот метод нередко ведет к большей инфляции расстояний, чем предыдущий метод, однако он изменяет расстояния в матрице равномерно.

Если матрица расстояний не нуждается в коррекции, т.к. уже евклидова, макрос не станет вычислять константу и сообщит об этом.

Константа Lingoes установлена как величина, чуть большая, чем , где есть последнее (минимальное) собственное число матрицы **C**, являющейся так называемым *двойным центратом* матрицы расстояний *D* (выдаваемых подкомандой REVERSE), возведенных в квадрат. (Двойной центрат это матрица расстояний, переведенная в матрицу скалярных произведений.)

Константа Cailliez установлена как величина, чуть большая, чем первое (максимальное) собственное число следующей асимметричной матрицы:

, где **P** есть двойной центрат матрицы расстояний, полученный так же, как **C**, но без возведения расстояний *D* в квадрат.

Коррекция успешна, если матрица-двойной центрат **C** потеряла отрицательные собственные числа.

Коррекция к евклидовости может быть калькуляторно емка при большой матрице и требовать времени. Поэтому не рекомендовано прибегать к ECORRECT, особенно методу CAILLIEZ, когда размер матрицы, т.е. число наблюдений, приближается к или превосходит 1000.

ПРИМЕР 3.

!KO\_gower ord= item3 item4 item5 item6 /bin= b1 b2 b3 b4 b5 /nom= n9 /reverse= ONEMINUS /ecorrect= LINGOES.

* REVERSE затребовало перевести сходство Гауэра в различие по формуле 1-S.
* Вероятно это расстояние не поддержит евклидово пространство, поэтому затребована коррекция Линго к евклидовости.

Источники:

* Gower, G.C., Legendre, P. Metric and euclidean properties of dissimilarity coefficients // *Journal of Classification*, 1986, 3, 5-48.
* Choi, H., Choi, S. Robust kernel Isomap // *Pattern Recognition*, 2007, 40, 853–862.

***Особые режимы***

Макрос игнорирует взвешивание массива (впрочем, он не берет в процедуру наблюдения с пропущенными и неположительными весами) и не рассчитан на расщепленность массива. Макрос слушается выбора/фильтрации наблюдений (команды FILTER, USE, SELECT IF, N OF CASES) и временных (под TEMPORARY) преобразований. Если вы используете выбор командами SELECT IF или N OF CASES не под TEMPORARY, то вы удалили часть наблюдений из исходного массива и значения столбца *CASENO\_* при матрице тогда – это номера наблюдений в том остаточном массиве. Во всех иных случаях значения *CASENO\_* – это номера наблюдений в исходном массиве, бывшие до просеивания вами наблюдений.

# МАКРОС !KO\_PROXQNT: РАЗНЫЕ МЕРЫ БЛИЗОСТИ ДЛЯ КОЛИЧЕСТВЕННЫХ ДАННЫХ

Version 2, Sep 2020 (Version 1, Feb 2018). Tested on SPSS Statistics 20, 22, 25.

!KO\_proxqnt vars= *v1 to v15* /\*Cписок числовых переменных (можно через to)

/view= CASE /\*Вычислять близости между наблюдениями(CASE, тж п/у) или

/\*между переменными (VARIABLE)

/id= /\*Опционально, при view=CASE: текстовая переменная-идентификатор наблюдений

/missing= /\*Как поступить user-missing данными: EXCLUDE (тж п/у) - исключить наблюдения списочно

/\*(как и system-missing); INCLUDE - взять user-missing как валидные данные

/normalize= NONE /\*Нормировать входящие векторы: нет (NONE, тж п/у), к сумме 1 (SUM),

/\*к сумме квадратов 1 (SS), z-стандартизация (Z), к диапазону 1 (RANGE),

/\*в диапазон 0-1 (RESCALE), sofmax (EXP), centered logratio (CLR),

/\*z-стандартизация поперечная (ZDIM)

/measure= CHORD /\*Мера близости: EUCLID, BLOCK, MINK p, CHEB, CANBERRA, CLARK, BRAYCURT,

/\*SOERGEL, KULCZ1, HEDGES, KULCZ2, PINKH i, ROBERTS, INTERSECT, HELLINGER, COS,

/\*CHORD, ARCCOS, SIN, SIMRATIO, IDENTC, HMEAN, GMEAN, BHATTACH, MORHORN,

/\*ELLENBERG, GLEASON, PANDEYA, MAHAL, MAHAL2, CHISQPR, PNDIV, SPNDIV,

/\*KLDIV, SKLDIV, KDIV, SKDIV, JDIFF, TANEJA, SIZE, SCATTER, SHAPE, ZSHAPE

/\*CATRP, MCRPA, ICC, BUTLER, BUTLER2

/matrix= *'d:\exercise\covmx.sav'* /\*Опционально, для MEASURE= MAHAL, MAHAL2, BUTLER, BUTLER2:

/\*Использовать эту матрицу (имя файла)

/noroot= /\*Не извлекать корень у мер, которые суть корень: YES или NO (тж п/у)

/abs= /\*У вычисленных близостей отменить знак: YES или NO (тж п/у)

/reverse= /\*Перевести близости: сходства - в различия, различия - в сходства:

/\*не надо (NONE, тж п/у), NEGSHIFT, ONEMINUS, LAWCOS, RECIP k (см.).

Минимум надо задать VARS, MEASURE.

Макрос вычисляет некоторые меры близости (сходства, различия/расстояния) между векторами данных, попарно, и выдает квадратную матрицу как новый безымянный рабочий массив. Предметами – векторами, между которыми вычисляется близость, могут быть наблюдения (ряды) или переменные (столбцы) входящего массива данных.

SPSS Statistics команда PROXIMITIES вычисляют некоторые основные меры близости для количественных данных (такие как евклидово расстояние, расстояние Минковского, манхэттенское, коэффициент корреляции, косинус, хи-квадрат и т.д.). Данный макрос предлагает значительно более богатый выбор мер близости.

Входящие данные должны быть количественные – мерные или счетные/долевые; двоичные данные (1 vs 0) тоже допустимы как частный случай. Во входящем массиве должно не быть переменных с именем *CASENO#$* или *STRID#$.*

Если вычислялись расстояния (различия), выходящая матрица имеет 0 на диагонали и неотрицательные внедиагональные значения. Если вычислялись сходства, на диагонали будут положительные значения (не всегда именно 1), а внедиагональные элементы могут иметь или не иметь отрицательного знака в зависимости от типа близости.

В матрице столбец *PROX* имеет первым словом ярлыка значений “DISSIMILARITY”, если близости – различия, и “SIMILARITY”, если близости – сходства. Это важно, т.к. процедуры SPSS Statistics, берущие матрицу в анализ (к примеру, кластерный), смотрят на это первое слово. Если вы захотите превратить различия в сходства или сходства в различия, не забудьте заменить и это слово в ярлыке столбца *PPOX*. Макрос имеет п/к REVERSE, переводящую различия и сходства друг в друга некоторыми популярными методами, однако методов такого перевода существует множество, и то, что предлагает REVERSE, вам может не подойти; вы должны сами решить, какой перевод оптимален для ваших данных и вашей меры близости.

Процедура вычисляет меру для всех пар векторов одновременно. Поэтому, если некоторые пары не позволяют вычислить заказанную меру, не будет получена и вся матрица. Макросом будет выдано сообщение об ошибке с указанием причины, по какой оказалось невозможно вычислить меру.

***Подкоманды***

**VARS**

Укажите списком и/или через “to” входящие количественные числовые переменные.

**VIEW**

Укажите, что является векторами, предметами, между которыми вычислять близость. Задайте VIEW=CASE, если близости нужны между наблюдениями, рядами данных (тж. по умолчанию), либо VIEW=VARIABLE, если близости нужны между переменными VARS, столбцами данных. В втором случае в выходящей матрице имена столбцов/рядов будут именами переменных VARS. Во первом случае имена столбцов/рядов будут последовательными VAR1, VAR2, VAR3…, и будет присутствовать столбец *CASENO*\_ - номер наблюдения в исходном массиве.

**MISSING**

Макрос исключает пропуски списочно: если наблюдение имеет пропущенное значение хотя бы в одной переменной VARS или идентификаторе, оно не берется в анализ. Заметьте, что на входе исключаются наблюдения, а не переменные, независимо от того, как задана п/к VIEW.

Подкоманда MISSING позволяет взять в анализ наблюдения с пользовательскими пропущенными значениями (user-missing), т.е. считать их валидными: MISSING=INCLUDE. По умолчанию же и MISSING=EXCLUDE они считаются пропусками и наблюдения исключаются, как и системные пропуски (system-missing).

**ID**

Эта подкоманда позволена только при VIEW=CASE. Вы можете указать текстовую переменную (не шире 8 байтов) на роль идентификатора наблюдений в выходящей матрице коэффициентов. По ее умолчанию/незаданию макрос создаст (для наблюдений) текстовый идентификатор с именем *ID*, значения которого состоят латинского “C” и номером наблюдения, совпадающим с *CASENO*\_.

**NORMALIZE**

Опциональная подкоманда, нормирует значения в векторах перед вычислением близостей.

NONE - (тж по умолчанию/незаданию) не делать нормирования, использовать значения как есть. Заметьте, однако, что некоторые меры (см. п/к MEASURE) сами выполняют линейное нормирование к сумме (SUM) как часть их формулы.

SUM - линейное нормирование к сумме: в каждом векторе значения делятся на сумму в векторе; сумма в векторе станет 1. Эта опция требует неотрицательных данных. Если вектор был нулевой, макрос оставляет его таким.

SS - нормирование к сумме квадратов: в каждом векторе значения делятся на корень квадратный суммы квадратов в векторе; сумма квадратов в векторе станет 1. Если вектор был нулевой, макрос оставляет его таким.

Z - z-стандартизация: в каждом векторе средняя будет 0 и стандартное отклонение 1. Если вектор был константа, он будет нулевым вектором.

RANGE - приведение к диапазону 1: в каждом векторе значения делятся на диапазон в векторе. Если вектор был константа, макрос оставляет его как есть.

RESCALE - линейное перешкалирование в диапазон [0,1]: в каждом векторе из значений вычитается минимальное значение, потом значения делятся на диапазон. Если вектор был константа, макрос делает его константой 0.5.

EXP - экспоненциальное нормирование, или softmax-преобразование. В каждом векторе значения будут лежать в диапазоне 0–1 и их сумма будет 1.

