IBM SPSS Statistics Base 24

Примечание

Прежде чем использовать эту информацию и продукт, описанный в ней, прочтите сведения в разделе “Замечания” на стр. 211.

Информация о продукте

Это издание применимо к версии 24, выпуск 0, модификация 0 IBM SPSS Statistics и ко всем последующим версиям и модификациям до тех пор, пока в новых изданиях не будет указано иное.

Содержание

Глава 1. Информация о данных

Процедура Информация о данных выводит информацию из словаря данных, такую как имена переменных, метки переменных, метки значений, пропущенные значения, а также итожащие статистики для всех заданных переменных и наборов множественных ответов в активном наборе данных. Для номинальных и порядковых переменных, а также наборов множественных ответов итожащие статистики включают количества и проценты. Для количественных переменных итожащие статистики включают среднее значение, стандартное отклонение и квартили.

Примечание: Процедура Информация о данных игнорирует состояние расщепленных файлов. Это включает группы расщепленных файлов, созданные для множественной импутации пропущенных значений (имеется в надстройке Пропущенные значения).

Доступ к процедуре Информация о данных

1. Выберите в меню:

   Анализ > Отчеты > Информация о данных
2. Откройте вкладку Переменные.
3. Выберите одну или несколько переменных и/или наборов множественных ответов.

Дополнительно вы можете:

- Управлять показанной информацией о переменных.
- Управлять выводом статистик (или исключить все итожащие статистики).
- Управлять порядком вывода переменных и наборов множественных ответов.
- Изменять шкалу измерений для любой переменной в списке исходных переменных, чтобы изменить выводимые итожащие статистики. Дополнительную информацию смотрите в разделе “Вкладка Информация о данных: Статистики” на стр. 3.

Изменение шкалы измерений

Можно временно изменить шкалу измерений для переменных. (Шкалу измерений нельзя изменить для наборов множественных ответов. Они всегда считаются номинальными.)

1. Щелкните правой кнопкой мыши по переменной в исходном списке.
2. Во всплывающем меню выберите шкалу измерений.

После этого шкала измерений будет временно изменена. С практической точки зрения это полезно только для числовых переменных. Шкала измерений для текстовых переменных может быть только номинальной или порядковой, причем в процедуре Информация о данных обе эти шкалы обрабатываются идентично.

Вкладка Информация о данных: Вывод

Вкладка Вывод управляет информацией о переменных, включаемой в вывод для всех переменных и наборов множественных ответов, порядком вывода переменных и наборов множественных ответов, а также содержимым дополнительной таблицы информации о файле.

Информация о переменной

Здесь задается информация из словаря данных, выводимая для всех переменных.

Положение. Целое число, представляющее положение переменной в порядке их расположения в файле. Этот параметр недоступен для наборов множественных ответов.

Метка. Описательная метка переменной или набора множественных ответов.

Тип. Основной тип данных. Тип может быть *Числовой*, *Текстовый* или *Набор множественных ответов*.

Формат. Формат вывода переменной, например *A4*, *F8.2* или *DATE11*. Этот параметр недоступен для наборов множественных ответов.

Шкала измерений. Возможные значения: *Номинальная*, *Порядковая*, *Количественная* и *Неизвестная*. Выводимым значением является шкала измерений, хранимая в словаре данных, и на нее не влияет никакое временное изменение шкалы измерений, сделанное в списке исходных переменных в представлении Переменные. Этот параметр недоступен для наборов множественных ответов.

Примечание: Шкала измерений для числовых переменных может быть "неизвестной" до первого прохода данных, если она не была задана явно, как, например, для данных, считанных из внешнего источника, или вновь создаваемых переменных. Дополнительную информацию смотрите в разделе .

Роль. Некоторые диалоговые окна поддерживают возможность предварительного выбора переменных для анализа, основанного на определенных ролях.

Метки значений. Описательные метки, связанные с определенными значениями данных.

- Если на вкладке Статистики выбрано Количество или Проценты , то заданные метки значений включаются в вывод, даже если они не были здесь выбраны для вывода.
- Для наборов множественных дихотомий метками значений являются метки переменных для элементарных переменных в наборе или метки подсчитываемых значений в зависимости от того, как определен набор. Дополнительную информацию смотрите в разделе .

Пропущенные значения. Пользовательские пропущенные значения. Если на вкладке Статистики выбрано Количество или Проценты, заданные метки значений включаются в вывод, даже если вы не выбрали здесь Отсутствующие значения. Этот параметр недоступен для наборов множественных ответов.

Настраиваемые атрибуты. Задаваемые пользователем атрибуты переменных. В вывод включаются и имена, и значения задаваемых пользователем атрибутов всех переменных. Дополнительную информацию смотрите в разделе . Этот параметр недоступен для наборов множественных ответов.

Зарезервированные атрибуты. Зарезервированные атрибуты системных переменных. Можно вывести системные атрибуты, но изменять их не следует. Имена системных атрибутов начинаются со знака доллара ($). Скрытые атрибуты с названиями, начинающимися с "@" или "$@", не включаются в вывод. В вывод включаются и имена, и значения системных атрибутов, связанных со всеми переменными. Этот параметр недоступен для наборов множественных ответов.

Информация о файле

Дополнительная таблица информации о файле может содержать любой из перечисленных ниже атрибутов файла:

Имя файла. Имя файла данных IBM® SPSS Statistics. Если набор данных никогда не был сохранен в формате IBM SPSS Statistics, то имя файла данных отсутствует. (Если в заголовке окна редактора данных нет имени файла, значит у активного набора данных нет имени файла.)

Положение. Каталог (папка), где расположен файл данных IBM SPSS Statistics. Если набор данных никогда не был сохранен в формате IBM SPSS Statistics, то местоположения у него нет.

Число наблюдений. Число наблюдений в активном наборе данных. Это общее число наблюдений, включая любые наблюдения, которые могли быть исключены при выводе итожащих статистик из-за условий фильтрации.

Метка. Это метка файла (если она есть), заданная командой FILE LABEL.

Документы. Текст документа файла данных.

Состояние взвешивания. Если взвешивание включено, отображается имя переменной взвешивания. Дополнительную информацию смотрите в разделе .

Настраиваемые атрибуты. Задаваемые пользователем атрибуты файла данных. Атрибуты файла данных, заданные командой DATAFILE ATTRIBUTE.

Зарезервированные атрибуты. Зарезервированные системные атрибуты файла данных. Можно вывести системные атрибуты, но изменять их не следует. Имена системных атрибутов начинаются со знака доллара ($). Скрытые атрибуты с названиями, начинающимися с "@" или "$@", не включаются в вывод. В вывод включаются и имена, и значения всех системных атрибутов файла данных.

Порядок вывода переменных

Имеются следующие альтернативны управления порядком, в котором выводятся переменные и наборы множественных ответов.

По алфавиту. Алфавитный порядок по именам переменных.

Файл. Порядок отображения переменных в наборе данных (порядок, в котором они отображаются в редакторе данных). При сортировке в порядке возрастания наборы множественных ответов выводятся последними, после всех выбранных переменных.

Шкала измерений. Сортировка по шкале измерений. При этом создаются четыре группы сортировки: номинальная, порядковая, количественная и неизвестная. Наборы множественных ответов рассматриваются как номинальные.

Примечание: Шкала измерений для числовых переменных может быть "неизвестной" до первого прохода данных, если она не была задана явно, как, например, для данных, считанных из внешнего источника, или вновь создаваемых переменных.

Список переменных. Порядок, в котором переменные и наборы множественных ответов показываются в списке выбранных переменных в представлении Переменные.

Имена атрибутов, задаваемые пользователем. В список параметров сортировки также входят имена любых определенных пользователем атрибутов переменных. При сортировке в порядке возрастания переменные без атрибутов показываются вверху, за ними следуют переменные с атрибутами, но без заданных значений атрибутов, и последними идут переменные с заданными значениями атрибутов в алфавитном порядке значений.

Максимальное количество категорий

Если в вывод включаются метки значений, количества или проценты для всех уникальных значений, то эта информация не будет выводиться в таблице, если число значений превышает указанное значение. По умолчанию эта информация не выводится, если число уникальных значений для переменной больше 200.

Вкладка Информация о данных: Статистики

На вкладке Статистики можно управлять выводом итожащих статистик и при желании не выводить их совсем.

Количества и проценты

Для номинальных и порядковых переменных, наборов множественных ответов, а также значений количественных переменных с метками доступны следующие статистики:

Количество. Количество наблюдений (объектов), имеющих каждое значение (или диапазон значений) переменной.

Проценты. Процент наблюдений, имеющих конкретное значение.

Положение центра распределения и разброс

Для количественных переменных доступны следующие статистики:

Mean. Мера центральной тенденции. Арифметическое среднее; сумма, деленная на число наблюдений.

Стандартное отклонение. Мера дисперсии вокруг среднего, выраженная в тех же единицах измерения, что и наблюдения. Равна корню квадратному из дисперсии. При нормальном распределении 68% наблюдений укладываются в одно стандартное отклонение от среднего, и 95% - в два стандартных отклонения. Если, например, средний возраст равен 45 годам со стандартным отклонением 10, то 95% наблюдений должны оказаться между 25 и 65 годами при нормальном распределении.

Квартили. Значения 25-го, 50-го и 75-го процентилей.

Примечание: Можно временно изменить шкалу измерений переменной (и, следовательно, изменить итожащие статистики, выводимые для этой переменной) в списке исходных переменных в представлении Переменные.

Глава 2. Частоты

Процедура Частоты дает возможность вычислять статистики и строить диаграммы, полезные для описания многих типов переменных. Процедура Частоты - это хорошее начало в исследовании данных.

При построении таблиц частот и столбчатых диаграмм можно задать порядок значений анализируемых переменных - по возрастанию или убыванию значений или частот. Если количество значений переменной слишком велико, вывод таблицы частот может быть запрещен. В диаграммах можно использовать частоты (по умолчанию) или проценты.

Пример. Как распределены клиенты по типу организаций, в которых они работают? Из вывода можно узнать, что 37.5% клиентов работают в государственных организациях, 24.9% работают в коммерческих организациях, 28.1% - в университетах и институтах, и 9.4% в сфере здравоохранения. Для непрерывных, количественных данных, например, дохода от продаж, можно определить, что средний доход одной продажи - $3.576, а стандартное отклонение - $1.078.

Статистики и графики. Частоты, проценты, кумулятивные проценты, среднее значение, медиана, мода, сумма, стандартное отклонение, дисперсия, размах, минимальное и максимальное значения переменных, стандартная ошибка среднего значения, асимметрия, эксцесс, стандартные ошибки оценок асимметрии и эксцесса, квартили, определяемые пользователем процентили, столбчатые диаграммы, круговые диаграммы и гистограммы.

Данные для процедуры Частоты

Данные. Для кодировки значений категориальных переменных (номинальных или порядковых) используйте числа или строки.

Допущения. Частоты и проценты дают полезные описания данных, независимо от вида распределения, особенно для переменных с упорядоченными и неупорядоченными категориями . Большинство необязательных итожащих статистик, например, среднее значение и стандартное отклонение, основаны на теории нормального распределения и применимы к количественным переменным с симметричным распределением. Робастные статистики, такие, как медиана, квартили и процентили, подходят для анализа числовых переменных, которые могут не удовлетворять предположению о нормальности распределения.

Как вывести частотную таблицу

1. Выберите в меню:

   Анализ > Описательные статистики > Частоты...
2. Выберите одну или несколько категориальных или количественных переменных.

Дополнительно вы можете:

- Щелкнуть мышью по кнопке **Статистики** , чтобы задать вычисление описательных статистик для количественных переменных.
- Щелкнуть мышью по кнопке **Диаграммы** , чтобы задать вывод столбчатых диаграмм, круговых диаграмм и гистограмм.
- Щелкнуть мышью по кнопке **Формат** , чтобы задать порядок, в котором будут выводиться результаты.

Статистики в процедуре Частоты

Значения процентилей. Значение процентили - это значение количественной переменной, которое разделяет упорядоченные данные на группы таким образом, что определенный процент наблюдений имеет значения этой количественной переменной меньше значения процентили, а другой процент наблюдений имеет значения этой количественной переменной больше значения процентили. Квартили - это 25%-е, 50%-е и 75%-е процентили, которые разделяют наблюдения на четыре группы одинакового объема. Если вы хотите получить разбивку на иное число равных групп, воспользуйтесь пунктом **Процентили для n равных групп**. Можно также задать отдельные процентили (например, 95%-й процентиль - значение, меньше которого значения 95% наблюдений).

Расположение (центральная тенденция). Статистики, описывающие расположения распределений, включают среднее, медиану, моду и сумму всех значений.

- *Mean*. Мера центральной тенденции. Арифметическое среднее; сумма, деленная на число наблюдений.
- *Медиана*. Значение, выше и ниже которого попадает по половине наблюдений, иначе 50-й процентиль. Если число наблюдений четно, медиана есть арифметическое среднее двух находящихся в середине значений, если выборку упорядочить по убыванию или по возрастанию. Медиана представляет собой меру центральной тенденции, которая нечувствительна к выбросам, в отличие от среднего значения, которое могут исказить несколько экстремально больших или малых значений.
- *Мода*. Чаще всего встречающееся значение. Если таких значений несколько, каждое из них является модой. Процедура Частоты выдает только наименьшее из этих значений.
- *Sum*. Сумма или итог для всех значений по всем наблюдениям, имеющим непропущенные значения.

Разброс. Статистики, которые измеряют вариацию или разброс в данных, включают стандартное отклонение, дисперсию, размах, минимальное значение, максимальное значение и стандартную ошибку среднего.

- *Среднеквадратичное отклонение*. Мера дисперсии вокруг среднего, выраженная в тех же единицах измерения, что и наблюдения. Равна корню квадратному из дисперсии. При нормальном распределении 68% наблюдений укладываются в одно стандартное отклонение от среднего, и 95% - в два стандартных отклонения. Если, например, средний возраст равен 45 годам со стандартным отклонением 10, то 95% наблюдений должны оказаться между 25 и 65 годами при нормальном распределении.
- *Дисперсия*. Мера дисперсии относительно среднего значения. Равна сумме квадратов отклонений от среднего, деленной на число, на единицу меньшее числа наблюдений. Дисперсия измеряется в единицах, которые равны квадратам единиц измерения самой переменной.
- *Range* . Разность между наибольшим и наименьшим значениями числовой переменной; максимум минус минимум.
- *Минимум*. Наименьшее значение числовой переменной.
- *Максимум*. Наибольшее значение числовой переменной.
- *Среднеквадратическая ошибка среднего*. Мера того, как сильно могут отличаться значения среднего от выборки к выборке, извлекаемых из одного и того же распределения. Можно применять для грубого сравнения наблюденного среднего с гипотетическим значением (то есть можно заключить, что два значения различаются, если отношение их разности к стандартному отклонению меньше -2 или больше +2).

Распределение. Асимметрия и эксцесс - это статистики, описывающие форму и симметричность распределения. Эти статистики выводятся вместе с их стандартными ошибками.

- *Асимметрия*. Мера асимметрии распределения. Нормальное распределение симметрично, и для него асимметрия равна 0. Распределение со значимой положительной асимметрией имеет длинный хвост справа. Распределение со значимой отрицательной асимметрией имеет длинный хвост слева. В качестве грубого правила можно сказать, что значение асимметрии, более чем вдвое превышающее ее стандартную ошибку, указывает на наличие асимметрии распределения.
- *Эксцесс*. Мера сгруппированности наблюдений вокруг центральной точки. Для нормального распределения значение эксцесса равно 0. Положительный эксцесс указывает на то, что по отношению к

нормальному распределению наблюдения для таких распределений сгруппированы более плотно около центра и имеют более тонкие хвосты до экстремумов распределения, и более толстые хвосты в области экстремальных значений. Отрицательный эксцесс указывает на то, что по отношению к нормальному распределению наблюдения для таких распределений сгруппированы менее плотно около центра и имеют более толстые хвосты до экстремумов распределения, и более тонкие хвосты в области экстремальных значений.

Значения - центры групп. Если значения анализируемых данных представлены средними точками групп (например, возраст всех людей от 30 до 40 лет закодирован числом 35), можно пометить этот элемент, чтобы получить оценки медианы и процентилей исходных, несгруппированных данных.

Диаграммы в процедуре Частоты

Тип диаграммы. Круговые диаграммы представляют вклад отдельных частей в целое. Каждый сектор круговой диаграммы соответствует группе, заданной одной группирующей переменной. Столбчатая диаграмма выводит число наблюдений для каждой категории, определяемой значением, в виде отдельного столбца, что позволяет визуально сравнивать категории. Гистограммы также состоят из столбцов; но каждый из них соответствует одинаковому интервалу значений исследуемой переменной. Высота каждого столбца отражает количество значений числовой переменной, попавших внутрь интервала, соответствующего этому столбцу. Гистограмма показывает форму, центр и разброс распределения. На гистограмму можно наложить кривую нормального распределения, которая поможет оценить, насколько распределение данных близко к нормальному.

Значения на диаграмме. Для столбчатых диаграмм можно помечать ось Y частотами или процентами.

Частоты: Формат

Упорядочить по. Данные в таблице частот могут быть расположены в порядке возрастания или убывания значений данных, либо в порядке возрастания или убывания частот этих значений. Однако, если задано построение гистограмм или вычисление процентилей, то процедура Частоты предполагает, что анализируемая переменная является количественной, и выводит ее значения в порядке возрастания.

Несколько переменных. Если вы строите таблицы статистик для нескольких переменных, можно либо вывести все переменные в одной таблице (**Сравнить переменные**), либо вывести отдельную таблицу для каждой переменной (**Выводить по переменным**).

Отключить таблицы со многими категориями. Этот параметр предотвращает вывод таблиц с числом категорий, большим заданного значения.

Глава 3. Описательные статистики

Процедура Описательные статистики осуществляет вывод одномерных итожащих статистик для нескольких переменных в одной таблице, а также вычисляет стандартизованные значения (*z* -значения) переменных. Переменные могут быть упорядочены по величине их средних значений (в порядке возрастания или убывания), по алфавиту или в порядке, в котором вы выбираете переменные (по умолчанию).

При сохранении *z*-оценок они добавляются в данные в редакторе данных и становятся доступны для диаграмм, списков данных и анализа. Если переменные измерены в разных единицах (например, валовой внутренний продукт на душу населения и процент грамотных), преобразование к *z* -значениям приводит переменные к единому масштабу, что облегчает их визуальное сравнение.

Пример. Если каждое наблюдение в анализируемых данных содержит итоги дневных объемов продаж для одного из членов коллектива продавцов (например, одно значение - для Алексея, одно - для Марии, одно - для Бориса) в течение нескольких месяцев, то процедура Описательные статистики может рассчитать средний дневной объем продаж для каждого продавца и расположить результаты в порядке от наиболее высоких средних ежедневных продаж к наиболее низким.

Статистика. Объем выборки, среднее значение, минимальное и максимальное значения, стандартное отклонение, дисперсия, размах, сумма, стандартная ошибка среднего, асимметрия, эксцесс, стандартные ошибки асимметрии и эксцесса.

Данные для процедуры Описательные статистики

Данные. Используйте числовые переменные после того, как вы исследовали их диаграммы на наличие ошибок записи, выбросов и аномалий в распределениях. Процедура Описательные статистики очень эффективно работает с файлами большого размера (содержащими тысячи наблюдений).

Допущения. Большинство статистик, которые могут быть вычислены при работе с данной процедурой (в том числе и *z* -значения), основаны на теории нормального распределения и подходят для количественных переменных (измеренных в интервальной шкале или шкале отношений), распределенных симметрично. Избегайте переменных с неупорядоченными категориями или несимметричными распределениями. Распределение *z*-значений имеет ту же форму, что и распределение исходных данных; поэтому переход к *z*-значениям не является средством исправления "недостатков" данных.

Как получить описательные статистики

1. Выберите в меню:

   Анализ > Описательные статистики > Описательные...
2. Выберите одну или несколько переменных.

Дополнительно вы можете:

- Выбрать параметр **Сохранить стандартизованные значения в переменных** , чтобы сохранить *z* -значения как новые переменные.
- Щелкнуть мышью по кнопке **Параметры** , чтобы выбрать дополнительные статистики и изменить порядок вывода результатов.

Параметры процедуры Описательные статистики

Среднее и сумма. Среднее значение или арифметическое среднее значение выводятся по умолчанию.

Разброс. Статистики, которые измеряют разброс данных, включают в себя стандартное отклонение, дисперсию, размах, минимальное и максимальное значения, а также стандартную ошибку среднего значения.

- *Стандартное отклонение*. Мера дисперсии вокруг среднего, выраженная в тех же единицах измерения, что и наблюдения. Равна корню квадратному из дисперсии. При нормальном распределении 68% наблюдений укладываются в одно стандартное отклонение от среднего, и 95% - в два стандартных отклонения. Если, например, средний возраст равен 45 годам со стандартным отклонением 10, то 95% наблюдений должны оказаться между 25 и 65 годами при нормальном распределении.
- *Дисперсия*. Мера дисперсии относительно среднего значения. Равна сумме квадратов отклонений от среднего, деленной на число, на единицу меньшее числа наблюдений. Дисперсия измеряется в единицах, которые равны квадратам единиц измерения самой переменной.
- *Range* . Разность между наибольшим и наименьшим значениями числовой переменной; максимум минус минимум.
- *Минимум*. Наименьшее значение числовой переменной.
- *Максимум*. Наибольшее значение числовой переменной.
- *Среднеквадратическая ошибка среднего*. Мера того, как сильно могут отличаться значения среднего от выборки к выборке, извлекаемых из одного и того же распределения. Можно применять для грубого сравнения наблюденного среднего с гипотетическим значением (то есть можно заключить, что два значения различаются, если отношение их разности к стандартному отклонению меньше -2 или больше +2).

Распределение. Эксцесс и асимметрия представляют собой статистики, описывающие форму и степень симметричности распределения. Эти статистики выводятся вместе с их стандартными ошибками.

- *Эксцесс*. Мера сгруппированности наблюдений вокруг центральной точки. Для нормального распределения значение эксцесса равно 0. Положительный эксцесс указывает на то, что по отношению к нормальному распределению наблюдения для таких распределений сгруппированы более плотно около центра и имеют более тонкие хвосты до экстремумов распределения, и более толстые хвосты в области экстремальных значений. Отрицательный эксцесс указывает на то, что по отношению к нормальному распределению наблюдения для таких распределений сгруппированы менее плотно около центра и имеют более толстые хвосты до экстремумов распределения, и более тонкие хвосты в области экстремальных значений.
- *Асимметрия*. Мера асимметрии распределения. Нормальное распределение симметрично, и для него асимметрия равна 0. Распределение со значимой положительной асимметрией имеет длинный хвост справа. Распределение со значимой отрицательной асимметрией имеет длинный хвост слева. В качестве грубого правила можно сказать, что значение асимметрии, более чем вдвое превышающее ее стандартную ошибку, указывает на наличие асимметрии распределения.

Порядок вывода. По умолчанию переменные выводятся в том порядке, в котором они выбирались пользователем. Вы также можете выводить переменные в алфавитном порядке, в порядке возрастания средних значений или в порядке убывания средних значений.

Команда DESCRIPTIVES: дополнительные возможности

Язык синтаксиса команд также позволяет:

- Сохранять стандартизованные значения (*z* -значения) для некоторых, но не всех переменных (с помощью подкоманды `VARIABLES`).
- Задавать имена новых переменных, содержащих стандартизованные значения (с помощью подкоманды `VARIABLES`).
- Исключать из анализа наблюдения с пропущенными значениями в какой-либо переменной (с помощью подкоманды `MISSING`).
- Сортировать переменные в выводе по значению любой статистики, а не только среднего (с помощью подкоманды `SORT`).

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Глава 4. Исследовать

Процедура Исследовать вычисляет итожащие статистики и выводит диаграммы как для всех наблюдений, так и отдельно для групп наблюдений. У этой процедуры много полезных способов применения: с ее помощью производится отслеживание данных, идентификация выбросов, описание, проверка предположений и описание различий между группами наблюдений. Отслеживание данных может показать наличие необычных значений, экстремальных значений, разрывов в данных или других особенностей. Процедура Исследовать позволяет определить, подходят ли для анализа ваших данных статистические методы, которые вы собираетесь использовать. Результаты процедуры Исследовать могут показать, что необходимо провести преобразование данных, если применение выбранного метода требует нормально распределенных данных. Или вы можете решить, что надо воспользоваться непараметрическими критериями.

Пример. Рассмотрим распределение времени, необходимого крысам на изучение лабиринта, при применении четырех различных схем кормления. Для каждой из четырех групп можно посмотреть, является ли распределение времени приближенно нормальным, и проверить, совпадают ли четыре дисперсии. Можно выделить наблюдения, которым соответствуют пять наименьших и пять наибольших значений времени. Ящичные диаграммы с усами и диаграммы "ствол-лист" графически подытоживают информацию о распределении времени на изучение для каждой группы.

Статистики и графики. Среднее значение, медиана, 5%-е усеченное среднее, стандартная ошибка, дисперсия, стандартное отклонение, минимальное и максимальное значения переменных, размах, межквартильный размах, асимметрия, эксцесс, стандартные ошибки асимметрии и эксцесса, доверительный интервал для среднего с задаваемым уровнем, процентили, робастные оценки центральной тенденции (М-оценки Хубера, Эндрюса, Хемпеля и Тьюки), пять наименьших и пять наибольших значений переменных, статистика Колмогорова-Смирнова с уровнем значимости Лильефорса для проверки на нормальность, статистика Шапиро-Уилкса. Ящичные диаграммы с усами, диаграммы "ствол-лист", гистограммы, нормальные вероятностные графики, диаграммы разброса по уровням с критерием Ливиня и возможностью задать преобразование данных.

Данные для процедуры Исследовать

Данные. Процедура Исследовать используется для анализа количественных переменных, заданных в интервальной шкале или шкале отношений. Факторная переменная (используемая для разбиения наблюдений на группы) должна иметь разумное число различных значений (категорий). Эти значения могут быть числовыми или короткими текстовыми. Переменная в поле Метить значениями используется для того, чтобы ее значениями метить выбросы в ящичных диаграммах с усами. Она может быть короткой текстовой, длинной текстовой (первые 15 байтов) или числовой.

Допущения. Распределение исследуемых данных не обязательно должно быть симметричным или нормальным.

Как Исследовать данные

1. Выберите в меню:

   Анализ > Описательные статистики > Исследовать...
2. Выберите одну или несколько зависимых переменных.

Дополнительно вы можете:

- Выбрать одну или несколько факторных переменных, значения которых зададут разбиение наблюдений на группы.
- Выбрать идентификационную переменную, чтобы метить наблюдения.

- Щелкнуть мышью по кнопке **Статистики** , чтобы задать вывод робастных оценок, выбросов, процентилей, частотных таблиц.
- Щелкнуть мышью по кнопке **Графики** и задать построение гистограмм, графиков и критериев для проверки нормальности, а также диаграмм разброса по уровням с критерием Ливиня.
- Щелкнуть мышью по кнопке **Параметры** и задать способ работы с пропущенными значениями.

Статистики процедуры Исследовать

Описательные статистики. Эти характеристики центральной тенденции и разброса выводятся по умолчанию. Характеристики положения центра распределения описывают положение распределения; они включают среднее значение, медиану и 5%-е усеченное среднее. Характеристики дисперсии отражают степень различия значений исследуемых данных; они включают стандартную ошибку, дисперсию, стандартное отклонение, минимальное и максимальное значения переменных, диапазон и межквартильный диапазон. Описательные статистики включают также характеристики формы распределения, такие как асимметрия и эксцесс, которые выводятся вместе со своими стандартными ошибками. Выводится также 95% доверительный интервал для среднего, можно задать иное значение доверительного уровня.

М-оценки. Робастные альтернативы выборочным среднему и медиане для оценивания положения. Они различаются весами, приписываемыми наблюдениям. Выводятся следующие оценки: М-оценка Хубера, волновая оценка Эндрюса, нисходящая М-оценка Хампеля, бивес-оценка Тьюки.

Выбросы. Выводятся пять наименьших и пять наибольших значений с метками наблюдений.

Процентили. Выводятся значения 5%-го, 10%-го, 25%-го, 50%-го, 75%-го, 90%-го и 95%-го процентилей.

Графики процедуры Исследовать

Ящичные диаграммы. Эти параметры управляют выводом ящичных диаграмм в случае, когда вы анализируете более одной зависимой переменной. Выбор **Уровни фактора вместе** формирует отдельный вывод для каждой зависимой переменной. В рамках производимого вывода ящичные диаграммы с усами выводятся для каждой из групп, определяемых значениями факторной переменной. Выбор **Зависимые вместе** формирует отдельный вывод для каждой из групп, определяемых факторной переменной. В рамках вывода ящичные диаграммы с усами показаны друг рядом с другом для каждой зависимой переменной. Это особенно удобно, когда различные переменные представляют одну и ту же характеристику, измеренную в разные моменты времени.

Описательные. Группа Описательные позволяет задать построение диаграмм "ствол-лист" и гистограмм.

Графики и критерии для проверки нормальности. Вывод нормального вероятностного графика и нормального вероятностного графика с удаленным трендом. Осуществляется также вывод значений статистики критерия Колмогорова-Смирнова с уровнем значимости Лильефорса для проверки на нормальность. Если заданы нецелочисленные веса, то статистика Шапиро-Уилкса вычисляется при взвешенном объеме выборки от 3 до 50. Если веса не заданы или целочисленны, то эта статистика рассчитывается, когда взвешенный объем выборки находится в пределах от 3 до 5 000.

Разброс по уровням с критерием Ливиня. Позволяет задать преобразование данных для диаграмм с разбросом (межквартильными размахами групп) и уровнем (медианами групп) по осям. Для всех диаграмм этого типа выводятся коэффициент наклона линии регрессии и значение робастного критерия однородности дисперсии Ливиня. Если выбрано преобразование данных, то критерий Ливиня вычисляется для преобразованных данных. Если не выбрана ни одна факторная переменная, то диаграммы не строятся. Выбор пункта **Оценка степени** позволяет изобразить на графике натуральные логарифмы межквартильных диапазонов против натуральных логарифмов медиан для всех групп вместе с оценкой степенного преобразования, которое делает равными дисперсии во всех группах. Диаграмма с разбросом и уровнем по осям помогает определить показатель степени для преобразования, которое стабилизирует (делает равными) дисперсии по группам. Выбор пункта **Преобразование** позволяет задать одно из степенных

преобразований (возможно, вы захотите последовать рекомендации пункта Оценка степени) и получить диаграммы, построенные для преобразованных данных. На график выводятся межквартильный диапазон и медиана преобразованных данных. Чтобы построить графики для исходных данных, выберите пункт **Без преобразования** . Это соответствует степенному преобразованию с показателем степени, равным 1.

Степенные преобразования в процедуре Исследовать

Для диаграмм с разбросом и уровнем по осям возможны степенные преобразования. Чтобы осуществить преобразование данных, вам необходимо выбрать степень производимого преобразования. Вы можете выбрать одну из следующих альтернатив:

- **Натуральный логарифм.** Натуральный логарифм (преобразование) Это вариант по умолчанию.
- **1/кв.корень.** Для каждого значения данных вычисляется величина, обратная квадратному корню из этого значения.
- **Обр. величина.** Для каждого значения данных вычисляется обратная ему величина.
- **Кв. корень.** Вычисляется квадратный корень каждого значения данных.
- **Квадрат.** Каждое значение данных возводится в квадрат.
- **Куб.** Каждое значение данных возводится в куб.

Параметры процедуры Исследовать

Пропущенные значения. Эта группа параметров позволяет управлять обработкой пропущенных значений.

- **Исключать целиком.** На всех этапах анализа исключаются наблюдения, имеющие пропущенные значения какой-либо зависимой или факторной переменной. Это вариант по умолчанию.
- **Исключать попарно.** Если наблюдения не имеют пропущенных значений для переменных в группе (ячейке), то они используются в анализе этой группы. Наблюдение может иметь пропущенные значения для переменных, которые используются в других группах.
- **Помещать в отчет.** Пропущенные значения для факторных переменных рассматриваются как отдельная категория. Для этой дополнительной категории выводится вся информация, как и для других категорий. Таблицы частот включают категории, соответствующие пропущенным значениям. Пропущенные значения для факторной переменной включаются в анализ, но отмечаются как пропущенные.

Команда EXAMINE: дополнительные возможности

Процедура Исследовать использует синтаксис команды `EXAMINE` . Язык синтаксиса команд также позволяет:

- Запросить итоговые вывод и графики в дополнение к выводу и графикам для групп, заданных факторными переменными (с помощью подкоманды `TOTAL`).
- Задать общую шкалу для группы ящичных диаграмм (с помощью подкоманды `SCALE`).
- Задать взаимодействия факторных переменных (с помощью подкоманды `VARIABLES`).
- Задать проценты, отличные от заданных по умолчанию (с помощью подкоманды `PERCENTILES`).
- Вычислить проценты, используя любой из пяти методов (с помощью подкоманды `PERCENTILES`).
- Задать любое степенное преобразование для диаграмм разброса по уровням (с помощью подкоманды `PLOT`).
- Задать число выводимых экстремальных значений (с помощью подкоманды `STATISTICS`).
- Задать параметры для М-оценок, робастных оценок положения (с помощью подкоманды `MESTIMATORS`).

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Глава 5. Таблицы сопряженности

Процедура Таблицы сопряженности формирует двумерные и многомерные таблицы, а также вычисляет целый ряд критериев и мер силы связи для двумерных таблиц. Структура таблицы и то, упорядочены категории или нет, определяет, какие меры и критерии использовать.

Статистики таблиц сопряженности и меры силы связи вычисляются только для двумерных таблиц. Если вы задали строку, столбец и фактор слоя (управляющую переменную), то процедура Таблицы сопряженности формирует панель соответствующих статистик и мер для каждого значения фактора слоя (или комбинации значений, если факторов два или более). Например, если *пол* - это фактор слоя для таблицы переменных *состоит в браке* (да, нет) и *жизнь* (как воспринимается жизнь - волнующая, обычная или скучная), то результаты двумерной таблицы будут вычисляться отдельно для женщин и отдельно для мужчин, и выводиться в виде двух панелей, расположенных одна за другой.

Пример. Верно ли, что клиенты мелких компаний приносят больший доход от продажи им услуг (например, консультации или тренинг), чем клиенты крупных компаний? Из таблицы сопряженности вы, возможно, увидите, что большинство мелких компаний (менее 500 работников) приносят высокий доход, тогда как большинство крупных компаний (более 2 500 работников) приносят низкий доход.

Статистики и меры силы связи. Хи-квадрат Пирсона, хи-квадрат отношение правдоподобия, критерий линейно-линейной связи, точный критерий Фишера, скорректированный хи-квадрат Йетса, *r* Пирсона, ро Спирмана, коэффициент сопряженности, фи, *V* Крамера, симметричное и несимметричное лямбда, тау Гудмана и Краскала, коэффициент неопределенности, гамма,*d* Somepca, тау-*b* Кендалла, тау-*c* Кендалла, коэффициент эта, каппа Коэна, оценка относительного риска, отношение шансов, критерий Макнемара, статистики Кокрена и Мантеля-Хенцеля, а также статистики пропорций столбцов.

Данные для процедуры Таблицы сопряженности

Данные. Для того чтобы задать категории каждой из использующихся в таблице переменных, используйте значения числовых или текстовых (длиной до восьми байт) переменных. Например, значения переменной *пол* можно закодировать как 1 и 2 или как *мужской* и *женский* .

Допущения. Для вычисления некоторых статистик и мер требуется, чтобы категории были упорядочены (порядковые данные) или чтобы значения были количественными (интервальные данные или данные, заданные в шкале отношений). Применение других статистик корректно и в том случае, когда категории переменных в таблице не упорядочены (номинальные данные). Для статистик, в основе которых лежит критерий хи-квадрат (статистика фи, статистика *V* Крамера и коэффициент сопряженности), данные должны представлять собой случайную выборку из полиномиального распределения.

Примечание: Порядковые переменные должны быть либо числовыми кодами, представляющими категории (например, 1= *низкий* , 2= *средний* , 3= *высокий*), либо строчными значениями. Однако предполагается, что алфавитный порядок строковых значений отражает истинный порядок категорий. Например, для строковой переменной со значениями *низкий* , *средний* , *высокий* интерпретируемый порядок категорий следующий: *высокий* , *низкий* , *средний* , что не соответствует правильному порядку. Вообще говоря, для представления порядковых данных надежнее использовать числовые коды.

Как построить таблицу сопряженности

1. Выберите в меню:

   Анализ > Описательные статистики > Таблицы сопряженности...
2. Выберите одну или несколько переменных для строк и одну или несколько переменных для столбцов.

Дополнительно вы можете:

- Выбрать одну или несколько управляющих (слоевых) переменных.
- Щелкнуть мышью по кнопке **Статистики** и выбрать нужные критерии и меры силы связи для двумерных таблиц или подтаблиц.
- Щелкнуть мышью по кнопке **Ячейки** , чтобы задать вывод наблюденных и ожидаемых значений, процентов, а также остатков.
- Щелкнуть мышью по кнопке **Формат** для задания порядка, в котором следует располагать категории.

Слои таблиц сопряженности

Если вы выбрали одну или несколько слоевых переменных, то для каждого значения каждой слоевой переменной (управляющей переменной) строится отдельная таблица сопряженности. Так, если у вас имеется одна переменная строки, одна переменная столбца и одна переменная слоя с двумя значениями, то вы получите по отдельной двухмерной таблице для каждой категории переменной слоя. Чтобы задать другие слои управляющих переменных, щелкните по **Далее** . Подтаблицы строятся для каждой комбинации категорий первой слоевой переменной и второй слоевой переменной и так далее. Если запрошен вывод статистик и мер силы связи, то они вычисляются только для двумерных подтаблиц.

Кластеризованные столбчатые диаграммы в процедуре Таблицы сопряженности

Вывести кластеризованные столбчатые диаграммы. Кластеризованная столбчатая диаграмма помогает подытожить данные для групп наблюдений. Каждому значению переменной, заданному в списке Строки, соответствует кластер столбцов диаграммы. Переменной, которая формирует столбцы в кластерах, является переменная, задаваемая в списке Столбцы. Каждому значению этой переменной соответствуют окрашенные одним цветом или одинаково заштрихованные столбцы диаграммы. Если в списках Строки или Столбцы задано более одной переменной, то кластеризованная столбчатая диаграмма строится для каждой комбинации переменных из этих двух списков.

Таблицы сопряженности, выводящие переменные слоев в слоях таблицы

Вывод переменных в слоях таблиц Можно задать вывод переменных слоев (управляющих переменных) в качестве переменных слоев в таблице сопряженности. Это дает возможность представлять таблицы таким образом, чтобы статистики выводились для переменных строк и столбцов, и при этом их можно было бы увидеть по категориям переменных слоев.

Ниже приведен пример, использующий файл данных *demo.sav* (он доступен в подкаталоге Samples каталога установки); для работы с ним выполните следующие действия:

1. Выберите *Категория дохода домохозяйства (inccat)* в качестве переменной строки, *Наличие персонального цифрового помощника (PDA) (ownpda)* в качестве переменной столбца и *Уровень образования (ed)* в качестве переменной слоя.
2. Выберите **Выводить переменные слоев в слоях таблицы**.
3. В диалоговом окне Вывод в ячейках выберите **По столбцу**.
4. Запустите процедуру Таблицы сопряженности, дважды щелкните по таблице сопряженности, и в раскрывающемся списке Уровень образования выберите **Высшее**.

В выбранном представлении таблицы сопряженности можно увидеть статистики для респондентов с высшим образованием.

Статистики, рассчитываемые для таблиц сопряженности

Хи-квадрат. Отметьте **Хи-квадрат** , чтобы получить значения критериев хи-квадрат Пирсона, хи-квадрат отношения правдоподобия, точного критерия Фишера и критерия хи-квадрат с поправкой Йетса (с поправкой на непрерывность) для таблиц, образованных двумя строками и двумя столбцами. Для таблиц 2 × 2: критерий Фишера вычисляется в том случае, когда таблица, которая не является результатом наличия пропущенных строк или столбцов в таблице большего размера, имеет ожидаемое значение меньше 5 хотя бы в одной ячейке. Для всех остальных таблиц размерности 2 × 2 рассчитывается критерий хи-квадрат с поправкой Йетса. Для таблиц с любым числом строк и столбцов отметьте **Хи-квадрат** , чтобы вывести значения хи-квадрата Пирсона и хи-квадрат отношения правдоподобия. Если обе переменные в таблице являются количественными, то при пометке элемента **Хи-квадрат** рассчитывается критерий линейно-линейной связи.

Корреляции. Для таблиц с упорядоченными переменными по строкам и столбцам при пометке элемента **Корреляции** вычисляются значения коэффициента корреляции Спирмана - ро (только для числовых данных). ро Спирмана является мерой связи между порядковыми переменными. Если обе переменные в таблице (факторы) являются числовыми, параметр **Корреляции** позволяет вычислить коэффициент корреляции Пирсона *r* , который характеризует силу линейной связи между переменными.

Номинальные. Для номинальных данных (которые не имеют естественного порядка - например, католическое, протестантское, иудейское вероисповедание) можно выбрать одну из следующих статистик: **Коэффициент сопряженности**, **Фи**(коэффициент) **и V Крамера**, **Лямбда** (симметричное и асимметричное значения лямбда, статистика и тау Гудмана и Краскала), а также **Коэффициент неопределенности**.

- *Коэффициент сопряженности.* Мера связи, основанная на хи-квадрат. Это значение меняется между 0 и 1, причем 0 означает отсутствие связи между переменными строки и столбца, а значение, близкое к 1, - высокую степень связи между этими переменными. Максимально возможное значение зависит от числа строк и столбцов в таблице.
- *Фи и V Крамера.* Мера связи, вычисляется делением статистики хи-квадрат на объем выборки и взятием корня квадратного из результата. V Крамера - это мера связи, основанная на статистике хи-квадрат.
- *Лямбда* . Мера связи, которая отражает относительное снижение ошибки, когда значения независимой переменной используются для предсказания значений зависимой переменной. Значение 1 означает, что независимая переменная точно предсказывает значения зависимой. Значение 0 означает, что независимая переменная абсолютно бесполезна для предсказания зависимой.
- *Коэффициент неопределенности.* Мера связи, указывающая относительное снижение ошибки в случае, когда значения одной переменной используются для предсказания значений другой. Например, значение 0.83 указывает на то, что знание одной переменной уменьшает ошибку в предсказании значений другой на 83%. Вычисляются как симметричная, так и несимметричная версии коэффициента неопределенности.

Порядковые. Для таблиц, в которых как строки, так и столбцы содержат упорядоченные значения, пометьте **Гамма** (нулевого порядка для двумерных таблиц и условное для таблиц размерности от 2 до 10), **тау-b Кендалла** и **тау-c Кендалла** . Для предсказания категорий столбца по категориям строки, пометьте **d Сомерса** .

- *Гамма.* Симметричная мера связи между двумя порядковыми переменными, значения которой меняются между -1 и 1. Значения, близкие по абсолютной величине к 1, указывают на сильную связь переменных. Значения, близкие к 0, говорят о слабой связи или ее отсутствии. Для таблиц сопряженности двух переменных вычисляется гамма нулевого порядка. Если же таблица сопряженности включает более двух переменных, для каждой подтаблицы вычисляется условная гамма.
- *d Сомерса.* Мера связи между двумя порядковыми переменными, изменяется между -1 и 1. Значения, близкие по абсолютной величине к 1, указывают на сильную связь между двумя переменными, а значения, близкие к 0, - на слабую связь или ее отсутствие. Это асимметричное расширение меры гамма, отличающееся только включением числа пар, не имеющих совпадений (связей) по независимой переменной. Вычисляется также симметричная версия этой статистики.
- *Тау-b Кендалла.* Непараметрическая мера корреляции для порядковых или ранговых переменных, которая учитывает возможные совпадения значений (связи). Знак коэффициента указывает направление

связи, а его модуль - силу связи, причем, чем он больше, тем связь сильнее. Значения изменяются в диапазоне между -1 и +1, однако -1 и +1 можно получить только для квадратных таблиц.

- *Тау-с Кендалла*. Непараметрическая мера связи для порядковых переменных, игнорирующая возможные совпадения значений (связи). Знак коэффициента указывает направление связи, а его модуль - силу связи, причем, чем он больше, тем связь сильнее. Значения изменяются в диапазоне между -1 и +1, однако -1 и +1 можно получить только для квадратных таблиц.

Номин./интерв. В ситуации, когда одна из переменных категориальная , а другая - количественная, выберите статистику **Эта** . Значения категориальной переменной должны быть закодированы числами.

- *Эта*. Мера связи между переменными строки и столбца, значения которой изменяются от 0 (отсутствие связи) до 1 (сильная связь). Индикатор Эта подходит для зависимой переменной, измеренной в интервальной шкале (такой, как доход) и независимой переменной с ограниченным числом категорий (такой, как возраст). Вычисляется два значения для эта: в одном случае переменная строк считается переменной интервала, а в другом переменная интервала - это переменная столбцов.

Каппа. Каппа Коэна измеряет согласие мнений двух экспертов, оценивающих одни и те же объекты. Значение 1 указывает на полное согласие. Значение 0 указывает на то, что согласие - не более чем случайность. Каппа основывается на квадратной таблице, в которой значения строк и столбцов измерены в одной и той же шкале. Любой ячейки, которая имеет наблюденные значения для одной переменной, но не имеет для другой, присваивается количество, равное 0. Каппа не вычисляется, если тип хранения данных (текстовый или числовой) не одинаков для обеих переменных. Для текстовых переменных, обе переменные должны иметь одинаковую заданную длину.

Риск. Для таблиц 2 x 2 мера силы связи между присутствием фактора и возникновением события. Если доверительный интервал для этой статистики включает 1, предположение о том, что фактор связан с событием, будет неверным. Если наличие фактора встречается редко, то в качестве оценки относительного риска можно использовать отношение шансов.

Макнемара. Непараметрический критерий для двух связанных дихотомических переменных. Проверяет изменения в откликах с помощью распределения хи-квадрат. Полезен для выявления изменений в откликах, обусловленных экспериментальным вмешательством в планах до-и-после. Для больших квадратных таблиц выдаются результаты критерия симметричности Макнемара - Боукера.

Статистики Кокрена и Мантеля-Хенцеля. Статистики Кокрена и Мантеля-Хенцеля могут использоваться для проверки условной независимости дихотомической факторной переменной и дихотомической переменной отклика при заданных ковариационных структурах, задаваемых одной или большим числом переменных слоя (управляющих переменных). Заметим, что в то время как другие статистики вычисляются послойно, статистики Кокрена и Мантеля-Хенцеля вычисляются сразу для всех слоев.

Вывод в ячейках для таблиц сопряженности

Чтобы помочь вам выявить структуры в данных, которые могут повлиять на результаты критерия хи-квадрат, процедура Таблицы сопряженности выводит ожидаемые значения частот и три типа остатков (отклонений), которые выступают как меры различия между ожидаемыми и наблюденными частотами. Каждая ячейка таблицы может содержать любую комбинацию выбранных количеств, процентов и остатков.

Количества. Число фактически наблюденных наблюдений и число наблюдений, ожидаемое при условии независимости переменных в строках и в столбцах. Можно выбрать не показывать частоты, которые меньше заданного целого. Скрытые значения будут выводиться как **<N**, где **N** - заданное целое. Заданное целое должно быть больше или равно 2, однако допускается значение 0, которое говорит о том, что скрытые количества отсутствуют.

Сравнить пропорции столбцов При выборе этого параметра выполняются попарные сравнения пропорций столбцов и указывается, какие пары столбцов (для данной строки) значимо различаются. Значимые различия в таблице сопряженности указываются с применением APA-стиля форматирования и

использованием букв подстрочного индекса, и вычисляются на уровне значимости 0,05. *Примечание*: Если данный параметр задан без выбора для вывода наблюденных количеств или процентов по столбцам, то наблюденные количества включаются в таблицу сопряженности с индексами в стиле APA, указывающими результаты применения критерия для сравнения пропорций столбцов.

- **Скорректировать p-значения (метод Бонферрони).** При попарных сравнениях пропорций столбцов используется коррекция Бонферрони, которая корректирует наблюденные уровни значимости, учитывая, что выполняются несколько сравнений.

Проценты. Проценты могут суммироваться по строкам и по столбцам. Также доступны проценты от общего числа наблюдений в таблице (один слой). *Примечание*: Если в группе Количества задать **Скрывать малые количества**, проценты, связанные со скрытыми количествами, тоже будут скрыты.

Остатки. Обычные нестандартизованные остатки вычисляются как разность между наблюденными и ожидаемыми значениями. Можно также получить значения стандартизованных и скорректированных стандартизованных остатков.

- *Нестандартизованные*. Разность между наблюдаемым и ожидаемым значениями. Ожидаемое значение - это количество наблюдений в ячейке при условии независимости переменных строки и столбца. Положительное значение остатка указывает на то, что в ячейке имеется больше наблюдений, чем в случае, если бы переменные строки и столбца были бы независимыми.
- *Стандартизованные*. Остаток, деленный на оценку его стандартного отклонения. Стандартизованные остатки, известные еще как пирсоновские, имеют среднее 0 и стандартное отклонение 1.
- *Скорректированные стандартизованные*. Остаток в некоторой ячейке (наблюдение минус ожидаемое значение), деленный на оценку его стандартной ошибки. Полученный стандартизованный остаток выражается в единицах стандартных отклонений выше или ниже среднего.

Нецелочисленные веса. Частоты в ячейках обычно являются целыми значениями, поскольку они представляют числа наблюдений в каждой ячейке. Но если наблюдения в файле данных взвешены с помощью переменной веса с нецелочисленными значениями (например, 1.25), то количества в ячейках могут также быть дробными. Округление и усечение можно применять как до, так и после вычислений количеств в ячейках, а также использовать дробные количества в ячейках как для вывода в таблицах, так и для вычисления статистик.

- *Округлять количества в ячейках*. Веса наблюдений используются как есть, но накопленные веса в ячейках перед вычислением любых статистик округляются.
- *Усекать количества в ячейках*. Веса наблюдений используются как есть, но накопленные веса в ячейках перед вычислением любых статистик усекаются.
- *Округлять веса наблюдений*. Перед применением веса наблюдений округляются.
- *Усекать веса значений*. Перед применением веса наблюдений урезаются.
- *Не корректировать*. Веса наблюдений используются как есть, также используются дробные частоты в ячейках. Однако когда запрашиваются Exact Statistics (доступные только при установке модуля Exact Tests), накопленные веса в ячейках перед вычислением статистик точных критериев либо усекаются, либо округляются.

Формат таблиц сопряженности

Вы можете расположить строки в порядке возрастания или убывания значений переменной строки.

Глава 6. Суммировать

Процедура Подытожить наблюдения вычисляет значения статистик для переменных по подгруппам, задаваемым категориями одной или нескольких группирующих переменных. Все уровни группирующей переменной представляются в таблице сопряженности. Вы можете выбрать порядок, в котором будут выводиться значения статистик. Выводятся также итожащие статистики для каждой переменной по всем категориям. Можно включить или выключить вывод списка значений данных в каждой категории. При работе с большими наборами данных вы можете выводить в списке только *n* первых наблюдений.

Пример. Каков средний объем одной продажи продукта по регионам и типам клиентов? Вы можете заметить, что средний объем одной продажи несколько выше в западном регионе, чем в других регионах, причем корпоративные клиенты в западном регионе обеспечивают наивысший средний объем одной продажи.

Статистика. Сумма, число наблюдений, среднее значение, медиана, групповая медиана, стандартная ошибка среднего значения, минимальное и максимальное значения, размах, значение группирующей переменной для первой категории, значение группирующей переменной для последней категории, стандартное отклонение, дисперсия, эксцесс, стандартная ошибка эксцесса, асимметрия, стандартная ошибка асимметрии, процент от общей суммы, процент от общего *N* , процент от суммы в, процент от *N* в, геометрическое среднее, гармоническое среднее.

Данные для процедуры Подытожить наблюдения

Данные. В качестве группирующих переменных используются категориальные переменные, значения которых могут быть числовыми или строковыми. Количество категорий должно быть разумно малым. Необходимо, чтобы остальные переменные могли быть упорядочены.

Допущения. Некоторые статистики для подгрупп, например, среднее и стандартное отклонение, основаны на теории нормального распределения и подходят для количественных переменных с симметричными распределениями. Робастные статистики (такие, как медиана и диапазон) подходят для количественных переменных, которые могут не удовлетворять предположению о нормальности.

Как получить итожащие статистики по наблюдениям

1. Выберите в меню:

   Анализ > Отчеты > Итожащие статистики...
2. Выберите одну или несколько переменных.

Дополнительно вы можете:

- Выбрать одну или несколько группирующих переменных, чтобы разделять ваши данные на подгруппы.
- Щелкнуть мышью по кнопке **Параметры** , чтобы изменить название отчета, добавить подпись под выведенными результатами или исключить наблюдения с пропущенными значениями.
- Щелкнуть мышью по кнопке **Статистики** , чтобы выбрать дополнительные статистики.
- Пометить переключателем пункт **Выводить наблюдения** , чтобы вывести список наблюдений в каждой подгруппе. По умолчанию система показывает в списке только первые 100 наблюдений из файла. Вы можете увеличить или уменьшить эту величину с помощью пункта **Ограничиться первыми** *n*, а также выключить этот переключатель для этого пункта, в результате чего в списке будут представлены все наблюдения.

Параметры процедуры Подытожить наблюдения

В процедуре Подытожить наблюдения можно изменить заголовок отчета или добавить подпись, которая будет выведена под таблицей вывода. Можно управлять переходом на следующую строку в заголовках и подписях, вводя \n там, где вы хотите разорвать строку.

Вы можете также выбрать или отменить вывод подзаголовков для итогов, а также управлять исключением и включением наблюдений с пропущенными значениями для любой из переменных, используемых в анализе. Часто оказывается желательным при выводе результатов отмечать пропущенные значения точками или звездочками. Можно ввести символ, фразу или код, которые будут появляться на месте пропущенных значений. Если этого не сделать, то пропущенные значения не будут учитываться специальным образом в выводе.

Статистики процедуры Подытожить наблюдения

Можно выбрать одну или несколько из следующих статистик для подгрупп, рассчитываемых для переменных внутри внутри каждой отдельной категории каждой группирующей переменной: сумма, число наблюдений, среднее значение, медиана, медиана группы, среднеквадратическая ошибка среднего значения, минимальное и максимальное значения, диапазон, значение группирующей переменной для первой категории, значение группирующей переменной для последней категории, стандартное отклонение, дисперсия, эксцесс, стандартная ошибка эксцесса, асимметрия, стандартная ошибка асимметрии, процент от общей суммы, процент от общего *N*, процент от суммы в, процент от *N* в, среднее геометрическое, среднее гармоническое. В выводе статистики располагаются в том порядке, в котором они указаны в списке Статистики в ячейках. Итожащие статистики также выводятся для каждой переменной по всем категориям.

Первое. Выводит первое значение данных, встреченное в файле данных.

Геометрическое среднее. Корень n-й степени из произведения n значений наблюдений.

Группированная медиана. Медианы, вычисленные для данных, закодированных по принадлежности к группам. Например, для данных о возрасте каждое значение для 30-летних кодируется как 35, каждое значение для 40-летних кодируется как 45 и т.д.; групповая медиана - это медиана, вычисленная по закодированным данным.

Гармоническое среднее. Используется для оценки среднего объема группы, когда объемы выборок в группах различаются. Гармоническое среднее - это общее число выборок, деленное на сумму величин, обратных объемам отдельных групп.

Эксцесс. Мера сгруппированности наблюдений вокруг центральной точки. Для нормального распределения значение эксцесса равно 0. Положительный эксцесс указывает на то, что по отношению к нормальному распределению наблюдения для таких распределений сгруппированы более плотно около центра и имеют более тонкие хвосты до экстремумов распределения, и более толстые хвосты в области экстремальных значений. Отрицательный эксцесс указывает на то, что по отношению к нормальному распределению наблюдения для таких распределений сгруппированы менее плотно около центра и имеют более толстые хвосты до экстремумов распределения, и более тонкие хвосты в области экстремальных значений.

Последнее. Выводит последнее значение в файле данных.

Максимум. Наибольшее значение числовой переменной.

Mean. Мера центральной тенденции. Арифметическое среднее; сумма, деленная на число наблюдений.

Медиана. Значение, выше и ниже которого попадает по половине наблюдений, иначе 50-й процентиль. Если число наблюдений четно, медиана есть арифметическое среднее двух находящихся в середине значений, если выборку упорядочить по убыванию или по возрастанию. Медиана представляет собой меру центральной

тенденции, которая нечувствительна к выбросам, в отличие от среднего значения, которое могут исказить несколько экстремально больших или малых значений.

Минимум. Наименьшее значение числовой переменной.

N. Число случаев (наблюдений или записей).

Процент от общего N. Процент от общего количества наблюдений в каждой категории.

Процент от общей суммы. Процент от общей суммы в каждой категории.

Range . Разность между наибольшим и наименьшим значениями числовой переменной; максимум минус минимум.

Асимметрия. Мера асимметрии распределения. Нормальное распределение симметрично, и для него асимметрия равна 0. Распределение со значимой положительной асимметрией имеет длинный хвост справа. Распределение со значимой отрицательной асимметрией имеет длинный хвост слева. В качестве грубого правила можно сказать, что значение асимметрии, более чем вдвое превышающее ее стандартную ошибку, указывает на наличие асимметрии распределения.

Стандартное отклонение. Мера дисперсии вокруг среднего, выраженная в тех же единицах измерения, что и наблюдения. Равна корню квадратному из дисперсии. При нормальном распределении 68% наблюдений укладываются в одно стандартное отклонение от среднего, и 95% - в два стандартных отклонения. Если, например, средний возраст равен 45 годам со стандартным отклонением 10, то 95% наблюдений должны оказаться между 25 и 65 годами при нормальном распределении.

Стандартная ошибка эксцесса . Отношение эксцесса к его стандартной ошибке можно использовать как критерий нормальности (то есть, можно отвергнуть нормальность, если это отношение меньше, чем -2, или больше, чем +2). Большое положительное значение эксцесса указывает, что хвосты распределения длиннее, чем у нормального; отрицательное значение эксцесса указывает на более короткие хвосты (как у равномерного распределения).

Стандартная ошибка среднего. Мера того, как сильно могут отличаться значения среднего от выборки к выборке, извлекаемых из одного и того же распределения. Можно применять для грубого сравнения наблюденного среднего с гипотетическим значением (то есть можно заключить, что два значения различаются, если отношение их разности к стандартному отклонению меньше -2 или больше +2).

Стандартная ошибка асимметрии . Отношение асимметрии к ее стандартной ошибке можно использовать как критерий нормальности (то есть, можно отвергнуть нормальность, если это отношение меньше, чем -2, или больше, чем +2). Большое положительное значение асимметрии указывает на длинный правый хвост (распределения); большое отрицательное значение - на длинный левый хвост.

Sum. Сумма или итог для всех значений по всем наблюдениям, имеющим непропущенные значения.

Дисперсия. Мера дисперсии относительно среднего значения. Равна сумме квадратов отклонений от среднего, деленной на число, на единицу меньшее числа наблюдений. Дисперсия измеряется в единицах, которые равны квадратам единиц измерения самой переменной.