CLR - centered logratio transformation. Сначала в каждом векторе делается нормирование к сумме (SUM). В полученном векторе значения делятся на геометрическую среднюю в нем и логарифмируются; сумма в векторе станет 0. Эта опция требует положительных входящих данных.

ZDIM - z-стандартизация «поперечная», т.е. внутри измерений, а не векторов: в каждом измерении средняя будет 0 и стандартное отклонение 1. Если измерение была константа, оно будет нулевым. В отличие от Z, нормирующее стандартное отклонение тут будет вычислено на “df=n”, не “df=n-1”. (При VIEW=CASE векторы это наблюдения, а измерения это переменные данных, а при VIEW=VARIABLE – наоборот.)

**MEASURE**

Выберите одну меру близости (сходство или различие) из списка. Обозначения *x* и *y* это значения двух векторов, между которыми вычисляется близость. Т.е. это *xi* и *yi*, но индексы “*i*” опущены в формулах ради скорописи. это , где *m* это число «измерений», т.е. число элементов вектора. Под словом «измерения» подразумеваются столбцы данных, когда векторы это ряды (VIEW=CASE), и ряды данных, когда векторы это столбцы (VIEW=VARIABLE).  означает, следовательно, ; означает .

Обозначения для линейно нормированных к сумме 1 неотрицательных векторов (т.е. композиционных данных): и .

EUCLID - **евклидово расстояние** (Euclidean distance),

Заметим, что евклидово расстояние после NORMALIZE=CLR известно как **Aitchison distance**. После NORMALIZE=SS известно как **chord distance**.

BLOCK - **манхэттенское** **расстояние** (Manhattan aka **city-block distance**),

Манхэттенское расстояние после NORMALIZE=SUM, деленное на 2, известно как **Whittaker distance**.

MINK *p* - **расстояние Минковского** (Minkowski distance),

где произвольная степень *p* – число в диапазоне (0,+∞]. *p*=1 дает **манхэттенское** расстояние, *p*=2 дает евклидово расстояние.

CHEB - **расстояние Чебышёва** (Chebyshev distance aka **chessboard distance** aka **supremum distance**),

CANBERRA - **канберрово расстояние** (Canberra distance),

(если знаменатель 0, все слагаемое принимается за 0).

CLARK - **расстояние Кларка** (Clark distance aka **Clark divergence coefficient**). Это L2 версия канберрова расстояния (которое в свою очередь есть L1 версия),

(если знаменатель 0, все слагаемое принимается за 0).

BRAYCURT - **расстояние Брея–Кертиса** (Bray–Curtice distance), для неотрицательных данных,

Сходство 1-*dxy* имеет множество синонимичных названий: **Czekanowski similarity**, **Bray–Curtice similarity**; в приложении к двоичным данным – **Sorensen** или **Dice similarity**; в анализе 2x2 таблиц беспорядка – **F1 index**; в анализе долей – **Renkonen index**. Уполовиненное 1-*dxy* называется **Motyka similarity**. Само *dxy* в случае двоичных данных носит название **Lance–Williams dissimilarity**.

ELLENBERG - **сходство Ellenberg** (Ellenberg similarity), для неотрицательных данных,

где функция равна 1, когда *xy* не равно 0, иначе она равна 0.

GLEASON - **сходство Gleason** (Gleason similarity), для неотрицательных данных. Та же формула, что Ellenberg, но без “” в знаменателе.

PANDEYA - **сходство Pandeya** (Pandeya similarity), для неотрицательных данных,

где функция как в Ellenberg выше.

KULCZ1 - **различие Кульчинкого 1** (Kulczynski 1 dissimilarity), для неотрицательных данных,

1-*dxy* это **сходство Кульчинского 1**.

SOERGEL - **различие Soergel** (Soergel dissimilarity aka **Tanimoto distance** aka **Levandowsky distance** aka **Marczewski–Steinhaus coefficient**), для неотрицательных данных,

1-*dxy* называется **сходством Ruzicka**, а также сходством **Jaccard** количественных данных (другой претендент это *SIMRATIO*).

HEDGES - **различие Hedges** (Hedges wave dissimilarity), для неотрицательных данных,

KULCZ2 - **различие Кульчинкого 2** (Kulczynski 2 dissimilarity), для неотрицательных данных,

1-*dxy* это **сходство Кульчинского 2**.

PINKH *i* - **сходство Pinkham–Pearson** (Pinkham–Pearson similarity aka **fuzzy Jaccard similarity**), для неотрицательных данных,

(если знаменатель 0, все слагаемое принимается за *i*; *i* это цифра либо 0, либо 1, следующая за ключевым словом)

ROBERTS - **сходство Roberts** (Roberts similarity), для неотрицательных данных,

INTERSECT - **сходство пересечения** (intersection similarity), для неотрицательных данных,

1-*dxy* есть **расстояние пересечения**.

MAHAL - **махаланобисово расстояние** (Mahalanobis distance aka **generalized distance**) с использованием ковариационной матрицы,

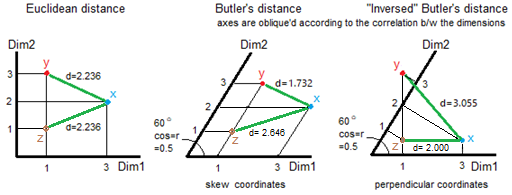
где *S* это ковариационная матрица между измерениями. Требует невырожденной *S*. Махаланобисово расстояние это евклидово расстояние, получающееся по «снятии» из данных информации о корреляции и о дисперсии (масштаба) измерений.

Чтобы ввести свою матрицу *S*, а не вычислять ее из данных – см. п/к MATRIX.

MAHAL2 - **махаланобисово расстояние** с использованием MSCP матрицы. Та же формула, что *MAHAL*, но *S* это матрица средних кросспроизведений между измерениями, т.е. это подобно ковариационной матрице, но без центрации данных при ее вычислении.

BUTLER - **батлерово расстояние** (Butler distance) с использованием корреляционной матрицы,

где *R* это корреляционная матрица между измерениями. Батлерово расстояние это евклидово расстояние в косоугольной системе координат: углы между осями задаются матрицей *R*. Входящие данные (*x* и *y*) принимаются за контравариантные (косые) координаты.



Чтобы принять входящие данные за ковариантные (перпендикулярные) координаты, добавьте кл. слово INV после слова BUTLER. Тогда будет использовано вместо в формуле ( должна быть тогда невырождена); это «инверсированное» батлерово расстояние есть махаланобисово без масшабирования.

Чтобы ввести свою матрицу *R*, а не вычислять ее из данных – см. п/к MATRIX.

BUTLER2 - **батлерово расстояние** с использованием косинусной матрицы. Та же формула, что *BUTLER*, но *R* это матрица косинусов между измерениями, она подобна корреляционной, но без центрации данных при ее вычислении. Добавьте кл. слово INV, чтобы трактовать ваши данные как ковариантные координаты.

COS - **сходство косинус** (cosine similarity aka **Tucker coefficient of congruence** aka **coefficient of proportionality** aka **Orchini similarity**),

1-*sxy* есть **косинусное расстояние**. После NORMALIZE=Z косинус это коэффициент корреляции Пирсона.

CHORD - **хордовое расстояние** (chord distance aka **Orloci distance**),

Это евклидово расстояние между NORMALIZE=SS векторами. Иначе говоря, это сходство косинус, переведенное в евклидово расстояние по теореме косинусов.

ARCCOS - косинус, переведенный в **угловое расстояние** (angular distance aka **geodesic metric**) – угол в радианах,

SIN - **различие синус** (sine dissimilarity),

IDENTC - **сходство коэффициент тождественности** (Zegers–ten Bergeidentity coefficient similarity aka **quantitative Dice similarity**),

SIMRATIO - **сходство отношение подобия** (similarity ratio aka **Kohonen similarity** aka **Kumar–Hassebrook** similarity),

Эту меру можно рассматривать как **сходство Жаккара для количественных** данных (другой претендент это 1-*SOERGEL*).

MORHORN - **сходство Morisita–Horn** (Morisita–Horn similarity), для неотрицательных данных,

HMEAN - **сходство** **гармоническая средняя** (harmonic mean similarity), для неотрицательных данных,

Используется преимущественно для векторов долей или вероятностей с единичной суммой в векторе. Тогда *sxx*=1.

GMEAN - **сходство** **геометрическая средняя** (geometric mean similarity), для неотрицательных данных,

Используется преимущественно для векторов долей или вероятностей с единичной суммой в векторе. Тогда *sxx*=1 и мера известна как **fidelity** или **Bhattacharyya coefficient** или **Hellinger affinity**. Чтобы получить **fidelity**, используйте NORMALIZE=SUM.

BHATTACH - **расстояние** **Bhattacharyya (**Bhattacharyya distance**)**, для неотрицательных данных,

Это отрицательный логарифм **fidelity**.

HELLINGER - **расстояние Хеллингера** (Hellinger distance), для неотрицательных данных,

, т.е. без множителя 2 под корнем, называется **Matusita** или **Jeffries–Matusita** **distance[[1]](#footnote-1)**.

CHISQPR - **хи-квадрат расстояние между двумя распределениями вероятностей** (chi-square distance between two probability distributions), для неотрицательных данных,

Это корень квадратный из статистики хи-квадрат для перекрестной таблицы 2 x *m* (векторы x категории), где суммы внутри векторов приведены к единице[[2]](#footnote-2). известно как **probabilistic symmetric chi-square**.

PNDIV - **пирсоновский/неймановский** **хи-квадрат несогласия вероятностей** (Pearson/Neyman chi-square of disagreement of probabilities, aka Pearson/Neyman divergence), для положительных данных,

Если это наблюдаемое распределение вероятностей и это ожидаемое (теоретическое) их распределение, тогда это корень квадратный из статистики хи-квадрат сравнения наблюдаемых вероятностей с теоретическими. Макрос выдает асимметричную матрицу, в которой элемент (r,c) соответствует случаю, когда r-й вектор находится в знаменателе.

SPNDIV - **симметричный** **пирсоновско-неймановский** **хи-квадрат несогласия вероятностей** (symmetric Pearson–Neyman chi-square of disagreement of probabilities), для положительных данных. Это . Макрос выдает симметричную матрицу.

KLDIV - **рассогласование** **Kullback–Leibler** (Kullback–Leibler divergence aka **KL-distance** aka **relative entropy** aka **information deviation** aka **information gain**), для неотрицательных данных,

(если или , все слагаемое принимается за 0). Макрос выдает асимметричную матрицу, в которой элемент (r,c) соответствует случаю, когда c-й вектор есть , r-й вектор есть .

SKLDIV - **симметричное** **рассогласование** **Kullback–Leibler** (aka **Jeffreys divergence** aka **J-distance**), для неотрицательных данных. Это , где *d* - рассогласование Kullback–Leibler выше. Макрос выдает симметричную матрицу.

KDIV - **K-рассогласование** (K-divergence), для неотрицательных данных,

(если или , все слагаемое принимается за 0). Макрос выдает асимметричную матрицу, в которой элемент (r,c) соответствует случаю, когда c-й вектор есть , r-й вектор есть .

SKDIV - **симметричное** **K-рассогласование** (aka **Topsoe distance**), для неотрицательных данных. Это , где *d* – K-рассогласование выше. Макрос выдает симметричную матрицу. Уполовиненное Topsoe distance известно как **Jensen–Shannon divergence**.

JDIFF - **разница Jensen** (Jensen difference aka **information radius**), для неотрицательных данных,

(если или , все слагаемое принимается за 0).

TANEJA - **расстояние Taneja** (Taneja distance), для неотрицательных данных,

(если или , все слагаемое принимается за 0).

SIZE - **разница** **Penrose size difference**,

Это (деленная на ) суммарная разница между векторами и выражает разницу *уровней* (поднятий) их профилей. В матрице симметричные элементы будут отличаться знаком. Элемент (r,c) соответствует случаю, когда c-й вектор есть *x*, r-й вектор есть *y*. П/к ASB=YES снимет знак с элементов.

SCATTER - **разница** **scatter difference**,

где *cx* и *cy* это центрованные векторы *x* и *y*. Это неодинаковость величины *рассеяния* в профилях. В матрице симметричные элементы будут отличаться знаком. Элемент (r,c) соответствует случаю, когда c-й вектор есть *x*, r-й вектор есть *y*. П/к ASB=YES снимет знак с элементов.

SHAPE - **расстояние** **Penrose shape distance**,

Эта величина равна квадратическому отклонению разниц *x-y* от 0, т.е. евклидовому расстоянию между центрованными векторами (*cx* и *cy*), и является мерой неодинаковости *формы* профилей *x* и *y*.

ZSHAPE - **расстояние** **standardized shape distance**,

Эта величина равна евклидовому расстоянию между z-стандартизованными векторами (*zx* и *zy*), и является мерой неодинаковости *формы* профилей *x* и *y*. Она линейно эквивалентна некоррелированности, т.е. величине .

CATRP - **сходство** **Cattell pattern similarity**, одна из мер подобия профилей,

где это евклидово расстояние между двумя векторами; *q* это медиана в распределении хи-квадрат с df=*m*, вычисляемое макросом по приближенной формуле . Сходство в норме ожидает, что значения входящих векторов z-стандартизованы внутри *измерений* (используя статистики данной выборки или внешние нормативы). Используйте NORMALIZE=ZDIM для стандартизации внутри измерений статистиками выборки.

MCRPA - **сходство** **McCrae profile agreement**, одна из мер подобия профилей,

где это евклидово расстояние между двумя векторами. Сходство в норме ожидает, что значения входящих векторов z-стандартизованы внутри *измерений* (используя статистики данной выборки или внешние нормативы). Используйте NORMALIZE=ZDIM для стандартизации внутри измерений статистиками выборки.

ICC - **сходство** **внутриклассовая корреляция** (intraclass correlation coefficient), одна из мер подобия профилей,

где есть общая средняя объединенного массива двух векторов.

При *двоичных данных* Similarity ratio = Jaccard sim = Ruzicka sim (= 1 - Soergel dis) = Ellenberg sim.

При *двоичных данных* Pinkham-Pearson (1) = Rand sim; Pinkham-Pearson (0) = Russel-Rao sim.

При *двоичных данных* Identity coef = Dice sim = Czekanowski sim (= 1 - Bray-Curtis \ Lance-Williams dis) = Gleason sim.

При *двоичных данных* Cosine sim = Ochiai sim = Fidelity.

Одна удобная, практическая **классификация** мер:

* **Меры для данных с любыми значениями**. Euclidean distance, Manhattan distance, Minkowski distance, Chebyshev distance, Canberra distance, Clark distance, Cosine similarity / Pearson correlation, Chord distance, Arc cosine distance, Sine dissimilarity, Identity coefficient similarity, Similarity ratio.
* **Меры для данных с неотрицательными значениями**. Bray–Curtice distance, Ellenberg similarity, Gleason similarity, Pandeya similarity, Kulczynski dissimilarity 1, Kulczynski dissimilarity 2, Soergel dissimilarity, Hedges wave dissimilarity, Pinkham–Pearson similarity (with parameter 0 or 1), Roberts similarity, Intersection similarity, Morisita–Horn similarity, Harmonic mean similarity, Geometric mean similarity.
* **Меры, часто используемые для векторов вероятностей** или других «композиционых данных», таких как счетные количества, с неотрицательными значениями, суммируемыми в векторе в 1 или некое Всего (макрос сначала нормирует сумму значений к 1). Bhattacharyya distance, Hellinger distance, Chi-square distance for probabilities, Pearson/Neyman chi-square divergence (asymmetric and symmetric), Kullback–Leibler asymmetric divergence (Information gain), Kullback–Leibler symmetric divergence (Jeffreys divergence), K-divergence (asymmetric), K-divergence (symmetric, Topsoe distance), Jensen difference (Information radius), Taneja distance. Harmonic mean similarity и Geometric mean similarity тоже применяются для вероятностных векторов в сочетании с NORMALIZE=SUM. Aitchison distance это евклидово расстояние на векторах, подвергнутых centered-logratio-преобразованию.
* **Меры, часто используемые в суждении подобия профилей**, сравнить “профили”. Size difference, Scatter difference, Shape distance, Standardized shape distance, Cattell pattern similarity, McCrae profile agreement, Intraclass correlation coefficient. Butler’s distance также используется в сравнении профилей.
* **Меры, рассматривающие корреляции**. Это расстояния, которые, будучи вычислены между, скажем, наблюдениями, одновременно учитывают корреляции между переменными. Mahalanobis distance, Butler’s distance. (Butler увеличивает дистанции, сонаправленные с корреляцией, и сжимает дистанции, противонаправленные с ней, Mahalanobis же искажает дистанции образом обратным этому, плюс масштабирует.)

Источники:

* Deza, M.M, Deza, E. Encyclopedia of distances. Springer, 2009.
* Cha, S.-H. Comprehensive survey on distance/similarity measures between probability density functions // *International Journal of Mathematical Models and Methods in Applied Sciences*, 2007, 4(1), 300-307.
* Podani, J. Introduction to the exploration of multivariate biological data. Backhuys Publishers, 2000. Chapter 3.
* McCrae, R.R. A Note on Some Measures of Profile Agreement // *Journal of Personality Assessment*, 2008, 90(2), 105-109.
* Rifkin, S. Comparing patterns: elevation, scatter, and shape / Paper for Executive Leadership Program course in the Graduate School of Education, George Washington University, 2006.
* Butler, J.K. A vector model for describing and comparing profiles // Educational and Psychological Measurement, 1983, 43(3), 747-758.

**MATRIX**

Эта опциональная подкоманда действует только при MEASURE= *MAHAL*, *MAHAL2*, *BUTLER* или *BUTLER2*. Вы можете ввести матрицу *S* или *R* (см. формулы в п/к MEASURE) для вычисления указанных расстояний. Будет использована введенная матрица, а не матрица, вычисляемая из данных. Укажите в кавычках или апострофах внешний .SAV-файл. Он должен представлять из себя квадратную симметричную матрицу с числом рядов и столбцов, равным числу измерений *m* данных. Имена *m* переменных файла могут быть любые. Когда вы вводите свою матрицу, а не вычисляете ее из данных, то нет разницы в результатах между *MAHAL* и *MAHAL2*, между *BUTLER* и *BUTLER2*.

**NOROOT**

NOROOT=YES выдает меру, которая суть корень, без извлечения корня. К примеру, MEASURE=*EUCLID* при NOROOT=YES выдаст квадратное евклидово расстояние, а не евклидово расстояние, потому что последнее действие в формуле евклидового расстояния – извлечение корня – будет снято. Эта подкоманда относится к следующим мерам: *EUCLID*, *CLARK*, *MAHAL*, *MAHAL2*, *BUTLER*, *BUTLER2*, *CHORD*, *SIN*, *CHISQPR*, *PNDIV*, *SPNDIV*, *SHAPE*, *ZSHAPE*, *MINK*. Прочие меры игнорируют ее.

**ABS**

Опциональная подкоманда, отменяющая отрицательный знак у вычисленных близостей (если данная мера близости может бывать отрицательной): ABS=YES.

**REVERSE**

Эта подкоманда переводит вычисленные близости: сходства – в различия, различия – в сходства. Выберите способ:

NONE - (тж по умолчанию/незаданию) не делать перевода.

NEGSHIFT - поменять знак и затем сместить в положительную сторону. Это универсальный линейный способ, он годится для любых близостей. Сходства переводятся в различия по формуле , где это наибольшее внедиагональное сходство в матрице, а на диагональ полученной матрицы различий ставится 0. Различия переводятся в сходства по формуле , где это наибольшее различие в матрице.

ONEMINUS - вычесть из единицы. Этот способ переводит сходства в различия по формуле и годится для сходств со значениями в пределах [0, 1] или [-1, 1]. Он переводит различия в сходства по формуле и годится для различий со значениями в пределах [0, 1].

LAWCOS - этот способ основан на законе косинусов. Он переводит сходства в различия по формуле и годится для сходств со значениями в пределах [0, 1] или [-1, 1]. Он переводит различия в сходства по формуле и годится для различий со значениями в пределах [0, 1].

RECIP *k* - через обращение. Этот способ переводит сходства в различия по формуле и годится для сходств со значениями в пределах (0, 1]. Он переводит различия в сходства по формуле и годится для любых различий. *k* – положительное число. По умолчанию числа *k* оно принимается за 1.

Когда ABS=YES, то REVERSE применяется после ABS.

***Особые режимы***

Макрос игнорирует взвешивание массива (впрочем, он не берет в процедуру наблюдения с пропущенными и неположительными весами) и не рассчитан на расщепленность массива. Макрос слушается выбора/фильтрации наблюдений (команды FILTER, USE, SELECT IF, N OF CASES) и временных (под TEMPORARY) преобразований. Если вы используете выбор командами SELECT IF или N OF CASES не под TEMPORARY, то вы удалили часть наблюдений из исходного массива и значения столбца CASENO\_ при матрице тогда – это номера наблюдений в том остаточном массиве. Во всех иных случаях значения CASENO\_ – это номера наблюдений в исходном массиве, бывшие до просеивания вами наблюдений.