Глава 7. Средние

Процедура Средние вычисляет средние значения для подгрупп и связанные с ними одномерные статистики для зависимых переменных внутри категорий одной или нескольких независимых переменных. Дополнительно вы можете провести однофакторный дисперсионный анализ, найти значения статистики эта (eta), а также выполнить тесты на линейность.

Пример. Измерим среднее поглощаемое количество жира для каждого из трех типов кулинарного жира, и проведем однофакторный дисперсионный анализ для проверки, различаются ли эти средние значения.

Статистика. Сумма, число наблюдений, среднее значение, медиана, групповая медиана, стандартная ошибка среднего значения, минимальное и максимальное значения, размах, значение группирующей переменной для первой категории, значение группирующей переменной для последней категории, стандартное отклонение, дисперсия, эксцесс, стандартная ошибка эксцесса, асимметрия, стандартная ошибка асимметрии, процент от общей суммы, процент от общего N , процент от суммы в, процент от N в, геометрическое среднее, гармоническое среднее. Дополнительные статистики включают дисперсионный анализ, значения эта (eta) и эта квадрат, а также критерий линейности, R и R^2 .

Данные для процедуры Средние

Данные. Зависимые переменные - количественные, независимые переменные - категориальные. Значения группирующих переменных могут быть числовыми и текстовыми.

Допущения. Некоторые статистики для подгрупп, например, среднее и стандартное отклонение, основаны на теории нормального распределения и подходят для количественных переменных с симметричными распределениями. Робастные статистики, такие как медиана, годятся и для количественных переменных, которые могут не удовлетворять условию нормальной распределенности. Дисперсионный анализ является робастным в отношении отклонений от нормальности, однако данные в каждой ячейке должны быть симметричными. При проведении дисперсионного анализа предполагается, что группы принадлежат совокупностям с одинаковыми дисперсиями. Для проверки этого предположения используйте критерий однородности дисперсии Ливиня, который выполняется в процедуре Однофакторный дисперсионный анализ.

Как выполнить процедуру Средние

1. Выберите в меню:

   Анализ > Сравнение средних > Средние...
2. Выберите одну или несколько зависимых переменных.
3. Используйте один из следующих методов для выбора категориальных независимых переменных:
   - Выберите одну или несколько независимых переменных. Для каждой независимой переменной результаты будут выведены отдельно.
   - Выберите один или несколько слоев независимых переменных. Каждый слой в дальнейшем делит выборку на подгруппы. Если одна из независимых переменных находится в слое 1, а вторая - в слое 2, то результаты будут выведены в одной таблице сопряженности, а не в отдельных таблицах для каждой независимой переменной.
4. Кроме того, можно щелкнуть **Параметры** для получения дополнительных статистических данных, таблицы дисперсионного анализа, значения эта (eta), эта квадрат, R и R^2 .

Параметры процедуры Средние

Можно выбрать одну или несколько из следующих статистик для подгрупп, рассчитываемых для переменных внутри внутри каждой отдельной категории каждой группирующей переменной: сумма, число наблюдений, среднее значение, медиана, медиана группы, стандартная ошибка среднего значения, минимальное и максимальное значения, диапазон, значение группирующей переменной для первой категории, значение группирующей переменной для последней категории, стандартное отклонение, дисперсия, эксцесс, стандартная ошибка эксцесса, асимметрия, стандартная ошибка асимметрии, процент от общей суммы, процент от общего *N* , процент от суммы в, процент от *N* в, геометрическое среднее, гармоническое среднее. Вы можете изменить порядок, в котором выводятся статистики подгрупп. Порядок, в котором статистики приведены в списке Статистики в ячейках, определяет их порядок при выводе. Итожащие статистики также выводятся для каждой переменной по всем категориям.

Первое. Выводит первое значение данных, встреченное в файле данных.

Геометрическое среднее. Корень n-й степени из произведения n значений наблюдений.

Группированная медиана. Медианы, вычисленные для данных, закодированных по принадлежности к группам. Например, для данных о возрасте каждое значение для 30-летних кодируется как 35, каждое значение для 40-летних кодируется как 45 и т.д.; групповая медиана - это медиана, вычисленная по закодированным данным.

Гармоническое среднее. Используется для оценки среднего объема группы, когда объемы выборок в группах различаются. Гармоническое среднее - это общее число выборок, деленное на сумму величин, обратных объемам отдельных групп.

Эксцесс. Мера сгруппированности наблюдений вокруг центральной точки. Для нормального распределения значение эксцесса равно 0. Положительный эксцесс указывает на то, что по отношению к нормальному распределению наблюдения для таких распределений сгруппированы более плотно около центра и имеют более тонкие хвосты до экстремумов распределения, и более толстые хвосты в области экстремальных значений. Отрицательный эксцесс указывает на то, что по отношению к нормальному распределению наблюдения для таких распределений сгруппированы менее плотно около центра и имеют более толстые хвосты до экстремумов распределения, и более тонкие хвосты в области экстремальных значений.

Последнее. Выводит последнее значение в файле данных.

Максимум. Наибольшее значение числовой переменной.

Mean. Мера центральной тенденции. Арифметическое среднее; сумма, деленная на число наблюдений.

Медиана. Значение, выше и ниже которого попадает по половине наблюдений, иначе 50-й процентиль. Если число наблюдений четно, медиана есть арифметическое среднее двух находящихся в середине значений, если выборку упорядочить по убыванию или по возрастанию. Медиана представляет собой меру центральной тенденции, которая нечувствительна к выбросам, в отличие от среднего значения, которое могут исказить несколько экстремально больших или малых значений.

Минимум. Наименьшее значение числовой переменной.

N. Число случаев (наблюдений или записей).

Процент от общего количества N. Процент от общего количества наблюдений в каждой категории.

Процент от общей суммы. Процент от общей суммы в каждой категории.

Range . Разность между наибольшим и наименьшим значениями числовой переменной; максимум минус минимум.

Асимметрия. Мера асимметрии распределения. Нормальное распределение симметрично, и для него асимметрия равна 0. Распределение со значимой положительной асимметрией имеет длинный хвост справа. Распределение со значимой отрицательной асимметрией имеет длинный хвост слева. В качестве грубого правила можно сказать, что значение асимметрии, более чем вдвое превышающее ее стандартную ошибку, указывает на наличие асимметрии распределения.

Стандартное отклонение. Мера дисперсии вокруг среднего, выраженная в тех же единицах измерения, что и наблюдения. Равна корню квадратному из дисперсии. При нормальном распределении 68% наблюдений укладываются в одно стандартное отклонение от среднего, и 95% - в два стандартных отклонения. Если, например, средний возраст равен 45 годам со стандартным отклонением 10, то 95% наблюдений должны оказаться между 25 и 65 годами при нормальном распределении.

Стандартная ошибка эксцесса . Отношение эксцесса к его стандартной ошибке можно использовать как критерий нормальности (то есть, можно отвергнуть нормальность, если это отношение меньше, чем -2, или больше, чем +2). Большое положительное значение эксцесса указывает, что хвосты распределения длиннее, чем у нормального; отрицательное значение эксцесса указывает на более короткие хвосты (как у равномерного распределения).

Стандартная ошибка среднего. Мера того, как сильно могут отличаться значения среднего от выборки к выборке, извлекаемых из одного и того же распределения. Можно применять для грубого сравнения наблюденного среднего с гипотетическим значением (то есть можно заключить, что два значения различаются, если отношение их разности к стандартному отклонению меньше -2 или больше +2).

Стандартная ошибка асимметрии . Отношение асимметрии к ее стандартной ошибке можно использовать как критерий нормальности (то есть, можно отвергнуть нормальность, если это отношение меньше, чем -2, или больше, чем +2). Большое положительное значение асимметрии указывает на длинный правый хвост (распределения); большое отрицательное значение - на длинный левый хвост.

Sum. Сумма или итог для всех значений по всем наблюдениям, имеющим непропущенные значения.

Дисперсия. Мера дисперсии относительно среднего значения. Равна сумме квадратов отклонений от среднего, деленной на число, на единицу меньшее числа наблюдений. Дисперсия измеряется в единицах, которые равны квадратам единиц измерения самой переменной.

Статистики для первого слоя

Таблица дисперсионного анализа и эта. Выводит таблицу однофакторного дисперсионного анализа и вычисляет значение эта и эта в квадрате (меры близости) для каждой независимой переменной в первом слое.

Критерий линейности. Вычисляет сумму квадратов, степени свободы и средний квадрат для линейных и нелинейных компонентов, а также F-отношение, значения R и R-квадрат. Линейность не вычисляется, если независимой объявлена короткая текстовая переменная.

Глава 8. Кубы OLAP

Процедура OLAP (Online Analytical Processing) Кубы вычисляет итоги, средние значения и другие одномерные статистики для количественных подытоживаемых переменных внутри категорий одной или нескольких категориальных группирующих переменных. Для каждой категории каждой группирующей переменной в таблице создается отдельный слой.

Пример. Суммарные продажи и средние объемы одной продажи для разных регионов и видов товаров внутри регионов.

Статистика. Сумма, число наблюдений, среднее значение, медиана, групповая медиана, стандартная ошибка среднего, минимум, максимум, размах, значение переменной для первой категории группирующей переменной, значение переменной для последней категории группирующей переменной, стандартное отклонение, дисперсия, эксцесс, стандартная ошибка эксцесса, асимметрия, стандартная ошибка асимметрии, процент от общего количества наблюдений, процент общей суммы, процент общего количества наблюдений в категориях группирующих переменных, процент общей суммы в категориях группирующих переменных, геометрическое среднее, гармоническое среднее.

Данные для процедуры OLAP Кубы

Данные. Подытоживаемые переменные являются количественными (непрерывными переменными, измеренными в интервальной шкале или шкале отношений), а группирующие переменные являются категориальными. Значения группирующих переменных могут быть числовыми и текстовыми.

Допущения. Некоторые статистики для подгрупп, например, среднее и стандартное отклонение, основаны на теории нормального распределения и подходят для количественных переменных с симметричными распределениями. Робастные статистики, такие как медиана и диапазон, годятся и для количественных переменных, которые могут не удовлетворять условию нормальной распределенности.

Как получить OLAP Кубы

1. Выберите в меню:

   Анализ > Отчеты > Кубы OLAP...
2. Выберите одну или несколько количественных подытоживаемых переменных.
3. Выберите одну или несколько категориальных группирующих переменных.

Дополнительно можно:

- Выбрать различные итожащие статистики (нажмите кнопку **Статистики**). Перед выбором статистик необходимо задать одну или более группирующих переменных.
- Вычислить разности между парами переменных и парами групп, заданных группирующей переменной (щелкните по **Разности**).
- Создать и отредактировать заголовки (нажмите кнопку **Заголовок**).
- Скрыть количества, меньшие заданного целого. Скрытые значения будут выводиться как **<N**, где **N** - заданное целое. Заданное целое должно быть больше или равно 2.

Статистики в процедуре OLAP Кубы

Можно выбрать одну или несколько из следующих статистик для подгрупп, рассчитываемых для итоговых переменных внутри внутри каждой отдельной категории каждой группирующей переменной: сумма, число наблюдений, среднее значение, медиана, медиана группы, среднеквадратическая ошибка среднего значения, минимальное и максимальное значения, диапазон, значение группирующей переменной для первой категории, значение группирующей переменной для последней категории, стандартное отклонение,

дисперсия, эксцесс, среднеквадратическая ошибка эксцесса, асимметрия, среднеквадратическая ошибка асимметрии, процент от всех наблюдений в группирующих переменных, процент от общей сумы в группирующих переменных, среднее геометрическое и среднее гармоническое.

Вы можете изменить порядок, в котором выводятся статистики подгрупп. Порядок, в котором статистики приведены в списке Статистики в ячейках, определяет их порядок при выводе. Итожащие статистики также выводятся для каждой переменной по всем категориям.

Первое. Выводит первое значение данных, встреченное в файле данных.

Геометрическое среднее. Корень n-й степени из произведения n значений наблюдений.

Группированная медиана. Медианы, вычисленные для данных, закодированных по принадлежности к группам. Например, для данных о возрасте каждое значение для 30-летних кодируется как 35, каждое значение для 40-летних кодируется как 45 и т.д.; групповая медиана - это медиана, вычисленная по закодированным данным.

Гармоническое среднее. Используется для оценки среднего объема группы, когда объемы выборок в группах различаются. Гармоническое среднее - это общее число выборок, деленное на сумму величин, обратных объемам отдельных групп.

Эксцесс. Мера сгруппированности наблюдений вокруг центральной точки. Для нормального распределения значение эксцесса равно 0. Положительный эксцесс указывает на то, что по отношению к нормальному распределению наблюдения для таких распределений сгруппированы более плотно около центра и имеют более тонкие хвосты до экстремумов распределения, и более толстые хвосты в области экстремальных значений. Отрицательный эксцесс указывает на то, что по отношению к нормальному распределению наблюдения для таких распределений сгруппированы менее плотно около центра и имеют более толстые хвосты до экстремумов распределения, и более тонкие хвосты в области экстремальных значений.

Последнее. Выводит последнее значение в файле данных.

Максимум. Наибольшее значение числовой переменной.

Mean. Мера центральной тенденции. Арифметическое среднее; сумма, деленная на число наблюдений.

Медиана. Значение, выше и ниже которого попадает по половине наблюдений, иначе 50-й процентиль. Если число наблюдений четно, медиана есть арифметическое среднее двух находящихся в середине значений, если выборку упорядочить по убыванию или по возрастанию. Медиана представляет собой меру центральной тенденции, которая нечувствительна к выбросам, в отличие от среднего значения, которое могут исказить несколько экстремально больших или малых значений.

Минимум. Наименьшее значение числовой переменной.

N. Число случаев (наблюдений или записей).

Процентная доля N в. Процент от количества наблюдений для указанной группирующей переменной внутри категорий другой группирующей переменной. Если имеется только одна группирующая переменная, это значение совпадает с процентом от общего числа наблюдений.

Процент от суммы в. Процент от суммы для указанной группирующей переменной внутри категорий другой группирующей переменной. Если имеется только одна группирующая переменная, это значение совпадает с процентом от общей суммы.

Процент от общего N. Процент от общего количества наблюдений в каждой категории.

Процент от общей суммы. Процент от общей суммы в каждой категории.

Range . Разность между наибольшим и наименьшим значениями числовой переменной; максимум минус минимум.

Асимметрия. Мера асимметрии распределения. Нормальное распределение симметрично, и для него асимметрия равна 0. Распределение со значимой положительной асимметрией имеет длинный хвост справа. Распределение со значимой отрицательной асимметрией имеет длинный хвост слева. В качестве грубого правила можно сказать, что значение асимметрии, более чем вдвое превышающее ее стандартную ошибку, указывает на наличие асимметрии распределения.

Стандартное отклонение. Мера дисперсии вокруг среднего, выраженная в тех же единицах измерения, что и наблюдения. Равна корню квадратному из дисперсии. При нормальном распределении 68% наблюдений укладываются в одно стандартное отклонение от среднего, и 95% - в два стандартных отклонения. Если, например, средний возраст равен 45 годам со стандартным отклонением 10, то 95% наблюдений должны оказаться между 25 и 65 годами при нормальном распределении.

Стандартная ошибка эксцесса . Отношение эксцесса к его стандартной ошибке можно использовать как критерий нормальности (то есть, можно отвергнуть нормальность, если это отношение меньше, чем -2, или больше, чем +2). Большое положительное значение эксцесса указывает, что хвосты распределения длиннее, чем у нормального; отрицательное значение эксцесса указывает на более короткие хвосты (как у равномерного распределения).

Стандартная ошибка среднего. Мера того, как сильно могут отличаться значения среднего от выборки к выборке, извлекаемых из одного и того же распределения. Можно применять для грубого сравнения наблюденного среднего с гипотетическим значением (то есть можно заключить, что два значения различаются, если отношение их разности к стандартному отклонению меньше -2 или больше +2).

Стандартная ошибка асимметрии . Отношение асимметрии к ее стандартной ошибке можно использовать как критерий нормальности (то есть, можно отвергнуть нормальность, если это отношение меньше, чем -2, или больше, чем +2). Большое положительное значение асимметрии указывает на длинный правый хвост (распределения); большое отрицательное значение - на длинный левый хвост.

Sum. Сумма или итог для всех значений по всем наблюдениям, имеющим непропущенные значения.

Дисперсия. Мера дисперсии относительно среднего значения. Равна сумме квадратов отклонений от среднего, деленной на число, на единицу меньшее числа наблюдений. Дисперсия измеряется в единицах, которые равны квадратам единиц измерения самой переменной.

OLAP Кубы: Разности

Это диалоговое окно позволяет вычислять разности в процентах и арифметические разности между подытоживаемыми переменными или между группами, задаваемыми группирующей переменной. Разности вычисляются для всех мер, выбранных в диалоговом окне OLAP Кубы: Статистики

Разность между переменными. Вычисляет разности между парами переменных. В каждой паре значения итожащих статистик для второй переменной (Минус переменная) вычитаются из значений итожащих статистик для первой переменной. Для разностей в процентах значение подытоживаемой переменной для Минус переменной используется в качестве знаменателя. Перед тем как задать разности между переменными, в главном диалоговом окне необходимо выбрать, по крайней мере, две подытоживаемые переменные.

Разность между группами наблюдений. Вычисляет разности между парой групп, заданной группирующей переменной. В каждой паре значения итожащих статистик для второй категории (Минус категория) вычитаются из значений итожащих статистик для первой категории. Разности в процентах используют значение итожащей статистики для Минус категории в качестве знаменателя. Перед тем как задать разности между группами, в главном диалоговом окне необходимо выбрать одну или несколько группирующих переменных.

OLAP Кубы: Заголовок

Вы можете изменить заголовок вывода или добавить подпись, которая появится ниже выведенной таблицы. Можно управлять переходом на следующую строку в заголовках и подписях, вводя `\n` там, где вы хотите разорвать строку.

Глава 9. Т-критерии

Т-критерии

Доступны *t* -критерии трех типов:

Т-критерий для независимых выборок (двухвыборочный t-критерий). Сравнивает средние значения одной переменной для двух групп наблюдений. Выдаются описательные статистики для каждой группы и критерий равенства дисперсий Ливиня, а также значения *t* как для предположительно равных, так и для предположительно неравных дисперсий, а также 95%-й доверительный интервал для разности средних значений.

Т-критерий для парных выборок (зависимый t-критерий). Сравнивает средние значения двух разных переменных для одной группы наблюдений. Этот критерий предназначен также для пар сочетаемых индивидуумов или планов исследования типа "случай-контроль". Выводятся описательные статистики для проверяемых переменных, корреляция между ними, описательные статистики для парных разностей, *t* -критерий и 95%-й доверительный интервал.

Одновыборочный t-критерий. Сравнивает среднее значение одной переменной с известным или гипотетическим значением. Помимо *t* -критерия, выдаются описательные статистики для проверяемых переменных. По умолчанию выдается 95%-й доверительный интервал для разности между средним значением проверяемой переменной и гипотетическим проверяемым значением.

Т-критерий для независимых выборок

Процедура Т-критерий для независимых выборок сравнивает средние значения для двух групп наблюдений. В идеале объекты для этого критерия должны быть случайным образом приписаны двум группам, чтобы любое различие в отклике определялось рассматриваемым воздействием, например лечением, (или его отсутствием), а не другими факторами. Это не выполняется, если вы сравниваете средний доход для мужчин и женщин. Пол не приписывается индивидууму случайным образом. В подобных ситуациях следует убедиться, что различия в других факторах не снижают и увеличивают значимые различия средних значений. На различие средних доходов может оказывать влияние такой фактор, как образование, а не только пол.

Пример. Пациенты с высоким давлением случайным образом делятся на контрольную группу и группу испытуемых. Пациенты в контрольной группе получают плацебо (фармакологически неактивные таблетки), а пациенты в группе испытуемых получают лекарство (исследуемые таблетки, которые предположительно понижают давление). Пациенты наблюдаются в течение двух месяцев, после чего для сравнения средних значений кровяного давления пациентов контрольной группы и группы испытуемых применяют двухвыборочный *t* -критерий. Давление каждого пациента измеряют один раз, и каждый пациент принадлежит только к одной группе.

Статистика. Для каждой переменной: размер выборки, среднее значение, среднеквадратичное отклонение и среднеквадратическая ошибка среднего значения. Для разности средних: среднее значение, среднеквадратическая ошибка и доверительный интервал (можно задать доверительный уровень). Критерии: Критерий равенства дисперсий Ливиня, а также *t*-критерий равенства средних как для объединенной, так и для раздельной дисперсии.

Данные для Т-критерия для независимых выборок

Данные. Значения изучаемой количественной переменной находятся в одном столбце файла данных. Чтобы разбить наблюдения на две группы, в процедуре используется группирующая переменная с двумя значениями. Эта переменная может быть числовой (например, со значениями 1 и 2 или 6.25 и 12.5) или короткой текстовой (например, со значениями *да* и *нет*). Возможно также использовать количественную

переменную, такую как *возраст* , чтобы разбить наблюдения на две группы путем задания пороговой точки (пороговая точка 21 разбивает *возраст* на группы: до 21 года и 21 год или более).

Допущения. Для *t* -критерия, предполагающего равенство дисперсий, наблюдения должны быть независимыми случайными выборками из нормальных распределений с одинаковыми дисперсиями. Для *t* -критерия, не предполагающего равенство дисперсий, наблюдения должны быть независимыми случайными выборками из нормальных распределений. Двухвыборочный *t* -критерий довольно устойчив к отклонениям от нормальности. Проверяя распределения графически, следите, чтобы они были симметричными и не содержали выбросов.

Чтобы получить t-критерий для независимых выборок

1. Выберите в меню:

   Анализ > Сравнение средних > Т-критерий для независимых выборок...
2. Выберите одну или несколько количественных переменных для проверки. *T* -критерий будет применен к каждой переменной в отдельности.
3. Выберите группирующую переменную и нажмите кнопку **Задать группы** , чтобы задать два кода для определения сравниваемых групп.
4. Можно щелкнуть мышью по кнопке **Параметры** и выбрать способ работы с пропущенными значениями, а также задать уровень для доверительного интервала.

Задание групп, сравниваемых процедурой Т-критерий для независимых выборок

Для числовых группирующих переменных две группы для *t* -критерия формируются путем задания двух значений или порога:

- **Заданные значения.** Введите одно значение в поле Группа 1, а другое значение - в поле Группа 2. Наблюдения с любыми иными значениями будут исключены из анализа. Числа не обязаны быть целыми (например, вполне подходят значения 6.25 и 12.5).
- **Порог.** Введите число, разбивающее значения группирующей переменной на два множества. Все наблюдения со значениями, меньшими значения порога, составляют одну группу, а наблюдения со значениями, большими или равными значению порога, составляют другую группу.

Для строковых группирующих переменных введите строковое значение в поле Группа 1, а другое строковое значение - в поле Группа 2, например, *да* и *нет*. Наблюдения со всеми прочими строками исключаются из анализа.

Параметры процедуры Т-критерий для независимых выборок

Доверительный интервал. По умолчанию для разности средних значений выводится 95%-й доверительный интервал. Чтобы задать другой доверительный уровень, введите значение между 1 и 99.

Пропущенные значения. Когда вы проверяете несколько переменных, и некоторые из них содержат пропущенные значения, вы можете указать, какие наблюдения следует включить (или исключить).

- **Исключать из каждого анализа.** При применении *t* -критерия используются все наблюдения, в которых проверяемая переменная имеет непропущенные значения. Объемы выборок могут меняться в зависимости от переменных, к которым применяется критерий.
- **Исключать целиком.** Каждый раз при применении *t* -критерия используются только те наблюдения, которые не имеют пропущенных значений для всех переменных, для которых запрошено применение *t* -критерия. Объем выборок одинаков для всех тестов.

T-критерий для парных выборок

Процедура Т-критерий для парных выборок сравнивает средние значения переменных для одной группы наблюдений. Для всех наблюдений вычисляются разности значений двух переменных, а затем проверяется, отличается ли среднее этих разностей от нуля.

Пример. При изучении проблемы повышенного артериального давления измеряют артериальное давление всем пациентам, проводят лечение, а затем повторно измеряют давление. Таким образом, для каждого пациента измерения проводят два раза (такие измерения часто называют измерениями *до* и *после*). Альтернативным планом эксперимента для применения этого критерия является исследование пар сочетаемых индивидуумов или исследование типа "случай-контроль". При изучении кровяного давления пациенты и соответствующие контрольные субъекты могут подбираться по возрасту (75-летнему пациенту соответствует 75-летний член контрольной группы).

Статистика. Для каждой переменной: среднее значение, объем выборки, среднеквадратичное отклонение и среднеквадратическая ошибка среднего значения. Для каждой пары переменных: корреляция, разность средних значений, *t* -критерий и доверительный интервал для разности средних (доверительный уровень вы можете задать сами). Стандартное отклонение и стандартная ошибка разности средних.

Данные для Т-критерия для парных выборок

Данные. Для каждого парного теста необходимо задать две количественные переменные (измеренные в интервальной шкале или шкале отношений). При исследовании пар сочетаемых индивидуумов или исследовании типа "случай-контроль" отклики для каждого тестируемого субъекта и для соответствующего ему контрольного субъекта должны содержаться в одном наблюдении (строке) файла данных.

Допущения. Наблюдения для каждой пары должны быть получены при одинаковых условиях. Средние разности должны быть нормально распределены. Дисперсии переменных могут быть как равными, так и неравными.

Чтобы получить t-критерий для парных выборок

1. Выберите в меню:

   Анализ > Сравнение средних > Т-критерий для парных выборок...
2. Выберите одну или несколько пар переменных
3. Можно щелкнуть мышью по кнопке **Параметры** и выбрать способ работы с пропущенными значениями, а также задать уровень для доверительного интервала.

Параметры процедуры Т-критерий для парных выборок

Доверительный интервал. По умолчанию для разности средних значений выводится 95%-й доверительный интервал. Чтобы задать другой доверительный уровень, введите значение между 1 и 99.

Пропущенные значения. Когда вы проверяете несколько переменных, и некоторые из них содержат пропущенные значения, вы можете указать, какие наблюдения следует включить (или исключить):

- **Исключать из каждого анализа.** При применении *t* -критерия используются все наблюдения, в которых пара проверяемых переменных имеют непропущенные значения. Объемы выборок могут меняться в зависимости от переменных, к которым применяется критерий.
- **Исключать целиком.** При применении *t* -критерия используются только те наблюдения, которые имеют непропущенные значения для всех пар проверяемых переменных. Объем выборок одинаков для всех тестов.

Команда T-TEST: дополнительные возможности

Язык синтаксиса команд также позволяет:

- Запускать одновыборочный t-критерий и t-критерий для независимых выборок при помощи одной команды.
- При расчете t-критерия для парных выборок проверять переменную вместе с каждой из переменных в списке (при помощи подкоманды PAIRS).

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Одновыборочный T-критерий

Процедура Одновыборочный T-критерий проверяет, отличается ли среднее одной переменной от заданной константы.

Примеры. Допустим, что требуется узнать, отличается ли средний IQ группы студентов от 100. Или, например, производитель хлопьев может взять выборку пачек с производственной линии и проверить, отличается ли средний вес выборки от 1.3 фунтов при 95% доверительном уровне.

Статистика. Для каждой проверяемой переменной: среднее значение, среднеквадратичное отклонение и среднеквадратическая ошибка среднего значения. Средняя разность между каждым значением данных и гипотетической проверяемой величиной, *t* -критерий для проверки равенства этой разности нулю, доверительный интервал для этой разности (доверительный уровень вы можете задать сами).

Данные для одновыборочного t-критерия

Данные. Чтобы выполнить тест для значений количественной переменной и гипотетического проверяемого значения, выберите количественную переменную и введите гипотетическое проверяемое значение.

Допущения. Этот критерий предполагает, что данные нормально распределены; однако этот критерий довольно устойчив к отклонениям от нормальности.

Как получить одновыборочный t-критерий

1. Выберите в меню:

   Анализ > Сравнение средних > Одновыборочный t-критерий...
2. Выберите одну или несколько переменных для проверки при одном и том же гипотетическом значении.
3. Введите значение, с которым будет сравниваться каждое выборочное среднее.
4. Можно щелкнуть мышью по кнопке **Параметры** и выбрать способ работы с пропущенными значениями, а также задать уровень для доверительного интервала.

Параметры процедуры Одновыборочный T-критерий

Доверительный интервал. По умолчанию для разности среднего и гипотетического проверяемого значения выводится 95%-й доверительный интервал. Чтобы задать другой доверительный уровень, введите значение между 1 и 99.

Пропущенные значения. Когда вы проверяете несколько переменных, и некоторые из них содержат пропущенные значения, вы можете указать, какие наблюдения следует включить (или исключить).

- **Исключать из каждого анализа.** При применении *t* -критерия используются все наблюдения, в которых проверяемые переменные имеют непропущенные значения. Объемы выборок могут меняться в зависимости от переменных, к которым применяется критерий.
- **Исключать целиком.** Каждый раз при применении *t* -критерия используются только те наблюдения, которые не имеют пропущенных значений для всех переменных, для которых запрошено применение *t* -критерия. Объем выборок одинаков для всех тестов.

Команда T-TEST: дополнительные возможности

Язык синтаксиса команд также позволяет:

- Запускать одновыборочный t-критерий и t-критерий для независимых выборок при помощи одной команды.
- При расчете t-критерия для парных выборок проверять переменную вместе с каждой из переменных в списке (при помощи подкоманды PAIRS).

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Команда T-TEST: дополнительные возможности

Язык синтаксиса команд также позволяет:

- Запускать одновыборочный t-критерий и t-критерий для независимых выборок при помощи одной команды.
- При расчете t-критерия для парных выборок проверять переменную вместе с каждой из переменных в списке (при помощи подкоманды PAIRS).

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Глава 10. Однофакторный дисперсионный анализ

Процедура Однофакторный дисперсионный анализ (ANOVA) выполняет однофакторный дисперсионный анализ для количественной зависимой переменной по единственной факторной (независимой) переменной. Дисперсионный анализ используется для проверки гипотезы о равенстве нескольких средних значений, соответствующих различным группам или уровням факторной переменной. Этот метод является расширением двухвыборочного *t* -критерия.

В дополнение к выявлению наличия различий между средними значениями, Вы, возможно, захотите узнать, какие именно групповые средние значения различаются. Есть два типа критериев для сравнения средних значений: априорные контрасты и апостериорные критерии. Контрасты это критерии, которые применяются *до* проведения эксперимента, апостериорные же критерии применяются *после* проведения эксперимента. Вы можете также осуществлять проверку наличия трендов по уровням (категориям).

Пример. Пончики впитывают различное количество жира в процессе их приготовления. В эксперименте используются три типа жиров: арахисовое масло, кукурузное масло и свиное сало. Арахисовое и кукурузное масло являются ненасыщенными жирами, а топленое сало - насыщенным жиром. Выясняя, зависит ли количество расходуемого жира от типа используемого жира, можно выбрать априорный контраст, позволяющий выяснить, различаются ли количества впитывающегося жира для насыщенных и ненасыщенных жиров.

Статистика. Для каждой группы: число наблюдений, среднее значение, стандартное отклонение, стандартная ошибка среднего значения, минимум, максимум и 95%-й доверительный интервал для среднего значения. Критерий Ливиня однородности дисперсий, таблица дисперсионного анализа и робастные критерии равенства средних значений для каждой зависимой переменной, задаваемые пользователем априорные контрасты, а также апостериорные критерии диапазона и множественные сравнения: Бонферрони, Шидака, критерий Тьюки достоверно значимой разности, GT2 Гохберга, Габриэля, Даннетта, *F*-критерий Райана-Эйнота-Габриэля-Уэлша (Р-Э-Г-У *F*), критерий диапазона Райана-Эйнота-Габриэля-Уэлша (Р-Э-Г-У *Q*), Тамхейна Т2, Даннетта Т3, Геймса-Хоуэлла, Даннетта *C*, критерий множественных сравнений Дункана, Стьюдента-Ньюмена-Келса (С-Н-К), Тьюки *b*, Уоллера-Дункана, Шеффе и наименьшей значимой разности.

Данные для однофакторного дисперсионного анализа

Данные. Факторные переменные должны быть целочисленными, а зависимая переменная - количественной (измерена по крайней мере в интервальной шкале).

Допущения. Каждая группа является независимой случайной выборкой из нормального распределения. Дисперсионный анализ робастен (устойчив) к отклонениям от нормальности, однако данные должны быть симметричны. Группы должны выбираться из совокупностей с одинаковыми дисперсиями. Для проверки последнего предположения используйте критерий Ливиня однородности дисперсий.

Чтобы выполнить Однофакторный дисперсионный анализ

1. Выберите в меню:

   Анализ > Сравнение средних > Однофакторный дисперсионный анализ...
2. Выберите одну или несколько зависимых переменных.
3. Выберите одну независимую факторную переменную.

Контрасты для однофакторного дисперсионного анализа

Вы можете разделить межгрупповые суммы квадратов на трендовые компоненты или задать априорные контрасты.

Полиномиальный. Разделяет межгрупповые суммы квадратов на трендовые компоненты. Вы можете выполнить проверку на наличие тренда зависимой переменной по упорядоченным уровням факторной переменной. Например, можно проверить наличие линейного тренда (возрастающего или убывающего) заработной платы по упорядоченным уровням переменной, характеризующей служебное положение или уровень образования.

- **Степень.** Вы можете выбрать полином степени 1, 2, 3, 4 или 5.

Коэффициенты. Задаваемые пользователем априорные контрасты, которые будут проверяться при помощи *t* -критерия. Введите значение коэффициента для каждой группы (уровня, категории) факторной переменной и после ввода очередного значения щелкайте мышью по кнопке **Добавить** . Каждое новое значение будет добавлено в конец списка коэффициентов. Задать дополнительные наборы контрастов можно, щелкая по кнопке **След.** . Пользуйтесь кнопками **След.** и **Предыд.** для перехода от одного набора контрастов к другому.

Порядок ввода коэффициентов важен, так как он соответствует возрастающему порядку значений категорий факторной переменной. Первый коэффициент в списке соответствует наименьшему значению факторной переменной, а последний - наибольшему. Например, если факторная переменная имеет шесть категорий, коэффициенты –1, 0, 0, 0, 0,5 и 0,5 сопоставляют первую группу с пятой и шестой группами. В большинстве случаев сумма коэффициентов должна быть равна нулю. Наборы с ненулевой суммой также могут быть использованы, однако в этом случае появится предупреждающее сообщение.

Апостериорные критерии для однофакторного дисперсионного анализа

Установив, что различия средних значений существуют, с помощью апостериорных критериев диапазона и парных множественных сравнений вы можете выяснить, какие именно средние различаются. Критерии диапазона выявляют однородные подмножества средних, не различающихся между собой. Парные множественные сравнения проверяют разности между каждой парой средних значений и выдают матрицу, в которой звездочками обозначены групповые средние, значимо различающиеся на уровне альфа, равном 0,05.

Предполагается равенство дисперсий

Критерии Тьюки достоверно значимой разности, GT2 Гохберга, Габриэля и Шеффе являются одновременно критериями диапазона и множественных сравнений. Кроме того, доступны следующие критерии диапазона: Тьюки *b* , С-Н-К (Стьюдента-Ньюмена-Келса), Дункана, Р-Э-Г-У *F* (*F* -критерий Райана-Эйнота-Габриэля-Уэлша), Р-Э-Г-У *Q* (критерий диапазона Райана-Эйнота-Габриэля-Уэлша) и Уоллера-Дункана. Доступными критериями множественных сравнений являются: Бонферрони, Тьюки достоверно значимой разности, Шидака, Габриэля, Гохберга, Даннетта, Шеффе и НЗР (наименьшей значимой разности).

- *НЗР*. Использует t-критерии для проведения всех парных сравнений групповых средних. Поправка для уровня ошибки на множественность сравнений не делается.
- *Бонферрони*. При проведении парных сравнений групповых средних используются t-критерии, но для управления общим уровнем ошибки по уровню ошибки каждой проверки вероятность ошибочного решения делится на общее число проверок. Доверительные интервалы и уровень значимости корректируются так, чтобы учесть проводимые множественные сравнения.
- *Шидак*. Критерий множественных попарных сравнений, основанный на t-статистике. Критерий Шидака изменяет величину уровня значимости в соответствии с числом множественных сравнений и обеспечивает более узкие границы, чем критерий Бонферрони.
- *Шеффе*. Производит одновременные сравнения совместных пар для всех возможных комбинаций пар средних. Использует выборочное F-распределение. Может применяться для проверки всех возможных линейных комбинаций групповых средних, а не только для парных сравнений.
- *R-E-G-W F*. Шаговая процедура множественных сравнений Райана-Эйнота-Габриэля-Уэлша, основанная на F-критерии.
- *R-E-G-W Q*. Шаговая процедура множественных сравнений Райана-Эйнота-Габриэля-Уэлша, основанная на стьюдентизированном размахе.

- *С-Н-К*. В соответствии с критерием Стьюдента-Ньюмена-Келса выполняются все попарные сравнения средних, используя распределение стьюдентизированного размаха. Если объемы выборок одинаковы, с помощью шаговой процедуры сравнивает также пары средних в однородных подмножествах. Средние упорядочиваются по убыванию, и вначале проверяются наибольшие разности.
- *Тьюки*. Использует статистику стьюдентизированного размаха для проведения всех парных сравнений между группами. Подгоняет уровень ошибки эксперимента к уровню ошибки совокупности всех парных сравнений.
- *Критерий Тьюки-b*. Использует статистику стьюдентизированного размаха для проведения всех парных сравнений между группами. Критической статистикой служит среднее из критических статистик двух критериев: достоверно значимой разности Тьюки и Стьюдента-Ньюмена-Келса.
- *Дункан*. Выполняются парные сравнения с использованием шагового порядка сравнений, как и в критерии Стьюдента-Ньюмена-Келса, но устанавливается защитный уровень доли ошибок для набора проверок, а не для доли ошибок отдельных проверок. Основан на статистике стьюдентизированного размаха.
- *GT2 Гохберга*. Критерий множественных сравнений и размахов, использующий стьюдентизированный максимум модуля. Аналогичен критерию достоверно значимой разности Тьюки.
- *Габриэль*. Критерий парных сравнений, использующий стьюдентизированный максимум модуля, обычно более мощный, чем критерий Гохберга GT2, когда размеры ячеек не равны. Критерий Габриэля может стать либеральным, когда размеры ячеек сильно различаются.
- *Уоллер-Дункан*. Процедура множественных сравнений, основанная на t-статистике; использует байесовский подход.
- *Даннетт*. t-критерий множественных парных сравнений, который сравнивает средние по группам (уровням фактора) с одним контрольным средним. Последняя категория (уровень фактора) по умолчанию служит контрольной. Как вариант можно выбрать первую категорию. **2-х сторонний** проверяет, что среднее на любом из уровней (за исключением контрольной категории) фактора не равно среднему для контрольной категории. **<Эталона** проверяет, не окажется ли среднее на каком-либо из уровней фактора меньше, чем в контрольной категории. **> Эталон** проверяет, не окажется ли среднее на каком-либо из уровней фактора больше, чем в контрольной категории.

Равенство дисперсий не предполагается

Критерии множественных сравнений Тамхейна T2, Даннетта T3, Геймса-Хоуэлла и Даннетта *C* не требуют равенства дисперсий.

- *Тамхейна T2*. Консервативный критерий попарных сравнений на основе t-критерия. Этот критерий подходит для случаев, когда дисперсии не равны.
- *Даннетта T3*. Критерий парных сравнений, основанный на стьюдентизированном максимуме модуля. Этот критерий подходит для случаев, когда дисперсии не равны.
- *Геймс-Хоуэлл*. Критерий парных сравнений, иногда являющийся либеральным. Этот критерий подходит для случаев, когда дисперсии не равны.
- *Даннетта C*. Критерий парных сравнений, основанный на стьюдентизированном размахе. Этот критерий подходит для случаев, когда дисперсии не равны.

Примечание: Возможно, вам будет легче интерпретировать результаты расчетов апостериорных критериев, если вы выключите переключатель **Скрыть пустые строки и столбцы** в диалоговом окне Свойства таблицы (при активированной сводной таблице в меню Формат выберите **Свойства таблицы**).

Параметры процедуры Однофакторный дисперсионный анализ

Статистика. Выберите одну или несколько из следующих возможностей:

- **Описательные.** Для каждой зависимой переменной и каждой группы вычисляются: количество наблюдений, среднее значение, стандартное отклонение, стандартная ошибка среднего значения, минимум, максимум и доверительные интервалы в 95%.

- **Фиксированные и случайные эффекты.** Выводит стандартное отклонение, стандартную ошибку и доверительный интервал в 95% для модели с фиксированными эффектами, а также стандартную ошибку, доверительный интервал в 95% и оценку межкомпонентной дисперсии для модели со случайными эффектами.
- **Проверка однородности дисперсии.** Вычисляется статистика Ливиня для проверки равенства дисперсий групп. Этот критерий не требует предположения о нормальности.
- **Брауна-Форсайта.** Вычисляется статистика Брауна-Форсайта для проверки равенства дисперсий групп. Эта статистика предпочтительнее *F* -статистики в случае, когда требование равенства дисперсий не выполняется.
- **Уэлч.** Вычисляется статистика Уэлча для проверки равенства дисперсий групп. Эта статистика предпочтительнее *F* -статистики в случае, когда требование равенства дисперсий не выполняется.

График средних. Выводит график, изображающий средние подгрупп (средние для всех групп, заданных значениями факторной переменной).

Пропущенные значения. Эта группа параметров позволяет управлять обработкой пропущенных значений.

- **Исключать по отдельности.** Наблюдение с пропущенным значением зависимой или факторной переменной не используется в анализе. Не будут также использоваться наблюдения со значениями вне заданного диапазона факторной переменной.
- **Исключать целиком.** Наблюдения с пропущенными значениями для факторной переменной или для любой из зависимых переменных, в списке зависимых переменных главного диалогового окна, не рассматриваются. Если не задано несколько независимых переменных, выбор этого параметра не играет роли.

Команда ONEWAY: дополнительные возможности

Язык синтаксиса команд также позволяет:

- Рассчитывать статистики для фиксированных и случайных эффектов. Стандартное отклонение, стандартную ошибку среднего и 95%-ный доверительные интервалы для моделей с фиксированными эффектами. Стандартную ошибку, 95%-ные доверительные интервалы и оценку межкомпонентной дисперсии для моделей со случайными эффектами (при помощи `STATISTICS=EFFECTS`).
- Задавать альфа-уровни для наименьшей значимой разности, критерием множественных сравнений Бонферрони, Дункана, Шеффе (при помощи подкоманды `RANGES`).
- Записывать матрицы средних значений, стандартных отклонений и частот, а также считывать матрицы средних значений, частот, объединенных дисперсий, и степеней свободы для объединенных дисперсий. Эти матрицы можно использовать в качестве исходных данных для однофакторного дисперсионного анализа (при помощи подкоманды `MATRIX`).

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Глава 11. Общая линейная модель: одномерный анализ

Процедура ОЛМ-одномерная выполняет регрессионный и дисперсионный анализы для одной зависимой переменной по одному или нескольким факторам и/или переменным. Факторная переменная делит генеральную совокупность на группы. Используя данную процедуру, реализующую общую линейную модель, вы можете проверять нулевую гипотезу о влиянии других переменных на средние различных групп значений единственной зависимой переменной. Вы можете исследовать как взаимодействие между факторами, так и эффекты отдельных факторов, некоторые из которых могут быть случайными. Дополнительно в модель могут быть включены эффекты ковариат и взаимодействия ковариат с факторами. Для регрессионного анализа независимые (предикторные) переменные задаются как ковариаты.

Проверка гипотез может осуществляться как для сбалансированных, так и для несбалансированных моделей. План является сбалансированным, если каждая ячейка в модели содержит одинаковое число наблюдений. Помимо проверки гипотез процедура ОЛМ-одномерная дает оценки параметров.

Для проверки гипотез в процедуре доступны обычно используемые априорные контрасты. После того как общий тест с использованием *F* -критерия показал значимость, вы можете использовать апостериорные критерии, чтобы оценить различия между конкретными средними. Оцененные маргинальные (групповые) средние дают оценки предсказанных средних значений для ячеек в модели, а графики профилей (графики взаимодействий) для этих средних позволяют легко визуализировать исследуемые взаимосвязи.

Для проверки допущений о модели в файле данных могут быть сохранены в качестве новых переменных остатки, предсказанные значения, расстояния Кука и величина плеча.

Поле Взвешенный МНК позволяет задать переменную, используемую для того, чтобы приписать неравные веса наблюдениям во взвешенном методе наименьших квадратов, возможно, для компенсации различий в точности измерений.

Пример. Данные собраны в течение нескольких лет для отдельных бегунов - участников Чикагского марафона. Зависимой переменной является время, за которое каждый бегун пробегает дистанцию. Остальные факторы включают погоду (холодная, хорошая или жаркая), число месяцев тренировки, число предшествующих марафонов и пол. Возраст рассматривается как ковариата. Возможно, что вы обнаружите, что эффект пола, а также взаимодействие пола и погоды являются значимыми.

Методы. При проверке различных гипотез могут использоваться суммы квадратов типа I, типа II, типа III и типа IV. Тип III задается по умолчанию.

Статистики. Апостериорные критерии диапазона и множественные сравнения: наименьшая значимая разность, Бонферрони, Шидака, Шеффе, множественный *F*-критерий Райана-Эйнота-Габриэля-Уэлша, множественный критерий диапазона Райана-Эйнота-Габриэля-Уэлша, Стьюдента-Ньюмена-Келса, критерий Тьюки достоверно значимой разности, Тьюки *b*, Дункана, Гохберга GT2, Габриэля, *t*-критерий Уоллера-Дункана, Даннетта (односторонний и двухсторонний), Тамхейна T2, Даннетта T3, Геймса-Хоуэлла и Даннетта *C*. Описательные статистики: наблюденные средние, среднеквадратические отклонения и частоты в ячейках для всех зависимых переменных. Критерий Ливиня (Levene) однородности дисперсии.

Графики. Разброс по уровням, остатки и профиль (взаимодействие).

Данные для процедуры ОЛМ-одномерная

Данные. Зависимая переменная является количественной. Факторы являются категориальными. Они могут принимать числовые или текстовые значения длиной до восьми символов. Ковариаты являются количественными переменными, связанными с зависимой переменной.

Допущения. Данные представляют собой случайную выборку из нормальной совокупности; дисперсия для всех ячеек должна быть одинаковой. Дисперсионный анализ робастен (устойчив) к отклонениям от нормальности, однако данные должны быть симметричны. Для проверки предположений вы можете использовать критерии однородности дисперсии и графики разброса по уровням. Вы можете также исследовать остатки и графики остатков.

Как запустить процедуру ОЛМ-одномерная

1. Выберите в меню:

   Анализ > Общая линейная модель > ОЛМ-одномерная...
2. Выберите зависимую переменную.
3. Выберите независимые переменные для списков Фиксированные факторы, Случайные факторы и Ковариаты в соответствии с вашими данными.
4. Дополнительно вы можете использовать поле Взвешенный МНК, чтобы задать переменную весов для анализа взвешенным методом наименьших квадратов. Если значение взвешивающей переменной равно нулю, отрицательно, или пропущено, наблюдение исключается из анализа. Переменная, используемая в модели, не может быть взвешивающей.

Общая линейная модель (ОЛМ)

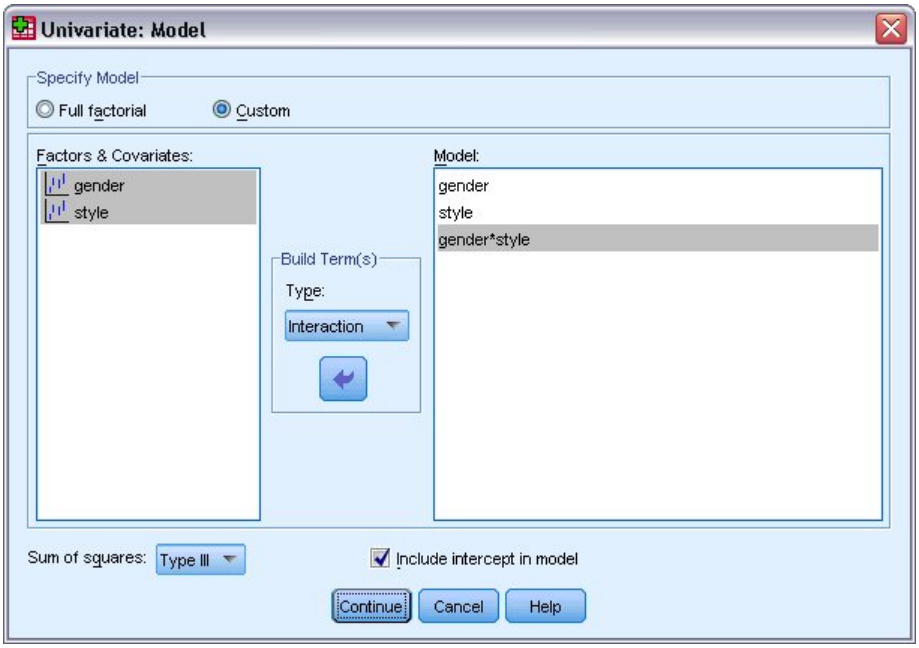


Рисунок 1. Диалоговое окно ОЛМ-одномерная: Модель

Задать модель. Полная факторная модель включает в себя все главные эффекты факторов и ковариат, а также все межфакторные взаимодействия. Она не содержит взаимодействий между ковариатами. Выберите **Настраиваемая** , чтобы задать только подмножество взаимодействий или взаимодействия типа фактор - ковариата. Необходимо указать все члены, включаемые в модель.

Факторы и ковариаты. Перечисляются факторы и ковариаты.

Модель. Модель зависит от природы ваших данных. Выбрав **Настраиваемая**, вы можете отобрать главные эффекты и взаимодействия, которые представляют интерес для анализа.

Сумма квадратов. Метод вычисления сумм квадратов. Для сбалансированных и несбалансированных моделей без пустых ячеек обычно используется метод сумм квадратов типа III.

Включить в модель свободный член. Обычно в модель включают свободный член. Если вы предполагаете, что данные проходят через начало координат, свободный член можно исключить.

Создать члены

Для выбранных факторов и ковариат:

Взаимодействие. Создается член взаимодействия наивысшего порядка всех выбранных переменных. Это вариант по умолчанию.

Главные эффекты. Создаются главные эффекты для всех выбранных переменных.

Все 2-факторные. Создаются все возможные двухфакторные взаимодействия выбранных переменных.

Все 3-факторные. Создаются все возможные трехфакторные взаимодействия выбранных переменных.

Все 4-факторные. Создаются все возможные четырехфакторные взаимодействия выбранных переменных.

Все 5-факторные. Создаются все возможные пятифакторные взаимодействия выбранных переменных.

Сумма квадратов

Для выбранной модели вы можете выбрать тип сумм квадратов. Тип III является наиболее часто используемым, и он задан по умолчанию.

Тип I. Этот метод также известен как метод иерархической декомпозиции сумм квадратов. Каждый член корректируется только по предшествующему ему члену модели. Тип l сумм квадратов обычно используется для:

- Сбалансированной модели дисперсионного анализа, в которой все главные эффекты определяются до эффектов взаимодействий первого порядка, все эффекты взаимодействий первого порядка определяются до эффектов взаимодействий второго порядка, и так далее.
- Полиномиальной регрессионной модели, в которой все члены более низкого порядка определяются раньше, чем любые члены более высокого порядка.
- Чисто гнездовой модели, в которой эффект, определенный первым, вложен в эффект, определенный вторым; эффект, определенный вторым, вложен в эффект, определенный третьим, и так далее. (Эту форму вложения можно задать только с помощью языка команд).

Тип II. Этот метод вычисляет суммы квадратов эффекта в модели, скорректированные по всем остальным "подходящим" эффектам. Под "подходящим" понимается тот эффект, который соответствует всем эффектам, не содержащим исследуемый эффект. Метод сумм квадратов типа II обычно используется для:

- Сбалансированной модели дисперсионного анализа.
- Любой модели, которая содержит только главные эффекты факторов.
- Любой регрессионной модели.
- Чисто гнездового плана. (Эту форму вложения можно задать с помощью языка команд.)

Тип III. Задается по умолчанию. Этот метод вычисляет суммы квадратов эффекта в плане как суммы квадратов, скорректированные по всем остальным эффектам, не содержащим данный эффект, и ортогональным к любому эффекту (если такие есть), содержащему данный эффект. Суммы квадратов типа III имеет одно главное преимущество, заключающееся в том, что они инвариантны относительно частот в ячейках, пока общая форма "оцениваемости" (estimability) остается неизменной. Таким образом, этот тип

сумм квадратов часто считается полезным для несбалансированной модели без пустых ячеек. В факторном плане без пустых ячеек этот метод эквивалентен методу Йетса взвешенных квадратов средних. Метод сумм квадратов типа III обычно используется для:

- Любых моделей, перечисленных для типа I и типа II.
- Любой сбалансированной или несбалансированной модели без пустых ячеек.

Тип IV. Этот метод разработан для случая, когда есть пустые ячейки. Для любого эффекта *F* в данном плане, если *F* не содержится в любом другом эффекте, то тип IV = тип III = тип II. Когда *F* содержится в других эффектах, тип IV распределяет контрасты, сформированные среди параметров в *F* , равноправно между всеми эффектами более высокого порядка. Метод сумм квадратов типа IV обычно используется для:

- Любых моделей, перечисленных для типа I и типа II.
- Любой сбалансированной или несбалансированной модели с пустыми ячейками.

Контрасты ОЛМ

Контрасты используются для проверки различий между уровнями фактора. Вы можете задать контраст для каждого фактора в модели (в модели повторных измерений для каждого межгруппового фактора). Контрасты представляют собой линейные комбинации параметров.

ОЛМ-одномерная. Проверка гипотез основывается на нулевой гипотезе **LB** =0, где **L** - матрица коэффициентов контрастов, а **B** - вектор параметров. При задании контраста создается **L** -матрица. Столбцы **L** -матрицы соответствуют фактору, сочетающемуся с контрастом. Оставшиеся столбцы корректируются так, чтобы матрица **L** допускала оценку.

Вывод включает *F* -статистику для каждого набора контрастов. Для разностей контрастов также выводятся совместные доверительные интервалы типа Бонферрони, основанные на *t* -распределении Стьюдента.

Имеющиеся контрасты

Доступны следующие контрасты: отклонения, простые, дифференциальные, Хелмерта, повторяемые и полиномиальные. Для контрастов типа отклонение и простых контрастов в качестве опорной категории можно указать первую или последнюю категории.

Типы контрастов

Отклонение. Сравнивает среднее значение каждого уровня (исключая опорную категорию) со средним значением всех уровней (генеральным средним). Уровни фактора могут быть расположены в произвольном порядке.

простые. Сравнивает среднее каждого уровня со средним заданного уровня. Этот тип контрастов полезен, когда есть контрольная группа. Вы можете выбрать первую или последнюю категорию в качестве опорной.

Разность. Сравнивает среднее каждого уровня (за исключением первого) со средним значением предыдущих уровней. (Иногда называются обратными контрастами Хелмерта.)

Хелмерт. Сравнивает среднее каждого уровня фактора (за исключением последнего) со средним последующих уровней.

Повторяемый. Сравнивает среднее каждого уровня (кроме последнего) со средним следующего уровня.

Полиномиальный. Сравнивает линейный эффект, квадратичный эффект, кубический эффект, и так далее. Первая степень свободы содержит линейный эффект по всем категориям, вторая степень свободы - квадратичный эффект, и так далее. Такие контрасты часто используются для оценки полиномиальных трендов.

Графики профилей в ОЛМ

Графики профилей (графики взаимодействий) полезны для сравнения маргинальных средних в модели. График профиля представляет собой линейный график, где каждая точка изображает оцененное маргинальное среднее зависимой переменной (скорректированное по всем ковариатам) для одного уровня фактора. Уровни второго фактора можно использовать для построения отдельных линий. Каждый уровень третьего фактора может быть использован для построения отдельного графика. Для графиков подходят все фиксированные и случайные факторы. В многомерном анализе графики профилей создаются для каждой зависимой переменной. В анализе с повторными измерениями, в графиках профилей можно использовать как межгрупповые, так и внутригрупповые факторы. Процедуры ОЛМ-многомерная и ОЛМ-повторные измерения доступны, только если у вас установлен модуль Расширенная статистика.

График профиля одного фактора показывает, возрастают или убывают оцененные маргинальные средние значения от уровня к уровню. Для двух или более факторов параллельность линий говорит о том, что между факторами нет взаимодействия, что означает, что вы можете исследовать уровни каждого фактора по отдельности. Непараллельные линии указывают на наличие факторного взаимодействия.

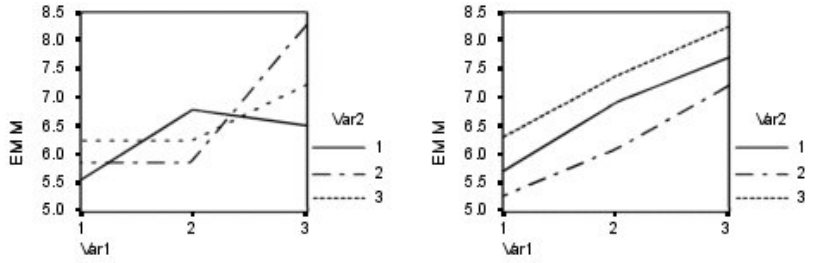


Рисунок 2. Непараллельный график (слева) и параллельный график (справа)

После того как выбраны факторы для горизонтальной оси и, возможно, факторы для отдельных линий и отдельных графиков, график нужно добавить к списку Графики.

Параметры процедуры ОЛМ

Это диалоговое окно позволяет задать дополнительные статистики. Статистики вычисляются с использованием модели с фиксированными эффектами.

Оцененные маргинальные средние. Выберите факторы и взаимодействия, для которых вы хотите получить оценки маргинальных средних значений популяций в ячейках. Эти средние корректируются с учетом ковариат, если они присутствуют в модели

- **Сравнить главные эффекты.** Дает не скорректированные парные сравнения между оцененными маргинальными средними для любых главных эффектов в модели, как для внутригрупповых, так и для межгрупповых факторов. Этот пункт доступен, только если главные эффекты заданы в списке Вывести средние для.
- **Корректировка доверительных интервалов.** Выберите одну из следующих корректировок доверительных интервалов и значимости: наименьшая значимая разность (НЗР), Бонферрони или Шидак. Этот пункт доступен, только если стоит переключатель **Сравнить главные эффекты** .

Вывод. Выберите **Описательные статистики** , чтобы получить наблюденные средние, стандартные отклонения и частоты в ячейках для всех зависимых переменных. Выбор **Оценки силы эффекта** дает значение частной эта-квадрат для каждого эффекта и каждой оценки параметра. Статистика эта-квадрат описывает долю суммарной вариабельности, приписываемую фактору. Выберите **Наблюденная мощность** , чтобы получить мощность критерия, когда альтернативная гипотеза формулируется на основе наблюденного значения. Выберите **Оценки параметров** , чтобы получить оценки параметров, стандартные ошибки, результаты *t* -критерия, доверительные интервалы и наблюденную мощность для каждого критерия. Выберите **Матрица коэфф. контрастов** , чтобы получить матрицу **L** .