# МАКРОС !KO\_PROXBIN: РАЗНЫЕ МЕРЫ БЛИЗОСТИ ДЛЯ ДВОИЧНЫХ ДАННЫХ

Version 1, Oct 2020. Tested on SPSS Statistics 20, 22, 25.

!KO\_proxbin vars= *v1 to v15* /\*Cписок числовых двоичных переменных (можно через to)

/view= CASE /\*Вычислять расстояния между наблюдениями(CASE, тж п/у) или

/\*между переменными (VARIABLE)

/id= /\*Опционально, при view=CASE: текстовая переменная-идентификатор наблюдений

/measure= SM /\*Мера близости: COUNT\_A, COUNT\_B, COUNT\_C, COUNT\_D,

/\*ECOUNT\_A, ECOUNT\_B, ECOUNT\_C, ECOUNT\_D,

/\*RR, SM, HAMANN, JACCARD, DICE, LW, RT, SS1, SS2, K1, SS3,

/\*K2, MCCON, OCHIAI, SS4, SS5, GOWER, CHISQ, CHISQY, CC, PHI, PRPHI, DISPER,

/\*SEUCLID, EUCLID, SIZE, SHAPE, PATTERN, VARIANCE, MN, TETRACH, LAMBDA,

/\*D, Y, Q, DIGBY, GK1, SIMPSON, BRAUN, FAI1, FAI2, BUB1, BUB2, MOUNTF, QDIFF,

/\*FCOS, CHORD, HELLINGER, MICHAEL, HD, FORBES, ALROY, PMI, DENNIS,

/\*TARWID, KAPPA, SPI, LOEV, COLE, MAXWELL, FLEISS, EYRAUD, GILBERT

/adjust= /\*Для сходств: перешкалировать вычисленные значения меры:

/\*не надо (NONE, тж п/у), INDBOUND, INDUPPER, MARGPOLE,

/\*INDMARGPOLE, INDMARGMAX (см.)

/abs= /\*У вычисленных близостей отменить знак: YES или NO (тж п/у)

/reverse= /\*Перевести близости: сходства - в различия, различия - в сходства:

/\*не надо (NONE, тж п/у), NEGSHIFT, ONEMINUS, LAWCOS, RECIP k (см.).

Минимум надо задать VARS, MEASURE.

Макрос вычисляет некоторые меры близости (сходства – в том числе коэффициенты сочетанности, различия/расстояния) между векторами двоичных данных, попарно, и выдает квадратную матрицу как новый безымянный рабочий массив. Предметами – векторами, между которыми вычисляется близость, могут быть наблюдения (ряды) или переменные (столбцы) входящего массива данных.

Во входящем массиве должно не быть переменных с именем *CASENO#$* или *STRID#$.*

Если вычислялись расстояния (различия), выходящая матрица имеет 0 на диагонали и неотрицательные внедиагональные значения. Если вычислялись сходства, на диагонали будут положительные значения[[3]](#footnote-3) (не всегда именно 1), а внедиагональные элементы могут иметь или не иметь отрицательного знака в зависимости от типа близости.

В матрице столбец *PROX* имеет первым словом ярлыка значений “DISSIMILARITY”, если близости – различия, и “SIMILARITY”, если близости – сходства. Это важно, т.к. процедуры SPSS Statistics, берущие матрицу в анализ (к примеру, кластерный), смотрят на это первое слово. Если вы захотите превратить различия в сходства или сходства в различия, не забудьте заменить и это слово в ярлыке столбца *PPOX*. Макрос имеет п/к REVERSE, переводящую различия и сходства друг в друга некоторыми популярными методами, однако методов такого перевода существует множество, и то, что предлагает REVERSE, вам может не подойти; вы должны сами решить, какой перевод оптимален для ваших данных и вашей меры близости.

Процедура вычисляет меру для всех пар векторов одновременно. Поэтому, если некоторые пары не позволяют вычислить заказанную меру, не будет получена и вся матрица. Макросом будет выдано сообщение об ошибке с указанием причины, по какой оказалось невозможно вычислить меру.

***Подкоманды***

**VARS**

Укажите списком и/или через “to” входящие числовые двоичные переменные, содержащие только два валидных значения: 1 и 0. Если ваши данные не дихотомические или имеют не двоичную кодировку, перекодируйте их сначала в двоичные командой RECODE, что можно сделать и временно (под TEMPORARY). Если ваши двоичные данные представляют двоичный/булев статус некоторого признака, тогда значение 1 должно отвечасть статусу “есть” или “истина”, тогда как 0 отвечать статусу “нету” или “ложь”.

Макрос исключает любые пропуски списочно: если наблюдение имеет пропущенное значение хотя бы в одной переменной VARS или идентификаторе, оно не берется в анализ. Заметьте, что на входе исключаются наблюдения, а не переменные, независимо от того, как задана п/к VIEW.

**VIEW**

Укажите, что является векторами, предметами, между которыми вычислять близость. Задайте VIEW=CASE, если близости нужны между наблюдениями, рядами данных (тж. по умолчанию), либо VIEW=VARIABLE, если близости нужны между переменными VARS, столбцами данных. В втором случае в выходящей матрице имена столбцов/рядов будут именами переменных VARS. Во первом случае имена столбцов/рядов будут последовательными VAR1, VAR2, VAR3…, и будет присутствовать столбец *CASENO*\_ - номер наблюдения в исходном массиве.

**ID**

Эта подкоманда позволена только при VIEW=CASE. Вы можете указать текстовую переменную (не шире 8 символов) на роль идентификатора наблюдений в выходящей матрице коэффициентов. По ее умолчанию/незаданию макрос создаст (для наблюдений) текстовый идентификатор с именем *ID*, значения которого состоят латинского “C” и номером наблюдения, совпадающим с *CASENO*\_.

**MEASURE**

Выберите одну меру близости (сходство или различие) из списка. Пусть два сравниваемых вектора – X и Y, оба длиной *m* – это число «измерений» (если X и Y это ряды данных, то измерения это столбцы данных, а если X и Y это столбцы данных, то измерения это ряды данных). Тогда обозначения в формулах следующие:

*a* число измерений, где X=1 и Y=1

*b* число измерений, где X=1 и Y=0

*c* число измерений, где X=0 и Y=1

*d* число измерений, где X=0 и Y=0

*a+b+c+d=m*

Т.е. имеем частотную таблицу:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Вектор Y | |
|  |  | 1 | 0 |
| Вектор X | 1 | *a* | *b* |
| 0 | *c* | *d* |

Табличку этих четырех частот макрос выдает в окно результатов, если число векторов в анализе равно двум.

Меры ниже, где диапазон при формуле не указан, колеблются в диапазоне [0,1]. На диагонали матрицы мер *RR*, *GOWER*, *CHISQY*, *DISPER*, *D*, *FAI2*, *FAI1*, *FCOS*, *MICHAEL*, *FORBES*, *PMI*, *TARWID*, *DENNIS*, *COLE*, *EYRAUD* стоит не константа, потому что у этих мер сходства величина *sii* зависит от формы распределения в векторе *i*. Причем у некоторых из этих мер на диагонали может встретиться 0.

*Замечание* о мерах Sokal–Sneath 1, 2, 3, 4, 5 (т.е. *SS1, SS2, SS3, SS4, SS5*): в разных литературных источниках существует путаница, какой номер какой формуле принадлежит, хотя формулы во всех источниках одни и те же. В данном макросе привязка формулы к номеру следует тому, как это было сделано в документации SPSS Statistics (команды PROXIMITIES, CLUSTER). Пользователю рекомендуется смотреть на формулу, а не на номер.

RR - **сходство Рассела–Рао** (Russel–Rao similarity aka **simple joint probability**),

SM - **сходство Рэнда** (Rand similarity aka **simple matching** aka **Sokal–Michener similarity**),

HAMANN - **сходство Hamann** (Hamann similarity), является пересчетом *SM* с диапазона [0, 1] на [-1, 1],

JACCARD - **сходство Жаккара** (Jaccard similarity aka **Tanimoto similarity** aka **community coefficient** aka **similarity ratio**),

1-*JACCARD* называется **Soergel distance**.

DICE - **сходство Дайса** (Dice similarity aka **Czekanowski similarity** aka **Sorenson similarity** aka **Gleason similarity** aka **identity coefficient** aka **coincidence index**),

Это среднегармоническая двух условных вероятностей: и (в задачах классификации известных как “Recall” и “Precision”).

LW - **различие Lance­–Williams** (Lance­–Williams dissimilarity),

Это **Bray–Curtis distance** в ситуации двоичных данных.

RT - **сходство** **Rogers–Tanimoto** (Rogers–Tanimoto similarity),

SS1 - **сходство** **Sokal–Sneath 1** (Sokal–Sneath 1 similarity aka **Gower–Legendre similarity**),

SS2 - **сходство Sokal–Sneath 2** (Sokal–Sneath 2 similarity),

K1 - **сходство Кульчинского 1** (Kulczynski 1 similarity). Колеблется в диапазоне [0, *m*-1]. Значения *sii* невычислимы и на диагональ результата ставится условное 999.

SS3 - **сходство Sokal–Sneath 3** (Sokal–Sneath 3 similarity). Колеблется в диапазоне [0, *m*-1]. Значения *sii* невычислимы и на диагональ результата ставится условное 999.

K2 - **сходство Кульчинского 2** (Kulczynski 2 similarity),

Это среднеарифметическая двух условных вероятностей: и (в задачах классификации известных как “Recall” и “Precision”).

MCCON - **сходство McConnaughey** (McConnaughey similarity), является пересчетом *K2* с диапазона [0, 1] на [-1, 1],

OCHIAI - **сходство Ochiai** (Ochiai similarity aka **cosine similarity** aka **Otsuka similarity**),

Это среднегеометрическая двух условных вероятностей: и (в задачах классификации известных как “Recall” и “Precision”). *OCHIAI*2 известно как **Sorgenfrei similarity** aka **correlation ratio**.

SS4 - **сходство Sokal–Sneath 4** (Sokal–Sneath 4 similarity aka **Anderberg 2 similarity**),

Это усреднение двух сходств Kulczynski 2 (*KK2*): одно соответствует кодировке данных “1=есть, 0=нету” (как обычно), а другое соответствует обратной кодировке.

SS5 - **сходство Sokal–Sneath 5** (Sokal–Sneath 5 similarity aka **Anderberg 1 similarity** aka **Ochiai 2 similarity**),

Это произведение двух косинусов (*OCHIAI*): один соответствует кодировке данных “1=есть, 0=нету” (как обычно), а другой соответствует обратной кодировке.

GOWER - **сходство binary Gower** (binary Gower similarity). Колеблется в диапазоне [0, *m*/(*m*-1)].

CHISQ - **сходство хи-квадрат Пирсона** (Pearson chi-square similarity), статистика хи-квадрат 2×2 таблицы. Колеблется в диапазоне [0, *m*].

CHISQY - **сходство хи-квадрат Пирсона с поправкой Йейтса** (Yates corrected Pearson chi-square similarity), статистика хи-квадрат 2×2 таблицы. Колеблется в диапазоне [0, *m*).

Логарифм *CHISQY* известен как **сходство Stiles**.

CC - **сходство коэффициент сопряженности** (Contingency coefficient similarity), колеблется (для таблицы 2×2) в диапазоне [0, √0.5].

PHI - **сходство фи-корреляция** (Phi correlation similarity), является коэффициентом **корреляции Пирсона** в ситуации двоичных данных. Колеблется в диапазоне [-1, 1],

где есть , если , иначе есть . *PHI*2 известно также как **Doolittle similarity**. *PHI* в ситуации двоичных данных совпадает с **V Крамера**, **Rho Спирмена**, **Tau-b Кендалла**.

PRPHI - **перешкалированная относительно полюса фи-корреляция** (pole-rescaled Phi correlation aka **Cole’s C7 similarity**), колеблется в диапазоне [-1, 1],

где это максимальная по абсолютной величине положительная (если *PHI* положительно) или отрицательная (если *PHI* отрицательно) фи-корреляция, достижимая в данной 2×2 таблице с фиксированными краевыми частотами; это величина хи-квадрат, соответствующая , и то же, что в *PHI*.

и , при

и , при (см. Warrens, 2008).

*PRPHI* это то же, что *PHI* с подкомандой ADJUST=MARGPOLE.

DISPER - **сходство dispersion** (dispersion similarity), колеблется в диапазоне [-0.25, 0.25].

Это (смещенная) **ковариация** в ситуации двоичных данных.

SEUCLID - **квадратное евклидово расстояние** (squared Euclidean distance aka **Hamming distance**), колеблется в диапазоне [0, ∞],

Для двоичных данных оно совпадает с **манхэттенским расстоянием** и **канберровым расстоянием**. 2*SEUCLID* называется **Mirkin distance**.

EUCLID - **евклидово расстояние** (Euclidean distance), колеблется в диапазоне [0, ∞],

SIZE - **различие** **size difference** (s**ize difference** dissimilarity aka **Baulieu dissimilarity**),

Это квадрат Penroze size difference, меры сравнения профилей.

SHAPE - **различие** **shape difference** (**shape difference** dissimilarity),

Это квадрат Penroze shape difference, меры сравнения профилей.

PATTERN - **различие** **pattern** (**pattern difference** dissimilarity),

2√*PATTERN* называется **Sneath dissimilarity**.

VARIANCE - **различие** **variance** (**variance** dissimilarity),

MN - **расстояние** **McNemar** (McNemar distance), является корнем квадратным статистики теста Макнемара. Колеблется в диапазоне [0, ∞].

TETRACH - **сходство тетрахорическая корреляция, пирсоновская аппроксимация** (tetrachoric correlation, pearsonian approximation), колеблется в диапазоне [-1, 1],

где есть , если , иначе есть . “Min” имеет целью сделать коэффициент симметричным при перекодировке данных 1 в 0 и 0 в 1, – что и ожидается от корреляции.

LAMBDA - **сходство лямбда Гудмана–Краскела** (Goodman–Kruskal lambda similarity),

D - **сходство Anderberg’s D** (Anderberg’s D similarity), колеблется в диапазоне [0, 0.5].

где и как в *LAMBDA.*

Q - **сходство Q Юла** (Yule’s Q similarity aka **Yule’s coefficient of association**), колеблется в диапазоне [-1, 1],

где *OR* это отношение шансов .

Это **гамма** **Гудмана–Краскела** для случая двоичных данных.

Y - **сходство Y Юла** (Yule’s Y similarity aka **Yule’s coefficient of colligation**), колеблется в диапазоне [-1, 1],

где *OR* как в *Q*.

DIGBY - **сходство Digby** (Digby similarity) является «компромиссом» между *Q* и *Y*, колеблется в диапазоне [-1, 1],

Этот коэффициент может служить аппроксимацией тетрахорической корреляции, альтернативной пирсоновской аппроксимации *TETRACH*.

GK1 - **сходство Goodman–Kruskal 1** (Goodman–Kruskal 1 similarity), колеблется в диапазоне [-1, 1],

SIMPSON - **сходство Simpson** (Simpson similarity aka **overlap similarity**),

BRAUN - **сходство Braun**–**Blanquet** (Braun–Blanquet similarity aka **Savage similarity**),

(*BRAUN + SIMPSON*)/2 *= K2*.

FAI2 - **сходство Faith 2** (Faith 2 similarity),

FAI1 - **сходство Faith 1** (Faith 1 similarity), является пересчетом *FAI2* с диапазона [0, 1] на [-1, 1],

BUB2 - **сходство Baroni-Urbani & Buser 2** (Baroni-Urbani & Buser 2 similarity),

Это модификация *SM* (Rand), где *d* заменено на . С приближением к нулю мера приближается к *JACCARD*.

BUB1 - **сходство Baroni-Urbani & Buser 1** (Baroni-Urbani & Buser 1 similarity), является пересчетом *BUB2* с диапазона [0, 1] на [-1, 1],

Это модификация *HAMANN*, где *d* заменено на .

MOUNTF - **сходство Mountford** (Mountford similarity). Значения *sii* невычислимы и на диагональ результата ставится условное 999.

Эта мера не вычисляется, когда *b* или *c* есть 0.

QDIFF - **различие Q0 difference** (Q0 difference dissimilarity), колеблется в диапазоне [0, ∞],

FCOS - **подправленный косинус Fager** (Fager’s corrected cosine). Это косинус (*OCHIAI*), уменьшенный на вычитаемую величину:

Если значение меньше 0, оно зануляется.

CHORD - **хордовое расстояние** (chord distance), колеблется в диапазоне [0, √2],

Это косинус, переведенный в соответствующее евклидово расстояние.

HELLINGER - **расстояние Хеллингера** (Hellinger distance), колеблется в диапазоне [0, 2],

MICHAEL - **сходство Michael** (Michael similarity), колеблется в диапазоне [-1, 1],

HD - **сходство Hawkins–Dotson** (Hawkins–Dotson similarity),

Это усреднение двух сходств Жаккара (*JACCARD*): одно соответствует кодировке данных “1=есть, 0=нету” (как обычно), а другое соответствует обратной кодировке.

FORBES - **сходство Forbes** (Forbes similarity, aka Mozley similarity [ошибочная атрибуция]), колеблется в диапазоне [0, *m*],

где .

PMI - точечное **взаимоинформационное** **сходство** **(**point aka binary mutual information aka **Gilbert–Wells similarity**), колеблется в диапазоне [-∞, ln *m*],

где и *.*

ALROY - **сходство Alroy** (Alroy similarity aka **modified Forbes similarity**),

где .

TARWID - **сходство Tarwid** (Tarwid similarity), колеблется в диапазоне [-1, (*m*-1)/(*m*+1)],

где как в *FORBES*.

DENNIS - **сходство Dennis** (Dennis similarity), колеблется в диапазоне [-*m*/(2√*m*), (*m*-1)/√*m*],

KAPPA - **сходство каппа Коэна** (Cohen’s kappa similarity), для двоичных данных колеблется в диапазоне [-1, 1],

Эта мера известна также как **Adjusted Rand** similarity и совпадает с *SM* под ADJUST= INDMARGMAX.

SPI - **сходство Scott’s Pi** (Scott’s Pi similarity), для двоичных данных колеблется в диапазоне [-1, 1],

LOEV - **сходство Loevinger** (Loevinger similarity aka **Benini similarity**), колеблется в диапазоне [-(*m*-1), 1],

COLE - **сходство Cole** (Cole similarity), колеблется в диапазоне [-1, *m*-1].

MAXWELL - **сходство Maxwell–Pilliner** (Maxwell–Pilliner similarity), колеблется в диапазоне [-1, 1],

FLEISS - **сходство Fleiss** (Fleiss similarity [не путать с каппой Флейсса]), колеблется в диапазоне [-((*m*-1)/2+1/(*m*-1)), 1],

EYRAUD - **сходство Eyraud** (Eyraud similarity), колеблется в диапазоне [-*m*2/(*m*-1),*m*2/(*m*-1)].

где как в *FORBES* и .

GILBERT - **сходство** **Gilbert** (Gilbert similarity), колеблется в диапазоне [-1, 1].

где как в *FORBES*.

Подкоманда MEASURE может выдать также *наблюдаемые* частоты *a*, *b*, *c* или *d*, или *ожидаемые* (под гипотезой независимости) частоты *E(a)*, *E(b)*, *E(c)* или *E(d)*. Ожидаемая частота есть произведение соответствующих краевых частот, деленное на полную частоту; например, .

COUNT\_A - матрица частот *a*.

COUNT\_B - матрица частот *b*. Это асимметричная матрица.

COUNT\_C - матрица частот *c*. Это асимметричная матрица, совпадающая с транспонатом *COUNT\_B*.

COUNT\_D - матрица частот *d*.

ECOUNT\_A - матрица ожидаемых частот *E(a)*.

ECOUNT\_B - матрица ожидаемых частот *E(b)*. Это асимметричная матрица.

ECOUNT\_C - матрица ожидаемых частот *E(c)*. Это асимметричная матрица, совпадающая с транспонатом *ECOUNT\_B*.

ECOUNT\_D - матрица ожидаемых частот *E(d)*.

Подкоманды ADJUST, ABS и REVERSE игнорируются при заказе наблюдаемых или ожидаемых частот.