Выбор **Критерии однородности** выводит критерий Ливиня однородности дисперсии для каждой зависимой переменной по всем комбинациям уровней межгрупповых факторов, только для межгрупповых факторов. Пункты График разброса по уровням и График остатков полезны для проверки предположений о данных. Этот пункт недоступен, если отсутствуют факторы. Выберите **График остатков** , чтобы для каждой зависимой переменной вывести двумерные графики всех возможных комбинаций наблюденных значений, предсказанных значений и стандартизованных остатков. Эти графики полезны для проверки предположения о равенстве дисперсии. Выберите **Отсутствие согласия** , чтобы проверить, может ли построенная модель адекватно описать связь между зависимой переменной и независимыми переменными. Выбор **Общая функция, допускающая оценку** позволяет конструировать и проверять гипотезы, основанные общей функции, допускающей оценку. Строки в любой матрице коэффициентов контрастов представляют собой линейные комбинации общей функции, допускающей оценку.

Уровень значимости. Возможно, вы захотите скорректировать уровень значимости, используемый в апостериорных критериях, и доверительный уровень, используемый при конструировании доверительных интервалов. Заданное значение используется также для вычисления наблюденной мощности критерия. Когда вы задаете уровень значимости, в диалоговом окне выводится соответствующий уровень доверительных интервалов.

Команда UNIANOVA: дополнительные возможности

Язык синтаксиса команд также позволяет:

- Задать вложенные (nested) эффекты в плане (используя подкоманду `DESIGN`).
- Задать тесты, сравнивающие эффекты с линейной комбинацией эффектов или некоторым значением (используя подкоманду `TEST`).
- Задать множественные контрасты (используя подкоманду `CONTRAST`).
- Включить пользовательские пропущенные значения (используя подкоманду `MISSING`).
- Задать EPS критерии (используя подкоманду `CRITERIA`).
- Сформировать свои собственные матрицу **L** , матрицу **M** и матрицу **K** (используя подкоманды `LMATRIX` , `MMATRIX` и `KMATRIX`).
- Для контрастов типа отклонение или простых контрастов задать промежуточную опорную категорию (используя подкоманду `CONTRAST`).
- Задать метрики для полиномиальных контрастов (используя подкоманду `CONTRAST`).
- Задать компоненты ошибки для апостериорных сравнений (используя подкоманду `POSTHOC`).
- Вычислить оцененные маргинальные средние для любого фактора или взаимодействия факторов среди факторов из списка факторов (используя подкоманду `EMMEANS`).
- Задать имена для временных переменных (используя подкоманду `SAVE`).
- Создать файл данных корреляционной матрицы (используя подкоманду `OUTFILE`).
- Создать матричный файл данных, содержащий статистики из межгрупповой таблицы дисперсионного анализа (используя подкоманду `OUTFILE`).
- Сохранить матрицу плана в новом файле данных (используя подкоманду `OUTFILE`).

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Апостериорные сравнения в ОЛМ

Апостериорные критерии множественных сравнений. Установив, что различия средних значений существуют, с помощью апостериорных критериев диапазона и парных множественных сравнений вы можете выяснить, какие именно средние различаются. Сравнения производятся на нескорректированных значениях. Эти критерии применяются только для фиксированных межгрупповых факторов. В процедуре ОЛМ-повторные измерения эти тесты не доступны, если нет межгрупповых факторов, и апостериорные тесты множественных сравнений проводятся для среднего значения по уровням внутригрупповых факторов. Для процедуры

ОЛМ-многомерная апостериорные тесты проводятся отдельно по каждой зависимой переменной. Процедуры ОЛМ-многомерная и ОЛМ-повторные измерения доступны, только если у вас установлен модуль Расширенная статистика.

Критерии Бонферрони и Тьюки достоверно значимой разности являются обычно используемыми критериями множественных сравнений. Критерий **Бонферрони** , основанный на *t* -критерии Стьюдента, корректирует наблюденный уровень значимости с учетом того факта, что выполняются множественные сравнения. **Т-критерий Шидака** также корректирует уровень значимости и дает более узкие границы, чем критерий Бонферрони. **Критерий Тьюки достоверно значимой разности** использует статистику стьюдентизированного диапазона для проведения всех парных сравнений между группами и устанавливает уровень ошибки эксперимента равным уровню ошибки для совокупности всех парных сравнений. При тестировании большого числа пар средних критерий Тьюки достоверно значимой разности является более мощным, чем критерий Бонферрони. Для малого числа пар более мощным становится критерий Бонферрони.

GT2 Гохберга подобен критерию Тьюки достоверно значимой разности, но использует стьюдентизированный максимальный модуль. Мощность критерия Тьюки обычно больше. **Критерий парных сравнений Габриэля** также использует стьюдентизированный максимальный модуль и обычно имеет большую мощность, чем GT2 Гохберга, при неравных объемах ячеек. Критерий Габриэля может стать либеральным, когда размеры ячеек сильно различаются.

Т-критерий парных множественных сравнений Даннетта сравнивает средние по уровням фактора с единственным контрольным средним. Последняя категория (уровень фактора) по умолчанию служит контрольной. Как вариант можно выбрать первую категорию. Вы также можете выбрать двухсторонний или односторонний критерий. Чтобы проверить, отличается ли среднее для некоторого уровня фактора (за исключением контрольной категории) от среднего для контрольной категории, используйте двухсторонний критерий. Для выяснения того, будет ли среднее для какого-либо уровня фактора меньше, чем среднее для контрольной категории, выберите **< Контр.** . Аналогично для проверки того, больше ли среднее для некоторого уровня фактора, чем среднее для контрольной категории, выберите **> Контр.** .

Райан, Эйнот, Габриэль и Уэлш (Р-Э-Г-У) разработали два множественных нисходящих (step-down) критерия диапазона. Множественная нисходящая процедура сначала проверяет, равны ли все средние. Если не все средние равны, на равенство проверяются подмножества средних значений. **F Р-Э-Г-У** основывается на *F* -критерии, а **Q Р-Э-Г-У** - на стьюдентизированном диапазоне. Эти критерии являются более мощными, чем множественный критерий диапазона Дункана и критерий Стьюдента-Ньюмена-Келса (которые также представляют собой множественные нисходящие процедуры), однако они не рекомендуются для ячеек неравного объема.

Если дисперсии не равны, используйте критерий **Тамхейна Т2** (консервативный критерий парных сравнений, основанный на *t* -критерии), критерий **Даннетта Т3** (критерий парных сравнений, основанный на стьюдентизированном максимальном модуле), **критерий парных сравнений Геймса-Хоуэлла** (иногда либеральный) или критерий **Даннетта С** (критерий парных сравнений, основанный на стьюдентизированном диапазоне). Следует заметить, что эти тесты недостоверны и не могут проводится при наличии в модели нескольких факторов.

Множественный критерий диапазона Дункана , критерии Стьюдента-Ньюмена-Келса (**С-Н-К**) и **Тьюки b** - это критерии диапазона, ранжирующие групповые средние и вычисляющие величину диапазона. Эти критерии используются реже, чем обсуждавшиеся выше.

Т-критерий Уоллера-Дункана использует Байесовский подход. Этот критерий диапазона использует гармоническое среднее объемов выборок, когда объемы выборок не равны.

Уровень значимости критерия **Шеффе** устанавливается так, чтобы можно было протестировать все возможные линейные комбинации групповых средних, а не только парные сравнения, доступные в этом качестве. В результате критерий Шеффе зачастую более консервативен, чем остальные, это означает, что для значимости требуется большая разность между средними.

Критерий наименьшей значимой разности (**НЗР**) парных множественных сравнений эквивалентен множеству отдельных *t* -критериев между всеми парами групп. Недостаток этого критерия в том, что не делается попытки скорректировать наблюденный уровень значимости для множественных сравнений.

Представленные тесты. Парные сравнения предусматриваются для НЗР, Шидака, Бонферрони, Геймса и Хоуэлла, Тамхейна Т2 и Т3, Даннетта *C* и Даннетта Т3. Однородные подмножества для критериев диапазона предусматриваются для С-Н-К, Тьюки *b* , Дункана, *F* Р-Э-Г-У, *Q* Р-Э-Г-У и Уоллера. Критерий Тьюки достоверно значимой разности, GT2 Гохберга, критерий Габриэля и критерий Шеффе являются одновременно критериями множественных сравнений и критериями диапазона.

Параметры процедуры ОЛМ

Это диалоговое окно позволяет задать дополнительные статистики. Статистики вычисляются с использованием модели с фиксированными эффектами.

Оцененные маргинальные средние. Выберите факторы и взаимодействия, для которых вы хотите получить оценки маргинальных средних значений популяций в ячейках. Эти средние корректируются с учетом ковариат, если они присутствуют в модели

- **Сравнить главные эффекты.** Дает не скорректированные парные сравнения между оцененными маргинальными средними для любых главных эффектов в модели, как для внутригрупповых, так и для межгрупповых факторов. Этот пункт доступен, только если главные эффекты заданы в списке Вывести средние для.
- **Корректировка доверительных интервалов.** Выберите одну из следующих корректировок доверительных интервалов и значимости: наименьшая значимая разность (НЗР), Бонферрони или Шидак. Этот пункт доступен, только если стоит переключатель **Сравнить главные эффекты** .

Вывод. Выберите **Описательные статистики** , чтобы получить наблюденные средние, стандартные отклонения и частоты в ячейках для всех зависимых переменных. Выбор **Оценки силы эффекта** дает значение частной эта-квадрат для каждого эффекта и каждой оценки параметра. Статистика эта-квадрат описывает долю суммарной вариабельности, приписываемую фактору. Выберите **Наблюденная мощность** , чтобы получить мощность критерия, когда альтернативная гипотеза формулируется на основе наблюденного значения. Выберите **Оценки параметров** , чтобы получить оценки параметров, стандартные ошибки, результаты *t* -критерия, доверительные интервалы и наблюденную мощность для каждого критерия. Выберите **Матрица коэфф. контрастов** , чтобы получить матрицу **L** .

Выбор **Критерии однородности** выводит критерий Ливиня однородности дисперсии для каждой зависимой переменной по всем комбинациям уровней межгрупповых факторов, только для межгрупповых факторов. Пункты График разброса по уровням и График остатков полезны для проверки предположений о данных. Этот пункт недоступен, если отсутствуют факторы. Выберите **График остатков** , чтобы для каждой зависимой переменной вывести двумерные графики всех возможных комбинаций наблюденных значений, предсказанных значений и стандартизованных остатков. Эти графики полезны для проверки предположения о равенстве дисперсии. Выберите **Отсутствие согласия** , чтобы проверить, может ли построенная модель адекватно описать связь между зависимой переменной и независимыми переменными. Выбор **Общая функция, допускающая оценку** позволяет конструировать и проверять гипотезы, основанные общей функции, допускающей оценку. Строки в любой матрице коэффициентов контрастов представляют собой линейные комбинации общей функции, допускающей оценку.

Уровень значимости. Возможно, вы захотите скорректировать уровень значимости, используемый в апостериорных критериях, и доверительный уровень, используемый при конструировании доверительных интервалов. Заданное значение используется также для вычисления наблюденной мощности критерия. Когда вы задаете уровень значимости, в диалоговом окне выводится соответствующий уровень доверительных интервалов.

Команда UNIANOVA: дополнительные возможности

Язык синтаксиса команд также позволяет:

- Задать вложенные (nested) эффекты в плане (используя подкоманду DESIGN).
- Задать тесты, сравнивающие эффекты с линейной комбинацией эффектов или некоторым значением (используя подкоманду TEST).
- Задать множественные контрасты (используя подкоманду CONTRAST).
- Включить пользовательские пропущенные значения (используя подкоманду MISSING).
- Задать EPS критерии (используя подкоманду CRITERIA).
- Сформировать свои собственные матрицу **L** , матрицу **M** и матрицу **K** (используя подкоманды LMATRIX , MMATRIX и KMATRIX).
- Для контрастов типа отклонение или простых контрастов задать промежуточную опорную категорию (используя подкоманду CONTRAST).
- Задать метрики для полиномиальных контрастов (используя подкоманду CONTRAST).
- Задать компоненты ошибки для апостериорных сравнений (используя подкоманду POSTHOC).
- Вычислить оцененные маргинальные средние для любого фактора или взаимодействия факторов среди факторов из списка факторов (используя подкоманду EMMEANS).
- Задать имена для временных переменных (используя подкоманду SAVE).
- Создать файл данных корреляционной матрицы (используя подкоманду OUTFILE).
- Создать матричный файл данных, содержащий статистики из межгрупповой таблицы дисперсионного анализа (используя подкоманду OUTFILE).
- Сохранить матрицу плана в новом файле данных (используя подкоманду OUTFILE).

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Сохранение новых переменных в ОЛМ

Вы можете сохранить значения, предсказанные моделью, остатки и связанные с моделью меры в качестве новых переменных в редакторе данных. Многие из этих переменных можно затем использовать для проверки предположений о данных. Для обращения к ним во время других сеансов работы с IBM SPSS Statistics, нужно сохранить этот файл данных.

Предсказанные значения. Значения, которые модель предсказывает для каждого наблюдения.

- *Нестандартизованные*. Значение зависимой переменной, предсказываемое в соответствии с моделью.
- *Взвешенные*. Взвешенные нестандартизованные предсказанные значения. Опция доступна только тогда, когда предварительно была выбрана ВМНК-переменная.
- *Среднеквадратическая ошибка*. Оценка стандартного отклонения среднего значения зависимой переменной для наблюдений с одинаковыми значениями независимых переменных.

Диагностики. Меры, выявляющие наблюдения с необычными комбинациями значений независимых переменных и наблюдения, которые могут оказать большое влияние на модель.

- *Расстояние Кука*. Для каждого наблюдения показывает насколько изменятся остатки всех наблюдений, если это наблюдение не использовать при вычислении коэффициентов регрессии. Большое расстояние Кука указывает на то, что исключение данного наблюдения из вычислений регрессии существенно меняет коэффициенты.
- *Значения разбалансировки*. Нецентрированные значения балансировки. Относительное влияние каждого наблюдения на согласие модели.

Остатки. Нестандартизованный остаток - это фактическое значение зависимой переменной минус значение, предсказанное моделью. Можно получить также стандартизованные, стьюдентизированные и "удаленные" остатки. Если выбрана переменная весов, можно вычислить взвешенные нестандартизованные остатки.

- *Нестандартизованные*. Разность между наблюдаемым и предсказанным моделью значением.
- *Взвешенные*. Взвешенные нестандартизованные остатки. Опция доступна только тогда, когда предварительно была выбрана ВМНК-переменная.
- *Стандартизованные*. Остаток, деленный на оценку его стандартного отклонения. Стандартизованные остатки, известные еще как пирсоновские, имеют среднее 0 и стандартное отклонение 1.
- *Стьюдентизированные*. Остаток, деленный на оценочное значение его среднеквадратичного отклонения, которое изменяется от наблюдения к наблюдению в зависимости от расстояния значений каждого наблюдения для независимых переменных от средних значений этих переменных.
- *Удалено*. Остаток для наблюдения, когда данное наблюдение исключается при вычислении регрессионных коэффициентов. Это разность между значением зависимой переменной и скорректированным предсказанным значением.

Статистики коэффициентов Ковариационная матрица оценок параметров модели сохраняется в новом наборе данных или во внешнем файле данных в формате IBM SPSS Statistics. Кроме того, для каждой зависимой переменной в нем содержится строка оценок параметров, строка уровней значимости *t* -статистик, соответствующих оценкам параметров, и строка степеней свободы остатков. В многомерной модели есть подобные строки для каждой зависимой переменной. Этот файл можно использовать в других процедурах, читающих матричные файлы.

Параметры процедуры ОЛМ

Это диалоговое окно позволяет задать дополнительные статистики. Статистики вычисляются с использованием модели с фиксированными эффектами.

Оцененные маргинальные средние. Выберите факторы и взаимодействия, для которых вы хотите получить оценки маргинальных средних значений популяций в ячейках. Эти средние корректируются с учетом ковариат, если они присутствуют в модели

- **Сравнить главные эффекты.** Дает не скорректированные парные сравнения между оцененными маргинальными средними для любых главных эффектов в модели, как для внутригрупповых, так и для межгрупповых факторов. Этот пункт доступен, только если главные эффекты заданы в списке Вывести средние для.
- **Корректировка доверительных интервалов.** Выберите одну из следующих корректировок доверительных интервалов и значимости: наименьшая значимая разность (НЗР), Бонферрони или Шидак. Этот пункт доступен, только если стоит переключатель **Сравнить главные эффекты** .

Вывод. Выберите **Описательные статистики** , чтобы получить наблюденные средние, стандартные отклонения и частоты в ячейках для всех зависимых переменных. Выбор **Оценки силы эффекта** дает значение частной эта-квадрат для каждого эффекта и каждой оценки параметра. Статистика эта-квадрат описывает долю суммарной вариабельности, приписываемую фактору. Выберите **Наблюденная мощность** , чтобы получить мощность критерия, когда альтернативная гипотеза формулируется на основе наблюденного значения. Выберите **Оценки параметров** , чтобы получить оценки параметров, стандартные ошибки, результаты *t* -критерия, доверительные интервалы и наблюденную мощность для каждого критерия. Выберите **Матрица коэфф. контрастов** , чтобы получить матрицу **L** .

Выбор **Критерии однородности** выводит критерий Ливиня однородности дисперсии для каждой зависимой переменной по всем комбинациям уровней межгрупповых факторов, только для межгрупповых факторов. Пункты График разброса по уровням и График остатков полезны для проверки предположений о данных. Этот пункт недоступен, если отсутствуют факторы. Выберите **График остатков** , чтобы для каждой зависимой переменной вывести двумерные графики всех возможных комбинаций наблюденных значений, предсказанных значений и стандартизованных остатков. Эти графики полезны для проверки предположения о равенстве дисперсии. Выберите **Отсутствие согласия** , чтобы проверить, может ли построенная модель адекватно описать связь между зависимой переменной и независимыми переменными. Выбор **Общая**

функция, допускающая оценку позволяет конструировать и проверять гипотезы, основанные общей функции, допускающей оценку. Строки в любой матрице коэффициентов контрастов представляют собой линейные комбинации общей функции, допускающей оценку.

Уровень значимости. Возможно, вы захотите скорректировать уровень значимости, используемый в апостериорных критериях, и доверительный уровень, используемый при конструировании доверительных интервалов. Заданное значение используется также для вычисления наблюденной мощности критерия. Когда вы задаете уровень значимости, в диалоговом окне выводится соответствующий уровень доверительных интервалов.

Команда UNIANOVA: дополнительные возможности

Язык синтаксиса команд также позволяет:

- Задать вложенные (nested) эффекты в плане (используя подкоманду DESIGN).
- Задать тесты, сравнивающие эффекты с линейной комбинацией эффектов или некоторым значением (используя подкоманду TEST).
- Задать множественные контрасты (используя подкоманду CONTRAST).
- Включить пользовательские пропущенные значения (используя подкоманду MISSING).
- Задать EPS критерии (используя подкоманду CRITERIA).
- Сформировать свои собственные матрицу **L** , матрицу **M** и матрицу **K** (используя подкоманды LMATRIX , MMATRIX и KMATRIX).
- Для контрастов типа отклонение или простых контрастов задать промежуточную опорную категорию (используя подкоманду CONTRAST).
- Задать метрики для полиномиальных контрастов (используя подкоманду CONTRAST).
- Задать компоненты ошибки для апостериорных сравнений (используя подкоманду POSTHOC).
- Вычислить оцененные маргинальные средние для любого фактора или взаимодействия факторов среди факторов из списка факторов (используя подкоманду EMMEANS).
- Задать имена для временных переменных (используя подкоманду SAVE).
- Создать файл данных корреляционной матрицы (используя подкоманду OUTFILE).
- Создать матричный файл данных, содержащий статистики из межгрупповой таблицы дисперсионного анализа (используя подкоманду OUTFILE).
- Сохранить матрицу плана в новом файле данных (используя подкоманду OUTFILE).

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Глава 12. Парные корреляции

Процедура Парные корреляции вычисляет коэффициент корреляции Пирсона, ро Спирмана и тау-*b* Кендалла, а также уровни значимости для них. Корреляции измеряют связь между переменными или рангами. Перед вычислением коэффициента корреляции проверьте данные на наличие выбросов (которые могут привести к вводящим в заблуждение результатам) и признаков наличия линейной связи. Коэффициент корреляции Пирсона является мерой линейной связи. Две переменные могут быть на 100% связаны, однако если эта связь нелинейная, коэффициент корреляции Пирсона не является подходящей статистикой для ее измерения.

Пример. Связано ли число выигранных баскетбольной командой игр со средним числом очков за игру? Диаграмма рассеяния показывает, что между ними имеется линейная связь. Анализ данных НБА о сезонах 1994–1995 годов выявил, что коэффициент корреляции Пирсона (0,581) значимо отличен от нуля на уровне значимости 0.01. Можно ожидать, что чем больше игр будет выиграно командой за сезон, тем меньше очков наберут соперники этой команды. Эти переменные отрицательно коррелированны (–0,401), и корреляция значима на уровне 0,05.

Статистика. Для каждой переменной: число наблюдений без пропущенных значений, среднее значение и стандартное отклонение. Для каждой пары переменных: коэффициент корреляции Пирсона, ро Спирмана, тау-*b* Кендалла, суммы перекрестных произведений отклонений, ковариация.

Данные для процедуры Парные корреляции

Данные. При работе с коэффициентом корреляции Пирсона используйте симметричные количественные переменные; при работе с ро Спирмана и тау-*b* Кендалла используйте количественные переменные или переменные с упорядоченными категориями (ранговые).

Допущения. Применение коэффициента корреляции Пирсона предполагает, что каждая пара переменных соответствует двумерному нормальному распределению.

Как запустить процедуру Парные корреляции

Выберите в меню:

Анализ > Корреляция > Парные...

1. Выберите две или более числовые переменные.

Доступны также следующие параметры:

- **Коэффициенты корреляции.** Для количественных нормально распределенных переменных выберите коэффициент корреляции **Пирсона** . Если данные не распределены нормально или имеют упорядоченные категории (являются ранговыми), выберите **тау-b Кендалла** или **Спирмана** , которые измеряют связь между рангами. Коэффициенты корреляции изменяются от –1 (полная отрицательная связь) до +1 (полная положительная связь). Значение 0 указывает на отсутствие линейной связи. При интерпретации полученных результатов тщательно следите за тем, чтобы не делать выводов о причинной связи на основе значимой корреляции.
- **Критерий значимости.** Вы можете выбрать двухсторонний или односторонний критерий. Если направление связи известно заранее, выберите **Односторонний** . В противном случае выберите **Двухсторонний** .
- **Метить значимые корреляции.** Коэффициенты корреляции, значимые на уровне 0.05, обозначены одной звездочкой, а значимые на уровне 0.01 - двумя звездочками.

Параметры процедуры Парные корреляции

Статистики. Для корреляции Пирсона вы можете выбрать один или оба из следующих пунктов:

- **Средние значения и стандартные отклонения.** Выводятся для каждой переменной. Выводится также число наблюдений без пропущенных значений. Пропущенные значения обрабатываются для каждой переменной по отдельности, вне зависимости от установки, выбранной в панели Пропущенные значения.
- **Суммы перекрестных произведений отклонений и ковариации.** Выводятся для каждой пары переменных. Сумма перекрестных произведений отклонений равна сумме произведений переменных, скорректированных по среднему. Это числитель в формуле коэффициента корреляции Пирсона. Ковариация - это ненормированная мера связи между двумя переменными, равная сумме перекрестных произведений отклонений, деленной на *N*–1.

Пропущенные значения. Вы можете выбрать один из следующих вариантов:

- **Исключать попарно.** Наблюдения с пропущенными значениями одной или обеих переменных пары, для которых вычисляется коэффициент корреляции, исключаются из анализа. Поскольку в вычислениях каждого коэффициента участвуют все наблюдения без пропущенных значений для данной пары переменных, то в каждом вычислении используется максимум доступной информации. Это может привести к тому, что набор коэффициентов будет вычислен для разного числа наблюдений.
- **Исключать целиком.** Наблюдения с пропущенными значениями для какой-либо переменной исключаются из вычислений всех корреляций.

Команды CORRELATIONS и NONPAR CORR: дополнительные возможности

Язык синтаксиса команд также позволяет:

- Записать корреляционную матрицу для корреляций Пирсона, которую можно использовать в качестве исходных данных в других процедурах, например, в факторном анализе (с использованием подкоманды `MATRIX`).
- Получить корреляции каждой переменной списка с каждой переменной другого списка (используя ключевое слово `WITH` в подкоманде `VARIABLES`).

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Глава 13. Частные корреляции

Процедура Частные корреляции вычисляет частные коэффициенты корреляции, которые описывают линейную связь между двумя переменными при устранении влияния одной или нескольких дополнительных переменных. Корреляции - это меры линейной связи. Две переменные могут иметь "полную" связь, однако если эта связь нелинейна, коэффициент корреляции не является подходящей статистикой для ее измерения.

Пример. Есть ли взаимосвязь между финансированием здравоохранения и уровнем заболеваемости? Хотя вы можете ожидать, что такая связь будет отрицательной, проведенное исследование показывает наличие значимой *положительной* корреляции: по мере увеличения финансирования здравоохранения увеличивается уровень заболеваемости. Фиксация уровня посещаемости медицинских учреждений, однако, устраняет эту наблюдаемую положительную корреляцию. Финансирование здравоохранения и уровень заболеваемости только кажутся положительно взаимосвязанными, поскольку при увеличении финансирования больше людей получают доступ к услугам здравоохранения, что приводит к выявлению большего числа случаев заболеваний.

Статистика. Для каждой переменной: число наблюдений без пропущенных значений, среднее значение и среднеквадратичное отклонение. Матрицы корреляций и частных корреляций со степенями свободы и уровнями значимости.

Данные для процедуры Частные корреляции

Данные. Используйте симметричные количественные переменные.

Допущения. Процедура Частные корреляции предполагает, что каждая пара переменных соответствует двумерному нормальному распределению.

Как запустить процедуру Частные корреляции

1. Выберите в меню:

   Анализ > Корреляция > Частные...
2. Выберите две или более числовые переменные, для которых будут вычисляться частные корреляции.
3. Выберите одну или несколько числовых переменных, влияние которых устраняется (Исключаемые).

Доступны также следующие параметры:

- **Критерий значимости.** Вы можете выбрать двухсторонний или односторонний критерий. Если направление связи известно заранее, выберите **Односторонний** . В противном случае выберите **Двухсторонний** .
- **Выводить истинный уровень значимости.** По умолчанию для каждого коэффициента корреляции выводятся вероятность и число степеней свободы. Если вы снимите пометку с этого элемента, коэффициенты корреляции, значимые на уровне 0.05, будут обозначаться одной звездочкой, а значимые на уровне 0.01 - двумя звездочками. При этом числа степеней свободы не выводятся. Данная установка относится как к частным корреляциям, так и к корреляциям нулевого порядка (т.е. обычным парным корреляциям).

Параметры процедуры Частные корреляции

Статистики. Вы можете выбрать один или оба из следующих пунктов:

- **Средние значения и стандартные отклонения.** Выводятся для каждой переменной. Выводится также число наблюдений без пропущенных значений.
- **Корреляции нулевого порядка.** Выводится матрица простых корреляций между всеми переменными, в том числе и теми, влияние которых будет устраняться.

Пропущенные значения. Вы можете выбрать одну из следующих альтернатив:

- **Исключать целиком.** Наблюдения с пропущенными значениями любой переменной, в том числе и переменной, влияние которой устраняется, исключаются из всех вычислений.
- **Исключать попарно.** Для вычисления корреляций нулевого порядка, на которых основывается вычисление частных корреляций, не будут использоваться наблюдения с пропущенными значениями для одной или обеих переменных пары. Попарное исключение использует данные в максимально возможной степени. Однако, в этом случае число используемых наблюдений может изменяться от одного коэффициента к другому. Когда задано попарное исключение, число степеней свободы для конкретного частного коэффициента основывается на наименьшем числе наблюдений, используемых при вычислении любой из корреляций нулевого порядка.

Команда PARTIAL CORR: дополнительные возможности

Язык синтаксиса команд также позволяет:

- Считывать корреляционные матрицы нулевого порядка и записывать матрицы частных корреляций (при помощи подкоманды `MATRIX`).
- Рассчитывать частные корреляции для переменных в двух списках (при помощи ключевого слова `WITH` в подкоманде `VARIABLES`).
- Анализировать несколько наборов переменных (при помощи нескольких подкоманд `VARIABLES`).
- Задавать порядок рассчитываемых корреляций (например частные корреляции первого и второго порядка), если имеется две контрольные переменные, (при помощи подкоманды `VARIABLES`).
- Выводить частные корреляции в компактном формате (при помощи подкоманды `FORMAT`).
- Выводить матрицу простых корреляций, если некоторые коэффициенты не могут быть рассчитаны (при помощи подкоманды `STATISTICS`).

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Глава 14. Расстояния

Эта процедура вычисляет любую статистику из широкого набора статистик, измеряющих либо сходства, либо различия (расстояния), причем либо между парами переменных, либо между парами наблюдений. Эти меры сходства или расстояния могут быть затем использованы в других процедурах, таких как факторный анализ, кластерный анализ или многомерное масштабирование, для того чтобы помочь анализировать сложные наборы данных.

Пример. Можно ли измерить сходство между парами автомобилей, основываясь на определенных характеристиках, таких как объем двигателя, расход топлива и мощность? Вычислив величины сходства между автомобилями, вы можете получить представление о том, какие автомобили похожи, а какие различаются. Для более формального анализа к величинам сходства можно применить иерархический кластерный анализ или многомерное масштабирование для того, чтобы исследовать скрытую структуру данных.

Статистика. Меры различия (расстояния) для интервальных данных: расстояние Евклида, квадрат расстояния Евклида, метрики Чебышева, блок, Минковского, а также задаваемые пользователем. Для частот: хи-квадрат и фи-квадрат. Для бинарных данных: расстояние Евклида, квадрат расстояния Евклида, различие размеров, различие структур, дисперсия, форма, Ланс и Уильямс. Мерами сходства для интервальных данных являются: коэффициент корреляции Пирсона и косинус. Для двоичных данных: Рассел и Рао, простая мера совпадений, Жаккар, дайс, Роджерс и Танимото, Сокал и Снит 1, Сокал и Снит 2, Сокал и Снит 3, Кульчинский 1, Кульчинский 2, Сокал и Снит 4, Хаманн, Лямбда, *D* Андерберга, *Y* Юла, *Q* Юла, Очиай, Сокал и Снит 5, четырехточечная корреляция фи, разброс.

Как получить матрицы расстояний

1. Выберите в меню:

   Анализ > Корреляция > Расстояния...
2. Выберите, по крайней мере, одну числовую переменную, чтобы вычислять расстояния между наблюдениями, или выберите, по крайней мере, две числовые переменные, чтобы вычислить расстояния между переменными.
3. Выберите одну из двух альтернатив в группе Вычислить расстояния между, чтобы вычислить расстояния либо между наблюдениями, либо между переменными.

Меры различия

В группе Мера выберите альтернативу, соответствующую типу данных (интервальным, количествам или двоичным); затем в выпадающем списке выберите одну из мер, которая соответствует этому типу данных. Доступными мерами в зависимости от типа данных являются следующие:

- **Интервальные данные.** Расстояние Евклида, квадрат расстояния Евклида, расстояние Чебышева, блок, Минковского или Настроенная (пользователем).
- **Частоты.** Меры хи-квадрат или фи-квадрат.
- **Двоичные данные.** Расстояние Евклида, квадрат расстояния Евклида, различие размеров, различие структур, дисперсия, форма, Ланс и Уильямс. (Введите значения в поля Наличие и Отсутствие, чтобы указать, какие два значения используются; остальные значения будут игнорироваться процедурой.)

Группа Преобразовать значения позволяет *перед* вычислением близостей стандартизировать значения данных либо для наблюдений, либо для переменных. Эти преобразования неприменимы к двоичным данным. Возможные методы стандартизации: *Z*–значения, Диапазон от –1 до 1, Диапазон от 0 до 1, Максимальная величина 1, Среднее 1 или Среднеквадратичное отклонение 1.

Группа Преобразовать меры позволяет преобразовать генерируемые значения меры расстояния. Преобразования выполняются после того, как вычислены значения меры расстояния. Возможные варианты преобразований: Взять модуль, Сменить знак, Привести к 0–1.

Меры сходства

В группе Мера выберите альтернативу, соответствующую типу данных (интервальная или двоичная); затем в выпадающем списке выберите одну из мер, которая соответствует этому типу данных. Доступными мерами в зависимости от типа данных являются следующие:

- **Интервальные данные.** Коэффициент корреляции Пирсона или косинус.
- **Двоичные данные.** Рассел и Рао, простая мера совпадений, Жаккар, дайс, Роджерс и Танимото, Сокал и Снит 1, Сокал и Снит 2, Сокал и Снит 3, Кульчинский 1, Кульчинский 2, Сокал и Снит 4, Хаманн, Лямбда, *D* Андерберга, *Y* Юла, *Q* Юла, Очиай, Сокал и Снит 5, четырехточечная корреляция фи, разброс. (Введите значения в поля Наличие и Отсутствие, чтобы указать, какие два значения используются; остальные значения будут игнорироваться процедурой.)

Группа Преобразовать значения позволяет перед вычислением расстояний стандартизировать значения данных либо для наблюдений, либо для переменных. Эти преобразования неприменимы к двоичным данным. Возможные методы стандартизации: *Z*–значения, Диапазон от –1 до 1, Диапазон от 0 до 1, Максимальная величина 1, Среднее 1 и Среднеквадратичное отклонение 1.

Группа Преобразовать меры позволяет преобразовать генерируемые значения меры расстояния. Преобразования выполняются после того, как вычислены значения меры расстояния. Возможные варианты преобразований: Взять модуль, Сменить знак, Привести к 0–1.

Команда PROXIMITIES: дополнительные возможности

Процедура Расстояния использует синтаксис команды `PROXIMITIES` . Язык синтаксиса команд также позволяет:

- Задать любое целое число в качестве степени для меры расстояния Минковского.
- Задать любое целое число в качестве корня для настраиваемой меры расстояния.

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Глава 15. Линейные модели

Линейные модели предсказывают значения непрерывных целевых переменных, основываясь на взаимосвязи между целевой переменной и одним или несколькими предикторами.

Линейные модели относительно просты и дают легко интерпретируемую математическую формулу для скоринга. Свойства этих моделей хорошо понятны, и их обычно можно построить очень быстро, по сравнению с моделями других типов (такими как нейронные сети или деревья решений) на том же наборе данных.

Пример. Страховая компания с ограниченными ресурсами для исследования страховых требований домовладельцев желает построить модель для оценки стоимости требований. Применяя эту модель в центрах обслуживания, сотрудники компании могут ввести информацию от требовании, разговаривая по телефону с клиентом, и немедленно получить "ожидаемую" стоимость требования, основываясь на прошлых данных.

Требования к полям Должны быть целевое и, по крайней мере, одно входное поля. По умолчанию не используются поля с предопределенными ролями Двойного назначения и Нет. Целевое поле должно быть непрерывным (количественным). Для предикторов (входов) отсутствуют ограничения на тип измерений; категориальные поля (номинальные и порядковые) используются в модели в качестве факторов, а непрерывные поля используются как ковариаты.

Примечание: Если категориальное поле содержит более 1000 категорий, эта процедура не выполняется и модель не строится.

Как запустить процедуру построения линейной модели

Для этой функциональной возможности требуется модуль База статистики.

Выберите в меню:

Анализ > Регрессия > Автоматические линейные модели...

1. Удостоверьтесь, что есть, по крайней мере, одна целевая и одна входная переменная.
2. Щелкните по **Параметры конструкции** , чтобы задать необязательные параметры сборки и модели.
3. Щелкните по **Параметры модели** , чтобы сохранить оценки в активном наборе данных и экспортировать модель во внешний файл.
4. Щелкните по **Запуск** , чтобы запустить процедуру и создать объекты модели.

Цели

Какова ваша главная цель? Выберите подходящую цель.

- **Создать стандартную модель.** Данный метод строит единичную модель для предсказания целевой переменной, используя предикторы. Вообще говоря, стандартные модели легче поддаются интерпретации и могут требовать меньше времени при скоринге, чем построенные с применением бустинга, бэггинга или ансамблей больших наборов данных.
- **Повысить точность модели (бустинг).** Данный метод строит модель ансамбля, используя бустинг, который генерирует последовательность моделей для получения более точных предсказаний. Ансамбли могут занять больше времени для их построения и скоринга, чем стандартная модель.

  Бустинг генерирует последовательность "компонентных моделей", каждая из которых строится по целому набору данных. Прежде чем строить каждую последовательную компонентную модель, записи взвешиваются на основе остатков для предшествующей компонентной модели. Наблюдениям с

большими остатками придаются относительно большие веса прецедентов, с тем чтобы следующая компонентная модель была сконцентрирована на том, чтобы хорошо предсказывать такие записи. Вместе такие компонентные модели образуют модель ансамбля. Модель ансамбля выполняет скоринг новых записей, пользуясь правилом объединения; доступные правила зависят от типа измерений целевой переменной.

- **Повысить стабильность модели (бэггинг).** Данный метод строит модель ансамбля, используя бэггинг (бутстреп-агрегирование), который генерирует множественные модели для получения более надежных предсказаний. Ансамбли могут занять больше времени для их построения и скоринга, чем стандартная модель.

  Бутстреп-агрегирование (бэггинг) формирует реплики обучающего набора данных путем выбора с возвращением из исходного набора данных. В результате создаются бутстреп-выборки исходного набора данных равного объема. Затем по каждой реплике формируется "компонентная модель". Вместе такие компонентные модели образуют модель ансамбля. Модель ансамбля выполняет скоринг новых записей, пользуясь правилом объединения; доступные правила зависят от типа измерений целевой переменной.
- **Создать модель для очень больших наборов данных (требует сервера IBM SPSS Statistics).** Данный метод строит модель ансамбля путем расщепления набора данных на отдельные блоки данных. Выберите этот вариант, если ваш набор данных слишком велик для построения моделей перечисленных выше, или для инкрементного построения модели. Данный вариант может потребовать меньше времени для построения, но больше времени для скоринга, чем стандартная модель. Для этой опции требуется соединение с сервером IBM SPSS Statistics.

Информацию о параметрах, связанных с бустингом, бэггингом и очень большими наборами данных, смотрите в разделе “Ансамбли” на стр. 66.

Основные параметры

Автоматически подготовить данные. Этот параметр позволяет процедуре выполнить внутренние преобразования целевой переменной и предикторов, чтобы максимизировать прогностическую силу модели. Все преобразования сохраняются вместе с моделью и применяются к новым данным при скоринге. Исходные версии преобразованных полей исключаются из модели. По умолчанию выполняются автоматические преобразования данных, описанные ниже.

- **Обработка дат и времени.** Каждый предиктор, являющейся переменной дат, преобразуется в новый непрерывный предиктор, содержащий время, прошедшее, начиная с опорной даты (1970-01-01). Каждый предиктор, являющийся переменной времени, преобразуется в новый непрерывный предиктор, содержащий время, прошедшее, начиная с опорного момента времени (00:00:00).
- **Корректировка шкалы измерений.** Непрерывные предикторы, содержащие менее 5 различных значений, преобразуются в порядковые предикторы. Порядковые предикторы, содержащие более 10 различных значений, преобразуются в непрерывные предикторы.
- **Обработка выбросов.** Значения непрерывных предикторов, которые лежат вне границ отсечения (определяемых тремя стандартными отклонениями от среднего значения), заменяются значением границы отсечения.
- **Обработка пропущенных значений.** Пропущенные значения номинальных предикторов заменяются модой обучающего разбиения. Пропущенные значения порядковых предикторов заменяются медианой обучающего разбиения. Пропущенные значения непрерывных предикторов заменяются средним значением обучающего разбиения.
- **Контролируемое объединение.** Эта операция делает модель более "экономной" путем уменьшения числа полей, обрабатываемых в связи с целевым полем. Идентифицируются подобные категории, основываясь на взаимосвязи между входным и целевым полями. Категории, которые не различаются значимо (т.е. имеющие p-значение больше 0,1), объединяются. Если все категории объединяются в одну, то исходная и полученная версии поля исключаются из модели, поскольку они не представляют ценности как предиктор.

Доверительный уровень. Это доверительный уровень, используемый при вычислении интервальных оценок коэффициентов модели, представленных на панели Коэффициенты. Задайте значение больше 0 и меньше 100. Значение по умолчанию - 95.

Подбор модели

Метод подбора модели. Выберите один из методов подбора модели (подробности ниже) или **Включить все предикторы**, когда все имеющиеся предикторы просто вводятся в модель как члены главных эффектов. По умолчанию используется **Прямой шаговый** .

Прямой шаговый отбор. Этот метод начинает работу с модели без эффектов, добавляя и удаляя эффекты по одному на каждом шаге до тех пор, пока ни один эффект нельзя будет добавить, руководствуясь критериями шагового отбора.

- **Критерии для включения/исключения.** Это статистика, используемая для определения того, следует ли эффект добавить в модель или исключить из нее. **Информационный критерий (AICC)** основывается на правдоподобии обучающего множества для данной модели и скорректирован с целью штрафовать излишне сложные модели. **F-статистики** основывается на статистическом критерии снижения модельной ошибки. **Скорректированный R-квадрат** основывается на точности подгонки для обучающего множества и скорректирован с целью штрафовать излишне сложные модели. **Критерий предотвращения сверхобучения (СКО)** основывается на точности подгонки (среднем квадрате ошибки или СКО) для множества предотвращения сверхобучения. Множество предотвращения сверхобучения представляет собой случайную подвыборку, содержащую приблизительно 30% наблюдений из исходного набора данных, которая не используется при обучении модели.

  Если выбран любой критерий, отличный от **F-статистики** , то на каждом шаге в модель добавляется эффект, соответствующий максимальному положительному приращению значения критерия. Все эффекты в модели, соответствующие уменьшению значения критерия, удаляются.

  Если в качестве критерия выбран **F-статистики** , то на каждом шаге в модель добавляется эффект, дающий наименьшее *p*-значение, при условии, что оно меньше порогового значения, заданного в **Включать эффекты с p-значениями, меньшими чем**. Значение по умолчанию - 0,05. Все эффекты в модели с *p*-значением, превосходящим пороговое значение, заданное в **Исключать эффекты с p-значениями, большими чем**, удаляются. Значение по умолчанию равно 0.10.
- **Задать максимальное число эффектов в окончательной модели.** По умолчанию все имеющиеся эффекты могут быть включены в модель. Как альтернатива, если шаговый алгоритм, заканчивая работу на некотором шаге, имеет заданное максимальное число эффектов в модели, то он останавливает работу, сохраняя текущий набор эффектов.
- **Задать максимальное число шагов.** Шаговый алгоритм останавливается после определенного числа шагов. По умолчанию это утроенное число имеющихся эффектов. Как альтернатива, задайте положительное целое для максимума числа шагов.

Выбор наилучших подмножеств. Проверяются "все возможные" модели или, по крайней мере, большая совокупность возможных моделей, чем при прямом пошаговом отборе, для выбора наилучших в соответствии с критерием наилучших подмножеств. **Информационный критерий (AICC)** основывается на правдоподобии обучающего множества для данной модели и скорректирован с целью штрафовать излишне сложные модели. **Скорректированный R-квадрат** основывается на точности подгонки для обучающего множества и скорректирован с целью штрафовать излишне сложные модели. **Критерий предотвращения сверхобучения (СКО)** основывается на точности подгонки (среднем квадрате ошибки или СКО) для множества предотвращения сверхобучения. Множество предотвращения сверхобучения представляет собой случайную подвыборку, содержащую приблизительно 30% наблюдений из исходного набора данных, которая не используется при обучении модели.

В качестве наилучшей модели выбирается модель с наибольшим значением критерия.

Примечание: Выбор наилучших подмножеств требует большего объема вычислений, чем прямой шаговый отбор. Когда выполняется выбор наилучших подмножеств в сочетании с бустингом, бэггингом или очень большими наборами данных, то для построения модели потребуется значительно больше времени, чем при построении стандартной модели с использованием прямого пошагового отбора.

Ансамбли

Данные параметры определяют поведение ансамбля, которое имеет место, когда на вкладке Цели запрашивается бэггинг, бустинг или очень большие наборы данных. Параметры, которые не применяются к выбранной цели, игнорируются.

Бэггинг и очень большие наборы данных. Это правило, которое применяется при скоринге ансамбля, чтобы объединить предсказанные значения для базовых моделей с целью вычисления значений скоринга для ансамбля.

- **Принятое по умолчанию правило объединения для непрерывных целевых полей.** Предсказанные значения для ансамбля в случае непрерывных целевых полей могут быть вычислены с использованием среднего значения или медианы предсказанных значений для базовых моделей.

Обратите внимание на то, что если цель состоит в повышении точности модели, выбор правила объединения игнорируется. При бустинге всегда используется взвешенное решение большинством голосов для скоринга категориальных целевых полей и взвешенная медиана для скоринга непрерывных целевых полей.

Бустинг и бэггинг. Задайте число базовых моделей для построения, когда целью является повышение точности или стабильности; для бэггинга это число бутстреп-выборок. Оно должно быть положительным целым.

Дополнительные параметры

Воспроизвести результаты. Задание стартового числа генератора псевдослучайных чисел позволяет воспроизвести результаты. Генератор псевдослучайных чисел используется для выбора записей, попадающих в множество предотвращения сверхобучения. Задайте целое число или щелкните по **Генерировать**, чтобы сгенерировать псевдослучайное целое число в диапазоне между 1 и 2147483647 включительно. Значение по умолчанию - 54752075.

Опции модели

Сохранить предсказанные значения в наборе данных. Именем переменной по умолчанию является *ПредсказанноеЗначение*.

Экспортировать модель. Модель записывается во внешний файл `.zip` . Этот файл модели можно использовать для применения информации о модели к другим файлам данных с целью скоринга. Задайте уникальное допустимое имя файла. Если файл с таким именем уже существует, то он перезаписывается.

Сводка для модели

{\f3 Вид Сводка для модели} {\f4 - } {\f3 это мгновенная визуальная сводка по модели и ее подгонке.}

Таблица. Данная таблица отображает некоторые установки высокого уровня для модели, включая:

- Имя назначения, указанного на вкладке Поля
- Выполнялась ли автоматическая подготовка, заданная в разделе Основные параметры
- Метод и критерий выбора модели, указанные в разделе параметров Выбор модели. Выводится также значение критерия отбора для окончательной модели и представляется в форме "меньше значит лучше".

Диаграмма. Данная диаграмма показывает точность окончательной модели, представленную в форме "больше значит лучше". Ее значение равно 100 × скорректированный R^2 для окончательной модели.

Автоматическая подготовка данных

Этот вид выводит информацию о том, какие поля были исключены и как преобразованные поля были получены на этапе автоматической подготовки данных (ADP). Для каждого поля, которое было преобразовано или исключено, в таблице перечисляется имя поля, его роль в анализе и действие, совершенное на этапе ADP. Поля сортируются в алфавитном порядке имен полей по возрастанию. Возможные действия, выполняемые для каждого поля, включают:

- **Вычислить продолжительность: в месяцах** вычисляет истекшее время в месяцах, исходя из значений в поле, содержащем даты, до текущей системной даты.
- **Вычислить продолжительность: в часах** вычисляет истекшее время в часах, исходя из значений в поле, содержащем время, до текущего системного времени.
- **Сменить тип измерений с непрерывного на порядковый** преобразует непрерывные поля с менее чем 5 различных значений в порядковые поля.
- **Сменить тип измерений с порядкового на непрерывный** преобразует порядковые поля с более чем 10 различных значений в непрерывные поля.
- **Урезать выбросы** заменяет значения непрерывных предикторов, которые лежат вне границ отсечения (определяемых тремя стандартными отклонениями от среднего значения), значением границы отсечения.
- **Заменить пропущенные значения** заменяет пропущенные значения номинальных полей модой, порядковых полей медианой, а непрерывных полей средним значением.
- **Объединить категории для максимизации взаимосвязи с целевым полем** выявляет "похожие" категории предикторов на основе взаимосвязи между входными и целевой переменными. Категории, которые не различаются значимо (т.е. имеющие *p*-значение больше 0,05), объединяются.
- **Исключить предиктор-константу / после обработки пропущенных значений / после объединения категорий** удаляет предикторы, которые имеют единственное значение, вероятно, в результате выполнения дополнительных действий автоматической подготовки данных.

Важность предикторов

Обычно при моделировании сосредотачивают внимание на наиболее важных предикторах и исключают или игнорируют наименее важные. Это помогает сделать диаграмма важности предикторов, показывая относительную важность каждого предиктора при оценке модели. Поскольку значения важности являются относительными, сумма этих значений для всех показанных предикторов равна 1,0. Важность переменных не связана с точностью модели. Она лишь связана с важностью каждого предиктора для предсказания, а не с точностью этого предсказания.

Предсказанные против наблюденных

Выводится диаграмма рассеяния с интервалами для предсказанных значений по вертикальной оси против наблюденных значений по горизонтальной оси. В идеале точки должны лежать на прямой, проведенной под углом 45 градусов. Такое представление позволяет определить, есть ли записи, которые плохо предсказываются моделью.

Остатки

Выводится диагностическая диаграмма модельных остатков.

Стили диаграммы. Имеются различные стили вывода, которые можно выбрать в выпадающем списке **Стиль**.

- **Гистограмма.** Это диаграмма рассеяния с интервалами для стьюдентизированных остатков с наложением нормального распределения. Для линейных моделей предполагается, что остатки имеют нормальное распределение, поэтому в идеале гистограмма должна хорошо аппроксимироваться этой гладкой линией.
- **P-P диаграмма.** Это диаграмма с интервалами типа вероятность-вероятность, сравнивающая распределение стьюдентизированных остатков с нормальным распределением. Если наклон выведенных точек менее крутой, чем наклон нормальной кривой, то остатки показывают большую изменчивость, чем

она должна быть для нормального распределения. Если этот наклон более крутой, то остатки показывают меньшую изменчивость, чем в случае нормального распределения. Если выведенные точки имеют форму S-образной кривой, то распределение остатков является скошенным.

Выбросы

Эта таблица выводит записи, которые оказывают чрезмерное влияние на модель, а также выводит ID записи (если это задано на вкладке Поля), значение целевого поля и расстояние Кука. Расстояние Кука - это мера того, насколько изменились бы остатки для всех записей, если конкретная запись не участвовала бы в вычислении коэффициентов модели. Большое расстояние Кука говорит о том, что исключение записи существенно изменяет коэффициенты, и должна рассматриваться как влияющая.

Влияющие записи должны быть тщательно исследованы, чтобы определить, нужно ли назначить им меньший вес при оценивании модели или урезать резко выделяющиеся значения (выбросы) до некоторого приемлемого порогового значения, или же полностью удалить влияющие записи.

Эффекты

Этот вид показывает величину каждого эффекта в модели.

Стили. Имеются различные стили вывода, которые можно выбрать в выпадающем списке **Стиль**.

- **Диаграмма.** Это диаграмма, в которой эффекты отсортированы сверху вниз по убыванию важности предикторов. Соединяющие линии на диаграмме являются взвешенными на основе значимости эффектов, с большей толщиной линии, соответствующей более значимым эффектам (меньшим *p*-значениям). При наведении указателя мыши на соединительную линию появляется всплывающая подсказка, выводящая *p*-значение и значение важности данного эффекта. Это задано по умолчанию.
- **Таблица.** Это таблица дисперсионного анализа для общих и индивидуальных эффектов модели. Индивидуальные эффекты отсортированы сверху вниз по убыванию важности предикторов. Обратите внимание на то, что по умолчанию таблица сворачивается, чтобы показать только результаты для модели в целом. Чтобы увидеть результаты для индивидуальных эффектов модели, щелкните по **Скорректированная модель** в ячейке таблице.

Важность предикторов. Имеется ползунок важности предикторов, который управляет тем, какие предикторы выводятся. Это не изменяет модели, а просто позволяет сосредоточить внимание на наиболее важных предикторах. По умолчанию выводятся 10 верхних эффектов.

Значимость. Имеется ползунок значимости, предоставляющий дополнительные возможности управлять тем, какие эффекты выводить, кроме тех, которые выводятся на основе значимости предикторов. Эффекты со значениями значимости, превосходящими заданное ползунком значение, скрыты. Это не изменяет модели, а просто позволяет сосредоточить внимание на наиболее важных эффектах. По умолчанию это значение равно 1,00, так что никакие эффекты не отфильтровываются на основе значимости.

Коэффициенты

Этот вид показывает значение каждого коэффициента в модели. Обратите внимание на то, что факторы (категориальные предикторы) имеют индикаторную кодировку в модели, так что **эффекты** , содержащие факторы, обычно будут иметь несколько связанных **коэффициентов**, по одному для каждой категории, исключая категорию, соответствующую избыточному (опорному) параметру.

Стили. Имеются различные стили вывода, которые можно выбрать в выпадающем списке **Стиль**.

- **Диаграмма.** Это диаграмма, в которой сначала выводится свободный член, а затем эффекты, отсортированные сверху вниз по убыванию важности предикторов. Внутри эффектов, содержащих факторы, коэффициенты сортируются в порядке возрастания значений данных. Соединяющие линии на диаграмме раскрашены в зависимости от знака коэффициента (см. ключ диаграммы) и взвешены в зависимости от значимости коэффициента, с большей толщиной линии, соответствующей более

значимым коэффициентам (меньшим *p*-значениям). При наведении указателя мыши на соединительную линию появляется всплывающая подсказка, выводящая значение коэффициента, *p*-значение для него, а также значение важности эффекта, с которым связан этот параметр. Это задано по умолчанию.

- **Таблица.** В этой таблице выводятся значения, результаты тестов на значимость и доверительные интервалы для индивидуальных коэффициентов модели. После свободного члена эффекты отсортированы сверху вниз по убыванию важности предикторов. Внутри эффектов, содержащих факторы, коэффициенты сортируются в порядке возрастания значений данных. Обратите внимание на то, что по умолчанию таблица сворачивается, чтобы вывести только коэффициент, значимость и важность для каждого параметра модели. Чтобы увидеть стандартную ошибку, *t*-статистику и доверительный интервал, щелкните по ячейке **Коэффициент** в таблице. При наведении указателя мыши на имя параметра модели в таблице появляется всплывающая подсказка, выводящая имя параметра, эффект, с которым связан этот параметр, и (для категориальных предикторов) метки значений, связанных с данным параметром модели. Это, в частности, позволяет увидеть новые категории, созданные, когда автоматическая подготовка данных привела к объединению сходных категорий категориального предиктора.

Важность предикторов. Есть ползунок важности предикторов, который управляет тем, какие предикторы выводятся. Это не изменяет модели, а просто позволяет сосредоточить внимание на наиболее важных предикторах. По умолчанию выводятся 10 верхних эффектов.

Значимость. Есть ползунок значимости, предоставляющий дополнительные возможности управлять тем, какие коэффициенты выводить, кроме тех, которые выводятся на основе значимости предикторов. Коэффициенты со значениями значимости, превосходящими заданное ползунком значение, скрыты. Это не изменяет модели, а просто позволяет сосредоточить внимание на наиболее важных коэффициентах. По умолчанию это значение равно 1,00, так что никакие коэффициенты не отфильтровываются на основе значимости.

Оцененные средние

Это диаграммы, выводимые для значимых предикторов. На диаграмме вдоль вертикальной оси выводится оцененное по модели значение целевой переменной для каждого значения предиктора на горизонтальной оси при сохранении значений всех остальных предикторов неизменными. Это дает полезную визуализацию того, какое влияние коэффициент каждого предиктора оказывает на целевую переменную.

Примечание: если нет значимых предикторов, оцененные средние не генерируются.

Сводка по построению модели

Эта панель предоставляет некоторые детали процесса построения модели, когда в группе параметров Подбор модели сделан выбор алгоритма отбора, отличный от **Включить все предикторы** .

Прямой шаговый. Если алгоритмом отбора является прямой шаговый, то в таблице выводятся последние 10 шагов шагового алгоритма. На каждом шаге показываются значение критерия отбора и эффекты в модели. Это дает понимание того, какой вклад в модель дает каждый шаг. В каждом столбце можно сортировать строки, чтобы было легче видеть, какие эффекты содержатся в модели на каждом шаге.

Наилучшие подмножества. Если алгоритмом отбора является "наилучшие подмножества", то таблица выводит 10 лучших моделей. Для каждой модели показываются значение критерия отбора и эффекты в модели. Это позволяет проверить стабильность лучших моделей. Если для них наблюдается тенденция иметь много схожих эффектов с небольшими различиями, то наилучшей модели можно вполне доверять. Если для них наблюдается тенденция иметь сильно различающиеся эффекты, то некоторые из этих эффектов могут быть слишком схожи между собой, и их следует объединить (или один удалить). В каждом столбце можно сортировать строки, чтобы было легче видеть, какие эффекты содержатся в модели на каждом шаге.

Глава 16. Линейная регрессия

Линейная регрессия оценивает коэффициенты линейного уравнения, содержащего одну или несколько независимых переменных, позволяющие наилучшим образом предсказать значение зависимой переменной. Например, вы можете попытаться предсказать объем годовых продаж для сотрудника отдела продаж (зависимая переменная) по таким независимым переменным, как возраст, образование и стаж работы.

Пример. Связано ли число матчей, выигранных за сезон баскетбольной командой, со средним количеством очков, набранных ей в каждом матче? Диаграмма рассеяния показывает, что эти переменные линейно связаны. Количество выигранных матчей и среднее число очков, набранное соперником, также линейно связаны между собой. Эти переменные имеют отрицательную связь. При росте количества выигранных матчей, среднее число очков, набранных соперником, уменьшается. С помощью линейной регрессии вы можете смоделировать зависимость этих переменных. Хорошую модель можно использовать для предсказания числа матчей, которые выиграют команды.

Статистика. Для каждой переменной: число наблюдений без пропущенных значений, среднее значение и среднеквадратичное отклонение. коэффициенты регрессии, матрица корреляций, частичные и частные корреляции, множественный R, R^2 скорректированный R^2, изменение R^2, среднеквадратическая ошибка оценки, таблица дисперсионного анализа, предсказанные значения и остатки. Также выдаются: 95%-е доверительные интервалы для каждого коэффициента регрессии, матрица ковариаций, коэффициент разбухания дисперсии (variance inflation factor), статистика допуска (толерантность), критерий Дарбина-Уотсона, меры расстояния (Махаланобиса, Кука и величина плеча), DfBeta, DfFit, интервалы предсказания, диагностическая информация по наблюдениям. Графики: диаграммы рассеяния, частные графики, гистограммы и нормальные вероятностные графики.

Данные для линейной регрессии

Данные. Зависимая и независимые переменные должны быть количественными. Категориальные переменные, такие как религия, основная область исследования, регион проживания, должны быть перекодированы в бинарные (фиктивные) переменные или в другие типы переменных контрастов.