Источники:

* Deza, M.M, Deza, E. Encyclopedia of distances. Springer, 2009.
* Choi, S.-S., Cha, S.-H., Tappert, C.C. A Survey of binary similarity and distance measures // *Systemics, cybernatics and informatics*, 2010, 8(1), 43-48.
* Podani, J. Introduction to the exploration of multivariate biological data. Backhuys Publishers, 2000. Chapter 3.
* Romesburg, H.C. Cluster analysis for researchers. Lifetime Learning Publications, 1984.
* Alroy, J. A new twist on a very old binary similarity coefficient // *Ecology*, 2015, 96(2), 575–586.
* Jeff. simbin(mat1,mat2,type,mask) // MATLAB Central File Exchange (https://www.mathworks.com/matlabcentral/fileexchange/55190-simbin-mat1-mat2-type-mask), Retrieved October 18, 2020.
* Warrens, M.J. On association coefficients for 2×2 tables and properties that do not depend on the marginal distributions // *Psychometrika*, 2008, 73 (4), 777–789.
* Warrens, M.J. On the indeterminacy of resemblance measures for binary (presence/absence) data // *Journal of classification*, 2008, 25, 125-136.
* Goodman, L.A., Kruskal, W.H. Measures of Association for Cross Classifications. II: Further Discussion and References // *Journal of the American Statistical Association*, 1959, 54(285), 123-163.
* Eidenberger, H. Handbook of multimedia information retrieval. Atpress, 2012.
* Hubalek, Z. Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation // *Biol. Rev*., 1982, 57, 669-689.

**О классификации двоичных мер сходства**. Многие или все меры сходства для двоичных данных часто называются коэффициентами сочетанности или совпадения (matching coefficients). Двумя фундаментальными классами их являются меры **порядковые** и **номинальные**; они различаются отношением к частоте “negative matches” *d*, т.е. частоте сочетания «нет признака – нет признака», или “0-0”.

* **«Порядковые»** или **«однополюсные»** сходства считают только частоту *a* (“1-1”), но не частоту *d* (“0-0”) основанием похожести двух векторов. Частота *d* для этих мер не говорит ни о сходстве, ни о несходстве, поэтому она отсутствует в их числителе. Это значит, что данные меры относятся к значениям 1 vs 0 *асимметрично*: 1 («есть») это «больше», чем 0 («нету»), – т.е. эти меры трактуют дихотомические данные как порядковые.
* **«Номинальные»** или **«двуполюсные»** сходства считают и *a* (“1-1”) и *d* (“0-0”) основанием похожести двух векторов. Если два объекта оба не имеют некоторого признака, то это тоже довод за их сходство, такой же, как если бы оба имели этот признак. Частота *d* поэтому присутствует в числителе меры наряду с частотой *a*. Следовательно, такие меры относятся к значениям 1 vs 0 *симметрично*: 1 ( «есть») и 0 («нету») это две равноправные категории, – т.е. эти меры трактуют дихотомические данные как номинальные. Однако некоторые из номинальных мер признают, что математически 1>0, и поэтому могут принимать не только положительные, но и отрицательные значения (наподобие корреляции).

К **порядковым** сходствам относятся: *RR*, *JACCARD*, *DICE*, *SS2*, *K1*, *K2*, *MCCON*, *OCHIAI*, *SIMPSON*, *BRAUN*, *MOUNTF*, *FCOS*, *FORBES*, *PMI*, *ALROY*. К **номинальным** сходствам относятся: *SM*, *HAMANN*, *RT*, *SS1*, *SS3*, *SS4*, *SS5*, *GOWER*, *CHISQ*, *CHISQY*, *CC*, *PHI*, *PRPHI*, *DISPER*, *TETRACH*, *LAMBDA*, *D*, *Q*, *Y*, *DIGBY*, *GK1*, *MICHAEL*, *HD*, *TARWID*, *DENNIS*, *KAPPA*, *SPI*, *LOEV*, *COLE*, *MAXWELL*, *FLEISS*, *EYRAUD*, *GILBERT*. Меры *FAI2*, *FAI1*, *BUB2*, *BUB1* являются промежуточными между порядковыми и номинальными, потому что они признают *d* основанием близости в меньшей степени, чем признают *a*.

Сходства *K2*, *MCCON*, *OCHIAI*, *DICE*, *SIMPSON*, *BRAUN*, *FCOS*, а также *SS4*, *SS5* можно назвать **мерами на условных вероятностях** (conditional probabilities measures), потому что их смысл – это доли и (в задачах классификации известных как “Recall” и “Precision”), а также иногда и . Ряд сходств тесно связан с **хи-квадрат статистикой и корреляцией**: *CHISQ*, *CHISQY*, *CC*, *PHI*, *PRPHI*, *TETRACH*, *DISPER*. Другой ряд связан с идеей **предсказываемости и отношения шансов** (predictability and odds ratio): *LAMBDA*, *D*, *Q*, *Y*, *DIGBY*.

**Отношения между двоичными мерами**. Линейно эквивалентны (r Пирсона = 1) меры внутри следующих классов:

* *SM*, *SEUCLID*, *HAMANN*, *VARIANCE*
* *DICE*, *LW*
* *K2*, *MCCON*
* *CHORD*, *HELLINGER*
* *FAI2*, *FAI1*
* *BUB2*, *BUB1*
* *PRPHI*, *LOEV* (только в области неотрицательных значений)

Меры *TETRACH*, *DIGBY* тесно коррелируют (r Пирсона > 0.99). Меры *SS4*, *PHI*, *MAXWELL* – тоже.

Монотонно эквивалентны (rho Спирмена = 1) меры внутри следующих классов:

* *SM*, *RT*, *SS1*, *EUCLID*
* *JACCARD*, *DICE*, *SS2*
* *OCHIAI*, *CHORD*
* *CHISQ*, *CC*
* *FORBES*, *PMI*, *TARWID*

Меры *TETRACH*, *Y*, *Q* тесно коррелируют (rho Спирмена > 0.99).

**ADJUST**

По умолчанию и ADJUST=NONE макрос выдает наблюдаемое в данных значение меры. П/к ADJUST позволяет заказать приведенное (скорректированное) значение меры сходства. Эта подкоманда только для сходств; она не применима к различиям. Выберите:

INDBOUND - поправка на базовую независимость (correction for baseline independence), она же поправка на связь за счет случайности (correction for association due to chance). Наблюдаемое значение меры перешкалируется так, что оно станет равно 0, если находится на уровне значения, ожидаемого в условиях отсутствия связи между векторами, т.е. на уровне, характерном для случайных неупорядоченных данных.

INDUPPER - другой вариант поправки на базовую независимость.

MARGPOLE - поправка на предельное значение, достижимое при данных краевых распределениях (correction for the pole attainable under the given marginal distributions). Наблюдаемое значение меры перешкалируется так, что оно станет равно 1, если оно совпадает с максимальным положительным значением, достижимым мерой при данных краевых частотах таблички 2×2; и оно станет равно -1, если оно совпадает с минимальным отрицательным значением, достижимым мерой в этих условиях.

INDMARGPOLE - поправка смешанная: одновременно и на базовую независимость, и на предельное значение, достижимое при данных краевых распределениях.

INDMARGMAX - другой вариант подобной смешанной поправки.

MARGPOLE, INDMARGPOLE и INDMARGMAX недоступны для меры *MOUNTF*. П/к ADJUST игнорируется, если MEASURE это частоты (*COUNT*\_... или *ECOUNT*\_...).

Все пять способов приведения являются перешкалированием наблюдаемого значения сходства по единой формуле:

где это наблюдаемое значение сходства, а остальные элементы определены ниже. Если числитель формулы положительный, то , а если отрицательный, то .

* При INDBOUND (correction for baseline independence):

есть значение сходства, посчитанное из таблички с ожидаемыми при независимости рядов и столбцов частотами

|  |  |
| --- | --- |
|  |  |
|  |  |

= верхняя граница генерального диапазона колебания сходства, если , и = нижняя граница генерального диапазона колебания сходства, если . Границы эти для каждого сходства см. в п/к MEASURE. *Примечание*: для меры *PMI* нижняя граница (которая есть -∞) принята за -10.

* При INDUPPER (correction for baseline independence):

как при INDBOUND

= верхняя граница генерального диапазона колебания сходства.

* При MARGPOLE (correction for the attainable pole):

= 0 (поэтому знак совпадёт со знаком )

, если , и , если .

есть верхняя, и есть нижняя граница колебания сходства *при данных* краевых распределениях, какие есть у двух сопоставляемых векторов, т.е. при фиксированных краевых частотах таблички 2×2. А именно:

есть значение сходства, вычисляемое из таблички, соответствующей со-сортированности векторов (т.е. соответствующее их сильнейшей позитивной связи):

|  |  |
| --- | --- |
|  |  |
|  |  |

есть значение сходства, вычисляемое из таблички, соответствующей противо-сортированности векторов (т.е. соответствующее их сильнейшей негативной связи):

|  |  |
| --- | --- |
|  |  |
|  |  |

* При INDMARGPOLE (двойная коррекция):

как при INDBOUND

как при MARGPOLE.

* При INDMARGMAX (двойная коррекция):

как при INDBOUND

.

Таблички вышеозначенных частот (*a’, b’, c’, d’* и т.д.) макрос выдает в окно результатов, если число векторов в анализе равно двум.

**ABS**

Опциональная подкоманда, отменяющая отрицательный знак у вычисленных близостей (если данная мера близости может бывать отрицательной): ABS=YES. Если задано ADJUST, то ABS применяется после нее. П/к ABS игнорируется, если MEASURE это частоты (*COUNT*\_... или *ECOUNT*\_...).

**REVERSE**

Эта подкоманда переводит вычисленные близости: сходства – в различия, различия – в сходства. Выберите способ:

NONE - (тж по умолчанию/незаданию) не делать перевода.

NEGSHIFT - поменять знак и затем сместить в положительную сторону. Это универсальный линейный способ, он годится для любых близостей. Сходства переводятся в различия по формуле , где это наибольшее внедиагональное сходство в матрице, а на диагональ полученной матрицы различий ставится 0. Различия переводятся в сходства по формуле , где это наибольшее различие в матрице.

ONEMINUS - вычесть из единицы. Этот способ переводит сходства в различия по формуле и годится для сходств со значениями в пределах [0, 1] или [-1, 1]. Он переводит различия в сходства по формуле и годится для различий со значениями в пределах [0, 1].

LAWCOS - этот способ основан на законе косинусов. Он переводит сходства в различия по формуле и годится для сходств со значениями в пределах [0, 1] или [-1, 1]. Он переводит различия в сходства по формуле и годится для различий со значениями в пределах [0, 1].

RECIP *k* - через обращение. Этот способ переводит сходства в различия по формуле и годится для сходств со значениями в пределах (0, 1]. Он переводит различия в сходства по формуле и годится для любых различий. *k* – положительное число. По умолчанию числа *k* оно принимается за 1.

REVERSE применяется после ADJUST и ABS. П/к REVERSE игнорируется, если MEASURE это частоты (*COUNT*\_... или *ECOUNT*\_...).