Допущения. Для каждого значения независимой переменной распределение зависимой переменной должно быть нормальным. Дисперсия распределения зависимой переменной должна быть постоянной для каждого значения независимой переменной. Взаимосвязи между зависимой и каждой из независимых переменных должны быть линейными, и все наблюдения должны быть независимыми.

Чтобы выполнить линейный регрессионный анализ

1. Выберите в меню:

   Анализ > Регрессия > Линейная...
2. В диалоговом окне Линейная регрессия выберите числовую зависимую переменную.
3. Выберите одну или несколько числовых независимых переменных.

Дополнительно вы можете:

- Объединять независимые переменные в блоки и задавать разные методы отбора переменных для разных подмножеств переменных.
- Выбирать переменную отбора наблюдений для того, чтобы ограничить анализ подмножеством наблюдений, имеющих конкретные значения этой переменной.
- Выбирать переменную для идентификации наблюдений (точек) на графиках.
- Выбрать числовую переменную весов для применения взвешенного метода наименьших квадратов.

ВМНК. Позволяет получить взвешенную модель методом наименьших квадратов. Вес точки данных равен обратной величине ее дисперсии. Это означает, что чем больше дисперсия наблюдения, тем слабее оно влияет на результат. Если значение взвешивающей переменной равно нулю, отрицательно, или пропущено, наблюдение исключается из анализа.

Методы отбора переменных для линейной регрессии

Выбор метода отбора позволяет задать то, каким образом независимые переменные включаются в анализ. Используя различные методы, вы можете построить целый ряд регрессионных моделей для одного и того же набора переменных.

- *Ввод (регрессия)*. Процедура для выбора переменной, когда все переменные в блоке вводятся на одном шаге.
- *Пошаговый*. На каждом шаге в уравнение включается новая независимая переменная с наименьшей вероятностью F, при условии, что эта вероятность достаточно мала. Переменные, уже введенные в регрессионное уравнение, исключаются из него, если их вероятность F становится достаточно большой. Алгоритм останавливается, когда не остается переменных, удовлетворяющих критерию включения или исключения.
- *Удалить*. Процедура отбора переменных, при которой все переменные блока исключаются на одном шаге.
- *Отбор исключением*. Процедура отбора переменных, при которой все переменные вводятся в уравнение, а затем последовательно исключаются из него. Первым кандидатом на удаление считается переменная, имеющая наименьшую частную корреляцию с зависимой переменной. Если она удовлетворяет критерию исключения, ее удаляют. Следующим кандидатом на исключение становится переменная, имеющая наименьшую среди оставшихся переменных частную корреляцию с зависимой переменной. Процедура останавливается, когда не остается переменных, удовлетворяющих критерию исключения.
- *Отбор включением*. Шаговая процедура отбора переменных, при которой переменные последовательно включаются в модель. Первым кандидатом на ввод служит переменная с наибольшим модулем корреляции с зависимой переменной. Если эта переменная удовлетворяет критерию ввода, она включается в модель. Если первая переменная включена в модель, то следующим кандидатом на включение среди оставшихся вне модели переменных становится переменная, имеющая наибольшую частную корреляцию. Процедура останавливается, когда не остается переменных, удовлетворяющих критерию ввода.

Значения значимостей в выводе результатов основаны на подгонке единственной модели. Поэтому значения значимостей, как правило, некорректны при применении шагового метода (Шаговый отбор, Включение или Исключение).

Вне зависимости от выбранного метода отбора, каждая переменная должна удовлетворять критерию допуска (толерантности) для того, чтобы быть введенной в уравнение. По умолчанию, значение уровня толерантности (допуска) равно 0.0001. Кроме того, переменная не будет введена в модель, если это повлечет за собой снижение толерантности переменной, уже введенной в уравнение, до величины, меньшей, чем значение критерия допуска.

Все отобранные независимые переменные будут добавлены в одну регрессионную модель. Однако, вы можете задавать различные методы ввода переменных для разных наборов переменных. Например, вы можете включить один блок переменных в регрессионную модель методом Шагового отбора, а другой блок - методом Включение. Чтобы добавить в регрессионную модель второй блок переменных, нажмите кнопку **След** .

Задание правила отбора наблюдений для линейной регрессии

В анализе используются наблюдения, отобранные с помощью правила отбора наблюдений. Например, если вы зададите переменную, выберете **равно** и введете 5 в качестве значения, то в анализе будут участвовать только те наблюдения, для которых значение заданной переменной равно 5. Допускается также текстовое значение.

Графики процедуры Линейная регрессия

Графики могут помочь при проверке предположений о нормальности, линейности и равенстве дисперсий. Графики полезны также для выявления выбросов, необычных наблюдений и влияющих наблюдений. Сохраненные в качестве новых переменных предсказанные значения, остатки и другая диагностическая информация становятся доступными в Редакторе данных. Их можно использовать в сочетании с независимыми переменными для построения графиков. Можно построить следующие графики:

Диаграммы рассеяния. Можно строить диаграммы для любой пары переменных из следующего списка: зависимая переменная, стандартизованные предсказанные значения, стандартизованные остатки, удаленные остатки, скорректированные предсказанные значения, стьюдентизированные остатки, стьюдентизированные удаленные остатки. Для проверки линейности и равенства дисперсий строится график стандартизованных остатков против стандартизованных предсказанных значений.

Список исходных переменных. Перечисляет для зависимой переменной (DEPENDNT) следующие предсказанные переменные и переменные остатка: стандартизованные предсказанные значения (*ZPRED), стандартизованные остатки (*ZRESID), удаленные остатки (*DRESID), скорректированные предсказанные значения (*ADJPRED), остатки по Стьюденту (*SRESID), удаленные остатки по Стьюденту (*SDRESID).

Выдать все частные графики. Выводятся диаграммы рассеяния остатков для всех пар переменных, состоящих из зависимой переменной и одной независимой переменной. Остатки получаются при раздельном построении регрессионных моделей для каждой переменной из пары по всем остальным независимым переменным. Чтобы был построен частный график, в регрессионное уравнение должны быть включены, по крайней мере, две независимые переменные.

Графики стандартизованных остатков. Вы можете построить гистограммы стандартизованных остатков и нормальные вероятностные графики, сравнивающие распределение стандартизованных остатков с нормальным распределением.

Если задан вывод каких-либо графиков, выдаются итожащие статистики для стандартизованных предсказанных значений и стандартизованных остатков (***ZPRED** и ***ZRESID**).

Линейная регрессия: Сохранение новых переменных

Предсказанные значения, остатки и другие статистики, полезные для диагностической информации, можно сохранить. Выбор каждого из перечисленных ниже пунктов добавляет к активному файлу данных одну или несколько переменных.

Предсказанные значения. Значения, которые регрессионная модель предсказывает для каждого наблюдения.

- *Нестандартизованные*. Значение зависимой переменной, предсказываемое в соответствии с моделью.
- *Стандартизованные*. Преобразование каждого предсказанного значения в стандартизованную форму. То есть, из каждого предсказанного значения вычитают среднее предсказанное значение, и полученную разность делят на стандартное отклонение предсказанного значения. Среднее стандартизованных предсказанных значений равно 0, а стандартное отклонение 1.
- *Скорректированное*. Предсказываемое значение для наблюдения, при условии, что это наблюдение не используется при вычислении коэффициентов регрессии.
- *Среднекв. ошибка средних*. Стандартные ошибки предсказанных значений. Оценка стандартного отклонения среднего значения зависимой переменной для наблюдений с одинаковыми значениями независимых переменных.

Расстояния. Меры, выявляющие наблюдения с необычными комбинациями значений независимых переменных и наблюдения, которые могут оказать большое влияние на регрессионную модель.

- *Махаланобиса*. Мера того, насколько значения наблюдений для независимых переменных отклоняются от среднего по всем наблюдениям. Большое расстояние Махаланобиса означает, что наблюдение содержит экстремальные значения в одной или более независимых переменных.

- *Кука*. Для каждого наблюдения показывает насколько изменятся остатки всех наблюдений, если это наблюдение не использовать при вычислении коэффициентов регрессии. Большое расстояние Кука указывает на то, что исключение данного наблюдения из вычислений регрессии существенно меняет коэффициенты.
- *Значения разбалансировки*. Измеряют влияние точки на согласие регрессионной модели. Центрированные балансировки изменяются от 0 (не влияет) до (N-1)/N.

Интервалы предсказания. Верхние и нижние границы интервалов предсказания для среднего и отдельного значения.

- *Mean*. Нижняя и верхняя границы (две переменные) интервала предсказания для среднего предсказываемого отклика.
- *Отдельное значение*. Нижняя и верхняя границы (две переменные) для интервала предсказания зависимой переменной для отдельного наблюдения.
- *Доверительный интервал*. Введите значение от 1 до 99,99, чтобы задать доверительный уровень для двух интервалов предсказания. Перед вводом этого значения необходимо выбрать Среднее или Отдельное значение. Типичные значения доверительного уровня - 90, 95 и 99.

Остатки. Фактическое значение зависимой переменной минус предсказанное регрессионным уравнением.

- *Нестандартизованные*. Разность между наблюдаемым и предсказанным моделью значением.
- *Стандартизованные*. Остаток, деленный на оценку его стандартного отклонения. Стандартизованные остатки, известные еще как пирсоновские, имеют среднее 0 и стандартное отклонение 1.
- *Стьюдентизированные*. Остаток, деленный на оценочное значение его среднеквадратичного отклонения, которое изменяется от наблюдения к наблюдению в зависимости от расстояния значений каждого наблюдения для независимых переменных от средних значений этих переменных.
- *Удалено*. Остаток для наблюдения, когда данное наблюдение исключается при вычислении регрессионных коэффициентов. Это разность между значением зависимой переменной и скорректированным предсказанным значением.
- *Стьюдентизированные удаленные*. Остаток для удаленного наблюдения, деленный на его стандартную ошибку. Разность между стьюдентизированным остатком с удалением и соответствующим ему стьюдентизированным остатком указывает, насколько сильно исключение наблюдения влияет на предсказание для него самого.

Статистики влияния. Изменение в регрессионных коэффициентах (DfBeta) и предсказанных значениях (DfFit), вызванное исключением из анализа конкретного наблюдения. Доступны также стандартизованные значения DfBeta и DfFit вместе с ковариационным отношением.

- *DfBeta(s)*. Разница в значении бета - это изменение регрессионного коэффициента в результате исключения отдельного наблюдения. Значение вычисляется для каждого компонента модели, включая свободный член.
- *Стандартизованные DfBeta* . Стандартизованная разность значений бета. Изменение коэффициента регрессии при исключении отдельного наблюдения. Имеет смысл исследовать наблюдения, у которых модуль этого значения, больше, чем 2/корень квадратный из (N), где N - число наблюдений. Значение вычисляется для каждого компонента модели, включая свободный член.
- *DfFit*. Разница в величине подгонки - это изменение предсказанного значения в результате исключения отдельного наблюдения.
- *Стандартизованные DfFit*. Стандартизованная разность предсказанных значений. Изменение предсказанного значения при исключении отдельного наблюдения. Имеет смысл исследовать наблюдения, у которых модуль этого значения больше, чем 2 * корень квадратный из (p/N), где p - число параметров в модели, а N - число наблюдений.
- *Ковариационное отношение*. Отношение определителя ковариационной матрицы, вычисленного без данного наблюдения, к определителю ковариационной матрицы, вычисленной для всей выборки. Если это отношение близко к 1, данное наблюдение не влияет на ковариационную матрицу существенно.

Статистики коэффициентов Сохраняет коэффициенты регрессии в наборе данных или файле данных. Наборы данных доступны для последующего использования в том же сеансе но не сохраняются как файлы до тех пор, пока они не будут сохранены явно до окончания текущего сеанса. Имена наборов данных должны удовлетворять требованиям к именам переменных.

Экспортировать модель в формате XML Оценки параметров и их ковариации (если помечено) экспортируются в специальный файл в формате XML (PMML). Этот файл модели можно использовать для применения информации о модели к другим файлам данных с целью скоринга.

Статистики процедуры Линейная регрессия

Доступны следующие статистики:

Коэффициенты регрессии. Оценки - Установка этого переключателя позволяет вывести коэффициент регрессии *B*, стандартную ошибку коэффициента *B*, стандартизованный коэффициент бета, *t*-значение для *B* и двусторонний уровень значимости для *t*. Установка переключателя **Доверительные интервалы** позволяет вывести доверительные интервалы с указанным уровнем доверия для каждого регрессионного коэффициента или ковариационной матрицы. Установка переключателя **Матрица ковариаций** выводит матрицу дисперсий-ковариаций оценок регрессионных коэффициентов с дисперсиями на диагонали и с ковариациями вне ее. Также выводится корреляционная матрица.

Согласие модели. Перечисляются переменные, включаемые в модель и исключаемые из нее, и выдаются следующие статистики согласия: множественный коэффициент *R*, R^2 и скорректированный R^2, стандартная ошибка оценки и таблица дисперсионного анализа.

Изменение R-квадрат. Изменение статистики R^2, вызванное добавлением или удалением независимой переменной. Если изменение R^2, связанное с переменной, велико, считается, что эта переменная - хороший предиктор зависимой переменной.

Описательные статистики. Выдается число наблюдений без пропущенных значений, среднее значение и стандартное отклонение для каждой анализируемой переменной. Выводятся также корреляционная матрица с односторонним уровнем значимости и числом наблюдений для каждой корреляции.

Частная корреляция. Корреляция между двумя переменными, оставшаяся после удаления корреляции, относящейся к их общей связи с другими переменными. Корреляция между зависимой и независимой переменной, когда из них исключены линейные эффекты других независимых переменных модели.

Частичная корреляция. Корреляция между зависимой переменной и независимой переменной, вычисленная после того, как из независимой переменной удалена линейная связь с остальными независимыми переменными в модели. Она связана с изменением R-квадрат, когда переменная добавляется в уравнение. Иногда она называется получастной корреляцией.

Диагностика коллинеарности. Коллинеарность (или мультиколлинеарность) - это нежелательная ситуация, когда одна независимая переменная является линейной комбинацией других независимых переменных. Выводятся собственные значения масштабированной и нецентрированной матрицы сумм перекрестных произведений, показатели обусловленности, доли в разложении дисперсии, а также коэффициенты разбухания дисперсии (VIF - variance inflation factor), толерантности (допуски) для отдельных переменных.

Остатки. Выводится критерий Дарбина-Уотсона сериальной корреляции остатков и поотчетная информация диагностики для наблюдений, удовлетворяющих критерию отбора (выбросы свыше *n* среднеквадратических отклонений).

Параметры процедуры Линейная регрессия

Доступны следующие параметры:

Критерий шагового метода. Эти параметры применяются, если в качестве метода отбора выбрано Включение, Исключение либо Шаговый отбор. Переменные могут быть введены в модель или исключены из модели на основе либо значимости (вероятности) *F* -статистики, либо самого значения *F* -статистики.

- *Использование вероятности F*. Переменная вводится в модель, если наблюдаемый уровень значимости ее F-значения меньше заданного порога включения, и исключается, если этот уровень значимости больше порога исключения. Порог включения должен быть меньше порога удаления, они оба должны быть положительными. Если необходимо включить в модель больше переменных, увеличьте порог включения. Чтобы исключить из модели большее число переменных, снизьте порог исключения.
- *Использование значения F*. Переменная вводится в модель, если ее F-значение превышает заданное значение включения, и исключается, если ее F-значение меньше значения исключения. Значение включения должно превосходить значение удаления, оба должны быть положительными. Если необходимо ввести в модель больше переменных, снизьте порог включения. Чтобы исключить из модели большее число переменных, увеличьте порог исключения.

Включить в уравнение константу. По умолчанию регрессионная модель содержит свободный член - константу. Если удалить этот переключатель, линия регрессии будет проходить через начало координат, что используется редко. Некоторые результаты для регрессии, проходящей через начало координат, несравнимы с результатами регрессии, содержащей константу. Например, R^2 для регрессии, проходящей через начало координат, невозможно интерпретировать обычным образом.

Пропущенные значения. Вы можете выбрать один из следующих вариантов:

- **Исключать целиком.** В анализ включаются только наблюдения без пропущенных значений для всех анализируемых переменных.
- **Исключать попарно.** При вычислении коэффициентов корреляции, применяемых в процедуре регрессии, используются только те наблюдения, у которых для данной пары переменных оба значения не пропущены. Числа степеней свободы основаны на минимальном попарном *N* .
- **Заменить средним.** Для вычислений используются все наблюдения, а пропущенные значения заменяются средним значением этой переменной.

Команда REGRESSION: дополнительные возможности

Язык синтаксиса команд также позволяет:

- Сохранять матрицу корреляций или считывать матрицу вместо исходных данных для выполнения регрессионного анализа (с помощью подкоманды `MATRIX`).
- Задавать уровни толерантности (с помощью подкоманды `CRITERIA`).
- Получать несколько моделей для одной и той же или разных зависимых переменных (с помощью подкоманд `METHOD` и `DEPENDENT` .)
- Получать дополнительные статистики (с помощью подкоманд `DESCRIPTIVES` и `STATISTICS` .)

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Глава 17. Порядковая

Порядковая регрессия позволяет моделировать зависимость политомического порядкового отклика от набора предикторов, которые могут быть факторами или ковариатами. Реализация процедуры Порядковая регрессия основывается на методологии Маккалага (McCullagh (1980, 1998)), и эта процедура в языке команд называется PLUM.

Стандартный линейный регрессионный анализ включает минимизацию суммы квадратов разностей между переменной отклика (зависимой) и взвешенной комбинацией предикторных (независимых) переменных. Оцененные коэффициенты отражают, насколько изменения значений предикторов влияет на отклик. Предполагается, что отклик является числовым в том смысле, что изменения уровня отклика эквивалентны для всего диапазона значений отклика. Например, различие в росте между человеком ростом 150 см и человеком ростом 140 см составляет 10 см, которое имеет то же значение, что и различие в росте между человеком ростом 210 см и человеком ростом 200 см. Это свойство необязательно справедливо для порядковых переменных, для которых выбор категорий отклика и их числа может быть весьма произвольным.

Пример. Порядковую регрессию можно использовать для изучения реакции пациента на дозировку лекарственного препарата. Возможные реакции можно классифицировать как *отсутствие*, *слабая*, *умеренная* или *сильная*. Различие между слабой и умеренной реакциями трудно либо невозможно выразить количественно, и оно зависит от восприятия. Более того, различие между слабой и умеренной реакциями может быть больше или меньше, чем различие между умеренной и сильной реакциями.

Статистики и графики. Наблюденные и ожидаемые частоты, а также накопленные частоты, остатки Пирсона для частот и накопленных частот, наблюденные и ожидаемые вероятности, наблюденные и ожидаемые накопленные вероятности каждой категории отклика по наборам значений, которые принимали ковариаты, асимптотические ковариационная и корреляционная матрицы оценок параметров, хи-квадрат Пирсона и хи-квадрат отношения правдоподобия, статистики согласия, хронология итераций, проверка предположения о параллельности линий, оценки параметров, стандартные ошибки, доверительные интервалы, а также статистики Кокса и Снелла, Нэйджелкерка и R^2 Макфаддена.

Данные для порядковой регрессии

Данные. Предполагается, что зависимая переменная является порядковой и может быть числовой или текстовой. Упорядочение определяется сортировкой значений зависимой переменной в порядке возрастания. Наименьшее значение задает первую категорию. Предполагается, что факторные переменные являются категориальными. Переменные ковариат должны быть числовыми. Обратите внимание на то, что использование более чем одной непрерывной ковариаты может легко привести к созданию очень большой таблицы вероятностей ячеек.

Допущения. Допускается только одна переменная отклика, и она должна быть задана. Кроме того, предполагается, что для всех различающихся наборов значений независимых переменных отклики являются независимыми полиномиальными переменными.

Родственные процедуры. Номинальная логистическая регрессия использует аналогичные модели для номинальных зависимых переменных.

Получение порядковой регрессии

1. Выберите в меню:

   Анализ > Регрессия > Порядковая...
2. Выберите одну зависимую переменную.
3. Щелкните по **OK**.

Порядковая регрессия: параметры

Диалоговое окно Параметры позволяет настроить параметры, используемые в итерационном алгоритме оценивания, выбрать уровень доверительных интервалов, а также функцию связи.

Итерации. Итерационный алгоритм можно настроить.

- **Максимум итераций.** Задайте неотрицательное целое число. Если задан 0, процедура возвращает начальные оценки.
- **Максимальное число шагов половинного деления.** Задайте целое положительное число.
- **Сходимость Log-правдоподобия.** Алгоритм останавливается, если абсолютное или относительное изменение log-правдоподобия меньше этого значения. Данный критерий не применяется, если задан 0.
- **Сходимость параметров.** Алгоритм останавливается, если абсолютное или относительное изменение каждой из оценок параметров меньше этого значения. Данный критерий не применяется, если задан 0.

Доверительный интервал. Задайте значение, большее или равное 0 и меньшее 100.

Дельта. Значение, прибавляемое к нулевым частотам в ячейках. Задайте неотрицательное значение, меньшее 1.

Допуск для вырожденности. Используется для проверки наличия сильной зависимости предикторов. Выберите значение из списка возможных значений.

Функция связи. Функция связи служит для преобразования кумулятивных вероятностей для расчета модели. Доступны следующие пять функций связи.

- **Логит.** $f(x)=\log(x / (1-x))$. Обычно используется для равномерно распределенных категорий.
- **Дополнительный логарифм-логарифм.** $f(x)=\log(-\log(1-x))$. Обычно используется, когда высшие категории более вероятны.
- **Отрицательный Log-log.** $f(x)=-\log(-\log(x))$. Обычно используется, когда низшие категории более вероятны.
- **Пробит.** $f(x)=\Phi^{-1}(x)$. Обычно используется, когда скрытая переменная равномерно распределена.
- **Коши (обратное Коши).** $f(x) = \tan(\pi (x - 0{,}5))$. Обычно используется, когда скрытая переменная имеет много экстремальных значений.

Порядковая регрессия: вывод

Диалоговое окно Вывод позволяет создать таблицы для просмотра в средстве просмотра и сохранить переменные в рабочем файле.

Вывод. Здесь можно задать вывод следующих таблиц:

- **Выводить историю итераций.** Печатаются log-правдоподобие и оценки параметров с заданной частотой повторения печати. Первая и последняя итерации печатаются всегда.
- **Статистики согласия.** Статистики хи-квадрат Пирсона и хи-квадрат отношения правдоподобия. Они вычисляются на основе классификации, заданной в списке переменных.
- **Итожащие статистики.** Статистики Кокса и Снелла, Нэйджелкерка, а также статистика R^2 Макфаддена.
- **Оценки параметров.** Оценки параметров, стандартные ошибки и доверительные интервалы.
- **Асимптотическая корреляция оценок параметров.** Матрица корреляций оценок параметров.
- **Асимптотическая ковариация оценок параметров.** Матрица ковариаций оценок параметров.
- **Информация по ячейкам.** Наблюденные и ожидаемые частоты, а также накопленные частоты, остатки Пирсона для частот и накопленных частот, наблюденные и ожидаемые вероятности, а также наблюденные и ожидаемые накопленные вероятности каждой категории отклика по наборам значений, которые принимали ковариаты. Обратите внимание на то, что при построении моделей с использованием

большого числа наблюдений с различающимися значениями ковариат (например, моделей с непрерывными ковариатами), применение данной возможности может привести к созданию очень большой, громоздкой таблицы.

- **Проверка параллельности линий.** Проверяется гипотеза о том, что параметры положения эквивалентны по всем уровням зависимой переменной. Это возможно для моделей, имеющих только компонент положения.

Сохраняемые переменные. В рабочем файле сохраняются следующие переменные:

- **Оцененные вероятности отклика.** Оцененные по модели вероятности классификации по категориям отклика для наборов значений, которые принимались факторами и ковариатами. Число вероятностей равно числу категорий отклика.
- **Предсказанная категория.** Категория отклика, имеющая наибольшую оцененную вероятность для набора значений, принимаемых факторами и ковариатами.
- **Вероятность предсказанной категории.** Оцененная вероятность для отклика попасть в предсказанную категорию для набора значений, принимаемых факторами и ковариатами. Эта вероятность также является максимумом оцененных вероятностей для данного набора значений факторов и ковариат.
- **Вероятность действительной категории.** Оцененная вероятность для отклика попасть в действительную категорию для набора значений, принимаемых факторами и ковариатами.

Выводить log-правдоподобие. Управляет выводом log-правдоподобия. **Включая полиноминальную константу** дает полное значение правдоподобия. Для того чтобы сравнить полученные результаты по произведениям, не включающим константу, можно выбрать ее исключение.

Порядковая регрессия: модель положения

Диалоговое окно Положение позволяет задать для анализа модель положения.

Задать модель. Модель главных эффектов включает главные эффекты ковариат и факторов, но не включает взаимодействия. Можно сформировать модель специального вида, включив в нее нужные подмножества взаимодействий факторов или взаимодействий ковариат.

Факторы/ковариаты. Перечисляются факторы и ковариаты.

Модель положения. Эта модель зависит от выбранных главных эффектов и эффектов взаимодействия.

Создать члены

Для выбранных факторов и ковариат:

Взаимодействие. Создается член взаимодействия наивысшего порядка всех выбранных переменных. Это вариант по умолчанию.

Главные эффекты. Создаются главные эффекты для всех выбранных переменных.

Все 2-факторные. Создаются все возможные двухфакторные взаимодействия выбранных переменных.

Все 3-факторные. Создаются все возможные трехфакторные взаимодействия выбранных переменных.

Все 4-факторные. Создаются все возможные четырехфакторные взаимодействия выбранных переменных.

Все 5-факторные. Создаются все возможные пятифакторные взаимодействия выбранных переменных.

Порядковая регрессия: модель масштаба

Диалоговое окно Масштаб позволяет задать для анализа модель масштаба.

Факторы/ковариаты. Перечисляются факторы и ковариаты.

Модель масштаба. Эта модель зависит от выбранных главных эффектов и эффектов взаимодействия.

Создать члены

Для выбранных факторов и ковариат:

Взаимодействие. Создается член взаимодействия наивысшего порядка всех выбранных переменных. Это вариант по умолчанию.

Главные эффекты. Создаются главные эффекты для всех выбранных переменных.

Все 2-факторные. Создаются все возможные двухфакторные взаимодействия выбранных переменных.

Все 3-факторные. Создаются все возможные трехфакторные взаимодействия выбранных переменных.

Все 4-факторные. Создаются все возможные четырехфакторные взаимодействия выбранных переменных.

Все 5-факторные. Создаются все возможные пятифакторные взаимодействия выбранных переменных.

Команда PLUM: дополнительные возможности

В задание на выполнение процедуры порядковой регрессии можно внести изменения путем передачи его в окно синтаксиса и редактирования полученного синтаксиса команды PLUM. Язык синтаксиса команд также позволяет:

- Формировать гипотезы для проверки путем задания нулевых гипотез, включающих линейные комбинации параметров.

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Глава 18. Подгонка кривых

Процедура Подгонка кривых позволяет вычислять статистики и строить сопутствующие графики для 11 различных регрессионных моделей оценки кривых. Для каждой зависимой переменной будет построена отдельная модель. Вы также можете сохранять предсказанные значения, остатки и интервалы прогноза в виде новых переменных.

Пример. Провайдер услуг Интернета отслеживает во времени процент зараженного вирусом почтового трафика в своих сетях. Диаграмма рассеивания обнаруживает нелинейную зависимость. Вы можете подогнать к данным квадратичную или кубическую модель, а также проверить выполнение предположений модели и степень ее согласия.

Статистика. Для каждой модели: коэффициенты регрессии, множественный коэффициент R , R^2 , скорректированный R^2 , стандартная ошибка оценки, таблица дисперсионного анализа, предсказанные значения, остатки и интервалы прогноза. Модели: линейная, логарифмическая, обратная, квадратичная, кубическая, степенная, составная, S-кривая, логистическая, роста и экспоненциальная.

Данные для процедуры Подгонка кривых

Данные. Зависимая и независимые переменные должны быть количественными. Если в качестве независимой переменной выбрано **Время** , а не переменная из активного набора данных, процедура Подгонка кривых создаст переменную типа время с одинаковыми периодами времени между наблюдениями. Если выбрано **Время** , то зависимая переменная должна представлять собой временной ряд. Для анализа временных рядов необходима такая структура файла данных, в которой каждое наблюдение (строка) представляет набор измерений, сделанных в момент времени, отличный от моментов времени других наблюдений, с одинаковыми периодами времени между соседними наблюдениями.

Допущения. Данные проверяются в графическом режиме, чтобы определить, как связаны между собой независимая и зависимая переменные (линейно, экспоненциально и т.д.). Остатки для хорошей модели должны быть распределены случайным образом и подчиняться нормальному распределению. При использовании линейной модели необходимо выполнение следующих условий: Для каждого значения независимой переменной распределение зависимой переменной должно быть нормальным. Дисперсия распределения зависимой переменной должна быть постоянной для каждого значения независимой переменной. Взаимосвязь между зависимой и независимой переменными должна быть линейной, а все наблюдения должны быть независимыми.

Чтобы запустить процедуру Подгонка кривых

1. Выберите в меню:

   Анализ > Регрессия > Подгонка кривых...
2. Выберите одну или несколько зависимых переменных. Для каждой зависимой переменной будет построена отдельная модель.
3. Выберите независимую переменную (либо переменную из активного набора данных, либо **Время**).
4. Дополнительно можно:

- Выбрать переменную, значения которой задают метки наблюдений в диаграммах рассеяния. Для каждой точки на диаграмме рассеяния использовать инструмент Идентификатор точек, чтобы вывести значение переменной, помещенной в поле Метки наблюдений.
- Щелкнуть мышью по кнопке **Сохранить** , чтобы сохранить предсказанные значения, остатки и интервалы прогноза в качестве новых переменных.

Доступны также следующие параметры:

- **Включить в уравнение константу.** Выполняется оценка свободного члена в уравнении регрессии. Свободный член включается в уравнение по умолчанию.
- **Графики моделей.** Для каждой выбранной модели выводится график значений зависимой переменной от значений независимой переменной. Для каждой зависимой переменной выводится отдельный график.
- **Вывести таблицу дисперсионного анализа.** Для каждой выбранной модели выводится сводная таблица дисперсионного анализа.

Модели подгонки кривых

Вы можете выбрать одну или несколько регрессионных моделей подгонки кривых. Чтобы определить, какую модель использовать, выведите данные графически. Если окажется, что переменные связаны линейно, используйте простую модель линейной регрессии. Если переменные не являются связанными линейно, попробуйте преобразовать ваши данные. Если преобразование не поможет, то, возможно, необходимо применение более сложной модели. Посмотрите на диаграмму рассеяния данных. Если диаграмма напоминает известную вам математическую функцию, используйте модель соответствующего типа для подгонки к данным. Например, если данные на диаграмме напоминают экспоненту, используйте экспоненциальную модель.

Линейная. Модель, задаваемая уравнением Y = b0 + (b1 * t). Значения ряда моделируются линейной функцией времени.

Логарифмическая. Модель с уравнением Y = b0 + (b1 * ln(t)).

Обратная. Модель, задаваемая уравнением Y = b0 + (b1 / t).

Квадратичная регрессия. Модель, задаваемая уравнением Y = b0 + (b1 * t) + (b2 * t**2). Квадратичная модель может применяться в качестве одной из альтернатив линейной модели, например, когда в ограниченном диапазоне значений наблюдается рост, более быстрый, чем линейный.

Кубическая регрессия. Модель, определяемая уравнением Y = b0 + (b1 * t) + (b2 * t**2) + (b3 * t**3).

Степенная. Модель с уравнением Y = b0 * (t**b1) или ln(Y) = ln(b0) + (b1 * ln(t)).

Составная. Модель, задаваемая уравнением Y = b0 * (b1**t) или ln(Y) = ln(b0) + (ln(b1) * t).

S-кривая. Модель, задаваемая уравнением Y = e**(b0 + (b1/t)) или ln(Y) = b0 + (b1/t).

Логистическая. Модель с уравнением Y = 1 / (1/u + (b0 * (b1**t))) или ln(1/Y - 1/u) = ln(b0) + (ln(b1) * t)) , где u есть ограничение сверху. Выбрав Логистическая, задайте границу сверху, которая будет использоваться в регрессионном уравнении. Это значение должно быть положительным числом, превышающим максимальное значение зависимой переменной.

Роста. Модель, задаваемая уравнением Y = e**(b0 + (b1 * t)) или ln(Y) = b0 + (b1 * t).

Экспоненциальная. Модель, задаваемая уравнением Y = b0 * (e**(b1 * t)) или ln(Y) = ln(b0) + (b1 * t).

Подгонка кривых: Сохранить

Сохранить переменные. Для каждой выбранной модели можно сохранить предсказанные значения, остатки (наблюденное значение зависимой переменной минус значение, предсказанные моделью) и интервалы прогноза (верхние и нижние границы). Имена и описательные метки новых переменных показываются в таблице в окне вывода.

Прогноз для наблюдений. Если вы выбрали **Время** , а не переменную из активного набора данных в качестве независимой переменной, вы можете задать период прогноза за концом временного ряда. Вы можете выбрать одну из следующих альтернатив:

- **Прогноз до последнего наблюдения.** Предсказывает значения для всех наблюдений в файле по наблюдениям из периода оценивания. Период оценивания, показанный внизу диалогового окна, задается при помощи диалогового окна Отобрать наблюдения: Диапазон, вызываемого из диалогового окна Отбор наблюдений (меню Данные, Отбор наблюдений). Если период оценивания не задан, для предсказания значений используются все наблюдения.
- **Прогноз до.** Прогнозирует значения до заданной даты, времени или номера наблюдения, на основании наблюдений за период оценивания. Эта альтернатива позволяет прогнозировать значения после последнего наблюдения временного ряда. То, какие поля доступны для задания конца интервала прогнозирования, зависит от того, какие переменные дат существуют в данных. Если переменные дат не заданы, вы можете указать номер последнего наблюдения.

Для создания переменных дат используйте пункт Задать данные в меню Данные.

Глава 19. Регрессия частично наименьших квадратов

Процедура Регрессия частично наименьших квадратов оценивает регрессионные модели частично наименьших квадратов (PLS), также известные как модели "проекции на скрытую структуру". PLS представляет собой метод для предсказания, который является альтернативой обычной регрессии наименьших квадратов (OLS), каноническим корреляциям или построению моделей с помощью структурных уравнений. Он особенно полезен, когда предикторные переменные сильно коррелированы или когда число предикторов превышает число наблюдений.

PLS соединяет свойства метода главных компонент и множественной регрессии. Сначала он выделяет набор скрытых факторов, которые объясняют как можно больше ковариации между независимыми и зависимыми переменными. Затем на шаге регрессии предсказываются значения зависимых переменных с использованием декомпозиции независимых переменных.

Таблицы. Доля объясненной дисперсии (по скрытым факторам), веса скрытых факторов, нагрузки скрытых факторов, важность независимой переменной в проекции (VIP - variable importance in projection), а также оценки параметров регрессии (по зависимым переменным) - всё выводится по умолчанию.

Диаграммы. Важность переменной в проекции (VIP), значения факторов, веса факторов для первых трех скрытых факторов и расстояние до модели - всё выводится с вкладки Параметры.

Данные для регрессии частично наименьших квадратов

Шкала измерений. Зависимые и независимые (предикторные) переменные могут быть количественными, номинальными или порядковыми. Данная процедура предполагает, что каждой переменной назначен подходящий тип измерений, хотя можно временно изменить тип измерений для переменной, щелкнув правой кнопки мыши по переменной в списке исходных переменных и выбрав тип измерений во всплывающем меню. Процедура одинаково трактует категориальные (номинальные и порядковые) переменные.

Кодировка категориальных переменных. Данная процедура на время выполнения процедуры перекодирует категориальные зависимые переменные, используя кодировку один из *c*>. Если переменная имеет *c* категорий, то значения этой переменной хранятся в виде *c* векторов, при этом первой категории приписывается (1,0,...,0), следующей категории - (0,1,0,...,0), ..., и последней категории - (0,0,...,0,1). Категориальные зависимые переменные представляются с использованием фиктивной кодировки; то есть просто опускается индикатор, соответствующий опорной категории.

Частотные веса Значения весов перед использованием округляются до ближайшего целого числа. Наблюдения с пропущенными весами или весами, меньшими 0,5, в анализе не используются.

Пропущенные значения. Пользовательские и системные пропущенные значения трактуются как недопустимые.

Изменение масштаба. Все переменные в модели, включая индикаторные переменные, представляющие категориальные переменные, центрируются и стандартизуются.

Для того чтобы получить регрессию частично наименьших квадратов

Выберите в меню:

Анализ > Регрессия > Частично наименьшие квадраты...

1. Выберите хотя бы одну зависимую переменную.
2. Выберите хотя бы одну независимую переменную.

Дополнительно вы можете:

- Задать опорную категорию для категориальных (номинальных и порядковых) зависимых переменных.
- Задать переменную для использования в качестве однозначного идентификатора для вывода по наблюдениям и сохраняемых наборов данных.
- Задать верхнюю границу для числа выделяемых скрытых факторов.

Необходимые условия работы

Процедура Регрессия частично наименьших квадратов - это команда расширения Python, и для нее требуется IBM SPSS Statistics - Essentials for Python, устанавливаемый по умолчанию с вашим продуктом IBM SPSS Statistics. Для нее требуются также свободно распространяемые библиотеки Python NumPy и SciPy.

Примечание: Для пользователей, работающих в режиме распределенного анализа (где требуется сервер IBM SPSS Statistics), NumPy и SciPy должны быть установлены на сервере. Обратитесь за помощью к администратору системы.

Пользователи Windows и Mac

Для Windows и Mac библиотеки NumPy и SciPy должны быть установлены в версии Python 2.7, отдельной от версии, установленной с IBM SPSS Statistics. Если у вас нет отдельной версии Python 2.7, ее можно загрузить с http://www.python.org. Затем установите NumPy и SciPy для Python версии 2.7. Программы установки доступны на странице http://www.scipy.org/Download.

Чтобы включить использование NumPy и SciPy, надо задать в качестве положения Python положение версии Python 2.7, в которой установлены NumPy и SciPy. Положение Python задается на вкладке Положение файлов диалогового окна Параметры (Правка > Параметры).

Пользователи Linux

Мы предлагаем самостоятельно получить исходные файлы и построить NumPy и SciPy. Эти исходные файлы доступны на http://www.scipy.org/Download. Вы можете установить NumPy и SciPy в версию Python 2.7, установленную с IBM SPSS Statistics. Она находится в подкаталоге `Python` каталога, где установлен IBM SPSS Statistics.

Если вы выбрали установку NumPy и SciPy не в версию Python, установленную с IBM SPSS Statistics, а в другую версию Python 2.7, надо задать положение Python, указывающее на эту версию. Положение Python задается на вкладке Положение файлов диалогового окна Параметры (Правка > Параметры).

Windows и Unix Server

Библиотеки NumPy и SciPy должны быть установлены на сервере в версии Python 2.7, отдельной от версии, установленной с IBM SPSS Statistics. Если на сервере нет отдельной версии Python 2.7, ее можно скачать с http://www.python.org. Библиотеки NumPy и SciPy для Python 2.7 доступны на http://www.scipy.org/Download. Чтобы включить использование NumPy и SciPy, надо задать в качестве положения Python положение версии Python 2.7, в которой установлены NumPy и SciPy. Для задания положения Python используется Консоль администрирования IBM SPSS Statistics.

Модель

Задать эффекты модели. Модель главных эффектов содержит все главные эффекты факторов и ковариат. Выберите **Настраиваемая**, чтобы задать взаимодействия. Необходимо указать все члены, включаемые в модель.

Факторы и ковариаты. Перечисляются факторы и ковариаты.

Модель. Модель зависит от природы ваших данных. Выбрав **Настраиваемая**, вы можете отобрать главные эффекты и взаимодействия, которые представляют интерес для анализа.

Создать члены

Для выбранных факторов и ковариат:

Взаимодействие. Создается член взаимодействия наивысшего порядка всех выбранных переменных. Это вариант по умолчанию.

Главные эффекты. Создаются главные эффекты для всех выбранных переменных.

Все 2-факторные. Создаются все возможные двухфакторные взаимодействия выбранных переменных.

Все 3-факторные. Создаются все возможные трехфакторные взаимодействия выбранных переменных.

Все 4-факторные. Создаются все возможные четырехфакторные взаимодействия выбранных переменных.

Все 5-факторные. Создаются все возможные пятифакторные взаимодействия выбранных переменных.

Параметры

Вкладка Параметры позволяет пользователю сохранить и представить графически модельные оценки для отдельных наблюдений скрытых факторов и предикторов.

Для каждого типа данных задайте имя набора данных. Имена наборов данных должны быть уникальными. Если задать имя существующего набора данных, его содержимое заменяется; в противном случае создается новый набор данных.

- **Сохранить оценки для отдельных наблюдений.** Сохраняются по наблюдениям следующие модельные оценки по наблюдениям: предсказанные значения, остатки, расстояние до модели скрытых факторов, а также значения скрытых факторов. Значения скрытых факторов также представляются графически.
- **Сохранить оценки для скрытых факторов.** Сохраняются нагрузки скрытых факторов и веса скрытых факторов. Веса скрытых факторов также представляются графически.
- **Сохранить оценки для независимых переменных.** Сохраняются оценки параметров регрессии и важность переменной в проекции (VIP). Значения VIP также представляются графически по скрытым факторам.

Глава 20. Метод ближайших соседей

Анализ ближайшего сходства представляет собой метод классификации наблюдений на основе сходства наблюдений. Этот метод машинного обучения был разработан в качестве способа распознавания структуры данных при неточном соответствии имеющих структур или наблюдений. Подобные наблюдения близки друг к другу, а непохожие наблюдения, наоборот, удалены друг от друга. Таким образом, дистанция между двумя наблюдениями является критерием их различия.

Близкие друг к другу наблюдения называются “соседи”. Когда представляется новое наблюдение, обозначенное знаком вопроса, вычисляется его расстояние от всех других наблюдений в модели. Определяется классификация наиболее похожих наблюдений (ближайшее сходство) и новое наблюдение помещается в категорию, в которой содержится наибольшее количество ближайшего сходства.

Пользователь может указать количество анализируемых ближайших соседей; это значение обозначается *k*.

Анализ ближайшего сходства также может использоваться для вычисления значений для непрерывного целевого объекта. В этой ситуации среднее целевое значение ближайшего сходства используется для получения предсказанного значения для нового наблюдения.

Данные для анализа методом ближайшего сходства

Цель и показатели. В качестве цели и показателей могут использоваться следующие переменные:

- *Номинальная*. Переменную можно рассматривать как номинальную, когда ее значения представляют категории без естественного упорядочения, например, подразделение компании, где работает наемный сотрудник. Примеры номинальных переменных включают регион, почтовый индекс или религию.
- *Порядковая*. Переменную можно рассматривать как порядковую, когда ее значения представляют категории с некоторым естественным для них упорядочением, например, уровни удовлетворенности обслуживанием от крайней неудовлетворенности до крайней удовлетворенности. Примеры порядковых переменных включают баллы, представляющие степень удовлетворенности или уверенности, или баллы, оценивающие предпочтение.
- *Шкалы*. Переменную можно рассматривать как количественную (непрерывную), когда ее значения представляют упорядоченные категории с осмысленной метрикой, так что уместно сравнивать расстояния между значениями. Примеры количественной переменной включают возраст в годах и доход в тысячах долларов.

  Процедура анализа методом ближайшего сходства одинаково трактует номинальные и порядковые переменные. Для данной процедуры предполагается, что каждой переменной присвоен подходящий тип шкалы измерений, хотя можно временно изменить тип шкалы измерений для переменной, щелкнув правой кнопкой мыши по переменной в списке исходных переменных и выбрав тип шкалы измерений во всплывающем меню.

Значок, расположенный рядом с каждой переменной в списке переменных, показывает тип шкалы измерений и тип данных:

Таблица 1. Значки уровня измерения

	Числовой	**Строка**	**Дата**	**Время**
Количественная (непрерывная)		(не задается)		
Порядковый				

Таблица 1. Значки уровня измерения (продолжение)

	Числовой	**Строка**	**Дата**	**Время**
Номинальный				

Кодировка категориальных переменных. Процедура на время своего выполнения перекодирует категориальные предикторные и зависимую переменные, используя кодировку один-из- *c*>. Если переменная имеет *c* категорий, то значения этой переменной хранятся как с векторов, при этом первой категории приписывается (1,0,...,0), следующей категории - (0,1,0,...,0), ..., и последней категории - (0,0,...,0,1).

Данная схема кодировки увеличивает размерность пространства показателей. В частности, общее число измерений равно числу количественных предикторов плюс число категорий по всем категориальным предикторам. Как результат, такая схема кодировки может привести к увеличению времени обучения. Если для метода ближайшего сходства обучение работает очень медленно, то можно попытаться уменьшить число категорий категориальных предикторов, прежде чем запустить процедуру, путем объединения похожих категорий или, отбрасывая наблюдения, которые имеют очень редко встречающиеся категории.

Все кодирование вида один-из-*c* основывается на обучающих данных, даже если задана контрольная выборка (смотрите раздел “Разделы” на стр. 92). Таким образом, если контрольная выборка содержит наблюдения с категориями предикторов, которые не присутствуют в обучающих данных, то такие наблюдения не учитываются. Если контрольная выборка содержит наблюдения с категориями зависимой переменной, которые не присутствуют в обучающих данных, то такие наблюдения учитываются.

Изменение масштаба. Количественные показатели нормализуются по умолчанию. Все изменение масштаба выполняется на основе обучающих данных, даже если задана опорная выборка (смотрите раздел “Разделы” на стр. 92). При задании переменной, определяющей группы, важно, чтобы показатели имели похожие распределения по обучающей и контрольной выборкам. Воспользуйтесь, например, процедурой Исследовать, чтобы проверить распределения по группам.

Частотные веса Частотные веса игнорируются данной процедурой.

Воспроизведение результатов. В процессе случайного формирования групп и слоев для перекрестной проверки данная процедура генерирует случайные числа. Если вы хотите точно воспроизвести полученные результаты, в дополнение к тем же установкам для процедуры задайте значение для генератора Твистер Мерсенна (смотрите раздел “Разделы” на стр. 92) или используйте переменные для задания групп и слоев для перекрестной проверки.

Как выполнить анализ методом ближайшего сходства

Выберите в меню:

Анализ > Классификация > Метод ближайшего сходства...

1. Задайте один или несколько показателей, которые при наличии целевой переменной могут рассматриваться как независимые переменные или предикторы.

   Цель (необязательно). Если не задана цель (зависимая переменная или отклик), то процедура находит только *k* ближайшее сходство – классификация и предсказание не выполняются.

   Нормализовать количественные показатели. Нормализованные показатели имеют один и тот же диапазон значений, что может повысить эффективность алгоритма оценивания. Используется скорректированная нормализация: [2*(x–min)/(max–min)]–1. Значения со скорректированной нормализацией лежат между –1 и 1.

   Идентификатор фокусного наблюдения (необязательно). Он позволяет отметить наблюдения, представляющие особый интерес. Например, исследователь хочет проверить, сопоставимы ли баллы оценок для одного школьного округа в США (фокусного наблюдения) с аналогичными для схожих

школьных округов. Он использует анализ методом ближайшего сходства, для того чтобы найти школьные округа, наиболее похожие по заданному набору показателей. Затем он сравнивает баллы оценок для фокусного школьного округа с баллами оценок для ближайшего сходства.

Фокусные наблюдения также можно использовать в клинических исследованиях для выбора контрольных наблюдений, подобных клиническим наблюдениям. Фокусные наблюдения выводятся в таблице *k* ближайших соседей и расстояний, на диаграмме пространства показателей, на диаграмме соседей и на диаграмме квадрантов. Информация о фокусных наблюдениях сохраняется в файлах, заданных на вкладке Вывод.

Наблюдения с положительным значением заданной переменной рассматриваются как фокусные наблюдения. Недопустимо задавать переменную, не имеющую положительных значений.

Метка наблюдения (необязательно). Наблюдения помечаются, используя эти значения, на диаграмме пространства показателей, на диаграмме соседей и на диаграмме квадрантов.

Поля с неизвестным типом измерений

В случае, когда тип измерений для одной или нескольких переменных (полей) в наборе данных неизвестен, выводится предупреждающее сообщение о типе измерений. Так как тип измерений влияет на вычисление результатов для этой процедуры, все переменные должны иметь заданный тип измерений.

Сканировать данные. Считывает данные в активном наборе данных и назначает тип измерений по умолчанию любым полям с неизвестным типом измерений. Это может занять некоторое время, если набор данных большой.

Назначить вручную. Открывает диалоговое окно, в котором перечисляются все поля с неизвестным типом измерений. Можно использовать это диалоговое окно, чтобы назначить тип измерений таким полям. Тип измерений можно также назначит в представлении Переменные Редактора данных.

Поскольку тип измерений важен для этой процедуры, нельзя получить доступ к диалоговому окну, позволяющему запустить эту процедуру, пока для всех полей не будет задан тип измерений.

Соседи

Количество ближайших соседей (k). Задайте число ближайшего сходства. Обратите внимание на то, что использование большего числа соседей необязательно приводит к более точной модели.

Если в представлении Переменные задана целевая переменная, то в качестве альтернативы можно задать диапазон значений и позволить процедуре выбрать наилучшее число соседей в этом диапазоне. Метод определения числа ближайших соседей зависит от того, запрошен ли отбор показателей на вкладке Показатели.

- Если задействован отбор показателей, то он выполняется для каждого значения *k* в заданном диапазоне, и выбирается *k*, а также набор показателей, дающие наименьший процент ошибок (или наименьшую сумму квадратов ошибок, если целевая переменная является количественной).
- Если отбор показателей не задействован, для выбора “наилучшего” числа соседей используется *V*-слойная перекрестная проверка. Для задания слоев перейдите на вкладку Группы.

Вычисление расстояний. Здесь задается метрика расстояния, используемая в качестве меры сходства наблюдений.

- **Метрика Евклида.** Расстояние между двумя наблюдениями x и y представляет собой квадратный корень из суммы квадратов разностей значений наблюдений по всем измерениям.
- **Метрика городского квартала.** Расстояние между двумя наблюдениями представляет собой сумму абсолютных разностей значений наблюдений по всем измерениям. Эта метрика также называется Манхэттенским расстоянием.

Дополнительно, если в представлении Переменные задана целевая переменная, то можно задать взвешивание показателей с помощью их нормализованной важности при вычислении расстояний. Важность показателя вычисляется для предиктора как отношение процента ошибок или ошибки в виде суммы квадратов для модели с удаленным рассматриваемым предиктором к проценту ошибок или ошибке в виде суммы квадратов для полной модели. Нормализованная важность вычисляется путем деления значений важностей показателей на одно и то же число, для того чтобы их сумма равнялась 1.

Предсказанные значения для количественной цели. Если в представлении Переменные задана количественная целевая переменная, то здесь указывается, будет ли предсказанное значение вычислено по значению среднего или медианы ближайшего сходства.

Показатели

Вкладка Показатели позволяет запросить и задать параметры для отбора показателей, когда в представлении Переменные задана целевая переменная. По умолчанию при отборе показателей рассматриваются все показатели, однако можно выделить часть показателей для принудительного включения в модель.

Критерий остановки. На каждом шаге в модель добавляется тот показатель, добавление которого в модель дает наименьшую ошибку (вычисляемую как процент ошибок для категориальной целевой переменной и как сумму квадратов ошибок для количественной целевой переменной). Отбор включением продолжается до тех пор, пока не выполнится заданное условие.

- **Заданное количество показателей.** Алгоритм отбирает фиксированное число показателей в дополнение к тем, которые принудительно включаются в модель. Задайте целое положительное число. Уменьшение числа отбираемых показателей создает более компактную модель, повышая риск упустить важные показатели. Увеличение числа отбираемых показателей приведет к включению всех важных показателей, повышая риск в итоге включить показатели, которые в действительности увеличивают модельную ошибку.
- **Минимум модуля относительного изменения ошибки.** Алгоритм останавливается, когда значение модуля относительного изменения ошибки указывает на то, что модель нельзя дальше улучшить путем добавления дополнительных показателей. Задайте положительное число. При уменьшении значения минимального изменения появляется тенденция включить больше показателей, при этом возникает риск включить показатели, которые не улучшают заметно качество модели. При увеличении значения минимального изменения появляется тенденция включить меньше показателей, при этом возникает риск потерять показатели, которые важны для модели. “Оптимальное” значение минимального изменения зависит от имеющихся данных и решаемой задачи. Смотрите диаграмму значений ошибок при отборе показателей в выводе, чтобы определить, какие показатели наиболее важны. Дополнительную информацию смотрите в разделе “Значения ошибок при отборе показателей” на стр. 97.

Разделы

Вкладка Группы позволяет разделить набор данных на обучающий и контрольный наборы и, когда это возможно, приписать наблюдения слоям для перекрестной проверки.

Обучающая и контрольная группы. Здесь задается метод разбиения активного набора данных на обучающую и контрольную выборки. **Обучающая выборка** содержит записи данных, используемые для обучения модели ближайшего сходства. Чтобы построить модель, необходимо некоторый процент наблюдений из набора данных включить в обучающую выборку. **Контрольная выборка** представляет собой независимый набор записей данных, используемый для проверки качества окончательной модели. Ошибка для контрольной выборки дает корректную оценку прогностической способности модели, поскольку контрольные наблюдения не использовались для построения модели.

- **Распределить наблюдения по группам случайным образом.** Задайте процент наблюдений, приписываемых к обучающей выборке. Остальные наблюдения приписываются к контрольной выборке.
- **Для распределения наблюдений использовать переменную.** Задайте числовую переменную, которая относит каждое наблюдение активного набора данных к обучающей или контрольной выборке. Наблюдения с

положительным значением этой переменной относятся к обучающей выборке, а наблюдения с отрицательным или нулевым значением - к контрольной выборке. Наблюдения с системными пропущенными значениями исключаются из анализа. Любые пользовательские пропущенные значения группирующей переменной всегда рассматриваются как не пропущенные.

Слои для перекрестной проверки. *V*-слойная перекрестная проверка используется для определения наилучшего числа соседей. Она недоступна совместно с отбором показателей по причинам, связанным с эффективностью работы процедуры.

Для выполнения перекрестной проверки выборка делится на некоторое число подвыборок или слоев. Затем формируются модели ближайшего сходства с поочередным исключением данных каждой подвыборки. Первая модель создается на основе всех наблюдений, кроме наблюдений из первого слоя выборки, вторая модель создается на основе всех наблюдений, кроме наблюдений из второго слоя выборки, и так далее. Для каждой модели оценивается ошибка путем применения модели к подвыборке, которая была исключена при ее создании. Наилучшее число ближайших соседей - это то, которое дает наименьшую среднюю ошибку по слоям.

- **Распределить наблюдения по слоям случайным образом.** Задайте число слоев, которое должно использоваться при перекрестной проверке. Процедура случайным образом распределяет наблюдения по слоям, пронумерованным от 1 до *V*, где V - число слоев.
- **Для распределения наблюдений использовать переменную.** Задайте числовую переменную, которая относит каждое наблюдение в активном наборе данных к некоторому слою. Эта переменная должна быть числовой и принимать значения от 1 до *V*. Если пропущены какие-либо значения в этом диапазоне, а также по каким-либо разбиениям, если используются файлы разбиения, это вызовет ошибку.

Задать начальное значение для Твистера Мерсенна. Установка начального значения позволяет воспроизводить результаты анализа. Применение этого элемента управления аналогично выбору Твистера Мерсенна в качестве активного генератора и заданию фиксированной начальной точки в диалоговом окне Генераторы случайных чисел с той существенной разницей, что задание значения в данном диалоговом окне запоминает текущее состояние генератора случайных чисел и восстанавливает это состояние после того, как анализ будет выполнен.

Сохранение

Имена сохраняемых переменных. Автоматическое формирование имен гарантирует, что будут сохранены все результаты вашей работы. Настраиваемые имена позволяют удалять/заменять результаты предыдущих прогонов без необходимости предварительно удалять сохраненные переменные в Редакторе данных.

Переменные для сохранения

- **Предсказанное значение или категория.** Это задает сохранение предсказанного значения для количественной целевой переменной или предсказанной категории для категориальной целевой переменной.
- **Предсказанная вероятность.** Это задает сохранение предсказанных вероятностей для категориальной целевой переменной. Для каждой из первых *n* категорий сохраняется отдельная переменная, где *n* задается с помощью управляющего элемента **Максимальное количество сохраняемых категорий для категориальной цели**.
- **Переменная обучающей/контрольной группы.** Если на вкладке Группы задано случайное распределение наблюдений между обучающей и контрольной выборками, то здесь сохраняется идентификатор группы (обучающей или контрольной), к которой наблюдение было отнесено.
- **Переменная слоя для перекрестной проверки.** Если на вкладке Группы задано случайное распределение наблюдений между слоями для перекрестной проверки, то здесь сохраняется идентификатор слоя, к которому наблюдение было отнесено.

Вывод

Вывод средства просмотра

- **Сводка обработки наблюдений.** Выводится сводная таблица обработки наблюдений, в которой приводятся числа наблюдений, включенных в анализ и исключенных из него, в целом, а также по обучающей и контрольной выборкам.
- **Диаграммы и таблицы.** Показывается вывод, относящийся к модели, включая таблицы и диаграммы. Таблицы, показанные в представлении моделей, включают *k* ближайшего сходства и расстояния для фокусных наблюдений, классификацию для категориальной переменной отклика, а также значения ошибок. Графический вывод, доступный в представлении моделей, включает значения ошибок отбора, диаграмму важности предикторов, диаграмму пространства показателей, диаграмму соседей и диаграмму квадрантов. Дополнительную информацию смотрите в разделе “Представление модели”.

Файлы

- **Экспортировать модель в файл XML.** Этот файл модели можно использовать для применения информации о модели к другим файлам данных с целью скоринга. Такая возможность отсутствует, если заданы файлы разбиения.
- **Экспортировать расстояния между фокусными наблюдениями и k ближайшими соседями.** В новом наборе данных формируются *k* переменных, в которых для каждого фокусного наблюдения содержится номер наблюдения (принадлежащего обучающей выборке), которое является соответствующим ближайшим соседом, а также *k* переменных с расстояниями до ближайших соседей.

Параметры

Пользовательские пропущенные значения. Категориальные переменные должны иметь допустимые значения, для того чтобы наблюдение было включено в анализ. Эти управляющие элементы позволяют решить, считать ли пользовательские пропущенные значения для категориальных переменных допустимыми.

Системные пропущенные значения и пропущенные значения для количественных переменных всегда рассматриваются как недопустимые.

Представление модели

Если на вкладке Вывод выбрано **Диаграммы и таблицы** то в средстве просмотра процедура создает объект Модель ближайшего сходства. Активация (двойным щелчком) этого объекта позволяет рассматривать модель в интерактивном режиме. Представление Модель имеет 2x-панельное окно:

- Первая панель выводит обзорное изображение модели, называемое главным видом.
- Вторая панель выводит изображение одного из двух типов:

  Дополнительное представление модели показывает дополнительную информацию о модели, но не концентрируется на самой модели.

  Связанный вид является видом, демонстрирующим один из элементов модели, когда пользователь углубляется в детали основного вида.

По умолчанию первая панель показывает пространство показателей, а вторая панель показывает диаграмму важности переменных. Если диаграмма важности недоступна, то есть на вкладке Соседи не было выбрано **При расчете расстояний взвешивать показатели значениями важности**, то показывается первый доступный элемент из раскрывающегося меню Вид.

Если изображение недоступно, то текст соответствующего ему элемента в раскрывающемся меню Вид отсутствует.

Пространство показателей

Диаграмма пространства показателей является интерактивной диаграммой пространства показателей (или подпространства, если имеется более 3 показателей). Каждая ось представляет показатель в модели, а расположение точек на диаграмме показывает значения этих показателей для наблюдений в обучающей и контрольной группах.

Ключи. Помимо значений показателей, точки на диаграмме содержат другую информацию.

- Форма показывает, к какой группе принадлежит точка: к обучающей или к контрольной.
- Цвет/оттенок точки показывает значение целевой переменной для данного наблюдения. Различающимися цветами обозначается принадлежность к различным категориям категориальной целевой переменной. Различными оттенками обозначаются различные диапазоны значений непрерывной целевой переменной. Показанное значение для обучающей группы является наблюденным значением; для контрольной группы это предсказанное значение. Если целевая переменная не задана, этот ключ не используется.
- Более жирный контур указывает на то, что наблюдение является фокусным. Фокусные наблюдения показываются соединенными с их *k* ближайшими соседями.

Элементы управления и интерактивность. С помощью ряда управляющих элементов, которые представлены на диаграмме, можно исследовать пространство показателей.

- Можно выбрать показатели, которые будут показаны на диаграмме, а также изменить соответствие между осями и показателями.
- “Фокусные наблюдения” - это всего лишь точки, выбранные на диаграмме пространства функций. Если задана переменная идентификации фокусных наблюдений, то точки, представляющие фокусные наблюдения, изначально будут выделены. Однако любая точка может временно стать фокусным наблюдением, если ее выделить. Применяются “обычный” способ выделения: щелчок по точке выделяет эту точку и снимает выделение всех остальных; щелчок по точке с нажатой клавишей Ctrl добавляет ее к набору выделенных точек. Связанные виды, такие, как Диаграмма сходства, автоматически обновятся в соответствии с выбором наблюдений в пространстве показателей.
- Можно изменить число ближайших соседей (*k*), выводимых для фокусных наблюдений.
- Наведение указателя мыши на точку вызовет вывод строки-подсказки со значением метки наблюдения или номера, если метки наблюдений не заданы, а также наблюденного и предсказанного значений целевой переменной.
- Кнопка “Сброс” позволяет вернуть пространство показателей в исходное состояние.

Добавление и удаление полей/переменных

К пространству показателей можно добавлять новые поля/переменные или удалять те, которые выведены.

Палитра переменных

Для того чтобы иметь возможность добавлять и удалять переменные, сначала необходимо вывести палитру переменных. Для того чтобы иметь возможность вывести палитру переменных, средство просмотра моделей должно находиться в режиме редактирования, и на диаграмме пространства показателей должно быть выбрано наблюдение.

1. Для того чтобы перевести средство просмотра моделей в режим редактирования, выберите в меню:

   Вид > Режим редактирования
2. Находясь в режиме редактирования, щелкните по любому наблюдению на диаграмме пространства показателей.
3. Для того чтобы вывести палитру переменных, выберите в меню:

   Вид > Палитры > Переменные

   Палитра переменных перечисляет все переменные в пространстве показателей. Значок рядом с именем переменной указывает шкалу измерений переменной.

4. Для того чтобы временно изменить шкалу измерений переменной, щелкните правой кнопкой мыши по переменной в палитре переменных и выберите вариант.

Зоны переменных

Переменные помещаются в зоны на диаграмме пространства показателей. Для того чтобы вывести зоны, начните перетаскивать переменную из палитры переменных или поставьте переключатель **Показать зоны**.

Данная диаграмма пространства показателей имеет зоны для осей *x*, *y* и *z*.

Перемещение переменных в зоны

Вот некоторые общие правила и подсказки, касающиеся перемещения переменных в зоны:

- Для того чтобы поместить переменную в зону, перетащите переменную из палитры переменных в эту зону. Если стоит переключатель **Показать зоны**, то можно также щелкнуть по зоне правой кнопкой мыши и в контекстном меню выбрать переменную, которую нужно поместить в зону.
- Если переменная из палитры переменных перетаскивается в зону, уже занятую другой переменной, то старая переменная заменяется новой.
- Если переменная из одной зоны перетаскивается в зону, уже занятую другой переменной, то переменные меняются местами.
- Щелчок по X в зоне удаляет переменную из этой зоны.
- Если визуально показано несколько графических элементов, то каждый графический элемент может иметь свои собственные зоны переменных. Сначала выберите графический элемент.

Важность переменных

Как правило, исследователь хочет сконцентрировать внимание на переменных, которые наиболее важны при построении модели, и отбросить малосущественные переменные. Диаграмма важности переменных помогает это сделать, показывая относительную важность каждой переменной для модели при ее оценивании. Поскольку эти значения являются относительными, в выводе их сумма по всем переменным полагается равной 1,0. Важность переменных не связана с точностью модели. Она означает важность каждой переменной для предсказания, безотносительно к тому, является ли предсказание точным или нет.

Соседи

Эта диаграмма показывает фокусные наблюдения и их *k* ближайших соседей по каждому показателю, а также целевой переменной. Она доступна, если на диаграмме пространства показателей выбирается фокусное наблюдение.

Связывающее поведение. Диаграмма соседей связана с пространством показателей двумя способами.

- Выбранные на диаграмме пространства показателей (фокусные) наблюдения выводятся вместе с их *k* ближайшими соседями на диаграмме соседей.
- Значение *k*, выбранное на диаграмме пространства показателей, используется на диаграмме соседей.

Расстояния до ближайших соседей

Эта таблица выводит *k* ближайших соседей и расстояния до них только для фокусных наблюдений. Она доступна, если на вкладке Переменные задана переменная идентификации фокусных наблюдений и выводит только фокусные наблюдения, идентифицированные этой переменной.

Каждая строка

- столбца **Фокусное наблюдение** содержит значение переменной меток для фокусного наблюдения. Если метки наблюдений не заданы, то этот столбец содержит номер фокусного наблюдения.

- *i*-того столбца в группе Ближайшие соседи содержит значение переменной меток для *i*-того ближайшего соседа фокусного наблюдения. Если метки наблюдений не заданы, то этот столбец содержит номер *i*-того ближайшего соседа фокусного наблюдения.
- *i*-того столбца в группе Наименьшие расстояния содержит расстояние от *i*-того ближайшего соседа до фокусного наблюдения.

Диаграмма квадрантов

Эта диаграмма выводит фокусные наблюдения и их *k* ближайших соседей на диаграмме рассеяния (или на точечной диаграмме, в зависимости от шкалы измерений целевой переменной) с целевой переменной по оси *y* и количественным показателем по оси *x*. Диаграмма разбита на панели по показателям. Она доступна, если задана целевая переменная и на диаграмме пространства показателей выбирается фокусное наблюдение.

- Для непрерывных переменных проводятся опорные линии через средние значения переменных для обучающей группы.

Значения ошибок при отборе показателей

Каждая точка на этой диаграмме по оси *y* показывает ошибку (либо долю ошибок, либо ошибку в виде суммы квадратов, в зависимости от шкалы измерений целевой переменной) для модели с показателем, указанным на оси *x* (и всеми показателями, указанными левее по оси *x*). Эта диаграмма доступна, если заданы целевая переменная и отбор показателей.

Значения ошибок при выборе k

Каждая точка на этой диаграмме по оси *y* показывает ошибку (либо долю ошибок, либо ошибку в виде суммы квадратов, в зависимости от шкалы измерений целевой переменной) для модели с числом ближайших соседей (*k*), указанным на оси *x*. Эта диаграмма доступна, если заданы целевая переменная и выбор *k*.

Значения ошибок при отборе показателей и выборе k

Эта диаграмма представляет собой диаграмму значений ошибок при отборе показателей (смотрите раздел “Значения ошибок при отборе показателей”), разбитую на панели по *k*. Эта диаграмма доступна, если заданы целевая переменная, а также отбор показателей и выбор *k*.

Таблица классификации

В этой таблице выводится перекрестная классификация наблюденных и предсказанных значений целевой переменной по группам. Она доступна, если задана категориальная целевая переменная.

- Строка **Пропущенные** в контрольной группе содержит число наблюдений из этой группы с пропущенными значениями целевой переменной. Для опорной выборки эти наблюдения дают вклад в общий процент, но не в процент правильно классифицированных наблюдений.

Сводка ошибок

Эта таблица доступна, если задана целевая переменная. В ней выводится ошибка модели: сумма квадратов для непрерывной целевой переменной и процент ошибок ((100% – общий процент правильно классифицированных наблюдений) для категориальной целевой переменной.

Глава 21. Дискриминантный анализ

При дискриминантном анализе происходит создание прогностической модели для принадлежности к группе. Данная модель строит дискриминантную функцию (или, когда групп больше двух, набор дискриминантных функций) в виде линейной комбинации предикторных переменных, обеспечивающую наилучшее разделение групп. Эти функции строятся по набору наблюдений, для которых их принадлежность к группам известна, и могут в дальнейшем применяться к новым наблюдениям с известными значениями предикторных переменных, но неизвестной групповой принадлежностью.

Примечание: У группирующей переменной не может быть больше двух значений. Коды для группирующей переменной должны быть целыми, однако вам необходимо задать их максимальное и минимальное значения. Наблюдения со значениями вне этих границ исключаются из анализа.

Пример. Люди в странах с умеренным климатом ежедневно потребляют в среднем больше калорий, чем живущие в тропиках, а большая часть населения в странах с умеренным климатом живет в городах. Исследователь желает построить на основе данной информации функцию для определения того, насколько хорошо можно разделить индивидуумов по этим двум группам стран (на основе данной информации). Исследователь считает, что также важными факторами могут явиться количество населения в стране и ее экономические показатели. Дискриминантный анализ позволяет оценить коэффициенты линейной дискриминантной функции, напоминающей правую часть уравнения множественной линейной регрессии. Если обозначить коэффициенты дискриминантной функции как *a* , *b* , *c* и *d* , то ее можно записать в следующем виде:

```
D = a * климат + b * горожанин ли + c* население + d * валовой внутренний продукт на душу населения
```

Если данные переменные являются существенными для разделения двух климатических зон, значения *D* будут различными для стран с умеренным и тропическим климатом. При использовании метода пошагового отбора переменных может оказаться, что нет необходимости включать в функцию все четыре переменные.

Статистика. Для каждой переменной: средние значения, стандартные отклонения, однофакторный дисперсионный анализ. Для каждого анализа: *M* - статистика Бокса, внутригрупповая корреляционная матрица, внутригрупповая ковариационная матрица, ковариационные матрицы для отдельных групп, общая ковариационная матрица. Для каждой канонической дискриминантной функции: собственное значение, процент дисперсии, каноническая корреляция, лямбда Уилкса, хи-квадрат. Для каждого шага: априорные вероятности, коэффициенты функции Фишера, нестандартизованные коэффициенты функции, лямбда Уилкса для каждой канонической функции.

Данные для дискриминантного анализа

Данные. Группирующая переменная должна иметь ограниченное число различных категорий, кодированных целыми числами. Независимые переменные, являющиеся номинальными, должны быть перекодированы в фиктивные переменные или переменные контрастов.

Допущения. Наблюдения должны быть независимыми. Предикторные переменные должны подчиняться многомерному нормальному распределению, а внутригрупповые ковариационные матрицы должны совпадать для всех групп. Групповая принадлежность предполагается взаимоисключающей (т.е. ни одно наблюдение не принадлежит более чем одной группе) и совместно исчерпывающей (т.е. каждое наблюдение принадлежит какой-либо группе). Процедура наиболее эффективна в ситуации, когда группирующая переменная является истинно категориальной; если принадлежность к группе определяется значениями непрерывной переменной (например, высокий IQ (коэффициент интеллекта) низкий IQ), то имеет смысл обратиться к линейной регрессии, чтобы воспользоваться преимуществом большей информативности непрерывной переменной.

Для выполнения дискриминантного анализа

1. Выберите в меню:

   Анализ > Классификация > Дискриминант...
2. Выберите целочисленную группирующую переменную и нажмите кнопку **Задать диапазон** , чтобы задать нужные категории.
3. Выберите независимые или предикторные переменные. (Если у группирующей переменной нет целых значений, то переменная с целыми значениями может быть создана с помощью пункта Автоматическая перекодировка меню Преобразовать.)
4. Выберите метод ввода независимых переменных.
   - **Вводить независимые вместе.** Одновременно вводятся все независимые переменные, удовлетворяющие критериям допуска (толерантности).
   - **Шаговый отбор.** Для включения и исключения переменных используется шаговый метод.
5. При желании вы можете осуществить отбор наблюдений при помощи переменной отбора.

Задание диапазона в процедуре Дискриминантный анализ

Укажите минимальное и максимальное значения группирующей переменной. Наблюдения со значениями вне заданного диапазона не будут использованы в дискриминантном анализе, но будут отнесены в одну из имеющихся групп на основании результатов анализа. Минимальное и максимальное значения должны быть целочисленными.

Отбор наблюдений для процедуры Дискриминантный анализ

Как отобрать наблюдения для анализа

1. В диалоговом окне Дискриминантный анализ выберите переменную отбора.
2. Щелкните по **Значение** , чтобы ввести целое число в качестве значения отбора.

При построении дискриминантных функций используются только наблюдения с заданным значением переменной отбора. Статистики и результаты классификации выводятся как для отобранных, так и не отобранных наблюдений. Это предоставляет механизм для классификации новых наблюдений на основе ранее существовавших данных или для разделения ваших данных на обучающее и контрольное подмножества, чтобы выполнить проверку адекватности построенной модели.

Статистики в процедуре Дискриминантный анализ

Описательные статистики. Доступны параметры: средние значения (включая стандартные отклонения), одномерный дисперсионный анализ, а также *M* -критерий Бокса.

- *Средние*. Выводятся общее и групповые средние, а также стандартные отклонения для независимых переменных.
- *Однофакторный дисперсионный анализ*. Проводит однофакторный дисперсионный анализ для проверки гипотезы о равенстве групповых средних для каждой независимой переменной.
- *M Бокса*. Критерий равенства групповых ковариационных матриц. Если p не значимо, а выборка достаточно велика, то нет достаточных свидетельств того, что матрицы различаются. Этот критерий чувствителен к отклонениям от многомерной нормальности.

Коэффициенты функции. Возможен вывод классификационных коэффициентов Фишера и нестандартизованных коэффициентов.

- *Фишера*. Коэффициенты классифицирующей функции Фишера, которые можно напрямую использовать для классификации. Для каждой группы создается отдельный набор коэффициентов, при этом наблюдение относится к группе, которой соответствует наибольшее значение дискриминантной функции (значение классифицирующей функции).
- *Нестандартизованные*. Выводит нестандартизованные коэффициенты дискриминантной функции.

Матрицы. Доступными матрицами коэффициентов для независимых переменных являются: внутригрупповая корреляционная матрица, внутригрупповая ковариационная матрица, ковариационные матрицы для отдельных групп и общая ковариационная матрица.

- *Внутригрупповая корреляция*. Выводится объединенная внутригрупповая корреляционная матрица, полученная путем усреднения ковариационных матриц отдельных групп перед вычислением корреляций.
- *Внутригрупповая ковариация*. Выводится объединенная внутригрупповая ковариационная матрица, которая может отличаться от общей ковариационной матрицы. Матрица вычисляется путем усреднения отдельных ковариационных матриц для всех групп.
- *Групповые ковариации*. Для каждой группы выводится отдельная ковариационная матрица.
- *Общая ковариация*. Выводится ковариационная матрица для всех наблюдений, как если бы они были из одной выборки.

Метод пошагового отбора процедуры Дискриминантный анализ

Метод. Выберите статистику, которая будет использоваться для введения или удаления новых переменных. Возможными альтернативами являются лямбда Уилкса, необъясненная дисперсия, расстояние Махаланобиса, наименьшее *F* отношение и *V* Рао. Выбрав *V* Рао, можно задать минимальное приращение *V* , необходимое для включения переменной.

- *Лямбда Уилкса*. Метод отбора переменных в шаговом дискриминантном анализе, отбирающий переменные для ввода в уравнение на основании того, насколько они уменьшают значение "лямбда" Уилкса. На каждом шаге вводится переменная, минимизирующая это значение.
- *Необъясненная дисперсия*. На каждом шаге вводится переменная, минимизирующая сумму необъясненной изменчивости между группами.
- *Расстояние Махаланобиса*. Мера того, насколько значения наблюдений для независимых переменных отклоняются от среднего по всем наблюдениям. Большое расстояние Махаланобиса означает, что наблюдение содержит экстремальные значения в одной или более независимых переменных.
- *Наименьшее F отношение*. Метод отбора переменных в шаговом анализе, основанный на максимизации F-отношения, вычисленного по расстоянию Махаланобиса между группами.
- *V Рао*. Мера различий между групповыми средними. Также называется следом Лоули-Хотеллинга. На каждом шаге вводится та переменная, которая максимизирует прирост индекса V Рао. Выбрав этот параметр, введите минимальное значение, которое должна иметь переменная, чтобы быть включенной в анализ.

Критерии. Возможные альтернативы: **Использовать F-значение** и **Использовать вероятность F**. Введите значения для включения и удаления переменных.

- *Использовать F-значение*. Переменная вводится в модель, если ее F-значение превышает заданное значение включения, и исключается, если ее F-значение меньше значения исключения. Значение включения должно превосходить значение удаления, оба должны быть положительными. Если необходимо ввести в модель больше переменных, снизьте порог включения. Чтобы исключить из модели большее число переменных, увеличьте порог исключения.
- *Использовать вероятность F*. Переменная вводится в модель, если наблюдаемый уровень значимости ее F-значения меньше заданного порога включения, и исключается, если этот уровень значимости больше порога исключения. Порог включения должен быть меньше порога удаления, они оба должны быть положительными. Если необходимо включить в модель больше переменных, увеличьте порог включения. Чтобы исключить из модели большее число переменных, снизьте порог исключения.

Выводить. Отчет о шагах выводит статистики для всех переменных после каждого шага; **F для попарных расстояний** выводит матрицу попарных *F*-отношений для каждой пары групп.

Дискриминантный анализ: классификация

Априорные вероятности. Эта функция определяет настройку классификационных коэффициентов в соответствии с априорным знанием принадлежности к группе.

- **Все группы равны.** Предполагаются равные вероятности для всех групп, что не оказывает влияния на коэффициенты.
- **Вычислить по размерам групп.** Априорные вероятности принадлежности к группе зависят от размера наблюдаемой группы в выборке. Например, если 50% наблюдений из области анализа попадает в первую группу, 25% во вторую и 25% в третью, классификационные коэффициенты настраиваются для увеличения правдоподобия принадлежности к первой группе по отношению ко второй и третьей.

Вывод. Доступные параметры: результаты по наблюдениям (Поточечные результаты), итоговая таблица, классификация методом скользящего контроля.

- *Поточечные результаты*. Коды для фактической группы, предсказанной группы, апостериорные вероятности и значения дискриминантной функции выводятся для каждого наблюдения.
- *Итоговая таблица*. Числа наблюдений, правильно и неправильно отнесенных к каждой из групп в дискриминантном анализе. Это иногда называют матрицей перекрестной классификации.
- *Классификация с удалением по одной точке*. Каждое наблюдение при анализе классифицируется с помощью функции, полученной по всем остальным наблюдениям, кроме данного. Используется также название "U-метод".

Заменить пропущенные значения средним. Выберите этот пункт, чтобы заменить средним независимой переменной пропущенные значения только на этапе классификации.

Ковариационная матрица. Вы можете выбрать один из двух способов классификации наблюдений - либо по внутригрупповой ковариационной матрице, либо по ковариационным матрицам для отдельных групп.

- *Внутри групп*. Для классификации наблюдений используется объединенная внутригрупповая ковариационная матрица.
- *Для отдельных групп*. Для классификации используются ковариационные матрицы для отдельных групп. Так как классификация производится на основе дискриминантных функций, а не на основе исходных переменных, выбор этого параметра не всегда равноценен квадратичной дискриминации.

Графики. Графические возможности: график для объединенных групп, графики для отдельных групп и территориальная карта.

- *Объединенные группы*. Строится диаграмма рассеяния значений первых двух дискриминантных функций для наблюдений из всех групп. Если есть только одна дискриминантная функция, вместо диаграммы рассеяния выводится гистограмма.
- *Для отдельных групп*. Диаграмма рассеяния значений первых двух дискриминантных функций строится для каждой группы в отдельности. Если есть только одна дискриминантная функция, вместо диаграммы рассеяния выводится гистограмма.
- *Территориальная карта*. График, на который нанесены границы, позволяющие отнести наблюдение к группе на основании значений функции. Числа соответствуют группам, по которым распределяют наблюдения. Среднее каждой группы обозначено звездочкой внутри границ этой группы. Если есть только одна дискриминантная функция, диаграмма не выводится.

Дискриминантный анализ: Сохранить

Вы можете добавить к активному файлу данных новые переменные. Можно сохранить: предсказанную принадлежность к группе (единственная переменная), дискриминантные оценки (одна переменная для каждой дискриминантной функции в решении), вероятности принадлежности к группе при данных дискриминантных баллах (одна переменная на каждую группу).

Вы можете также экспортировать информацию о модели в заданный файл в формате XML. Этот файл модели можно использовать для применения информации о модели к другим файлам данных с целью скоринга.

Команда DISCRIMINANT: дополнительные возможности

Язык синтаксиса команд также позволяет:

- Выполнить дискриминантный анализ несколько раз (с помощью одной команды), а также управлять порядком, в котором добавляются переменные (с помощью подкоманды ANALYSIS).
- Задать априорные вероятности для классификации (с помощью подкоманды PRIORS).
- Вывести повернутые матрицу коэффициентов дискриминантных функций и структурную матрицу (с помощью подкоманды ROTATE).
- Ограничить число формируемых дискриминантных функций (с помощью подкоманды FUNCTIONS).
- Ограничить классификацию наблюдениями, которые отобраны (не отобраны) для анализа (с помощью подкоманды SELECT).
- Считать и анализировать корреляционную матрицу (с помощью подкоманды MATRIX).
- Сохранить корреляционную матрицу для дальнейшего анализа (с помощью подкоманды MATRIX).

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Глава 22. Факторный анализ

Целью факторного анализа является выявление скрытых переменных или **факторов** , объясняющих структуру корреляций внутри набора наблюденных переменных. Факторный анализ часто используется для снижения размерности данных, чтобы найти небольшое число факторов, которые объясняют большую часть дисперсии, наблюденной для значительно большего числа явных переменных. Факторный анализ может также использоваться для формирования гипотез относительно механизмов причинных связей или с целью проверки переменных перед дальнейшим анализом (например, чтобы выявить коллинеарность перед проведением линейного регрессионного анализа).

Рассматриваемая процедура факторного анализа обеспечивает большую гибкость:

- Доступны семь методов выделения факторов.
- Доступны пять методов вращения, в том числе прямой облимин и промакс для не ортогональных вращений.
- Доступны три метода вычисления значений факторов, которые можно сохранить в виде переменных для дальнейшего анализа.

Пример. Какие внутренние побуждения определяют ответы людей на вопросы обследования, касающегося политики? Исследование корреляций между вопросами обследования обнаруживает значительные пересечения в подгруппах вопросов -- вопросы о налогах имеют тенденцию коррелировать между собой, вопросы касающиеся обороны также коррелируют между собой и т.д. С помощью факторного анализа можно выявить некоторое число основополагающих факторов и определить, что эти факторы представляют собой концептуально. Помимо этого, для каждого респондента можно вычислить значения факторов, которые можно использовать в последующем анализе. Например, основываясь на значениях факторов, вы можете построить модель логистической регрессии для прогнозирования поведения людей на выборах.

Статистика. Для каждой переменной: число наблюдений без пропущенных значений, среднее значение и стандартное отклонение. Для каждого случая применения факторного анализа: корреляционная матрица переменных, включая уровни значимости, определитель и обратную матрицу; воспроизведенная корреляционная матрица, включая антиобраз; начальное решение (общности, собственные числа и процент объясненной дисперсии); показатель выборочной адекватности Кайзера-Мейера-Олкина и критерий сферичности Бартлетта; неповернутое решение, включая факторные нагрузки, общности и собственные числа; повернутое решение, включая матрицу факторного отображения после вращения и матрицу преобразования факторов. Для косоугольных вращений: матрицы факторного отображения и факторной структуры после вращения; матрица коэффициентов значений факторов и матрица ковариаций факторов. Графики: график типа "осыпь" собственных значений, диаграмма нагрузок первых двух или трех факторов.

Данные для факторного анализа

Данные. Переменные должны быть количественными, измеренными в *интервальной* шкале или шкале *отношений* . Категориальные данные (такие как исповедуемая религия или место рождения) не подходят для факторного анализа. Данные, для которых вычисление коэффициента корреляции Пирсона представляется осмысленным, пригодны также и для факторного анализа.

Допущения. Для каждой пары переменных данные должны представлять собой выборку из двумерного нормального распределения, а наблюдения должны быть независимыми. Модель факторного анализа предполагает, что переменные определяются общими факторами (факторами, оцененными моделью) и характерными или специфическими факторами (не перекрывающимися между наблюденными переменными); вычисляемые оценки основаны на том, что все характерные факторы не коррелированны друг с другом и с общими факторами.

Как запустить процедуру Факторный анализ

1. Выберите в меню:

   Анализ > Снижение размерности > Фактор...
2. Выберите переменные для факторного анализа.

Отбор наблюдений для факторного анализа

Как отобрать наблюдения для анализа

1. Задайте переменную отбора.
2. Щелкните по **Значение** , чтобы ввести целое число в качестве значения отбора.

Только наблюдения с этим значением переменной отбора будут использованы в факторном анализе.

Описательные статистики факторного анализа

Статистики. Одномерные описательные статистики включают среднее значение, среднеквадратичное отклонение и количество наблюдений без пропущенных значений для каждой переменной. **Начальное решение** выводит начальные общности, собственные значения и доли объясненной дисперсии, выраженные в процентах.

Корреляционная матрица. Возможности для вывода: коэффициенты, уровни значимости, детерминант, КМО и критерий сферичности Бартлетта, обратная, воспроизведенная и антиобраз.

- *КМО и критерий сферичности Бартлетта*. Мера выборочной адекватности Кайзера-Мейера-Олкина (КМО), используемая для проверки гипотезы о том, что частные корреляции между переменными малы. Критерий сферичности Бартлетта проверяет гипотезу о том, что корреляционная матрица является единичной матрицей. Если гипотеза верна, факторная модель непригодна.
- *Воспроизведенная*. Корреляционная матрица, оцененная по факторному решению. Выводятся также остатки (разность между оцененными и наблюденными корреляциями).
- *Антиобраз*. Корреляционная матрица антиобразов содержит коэффициенты частных корреляций с обратными знаками, а ковариационная матрица антиобразов содержит частные ковариации с обратными знаками. В хорошей факторной модели большинство внедиагональных элементов будут малы. Мера выборочной адекватности некоторого фактора лежит на диагонали матрицы корреляций антиобразов.

Выделение факторов в процедуре Факторный анализ

Метод. Позволяет задать метод извлечения факторов. Доступные методы: главные компоненты, невзвешенный МНК, обобщенный МНК, максимальное правдоподобие, факторизация главной оси, альфа факторизация и анализ образов.

- *Анализ главных компонентов*. Метод выделения факторов, используемый для формирования некоррелированных линейных комбинаций наблюденных переменных. Первый компонент имеет максимальную дисперсию. Последовательно получаемые компоненты объясняют все меньшие доли дисперсии, и все они не коррелированы между собой. Анализ методом главных компонент применяется для получения начального факторного решения. Может использоваться для сингулярных (вырожденных) корреляционных матриц.
- *Метод невзвешенных наименьших квадратов*. Метод выделения факторов, минимизирующий сумму квадратов разностей между наблюдаемой и воспроизведенной корреляционной матрицами без учета диагоналей.
- *Обобщенный метод наименьших квадратов*. Метод выделения факторов, минимизирующий сумму квадратов разностей между наблюденной и воспроизведенной корреляционными матрицами. Корреляции взвешиваются величинами, обратными характерностям, так что переменные с высокой характерностью получают меньшие веса, чем переменные с низкой.
- *Метод максимального правдоподобия*. Метод выделения факторов. В качестве оценок параметров выбираются те, для которых наблюденная корреляционная матрица наиболее правдоподобна, если

выборка взята из многомерного нормального распределения. Корреляции взвешиваются значениями, обратными к характерностям переменных, и применяется итеративный алгоритм.

- *Факторизация главных осей.* Метод выделения факторов из исходной корреляционной матрицы с квадратами коэффициентов множественных корреляций по диагонали в качестве начальных оценок общностей. Эти факторные нагрузки используют для оценки новых общностей, замещающих старые оценки общностей на диагонали. Итерации будут продолжаться до тех пор, пока изменения общностей от одной итерации к другой не удовлетворят критерию сходимости.
- *Альфа.* Метод выделения факторов, рассматривающий анализируемые переменные как выборку из пространства всех возможных переменных. Он максимизирует альфа пригодность факторов.
- *Анализ образов.* Метод выделения факторов, разработанный Гуттманом и основанный на теории образов. Общая часть переменной, частный образ, определяется как ее линейная регрессия на остальные переменные, а не как функция гипотетических факторов.

Анализ. Позволяет задать для анализа либо корреляционную матрицу, либо ковариационную матрицу.

- **Матрица корреляций** Этот выбор оправдан, если анализируемые переменные измерены в разном масштабе.
- **Матрица ковариаций.** Это полезно, когда необходимо применить факторный анализ к большому числу групп с различными дисперсиями для каждой переменной.

Выделить. Возможно сохранение либо всех тех факторов, собственные числа для которых превосходят заданное значение, либо сохранение заданного количества факторов.

Вывод. Позволяет запросить вывод неповернутого факторного решения, а также график типа "осыпь" для собственных значений.

- *Неповернутое факторное решение.* Выводятся факторные нагрузки (матрица факторного отображения), общности и собственные значения факторного решения без вращения.
- *График собственных значений.* График, на котором изображены дисперсии, связанные с каждым фактором. Используется для определения того, сколько факторов следует сохранить. Обычно график показывает явный разрыв между крутым наклоном больших факторов и постепенным уменьшением остальных ("осыпь").

Максимум итераций до сходимости. Позволяет задать максимальное число шагов, которое может использовать алгоритм для получения решения.

Вращение факторов для факторного анализа

Метод. Позволяет выбрать метод вращения факторов. Доступные методы: варимакс, прямой облимин, квартимакс, эквимакс и промакс.

- *Метод варимакс.* Ортогональный метод вращения, минимизирующий число переменных с высокими нагрузками на каждый фактор. Этот метод упрощает интерпретацию факторов.
- *Метод Прямой облимин.* Метод косоугольного (неортогонального) вращения. Самое косоугольное решение соответствует дельте, равной 0 (по умолчанию). По мере того, как дельта отклоняется в отрицательную сторону, факторы становятся более ортогональными. Чтобы изменить задаваемое по умолчанию дельта (равное 0), введите число, меньшее или равное 0,8.
- *Метод квартимакс.* Метод вращения, который минимизирует число факторов, необходимых для объяснения каждой переменной. Этот метод упрощает интерпретацию наблюденных переменных.
- *Метод эквимакс.* Метод вращения, объединяющий методы варимакс, упрощающий факторы, и квартимакс, упрощающий переменные. Минимизируется число переменных с большими факторными нагрузками и число факторов, требуемых для объяснения переменной.
- *Вращение типа промакс.* Косоугольное вращение в предположении, что факторы могут коррелировать между собой. Оно производится быстрее, чем вращение типа прямой облимин, поэтому оно полезно для больших наборов данных.

Вывод. Позволяет запросить вывод повернутого решения, а также графиков нагрузок для первых двух или трех факторов.

- *Повернутое решение*. Чтобы получить повернутое решение, необходимо выбрать метод вращения. Для ортогонального вращения выдаются матрица факторных нагрузок после вращения и матрица преобразования факторов. Для косоугольного вращения выводятся следующие матрицы: факторных нагрузок после вращения, структурная и корреляций факторов.
- *График факторных нагрузок*. Трехмерный график факторных нагрузок для трех первых факторов. Для двухфакторного решения выдается двумерный график. Если выделен только один фактор, график не выдается. Если задано вращение, график выдается для повернутого решения.

Максимум итераций до сходимости. Позволяет задать максимальное число шагов, которое может использовать алгоритм для выполнения вращения.

Значения факторов в процедуре факторного анализа

Сохранить как переменные. Создает по одной новой переменной для каждого фактора в окончательном решении.

Метод. Альтернативные методы вычисления факторных значений - Бартлетта и Андерсона-Рубина.

- *Регрессионный метод*. Метод оценивания коэффициентов факторных значений. Получающиеся оценки факторных значений имеют среднее, равное нулю, и дисперсию, равную квадрату множественного коэффициента корреляции между оцененными значениями фактора и истинными. Эти факторные значения могут быть коррелированы, даже если факторы ортогональны.
- *Значения Бартлетта*. Метод оценивания коэффициентов факторных значений. Получаемые значения имеют среднее, равное 0. Минимизируется сумма квадратов характерных факторов по всем переменным.
- *Метод Андерсона-Рубина*. Метод оценивания коэффициентов факторных значений; модификация метода Бартлетта, гарантирующая ортогональность оцененных факторов. Получаемые значения некоррелированы, имеют среднее 0 и стандартное отклонение 1.

Вывести матрицу коэффициентов значений факторов. Выводит коэффициенты, на которые умножаются переменные для получения значений факторов. Выводятся также корреляции между факторными значениями.

Параметры процедуры Факторный анализ

Пропущенные значения. Позволяет задать режим обработки пропущенных значений. Возможными альтернативами для наблюдений с пропущенными значениями являются исключение *целиком* , исключение *попарно* или замена пропущенного значения средним.

Формат вывода коэффициентов. Позволяет задать режим вывода матриц. Вы можете отсортировать коэффициенты по величине и не выводить коэффициенты, которые по модулю меньше заданного значения.

Команда FACTOR: дополнительные возможности

Язык синтаксиса команд также позволяет:

- Задать критерии сходимости итераций для выделения факторов и вращения.
- Задать отдельные графики вращения факторов.
- Задать, сколько значений факторов нужно сохранять.
- Задать диагональные значения для метода факторизации главной оси.
- Сохранить на диске корреляционные матрицы и матрицы факторных нагрузок для дальнейшего анализа.
- Считать и анализировать корреляционные матрицы и матрицы факторных нагрузок.

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Глава 23. Выбор процедуры кластеризации

Кластерный анализ можно выполнить, используя процедуры двухэтапного, иерархического кластерного анализа или метода k-средних. Каждая процедура использует разные алгоритмы для формирования кластеров, и каждая имеет параметры, недоступные для других.

Двухэтапный кластерный анализ. Для многих приложений процедура Двухэтапный кластерный анализ окажется подходящим выбором. Она дает следующие уникальные возможности:

- Автоматический выбор наилучшего числа кластеров и мер для выбора кластерной модели.
- Кластерные модели можно строить одновременно на основе и категориальных, и непрерывных переменных.
- Сохранение модели кластеров во внешнем XML файле для дальнейшего считывания этого файла и обновления модели кластеров на основе новых данных.

Кроме того, процедура Двухэтапный кластерный анализ может анализировать большие файлы данных.

Иерархический кластерный анализ. Применение процедуры Иерархический кластерный анализ огранивается небольшими файлами данных (сотни объектов для кластеризации), однако она обладает следующими уникальными возможностями:

- Способность разбивать на кластеры как наблюдения, так и переменные.
- Способность формировать диапазон возможных решений и сохранять принадлежность к кластерам для каждого из этих решений.
- Наличие нескольких методов формирования кластеров, преобразования переменных и измерения расстояний между кластерами.

Процедура Иерархический кластерный анализ может анализировать интервальные (непрерывные), двоичные переменные или количества, если все переменные имеют один и тот же тип.

Кластерный анализ методом k-средних. Применение процедуры Кластерный анализ методом k-средних ограничивается непрерывными данными и требует задания числа классов заранее, но она имеет следующие уникальные возможности:

- Способность сохранять расстояния от центра кластера до каждого объекта.
- Способность считывать начальные центры кластеров из внешнего файла IBM SPSS Statistics и сохранять в нем окончательные центры кластеров.

Кроме того, процедура Кластерный анализ методом k-средних может анализировать большие файлы данных.

Глава 24. Двухэтапный кластерный анализ

Процедура Двухэтапный кластерный анализ представляет собой средство разведочного анализа для выявления естественного разбиения набора данных на группы (или кластеры), которое без ее применения трудно обнаружить. Алгоритм, используемый этой процедурой, имеет несколько привлекательных особенностей, которые отличают его от традиционных методов кластерного анализа:

- **Работа с категориальными и непрерывными переменными.** Предполагая независимость переменных, можно считать, что категориальные и непрерывные переменные имеют совместное полиномиально-нормальное распределение.
- **Автоматический выбор числа кластеров.** Сравнивая значения критерия отбора модели для различных кластерных решений, процедура может автоматически определить оптимальное число кластеров.
- **Масштабируемость.** Формируя дерево свойств кластеров (СК), которое является компактным представлением информации о наблюдениях, двухэтапный алгоритм позволяет анализировать большие файлы данных.

Пример. Компании производства потребительских товаров и розничной торговли регулярно применяют методы кластерного анализа к данным, описывающим покупательские привычки их клиентов, а также их пол, возраст, уровень доходов и т.д. Эти компании настраивают стратегии маркетинга и развития производства на каждую из групп потребителей, чтобы увеличить продажи и повысить приверженность потребителей маркам товаров.

Мера расстояния. Выбор в этой группе определяет, как вычисляется сходство между двумя кластерами.

- **Log-правдоподобия.** Мера правдоподобия приписывает переменным вероятностное распределение. Предполагается, что непрерывные переменные имеют нормальное распределение, а категориальные переменные - полиномиальное. Все переменные предполагаются независимыми.
- **Евклидова.** Евклидова мера является расстоянием "по прямой линии" между двумя кластерами. Она может быть использована, только когда все переменные являются непрерывными.

Число кластеров. Выбор в этой группе позволяет задать, как будет определяться число классов.

- **Определять автоматически.** Процедура автоматически определит "наилучшее" число классов, используя критерий, заданный в группе Критерий кластеризации. Дополнительно вы можете ввести положительное целое число, задающее максимальное число кластеров, которое должна рассмотреть процедура.
- **Задать.** Позволяет зафиксировать число кластеров в решении. Введите положительное целое число.

Количество непрерывных переменных. Эта группа дает сводную информацию об установках, касающихся стандартизации непрерывных переменных, заданных в диалоговом окне Параметры. Дополнительную информацию смотрите в разделе “Параметры процедуры Двухэтапный кластерный анализ” на стр. 112.

Критерий кластеризации. Выбор в этой группе задает способ, которым автоматический алгоритм кластеризации определяет число кластеров. Можно задать либо Байесовский информационный критерий (BIC), либо Информационный критерий Акаике (AIC).

Данные для двухэтапного кластерного анализа

Данные. Данная процедура работает как с непрерывными, так и с категориальными переменными. Наблюдения представляют собой объекты кластеризации, а переменные являются атрибутами, на которых основывается кластеризация.

Порядок наблюдений. Обратите внимание на то, что дерево свойств кластеров и окончательное решение могут зависеть от порядка наблюдений. Чтобы минимизировать эффект порядка наблюдений, расположите их в случайном порядке. Возможно, что вы захотите получить несколько различных решений с

наблюдениями, упорядоченными случайным образом, чтобы проверить стабильность данного решения. В ситуациях, когда это трудно сделать в силу чрезвычайно больших размеров файлов, можно в качестве альтернативы несколько раз выполнить процедуру с выборкой наблюдений, отсортировывая ее в случайном порядке.

Предположения. Мера расстояния, основанная на правдоподобии, предполагает, что переменные в кластерной модели являются независимыми. Кроме того предполагается, что каждая непрерывная переменная имеет нормальное (гауссово) распределение, а каждая категориальная переменная - полиномиальное распределение. Эмпирические исследования показывают, что эта процедура вполне устойчива к нарушениям предположений как о независимости, так и о распределениях, однако следует проверить, насколько эти предположения выполняются.

Для проверки независимости двух непрерывных переменных воспользуйтесь процедурой Парные корреляции. Для проверки независимости двух категориальных переменных воспользуйтесь процедурой Таблицы сопряженности. Для проверки независимости между непрерывной переменной и категориальной переменной воспользуйтесь процедурой Средние. Для проверки нормальности непрерывной переменной воспользуйтесь процедурой Исследовать. Для проверки того, что категориальная переменная имеет заданное полиномиальное распределение, воспользуйтесь процедурой Критерий хи-квадрат.

Как запустить процедуру Двухэтапный кластерный анализ

1. Выберите в меню:

   Анализ > Классификация > Двухэтапный кластерный анализ...

2. Выберите одну или несколько категориальных или непрерывных переменных.

Дополнительно вы можете:

- Установить критерии, по которым формируются кластеры.
- Выбрать установки для обработки шумов, выделения памяти, стандартизации переменных и ввода кластерной модели.
- Запрос вывода средства просмотра моделей.
- Сохранить результаты построения модели в рабочем файле или внешнем XML файле.

Параметры процедуры Двухэтапный кластерный анализ

Обработка выбросов. Эта группа позволяет обрабатывать выбросы специальным образом во время кластеризации, если заполняется дерево свойств кластеров (СК). Дерево свойств кластеров (СК) является полным, если оно не может больше принимать наблюдения в терминальный узел и никакой терминальный узел не может быть разделен.

- Если вы задали обработку шумов и дерево свойств (СК) кластеров заполняется, то оно будет перестроено после того, как наблюдения в разреженных листьях будут помещены в лист шума. Лист считается разреженным, если он содержит меньше наблюдений, чем заданный процент от максимального размера листа. После того как дерево перестроено, выбросы будут помещены в дерево свойств кластеров (СК), если это возможно. В противном случае выбросы будут отброшены.
- Если вы не выберете обработку шумов и дерево свойств кластеров (СК) заполняется, то оно будет перестроено с использованием большего порога изменения расстояния. После окончательного разбиения на кластеры, значения, которые не могут быть приписаны к кластерам, помечаются как выбросы. Кластеру выбросов дается идентификационный номер –1, и он не включается в подсчет числа кластеров.

Выделение памяти. Эта группа позволяет задать максимальное количество памяти в мегабайтах (MB), которую должен использовать алгоритм кластеризации. Если процедура превысит этот максимум, то она использует диск для хранения информации, которая не умещается в памяти. Задайте число, большее или равное 4.

- Проконсультируйтесь с вашим системным администратором по поводу максимального значения, которое может быть задано для вашей системы.

- Алгоритм может не найти подходящее или заданное число кластеров, если это значение слишком мало.

Стандартизация переменных. Алгоритм кластеризации работает со стандартизованными непрерывными переменными. Все непрерывные переменные, которые не стандартизованы, должны быть оставлены в списке Подлежат стандартизации. Чтобы несколько сэкономить время и снизить вычислительные затраты, можно поместить все непрерывные переменные, которые уже стандартизованы, в список Считаются стандартизованными.

Дополнительные опции

Критерии настройки дерева свойств кластеров (СК). Следующие установки алгоритма кластеризации относятся непосредственно к к дереву свойств кластеров (СК), и их следует изменять с осторожностью:

- **Начальный порог изменения расстояния.** Это начальный порог, используемый для построения дерева СК. Если включение данного наблюдения в лист дерева СК даст плотность, меньшую, чем порог, то лист не разделяется. Если плотность превосходит порог, то лист разделяется.
- **Максимальное число ветвей (на узел).** Максимальное число узлов, являющихся непосредственными потомками, которое может иметь узел.
- **Максимальная глубина дерева.** Максимальное число уровней, которое может иметь дерево СК.
- **Максимально возможное число узлов.** Это указывает максимальное число узлов в дереве СК, которые могут быть созданы процедурой, на основе функции $(b^{d+1} - 1) / (b - 1)$, где b есть максимальное число ветвей, а d есть максимальная глубина дерева. Отдавайте себе отчет в том, что чрезмерно большое дерево СК может вызвать перерасход системных ресурсов и неблагоприятно повлиять на эффективность процедуры. Каждый узел требует, как минимум, 16 байт.

Обновление модели кластеров. Эта группа позволяет импортировать и обновлять модель кластеров, полученную в результате проведенного ранее анализа. Входной файл содержит дерево СК в формате XML. Позже эта модель будет обновлена с помощью данных, содержащихся в активном файле. В главном диалоговом окне имена переменных должны быть выбраны в том же порядке, в котором они были заданы во время проведенного ранее анализа. Файл XML остается неизменным до тех пор, пока вы не сохраните информацию о новой модели под тем же именем. Дополнительную информацию смотрите в разделе “Вывод процедуры Двухэтапный кластерный анализ”.

Если задано обновление модели кластеров, используются те параметры, относящиеся к формированию дерева СК, которые были заданы для исходной модели. Более конкретно, используются мера расстояния, выделение памяти и установки в критериях настройки дерева СК для сохраненной модели, а любые установки для этих параметров, заданные в диалоговых окнах, игнорируются.

Примечание: При выполнении обновления модели кластеров процедура предполагает, что никакие из выбранных в активном наборе данных наблюдений, не были использованы для создания исходной модели кластеров. Процедура также предполагает, что наблюдения, используемые при обновлении модели, извлечены из той же генеральной совокупности, что и наблюдения, использованные при создании исходной модели; т.е. средние значения и дисперсии непрерывных переменных и уровни категориальных переменных предполагаются одинаковыми по обоим наборам наблюдений. Если "новый" и "старый" наборы наблюдений извлечены из неоднородных генеральных совокупностей, то для получения наилучших результатов следует запустить процедуру Двухэтапный кластерный анализ для объединенного набора наблюдений.

Вывод процедуры Двухэтапный кластерный анализ

Вывод. Эта группа предоставляет параметры для вывода таблиц результатов кластеризации.

- **Сводные таблицы.** Результаты выводятся в сводных таблицах.
- **Диаграммы и таблицы в средстве просмотра моделей.** Результаты выводятся в окне средства просмотра моделей.

- **Поля нормирования.** Здесь вычисляются данные кластера для переменных, которые не использовались в создании кластера. Поля нормирования могут отображаться вместе с входными функциями, если их выбрать в диалоговом окне Вывод. Поля с пропущенными значениями игнорируются.

Рабочий файл данных. Эта группа позволяет сохранить переменные в активном наборе данных.

- **Создать переменную принадлежности к кластерам.** Эта переменная содержит идентификационный номер кластера для каждого наблюдения. Эта переменная имеет имя *tsc_n* , где *n* является положительным целым числом, обозначающим порядковый номер операции сохранения активного набора данных, выполненной этой процедурой в течение данного сеанса работы.

Файлы XML. Окончательная модель кластеров и дерево СК являются двумя типами выходных файлов, которые можно экспортировать в формате XML.

- **Экспортировать окончательную модель.** Окончательная модель кластеров экспортируется в заданном файле в формате XML (PMML). Этот файл модели можно использовать для применения информации о модели к другим файлам данных с целью скоринга.
- **Экспортировать дерево свойств кластеров (СК).** Этот параметр позволяет сохранить текущее состояние дерева кластеров и обновить его позже, используя новые данные.

Средство просмотра кластеров

Кластерные модели обычно используются для выявления групп (или кластеров) похожих записей путем исследования переменных, в которых сходство членов одной группы велико, а сходство представителей разных групп мало. Полученные результаты можно использовать для идентификации взаимосвязей, которые другим путем было бы трудно обнаружить. Например, с помощью кластерного анализа предпочтений покупателей, уровня доходов и покупательских привычек можно идентифицировать типы клиентов, которые с большей вероятностью откликнуться на проводимую маркетинговую кампанию.

Имеются два подхода к интерпретации выведенных результатов кластерного анализа:

- Исследовать кластеры с целью выявления уникальных особенностей отдельных кластеров. *Содержит ли один кластер всех заемщиков с высоким доходом? Содержит ли данный кластер больше записей, чем остальные?*
- Исследовать поля по кластерам, чтобы определить, как распределяются значения среди кластеров. *Определяет ли уровень образования конкретного лица принадлежность к кластеру? Определяет ли высокая кредитная оценка принадлежность к тому или иному кластеру?*

Основная и дополнительная панель Средства просмотра кластеров, а также различные виды представления моделей могут помочь получить ответы на эти вопросы.

Чтобы получить информацию о кластерной модели, активируйте (двойным щелчком) в окне средства просмотра объект Средства просмотра моделей.

Средство просмотра кластеров

Средство просмотра кластеров состоит из двух панелей: основной, находящейся слева, и дополнительной, находящейся справа. Имеется два основных представления:

- Сводка для модели (по умолчанию). Дополнительную информацию смотрите в разделе “Вид представления Сводка для модели” на стр. 115.
- Кластеры. Дополнительную информацию смотрите в разделе “Вид представления Кластеры” на стр. 115.

В дополнительной панели доступны четыре вида представления:

- Важность предикторов. Дополнительную информацию смотрите в разделе “Вид представления Важность предикторов в кластерах” на стр. 117.
- Размеры кластеров (по умолчанию). Дополнительную информацию смотрите в разделе “Вид представления Размеры кластеров” на стр. 117.

- Распределение ячеек. Дополнительную информацию смотрите в разделе “Вид представления Распределение в ячейке” на стр. 117.
- Сравнение кластеров. Дополнительную информацию смотрите в разделе “Вид представления Сравнение кластеров” на стр. 117.

Вид представления Сводка для модели

В представлении Сводка для модели показан "мгновенный снимок" или сводка для кластерной модели, включая силуэтную меру связности и разделения кластеров, с использованием затенения для индикации низкого, среднего и хорошего качества полученных результатов. "Мгновенный снимок" дает возможность быстро понять, является ли качество разбиения на кластеры низким. В этом случае, возможно, стоит вернуться к узлу моделирования, чтобы скорректировать параметры для построения модели с целью получения более приемлемых результатов.

Решение вопроса о том, являются качество разбиения на кластеры низким, средним или хорошими основывается на работе Кауфмана и Rousseeuw (Kaufman and Rousseeuw (1990)), касающейся интерпретации кластерных структур. Показанное в сводке для модели качество разбиения считается хорошим, если согласно оценке Кауфмана и Rousseeuw имеется обоснованное или сильное свидетельство наличия кластерной структуры в данных. Среднее качество разбиения соответствует их оценке иметь слабое свидетельство, а низкое соответствует оценке не иметь значимого свидетельства наличия кластерной структуры.

Силуэтная мера усредняет по всем записям величину (B–A) / max(A,B), где A - это расстояние от записи до центра ее кластера, а B - расстояние от записи до центра ближайшего кластера, к которому она не принадлежит. Силуэтный коэффициент, равный 1, означал бы, что все наблюдения расположены точно в центрах их кластеров. Значение –1 означало бы, что все наблюдения расположены в центрах некоторого другого кластера. Значение 0 означает, что наблюдения расположены в среднем на равных расстояниях от центра их кластера и центра ближайшего кластера.

Сводка включает таблицу, которая содержит следующую информацию:

- **Алгоритм.** Используемый алгоритм кластеризации, например, "Двухэтапный".
- **Исходные показатели.** Число полей, также называемых **входными** или **предикторами**.
- **Кластеры.** Число кластеров в решении.

Вид представления Кластеры

Представление Кластеры содержит "сетку" кластеров по показателям, которая включает имена кластеров, объемы (размеры) и профили каждого кластера.

Столбцы в сетке содержат следующую информацию:

- **Кластер.** Номера кластеров, созданных в результате работы алгоритма.
- **Метка.** Любые метки, заданные для кластеров (по умолчанию они пустые). Дважды щелкните по ячейке, чтобы ввести метку, описывающую содержимое кластера, например, "Покупатели престижных автомобилей".
- **Описание.** Описание содержимого кластеров (по умолчанию оно пустое). Дважды щелкните по ячейке, чтобы ввести описание кластера, например, "возраст 55+ лет, профессионалы, доход превосходит $100000".
- **Размер.** Размер каждого кластера в виде процента от общего размера выборки, которая использовалась для построения модели кластеризации. В каждой ячейке размера внутри сетки выводится вертикальный столбец, показывающий размер кластера в процентах, размер кластера в процентах в числовом виде и число наблюдений в кластере.
- **Элементы.** Отдельные предикторы, по умолчанию отсортированные по общей важности. Если какие-либо столбцы имеют одинаковые размеры, они выводятся в возрастающем порядке номеров кластеров.

  Общая важность показателей обозначается интенсивностью цвет фона ячейки: наиболее важный показатель является наиболее темным. Легенда над таблицей показывает соответствие между важностью и интенсивностью цвета.

Если поместить указатель мыши на ячейку, то будет выведено полное имя/метка показателя и значение важности для этой ячейки. В зависимости от типа показателя и вида представления может быть выведена дополнительная информация. Для представления Центры кластеров такая информация будет включать статистику ячейки и значение ячейки, например: “Среднее: 4,32”. Для категориальных показателей в ячейке выводится имя наиболее часто встречающейся (модальной) категории и соответствующий ей процент.

Внутри представления Кластеры можно выбрать различные способы вывода информации о кластерах:

- Транспонировать кластеры и показатели. Дополнительную информацию смотрите в разделе “Транспонировать кластеры и показатели”.
- Сортировать показатели. Дополнительную информацию смотрите в разделе “Сортировать показатели”.
- Сортировать кластеры. Дополнительную информацию смотрите в разделе “Сортировать кластеры”.
- Выбрать содержимое ячеек. Дополнительную информацию смотрите в разделе “Содержимое ячеек.”.

Транспонировать кластеры и показатели: По умолчанию, кластеры выводятся как столбцы, а показатели выводятся как строки. Чтобы поменять местами строки и столбцы в выводе, нажмите кнопку **Транспонировать кластеры и показатели**, расположенной слева от кнопки **Сортировать показатели по**. Например, это можно сделать, чтобы реже пользоваться горизонтальной прокруткой при просмотре данных, когда выведено много кластеров.

Сортировать показатели: Кнопка **Сортировать показатели по** позволяет выбрать, как выводить ячейки показателей:

- **Общая важность.** Этот порядок сортировки задан по умолчанию. Показатели сортируются в убывающем порядке общей важности, и порядок сортировки один и тот же по всем кластерам. Если какие-либо показатели имеют совпадающие значения важности, то такие показатели перечисляются в возрастающем порядке имен показателей.
- **Важность для кластера.** Показатели сортируются по их важности для каждого кластера. Если какие-либо показатели имеют совпадающие значения важности, то такие показатели перечисляются в возрастающем порядке имен показателей. Если выбран этот вариант, порядок сортировки в кластерах обычно различается.
- **Имя.** Показатели сортируются по именам в алфавитном порядке.
- **Порядок следования в данных.** Показатели сортируются по порядку их расположения в наборе данных.

Сортировать кластеры: По умолчанию кластеры сортируются в убывающем порядке их размеров. Кнопка **Сортировать кластеры по** позволяет сортировать кластеры по именам в алфавитном порядке или, если заданы уникальные метки, в алфавитном порядке меток.

Показатели, которые имеют одну и ту же метку, сортируются по именам кластеров. Если кластеры отсортированы по метками и метки редактируются, то порядок сортировки автоматически меняется.

Содержимое ячеек.: Кнопки **Ячейки** позволяют изменить вывод содержимого ячеек для показателей и полей оценивания.

- **Центры кластеров.** По умолчанию ячейки выводят имена/метки показателей и показатель положения центра распределения для каждой комбинации кластера и показателя. Для непрерывных полей показывается среднее значение, а для категориальных полей - мода (категория, которая встречается наиболее часто) вместе с процентами по категориям.
- **Абсолютные распределения.** Показываются имена/метки показателей и абсолютные распределения показателей внутри каждого кластера. Для категориальных показателей в выводе показываются столбчатые диаграммы для категорий, упорядоченных по возрастанию значений данных. Для непрерывных полей в выводе показывается диаграмма сглаженной плотности, в которой используются конечные точки и интервалы, одинаковые для всех кластеров.

  Вывод, окрашенный в насыщенный красный цвет, показывает распределение для кластеров, тогда как бледный вывод представляет полные данные.

- **Относительные распределения.** Показываются имена/метки показателей и относительные распределения в ячейках. Вообще эти выводы подобны тем, в которых показываются абсолютные распределения, за исключением того, что на них выводятся относительные распределения.

  Вывод, окрашенный в насыщенный красный цвет, показывает распределение для кластеров, тогда как бледный вывод представляет полные данные.
- **Базовое представление.** Когда имеется много кластеров, бывает трудно увидеть все детали, не используя прокрутку. Чтобы снизить потребность в использовании прокрутки, выберите этот вид представления для вывода таблицы в более компактном виде.

Вид представления Важность предикторов в кластерах

Представление Важность предикторов показывает относительную важность каждого поля при оценивании модели.

Вид представления Размеры кластеров

Представление Размеры кластеров показывает круговую диаграмму, содержащую все кластеры. В каждом секторе показывается относительный размер каждого кластера в процентах. Поместите указатель мыши на сектор, чтобы вывести частоту в этом секторе.

Ниже этой диаграммы расположена таблица, выводящая следующую информацию о размерах:

- Размер наименьшего кластера (как частота и как процент от целого).
- Размер наибольшего кластера (как частота и как процент от целого).
- Отношение размера наибольшего кластера к размеру наименьшего кластера.

Вид представления Распределение в ячейке

Представление Распределение в ячейке выводит расширенную, более детальную диаграмму распределения данных для любой ячейки показателя, выбранной в таблице в представлении Кластеры в основной панели.

Вид представления Сравнение кластеров

Представление Сравнение кластеров имеет форму сетки с показателями в строках и выбранными кластерами в столбцах. Этот вид представления помогает лучше понять, какие факторы формируют кластер. Он также позволяет увидеть различие между кластерами, не только в сравнении со всеми данными, но и в сравнении между собой.

Чтобы выбрать кластеры для вывода, щелкните по верху столбца кластера в основной панели в представлении Кластеры. Пользуйтесь клавишами Ctrl и Shift совместно с щелчком мышью для выбора или отмены выбора нескольких кластеров для сравнения.

Примечание: Можно выбрать для вывода до пяти кластеров.

Кластеры выводятся в том порядке, в котором они были выбраны, тогда как порядок полей определяется параметром **Сортировать показатели по**. При выборе **по важности для кластера** поля всегда сортируются по общей важности.

Диаграммы на заднем плане показывают общие распределения каждого показателя:

- Категориальные показатели выводятся в виде точечных диаграмм, где для указания наиболее часто встречающейся (модальной) категории в каждом кластере (по показателям) используется размер точки.
- Непрерывные показатели выводятся в виде ящичных диаграмм с усами, которые показывают общие медианы и межквартильные диапазоны.

На эти изображения заднего плана накладываются ящичные диаграммы с усами для выбранных кластеров:

- Для непрерывных показателей квадратные точечные маркеры и горизонтальные линии показывают медиану и межквартильный диапазон для каждого кластера.
- Каждый кластер представляется своим цветом, показанным в верхней части изображения.