***Особые режимы***

Макрос игнорирует взвешивание массива (впрочем, он не берет в процедуру наблюдения с пропущенными и неположительными весами) и не рассчитан на расщепленность массива. Макрос слушается выбора/фильтрации наблюдений (команды FILTER, USE, SELECT IF, N OF CASES) и временных (под TEMPORARY) преобразований. Если вы используете выбор командами SELECT IF или N OF CASES не под TEMPORARY, то вы удалили часть наблюдений из исходного массива и значения столбца CASENO\_ при матрице тогда – это номера наблюдений в том остаточном массиве. Во всех иных случаях значения CASENO\_ – это номера наблюдений в исходном массиве, бывшие до просеивания вами наблюдений.

# МАКРОС !KO\_TETRACH: ТЕТРАХОРИЧЕСКИЙ КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

Version 1, Dec 2004. Tested on SPSS 11, 11.5, 13, 14.

!KO\_tetrach vars= v1 to v4 /\*Двоичные (0 и 1) переменные, между которыми вычислять корреляции, список поименно и/или ч-з to

/missing= PAIRWISE /\*Исключать пропущенные значения списочно (LISTWISE) или попарно (PAIRWISE, тж п/у).

Минимум надо задать VARS.

Макрос вычисляет и выдает как новый рабочий файл матрицу тетрахорических коэффициентов корреляции между входящими двоичными переменными. Тетрахорический коэффициент корреляции предназначен для таких дихотомических переменных, за которыми скрываются величины мерные, т.е. количественные и непрерывные (для которых обычно подходит интервальная или отношенческая шкала), причем имеющие в популяции нормальное распределение; дихотомичность же их представляется следствием огрубления измерительной шкалы до 2-х градаций. При таком (довольно смелом) допущении тетрахорический коэффициент есть предположительное значение коэффициента корреляции Пирсона, которое было бы, останься переменные мерными, не огрубленными до дихотомических.

Вычисление идет по «формуле косинуса», предложенной Пирсоном (Wherry R. J. Contributions to correlational analysis., 1984), дающей хорошее приближение для переменных не с очень перекошенным распределением (не более 90-95% наблюдений в одной из двух категорий). Макрос не вычисляет значимостей коэффициентов.

R = cos [ pi\*sqrt(BC) / [sqrt(AD)+sqrt(BC)] ]

A, B, C, D образуют частотную таблицу:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | VarY | |
|  |  | 0 | 1 |
| VarX | 0 | A | B |
| 1 | C | D |

***Подкоманды***

**VARS**

Укажите, списком и/или через “to”, входящие переменные. Переменные должны быть двоичной (0 и 1) кодировки. Если ваши дихотомические переменные иной кодировки, переделайте их в двоичные.

**MISSING**

Если есть пропущенные данные, то как исключать их:

PAIRWISE – исключить попарно, т.е. из каждой пары переменных отдельно (тж. по умолчинию или незаданию подкоманды).

LISTWISE – исключить списочно, т.е. из всех переменных, если хотя бы в одной наблюдение есть пропуск.

***Особые режимы***

Макрос игнорирует взвешивание файла и не рассчитан на расщепленность файла.

# МАКРОС !KO\_BISER: БИСЕРИАЛЬНЫЙ КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

Version 1, Aug 2005. Tested on SPSS 11, 11.5, 13, 14.

!KO\_biser scale= v1 to v5 /\*Мерные переменные, списочно и/или ч-з to

/binar= b1 to b3 /\*Двоичные (0 и 1) переменные, полагаемые как исходно

/\*нормально распределенные, списочно и/или ч-з to

/missing= /\*Исключать пропущенные значения списочно (LISTWISE)

/\*или попарно (PAIRWISE, тж п/у)

/alpha= /\*Опционально: выдать только коэф-ты, значимые на этом 2-стороннем уровне

/\*(незначимые занулить); п/у выдаются все коэф-ты и их значимости.

Минимум надо задать SCALE, BINAR.

Макрос вычисляет и выдает в окно результатов и в новый безымянный массив данных бисериальные коэффициенты корреляции между мерными и двоичными переменными, и также в окно результатов - их значимости. Есть опция вместо показа значимостей коэффициентов показать только значимые на заданном уровне альфа коэффициенты.

Бисериальный коэффициент корреляции предназначен для коррелирования переменной мерной (т.е. количественной непрерывной, меряемой обычно интервальной или отношенческой шкалой) с переменной дихотомической, за которой скрывается мерная величина, имеющая притом в популяции нормальное распределение; дихотомичность же переменной представляется следствием огрубления измерительной шкалы до 2-х градаций. Если такое (довольно смелое) допущение принять, то бисериальный коэффициент суть предположительное (выведенное) значение коэффициента корреляции Пирсона, которое было бы получено, останься вторая переменная мерной, не огрубленной до дихотомии. Формулы (Wherry R. J. Contributions to correlational analysis., 1984):

Rbis = (M1-M0)/S \* PQ/H

M1 и M0 – средние мерной переменной в двух градациях дихотомической переменной;

S – ст. отклонение мерной переменной;

P и Q – доли в распределении дихотомической переменной, P+Q=1;

H – ордината (плотность вероятности) ст. нормальной кривой в точке деления площади под кривой на P и Q.

Приближенная значимость (можно пользоваться если распределение дихотомической переменной не слишком скошенное: P или Q <= 0.9 - 0.95):

Z = Rbis\*sqrt(N) / (sqrt(PQ)/H – Rbis^2)

Z подставляется в ст. нормальное распределение для получения 2-стороннего *p*-значения.

Вычисленный бисериальный коэффициент в редких случаях может превзойти 1 по абсолютной величине. Теоретически коэф-т корреляции не может быть выше 1, и макрос предупреждает, если получены коэффициенты абсолютно выше 1. Такие коэффициенты являются признаком (одним из), что допущение о под-лежащей нормальности дихотомической переменной ложно.

***Подкоманды***

**SCALE**

Мерные переменные (непрерывные или с шкалой с достаточным числом градаций), поименно и/или через “to”. Имена переменных - до 8 байтов длиной. В данных должно не быть значения: -999.

**BINAR**

Дихотомические переменные в двоичной (0 и 1) кодировке, поименно и/или через “to”. Если ваши дихотомические переменные иной кодировки, переделайте их в двоичные. Имена переменных - до 8 байтов длиной.

**MISSING**

Если есть пропущенные данные, то как исключать их:

PAIRWISE - (тж. умолчанию/незаданию) исключить наблюдения попарно, т.е. из каждой пары (коррелируемых) переменных отдельно.

LISTWISE - исключить наблюдения списочно, т.е. из всех переменных, если хотя бы в одной наблюдение есть пропуск.

**ALPHA**

По умолчанию/незаданию этой подкоманды макрос выдает все вычисленные коэффициенты и их 2-сторонние значимости. Вместо этого вы можете задать критический уровень значимости (альфа), - тогда все незначимые на этом уровне коэффициенты окажутся в результатах (включая коэффициенты, выводимые в рабочий файл) занулены, а значимостей в результатах не будет.

***Особые режимы***

Макрос игнорирует взвешивание массива (впрочем, он не берет в процедуру наблюдения с пропущенными и неположительными весами) и не рассчитан на расщепленность массива. Макрос слушается выбора/фильтрации наблюдений (команды FILTER, USE, SELECT IF, N OF CASES) и временных (под TEMPORARY) преобразований.

# МАКРОС !KO\_ACORRD: АВТОКОРРЕЛЯТИВНОЕ РАССТОЯНИЕ

Version 1, Feb 2006. Tested on SPSS 11, 11.5, 13, 14.

!KO\_acorrd vars= *v1 to v5* /\*Cписок числовых переменных (можно ч-з to); это предметы,

/\*между которыми вычислить расстояния

/negat= ASIS /\*Полученные отрицательные расстояния: оставить как есть (ASIS);

/\*отменить знак (ABS, тж п/у); обнулить (ZERO)

/missing= LINT /\*Как поступить с пропущенными значениями:

/\*DELETE - исключить наблюдения списочно (тж п/у);

/\*LAG - заменить пропущ значение на значение ближайшего предшествующего

/\*валидного наблюдения; MEAN - заменить пропущ значение на среднюю 2-х

/\*ближайших окружающих его валидных значений; LINT - заменить пропущ

/\*значение на линейную интерполяцию 2-х ближайших окружающих

/\*его валидных значений.

Минимум надо задать VARS.

Это расстояние придает значения различиям, которые систематичны в смысле (1-лаговой) автокорреляции, т.е. продолжаются при переходе от наблюдения к следующему наблюдению (в том случае если вы сравниваете между собой переменные) или от переменной к следующей переменной (если сравниваете между собой наблюдения). Данный макрос сравнивает между собой переменные, поэтому если стоит цель сравнить наблюдения, транспонируйте прежде данные (меню Data – Transpose). Макрос выводит матрицу расстояний как новый безымянный массив данных.

Для автокоррелятивного расстояния принципиально, в каком порядке расположены наблюдения в массиве, поскольку оно относится к череде наблюдений как к «серии» или «временнóму ряду».

N

DXY = sum [ (X-Y)i \* (X-Y)i-1 ], X, Y – значения сравниваемых переменных в данном наблюдении i или в i-1

i=2

Автокоррелятивное расстояние, как видно из формулы, это ненормированный коэффициент автокорреляции разниц между двумя сериями, X и Y, и поэтому оно может принимать значения и отрицательные. Отрицательная автокорреляция соответствует «переключающейся» систематичности различий: в одном наблюдении X>Y, в следующем X<Y, потом обратно, и т.д. Этот паттерн редко встречается в собираемых социальными науками данных. В них преобладает как раз «продолжающаяся» систематичность различий: X>Y (или X<Y) на протяжении значительного отрезка изучаемой серии (как то: пункты времени, уровни цены и пр.); указанная систематичность соответствует положительной автокорреляции. Чем больше различие и чем «дольше» оно держится вдоль серии, тем будет выше автокоррелятивное расстояние. Большие различия, если они не стойкие, будут давать близкие к 0 значения расстояния, точно так же как почти отсутствующие различия.

***Подкоманды***

**VARS**

Укажите поименно и/или ч-з to числовые количественные (не номинальные) переменные, расстояния между которыми вас интересуют. Имена этих переменных - до 8 байтов длиной. В их данных должно не быть значения: ровно -999. Если величины во входящих данных большие, а наблюдений много (тысячи), расстояния в целом по матрице могут выдаться тоже очень большие. Если это кажется неудобным, поделите данные предварительно на какую-нибудь константу.