Перемещение по средству просмотра кластеров

Средство просмотра кластеров представляет собой интерактивный вывод. Вы можете:

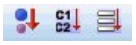
- Выбрать поле или кластер, чтобы увидеть больше деталей.
- Сравнить кластеры, чтобы выбрать элементы, представляющие интерес.
- Видоизменить вывод.
- Транспонировать оси.

Использование панели инструментов.

С помощью панели инструментов можно управлять выводом информации на левой и правой панелях. Пользуясь элементами управления панели инструментов, можно изменять ориентацию вывода (сверху вниз, слева направо или справа налево). Кроме того, параметрам средства просмотра можно возвратить значения, установленные по умолчанию, и открыть диалоговое окно, чтобы задать содержимое представления Кластеры в основной панели.

Возможность выбрать **Сортировать показатели по**, **Сортировать кластеры по**, **Ячейки** и **Показать** появляется, только если выбрать представление **Кластеры** в основной панели. Дополнительную информацию смотрите в разделе “Вид представления Кластеры” на стр. 115.

Таблица 2. Значки панели инструментов.

Значок	Тема
	Смотрите Транспонировать кластеры и показатели
	Смотрите опцию Сортировать показатели по
	Смотрите опцию Сортировать кластеры по
	Смотрите опцию Ячейки

Управление выводом для представления Кластеры

Чтобы получить доступ к управлению тем, что показано в представлении Кластеры в основной панели, нажмите кнопку **Показать**. Откроется диалоговое окно Показать.

Характеристики. Выбрано по умолчанию. Чтобы скрыть все входные показатели, снимите этот переключатель.

Поля для оценки. Выберите поля для оценки (поля, которые не используются для создания модели кластеров, но посылаются в средство просмотра моделей, чтобы оценить качество кластеров), которые будут выведены. По умолчанию ни одно не выводится. *Примечание* Поле оценки должно быть строкой с несколькими значениями. Этот переключатель недоступен, если нет ни одного поля для оценки.

Описания кластеров. Выбрано по умолчанию. Чтобы скрыть все ячейки описания кластеров, снимите этот переключатель.

Размеры кластеров. Выбрано по умолчанию. Чтобы скрыть все ячейки размеров кластеров, снимите этот переключатель.

Максимальное число категорий. Задайте максимальное число категорий для вывода на диаграммах категориальных показателей. Значение по умолчанию равно 20.

Фильтрация записей

При необходимости узнать больше о наблюдениях в отдельном кластере или группе кластеров можно выбрать подмножество записей для дальнейшего анализа на основе выбранных кластеров.

1. Выберите кластеры на панели представления Кластеры Средства просмотра кластеров. Чтобы выбрать несколько кластеров, щелкните мышью с нажатием клавиши Ctrl.
2. Выберите в меню:

   Генерировать > Записи фильтра...
3. Введите имя фильтрующей переменной. Записям из выбранных кластеров в этом поле будет присвоено значение 1. Всем остальным записям будет присвоено значение 0, и они будут исключены из дальнейшего анализа до тех пор, пока не будет изменено состояние фильтра.
4. Щелкните по **OK**.

Глава 25. Иерархический кластерный анализ

Эта процедура предназначена для выявления относительно однородных групп наблюдений (или переменных) по заданным характеристикам при помощи алгоритма, который вначале рассматривает каждое наблюдение (переменную) как отдельный кластер, а затем последовательно объединяет кластеры, пока не останется только один. Можно анализировать исходные переменные или воспользоваться набором стандартизирующих преобразований. Расстояния или меры сходства формируются процедурой Расстояния (Proximities). Чтобы помочь в выборе наилучшего решения, на каждом шаге выводятся разнообразные статистики.

Пример. Можно ли разбить телевизионные шоу на группы, так чтобы в каждой группе зрители, которых они привлекают, были схожи? С помощью иерархического кластерного анализа вы можете разделить (кластеризовать) телевизионные шоу (наблюдения) на однородные группы, исходя из характеристик их зрителей. Это можно использовать при сегментации рынка. Или вы можете разбить города (наблюдения) на однородные группы, что позволит отбирать сравнимые города для проверки различных маркетинговых стратегий.

Статистика. Порядок агломерации, матрица расстояний (или сходств) и состав кластеров для одного решения или диапазона решений. Диаграммы: дендрограммы и сосульчатые диаграммы.

Данные для иерархического кластерного анализа

Данные. Переменные могут быть количественными, бинарными или частотами. Масштаб измерения переменных важен - различия в масштабах могут повлиять на полученные кластерные решения. Если масштаб переменных сильно различается (например, одна переменная измерена в долларах, а другая - в годах), то следует подумать об их стандартизации (она может быть проведена автоматически с помощью процедуры Иерархическая кластерный анализ).

Порядок наблюдений. Если во входных данных существуют совпадающие расстояния или сходства или они появляются в обновленных кластерах в процессе объединения, то результирующее кластерное решение может зависеть от порядка наблюдений в файле Возможно, что вы захотите получить несколько различных решений с наблюдениями, упорядоченными случайным образом, чтобы проверить стабильность данного решения.

Допущения. Используемые расстояния или меры сходства должны соответствовать анализируемым данным (более полную информацию относительно выбора расстояний и мер сходства можно найти в описании процедуры Proximities (Расстояния)). Кроме того, в анализ необходимо включать все переменные, имеющие отношение к проблеме. Игнорирование важных переменных может привести к решению, вводящему в заблуждение. Поскольку иерархический кластерный анализ является разведочным методом, его результаты следует считать предварительными, пока они не будут подтверждены на независимой выборке.

Как запустить процедуру Иерархический кластерный анализ

1. Выберите в меню:

   Анализ > Классификация > Иерархическая кластеризация...

2. Если вы кластеризуете наблюдения, выберите, по крайней мере, одну числовую переменную. При кластеризации переменных выберите, по крайней мере, три числовые переменные.

По желанию можно выбрать идентифицирующую переменную для вывода меток наблюдений.

Задание метода иерархического кластерного анализа

Метод кластеризации. Возможные альтернативы: Межгрупповые связи, Внутригрупповые связи, Ближайший сосед, Дальний сосед, Центроидная кластеризация, Медианная кластеризация, Метод Варда.

Мера. Позволяет задать расстояние или меру сходства, которые будет использованы при кластеризации. Выберите тип данных и соответствующее расстояние или меру сходства:

- **Интервальная.** Возможные альтернативы: Евклидово расстояние, Квадрат расстояния Евклида, Косинус, Корреляция Пирсона, Чебышев, Блок, Минковского, Настроенная.
- **Количества.** Возможные альтернативы: Мера хи-квадрат и Мера фи-квадрат.
- **Бинарная.** Возможные альтернативы: Евклидово расстояние, Квадрат расстояния Евклида, Различие размеров, Различие структур, Дисперсия, Разброс, Форма, Простая совпадений, 4-точечная корреляция фи, Лямбда, *D* Андерберга, Дайс, Хаманн, Жаккар, Кульчинский 1, Кульчинский 2, Ланс и Уильямс, Очиай, Роджерс и Танимото, Рассел и Рао, Сокал и Снит 1, Сокал и Снит 2, Сокал и Снит 3, Сокал и Снит 4, Сокал и Снит 5, *Y* Юла и *Q* Юла.

Преобразовать значения. Позволяет стандартизировать значения данных либо для наблюдений, либо для переменных до вычисления близостей (недоступно для бинарных данных). Возможные методы стандартизации: *Z*–значения, Диапазон от –1 до 1, Диапазон от 0 до 1, Максимальная величина 1, Среднее 1 и Стандартное отклонение 1.

Преобразовать меры. Позволяет преобразовать значения, порожденные мерой расстояния. Преобразования выполняются после того, как вычислены значения меры расстояния. Возможные варианты преобразований: Взять модуль, Сменить знак, Привести к 0–1.

Статистики для процедуры Иерархический кластерный анализ

Порядок агломерации. Выводятся наблюдения или кластеры, объединяемые на каждом этапе, расстояния между объединяемыми наблюдениями или кластерами и уровень кластеризации, на котором к кластеру последний раз добавлялось наблюдение (или переменная).

Матрица близостей. Выводятся расстояния или сходства между объектами.

Принадлежность к кластерам. Выводится кластер, к которому отнесено каждое наблюдение для одного или нескольких этапов объединения кластеров. Возможными вариантами являются одно решение и диапазон решений.

Графики для процедуры Иерархический кластерный анализ

Дендрограмма. Выводится *дендрограмма* . Дендрограммы могут использоваться при исследовании взаимного притяжения формируемых кластеров и предоставить информацию о том, какое число кластеров сохранить.

Сосульчатый. Выводится *сосульчатая диаграмма* для всех кластеров или кластеров из заданного диапазона. Сосульчатые диаграммы дают информацию о том, как наблюдения объединяются в кластеры на каждой итерации анализа. Панель Ориентация позволяет выбрать между вертикальной и горизонтальной диаграммами.

Сохранение новых переменных в процедуре Иерархический кластерный анализ

Принадлежность к кластерам. Позволяет сохранить принадлежность к кластерам для одного решения или диапазона решений. Сохраненные переменные можно затем использовать в последующем анализе для изучения других различий между группами.

Дополнительные возможности синтаксиса команды CLUSTER

Процедура иерархической кластеризации использует синтаксис команды `CLUSTER` . Язык синтаксиса команд также позволяет:

- Использовать несколько методов кластеризации за один прогон процедуры.
- Считывать и анализировать матрицу близостей.
- Сохранять матрицу близостей для дальнейшего анализа.
- Задавать любые значения порядков и корней для настраиваемой (степенной) меры расстояния.
- Задавать имена сохраняемых переменных.

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Глава 26. Кластерный анализ методом K средних

Эта процедура пытается выявить относительно однородные группы наблюдений на основе выбранных характеристик, используя алгоритм, позволяющий обработать большое число наблюдений. Однако этот алгоритм требует указания числа кластеров. Вы можете задать начальные центры кластеров, если такая информация вам доступна. Вы можете выбрать один из двух методов классификации наблюдений, либо итеративно обновляя центры кластеров, либо ограничиваясь только классификацией. Вы можете сохранить принадлежность к кластерам, информацию о расстояниях и окончательные центры кластеров. Дополнительно вы можете задать переменную, значения которой будут использоваться в качестве меток наблюдений при выводе результатов. Вы можете также запросить вывод *F* -статистик дисперсионного анализа. Относительные величины этих статистик дают информацию о вкладе каждой переменной в разделение групп.

Пример. Можно ли разбить телевизионные шоу на группы, так чтобы в каждой группе зрители, которых они привлекают, были схожи? С помощью кластерного анализа методом *k* -средних вы можете разделить (кластеризовать) телевизионные шоу (наблюдения) на *k* однородных групп, исходя из характеристик их зрителей. Это можно использовать при сегментации рынка. Или вы можете разбить города (наблюдения) на однородные группы, что позволит отбирать сравнимые города для проверки различных маркетинговых стратегий.

Статистика. Полное решение: начальные центры кластеров, таблица дисперсионного анализа. Для каждого наблюдения: информация о кластерах, расстояние от центра кластера.

Данные для кластерного анализа методом k-средних

Данные. Переменные должны быть количественными и измеренными в интервальной шкале или шкале отношений. Если переменные являются двоичными или количествами, воспользуйтесь процедурой Иерархический кластерный анализ.

Порядок наблюдений и начальных центров кластеров. Алгоритм, используемый по умолчанию для выбора начальных центров кластеров, не является инвариантным относительно порядка наблюдений. Параметр **Использовать скользящие средние** в диалоговом окне Итерации делает получающееся в результате решение потенциально зависимым от порядка наблюдений, независимо от того, как выбираются начальные центры кластеров. При использовании любого из этих методов, вы, возможно, захотите получить несколько различных решений с наблюдениями, расположенными в случайном порядке, чтобы удостовериться в стабильности данного решения. Задание начальных центров кластеров и не использование параметра **Использовать скользящие средние** позволит избежать проблем, связанных с порядком наблюдений. Однако упорядочение начальных центров кластеров может повлиять на решение, если имеются совпадающие расстояния от наблюдений до центров кластеров. Чтобы оценить стабильность данного решения, можно сравнить результаты анализа с различными перестановками значений начальных центров.

Допущения. Для вычисления расстояний используется простое евклидово расстояние. Если необходимо задать другой тип расстояния или меры сходства, обратитесь к процедуре Иерархический кластерный анализ. Масштабирование переменных играет важную роль. Если ваши переменные имеют различный масштаб измерений (например, одна переменная измерена в долларах, а вторая - в годах), то результаты могут быть некорректными. В этой ситуации необходимо подумать о стандартизации ваших переменных до выполнения кластерного анализа методом *k* -средних (это можно сделать при помощи процедуры Описательные статистики). Предполагается, что выбрано подходящее число кластеров, а в анализ включены все существенные переменные. Если вы неправильно выбрали число кластеров или не включили важные переменные, то полученные результаты также могут ввести вас в заблуждение.

Как запустить Кластерный анализ методом k-средних

1. Выберите в меню:

Анализ > Классификация > Кластерный методом K средних...

2. Выберите переменные для использования в кластерном анализе.
3. Задайте число кластеров. (Оно должно быть не меньше двух и не больше числа наблюдений в файле данных.)
4. Выберите либо метод **Итерации и классификация** , либо метод **Только классификация** .
5. Дополнительно можно выбрать идентификационную переменную, чтобы метить наблюдения.

Эффективность кластерного анализа методом k-средних

Алгоритм *k* -средних эффективен прежде всего потому, что он не нуждается в вычислении всех попарных расстояний между наблюдениями, в отличие от большинства других алгоритмов кластеризации, включая тот, что используется в процедуре иерархического кластерного анализа.

Для достижения максимальной эффективности возьмите выборку из наблюдений и используйте метод **Итерации и классификация** , чтобы определить центры кластеров. Выберите **Записать окончательные в** . Затем вернитесь к полному файлу данных и выберите **Только классификация** в качестве метода и выберите **Прочитать начальные из** , чтобы классифицировать весь файл с использованием центров, оцененных по выборке. Вы можете записывать в файл или набор данных, а также считывать из них. Наборы данных доступны для последующего использования в том же сеансе но не сохраняются как файлы до тех пор, пока они не будут сохранены явно до окончания текущего сеанса. Имена наборов данных должны удовлетворять требованиям к именам переменных. Дополнительную информацию смотрите в разделе .

Итерации в кластерном анализе методом k-средних

Примечание: Эти опции доступны, только если вы выберите метод **Итерации и классификация** в диалоговом окне Кластерный анализ методом K средних.

Максимум итераций. Ограничивает число итераций для алгоритма *k* -средних. Алгоритм останавливается после заданного здесь числа итераций, даже если не выполняется критерий сходимости. Это число должно быть от 1 до 999.

Если необходимо воспроизвести алгоритм, использовавшийся командой QUICK CLUSTER в старых версиях (до 5.0), установите **Максимум итераций** равным 1 .

Критерий сходимости. Задает условие прекращения итераций. Оно выражает долю минимального расстояния между начальными центрами кластеров, поэтому должно быть больше 0, но не превышать 1. Если значение критерия равно, например, 0.02, итерации прекращаются, когда полная итерация не сдвигает ни один из центров кластеров на расстояние, превышающее 2% от наименьшего расстояния между центрами любых начальных кластеров.

Использовать скользящие средние. Позволяет запросить обновление центров кластеров после классификации очередного наблюдения. Если этот пункт не отмечен, новые центры кластеров вычисляются после распределения по кластерам всех наблюдений.

Сохранение новых переменных в кластерном анализе методом k-средних

Вы можете сохранить следующую информацию о решении в виде новых переменных для использования в последующем анализе:

Принадлежность к кластеру. Создается новая переменная, показывающая окончательную принадлежность каждого наблюдения к кластеру. Значения этой новой переменной могут меняться от 1 до числа кластеров.

Расстояние от центра кластера. Создается новая переменная, показывающая евклидово расстояние между каждым наблюдением и центром кластера, куда оно было отнесено.

Параметры процедуры Кластерный анализ методом К-средних

Статистика. Вы можете выбрать следующие статистики: начальные центры кластеров, таблица дисперсионного анализа, а также информация о принадлежности к кластерам для каждого наблюдения.

- *Начальные центры кластеров*. Начальная оценка положения средних для каждого кластера. По умолчанию, отбираются объекты, находящиеся на значительном расстоянии друг от друга, причем столько, сколько задано кластеров. Начальные центры кластеров используются на первом этапе грубой классификации, а затем обновляются.
- *Таблица дисперсионного анализа*. Выводится таблица дисперсионного анализа, включающая одномерный F-критерий для каждой кластерной переменной. F-критерий приводится для чисто ориентировочных целей, и выдаваемые вероятности не подлежат интерпретации. Таблица не выдается, если все наблюдения попадают в один кластер.
- *Конечный кластер для каждого наблюдения*. Для каждого наблюдения указывается финальный кластер, к которому оно отнесено, и евклидово расстояние до центра этого кластера. Выводится также евклидово расстояние между центрами финальных кластеров.

Пропущенные значения. Возможными альтернативами являются **Исключать целиком** и **Исключать наблюдения попарно** .

- **Исключать целиком.** Наблюдения с пропущенными значениями в любой из кластерных переменных исключаются из анализа.
- **Исключать попарно.** Наблюдения относятся к кластерам на основании расстояний, вычисленных по всем переменным с непропущенными значениями.

Команда QUICK CLUSTER: дополнительные возможности

Процедура Кластерный анализ методом k-средних использует синтаксис команды `QUICK CLUSTER` . Язык синтаксиса команд также позволяет:

- Использовать первые *k* наблюдений в качестве начальных центров кластеров, тем самым избегая прохода по данным, обычно применяемого, чтобы их оценить.
- Задать начальные центры кластеров напрямую, как часть командного синтаксиса.
- Задавать имена сохраняемых переменных.

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Глава 27. Непараметрические критерии

Непараметрические критерии требуют самых минимальных допущений о распределении данных. Критерии, доступные с помощью данных диалоговых окон, можно разделить на три общие категории в зависимости от организации данных:

- Одновыборочный критерий анализирует единственное поле.
- Критерий для связанных выборок сравнивает два или большее число полей для одного и того же набора наблюдений.
- Критерий для независимых выборок анализирует единственное поле, разбитое на группы категориями другого поля.

Одновыборочные непараметрические критерии

Процедура Одновыборочные непараметрические критерии выявляет различия в единичных полях, используя один или несколько непараметрических критериев. Непараметрические критерии не предполагают, что данные соответствуют нормальному распределению.

Какова ваша цель? Вкладка Цель позволяет быстро задать параметры для решения различных и в то же время наиболее типичных задач проверки гипотез.

- **Автоматически сравнить наблюденные данные с гипотетическими** Для этой цели к категориальным полям, имеющим только две категории, применяется биномиальный критерий. Ко всем остальным категориальным полям применяется критерий хи-квадрат. К непрерывным полям применяется критерий Колмогорова-Смирнова.
- **Проверить последовательность на случайность.** Для проверки наблюденной последовательности данных на случайность используется критерий серий.
- **Настроить анализ.** Выберите этот вариант при желании вручную внести коррективы в параметры тестирования на вкладке Параметры. Обратите внимание на то, что этот выбор производится автоматически, если на вкладке Параметры сделать изменения, несовместимые с выбранной целью.

Чтобы получить одновыборочные непараметрические критерии

Выберите в меню:

Анализ > Непараметрические критерии > Одна выборка...

1. Нажмите кнопку **Выполнить**.

Дополнительно вы можете:

- Задать цель на вкладке Цель.
- Задать назначение полей на вкладке Поля.
- Самостоятельно выбрать параметры на вкладке Параметры.

Вкладка Поля

На вкладке Поля задаются проверяемые поля.

Использовать заранее заданные роли. При этом варианте выбора используется имеющаяся информация о полях. Все поля с предопределенными ролями, такими как Входная, Целевая или Двойного назначения, будут использованы как проверяемые поля. Необходимо задать, по крайней мере, одно поле для проверки.

Настроить назначения полей. Этот вариант выбора позволяет не принимать во внимание роли, назначенные полям. После выбора этого варианта задайте поля:

- **Проверяемые поля.** Выберите одно или несколько полей.

Вкладка Параметры

Вкладка Параметры содержит несколько различных групп параметров, которые можно изменять, чтобы точно настроить то, как алгоритм будет обрабатывать имеющиеся данные. Если в настройку параметров по умолчанию внести изменения, которые несовместимы с выбранной целью, то выбор на вкладке Цели будет автоматически изменен на **Настроить анализ**.

Выберите критерии

Эти параметры определяют, какие критерии будут применяться к полям, заданным на вкладке Поля.

Автоматически выбрать критерии на основе данных. При выборе этого варианта к категориальным полям, имеющим только две категории (с не пропущенными значениями), применяется биномиальный критерий. Ко всем остальным категориальным полям применяется критерий хи-квадрат. К непрерывным полям применяется критерий Колмогорова-Смирнова.

Настроить критерии. Этот вариант дает возможность выбрать применяемые критерии.

- **Сравнить наблюденную двоичную вероятность с гипотетической (Биномиальный критерий).** Биномиальный критерий можно применить ко всем полям. Применяется одновыборочный критерий для проверки того, соответствует ли выборочное распределение поля признака (категориальное поле с двумя категориями) заданному биномиальному распределению. Дополнительно можно запросить вывод доверительных интервалов. Подробности этих параметров критериев смотрите в разделе “Вкладка Параметры биномиального критерия”.
- **Сравнить наблюденные вероятности с гипотетическими (критерий Хи-квадрат).** Критерий хи-квадрат применяется к номинальным и порядковым полям. Применяется одновыборочный критерий, который вычисляет статистику хи-квадрат на основе разностей между наблюденными и ожидаемыми частотами категорий поля. Подробности этих параметров критериев смотрите в разделе “Вкладка Параметры критерия Хи-квадрат” на стр. 131.
- **Сравнить наблюденное распределение с гипотетическим (критерий Колмогорова-Смирнова).** Критерий Колмогорова-Смирнова применяется к непрерывным и порядковым полям. Применяется одновыборочный критерий для проверки того, что выборочная функция распределения для поля согласуется с равномерным, нормальным или экспоненциальным распределением, а также распределением Пуассона. Подробности о задании параметров критериев смотрите в разделе “Параметры критерия Колмогорова-Смирнова” на стр. 131.
- **Сравнить медиану с гипотетической (критерий знаковых рангов Уилкоксона).** Критерий знаковых рангов Уилкоксона применяется к непрерывным и порядковым полям. Для проверки медианы значений поля применяется одновыборочный критерий. Задайте число в качестве гипотетического значения медианы.
- **Проверить последовательность на случайность (критерий серий).** Критерий серий применяется ко всем полям. Применяется одновыборочный критерий для проверки того, что последовательность значений дихотомизированного поля является случайной. Подробности этих параметров критериев смотрите в разделе “Опции критерия серий” на стр. 131.

Вкладка Параметры биномиального критерия: Биномиальный критерий предназначен для полей признаков (категориальных полей только с двумя категориями), однако он применяется ко всем полям, используя правило задания "успеха".

Гипотетическая доля. Здесь задается ожидаемая доля записей, заданных как "успех", или *p*. Задайте значение больше 0 и меньше 1. Значение по умолчанию равно 0,5.

Доверительный интервал. Доступны следующие методы вычисления доверительных интервалов для двоичных данных:

- **Клоппер-Пирсон (точный).** Точный интервал, основанный на функции распределения биномиального распределения.

- **Джеффриз.** Байесовский интервал, основанный на апостериорном распределении *p* при использовании априорного распределения вероятностей Джеффриза.
- **Отношение правдоподобия.** Интервал, основанный на функции правдоподобия для *p*.

Задать "успех" для категориальных полей. Здесь задается, как для категориальных полей определяется "успех", т.е. значение или значения, доля которых сравнивается с гипотетической долей.

- **Использовать первую категорию, встретившуюся в данных**. В качестве "успеха" для биномиального критерия используется первое значение, найденное в выборке. Этот выбор применим только к номинальным и порядковым полям и только с двумя категориями. Все остальные категориальные поля, заданные на вкладке Поля, проверяться не будут. Это задано по умолчанию.
- **Задать значения "успеха"**. Биномиальный критерий применяется с целым списком значений, заданных в качестве "успеха". Задайте список текстовых или числовых значений. Значения из этого списка необязательно должны присутствовать в выборке.

Задать "успех" для количественных полей. Здесь задается, как для непрерывных полей определяется "успех", т.е. значение или значения, доля которых сравнивается с тестовым значением. Успех задается как значения, равные или меньшие, чем точка отсечения.

- **Средняя точка выборки** задает в качестве точки отсечения среднее значение минимального и максимального значений.
- **Заданная точка отсечения** позволяет задать значение точки отсечения.

Вкладка Параметры критерия Хи-квадрат: У всех категорий равные вероятности. Это дает равные частоты всем категориям из выборки. Это вариант по умолчанию.

Задать ожидаемую вероятность. Это позволяет задать неравные частоты для заданного списка категорий. Задайте список текстовых или числовых значений. Значения из этого списка необязательно должны присутствовать в выборке. В столбце **Категория** задайте значения категорий. В столбце **Относительная частота** для каждой категории задайте положительное значение. Задаваемые частоты рассматриваются как относительные частоты, так что, например, задание частот 1, 2 и 3 эквивалентно заданию частот 10, 20 и 30, причем оба эти набора частот говорят о том, что ожидается, что 1/6 записей попадет в первую категорию, 1/3 - во вторую и 1/2 - в третью. Когда задаются ожидаемые вероятности, задаваемые значения категорий должны включать все значения полей в данных. В противном случае для соответствующего поля тест не будет выполнен.

Параметры критерия Колмогорова-Смирнова: В этом диалоговом окне задается, какие распределения должны быть проверены, а также параметры предполагаемых распределений.

Нормальное. Использовать данные выборки использует наблюденные среднее и стандартное отклонение, **Задать** позволяет задать значения.

Равномерное. Использовать данные выборки использует наблюденные минимум и максимум, **Задать** позволяет задать значения.

Экспоненциальные. Выборочное среднее использует наблюденное среднее значение, **Задать** позволяет задать значения.

Пуассона. Выборочное среднее использует наблюденное среднее значение, **Задать** позволяет задать значения.

Опции критерия серий: Критерий серий предназначен для полей признаков (категориальных полей только с двумя категориями), однако его можно применить ко всем полям, используя правило задания групп.

Задать группы для категориальных полей. Доступны следующие параметры:

- **В выборке имеется только две категории**. Критерий серий применяется с использованием значений для задания групп, найденных в выборке. Этот выбор применим только к номинальным и порядковым полям и только с двумя категориями. Все остальные категориальные поля, заданные на вкладке Поля, проверяться не будут.
- **Перекодировать данные в 2 категории**. Критерий серий применяется с использованием целого заданного списка значений для задания одной из групп. Все остальные значения из выборки задают другую группу. В выборке необязательно должны присутствовать все значения из списка, но, по крайней мере, одна запись должна быть в каждой группе.

Задать точку отсечения для количественных полей. Здесь задается, как формируются группы для непрерывных полей. К первой группе относятся значения, равные или меньшие, чем точка отсечения.

- **Выборочная медиана** задает точку отсечения равной выборочной медиане.
- **Выборочное среднее** задает точку отсечения равной выборочному среднему.
- **Задать** позволяет задать значение точки отсечения.

Параметры критериев

Уровень значимости. Здесь задается уровень значимости (альфа) для всех критериев. Задайте числовое значение между 0 и 1. 0,05 является значением по умолчанию.

Доверительный интервал (%). Здесь задается доверительный уровень для всех рассчитываемых доверительных интервалов. Укажите числовое значение от 0 до 100. Значение по умолчанию - 95.

Исключенные наблюдения. Здесь задается, какие наблюдения используются при выполнении тестов.

- **Исключать наблюдения целиком** означает, что записи с пропущенными значениями в любых полях, указанных на вкладке Поля, исключаются из анализа.
- **Исключать по отдельности** означает, что записи с пропущенными значениями в поле, используемом при выполнении конкретного теста, не используются при выполнении этого теста. Когда задано одновременно несколько тестов, для каждого из них вопрос об использовании записей с пропущенными значениями решается независимо от других.

Пользовательские значения отсутствия

Пользовательские пропущенные значения для категориальных полей. Категориальные поля должны иметь допустимые значения, для того чтобы запись была включена в анализ. С помощью этих управляющих элементов можно определить, рассматривать ли пользовательские пропущенные значения в категориальных полях как допустимые. Системные пропущенные значения и пропущенные значения для количественных полей всегда рассматриваются как недопустимые.

Команда NPTESTS: дополнительные возможности

Язык синтаксиса команд также позволяет:

- Задать применение одновыборочного критерия, а также критериев для независимых и связанных выборок, запуская процедуру один раз.

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Непараметрические критерии для независимых выборок

Процедура Непараметрические критерии для независимых выборок выявляет различия между двумя или большим числом групп, используя один или несколько непараметрических критериев. Непараметрические критерии не предполагают, что данные соответствуют нормальному распределению.

Какова ваша цель? Вкладка Цель позволяет быстро задать параметры для решения различных и в то же время наиболее типичных задач проверки гипотез.

- **Автоматически сравнить распределения для групп.** Для этой цели применяется U-критерий Манна-Уитни к данным с 2 группами или однофакторный дисперсионный анализ Краскала-Уоллиса к данным с *k* группами.
- **Сравнить медианы для групп.** Для этой цели применяется медианный критерий, сравнивающий наблюденные медианы в группах.
- **Настроить анализ.** Выберите этот вариант при желании вручную внести коррективы в параметры тестирования на вкладке Параметры. Обратите внимание на то, что этот выбор производится автоматически, если на вкладке Параметры сделать изменения, несовместимые с выбранной целью.

Чтобы получить непараметрические критерии для независимых выборок

Выберите в меню:

Анализ > Непараметрические критерии > Независимые выборки...

1. Нажмите кнопку **Выполнить**.

Дополнительно вы можете:

- Задать цель на вкладке Цель.
- Задать назначение полей на вкладке Поля.
- Самостоятельно выбрать параметры на вкладке Параметры.

Вкладка Поля

На вкладке Поля задается, какие поля сравниваются и какие поля задают группы.

Использовать заранее заданные роли. При этом варианте выбора используется имеющаяся информация о полях. Все непрерывные и порядковые поля с предопределенными ролями, такими как Целевая или Двойного назначения, будут использованы как проверяемые поля. Если имеется единственное категориальное поле с предопределенной ролью Входная, то оно будет использовано в качестве группирующего поля. В противном случае по умолчанию не будут использоваться группирующие поля, и назначения полей необходимо задать самостоятельно. Требуется, по крайней мере, одно проверяемое поле и одно группирующее поле.

Настроить назначения полей. Этот вариант выбора позволяет не принимать во внимание роли, назначенные полям. После выбора этого варианта задайте поля:

- **Проверяемые поля.** Выберите одно или несколько непрерывных или порядковых полей.
- **Группы.** Выберите категориальное поле.

Вкладка Параметры

Вкладка Параметры содержит несколько различных групп параметров, которые можно изменять, чтобы точно настроить то, как алгоритм будет обрабатывать имеющиеся данные. Если в настройку параметров по умолчанию внести изменения, которые несовместимы с выбранной целью, то выбор на вкладке Цели будет автоматически изменен на **Настроить анализ**.

Выберите критерии

Эти параметры определяют, какие критерии будут применяться к полям, заданным на вкладке Поля.

Автоматически выбрать критерии на основе данных. При выборе этого варианта применяется U-критерий Манна-Уитни к данным с 2 группами или однофакторный дисперсионный анализ Краскала-Уоллиса к данным с *k* группами.

Настроить критерии. Этот вариант дает возможность выбрать применяемые критерии.

- **Сравнить распределения для групп.** Здесь представлены критерии для независимых выборок для проверки того, извлечены ли выборки из одной и той же генеральной совокупности.

  U Манна-Уитни (для 2-х выборок) использует ранги всех наблюдений, чтобы проверить, извлечены ли группы из одной и той же генеральной совокупности. Первое в порядке по возрастанию значение группирующего поля задает первую группу, а второе задает вторую группу. Если группирующее поле имеет более двух значений, то этот тест не выполняется.

  Колмогорова-Смирнова (для 2-х выборок) чувствителен к любым различиям двух распределений в медианах, разбросе, скошенности и т.д. Если группирующее поле имеет более двух значений, то этот тест не выполняется.

  Проверить последовательность на случайность (Вальда-Вольфовица для 2-х выборок) задает применение критерия серий с групповой принадлежностью в качестве признака. Если группирующее поле имеет более двух значений, то этот тест не выполняется.

  Однофакторный дисперсионный анализ Краскала-Уоллиса (для k выборок) является обобщением U-критерия Манна-Уитни и непараметрическим аналогом одномерного дисперсионного анализа. Дополнительно можно запросить множественные сравнения *k* выборок, выбрав либо **Все попарно**, либо **Пошагово вниз**.

  Критерий для упорядоченных альтернатив (Джонкхира-Терпстры для k выборок) является более мощной альтернативой критерию Краскала-Уоллиса, когда *k* выборок имеют естественное упорядочение. Например, *k* совокупностей могут представлять собой *k* возрастающих температур. Проверяется гипотеза о том, что разные температуры дают одинаковое распределение откликов, против альтернативной гипотезы о том, что при увеличении температуры возрастает и величина отклика. Здесь альтернативная гипотеза упорядочена; следовательно, наиболее подходящим будет критерий Джонкхира-Терпстры. **От наименьшего к наибольшему** задает альтернативную гипотезу, что параметр положения первой группы меньше или равен параметру во второй группе, который меньше или равен параметру третьей группы и так далее. **От наибольшего к наименьшему** задает альтернативную гипотезу, что параметр положения первой группы больше или равен параметру во второй группе, который больше или равен параметру третьей группы и так далее. Для обеих опций альтернативная гипотеза предполагает также, что не все положения равны. Дополнительно можно запросить множественные сравнения *k* выборок, выбрав либо **Все попарно**, либо **Пошагово вниз**.
- **Сравнить диапазоны для групп.** Здесь представлены критерии для независимых выборок для проверки того, что группы имеют одинаковый разброс. **Экстремальной реакции Мозеса (для 2-х выборок)** сравнивает контрольную группу с группой сравнения. Первое в порядке по возрастанию значение группирующего поля задает контрольную группу, а второе задает группу сравнения. Если группирующее поле имеет более двух значений, то этот тест не выполняется.
- **Сравнить медианы для групп.** Здесь представлены критерии для независимых выборок для проверки того, что группы имеют одинаковые медианы. **Медианный критерий (для k выборок)** может использовать либо объединенную выборочную медиану (вычисленную по всем записям в наборе данных), либо заданное в качестве гипотетического значение медианы. Дополнительно можно запросить множественные сравнения *k* выборок, выбрав либо **Все попарно**, либо **Пошагово вниз**.
- **Оценить доверительный интервал для групп. Оценка Ходжеса-Лемана (для 2-х выборок)** вычисляет оценку по независимым выборкам и доверительный интервал для разности медиан двух групп. Если группирующее поле имеет более двух значений, то этот тест не выполняется.

Параметры критериев

Уровень значимости. Здесь задается уровень значимости (альфа) для всех критериев. Задайте числовое значение между 0 и 1. 0,05 является значением по умолчанию.

Доверительный интервал (%). Здесь задается доверительный уровень для всех рассчитываемых доверительных интервалов. Укажите числовое значение от 0 до 100. Значение по умолчанию - 95.

Исключенные наблюдения. Здесь задается, какие наблюдения используются при выполнении тестов. **Исключать наблюдения целиком** означает, что записи с пропущенными значениями в любых полях, указанных в любой подкоманде, исключаются из анализа. **Исключать по отдельности** означает, что записи с пропущенными значениями в поле, используемом при выполнении конкретного теста, не используются при

выполнении этого теста. Когда задано одновременно несколько тестов, для каждого из них вопрос об использовании записей с пропущенными значениями решается независимо от других.

Пользовательские пропущенные значения

Пользовательские пропущенные значения для категориальных полей. Категориальные поля должны иметь допустимые значения, для того чтобы запись была включена в анализ. С помощью этих управляющих элементов можно определить, рассматривать ли пользовательские пропущенные значения в категориальных полях как допустимые. Системные пропущенные значения и пропущенные значения для количественных полей всегда рассматриваются как недопустимые.

Команда NPTESTS: дополнительные возможности

Язык синтаксиса команд также позволяет:

- Задать применение одновыборочного критерия, а также критериев для независимых и связанных выборок, запуская процедуру один раз.

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Непараметрические критерии для связанных выборок

Выявляются различия между двумя или большим числом связанных полей при помощи одного или нескольких непараметрических критериев. Непараметрические критерии не предполагают, что данные соответствуют нормальному распределению.

Данные. Каждая запись соответствует конкретному объекту, для которого два или более связанных измерений сохраняются в отдельных полях в наборе данных. Например, исследование эффективности диеты можно проводить, используя непараметрические критерии для связанных выборок, если вес каждого объекта измеряется через равные интервалы времени и сохраняется в полях с метками *Вес до начала диеты*, *Вес в середине диеты* и *Вес по окончании диеты*. Эти поля являются "связанными".

Какова ваша цель? Вкладка Цель позволяет быстро задать параметры для решения различных и в то же время наиболее типичных задач проверки гипотез.

- **Автоматически сравнить наблюденные данные с гипотетическими.** При выборе этой цели к категориальным данным применяется критерий Макнемара, если заданы два поля, и критерий Q Кокрена, если задано более двух полей. К количественным данным в этом случае применяется парный критерий знаковых рангов Уилкоксона, если заданы два поля, и двухфакторный дисперсионный анализ Фридмана по рангам, если задано более двух полей.
- **Настроить анализ.** Выберите этот вариант при желании вручную внести коррективы в параметры тестирования на вкладке Параметры. Обратите внимание на то, что этот выбор производится автоматически, если на вкладке Параметры сделать изменения, несовместимые с выбранной целью.

Если задаются поля с различающимися шкалами измерений, то они сначала разделяются по шкалам измерений, а затем к каждой группе применяется подходящий критерий. Например, если в качестве цели выбрать **Автоматически сравнить наблюденные данные с гипотетическими**, и задать 3 количественных, а также 2 номинальных поля, то к количественным полям будет применен критерий Фридмана, а к номинальным полям будет применен критерий Макнемара.

Чтобы применить непараметрические критерии для связанных выборок

Выберите в меню:

Анализ > Непараметрические критерии > Связанные выборки...

1. Нажмите кнопку **Выполнить**.

Дополнительно вы можете:

- Задать цель на вкладке Цель.
- Задать назначение полей на вкладке Поля.
- Самостоятельно выбрать параметры на вкладке Параметры.

Вкладка Поля

На вкладке Поля задаются проверяемые поля.

Использовать заранее заданные роли. При этом варианте выбора используется имеющаяся информация о полях. Все поля с предопределенными ролями, такими как Целевая или Двойного назначения, будут использованы как проверяемые поля. Необходимо задать, по крайней мере, два поля для проверки.

Настроить назначения полей. Этот вариант выбора позволяет не принимать во внимание роли, назначенные полям. После выбора этого варианта задайте поля:

- **Проверяемые поля.** Выберите два поля или более. Каждое поле соответствует отдельной связанной выборке.

Вкладка Параметры

Вкладка Параметры содержит несколько различных групп параметров, которые можно изменять, чтобы точно настроить то, как процедура будет обрабатывать имеющиеся данные. Если в настройку параметров по умолчанию внести изменения, которые несовместимы с другими целями, то выбор на вкладке Цель будет автоматически изменен на **Настроить анализ**.

Выберите критерии

Эти параметры определяют, какие критерии будут применяться к полям, заданным на вкладке Поля.

Автоматически выбрать критерии на основе данных. При выборе этого варианта к категориальным данным применяется критерий Макнемара, если заданы два поля, и критерий Q Кокрена, если задано более двух полей. К количественным данным в этом случае применяется парный критерий знаковых рангов Уилкоксона, если заданы два поля, и двухфакторный дисперсионный анализ Фридмана по рангам, если задано более двух полей.

Настроить критерии. Этот вариант дает возможность выбрать применяемые критерии.

- **Проверить наличие изменений в двоичных данных. Критерий Макнемара (для 2 выборок)**можно применить к категориальным полям. При этом применяется критерий для связанных выборок, который проверяет, являются ли равновероятными комбинации значений двух флаговых полей (категориальных полей только с двумя значениями). Если на вкладке Поля задано более двух полей, этот критерий не применяется. Подробности о задании параметров критериев смотрите в разделе “Критерий Макнемара: определить успех” на стр. 137. **Q Кокрена (для k выборок)** можно сделать для категориальных полей. При этом применяется критерий для связанных выборок, который проверяет, являются ли равновероятными комбинации значений *k* флаговых полей (категориальных полей только с двумя значениями). Дополнительно можно запросить множественные сравнения *k* выборок, выбрав либо **Все попарно**, либо **Пошагово вниз**. Подробности о задании параметров критериев смотрите в разделе “Критерий Q Кокрена: определить успех” на стр. 137.
- **Проверить наличие изменений в полиномиальных данных. Критерий маргинальной однородности (для 2 выборок)** позволяет применить критерий для связанных выборок, который проверяет, являются ли равновероятными комбинации значений двух парных порядковых полей. Критерий маргинальной однородности обычно применяется при наличии повторных измерений. Этот критерий обобщает критерий Макнемара для двоичных откликов на случай полиномиальных откликов. Если на вкладке Поля задано более двух полей, этот критерий не применяется.
- **Сравнить медианную разность с гипотетической.** Каждый из этих критериев проверяет, отлична ли от 0 медиана разностей между двумя полями. Этот критерий применяется к непрерывным и порядковым полям. Если на вкладке Поля задано более двух полей, эти критерии не применяются.

- **Оценить доверительный интервал.** Здесь можно запросить оценку и доверительный интервал для медианы разностей двух парных полей. Этот критерий применяется к непрерывным и порядковым полям. Если на вкладке Поля задано более двух полей, этот критерий не применяется.
- **Количественно измерить связи.** Выбор **Коэффициент согласия Кендалла (для k выборок)** позволяет к вычислить меру согласия мнений экспертов или респондентов, и каждая запись содержит мнения одного опрашиваемого по нескольким пунктам (занимающим несколько полей). Дополнительно можно запросить множественные сравнения *k* выборок, выбрав либо **Все попарно**, либо **Пошагово вниз**.
- **Сравнить распределения. Двухфакторный дисперсионный анализ Фридмана по рангам (для k выборок)** позволяет применить критерий, который проверяет, извлечены ли *k* связанных выборок из одной генеральной совокупности. Дополнительно можно запросить множественные сравнения *k* выборок, выбрав либо **Все попарно**, либо **Пошагово вниз**.

Критерий Макнемара: определить успех: Критерий Макнемара предназначен для флаговых полей (категориальных полей только с двумя категориями), однако он применяется ко всем категориальным полям, используя правило задания "успеха".

Задать "успех" для категориальных полей. Здесь задается, что является "успехом" для категориальных полей.

- Выбор **Первое значение, встретившееся в данных** приведет к тому, что в качестве "успеха" в критерии будет использоваться первое значение, обнаруженное в выборке. Этот выбор применим только к номинальным и порядковым полям и только с двумя категориями. Все остальные категориальные поля, заданные на вкладке Поля, проверяться не будут. Это задано по умолчанию.
- Выбор **Объединить значения в категорию "успеха"** приведет к тому, что в качестве "успеха" в критерии будут использоваться все значения из заданного списка. Задайте список текстовых или числовых значений. Значения из этого списка необязательно должны присутствовать в выборке.

Критерий Q Кокрена: определить успех: Критерий Q Кокрена предназначен для флаговых полей (категориальных полей только с двумя категориями), однако он применяется ко всем категориальным полям, используя правило задания "успеха".

Задать "успех" для категориальных полей. Здесь задается, что является "успехом" для категориальных полей.

- Выбор **Первое значение, встретившееся в данных** приведет к тому, что в качестве "успеха" в критерии будет использоваться первое значение, обнаруженное в выборке. Этот выбор применим только к номинальным и порядковым полям и только с двумя категориями. Все остальные категориальные поля, заданные на вкладке Поля, проверяться не будут. Это вариант по умолчанию.
- Выбор **Объединить значения в категорию "успеха"** приведет к тому, что в качестве "успеха" в критерии будут использоваться все значения из заданного списка. Задайте список текстовых или числовых значений. Значения из этого списка необязательно должны присутствовать в выборке.

Параметры критериев

Уровень значимости. Здесь задается уровень значимости (альфа) для всех критериев. Задайте числовое значение между 0 и 1. 0,05 является значением по умолчанию.

Доверительный интервал (%). Здесь задается доверительный уровень для всех рассчитываемых доверительных интервалов. Укажите числовое значение от 0 до 100. Значение по умолчанию - 95.

Исключенные наблюдения. Здесь задается, какие наблюдения используются при выполнении тестов.

- **Исключать наблюдения целиком** означает, что записи с пропущенными значениями в любых полях, указанных в любой подкоманде, исключаются из анализа.
- **Исключать по отдельности** означает, что записи с пропущенными значениями в поле, используемом при выполнении конкретного теста, не используются при выполнении этого теста. Когда задано одновременно несколько тестов, для каждого из них вопрос об использовании записей с пропущенными значениями решается независимо от других.

Пользовательские значения отсутствия

Пользовательские пропущенные значения для категориальных полей. Категориальные поля должны иметь допустимые значения, для того чтобы запись была включена в анализ. С помощью этих управляющих элементов можно определить, рассматривать ли пользовательские пропущенные значения в категориальных полях как допустимые. Системные пропущенные значения и пропущенные значения для количественных полей всегда рассматриваются как недопустимые.

Команда NPTESTS: дополнительные возможности

Язык синтаксиса команд также позволяет:

- Задать применение одновыборочного критерия, а также критериев для независимых и связанных выборок, запуская процедуру один раз.

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Средство просмотра моделей

Представление модели

Данная процедура создает объект для средства просмотра моделей в средстве просмотра. Активация (двойным щелчком) этого объекта позволяет рассматривать модель в интерактивном режиме. Представление модели состоит из двух панелей: основного представления слева и связанного с ним вспомогательного представления справа.

Имеется два основных представления:

- Сводка по проверке гипотез. Это представление по умолчанию. Дополнительную информацию смотрите в разделе “Сводка по проверке гипотез”.
- Сводка по доверительным интервалам. Дополнительную информацию смотрите в разделе “Сводка по доверительным интервалам” на стр. 139.

Имеется семь связанных/вспомогательных представлений:

- Одновыборочный критерий. Если запрошены одновыборочные критерии, то это представление показывается по умолчанию. Дополнительную информацию смотрите в разделе “Одновыборочный критерий” на стр. 139.
- Критерий для связанных выборок. Если запрошены критерии для связанных выборок и не запрошены одновыборочные критерии, то это представление показывается по умолчанию. Дополнительную информацию смотрите в разделе “Критерии для связанных выборок” на стр. 140.
- Критерий для независимых выборок. Если не запрошены критерии для связанных выборок или одновыборочные критерии, то это представление показывается по умолчанию. Дополнительную информацию смотрите в разделе “Критерий для независимых выборок” на стр. 141.
- Информация по категориальным полям. Дополнительную информацию смотрите в разделе “Информация по категориальным полям” на стр. 142.
- Информация по количественным полям. Дополнительную информацию смотрите в разделе “Информация по количественным полям” на стр. 142.
- Парные сравнения. Дополнительную информацию смотрите в разделе “Парные сравнения” на стр. 142.
- Однородные подмножества. Дополнительную информацию смотрите в разделе “Однородные подмножества” на стр. 143.

Сводка по проверке гипотез

Представление Сводка по модели- это мгновенная визуальная сводка по результатам применения непараметрических критериев. На ней внимание акцентируется на нулевых гипотезах и выводах, а также значимых *p*-значениях.

- Каждая строка соответствует отдельному тесту. Щелкнув по строке, можно получить дополнительную информацию о результатах теста на панели связанного представления.
- Щелкнув по заголовку любого столбца, можно отсортировать строки по значениям данного столбца.
- Кнопка **Сброс** позволяет вернуть средство просмотра моделей в исходное состояние.
- Раскрывающийся список **Фильтр полей** позволяет вывести результаты только тех тестов, в которые включены выбранные поля.

Сводка по доверительным интервалам

Сводка по доверительным интервалам выводит все доверительные интервалы, сформированные процедурами непараметрических критериев.

- Каждая строка соответствует отдельному доверительному интервалу.
- Щелкнув по заголовку любого столбца, можно отсортировать строки по значениям данного столбца.

Одновыборочный критерий

Представление Одновыборочный критерий отображает детальную информацию обо всех запрошенных одновыборочных непараметрических критериях. Эта информация зависит от того, какие критерии выбраны.

- Раскрывающийся список **Критерий** позволяет выбрать нужный тип одновыборочного критерия.
- Раскрывающийся список **Поля** позволяет выбрать поле, для которого был выполнен тест с помощью критерия, выбранного в раскрывающемся списке **Критерий** .

Биномиальный критерий

Для биномиального критерия выводится составная столбчатая диаграмма и таблица результатов теста.

- На составной столбчатой диаграмме выводятся наблюденные и гипотетические частоты для категорий "успеха" и "неуспеха" проверяемых полей, причем "неуспехи" пристыкованы к "успехам" сверху. Наведение указателя мыши на столбец приведет к выводу в контекстной строке процента для данной категории. Видимые различия размеров столбцов указывают на то, что распределение проверяемого поля может не соответствовать гипотетическому биномиальному распределению.
- В таблице показаны подробные результаты теста.

Критерий хи-квадрат

Представление Критерий хи-квадрат выводит кластеризованную столбчатую диаграмму и таблицу результатов теста.

- На кластеризованной столбчатой диаграмме выводятся наблюденные и гипотетические частоты для каждой категории проверяемого поля. Наведение указателя мыши на столбец приведет к выводу в контекстной строке наблюденной и гипотетической частот, а также их разности (остатка). Видимые различия размеров наблюденных и гипотетических столбцов указывают на то, что распределение проверяемого поля может не соответствовать гипотетическому.
- В таблице показаны подробные результаты теста.

Знаковых рангов Уилкоксона

Представление Критерий знаковых рангов Уилкоксона выводит гистограмму и таблицу результатов теста.

- Гистограмма содержит вертикальные линии, которые показывают наблюденные и гипотетические медианы.
- В таблице показаны подробные результаты теста.

Критерий серий

Представление Критерий серий выводит диаграмму и таблицу результатов теста.

- На диаграмме выводится нормальное распределение с наблюденным числом серий, отмеченным вертикальной линией. Обратите внимание на то, что при применении точного критерия соответствующий тест не основывается на нормальном распределении.
- В таблице показаны подробные результаты теста.

Критерий Колмогорова-Смирнова

Представление Критерий Колмогорова-Смирнова выводит гистограмму и таблицу результатов теста.

- Гистограмма включает наложение функции плотности вероятностей для гипотетического, равномерного, нормального, экспоненциального распределений или распределения Пуассона. Обратите внимание на то, что тест основывается на (накопленных) функциях распределения, и представленные в таблице Наиболее экстремальные различия нужно интерпретировать в терминах (накопленных) функций распределения.
- В таблице показаны подробные результаты теста.

Критерии для связанных выборок

Представление Одновыборочный критерий показывает детальную информацию обо всех запрошенных одновыборочных непараметрических критериях. Эта информация зависит от того, какие критерии выбраны.

- Раскрывающийся список **Критерий** позволяет выбрать нужный тип одновыборочного критерия.
- Раскрывающийся список **Поля** позволяет выбрать поле, для которого был выполнен тест с помощью критерия, выбранного в раскрывающемся списке **Критерий** .

Критерий Макнемара

Представление Критерий Макнемара выводит кластеризованную столбчатую диаграмму и таблицу результатов теста.

- На кластеризованной столбчатой диаграмме выводятся наблюденные и гипотетические частоты для недиагональных ячеек таблицы 2×2, определяемой проверяемыми полями.
- Таблица выводит детальную информацию о результатах теста.

Критерий знаков

Представление Критерий знаков выводит составную гистограмму и таблицу результатов теста.

- На составной гистограмме выводятся различия между полями с использованием знака разности в качестве стыкующего поля.
- Таблица выводит детальную информацию о результатах теста.

Критерий знаковых рангов Уилкоксона

Представление Критерий знаковых рангов Уилкоксона выводит составную гистограмму и таблицу результатов теста.

- На составной гистограмме выводятся различия между полями с использованием знака разности в качестве стыкующего поля.
- Таблица выводит детальную информацию о результатах теста.

Критерий маргинальной однородности

Представление Критерий маргинальной однородности выводит кластеризованную столбчатую диаграмму и таблицу результатов теста.

- На кластеризованной столбчатой диаграмме выводятся наблюденные частоты для недиагональных ячеек таблицы, определяемой проверяемыми полями.
- Таблица выводит детальную информацию о результатах теста.

Критерий Q Кокрена

Представление Критерий Q Кокрена выводит составную столбчатую диаграмму и таблицу результатов теста.

- На составной столбчатой диаграмме выводятся наблюденные частоты для категорий "успеха" и "неуспеха" проверяемых полей, причем "неуспехи" пристыкованы к "успехам" сверху. Наведение указателя мыши на столбец приведет к выводу в контекстной строке процента для данной категории.
- Таблица выводит детальную информацию о результатах теста.

Двухфакторный дисперсионный анализ Фридмана по рангам

Представление Двухфакторный дисперсионный анализ Фридмана по рангам выводит гистограммы с панелями и таблицу результатов теста.

- На гистограммах выводятся наблюденные распределения рангов, разбитые на панели по проверяемым полям.
- Таблица выводит детальную информацию о результатах теста.

Коэффициент согласия Кендалла

Представление Коэффициент согласия Кендалла выводит гистограммы с панелями и таблицу результатов теста.

- На гистограммах выводятся наблюденные распределения рангов, разбитые на панели по проверяемым полям.
- Таблица выводит детальную информацию о результатах теста.

Критерий для независимых выборок

Представление Критерий для независимых выборок отображает детальную информацию обо всех запрошенных непараметрических критериях для независимых выборок. Эта информация зависит от того, какие критерии выбраны.

- Раскрывающийся список **Критерий** позволяет выбрать нужный тип критерия для независимых выборок.
- Раскрывающийся список **Поля** позволяет выбрать комбинацию критерия и группирующего поля, для которой был выполнен тест с помощью критерия, выбранного в раскрывающемся списке **Критерий** .

Критерий Манна-Уитни

Представление Критерия Манна-Уитни выводит диаграмму пирамиды населения и таблицу результатов теста.

- На диаграмме пирамиды населения последовательно по категориям группирующего поля выводятся гистограммы с указанием числа записей в каждой группе и среднего ранга для группы.
- В таблице показаны подробные результаты теста.

Критерий Колмогорова-Смирнова

Представление Критерий Колмогорова-Смирнова выводит диаграмму пирамиды населения и таблицу результатов теста.

- На диаграмме пирамиды населения последовательно по категориям группирующего поля выводятся гистограммы с указанием числа записей в каждой группе. Линии эмпирической функции распределения могут быть выведены или скрыты щелчком по кнопке **Cumulative** .
- В таблице показаны подробные результаты теста.

Критерий серий Вальда-Вольфовица

Представление Критерий серий Вальда-Вольфовица выводит составную столбчатую диаграмму и таблицу результатов теста.

- На диаграмме пирамиды населения последовательно по категориям группирующего поля выводятся гистограммы с указанием числа записей в каждой группе.
- В таблице показаны подробные результаты теста.

Критерий Краскала-Уоллиса

Представление Критерий Краскала-Уоллиса выводит ящичные диаграммы и таблицу результатов теста.

- Для каждой категории группирующего поля выводится отдельная ящичная диаграмма. Наведение указателя мыши на ящик приведет к выводу в контекстной строке среднего ранга.
- В таблице показаны подробные результаты теста.

Критерий Джонкхира-Терпстры

Представление Критерий Джонкхира-Терпстры выводит ящичные диаграммы и таблицу результатов теста.

- Для каждой категории группирующего поля выводится отдельная ящичная диаграмма.
- В таблице показаны подробные результаты теста.

Критерий экстремальной реакции Мозеса

Представление Критерий экстремальной реакции Мозеса выводит ящичные диаграммы и таблицу результатов теста.

- Для каждой категории группирующего поля выводится отдельная ящичная диаграмма. Метки точек могут быть выведены или скрыты щелчком по кнопке **ID записи**.
- В таблице показаны подробные результаты теста.

Медианный критерий

Представление Медианный критерий выводит ящичные диаграммы и таблицу результатов теста.

- Для каждой категории группирующего поля выводится отдельная ящичная диаграмма.
- В таблице показаны подробные результаты теста.

Информация по категориальным полям

Представление Информация по категориальным полям выводит столбчатую диаграмму для категориального поля, выбранного в раскрывающемся списке **Поля** . Список доступных полей ограничен категориальными полями, использованными тестом, выбранным в качестве текущего в представлении Сводка по проверке гипотез.

- Наведение указателя мыши на столбец приведет к выводу в контекстной строке процента для данной категории.

Информация по количественным полям

Представление Информация по количественным полям выводит гистограмму для количественного поля, выбранного в раскрывающемся списке **Поля** . Список доступных полей ограничен количественными полями, использованными тестом, выбранным в качестве текущего в представлении Сводка по проверке гипотез.

Парные сравнения

Представление Парные сравнения выводит сетевой график расстояний и таблицу сравнений, которые формируются процедурами *k*-выборочных непараметрических критериев в случае, если запрашиваются парные множественные сравнения.

- Сетевая диаграмма расстояний является графическим представлением таблицы сравнений, в котором расстояния между узлами сети соответствуют различиям между выборками. Желтые линии соответствуют статистически значимым различиям; черные линии соответствуют незначимым различиям. Наведение указателя мыши на линию в сети приведет к выводу контекстной строки со скорректированным значением значимости различия между узлами, соединенными данной линией.

- Таблица сравнений выводит численные результаты всех парных сравнений. Каждая строка соответствует отдельному парному сравнению. Щелкнув по заголовку столбца, можно отсортировать строки по значениям данного столбца.

Однородные подмножества

Представление Однородные подмножества выводит таблицу сравнений, которая формируется процедурами *k*-выборочных непараметрических критериев в случае, когда запрашиваются пошаговые нисходящие множественные сравнения.

- Каждая строка в группе выборки соответствует отдельной связанной выборке (представленной в данных отдельным полем). Выборки, которые статистически значимо не различаются, объединяются в подмножества, элементы которых выделяются одним цветом. Для каждого выявленного подмножества имеется отдельный столбец. Если все выборки статистически значимо различаются, то каждой выборка представляет собой отдельное подмножество. Если ни одна из выборок статистически значимо не отличается от остальных, то имеется единственное подмножество.
- Для каждого подмножества, содержащего более одной выборки, вычисляются статистика критерия, значение значимости и скорректированное значение значимости.

Команда NPTESTS: дополнительные возможности

Язык синтаксиса команд также позволяет:

- Задать применение одновыборочного критерия, а также критериев для независимых и связанных выборок, запуская процедуру один раз.

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Устаревшие диалоговые окна

Имеется несколько "устаревших" диалоговых окон, которые также позволяют применить непараметрические критерии. Эти диалоговые окна поддерживают функциональные возможности, предоставляемые модулем Exact Tests.

Критерий хи-квадрат. Табулирует переменную по категориям и рассчитывает статистику хи-квадрат, основываясь на разностях между наблюденными и ожидаемыми частотами.

Биномиальный критерий. Сравнивает наблюденную частоту для каждой категории дихотомической переменной с ожидаемыми частотами для данного биномиального распределения.

Критерий серий. Проверяет, является ли случайным порядок появления двух значений переменной.

Одновыборочный критерий Колмогорова-Смирнова. Сравнивает эмпирическую функцию распределения переменной с заданным теоретическим распределением, которое может быть нормальным, равномерным, экспоненциальным или пуассоновским.

Критерии для двух независимых выборок. Сравнивают две группы наблюдений для одной переменной. Доступны следующие критерии: *U* критерий Манна-Уитни, двухвыборочный критерий Колмогорова-Смирнова, критерий экстремальных реакций Мозеса и критерий серий Вальда-Вольфовица.

Критерии для двух связанных выборок. Сравнивают распределения двух переменных. Доступны следующие критерии: критерий знаковых рангов Уилкоксона, критерий знаков и критерий Макнемара.

Критерии для нескольких независимых выборок. Сравнивают две или большее число групп наблюдений для одной переменной. Доступны следующие критерии: критерий Краскала-Уоллиса, медианный критерий, критерий Джонкхира-Терпстры.

Критерии для нескольких связанных выборок. Сравнивает распределения двух или большего числа переменных. Доступны следующие критерии: критерий Фридмана, критерий *W* Кендалла и критерий *Q* Кокрена.

Для всех вышеперечисленных критериев предусмотрена возможность вывода квартилей, средних значений, стандартных отклонений, минимумов, максимумов и числа непропущенных наблюдений.

Критерий хи-квадрат

Процедура Критерий хи-квадрат табулирует переменную по категориям и рассчитывает статистику хи-квадрат. Данный критерий согласия сравнивает наблюденные и ожидаемые частоты в каждой категории, чтобы проверить, что либо все категории содержат одинаковые доли значений, либо каждая категория содержит заданную пользователем долю значений.

Примеры. Критерий хи-квадрат можно использовать для проверки того, равны ли доли синих, коричневых, зеленых, оранжевых, красных и желтых конфет в пакете. Также можно проверить, содержится ли в этом пакете 5% синих, 30% коричневых, 10% зеленых, 20% оранжевых, 15% красных и 15% желтых конфет.

Статистика. Среднее значение, стандартное отклонение, минимум, максимум и квартили. Количество и процент непропущенных и пропущенных наблюдений, количество наблюдаемых и ожидаемых наблюдений для каждой категории, остатки и статистика хи-квадрат.

Данные для критерия хи-квадрат

Данные. Используйте упорядоченные или неупорядоченные числовые категориальные переменные (порядковые или номинальные). Для преобразования текстовых переменных в числовые используйте процедуру Автоматическая перекодировка, вызываемую в меню Преобразовать.

Допущения. Непараметрические критерии не требуют выполнения каких-либо предположений относительно формы распределения, из которого взяты данные. Предполагается, что данные являются случайной выборкой. Ожидаемые частоты для каждой категории должны быть не меньше 1. Не более 20% категорий могут иметь ожидаемые частоты, меньшие 5.

Как запустить процедуру Непараметрический критерий хи-квадрат

1. Выберите в меню:

   Анализ > Непараметрические критерии > Устаревшие диалоговые окна > Хи-квадрат...
2. Выберите одну или несколько переменных для проверки. Для каждой переменной критерий будет рассчитываться отдельно.
3. По желанию можно щелкнуть по кнопке **Параметры** , чтобы задать вывод описательных статистик и квартилей, а также параметры обработки пропущенных данных.

Ожидаемый диапазон и ожидаемые значения для непараметрического критерия хи-квадрат

Ожидаемый диапазон. По умолчанию, каждое встречающееся значение переменной задает категорию. Чтобы использовать категории только из заданного диапазона, выберите вариант **Использовать указанный диапазон** и введите целочисленные значения для верхней и нижней границ диапазона. Категориями будут все целочисленные значения в этом диапазоне, включая границы, а наблюдения со значениями вне диапазона будут исключены из анализа. Например, если в качестве нижней границы задана 1, а в качестве верхней - 4, для критерия хи-квадрат будут использоваться только целочисленные значения от 1 до 4.

Ожидаемые значения. По умолчанию ожидаемые значения для всех категорий равны между собой. Категории могут также иметь задаваемые пользователем ожидаемые доли. Выберите вариант **Значения** и для каждой категории проверяемой переменной введите значение большее 0 и щелкните по **Добавить** . Каждый раз, когда вы добавляете значение, оно появляется внизу списка. Порядок значений существен; он соответствует возрастающему порядку значений категорий проверяемой переменной. Первое значение в списке соответствует наименьшему значению проверяемой переменной, а последнее значение - наибольшему.

Значения в списке суммируются, затем каждое значение делится на эту сумму. В результате для каждой категории получается доля ожидаемых в ней наблюдений. Например, список значений 3, 4, 5, 4 задает следующие ожидаемые доли: 3/16, 4/16, 5/16 и 4/16.

Параметры процедуры Непараметрический критерий хи-квадрат

Статистики. Можно выбрать один или оба параметра вывода итожащих статистик.

- **Описательные.** Вывод среднего значения, стандартного отклонения, минимума, максимума и количества непропущенных наблюдений.
- **Квартили.** Значения 25-го, 50-го и 75-го процентилей.

Пропущенные значения. Эта группа параметров позволяет управлять обработкой пропущенных значений.

- **Исключать по отдельности.** Если задан расчет нескольких критериев, то в каждом из них пропущенные значения обрабатываются отдельно.
- **Исключать целиком.** Наблюдения, имеющие пропущенное значения хотя бы в одной участвующей в анализе переменной, исключаются из всех расчетов.

Команда NPAR TESTS: дополнительные возможности (при расчете критерия хи-квадрат)

Язык синтаксиса команд также позволяет:

- Задавать различные минимальные и максимальные значения или ожидаемые частоты для разных переменных (подкоманда `CHISQUARE`).
- Проверять одну и ту же переменную для разных ожидаемых частот или использовать разные диапазоны. (подкоманда `EXPECTED`).

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Биномиальный критерий

Процедура Биномиальный критерий сравнивает наблюденные частоты для двух категорий дихотомической переменной с частотами, ожидаемыми для биномиального распределения с заданным значением параметра вероятности. По умолчанию значение параметра вероятности для обеих групп равно 0.5. Чтобы изменить эти вероятности, можно ввести значение проверяемой доли для первой группы. Значение вероятности для второй группы будет равно 1 минус заданное значение вероятности для первой группы.

Пример. При бросании монетки вероятность выпадения орла равна 1/2. Исходя из этой гипотезы, монетка подбрасывается 40 раз, и результаты бросания (орел/решетка) записываются. С помощью биномиального критерия получаем, что при выпадении орла для 3/4 подбрасываний наблюдаемый уровень значимости мал (0.0027). Это означает, что вряд ли вероятность выпадения орла равна 1/2; по всей видимости, монета несколько асимметрична.

Статистика. Среднее значение, стандартное отклонение, минимум, максимум, количество непропущенных наблюдений и квартили.

Данные для биномиального критерия

Данные. Проверяемые переменные должны быть числовыми и дихотомическими. Для преобразования текстовых переменных в числовые используйте процедуру Автоматическая перекодировка, вызываемую в меню Преобразовать. **Дихотомическая переменная** - это переменная, которая может принимать только два возможных значения: *да* или *нет*, *истина* или *ложь*, 0 или 1 и так далее. Первое встреченное значение в наборе данных определяет первую группу, а остальные значения определяют вторую группу. Если переменные не дихотомические, необходимо задать пороговое значение. Наблюдения со значениями, равными или меньшими порогового, попадают в одну группу, а остальные наблюдения - в другую группу.

Допущения. Непараметрические критерии не требуют выполнения каких-либо предположений относительно формы распределения, из которого взяты данные. Предполагается, что данные являются случайной выборкой.

Как запустить процедуру Биномиальный критерий

1. Выберите в меню:

   Анализ > Непараметрические критерии > Устаревшие диалоговые окна > Биномиальный...
2. Выберите одну или несколько числовых переменных для проверки.
3. По желанию можно щелкнуть по кнопке **Параметры** , чтобы задать вывод описательных статистик и квартилей, а также параметры обработки пропущенных данных.

Параметры процедуры Биномиальный критерий

Статистики. Можно выбрать один или оба параметра вывода итожащих статистик.

- **Описательные.** Вывод среднего значения, стандартного отклонения, минимума, максимума и количества непропущенных наблюдений.
- **Квартили.** Значения 25-го, 50-го и 75-го процентилей.

Пропущенные значения. Эта группа параметров позволяет управлять обработкой пропущенных значений.

- **Исключать по отдельности.** Если задан расчет нескольких критериев, то в каждом из них пропущенные значения обрабатываются отдельно.
- **Исключать целиком.** Наблюдения с пропущенными значениями для какой-либо проверяемой переменной исключаются из всех вычислений.

Команда NPAR TESTS: дополнительные возможности (при вычислении биномиального критерия)

Язык синтаксиса команд также позволяет:

- Выбирать отдельные группы значений (исключая остальные), если у переменной имеется более двух категорий (подкоманда `BINOMIAL`).
- Задавать различные пороговые значения или вероятности для разных переменных (подкоманда `BINOMIAL`).
- Проверять одну и ту же переменную для различных пороговых значений или вероятностей (подкоманда `EXPECTED`).

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Критерий серий

Процедура Критерий серий позволяет проверить, является ли случайным порядок появления двух значений переменной. Серия - это последовательность похожих наблюдений. Если в выборке либо слишком много серий, либо слишком мало, то эта выборка не является случайной.

Примеры. Предположим, что мы отобрали 20 человек, чтобы выяснить, собираются ли они приобрести некоторый товар. Если все 20 человек окажутся одного пола, случайность этой выборки довольно сомнительна. Критерий серий можно использовать для того, чтобы выяснить, является ли выборка случайной.

Статистика. Среднее значение, стандартное отклонение, минимум, максимум, количество непропущенных наблюдений и квартили.

Данные для критерия серий

Данные. Переменные должны быть числовыми. Для преобразования текстовых переменных в числовые используйте процедуру Автоматическая перекодировка, вызываемую в меню Преобразовать.

Допущения. Непараметрические критерии не требуют выполнения каких-либо предположений относительно формы распределения, из которого взяты данные. Используйте выборки из непрерывных вероятностных распределений.

Как запустить процедуру Критерий серий

1. Выберите в меню:

   Анализ > Непараметрические критерии > Устаревшие диалоговые окна > Серии...
2. Выберите одну или несколько числовых переменных для проверки.
3. По желанию можно щелкнуть по кнопке **Параметры** , чтобы задать вывод описательных статистик и квартилей, а также параметры обработки пропущенных данных.

Пороговое значение для процедуры Критерий серий

Пороговое значение. Задает пороговое значение для разбиения на две части (дихотомизации) значений выбранных переменных. В качестве порогового значения можно использовать наблюденное среднее значение или моду, либо можно задать пороговое значение. Наблюдения со значениями, меньшими порогового, попадут в одну группу, а наблюдения со значениями, большими или равными пороговому, попадут в другую группу. Для каждого заданного порогового значения рассчитывается отдельный критерий.

Параметры критерия серий

Статистики. Можно выбрать один или оба параметра вывода итожащих статистик.

- **Описательные.** Вывод среднего значения, стандартного отклонения, минимума, максимума и количества непропущенных наблюдений.
- **Квартили.** Значения 25-го, 50-го и 75-го процентилей.

Пропущенные значения. Эта группа параметров позволяет управлять обработкой пропущенных значений.

- **Исключать по отдельности.** Если задан расчет нескольких критериев, то в каждом из них пропущенные значения обрабатываются отдельно.
- **Исключать целиком.** Наблюдения, имеющие пропущенное значения хотя бы в одной участвующей в анализе переменной, исключаются из всех расчетов.

Команда NPAR TESTS: дополнительные возможности (при расчете критерия серий)

Язык синтаксиса команд также позволяет:

- Задавать различные пороговые значения для разных переменных (подкоманда RUNS).
- Рассчитать критерии для одной и ту же переменной, но для разных пороговых значений (подкоманда RUNS).

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Одновыборочный критерий Колмогорова-Смирнова

Процедура Одновыборочный критерий Колмогорова-Смирнова сравнивает эмпирическую функцию распределения переменной с заданным теоретическим распределением, которое может быть нормальным, равномерным, пуассоновским или экспоненциальным. Статистика *Z* Колмогорова-Смирнова вычисляется как максимум модуля разности между эмпирической и теоретической функциями распределения. Эта статистика критерия согласия используется для проверки гипотезы о том, что наблюдения взяты из указанного распределения.

Пример. Многие параметрические критерии требуют, чтобы переменные были распределены нормально. Одновыборочный критерий Колмогорова-Смирнова можно использовать для проверки гипотезы о том, что переменная (например, *доход*) имеет нормальное распределение.

Статистика. Среднее значение, стандартное отклонение, минимум, максимум, количество непропущенных наблюдений и квартили.

Данные для одновыборочного критерия Колмогорова-Смирнова

Данные. Используйте количественные переменные (измеренные в интервальной шкале или шкале отношений).

Допущения. При использовании критерия Колмогорова-Смирнова предполагается, что параметры проверяемого распределения заданы заранее. В данной процедуре эти параметры оцениваются по выборке. Выборочные среднее значение и стандартное отклонение используются в качестве параметров для нормального распределения, выборочные минимум и максимум задают диапазон равномерного распределения, наконец, выборочное среднее используется как параметр для пуассоновского и экспоненциального распределений. Способность критерия определить отклонение от предполагаемого распределения может быть значительно снижена. Для проверки нормального распределения с оцененными параметрами рассмотрите модифицированный критерий Колмогорова-Смирнова - критерий Лильефорса (доступен в процедуре Исследовать).

Как запустить одновыборочный критерий Колмогорова-Смирнова

1. Выберите в меню:

   Анализ > Непараметрические критерии > Устаревшие диалоговые окна > Одновыборочный К-С...
2. Выберите одну или несколько числовых переменных для проверки. Для каждой переменной критерий будет рассчитываться отдельно.
3. По желанию можно щелкнуть по кнопке **Параметры** , чтобы задать вывод описательных статистик и квартилей, а также параметры обработки пропущенных данных.

Параметры процедуры Одновыборочный критерий Колмогорова-Смирнова

Статистики. Можно выбрать один или оба параметра вывода итожащих статистик.

- **Описательные.** Вывод среднего значения, стандартного отклонения, минимума, максимума и количества непропущенных наблюдений.
- **Квартили.** Значения 25-го, 50-го и 75-го процентилей.

Пропущенные значения. Эта группа параметров позволяет управлять обработкой пропущенных значений.

- **Исключать по отдельности.** Если задан расчет нескольких критериев, то в каждом из них пропущенные значения обрабатываются отдельно.
- **Исключать целиком.** Наблюдения, имеющие пропущенное значения хотя бы в одной участвующей в анализе переменной, исключаются из всех расчетов.

Команда NPAR TESTS: дополнительные возможности (при вычислении одновыборочного критерия Колмогорова-Смирнова)

Язык командного синтаксиса также позволяет задавать параметры распределения критериев (с помощью подкоманды `K-S`).

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Критерии для двух независимых выборок

Процедура Критерии для двух независимых выборок сравнивает две группы наблюдений одной переменной.

Пример. Разработана новая разновидность зубных пластинок, которые, по замыслу их создателей, должны быть более удобными, лучше выглядеть и быстрее выравнивать зубы. Чтобы понять, необходимо ли носить новые зубные пластинки также долго, как и старые зубные пластинки, для ношения новых зубных пластинок были случайно отобраны 10 детей. Применив *U*-критерий Манна-Уитни можно обнаружить, что в среднем детям, носившим новые пластинки, не приходилось носить их так же долго, как и детям, носившим старые пластинки.

Статистика. Среднее значение, стандартное отклонение, минимум, максимум, количество непропущенных наблюдений и квартили. Критерии: *U*-критерий Манна-Уитни, критерий экстремальных реакций Мозеса, *Z*-критерий Колмогорова-Смирнова, критерий серий Вальда-Вольфовица.

Данные для непараметрических критериев для двух независимых выборок

Данные. Используйте количественные переменные с упорядоченными значениями.

Допущения. Используйте независимые случайные выборки. *U*-критерий Манна-Уитни проверяет равенство двух распределений. Для того, чтобы использовать его для оценки различий между двумя распределениями, необходимо допустить, что распределения имеют одинаковую форму.

Как запустить процедуру Критерии для двух независимых выборок

1. Выберите в меню:

   Анализ > Непараметрические критерии > Устаревшие диалоговые окна > Для двух независимых выборок...
2. Выберите одну или несколько числовых переменных.
3. Выберите группирующую переменную и щелкните мышью по **Задать группы** , чтобы разделить файл на две группы или выборки.

Типы непараметрических критериев для двух независимых выборок

Тип критерия. Для проверки гипотезы о том, что две независимые выборки (группы) взяты из одной и той же генеральной совокупности, можно воспользоваться четырьмя критериями.

U критерий Манна-Уитни - наиболее популярный среди непараметрических критериев для двух независимых выборок. Он эквивалентен критерию ранговых сумм Уилкоксона и критерию Краскала-Уоллеса для двух групп. Критерий Манна-Уитни проверяет гипотезу о том, что две генеральные совокупности, из которых были отобраны выборки, эквивалентны по расположению. Наблюдения из обеих групп объединяются и ранжируются, причем совпадающим значениям назначается средний ранг. Количество совпадающих значений должно быть мало по сравнению с общим количеством наблюдений. Если проверяемые совокупности эквивалентны по расположению, то ранги должны быть распределены между двумя выборками случайным образом. При расчете критерия подсчитываются число раз, когда значение из группы 1 предшествует значению из группы 2, и число раз, когда значение из группы 2 предшествует значению из группы 1. *U*-статистикой Манна-Уитни является меньшее из этих двух чисел. Также отображается статистика ранговой суммы Уилкоксона *W*. *W* представляет собой сумму рангов для группы с меньшим средним рангом, если у групп средние ранги не равны, а если равны то это сумма рангов для группы, указанной последней в диалоговом окне Две независимые выборки: Задать группы.

Критерий Z Колмогорова-Смирнова и **критерий серий Вальда-Вольфовица** носят более общий характер и выявляют различия между распределениями как в расположении, так и в форме. Критерий Колмогорова-Смирнова основан на максимуме модуля разности между эмпирическими функциями распределения для обеих выборок. Если эта разность значимо велика, распределения считаются различными. Критерий серий Вальда-Вольфовица объединяет и ранжирует наблюдения из обеих групп. Если обе выборки взяты из одной генеральной совокупности, то обе группы должны быть разбросаны по проранжированным данным случайным образом.

Критерий экстремальных реакций Мозеса предполагает, что экспериментальная переменная воздействует на некоторые объекты в одном направлении, а на другие объекты в противоположном. Критерий выявляет экстремальные отклики в сравнении с контрольной группой. Он сосредотачивается на диапазоне контрольной группы и является показателем того, сколь сильно экстремальные значения из экспериментальной группы влияют на этот диапазон, когда экспериментальной группа объединена с контрольной группой. Контрольная группа задается значением для группы 1 в диалоговом окне Две независимые выборки: Задать группы. Наблюдения из обеих групп объединяются и ранжируются. Размах контрольной группы вычисляется как разность между рангами наибольшего и наименьшего значений в контрольной группе плюс 1. Поскольку случайные выбросы могут легко исказить величину диапазона, 5% наблюдений с каждого конца контрольной группы автоматически усекаются.

Задание групп в процедуре Критерии для двух независимых выборок

Чтобы разбить файл на две группы или выборки, введите одно целое значение в поле Группа 1, а другое целое значение - в поле Группа 2. Наблюдения со всеми прочими значениями исключаются из анализа.

Параметры процедуры Критерии для двух независимых выборок

Статистики. Можно выбрать один или оба параметра вывода итожащих статистик.

- **Описательные.** Среднее значение, стандартное отклонение, минимум, максимум и количество непропущенных наблюдений.
- **Квартили.** Значения 25-го, 50-го и 75-го процентилей.

Пропущенные значения. Эта группа параметров позволяет управлять обработкой пропущенных значений.

- **Исключать по отдельности.** Если задан расчет нескольких критериев, то в каждом из них пропущенные значения обрабатываются отдельно.
- **Исключать целиком.** Наблюдения, имеющие пропущенное значения хотя бы в одной участвующей в анализе переменной, исключаются из всех расчетов.

Команда NPAR TESTS - дополнительные возможности (Непараметрические критерии для двух независимых выборок)

Синтаксис команды также позволяет задавать количество наблюдений, удаляемых при расчете критерия Мозеса (при помощи подкоманды `MOSES`).

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Критерии для двух связанных выборок

Процедура Критерии для двух связанных выборок сравнивает распределения двух переменных.

Пример. Получают ли обычно семьи запрошенную цену при продаже своих домов? Применив для анализа данных по 10-ти домам критерий знаковых рангов Уилкоксона, можно обнаружить, что семь семей получают меньше запрошенного, одна семья - больше и две семьи - запрошенную цену.

Статистика. Среднее значение, стандартное отклонение, минимум, максимум, количество непропущенных наблюдений и квартили. Критерии: знаковых рангов Уилкоксона, знаков, Макнемара. Если установлен модуль Exact Tests (имеется только для операционных систем Windows), также доступен тест маргинальной неоднородности.

Данные для критериев для двух связанных выборок

Данные. Используйте количественные переменные с упорядоченными значениями.

Допущения. Хотя наличия определенных распределений у двух анализируемых переменных не требуется, теоретическое распределение парных разностей предполагается симметричным.

Как запустить процедуру Критерии для двух связанных выборок

1. Выберите в меню:

   Анализ > Непараметрические критерии > Устаревшие диалоговые окна > Для двух связанных выборок...
2. Выберите одну или несколько пар переменных.

Типы критериев, доступные в процедуре Критерии для двух связанных выборок

Критерии, описываемые в настоящем разделе, сравнивают распределения двух связанных переменных. Применяемый критерий зависит от типа данных.

Если данные являются непрерывными, используйте критерий знаков или критерий знаковых рангов Уилкоксона. **Критерий знаков** рассчитывает разности между двумя переменными для всех наблюдений и

классифицирует их как положительные, отрицательные или совпадения (нулевые). Если обе переменные одинаково распределены, число положительных и отрицательных разностей не будет значимо различным. **Критерий знаковых рангов Уилкоксона** учитывает информацию как о знаке разности между парами, так и о величине этой разности. Поскольку критерий знаковых рангов Уилкоксона использует больше информации о данных, он является более мощным, чем критерий знаков.

Если данные являются бинарными, следует использовать **критерий Макнемара** . Этот критерий, как правило, применяют при наличии повторных измерений, когда реакция (отклик) каждого объекта фиксируется дважды: один раз до, а другой - после наступления некоторого события. При помощи критерия Макнемара определяют, совпадает ли начальный уровень отклика (до события) с итоговым (после события). Этот критерий полезен при выявлении изменений в откликах, вызванных экспериментальным вмешательством, в планах исследований типа "до-и-после".

Если данные являются категориальными, используйте **критерий маргинальной однородности** . Этот критерий обобщает критерий Макнемара для двоичных откликов на случай полиномиальных откликов. Он проверяет наличие изменений в отклике, используя распределение хи-квадрат, и полезен для обнаружения изменений в откликах, вызванных экспериментальным вмешательством, в планах исследований типа "до-и-после". Критерий маргинальной однородности доступен, только если установлен модуль Exact Tests.

Параметры процедуры Критерии для двух связанных выборок

Статистики. Можно выбрать один или оба параметра вывода итожащих статистик.

- **Описательные.** Среднее значение, стандартное отклонение, минимум, максимум и количество непропущенных наблюдений.
- **Квартили.** Значения 25-го, 50-го и 75-го процентилей.

Пропущенные значения. Эта группа параметров позволяет управлять обработкой пропущенных значений.

- **Исключать по отдельности.** Если задан расчет нескольких критериев, то в каждом из них пропущенные значения обрабатываются отдельно.
- **Исключать целиком.** Наблюдения, имеющие пропущенное значения хотя бы в одной участвующей в анализе переменной, исключаются из всех расчетов.

Команда NPAR TESTS: дополнительные возможности (при расчете непараметрических критериев для двух связанных выборок)

Синтаксис команд также позволяет рассчитывать критерии для переменной с каждой из переменных в списке.

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Критерии для нескольких независимых выборок

Процедура Непараметрические критерии для нескольких независимых выборок сравнивает две или большее количество групп наблюдений по одной переменной.

Пример. Существуют ли различия в среднем времени работы между тремя разновидностями электрических ламп мощностью 100 ватт? Выполнив однофакторный дисперсионный анализ Краскела—Уоллиса, мы увидим, что такое различие действительно имеет место.

Статистика. Среднее значение, стандартное отклонение, минимум, максимум, количество непропущенных наблюдений и квартили. Критерии: Краскела—Уоллиса *H*, медианный.

Данные для непараметрических критериев для нескольких независимых выборок

Данные. Используйте количественные переменные с упорядоченными значениями.

Допущения. Используйте независимые случайные выборки. Критерий *H* Краскела—Уоллиса требует, чтобы форма распределений проверяемых выборок были схожими.

Как запустить процедуру Непараметрические критерии для нескольких независимых выборок

1. Выберите в меню:

   Анализ > Непараметрические критерии > Устаревшие диалоговые окна > Для K независимых выборок...

2. Выберите одну или несколько числовых переменных.
3. Выберите группирующую переменную и нажмите кнопку **Задать диапазон** , чтобы указать минимальное и максимальное целые значения для группирующей переменной.

Типы критериев в процедуре Критерии для нескольких независимых выборок

Для проверки гипотезы о том, что несколько независимых выборок взяты из одной и той же генеральной совокупности, можно воспользоваться тремя критериями. Каждый из критериев: критерий *H* Краскела—Уоллиса, медианный критерий и критерий Джонкхира-Терпстры проверяют, взяты ли несколько независимых выборок из одной и той же генеральной совокупности.

Критерий H Краскела—Уоллиса , являющийся расширением *критерия U Манна-Уитни* , представляет собой непараметрический аналог однофакторного дисперсионного анализа и используется для выявления различий в расположении распределений выборок. **Медианный критерий** , который является более общим, но не столь мощным критерием, используется для выявления различий между распределениями и в расположении, и в форме. Критерий *H* Краскела—Уоллиса и медианный критерий предполагают, что *k* генеральных совокупностей, из которых взяты выборки, *априори* не упорядочены.

При *наличии* естественной *априорной* упорядоченности (по возрастанию или по убыванию) *k* совокупностей более мощным является **критерий Джонкхира-Терпстры** . Например, *k* совокупностей могут представлять собой *k* возрастающих температур. Проверяется гипотеза о том, что разные температуры дают одинаковое распределение откликов, против альтернативной гипотезы о том, что при увеличении температуры возрастает и величина отклика. Здесь альтернативная гипотеза упорядочена; следовательно, наиболее подходящим будет критерий Джонкхира-Терпстры. Критерий Джонкхира-Терпстры доступен, только если установлена надстройка Exact Tests.

Задание диапазона в процедуре Непараметрические критерии для нескольких независимых выборок

Чтобы задать диапазон, введите целые значения для **Минимума** и **Максимума** , соответствующие наименьшей и наибольшей категориям группирующей переменной. Наблюдения со значениями вне заданного диапазона исключаются из анализа. Например, если заданы минимальное значение, равное 1, и максимальное значение, равное 3, то будут использоваться только целые значения от 1 до 3. Минимальное значение должно быть меньше максимального, и оба значения должны быть заданы.

Параметры процедуры Непараметрические критерии для нескольких независимых выборок

Статистики. Можно выбрать один или оба параметра вывода итожащих статистик.

- **Описательные.** Среднее значение, стандартное отклонение, минимум, максимум и количество непропущенных наблюдений.
- **Квартили.** Значения 25-го, 50-го и 75-го процентилей.

Пропущенные значения. Эта группа параметров позволяет управлять обработкой пропущенных значений.

- **Исключать по отдельности.** Если задан расчет нескольких критериев, то в каждом из них пропущенные значения обрабатываются отдельно.
- **Исключать целиком.** Наблюдения, имеющие пропущенное значения хотя бы в одной участвующей в анализе переменной, исключаются из всех расчетов.

Команда NPAR TESTS: дополнительные возможности (при расчете критериев для нескольких независимых выборок)

Синтаксис языка команд позволяет задавать для медианного критерия значение, отличное от наблюденной медианы (подкоманда `MEDIAN`).

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Критерии для нескольких связанных выборок

Процедура Непараметрические критерии для нескольких связанных выборок позволяет сравнить распределения двух или большего количества переменных.

Пример. Различается ли престиж профессии врача, адвоката, офицера полиции и учителя? Десятерых респондентов попросили расположить эти четыре профессии в порядке возрастания их престижности. Критерий Фридмана показывает, что в общественном мнении престижность этих профессий действительно различна.

Статистика. Среднее значение, стандартное отклонение, минимум, максимум, количество непропущенных наблюдений и квартили. Критерии: Фридмана, *W* Кэндалла и *Q* Кокрена.

Данные для непараметрических критериев для нескольких связанных выборок

Данные. Используйте количественные переменные с упорядоченными значениями.

Допущения. Непараметрические критерии не требуют выполнения каких-либо предположений относительно формы распределения, из которого взяты данные. Используйте зависимые случайные выборки.

Как запустить процедуру Непараметрический критерии для нескольких связанных выборок

1. Выберите в меню:

   Анализ > Непараметрические критерии > Устаревшие диалоговые окна > Для K связанных выборок...
2. Выберите две или большее количество числовых переменных для тестирования.

Типы критериев, используемых в процедуре Непараметрические критерии для нескольких связанных выборок

Чтобы сравнить распределения нескольких связанных выборок, можно воспользоваться тремя критериями.

Критерий Фридмана - это непараметрический эквивалент одновыборочного плана с повторными измерениями или двухфакторного дисперсионного анализа с одним наблюдением на ячейку. Критерия Фридмана проверяют нулевую гипотезу о том, что *k* связанных переменных взяты из одной и той же генеральной совокупности. Для каждого наблюдения *k* переменных ранжируются от 1 до *k*. Статистика критерия основывается на этих рангах.

Критерий **W Кендалла** является нормализацией статистики Фридмана. Критерий *W* Кендалла интерпретируется как коэффициент конкордации (согласованности), который является показателем согласия среди респондентов (экспертов). Каждый наблюдение представляет эксперта, каждая переменная - оцениваемый объект. Для каждой переменной вычисляется сумма рангов. Значение *W* Кендалла изменяется от 0 (нет согласия) до 1 (полное согласие).

Критерий Q Кокрена идентичен критерию Фридмана, но применяется, когда все отклики являются бинарными. Этот критерий является развитием критерия Макнемара для *k* выборок. При помощи критерия *Q* Кокрена проверяют гипотезу о том, что несколько связанных дихотомических переменных имеют одинаковые средние значения. Переменные измеряются на одном и том же объекте или на эквивалентных объектах.

Статистики критериев для нескольких связанных выборок

Можно задать вывод следующих статистик.

- **Описательные.** Среднее значение, стандартное отклонение, минимум, максимум и количество непропущенных наблюдений.
- **Квартили.** Значения 25-го, 50-го и 75-го процентилей.

Команда NPAR TESTS: дополнительные возможности (при расчете критериев для K связанных выборок)

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Глава 28. Анализ множественных ответов

Анализ множественных ответов

Для анализа наборов множественных дихотомий и наборов множественных категорий предназначены две процедуры. Процедура Частоты множественных ответов выводит частотные таблицы. Процедура Таблицы сопряженности множественных ответов выводит двух- и трехмерные таблицы сопряженности. Перед использованием любой из этих процедур необходимо задать анализируемые наборы данных с множественными ответами.

Пример. Описываемый пример иллюстрирует использование модели данных с множественными ответами в маркетинговом исследовании. Приведенные здесь данные являются вымышленными и не должны восприниматься как реальные. Итак, некая авиакомпания собирается провести опрос пассажиров, летящих по определенному маршруту, с целью оценки конкурирующих авиакомпаний. Пусть авиакомпанию "American Airlines" интересует, пользуются ли ее пассажиры услугами других авиакомпаний на маршруте Чикаго-Нью-Йорк, а также относительная важность расписания полетов и качества обслуживания при выборе авиакомпании. Во время посадки на самолет стюардесса вручает каждому пассажиру краткий вопросник. Первый вопрос звучит следующим образом: "Обведите названия всех авиакомпаний из следующего списка, самолетами которых вы летали хотя бы один раз в течение последних шести месяцев: American, United, TWA, USAir, Другие. Этот вопрос является вопросом с множественными ответами, поскольку пассажир может отметить более одного ответа. Ответы на этот вопрос нельзя закодировать непосредственно, поскольку для каждого наблюдения переменная может принимать только одно значение. Чтобы зафиксировать ответы на каждый из вопросов, вам придется использовать несколько переменных. Это можно сделать двумя способами. Первый определить переменную, соответствующую каждому возможному выбору (например, переменные American , United, TWA, USAir и другие). Если пассажир отмечает в вопроснике авиакомпанию United, переменной *united* присваивается значение 1, в противном случае 0. Такой подход к кодированию ответов называют **методом множественных дихотомий** . Ответы можно представить и другим способом с помощью **метода множественных категорий** , при использовании которого оценивается максимальное число возможных ответов на вопрос и вводится такое же число переменных со значениями, указывающими на компанию, услугами которой пользовался пассажир. Внимательно просматривая заполненные вопросники, Вы, возможно, обнаружите, что в течение последних шести месяцев никто из пассажиров не летал по этому маршруту самолетами более чем трех различных авиакомпаний. Далее вы увидите, что благодаря сокращению государственного вмешательства в деятельность авиакомпаний в категории "Другие" были названы 10 авиакомпаний. Используя метод множественных категорий, вы могли бы задать три переменные со значениями 1= *american* , 2= *united* , 3= *twa* , 4= *usair* , 5= *delta* и так далее. Если данный пассажир отмечает авиакомпании American и TWA, то первой переменной присваивается значение 1, второй - значение 3, а третьей - код пропущенного значения. Другой пассажир мог отметить авиакомпании American и Delta. Тогда первой переменной присваивается значение 1, второй - значение 5, а третьей - код пропущенного значения. Если бы в приведенном примере вы пользовались для записи данных методом множественных дихотомий, то в результате получили бы 14 отдельных переменных. Итак, хотя для этого опроса применимы оба метода представления данных, выбор конкретного метода зависит от того, как распределяются ответы.

Задание наборов множественных ответов

Процедура Задать наборы множественных ответов группирует элементарные переменные в наборы множественных дихотомий и множественных категорий, для которых можно затем построить частотные таблицы и таблицы сопряженности. Можно задать до 20 наборов множественных ответов. Каждый набор должен иметь свое имя. Чтобы удалить набор, выделите его в списке наборов множественных ответов и нажмите кнопку **Удалить** . Чтобы изменить набор, выделите его в списке, модифицируйте любые характеристики набора и нажмите кнопку **Изменить** .

Вы можете закодировать элементарные переменные либо как дихотомии, либо как категории. Чтобы использовать дихотомические переменные, установите переключатель в положение **Дихотомии** для создания набора множественных дихотомий. Введите целое число в поле Подсчитываемое значение. Каждая переменная, хотя бы один раз принимающая это значение, становится категорией набора множественных дихотомий. Установите переключатель в положение **Категории** для создания набора множественных категорий, имеющего тот же диапазон значений, что и составляющие его переменные. Введите целые числа для нижней и верхней границ диапазона значений набора множественных категорий. Процедура подсчитывает встречаемость каждого отдельного целого значения в рамках указанного диапазона по всем переменным, составляющим данный набор. Пустые категории в таблицах не приводятся.

Каждому набору множественных ответов необходимо присвоить уникальное имя длиной до 7 символов. Процедура присоединяет спереди к выбранному вами имени знак доллара ($). Следующие зарезервированные имена использовать нельзя: *casenum*, *sysmis*, *jdate*, *date*, *time*, *length* и *width*. Имя набора множественных ответов доступно только в процедурах анализа множественных ответов. Эти имена нельзя использовать в других процедурах. По желанию для набора множественных ответов можно ввести описательную метку. Ее длина не должна превышать 40 символов.

Чтобы задать наборы множественных ответов

1. Выберите в меню:

   Анализ > Множественные ответы > Задать наборы переменных...
2. Выберите две или более переменных.
3. Если переменные являются дихотомическими, укажите подсчитываемое значение. Если переменные закодированы как категории, задайте диапазон категорий.
4. Введите уникальное имя для каждого набора множественных ответов.
5. Нажмите кнопку **Добавить** , чтобы добавить набор множественных ответов к списку заданных наборов.

Частоты для множественных ответов

Процедура Частоты для множественных ответов позволяет построить частотные таблицы для наборов множественных ответов. Сначала вы должны задать один или несколько наборов множественных ответов (смотрите раздел "Задание наборов множественных ответов").

При выводе результатов для наборов множественных дихотомий в качестве названий категорий используются метки, заданные для элементарных переменных группы. Если эти метки не заданы, то в качестве меток используются имена переменных. Для наборов множественных категорий в качестве меток категорий используются метки значений первой переменной в группе. Если категории, пропущенные для первой переменной, присутствуют в других переменных группы, то необходимо задать метку значений для пропущенных категорий.

Пропущенные значения. Наблюдения с пропущенными значениями исключаются отдельно для каждой таблицы. В качестве альтернативы можно выбрать один или оба из следующих пунктов:

- **Исключать наблюдения целиком в дихотомиях.** Из таблицы для набора множественных дихотомий исключаются наблюдения, у которых пропущено значение хотя бы для одной переменной набора. Применяется только к наборам множественных ответов, заданным как наборы дихотомий. По умолчанию наблюдение считается пропущенным для набора множественных дихотомий, если ни одна из входящих в набор переменных не содержит подсчитываемого значения. Наблюдения с пропущенными значениями для некоторых (но не для всех) переменных набора включаются в таблицу, если, по крайней мере, одна переменная набора содержит подсчитываемое значение.
- **Исключать наблюдения целиком в категориях.** Из таблицы для набора множественных категорий исключаются наблюдения, у которых пропущено значение хотя бы для одной переменной. Этот параметр применяется только к наборам множественных ответов, заданным как наборы категорий. По умолчанию наблюдение считается пропущенным для набора множественных категорий, только если ни одна из входящих в набор переменных не принимает значений в заданном диапазоне.

Пример. Любая переменная, созданная для записи ответа на вопрос обследования является элементарной переменной. Чтобы осуществить анализ группы элементарных данных, представляющих множественные ответы, необходимо объединить переменные в один из двух типов наборов множественных ответов: набор множественных дихотомий или набор множественных категорий. Например, если бы в опросе, проводимом некоей авиакомпанией, спрашивалось, самолетами какой из трех авиакомпаний (American, United, TWA) летали респонденты в течение последних шести месяцев, а для ввода данных использовались дихотомические переменные, а также был задан **набор множественных дихотомий** , то каждая из трех переменных вошедших в набор стала бы категорией групповой переменной. Количества и проценты для трех указанных авиакомпаний представлены в одной частотной таблице. Если обнаружится, что ни один из опрошенных не отметил более двух авиакомпаний, то можно сформировать две переменные, каждая из которых имеет три значения (по одному для каждой из авиакомпаний). Если вы задаете **набор множественных категорий** , значения сводятся в таблицу путем сложения вместе одинакового кода по всем элементарным переменным. Результирующий набор значений является таким же, как и для каждой элементарной переменной. Например, 30 ответов United представляют собой сумму 5 ответов United в переменной авиакомпания 1 и 25 ответов United в переменной авиакомпания 2. Количества (количества наблюдений) и проценты для трех указанных авиакомпаний представляются в одной частотной таблице.

Статистика. В частотных таблицах отображаются частоты (количества наблюдений), проценты ответов, проценты наблюдений, число наблюдений без пропущенных значений и число пропущенных наблюдений.

Данные для процедуры Частоты для множественных ответов

Данные. Используйте наборы множественных ответов.

Допущения. Частоты и проценты полезны при описании данных, какому бы распределению они ни соответствовали.

Родственные процедуры. Процедура Задать наборы множественных ответов позволяет вам задать наборы множественных ответов.

Как построить частотные таблицы для наборов множественных ответов

1. Выберите в меню:

   Анализ > Множественные ответы > Частоты...
2. Выберите один или несколько наборов множественных ответов.

Таблицы сопряженности для множественных ответов

Процедура Таблицы сопряженности для множественных ответов осуществляет построение таблиц сопряженности для заданных наборов множественных ответов, элементарных переменных или их комбинации. Вы можете также рассчитать проценты в ячейках, основанные на наблюдениях или ответах, изменить режим обработки пропущенных значений и получить парные таблицы сопряженности. Сначала вы должны задать один или несколько наборов множественных ответов (смотрите раздел "Задание наборов множественных ответов").

При выводе результатов для наборов множественных дихотомий в качестве названий категорий используются метки, заданные для элементарных переменных группы. Если эти метки не заданы, то в качестве меток используются имена переменных. Для наборов множественных категорий в качестве меток категорий используются метки значений первой переменной в группе. Если категории, пропущенные для первой переменной, присутствуют в других переменных группы, то необходимо задать метку значений для пропущенных категорий. Процедура выводит метки категорий для столбцов в три строки, содержащих до 8 символов на строку. Чтобы избежать нежелательной разбивки слов, можно поменять местами элементы столбцов и строк или переопределить метки.

Пример. Эта процедура позволяет строить таблицы сопряженности с другими переменными как для наборов множественных дихотомий, так и для наборов множественных категорий. При проведении опроса

авиапассажиров задаются следующие вопросы: Обведите названия всех авиакомпаний из следующего списка, самолетами которых вы летали хотя бы один раз в течение последних шести месяцев (American, United, TWA). Что важнее при выборе авиакомпании - расписание или качество обслуживания? Выберите только один вариант ответа. После ввода данных в виде дихотомий или множественных категорий и объединения их в набор можно построить таблицу сопряженности предпочтений авиакомпаний и ответа на вопрос, затрагивающий расписание и качество обслуживания.

Статистика. Таблицы сопряженности с частотами в ячейках, строках и столбцах и общим итогом, а также процентами для ячеек, строк, столбцов и таблицы в целом. Проценты для ячеек могут основываться на наблюдениях или ответах.

Данные для процедуры Таблицы сопряженности для множественных ответов

Данные. Используйте наборы множественных ответов или числовые категориальные переменные.

Допущения. Частоты и проценты полезны при описании данных, порожденных любыми распределениями.

Родственные процедуры. Процедура Задать наборы множественных ответов позволяет вам задать наборы множественных ответов.

Как построить таблицы сопряженности для множественных ответов

1. Выберите в меню:

   Анализ > Множественные ответы > Таблицы сопряженности...
2. Выберите одну или несколько числовых переменных или наборов множественных ответов для каждого измерения таблицы сопряженности.
3. Задайте диапазон для каждой элементарной переменной.

По желанию можно построить двумерную таблицу сопряженности для каждой категории управляющей переменной или набора множественных ответов. Выберите один или несколько объектов для списка слоев.

Задание диапазонов переменных в таблицах сопряженности для наборов множественных ответов

Для каждой элементарной переменной в таблице сопряженности должен быть определен диапазон значений. Введите целые минимальное и максимальное значения категорий, которые вы хотите использовать в таблице. Категории, значения которых выходят за указанные границы диапазона, исключаются из анализа. Предполагается, что внутри диапазона значения являются целыми (дробные значения усекаются).

Параметры процедуры Таблицы сопряженности для множественных ответов

Проценты в ячейках. Количества в ячейках выводятся всегда. Вы можете задать вывод процентов по отношению к строкам, столбцам и к итогу по двумерной таблице.

База для расчета процентов. Вы можете вычислять проценты в ячейках по отношению к наблюдениям (или респондентам). Данной возможностью нельзя воспользоваться, если вы выбрали сопоставление переменных по наборам множественных категорий. Вы можете также вычислять проценты в ячейках по отношению к ответам. При использовании наборов множественных дихотомий число ответов равно числу подсчитываемых значений по всем наблюдениям. При использовании множественных категорий число ответов равно числу значений в заданном диапазоне.

Пропущенные значения. Вы можете выбрать один или оба из следующих пунктов:

- **Исключать наблюдения целиком в дихотомиях.** Из таблицы для набора множественных дихотомий исключаются наблюдения, у которых пропущено значение хотя бы для одной переменной набора. Применяется только к наборам множественных ответов, заданным как наборы дихотомий. По

умолчанию наблюдение считается пропущенным для набора множественных дихотомий, если ни одна из входящих в набор переменных не содержит подсчитываемого значения. Наблюдения с пропущенными значениями для некоторых (но не для всех) переменных набора включаются в таблицу, если, по крайней мере, одна переменная набора содержит подсчитываемое значение.

- **Исключать наблюдения целиком в категориях.** Из таблицы для набора множественных категорий исключаются наблюдения, у которых пропущено значение хотя бы для одной переменной. Этот параметр применяется только к наборам множественных ответов, заданным как наборы категорий. По умолчанию наблюдение считается пропущенным для набора множественных категорий, только если ни одна из входящих в набор переменных не принимает значений в заданном диапазоне.

По умолчанию при создании таблицы сопряженности двух наборов множественных категорий процедура соотносит каждую переменную первой группы с каждой переменной второй группы и суммирует частоты (количества наблюдений) в каждой ячейке; поэтому некоторые ответы могут появиться в таблице более одного раза. Вы можете выбрать следующую возможность:

Сопоставить переменные по наборам ответов. Эта возможность сопоставляет первую переменную первой группы с первой переменной второй группы, вторую переменную первой группы - со второй переменной второй группы и так далее. Если вы выберете эту возможность, процедура будет основывать вычисление процентов в ячейках не на респондентах, а на ответах. Объединение в пары невозможно для наборов множественных дихотомий или для элементарных переменных.

Команда MULT RESPONSE: дополнительные возможности

Язык синтаксиса команд также позволяет:

- Создавать таблицы сопряженности, имеющие до пяти измерений (подкоманда `BY`).
- Изменять спецификации формата вывода, включая подавление вывода меток значений (подкоманда `FORMAT`).

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Глава 29. Создание отчетов

Создание отчетов

Основными инструментами изучения и представления данных служат списки наблюдений и описательные статистики. Списки наблюдений можно получить при помощи Редактора данных или процедуры Итоги; частоты и описательные статистики - при помощи процедуры Частоты; групповые статистики - при помощи процедуры Средние. Формат вывода каждой из этих процедур подобран таким образом, чтобы сделать информацию как можно более ясной. Если желательно отобразить информацию в ином формате, процедуры Итоги по строкам и Итоги по столбцам обеспечат необходимый контроль над представлением данных.

Итоги по строкам

Процедура Итоги по строкам позволяет создать отчеты, в которых различные итожащие статистики располагаются по строкам. Возможен также вывод списка наблюдений вместе с итожащими статистиками или без них.

Пример. Компания с сетью магазинов розничной торговли ведет запись информации о служащих, включая размер оклада, продолжительность работы в занимаемой должности, а также магазин и отдел, в котором служащий работает. Вы могли бы создать отчет, содержащий информацию по каждому служащему (список наблюдений), сгруппировав его по магазину и отделу (группирующие переменные), а также включить в него итожащие статистики (например, среднюю зарплату) для каждого магазина, отдела или отдела внутри каждого магазина.

Столбцы данных. В этой группе задается список переменных, для которых вы хотите получить список значений наблюдений или итожащие статистики, а также предоставляется возможность управлять форматом вывода столбцов данных.

Столбцы группировки. Эта группа позволяет задать список необязательных переменных, разбивающих отчет на группы, а также управлять выводом итожащих статистик и форматом вывода группирующих столбцов. При наличии нескольких группирующих переменных, для каждой категории каждой группирующей переменной будет создана отдельная группа внутри категорий предшествующей в списке группирующей переменной. Группирующие переменные должны представлять собой дискретные категориальные переменные, делящие наблюдения на ограниченное число имеющих смысл категорий. Индивидуальные значения каждой группирующей переменной выводятся в отсортированном виде в отдельном столбце слева от всех столбцов данных.

Отчет. Эта группа предназначена для управления общими характеристиками отчета, в том числе итожащими статистиками для всей совокупности данных, отображением пропущенных значений, нумерацией страниц и заголовками.

Выводить наблюдения. Для каждого наблюдения выводятся фактические значения (или метки значений) переменных, указанных в группе Столбцы данных. Этот параметр создает отчет со списком наблюдений, который может быть намного длиннее сводного отчета.

Просмотр. Выводится только первая страница отчета. Этот параметр полезен для предварительного просмотра форматов, использованных в отчете, до момента генерации всего отчета.

Данные уже отсортированы. Для создания отчетов с группирующими переменными необходимо перед созданием отчета отсортировать файл данных по значениям группирующих переменных. Можно сберечь время обработки, выбрав этот параметр, если файл данных уже отсортирован по значениям группирующих переменных. Эта возможность особенно полезна после выполнения предварительного просмотра отчета.

Получение сводного отчета: итоги по строкам

1. Выберите в меню:

   Анализ > Отчеты > Отчет Итоги по строкам...

2. Выберите одну или несколько переменных для списка Столбцы данных. Для каждой отобранной переменной в отчете будет создан свой столбец.
3. Для отчетов, сортируемых и выводимых по подгруппам, выберите одну или несколько переменных для списка Группировать по.
4. Для отчетов с итожащими статистиками для подгрупп, задаваемых группирующими переменными, выберите группирующую переменную в списке Переменные группировки по столбцам и нажмите кнопку **Итоги** в панели Столбцы, чтобы задать необходимые итожащие показатели.
5. Для отчетов с итожащими статистиками для всей совокупности данных нажмите кнопку **Итоги**, чтобы задать необходимые итожащие показатели.

Формат столбцов данных / группирующих столбцов отчета

Диалоговые окна формата позволяют управлять заголовками столбцов, шириной столбцов, выравниванием текста и выбирать между выводом значений данных или меток значений. Диалоговое окно Формат столбца данных позволяет управлять форматом столбцов данных, располагающихся на правой стороне страницы отчета. Диалоговое окно Формат группировки позволяет управлять форматом группирующих столбцов, располагающихся слева.

Заголовок столбца. В этом текстовом поле задается заголовок столбца для выбранной переменной. Для длинных заголовков осуществляется автоматический переход на следующую строку в границах столбца. Пользуйтесь клавишей , чтобы вручную разорвать строку в том месте, где вы хотите продолжить вывод заголовка со следующей строки.

Положение значения в столбце. Для выбранной переменной можно управлять выравниванием значений или меток данных внутри столбца. Выравнивание значений или меток не влияет на выравнивание заголовков столбцов. Вы можете либо задать отступ содержимого столбца на заданное число символов, либо центрировать его.

Содержимое столбца. Для выбранной переменной этот переключатель позволяет задать вывод либо значений данных, либо заданных меток значений. Всегда, при отсутствии заданных меток значений показываются значения данных. (Переключатель не доступен для столбцов данных в отчетах по столбцам.)

Строки итогов для / строки с заключительными итогами в отчете

Два диалоговых окна задания строк итогов позволяют управлять отображением итожащих статистик для групп разбивки и для всего отчета в целом. Диалоговое окно Строки итожащих для позволяет управлять отображением групповых статистик для каждой категории, задаваемой группирующими переменными. Диалоговое окно Строки с заключительными итогами позволяет управлять отображением статистик для всей совокупности данных, выводимых в конце отчета.

Доступны следующие итожащие статистики: сумма, среднее значение, минимум, максимум, число наблюдений, процент наблюдений со значениями, меньшими или большими, чем заданное, процент наблюдений со значениями в заданном диапазоне, стандартное отклонение, эксцесс, дисперсия и асимметрия.

Параметры группировки отчета

Диалоговое окно параметров группировки позволяет управлять интервалами и распределением по страницам информации, сгруппированной по категориям.

Управление страницей. Эта группа позволяет управлять интервалами и распределением по страницам категорий выбранной группирующей переменной. Вы можете задать число пустых строк между группами или запросить вывод каждой группы с новой страницы.

Пустых строк перед итожащими статистиками. При помощи этого параметра можно управлять количеством пустых строк между метками групп или данными и итожащими статистиками. Эта возможность особенно полезна для комбинированных отчетов, включающих как списки отдельных наблюдений, так и итожащие статистики для групп; в таких отчетах можно вставлять пустые строки между списками наблюдений и итожащими статистиками.

Параметры отчета

Диалоговое окно параметров отчета позволяет управлять режимом обработки и вывода пропущенных значений, а также нумерацией страниц.

Исключать наблюдения с пропущенными значениями целиком. Исключает из отчета любое наблюдение с пропущенными значениями для какой-либо из переменных отчета.

Выводить пропущенные значения как. Этот параметр позволяет указать символ, который будет изображать значение, пропущенное в файле данных. Можно указать только один символ. Символ используется для представления как *системных пропущенных* значений, так и *задаваемых пользователем пропущенных* значений.

Начать нумерацию страниц с. Этот параметр позволяет указать номер для первой страницы отчета.

Компоновка отчета

Диалоговое окно компоновки отчета позволяет управлять шириной и высотой каждой страницы отчета, расположением отчета на странице и вставкой пустых строк и меток.

Компоновка страницы. Эта группа позволяет управлять отступами на странице, выраженными в строках (сверху и снизу) и символах (слева и справа), а также выравниванием отчета в границах этих отступов.

Заголовки и колонтитулы. Эта группа позволяет управлять количеством строк, отделяющих заголовки и колонтитулы от собственно отчета.

Столбцы группировки. Эта группа позволяет управлять выводом группирующих столбцов. Если задано несколько группирующих переменных, они могут находиться либо в отдельных столбцах, либо в первом столбце. При размещении всех группирующих переменных в первом столбце отчет получается более узким.

Заголовки столбцов. Эта группа позволяет управлять выводом заголовков столбцов, в том числе подчеркиванием, пропуском между заголовками и собственно отчетом, а также вертикальным выравниванием заголовков столбцов.

Строки данных и метки групп. Эта группа позволяет управлять расположением информации в столбцах данных (значения данных и/или итожащие статистики) относительно меток группировки, выводимых в начале каждой категории группировки. Первая строка информации в столбцах данных может либо начинаться на той же строке, что и метка категории группировки, либо отстоять от нее на заданное число строк. (Панель не задействована для отчетов по столбцам.)

Заголовки отчета

Диалоговое окно задания заголовков позволяет управлять содержанием и расположением заголовков и нижних колонтитулов. Вы можете задать заголовки и колонтитулы величиной до 10-ти строк с компонентами, выровненными на каждой строке влево, вправо или по центру.

Если в поля заголовков или колонтитулов вставлены переменные, то в заголовках или колонтитулах будут показаны их текущие значения или метки значений. В заголовках показывается метка, соответствующая значению переменной в начале страницы. В колонтитулах показывается метка, соответствующая значению переменной в конце страницы. Если у значения нет метки, показывается само значение.

Специальные переменные. Специальные переменные *DATE* и *PAGE* позволяют вставить текущую дату или номер страницы в любую строку заголовка или колонтитула. Если ваш файл данных содержит переменную *DATE* или *PAGE* , то вы не сможете использовать значения этих переменных в заголовках и колонтитулах.

Итоги по столбцам

Процедура Итоги по столбцам создает отчеты, в которых различные итожащие статистики располагаются в отдельных столбцах.

Пример. Компания с сетью магазинов розничной торговли ведет запись информации о служащих, включая размер оклада, продолжительность работы в занимаемой должности, а также магазин и отдел, в котором служащий работает. Вы могли бы создать отчет, содержащий итожащие статистики по продажам (например, среднее, минимум и максимум) для каждого отдела.

Столбцы данных. В этой группе задается список переменных, по которым необходимо получить итожащие статистики, а также предоставляется возможность управления форматом отображения и итожащими статистиками, выводимыми для каждой переменной.

Столбцы группировки. Эта группа позволяет задать список необязательных переменных, разбивающих отчет на группы, а также управлять форматом вывода группирующих столбцов. При наличии нескольких группирующих переменных для каждой категории каждой группирующей переменной будет создана отдельная группа внутри категорий предшествующей в списке группирующей переменной. Группирующие переменные должны представлять собой дискретные категориальные переменные, делящие наблюдения на ограниченное число имеющих смысл категорий.

Отчет. Эта группа предназначена для управления общими характеристиками отчета, в том числе отображением пропущенных значений, нумерацией страниц и заголовками.

Просмотр. Выводится только первая страница отчета. Этот параметр полезен для предварительного просмотра форматов, использованных в отчете, до момента генерации всего отчета.

Данные уже отсортированы. Для создания отчетов с группирующими переменными необходимо перед созданием отчета отсортировать файл данных по значениям группирующих переменных. Можно сберечь время обработки, выбрав этот параметр, если файл данных уже отсортирован по значениям группирующих переменных. Эта возможность особенно полезна после выполнения предварительного просмотра отчета.

Получение сводного отчета: Итоги по столбцам

1. Выберите в меню:

   Анализ > Отчеты > Отчет Итоги по столбцам...
2. Выберите одну или несколько переменных для списка Столбцы данных. Для каждой отобранной переменной в отчете будет создан свой столбец.
3. Для изменения итожащих показателей, показанных для переменной, выберите нужную переменную в списке Переменные столбцов данных и нажмите кнопку **Итоги** .
4. Чтобы получить несколько итожащих мер для одной переменной, выберите эту переменную в исходном списке и поместите ее в список Переменные столбцов данных несколько раз, по одному разу для каждой итожащей меры.
5. Для просмотра столбца, содержащего сумму, среднее значение, отношение или другую функцию от имеющихся столбцов, щелкните по **Вставить Итог** . При этом в списке Столбцы данных появится переменная *Итог* .

6. Для отчетов, сортируемых и выводимых по подгруппам, выберите одну или несколько переменных для списка Группировать по.

Итожащие функции столбцов данных

Диалоговое окно Строки итожащих для управляет итожащими статистиками, отображаемыми для переменной, выбранной в списке Столбцы данных.

Доступны следующие итожащие статистики: сумма, среднее значение, минимум, максимум, число наблюдений, процент наблюдений со значениями, меньшими или большими, чем заданное, процент наблюдений со значениями в заданном диапазоне, стандартное отклонение, эксцесс, дисперсия и асимметрия.

Итожащие статистики для столбцов данных, формирующие столбец итогов

Диалоговое окно Столбец итогов позволяет выбрать общие итожащие статистики, вычисляемые по двум или большему числу столбцов данных.

Вы можете выбирать среди следующих общих итожащих статистик: сумма столбцов, среднее столбцов, минимум столбцов, максимум столбцов, разность между значениями двух столбцов, частное от деления значений в одном столбце на значения в другом столбце, произведение столбцов.

Сумма столбцов. Столбец *итогов* представляет собой сумму столбцов, указанных в списке Столбец итожащих.

Среднее столбцов. Столбец *итогов* представляет собой столбец средних значений столбцов, указанных в списке Столбец итожащих.

Минимум столбцов. Столбец *итогов* представляет собой столбец минимальных значений столбцов, указанных в списке Столбец итожащих.