**MISSING**

Если у первого наблюдения в массиве есть хотя бы одно пропущенное значение, макрос исключит это наблюдение. То же касается последнего наблюдения в массиве, кроме случая MISSING=LAG, при котором оно не исключается, а заменяется. Для остальных наблюдений с пропущенными данными вы можете заказать, как поступить:

DELETE - (тж. умолчанию/незаданию) исключить наблюдение, если хотя бы в одной переменной оно есть пропуск (system или user-missing).

LAG - заменить пропущенное значение на ближайшее предшествующее (выше в массиве) валидное значение.

MEAN - заменить пропущенное значение на среднюю из пары ближайших, выше и ниже в массиве, валидных значений.

LINT - заменить пропущенное значение на линейную интерполяцию между парой ближайших, выше и ниже в массиве, валидных значений. Интерполяция дает другой эффект, чем MEAN, в смежных пропущенных наблюдениях.

**NEGAT**

Отрицательные расстояния, если такие получатся, можно изменить перед выводом матрицы как массив.

ASIS - оставить все расстояния как есть.

ABS - (тж. по умолчанию/незаданию) отменить отрицательный знак у расстояний.

MEAN - обратить отрицательные расстояния в 0.

Если отрицательная автокорреляция разниц для вас означает наличие различия, как и положительная автокорреляция, используйте ABS. Если же она для вас артефакт или говорит об отсутствии различия, употребите ZERO.

***Особые режимы***

Макрос игнорирует взвешивание массива (впрочем, он не берет в процедуру наблюдения с пропущенными и неположительными весами) и не рассчитан на расщепленность массива. Макрос слушается выбора/фильтрации наблюдений (команды FILTER, USE, SELECT IF, N OF CASES) и временных (под TEMPORARY) преобразований.

# МАКРОС !KO\_RESCR: ПЕРЕШКАЛИРОВАННЫЙ КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ ПИРСОНА

Version 1, Dec 2009. Tested on SPSS Statistics 13, 15, 17, 26.

!KO\_rescr vars= *v1 to v10* /\*Cписок числовых переменных (можно через to), между которыми

/\*вычислять корреляции

/rescale= POLE /\*Перешкалировать к ближнему пределу (POLE, тж п/у) или

/\*в диапазон между пределами (RANGE)

/eigen= YES /\*Контролировать, чтобы матрица перешкалированных r была грамовой:

/\*YES (тж п/у) или NO

/save= *'d:\exercise\newdata.sav'* /\*Опционально, при EIGEN=YES: подогнать переменные под

/\*перешкалированные r: имя внешнего файла или заявленого массива

/id= *id* /\*Опционально: переменная-идентификатор наблюдений

/print= YES /\*Показать корреляции в Output Viewer: YES или NO (тж п/у).

Минимум надо задать VARS.

Макрос вычисляет *r* Пирсона между входящими переменными и перешкалирует его значение относительно предела или пределов, которые *r* мог бы реально достичь в условиях наблюдаемых данных: в условиях переменных с именно такими эмпирическими краевыми распределениями, как у нас в данных.

Макрос выводит матрицу перешкалированных *r* в новый безымянный массив данных. В окне результатов макрос сообщает об усредненной (между всеми *r*) абсолютной величине отличия перешкалированных *r* от соответствующих обычных (т.е. до перешкалирования) *r*. В макросе есть еще возможности (см. п/к EIGEN и SAVE).

Коэффициент корреляции Пирсона теоретически варьирует от -1 до +1. Однако в реальных данных эмпирический диапазон, за который не может выйти *r* между двумя конкретными переменными, обычно сужен по сравнению с теоретическим. Линейная корреляция между X и Y имеет шанс достигнуть +1 тогда и только тогда, когда форма распределения полностью тождественна у X и Y (или, другими словами, распределения X и Y различаются не более чем линейно). В противном случае верхний предел для *r* будет ниже, чем +1. Аналогично, *r* может достигнуть -1 тогда и только тогда, когда форма распределения полностью тождественна у X и -Y. В противном случае нижний предел для *r* будет выше, чем -1. Чтобы корреляция имела одновременно пределы верхний +1 и нижний -1, два распределения должны быть не только идентичны по форме, по и совершенно симметричны по форме. У реальных переменных, количественных или дискретных признаков (включая ликертовские и двоичные данные) , форма распределения часто разная и часто асимметричная; значит, эмпирический диапазон для *r* в той или иной степени сужен.

Большое сужение эмпирического диапазона *r* означает, что сила связи между двумя переменными недооценена коэффициентом корреляции Пирсона, и недооценена тем больше, чем связь сильнее. Ибо сильные отношения тогда неминуемо нелинейны; *r* же «снимает» только линейную порцию отношений. Если исследователь захочет в такой ситуации «снять» своим коэффициентом больше связи, ему стоит подумать о применении какого-нибудь нелинейного коэффициента корреляции, например непараметрического коэффициента *rho* Спирмена. Однако есть и другой вариант решения: перешкалировать наблюдаемый *r* Пирсона относительно своих эмпирических пределов. Скажем, если наблюдается *r* = 0.4, и при этом для данных двух наблюденных переменных *r* не может быть выше 0.95, то *r* подправляется: 0.4/0.95 = 0.42. Это перешкалирование *r* к одному, ближайшему, пределу (полюсу). Можно перешкалировать *r* и к обоим пределам, положительному и отрицательному, вместе; это менее употребительно. Макрос предлагает тот и другой способ перешкалирования на выбор.

***Подкоманды***

**VARS**

Укажите поименно и/или ч-з to числовые переменные, корреляции между которыми вас интересуют. Имена этих переменных - до 8 байтов длиной. Макрос удаляет пропущенные данные списочно, т.е. удаляется целиком наблюдение, если оно пропуск хотя бы в одной из VARS.

**RESCALE**

Вид перешкалирования вычисленного коэффициента корреляции:

POLE - (тж. по умолчанию/незаданию) наблюдаемое (обычное) значение *r* перешкалируется к своему ближайшему эмпирическому пределу. Другими словами, это перешкалирование из диапазона {0 : +1} (если данное значение *r* положительно) или из {0 : -1} (если данное значение *r* отрицательно) в диапазон {0 : предел того же знака, что у данного значения *r*}. Перешкалированный ненулевой *r* всегда будет больше по абс. величине, чем обычный *r*. Нулевой *r* останется нулевым.

RANGE - наблюдаемое значение *r* перешкалируется к обоим эмпирическим пределам вместе, т.е. из диапазона {-1 : +1} в диапазон {отрицательный предел : положительный предел}, так что в итоге *r* займет в последнем то же относительное положение, какое он занимал в диапазоне {-1 : +1}. Перешкалированный *r* по абс. величине может оказаться больше или меньше, чем обычный *r*. Близкие к нулю *r* могут стать нулевыми и даже поменять знак, нулевой *r* может стать ненулевым.

**EIGEN**

По умолчанию/незаданию подкоманды и при EIGEN=YES макрос следит, чтобы матрица перешкалированных *r* была положительно (полу)определенной (синоним: грамовой), т.е. при разложении не давала отрицательных собственных чисел; и если она таковой не окажется, то макрос будет уменьшать (понемногу отменять) величину отличия перешкалированных *r* от обычных *r* до тех пор, пока матрица не станет грамовой. Величина отличия будет уменьшена для всех элементов матрицы в одно и то же число раз. (Некоторые виды анализа, например факторный, требуют, чтобы входящая матрица была положительно определенной или полуопределенной.)

Матрица перешкалированных *r* оказывается не положительно (полу)определенной (не грамовой) обычно при сильно дискретных входящих данных, например дихотомических.

Если среди входящих переменных VARS наблюдается полная коллинеарность, матрица исходных *r* является вырожденной; тогда матрица перешкалированных *r* почти наверняка будет неграмовой и, более того, подкоманда EIGEN скорей всего не справится. Вводите данные без коллинеарностей.

При EIGEN=NO макрос всегда оставляет *r* такими, какими они получились при перешкалировании. При EIGEN=NO п/к SAVE и ID игнорируются.

**SAVE**

Эта подкоманда не действует при EIGEN=NO. Вы можете, указав путь/имя внешнего .SAV-файла или имя внутреннего существующего/объявленного массива данных, затребовать, чтобы макрос сохранил туда входящие переменные, подогнанные им к тому, чтобы воспроизводить итоговые перешкалированные *r*. Т.е. обычные коэффициенты *r*, вычисленные на подогнанных переменных, будут равны перешкалированным *r* (тем, что выведены в новый массив данных). Подгонка переменных осуществляется «главнокомпонентным методом с прокрустовым доворотом»[[4]](#footnote-4). Этот метод старается изменять значения переменных щадяще.

Переменные будут сохранены в стандартизованном виде (средняя = 0, ст. откл. = 1); вы можете потом сами перешкалировать их так, как хотите[[5]](#footnote-5).

**ID**

Подкоманда имеет смысл, если задано SAVE. Укажите одну переменную, идентифицирующую наблюдения во входящих данных. Т.к. наблюдения с пропусками, если они есть, макрос исключит, то ID-переменная позволит вам впоследствии объединить сохраненные подкомандой SAVE переменные с массивом данных, входящим в макрос.

**PRINT**

PRINT=YES печатает в окно результатов матрицу r входящих переменных и матрицу перешкалированных r. По умолчанию, PRINT=NO.

***Особые режимы***

Макрос игнорирует взвешивание массива (впрочем, он не берет в процедуру наблюдения с пропущенными и неположительными весами) и не рассчитан на расщепленность массива. Макрос слушается выбора/фильтрации наблюдений (команды FILTER, USE, SELECT IF, N OF CASES) и временных (под TEMPORARY) преобразований.

1. В литературе встречается также вариант формулы расстояния Хеллингера, где вместо множителя 2 под корнем используется множитель ½. [↑](#footnote-ref-1)
2. SPSS Statistics, команда PROXIMITIES, вычисляет это хи-квадрат расстояние, но рассчитанное на любые неотрицательные данные, например сырые частоты. Макрос делает это только для долей. [↑](#footnote-ref-2)
3. Некоторые меры сходства могут давать диагональный 0, если вектор – константа. [↑](#footnote-ref-3)
4. См. функцию !KO\_TOCOV из коллекции “MATRIX - END MATRIX functions”. [↑](#footnote-ref-4)
5. Для этой цели удобна функция !KO\_RESCALE из коллекции “MATRIX - END MATRIX functions”. [↑](#footnote-ref-5)