Максимум столбцов. Столбец *итогов* представляет собой столбец максимальных значений столбцов, указанных в списке Столбец итожащих.

1-й столбец - 2-й столбец. Столбец *итогов* представляет собой разность столбцов из списка Столбец итожащих. В списке Столбец итожащих должны присутствовать ровно два столбца.

1-й столбец / 2-й столбец. Столбец *итогов* представляет собой частное от деления столбцов, указанных в списке Столбец итожащих. В списке Столбец итожащих должны присутствовать ровно два столбца.

% в 1-й столб. / 2-й столб. Столбец *итогов* показывает, сколько процентов составляет значение первого столбца по отношению к значению второго столбца из списка Столбец итожащих. В списке Столбец итожащих должны присутствовать ровно два столбца.

Произведение столбцов. Столбец *итогов* представляет собой произведение столбцов, указанных в списке Столбец итожащих.

Формат столбцов отчета

Параметры форматирования столбцов данных и группирующих столбцов для процедуры Итоги по столбцам аналогичны описанным параметрам процедуры Итоги по строкам.

Параметры группировки отчета с итогами по столбцам

Диалоговое окно параметров группировки отчета позволяет управлять выводом на экран групповых итогов, интервалами и распределением по страницам информации, разбитой по категориям.

Групповой итог. Управляет отображением групповых итогов для категорий разбивки.

Управление страницей. Эта группа позволяет управлять интервалами и распределением по страницам категорий выбранной группирующей переменной. Вы можете задать число пустых строк между группами или запросить вывод каждой группы с новой страницы.

Пустых строк перед групповым итогом. Управляет количеством пустых строк между данными группы и групповыми итогами.

Параметры отчета для итогов по столбцам

Диалоговое окно параметров отчета позволяет управлять выводом на экран общих итогов, выводом на экран пропущенных значений, а также нумерацией страниц.

Общий итог. Эта панель позволяет управлять отображением общего итога и задавать его метку; общий итог выводится внизу столбца.

Пропущенные значения. Вы можете исключить пропущенные значения из отчета или указать один символ, который будет изображать пропущенные значения в отчете.

Компоновка отчета с итогами по столбцам

Параметры компоновки отчета для процедуры Итоги по столбцам аналогичны параметрам для процедуры Итоги по строкам.

Команда REPORT: дополнительные возможности

Язык синтаксиса команд также позволяет:

- Выводить различные итожащие функции в столбцах единственной итожащей строки.
- Вставлять итожащие строки в столбцы данных для переменных, отличных от переменной рассматриваемого столбца данных, или для различных комбинаций (сложных функций) итожащих функций.
- Использовать медиану, моду, частоту и процент в качестве итожащих функций.
- Более точно управлять форматом вывода итожащих статистик.
- Вставлять пустые строки в различные места отчета.
- Вставлять пустые строки после каждого *n* -го наблюдения в листинге.

Ввиду сложности синтаксиса команды `REPORT` , Вы, возможно, найдете удобным при составлении нового отчета с помощью синтаксиса приблизительно задать его форму с помощью диалоговых окон, затем скопировать и вставить соответствующий синтаксис, а затем уточнить синтаксис, чтобы вывести отчет в точности в той форме, в какой вы хотите.

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Глава 30. Анализ надежности

Анализ надежности позволяет изучить свойства шкал измерений и пунктов (items), которые их формируют. Процедура Анализ надежности вычисляет набор широко используемых мер надежности шкал, а также дает информацию о связях между отдельными пунктами на шкале. Для вычисления "межреспондентных" (interrater) оценок надежности могут использоваться внутриклассовые (intraclass) коэффициенты корреляции.

Пример. Измеряет ли моя анкета удовлетворенность клиентов надлежащим образом? Используя анализ надежности, вы можете определить степень, до которой пункты вашей анкеты связаны друг с другом. Вы можете получить общий индекс повторяемости или внутренней согласованности (internal consistency) шкалы в целом, а также можете идентифицировать проблемные пункты, которые следует удалить из шкалы.

Статистика. Описательные статистики для каждой переменной и для шкалы, итожащие статистики по пунктам, межпунктовые (inter-item) корреляции и ковариации, оценки надежности, таблица дисперсионного анализа (ANOVA), внутриклассовые коэффициенты корреляции, T^2 Хотеллинга и тест Тьюки на аддитивность.

Модели. Доступны следующие модели пригодности:

- **Альфа (Кронбаха).** Это модель внутренней согласованности, основанная на средней межпунктовой корреляции.
- **Расщепления пополам.** Эта модель делит шкалу на две части и исследует корреляцию между частями.
- **Гуттмана.** Эта модель вычисляет нижние границы Гуттмана для истинной пригодности.
- **Параллельная.** Эта модель предполагает, что все пункты имеют равные дисперсии и равные дисперсии ошибок по повторениям.
- **Строго параллельная.** Эта модель предполагает выполненными условия параллельной модели и, кроме того, требует равенства средних значений по пунктам.

Данные для анализа надежности

Данные. Данные могут быть дихотомическими, порядковыми или интервальными, но они должны быть закодированными в числовой форме.

Допущения. Наблюдения должны быть независимыми, а ошибки должны быть некоррелированными между пунктами. Каждая пара пунктов должна иметь двумерное нормальное распределение. Шкалы должны быть аддитивными, так что каждый пункт линейно связан с суммарной оценкой (total score).

Родственные процедуры. Если вы хотите выяснить размерность пунктов шкалы, чтобы определить, требуется ли более одной характеристики (construct) для объяснения структуры баллов пунктов, используйте Факторный анализ или Многомерное масштабирование. Чтобы выявить однородные группы переменных, используйте иерархический кластерный анализ для кластеризации переменных.

Как запустить анализ надежности

1. Выберите в меню:

   Анализ > Шкала > Анализ надежности...
2. Выберите две или более переменных в качестве потенциальных компонентов аддитивной шкалы.
3. Выберите модель из выпадающего списка Модель.

Статистики процедуры Анализ надежности

Вы можете выбрать различные статистики, описывающие вашу шкалу и пункты. Статистики, выводимые по умолчанию, включают число наблюдений, число пунктов и следующие оценки надежности:

- **Альфа модели.** Для дихотомических данных он эквивалентен коэффициенту Кьюдера-Ричардсона 20 (KR20).
- **Модели расщепления пополам:** Корреляция между формами, пригодность при расщеплении пополам Гуттмана, пригодность по Спирману-Брауну (равная и неравная длина) и коэффициент альфа для каждой половины.
- **Модели Гуттмана:** Коэффициенты пригодности от лямбда 1 до лямбда 6.
- **Параллельная и Строго параллельная модели:** Тест на согласие модели, оценки дисперсии ошибки, общая дисперсия и истинная дисперсия, оцененная общая межпунктовая корреляция, оцененная пригодность и несмещенная оценка пригодности.

Описательные для. Выдает описательные статистики для шкал или пунктов по наблюдениям.

- **Пункта.** Выдает описательные статистики для пунктов по наблюдениям.
- **Масштаб.** Выдает описательные статистики для шкал.
- **Шкалы, если пункт удален.** Выводит итожащие статистики, сравнивающие каждый пункт со шкалой, построенной по другим пунктам. Статистики включают среднее и дисперсию шкалы, когда из нее удален этот пункт, корреляцию между пунктом и шкалой, построенной по другим пунктам и значение альфа Кронбаха, если пункт удален из шкалы.

Итожащие статистики. Выводит описательные статистики распределений пунктов по всем пунктам шкалы.

- *Средние*. Итожащие статистики для средних пунктов. Выводятся наименьшее, наибольшее и среднее средних пунктов, диапазон и дисперсия средних для пунктов, а также отношение наибольшего среднего к наименьшему.
- *Дисперсии*. Итожащие статистики для дисперсий пунктов. Выводятся максимальная, минимальная и средняя дисперсии пунктов, размах и дисперсия для дисперсий пунктов, а также отношение максимальной дисперсии пунктов к минимальной.
- *Ковариации*. Итожащие статистики для межпунктовых корреляций. Выводятся наименьшее, наибольшее и среднее значения межпунктовых ковариаций, их диапазон и дисперсия, а также отношение наибольшей ковариации к наименьшей.
- *Корреляции*. Итожащие статистики для межпунктовых корреляций. Выводятся наименьшее, наибольшее и среднее значения межпунктовых корреляций, их диапазон и дисперсия, а также отношение наибольшей корреляции к наименьшей.

Межпунктовые. Выводит матрицы корреляций или ковариаций между пунктами.

Таблица дисперсионного анализа. Выводит результаты тестов на равенство средних.

- *F критерий*. Выводит таблицу дисперсионного анализа повторяющихся измерений.
- *Хи-квадрат Фридмана*. Выводит хи-квадрат Фридмана и коэффициент согласия Кендалла. Этот параметр подходит для ранговых данных. Критерий хи-квадрат заменяет обычный F-критерий в таблице ДА (ANOVA).
- *Хи-квадрат Кокрена*. Выводится Q Кокрена. Этот параметр подходит для дихотомических данных. Q статистика выдается в таблице ДА (ANOVA) вместо F-статистики.

Т-квадрат Хотеллинга. Выводит результаты многомерного теста для проверки нулевой гипотезы о том, что все пункты шкалы имеют одинаковые средние.

Критерий аддитивности Тьюки. Выводит результаты теста для проверки предположения об отсутствии мультипликативных взаимодействий между пунктами.

Внутриклассовые коэффициенты корреляции. Выводит меры согласованности значений внутри наблюдений.

- **Модель.** Выберите модель для вычисления внутриклассового коэффициента корреляции. Доступными моделями являются Двухфакторная смешанная, Двухфакторная случайная и Однофакторная случайная. Выбирайте **Двухфакторная смешанная** , если эффекты индивидуумов случайны, а эффекты пунктов фиксированы; **Двухфакторная случайная** , если эффекты индивидуумов и пунктов случайны, или **Однофакторная случайная** , если эффекты индивидуумов случайны.
- **Тип.** Выберите тип индекса. Доступными типами являются Согласованность и Абсолютное согласие.
- **Доверительный интервал.** Задайте уровень для доверительного интервала. Значение по умолчанию - 95%.
- **Проверяемое значение.** Задайте предполагаемое значение коэффициента для проверки гипотезы. Это значение, с которым сравнивается наблюденное значение. Значение по умолчанию равно 0.

Команда RELIABILITY: дополнительные возможности

Язык синтаксиса команд также позволяет:

- Считывать и анализировать корреляционную матрицу.
- Сохранять корреляционную матрицу для дальнейшего анализа.
- Для метода расщепления пополам задать расщепление на неравные части.

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Глава 31. Многомерное масштабирование

Целью Многомерного масштабирования (ММ) является обнаружение структуры в наборе значений некоторой меры расстояния между объектами или наблюдениями. Это осуществляется путем приписывания наблюдениям положения в некотором многомерном пространстве (обычно размерности два или три) таким образом, чтобы расстояния между полученными точками в этом пространстве как можно более точно аппроксимировали исходные различия. Во многих случаях размерности (измерения) этого пространства могут быть интерпретированы и использованы для дальнейшего осмысления ваших данных.

Если вы имеете переменные, полученные в результате реальных измерений, вы можете использовать многомерное масштабирование для снижения размерности данных (если необходимо, процедура Многомерного масштабирования может вычислить расстояния по многомерным данным). Многомерное масштабирование может также применяться к данным, представляющим собой субъективные оценки различий между объектами или понятиями. Дополнительно процедура Многомерного масштабирования может манипулировать данными типа различий из нескольких источников, которые могут появиться в случае наличия нескольких индивидуумов, производящих оценку, или респондентов, отвечающих на вопросы анкеты.

Пример. Как люди воспринимают сходство между различными марками и моделями автомобилей? Если у вас есть данные от респондентов, представляющие рейтинги сходства между различными марками и моделями автомобилей, то многомерное масштабирование может быть использовано для идентификации размерностей (измерений), описывающих восприятие потребителей. Например, вам, возможно, удастся показать, что цена и размер автомобиля определяют двумерное пространство, которое объясняет сходства, определенные вашими респондентами.

Статистика. Для каждой модели: матрица данных, матрица данных, полученная в результате оптимального масштабирования, S - стресс (Юнга), стресс (Краскала), RSQ, координаты стимулов, средний стресс и RSQ для каждого стимула в модели Повторяемого ММ (Replicated MDS). Для моделей индивидуальных различий (INDSCAL): веса субъекта и индекс отклонения направления вектора весов от средней тенденции (weirdness index). Для каждой матрицы в моделях повторяемого многомерного масштабирования: стресс и RSQ для каждого стимула. Графики: координаты стимулов (двумерные или трехмерные), диаграммы рассеяния преобразованных исходных близостей (disparities) против расстояний.

Данные для многомерного масштабирования

Данные. Если ваши данные - различия, то все они должны быть количественными и измеренными в одной и той же метрике. Если у вас многомерные данные, то переменные могут быть количественными, двоичными или количествами. Масштаб переменных является важным моментом - различия в масштабах могут повлиять на решение. Если ваши данные имеют существенные различия в масштабах (например, одна переменная измерена в долларах, а другая в годах), то вам следует подумать об их стандартизации (это может быть выполнено автоматически процедурой Многомерного масштабирования).

Предположения. Процедура Многомерного масштабирования не накладывает жестких ограничений на распределение вероятностей. Не забудьте выбрать подходящий уровень измерений (порядковый, интервальный или отношения) в диалоговом окне Многомерное масштабирование: Параметры, чтобы получить корректные результаты.

Родственные процедуры. Если вашей целью является снижение размерности, то альтернативным методом может быть факторный анализ, особенно в случае, когда ваши данные количественные. Если вы хотите идентифицировать группы сходных наблюдений, то дополните многомерное масштабирование применением одного из методов кластерного анализа: иерархического или *k* -средних.

Как запустить процедуру многомерного масштабирования

1. Выберите в меню:

   Анализ > Шкала > Многомерное масштабирование...
2. Для анализа выберите по крайней мере четыре числовых значения.
3. В группе Расстояния выберите пункты **Данные содержат расстояния** или **Вычислить расстояния по данным**.
4. Если выбран пункт **Вычислить расстояния по данным**, можно также выбрать группирующую переменную для индивидуальных метрик. Группирующая переменная может быть как числовой, так и строковой.

Дополнительно можно выполнить следующие действия.

- Указать форму матрицы расстояния, если даты являются расстояниями.
- Укажите меру расстояния для использования при создании расстояний из данных.

Многомерное масштабирование: Форма данных

Если ваш активный набор данных представляет расстояния между объектами для некоторого набора объектов или расстояния между двумя наборами объектов, задайте форму матрицы ваших данных, чтобы получить корректные результаты.

Примечание: Если в диалоговом окне Модель задана построчная обусловленность, выбор формы **Квадратная симметричная** невозможен.

Создание меры для многомерного масштабирования

Многомерное масштабирование использует данные типа различий для получения решения задачи масштабирования. Если вы имеете многомерные данные (значения измеренных переменных), вы должны сформировать данные типа различий для получения решения задачи масштабирования. Вы можете задать детали формирования мер различия по вашим данным.

Мера. В этой группе вы можете задать меру различия для предстоящего анализа. Выберите одну из альтернатив в группе Мера в соответствии с типом ваших данных и затем выберите одну из мер из выпадающего списка мер указанного типа. Доступны следующие альтернативы:

- **Интервальная.** Расстояние Евклида, квадрат расстояния Евклида, Чебышев, Блок, Минковского или Настроенная.
- **Количества**. Мера хи-квадрат или мера фи-квадрат.
- **Двоичная.** Расстояние Евклида, квадрат расстояния Евклида, Различие размеров, Различие структур, Дисперсия, Ланс и Уильямс.

Создать матрицу расстояний. Позволяет выбрать элемент анализа. Альтернативами являются Между переменными и Между наблюдениями.

Преобразовать значения. В определенных случаях, когда масштабы значений переменных сильно различаются, Вы, возможно, захотите стандартизировать значения, перед тем как вычислять близости (неприменимо к двоичным данным). Выберите метод стандартизации из выпадающего списка Стандартизация. Если стандартизация не требуется, выберите **Нет**.

Модель многомерного масштабирования

Корректность оценивания модели многомерного масштабирования зависит от данных и выбора модели.

Шкала измерения. Эта группа позволяет задать тип шкалы ваших данных. Альтернативами являются Порядковая, Интервальная и Отношений. Если ваши переменные измерены в порядковой шкале, то выбор **Развязывать связанные** позволит рассматривать переменные как непрерывные, так что проблема совпадений или связей (равных значений для разных наблюдений) будет решена оптимальным образом.

Обусловленность. Эта группа позволяет определить, какие сравнения осмысленны. Альтернативами являются Матричная, Построчная и Безусловно.

Размерность. Эта группа позволяет задать размерности (числа измерений) решений задачи масштабирования. Для каждого числа в заданном диапазоне находится одно решение. Задайте целые между 1 и 6. Минимум, равный 1, допустим, только если вы выбрали **Расстояние Евклида** в качестве модели масштабирования. Если вам требуется одно решение, задайте в качестве минимума и максимума одинаковые значения.

Модель масштабирования. Эта группа позволяет задать предположения, в которых осуществляется масштабирование. Возможными альтернативами являются Расстояние Евклида и Евклидово расстояние индивидуальных различий (эта модель иначе называется INDSCAL). Для модели индивидуальных различий с расстоянием Евклида вы можете пометить элемент **Допускать отрицательные веса субъектов** , если это подходит для ваших данных.

Параметры процедуры Многомерное масштабирование

Вы можете задать параметры для задачи многомерного масштабирования:

Вывод. Эта группа позволяет задать вывод различной выходной информации. Можно выбрать Групповые графики, Индивидуальные графики для субъектов, Матрица данных и Сводка по модели и параметрам.

Критерии. Эта группа позволяет определить, когда следует остановить итерации. Чтобы изменить значения по умолчанию, введите значения для **Сходимость s-стресса** , **Минимум s-стресса** и **Максимум итераций** .

Считать расстояния, меньшие n, пропущенными. Расстояния, меньшие, чем это значение, исключаются из анализа.

Команда ALSCAL: дополнительные возможности

Язык синтаксиса команд также позволяет:

- Применить модели трех дополнительных типов, известные как ASCAL, AINDS и GEMSCAL в литературе по многомерному масштабированию.
- Выполнить полиномиальные преобразования для данных, измеренных в интервальной шкале или шкале отношений.
- Анализировать сходства (вместо расстояний) для порядковых данных.
- Анализировать номинальные данные.
- Сохранять в файлах различные матрицы координат и весов и затем считывать их для анализа.
- Ввести ограничения для многомерной развертки.

Обратитесь к *Command Syntax Reference* за полной информацией о синтаксисе языка команд.

Глава 32. Статистики отношений

Процедура Статистики отношений предоставляет полный список итожащих статистик для описания отношения двух количественных переменных.

Вы можете отсортировать выводимые результаты по значениям группирующей переменной в возрастающем или убывающем порядке. Можно отменить вывод результатов процедуры вычисления статистик отношений, а сохранить их во внешнем файле.

Пример. Можно ли считать одинаковым отношение оценочной и продажной цен домов в каждой из пяти стран? Глядя на вывод процедуры, можно увидеть, что распределение отношений изменяется значительно при переходе от одной страны к другой.

Статистика. Медиана, среднее, взвешенное среднее, доверительный интервалы, коэффициент разброса (КР), центрированный к медиане коэффициент вариации, центрированный к среднему коэффициент вариации, индекс регрессивности (ИР), стандартное отклонение, среднее абсолютное отклонение (САО), диапазон, минимальное и максимальное значения, а также индекс концентрации для задаваемого пользователем диапазона в явном виде или как процент от медианы отношений, определяющий интервал вокруг медианы.

Данные для статистик отношений

Данные. Для кодировки значений группирующих переменных (номинальных или порядковых) используйте числа или строки (до 8 символов).

Допущения. Переменные, которые задают числитель и знаменатель отношения, должны быть количественными переменными, принимающими положительные значения.

Как получить статистики отношений

1. Выберите в меню:

   Анализ > Описательные статистики > Отношение...
2. Выберите переменную числителя.
3. Выберите переменную знаменателя.

Дополнительно можно:

- Выбрать группирующую переменную и задать порядок групп в выводе результатов.
- Выбрать, выводить ли результаты в окне средства просмотра.
- Выбрать, сохранить или нет результаты во внешнем файле для дальнейшего использования, а также задать имя файла, где результаты будут сохранены.

Статистики отношений

Расположение. Мерами положения центра распределения являются статистики, которые описывают распределение отношений.

- **Медиана.** Значение, такое, что число отношений, которые меньше данного значения, и число отношений, которые больше данного значения, одинаковы.
- **Среднее.** Результат суммирования отношений с делением результата на общее число отношений.
- **Взвешенное среднее.** Результат деления среднего значения числителя на среднее значение знаменателя. Взвешенное среднее также является средним значением отношений, взвешенных с помощью знаменателя.

- **Доверительные интервалы.** Это позволяет вывести доверительные интервалы для среднего, медианы и взвешенного среднего. В качестве доверительного уровня задайте значение, большее или равное 0 и меньшее 100.

Разброс. Эти статистики измеряют величину разброса наблюденных значений.

- **САО.** Среднее абсолютное отклонение является результатом суммирования абсолютных отклонений отношений от медианы с делением результата на общее число отношений.
- **КР.** Коэффициент разброса является результатом представления среднего абсолютного отклонения в виде процента от медианы.
- **ИР.** Индекс регрессивности, является результатом деления среднего на взвешенное среднее.
- **Ковариат, центрированный по медиане.** Центрированный к медиане коэффициент вариации является результатом представления квадратного корня из среднего квадрата отклонений от медианы в виде процента от медианы.
- **Ковариат, центрированный по среднему.** Центрированный к среднему коэффициент вариации является результатом представления стандартного отклонения в виде процента от медианы.
- **Стандартное отклонение.** Результат суммирования квадратов отклонений отношений от среднего, деления этой суммы на число общее отношений без единицы и взятия положительного квадратного корня.
- **Диапазон.** Диапазон является результатом вычитания минимального отношения из максимального отношения.
- **Минимум.** Минимум является наименьшим отношением.
- **Максимум.** Максимум является наибольшим отношением.

Индекс концентрации. Коэффициент концентрации измеряет процент отношений, которые попадают в некоторый интервал. Он может быть вычислен двумя различными способами:

- **Отношения между.** Здесь интервал задается явно указанием нижней и верхней границ интервала. Введите значения минимальной и максимальной долей и щелкните по **Добавить** , чтобы задать интервал.
- **Отношения в пределах.** Здесь интервал задается неявно, указанием процента от медианы. Введите значение между 0 и 100, затем щелкните по **Добавить** . Нижний конец интервала равен (1 – 0,01 × значение) × медиана, а верхний конец равен (1 + 0,01 × значение) × медиана.

Глава 33. Кривые ROC

Эта процедура полезна для оценки эффективности схем классификации, в которых есть одна переменная с двумя категориями, по которым классифицируются объекты.

Пример. Банк заинтересован в том, чтобы правильно классифицировать заемщиков по признаку возврата или не возврата предоставляемого им кредита. Для такой классификации разработаны различные методы. ROC кривые могут использоваться для оценки того, как хорошо работают эти методы.

Статистика. Площадь под ROC кривой с доверительным интервалом и точками координат ROC кривой. Диаграммы: кривая ROC.

Методы. Оценка площади под ROC кривой может быть вычислена или непараметрически, или параметрически с использованием дважды отрицательной экспоненциальной (binegative exponential) модели.

Данные для ROC кривой

Данные. Тестируемые переменные являются числовыми. Они нередко представляют собой вероятности, полученные из дискриминантного анализа или логистической регрессии, или оценки в произвольной шкале, обозначающие "степень уверенности" эксперта или оценивающего в том, что субъект попадает в ту или иную категорию. Переменная состояния может быть любого типа и указывает истинную категорию, к которой принадлежит субъект. Значение переменной состояния обозначает категорию, которую следует рассматривать как *положительную* .

Допущения. Предполагается, что возрастающие значения на шкале эксперта или оценивающего представляют возрастающую уверенность в том, что субъект принадлежит одной категории, тогда как убывающие значения на шкале представляют возрастающую уверенность в том, что субъект принадлежит другой категории. Пользователь должен выбрать направление, которое будет считаться *положительным* . Предполагается также, что известна *истинная* категория, к которой принадлежит каждый субъект.

Как запустить процедуру ROC Кривые

1. Выберите в меню:

   Анализ > ROC Кривые...
2. Выберите одну или несколько тестируемых переменных с вероятностями в качестве значений.
3. Выберите одну переменную состояния.
4. Задайте *положительное* значение для переменной состояния.

Параметры процедуры ROC Кривые

Вы можете задать следующие спецификации для ROC анализа:

Классификация. Позволяет определить, следует ли при классификации включать значение отсечения в группу, идентифицируемую как *положительную* , или нет. В настоящее время это не влияет на вывод результатов.

Направление теста. Позволяет задать направление шкалы по отношению к *положительной* категории.

Параметры для стандартной ошибки площади. Позволяет задать метод оценивания стандартной ошибки площади под кривой. Доступными методами являются непараметрический и основанный на дважды отрицательном экспоненциальном распределении. Также можно задать уровень для доверительного интервала. Доступным является диапазон от 50.1% до 99.9%.

Пропущенные значения. Позволяет задать режим обработки пропущенных значений.

Глава 34. Имитация

Прогнозные модели, например, линейная регрессия, требуют набора входных данных для прогноза исхода или целевого значения. Во многих реальных применениях значения входных данных не являются определенными. Имитация позволяет учесть неопределенность входных данных прогнозных моделей и оценить вероятность различных исходов модели в присутствии этой неопределенности. Например, у вас имеется модель прибыли, которая включает стоимость материалов в качестве входных данных, однако существует неопределенность в цене из-за волатильности рынка. Для моделирования этой неопределенности и определения ее влияния на прибыль можно воспользоваться имитацией.

Для имитации в IBM SPSS Statistics используется метод Монте-Карло. Неопределенные входные данные моделируются с распределениями вероятности (например, с треугольным распределением). Имитированные значения этих входных данных создаются, исходя из этих распределений. Входные данные, значения которых известны, остаются постоянными. Прогнозная модель оценивается при помощи имитированного значения для всех неопределенных входных данных и фиксированных значений для известных входных данных. На их основе рассчитывается целевое значение (или целевые значения) модели. Процесс повторяется множество раз (обычно десятки тысяч или сотни тысяч раз). В результате получается распределение целевых значений, которое можно использовать для ответа на вопросы о вероятностях. В контексте IBM SPSS Statistics при каждом повторе процесса создается отдельное наблюдение (запись) данных, которое состоит из набора имитированных значений для неопределенных входных данных, фиксированных значений и прогнозного целевого значения (или значений) модели.

Вы можете также имитировать данные при отсутствии прогнозной модели, задав распределения вероятностей для переменных, которые будут имитироваться. Каждое сгенерированное наблюдение данных состоит из набора имитированных значений для указанных переменных.

Чтобы выполнить имитацию, необходимо указать подробные сведения, такие как прогнозную модель, распределения вероятности для неопределенных входными данных, корреляции между этими входными значениями и фиксированными значениями. После указания всех сведений для имитации можно выполнить ее и дополнительно сохранить ее характеристики в файл **плана имитации**. Можно поделиться этим планом с другими пользователями, которые затем могут запустить имитацию без необходимости вникать в подробности ее создания.

Для работы с имитациями доступны два интерфейса. Мастер имитаций (Simulation Builder) представляет собой расширенный интерфейс для пользователей, которые разрабатывают и выполняют имитации. Он обеспечивает полный набор возможностей: разработка имитации, сохранение ее характеристик в файл плана имитации, указание вывода и запуск имитации. Можно создать имитацию на основе файла модели IBM SPSS или на основе набора определяемых пользователем уравнений в мастере имитаций. Кроме того, можно загрузить имеющийся план имитации в мастер имитаций, изменить любые настройки и запустить имитацию, при необходимости сохранив ее обновленный план. Также доступен упрощенный интерфейс для тех случаев, когда план имитации уже имеется, и нужно просто запустить ее. Он позволяет изменять настройки, чтобы выполнять имитацию при разных условиях, однако не обеспечивает полный набор возможностей мастера имитаций для их создания.

Порядок разработки имитации на основе файла модели

1. Выберите в меню:

   Анализ > Имитация...
2. Выберите **Выбрать файл модели SPSS** и нажмите кнопку **Продолжить**.
3. Откройте нужный файл модели.

   Файл модели - это файл XML, содержащий модель PMML, созданную из IBM SPSS Statistics или IBM SPSS Modeler. Дополнительную информацию смотрите в разделе “Вкладка Модель” на стр. 182.

4. На вкладке Имитация (в мастере имитаций) укажите распределения вероятности для имитированных входящих данных и фиксированных значений. Если в активном наборе данных содержатся хронологические данные для имитированных входных данных, нажмите кнопку **Подогнать все** для автоматического определения распределения, которое наиболее точно соответствует данным для каждого входящего значения, а также для определения корреляций между ними. Для каждого имитированного входного значения, не соответствующего данным хронологии, вы должны явно указать распределение, выбрав тип распределения и введя обязательные параметры.
5. Нажмите кнопку **Выполнить**, чтобы выполнить имитацию. По умолчанию план имитации с подробными сведениями о ней сохраняется в место, указанное в настройках сохранения.

Доступны следующие параметры:

- Измените расположение сохраненного плана имитации.
- Укажите известные корреляции между имитированными входными данными.
- Вычислите автоматически таблицу сопряженности связей между категориальными входными полями и используйте эти связи при генерировании данных для этих входных полей.
- Укажите анализ чувствительности для изучения эффекта вариации фиксированных значений или вариации параметра распределения для имитированных входных данных.
- Укажите дополнительные параметры, например, настройку максимального количества наблюдений для формирования или запроса хвостовой выборки.
- Настройка вывода.
- Сохранение имитированных данных в файл данных.

Порядок разработки имитации на основе пользовательских уравнений

1. Выберите в меню:

   Анализ > Имитация...
2. Выберите **Ввести уравнения** и нажмите кнопку **Продолжить**.
3. Чтобы определить каждое уравнение для прогнозной модели, на вкладке Модель мастера имитаций нажмите кнопку **Новое уравнение**.
4. Щелкните по вкладке Имитация и задайте распределения вероятности для имитированных и фиксированных входящих значений. Если в активном наборе данных содержатся хронологические данные для имитированных входных данных, нажмите кнопку **Подогнать все** для автоматического определения распределения, которое наиболее точно соответствует данным для каждого входящего значения, а также для определения корреляций между ними. Для каждого имитированного входного значения, не соответствующего данным хронологии, вы должны явно указать распределение, выбрав тип распределения и введя обязательные параметры.
5. Нажмите кнопку **Выполнить**, чтобы выполнить имитацию. По умолчанию план имитации с подробными сведениями о ней сохраняется в место, указанное в настройках сохранения.

Доступны следующие параметры:

- Измените расположение сохраненного плана имитации.
- Укажите известные корреляции между имитированными входными данными.
- Вычислите автоматически таблицу сопряженности связей между категориальными входными полями и используйте эти связи при генерировании данных для этих входных полей.
- Укажите анализ чувствительности для изучения эффекта вариации фиксированных значений или вариации параметра распределения для имитированных входных данных.
- Укажите дополнительные параметры, например, настройку максимального количества наблюдений для формирования или запроса хвостовой выборки.
- Настройка вывода.
- Сохранение имитированных данных в файл данных.

Порядок разработки имитации без прогнозной модели

1. Выберите в меню:

   Анализ > Имитация...
2. Щелкните по **Создать имитированные данные** и нажмите кнопку **Продолжить**.
3. На вкладке Модель (в мастере имитаций) выберите поля, которые вы хотите имитировать. Можно выбрать поля из активного набора данных или определить новые поля, нажав кнопку **Создать**.
4. Щелкните по вкладке Имитация и задайте распределения вероятности для полей, для которых хотите выполнить имитацию. Если в активном наборе данных содержатся хронологические данные для каких-либо из этих полей, нажмите кнопку **Подогнать все** для автоматического определения распределения, которое наиболее точно соответствует данным и для определения корреляций между этими полями. Для полей, не соответствующих данным хронологии, вы должны явно указать распределение, выбрав тип распределения и введя обязательные параметры.
5. Нажмите кнопку **Выполнить**, чтобы выполнить имитацию. По умолчанию имитированные данные сохраняются в новом наборе данных, заданном в параметрах сохранения. Кроме того, план имитации с подробными сведениями о ней сохраняется в место, указанное в настройках сохранения.

Доступны следующие параметры:

- Измените положение данных имитации или сохраненного плана имитации.
- Укажите известные корреляции между имитированными полями.
- Вычислите автоматически таблицу сопряженности связей между категориальными полями и используйте эти связи при генерировании данных для этих полей.
- Укажите анализ чувствительности для изучения эффекта параметра распределения для имитированного поля.
- Укажите дополнительные параметры, например, число генерируемых случаев.

Порядок выполнения имитации из плана

Доступны два способа выполнения имитации из плана. Можно воспользоваться диалоговым окном Выполнение имитации, которое в основном предназначено для выполнения имитации из плана. Кроме этого, можно воспользоваться мастером имитаций.

Порядок использования диалогового окна Выполнение имитации .

1. Выберите в меню:

   Анализ > Имитация...
2. Выберите **Открыть существующий план имитации**.
3. Убедитесь, что переключатель **Открыть в мастере имитаций** выключен, и нажмите кнопку **Продолжить**.
4. Откройте нужный план имитации.
5. В диалоговом окне Выполнение имитации нажмите кнопку **Выполнить**.

Порядок выполнения имитации из мастера имитации.

1. Выберите в меню:

   Анализ > Имитация...
2. Выберите **Открыть существующий план имитации**.
3. Включите переключатель **Открыть в мастере имитаций** и нажмите кнопку **Продолжить**.
4. Откройте нужный план имитации.
5. На вкладке Имитация измените все значения параметров, которые вы хотите изменить.
6. Нажмите кнопку **Выполнить**, чтобы выполнить имитацию.

Кроме того, можно выполнить действия, которые указаны ниже.

- Настройка или изменение анализа чувствительности для изучения эффекта вариации фиксированных значений или вариации параметра распределения для имитированных входных данных.
- Изменение распределений и корреляций для имитированных входных данных в соответствии с новыми данными.
- Изменение распределения для имитированных входных данных.
- Настройка вывода.
- Сохранение имитированных данных в файл данных.

Мастер имитаций

Мастер имитаций предоставляет полный набор возможностей для разработки и выполнения имитаций. Он позволяет выполнить общие задачи, которые перечислены ниже.

- Разработка и выполнение плана имитации для модели IBM SPSS, определенной в файле модели PMML.
- Разработка и выполнение имитации для прогнозной модели, определенной набором настраиваемых уравнений, определенных пользователем.
- Разработка и выполнение имитации, которая генерирует данные при отсутствии прогнозной модели.
- Выполнение имитации на основе существующего плана с дополнительным изменением настроек плана.

Вкладка Модель

Для имитаций, основанных на прогнозной модели, вкладка Модель задает источник модели. Для имитаций, не содержащих прогнозную модель, вкладка Модель задает поля для имитации.

Выбрать файл модели SPSS. Эта опция указывает, что прогнозная модель определяется в файле модели IBM SPSS. Файл модели IBM SPSS - это файл XML или сжатый архивный файл (файл .zip), содержащий модель PMML, созданную из IBM SPSS Statistics или IBM SPSS Modeler. Прогнозные модели создаются процедурами, такими как линейная регрессия и деревья решений в IBM SPSS Statistics, и могут экспортироваться в файл модели. Можно использовать другую модель; нажмите для этого кнопку **Обзор** и выберите нужный файл.

В мастере имитаций поддерживаются модели PMML

- Линейная регрессия
- Автоматизированная линейная модель
- Обобщенная линейная модель
- Обобщенная линейная смешанная модель
- Общая линейная модель
- Бинарная логистическая регрессия
- Полиномиальная логистическая регрессия
- Порядковая полиномиальная регрессия
- Регрессия Кокса
- Дерево
- Дерево с применением бустинга (C5)
- Дискриминантный
- Двухэтапный кластерный анализ
- Кластеризация К-средними
- Нейронная сеть
- Набор правил (список решений)

Примечание:

- Использование моделей PMML, у которых несколько полей (переменных) назначения или несколько разбиений, в имитации не поддерживается.
- Значения строковых входных полей моделей логистической регрессии ограничены в модели 8 байтами. Если вы заносите такие строчные значения в активный набор данных, убедитесь, что их длина не превышает 8 байт. Значения данных длиннее 8 байт исключаются из связанного категориального распределения при вводе и выводятся как несоответствующие в выходной таблице Несоответствующие категории.

Ввести уравнения для модели. Этот параметр указывает, что прогнозная модель состоит из одного или нескольких настраиваемых уравнений, созданных пользователем. Чтобы создать уравнения, нажмите кнопку **Новое уравнение**. Откроется редактор уравнений. В этом редакторе можно изменять существующие уравнения, копировать их для использования в качестве шаблонов для новых уравнений, изменять их порядок и удалять их.

- Мастер имитаций не поддерживает системы совместных уравнений или уравнений, целевое значение которых не является линейным.
- Настраиваемые уравнения оцениваются в том порядке, в котором они указаны. Если уравнение для данного целевого значения зависит от другого целевого значения, то последнее должно быть определено до первого.

  Рассмотрим набор из трех уравнений ниже; уравнение для *profit* зависит от значений *revenue* и *expenses*, поэтому уравнения для *revenue* и *expenses* должны предшествовать уравнению для *profit*.

```
revenue = price*volume
expenses = fixed + volume*(unit_cost_materials + unit_cost_labor)
profit = revenue - expenses
```

Создать имитированные данные без модели. Выберите эту опцию, чтобы имитировать данные без прогнозной модели. Задайте поля для имитирования, выбрав их из активного набора данных или нажав кнопку **Создать** для определения новых полей.

Редактор уравнений

Редактор уравнений позволяет создавать или изменять настраиваемые уравнения для прогнозной модели.

- Выражение для уравнения может содержать поля из активного набора данных или новых полей входных данных, которые определены в редакторе уравнений.
- Можно указать свойства цели, такие как шкалу измерения, метки значений и создание вывода для цели.
- Целевые значение раннее определенных моделей можно использовать как входящие значения для текущего уравнения, что позволяет создавать связанные уравнения.
- К уравнению можно приложить описательный комментарий. Комментарии показаны рядом с уравнением на вкладке Модель.

1. Введите название цели. (Необязательно) Выберите **Правка**, чтобы открыть диалоговое окно Определенные входные данные, позволяющее изменить свойства по умолчанию для назначения.
2. Для создания выражения можно вставлять компоненты в поле Числовое выражение или ввести в него условие вручную.

- Можно построить выражение при помощи полей активного набора данных или определить новые входные данные, нажав кнопку **Создать**. Это откроет диалоговое окно Определение входящих данных.
- Вы можете вставлять функции, выбрав группу функций из списка Группы функций и дважды щелкнув затем на функции в списке Функции (или выбрав функцию и затем щелкнув на кнопке со стрелкой). Введите все параметры, отмеченные знаками вопроса. Выбор группы **Все** обеспечивает вывод списка всех доступных функций. В специально выделенной области диалогового окна показано краткое описание выбранной функции.
- Текстовые константы должны быть заключены в апострофы.
- В значениях с десятичными знаками в качестве десятичного разделителя должна использоваться точка (.).

Примечание: В имитации не поддерживаются пользовательские уравнения со строковыми целевыми значениями.

Определенные входные данные: Диалоговое окно Определенные входные данные позволяет определить новые входные данные и задать свойства для целевых значений.

- Если входные данные, которые нужно использовать в уравнении, не существуют в активном наборе данных, их надо определить перед использованием в этом уравнении.
- Если вы имитируете данные без прогнозной модели, надо определить все имитируемые входные поля, существующие в активном наборе данных.

Имя. Укажите имя целевого или входного значения.

Назначение. Укажите шкалу измерений целевого значения. По умолчанию шкала измерений является количественной. Также можно определить создание вывода для этого целевого значения. Например, для набора связанных уравнений вас может интересовать вывод только из целевого значения для последнего уравнения и подавление вывода из других целевых значений.

Входные данные для имитации. Указывается, какие входные данные будут имитированы в соответствии с указанным распределением вероятности (распределение вероятности указано на вкладке Имитация). Шкала измерений определяет набор распределений по умолчанию, которые рассматриваются при поиске распределения, наиболее точно соответствующего данным для ввода (на вкладке Имитация посредством включения опции **Подогнать** или **Подогнать все**). Например, если шкала измерений непрерывна, будет рассматриваться нормальное распределение (подходящее для непрерывных данных), но биномиальное распределение рассматриваться не будет.

Примечание: Выберите шкалу измерений Строковая для строковых значений. Имитация строковых значений ограничена категорийным распределением.

Фиксированное входящее значение. Указывает на то, что значение входящего параметра известно и будет фиксированным. Фиксированные входящие значения могут быть числовыми или текстовыми. Укажите фиксированное входящее значение. Текстовые переменные должны быть заключены в апострофы.

Метки значений. Метки значений можно указать для целевых, имитированных и фиксированных входных данных. Метки значений используются при выводе диаграмм и таблиц.

Вкладка Имитация

На вкладке Имитация определены все свойства имитации, отличные от свойств прогнозной модели. На вкладке Имитация можно выполнить общие задачи, которые перечислены ниже.

- Указание распределений вероятности для имитированных входных данных и значений для фиксированных входных данных.
- Указание корреляций между имитированными входными данными. Для категорийных входных полей вы можете задать, что ассоциации, существующие между этими входными полями в активном наборе данных, используются при генерировании данных для этих входных полей.
- Указание дополнительных параметров, например, хвостовых выборок и критерия для соответствия распределений хронологическим данным.
- Настройка вывода.
- Укажите, где сохранять план имитации и имитированные данные.

Имитированные поля

Чтобы выполнить имитацию, каждое входное поле должно быть указано как фиксированное или имитированное. Имитированные входные значения являются неопределенными и создаются на основе указанного распределения вероятностей. Если для подлежащих имитации входных данных доступны данные хронологии, распределения, наиболее точно соответствующие данным, могут быть определены

автоматически наряду со всеми корреляциями между этими входными данными. Также можно указать распределения или корреляции вручную, если хронологические данные недоступны или необходимо использовать особые распределения или корреляции.

Фиксированные входные значения известны и остаются постоянными при каждом генерировании имитации. Например, у вас имеется линейная регрессионная модель продаж как функции количества входных данных, включая цену. Необходимо зафиксировать цену на уровне текущей рыночной цены. Вы укажите цену как фиксированное входящее значение.

Для имитаций, основанных на прогнозной модели, каждый предиктор в модели - это входное поле для имитации. Для имитаций, не содержащих прогнозную модель, входные поля - это поля, заданные на вкладке Модель.

Автоматическая подгонка распределений и вычисление корреляций для имитированных входных данных. Если активный набор данных содержит данные хронологии для входных данных, которые вы хотите имитировать, можно автоматически найти распределения, наиболее точно соответствующие этим входным данным, а также определить все корреляции между ними. Порядок выполнения действий описан далее.

1. Проверьте, что каждый элемент входных данных, который необходимо имитировать, соответствует корректному полю в активном наборе данных. Входные данные перечислены в столбце Входные данные. Подогнать по столбцу показывает соответствующее поле в активном наборе данных. Можно сопоставить входные данные с другим полем в активном наборе данных. Для этого выберите элемент Подогнать по раскрывающемуся списку .

   Значение -*Нет*- в столбце Подогнать свидетельствует о невозможности автоматического сопоставления входных данных с полем в активном наборе данных. По умолчанию входные данные сопоставляются с полями набора данных по имени, шкале измерения и типу (числовой или строковый). Если активный набор данных не содержит хронологических входных данных, то необходимо вручную задать распределение для них или указать фиксированные входные данные как описано ниже.
2. Нажмите кнопку **Подогнать все**.

Наиболее точно соответствующее распределение и связанные с ним параметры выводятся в столбце Распределение наряду с диаграммой распределения поверх гистограммы (или столбчатой диаграммы) хронологических данных. Корреляции между имитированными входными данными показаны в настройках корреляций. Можно проанализировать результаты подгонки и настроить автоматическую подгонку распределения для конкретных входных данных, выбрав для них строку и нажав кнопку **Детали подгонки**. Дополнительную информацию смотрите в разделе “Настройка подгонки” на стр. 187.

Можно выполнить автоматическую подгонку распределения для конкретных входных данных, выбрав для них строку и нажав кнопку **Подогнать**. Корреляции для всех имитированных входных данных, которые соответствуют полям в активном наборе данных, также вычисляются автоматически.

Примечание:

- Наблюдения с пропущенными значениями для имитированных входных данных исключаются из подгонки распределения, вычисления корреляций и вычисления необязательной таблицы сопряженности (для входных данных с категориальным распределением). Есть возможность указать, считать ли допустимыми пользовательские пропущенные значения во входных данных с категориальным распределением. По умолчанию они рассматриваются как пропущенные. Более подробную информацию смотрите в теме “Дополнительные параметры” на стр. 189.
- Если для количественных и порядковых входных данных не удается найти приемлемое соответствие среди всех протестированных распределений, в качестве наиболее точного соответствия предлагается эмпирическое распределение. Для количественных входных данных эмпирическое распределение является кумулятивной функцией распределения хронологических данных. Для порядковых входных данных эмпирическое распределение является категориальным распределением хронологических данных.

Указание распределений вручную. Распределение вероятностей для любых имитированных входных данных можно указать вручную, выбрав нужное распределение в выпадающем списке **Тип** и введя параметры

распределения в сетке Параметры. После ввода параметров для распределения рядом с сеткой Параметры будет выведен образец диаграммы распределения на основе указанных параметров. Далее изложены некоторые примечания по некоторым распределениям.

- **Категориальное.** Категориальные распределения описывают входное поле с фиксированным количеством значений, называемых категориями. Каждая категория имеет связанную с ней вероятность. Сумма вероятностей всех категорий равняется единице. Чтобы ввести категорию, щелкните по левому столбцу в сетке Параметры и задайте значение категории. Введите вероятность, связанную с категорией, в правый столбец.

  Примечание: Для категорийных входных полей из модели PMML категории определяются из модели, и изменить их нельзя.
- **Негативное биноминальное - ошибки.** Описывает распределение количества ошибок в последовательности испытаний перед обзором количества успешных исходов. Параметр *thresh* - указанное количество успешных исходов; параметр *prob* - вероятность успешного исхода в любых испытаниях.
- **Негативное биноминальное - испытания.** Описывает распределение количества испытаний, требуемых перед обзором количества успешных исходов. Параметр *thresh* - указанное количество успешных исходов; параметр *prob* - вероятность успешного исхода в любых испытаниях.
- **Диапазон.** Это распределение состоит из набора интервалов с вероятностью, назначенной каждому интервалу. Сумма вероятностей всех интервалов равна 1. Значения с заданным интервалом извлекаются из равномерного распределения, определенного на этом интервале. Интервалы указываются вводом минимального значения, максимального значения и связанной с ними вероятности.

  Например, вы полагаете что стоимость за единицу материала имеет 40%-ую вероятность попадания в диапазон $10 - $15 и 40%-ую вероятность попадания в диапазон $15 - $20. Вы смоделируете стоимость при помощи распределения Диапазон, которое состоит из двух интервалов - [10 - 15] и [15 - 20]. Для первого интервала вероятность составляет 0,4, для второго - 0,6. Интервалы не обязательно должны быть количественными; они могут даже пересекаться. Например, можно указать интервалы $10 - $15 и $20 - $25 или $10 - $15 и $13 - $16.
- **Распределение Вейбулла.** Параметр *c* является необязательным параметром положения, указывающим, где находится источник распределения.

У параметров для следующих распределений то же смысловое значение, что и у связанных функций генерации случайных переменных, доступных в диалоговом окне Вычисление переменной: Бернулли, бета, биноминальное, экспоненциальное, гамма, логнормальное, негативное биноминальное (испытания и ошибки), нормальное, пуассоновское и равномерное.

Указание фиксированных входных данных. Чтобы указать фиксированные входные данные, в выпадающем списке **Тип** столбца Распределение выберите Фиксированные и введите фиксированное значение. Данное значение может быть числовым или строковым в зависимости от того, является ли входное значение числовым или строковым. Текстовые переменные должны быть заключены в апострофы.

Указание границ имитированных значений. Большинство распределений поддерживают указание верхней и нижней границ имитированных значений. Чтобы указать нижнюю границу, введите значение в текстовое поле **Мин**; чтобы указать нижнюю границу, введите значение в текстовое поле **Макс**.

Блокирование входных данных. Блокирование входных данных, которое выполняется при помощи установки переключателя в таблице со значком блокировки, исключает их из автоматической подгонки распределения. Это особенно полезно при определении распределения или фиксированного значения вручную и необходимости устранить воздействие автоматической подгонки распределения. Блокирование также полезно, если вы собираетесь предоставить свой план имитации другим пользователям, которые запустят его в диалоговом окне Выполнение имитации, при необходимости предотвратить любые изменения в определенные входные данные. В этом отношении спецификации для заблокированных входных данных невозможно изменить в диалоговом окне Выполнение имитации.

Анализ чувствительности. Анализ чувствительности позволяет изучить влияние изменения систематических изменений фиксированных входных данных или параметра распределения для имитированных входных данных посредством формирования независимого набора имитированных наблюдений (то есть фактически отдельной имитации) для каждого указанного значения. Чтобы определить анализ чувствительности, выберите фиксированные или имитированные входные данные и нажмите кнопку **Анализ чувствительности**. Анализ чувствительности ограничен единым фиксированным входным параметром или единым параметром распределения для имитированного входного параметра. Дополнительную информацию смотрите в разделе “Анализ чувствительности” на стр. 188.

Значки состояния подгонки

Значки в Подогнать по столбцу указывают состояние подгонки для каждого поля входных данных.

Таблица 3. Значки состояния.

Значок	Описание
	Для входных данных не указано распределение и входные данные не указаны как фиксированные. Чтобы выполнить имитацию, необходимо указать распределение для этих входных данных или определить их как фиксированные и указать значение.
	Входные данные были раннее подогнаны по полю, которое не существует в активном наборе данных. Нет необходимости предпринимать какие-либо действия за исключением случаев, когда необходимо изменить распределение для входных данных в активном наборе данных.
	Наиболее точное распределение заменено альтернативным распределением из диалогового окна Детали подгонки.
	Входные данные задаются для наиболее точного распределения.
	Распределение указано вручную или итерации анализа чувствительности указаны для этих входных данных.

Настройка подгонки: В диалоговом окне Детали подгонки показаны результаты автоматической подгонки распределения для конкретных входных данных. Распределения упорядочиваются по степени согласия; наиболее точно соответствующее распределение указывается первым. Наиболее точно соответствующее распределение можно переопределить, включив радиокнопку для нужного распределения в столбце Использование. При выборе радиокнопки в столбце Использование также показана диаграмма распределения поверх гистограммы (или столбчатой диаграммы) хронологических данных для этих входных данных.

Статистика согласия. По умолчанию, а также для количественных полей, для определения статистики согласия применяется тест Андерсона-Дарлинга. Помимо этого, а также только для количественных полей можно указать тест Колмогорова-Смирнова для статистики согласия. Для этого нужно сделать соответствующий выбор в настройках Дополнительные параметры. Для количественных входных данных результаты обоих тестов показаны в столбце Статистика согласия (столбец A для теста Андерсона-Дарлинга и столбец K для теста Колмогорова-Смирнова) с выбранным тестом, который используется для упорядочивания распределений. Для порядковых и номинальных входных данных используется тест хи-квадрат. Также показаны p-значения, связанные с тестами.

Параметры. Параметры распределения, связанные с каждым подогнанным распределением, показаны в столбце Параметры. У параметров для следующих распределений то же смысловое значение, что и у связанных функций генерации случайных переменных, доступных в диалоговом окне Вычисление переменной: Бернулли, бета, биноминальное, экспоненциальное, гамма, логнормальное, негативное биноминальное (испытания и ошибки), нормальное, пуассоновское и равномерное. Дополнительную

информацию смотрите в разделе . Для категориального распределения имена параметров являются категориями, а значения параметров являются связанными с ними вероятностями.

Изменение при помощи настраиваемого набора распределения. Для автоматической подгонки распределения по умолчанию применяется шкала измерений входных данных, которая используется для определения набора распределений. Например, количественные распределения, такие как логнормальное и гамма, применяются при подгонке количественных входных данных, но дискретные распределения, такие как Пуассона и бинормальное, при этом не применяются. Можно выбрать подмножество распределений по умолчанию; для этого надо выбрать нужные распределения в столбце Изменение (подгонки). Можно также переопределить набор распределений по умолчанию, выбрав другую шкалу измерений в выпадающем списке **Рассматривать как (Шкала)** и выбрав распределения в столбце Изменение (подгонки). Нажмите кнопку **Выполнить изменение (подгонки)**, чтобы изменить настраиваемый набор распределения.

Примечание:

- Наблюдения с пропущенными значениями для имитированных входных данных исключаются из подгонки распределения, вычисления корреляций и вычисления необязательной таблицы сопряженности (для входных данных с категориальным распределением). Есть возможность указать, считать ли допустимыми пользовательские пропущенные значения во входных данных с категориальным распределением. По умолчанию они рассматриваются как пропущенные. Более подробную информацию смотрите в теме “Дополнительные параметры” на стр. 189.
- Если для количественных и порядковых входных данных не удается найти приемлемое соответствие среди всех протестированных распределений, в качестве наиболее точного соответствия предлагается эмпирическое распределение. Для количественных входных данных эмпирическое распределение является кумулятивной функцией распределения хронологических данных. Для порядковых входных данных эмпирическое распределение является категориальным распределением хронологических данных.

Анализ чувствительности: Анализ чувствительности позволяет изучить эффект изменения фиксированных входных данных или параметра распределения для имитированных входных данных по указанным наборам значений. Для каждого указанного значения формируется независимый набор имитированных наблюдений, т. е. фактически отдельная имитация. Каждый набор имитированных наблюдений называется **итерацией**.

Итерировать. Этот выбор позволяет указать набор значений, по которым будет изменяться входной параметр.

- Если вы изменяете значение параметра распределения, выберите нужный параметр в выпадающем списке. Введите набор значений в значение Параметр по сетке итераций. После нажатия кнопки **Продолжить** заданные значения будут добавлены в сетку Параметры связанного входного параметра с индексом, указывающим номер итерации значения.
- Для категориальных распределений или распределений диапазона могут быть изменены вероятности категорий или интервалов (соответственно), однако значения категорий и конечных точек интервалов не могут быть изменены. Выберите категорию или интервал из выпадающего списка и укажите набор вероятностей в значении Параметр по сетке итераций. Вероятности для других категорий или интервалов будут автоматически настроены соответственно.

Без итераций. Используйте этот параметр для отмены итераций для входных данных. После нажатия кнопки **Продолжить** итерации будут удалены.

Корреляции

Входные поля для имитации часто коррелируют - например, рост и вес. Корреляции между входными данными, которые будут имитированы, должны быть учтены, чтобы обеспечить их сохранение в имитированных значениях.

Пересчитать корреляции при подгонке. Этот вариант выбора позволяет автоматически рассчитать корреляции между имитированными входными данными при подгонке распределений к активному набору данных посредством действий **Подогнать все** или **Подогнать** в настройках Имитированные поля.

Не пересчитывать корреляции при подгонке. Выберите этот параметр, если необходимо вручную указать корреляции и не допустить их перезаписи при автоматической подгонке распределений в активном наборе данных. Значения, введенные в сетку Корреляции, должны быть в диапазоне между -1 и 1. Значение 0 указывает на отсутствие корреляции между связанными парами входных данных.

Сброс. Обнуление всех корреляций.

Использовать подогнанную многостороннюю таблицу сопряженности для вводов с категориальным распределением. Для входных полей с категорийным распределением вы можете автоматически вычислить многостороннюю таблицу сопряженности из активного набора данных, который описывает связи между этими входными полями. Эта таблица сопряженности затем используется при генерировании данных для этих входных полей. Если вы выбрали сохранение плана имитации, таблица сопряженности сохраняется в файле плана и используется, когда вы запускаете этот план.

- **Рассчитать таблицу сопряженности из активного набора данных.** Если вы работаете с существующим планом имитации, который содержит таблицу сопряженности, может пересчитать таблицу сопряженности по активному набору данных. Это действие переопределяет таблицу сопряженности из загруженного файла плана.
- **Использовать таблицу сопряженности из загруженного плана имитации.** По умолчанию, когда вы загружаете план имитации, который содержит таблицу сопряженности, используется таблица из этого плана. Вы можете пересчитать таблицу сопряженности из активного набора данных, выбрав **Рассчитать таблицу сопряженности из активного набора данных**.

Дополнительные параметры

Максимальное количество наблюдений. Указывает максимальное количество наблюдений имитированных данных, а также связанных целевых значений для создания. Если указан анализ чувствительности, это значение является максимальным значением для каждой итерации.

Цель для критерия останова. Если прогнозная модель содержит больше одного целевого значения, то можно выбрать цель, для которой будут применяться критерии остановки.

Критерий остановки Эти выборы определяют критерий для остановки имитации, потенциально до генерации максимально разрешенного количества наблюдений.

- **Продолжать до достижения максимума.** Указывает на то, что имитированные наблюдения будут сформированы до достижения максимального количества.
- **Остановить при выборке хвостов.** Воспользуйтесь этим параметром для гарантии адекватной выборки одного из хвостов указанного целевого распределения. Имитированные наблюдения будут созданы до завершения выборки хвоста или до достижения максимального количества наблюдений. Если прогнозная модель содержит несколько целевых значений, то выберите целевое значение, к которому будет применен этот критерий из списка **Целевое значение для критерия остановки**.

  Тип. Можно определить границы региона хвоста, указав целевое значение, например, 10000000 или процентиль, например, 99-ый. Если в раскрывающемся списке **Тип** выбрано Значение, введите значение границы в текстовое поле Значение и воспользуйтесь раскрывающимся списком **Сторона** для определения правой или левой области хвоста. Если в раскрывающемся списке **Тип** выбрано Процентиль, введите значение в текстовом поле Процентиль.

  Частота. Укажите количество целевых значений, которые должны лежать в области хвоста, чтобы обеспечить адекватную выборку хвоста. Генерирование наблюдений остановится, когда это количество будет достигнуто.
- **Остановиться, когда доверительный интервал среднего в пределах указанного порогового значения.** Воспользуйтесь этим параметром, чтобы обеспечить заданную степень точности среднего целевого значения. Имитированные наблюдения будут созданы до достижения указанной степени точности или максимального количества наблюдений. Чтобы воспользоваться этим параметром, укажите доверительный интервал и пороговое значение. Имитированные наблюдения будут генерироваться до тех пор, пока доверительный интервал, связанный с указанным уровнем, находится в пределах порогового значения. Например, можно воспользоваться этим параметром, чтобы определить формирование

наблюдений до тех пор, пока доверительный интервал среднего с доверительным уровнем 95% находится в пределах 5%-го отклонения от среднего значения. Если прогнозная модель содержит несколько целевых значений, то выберите целевое значение, к которому будет применен этот критерий из списка **Целевое значение для критерия остановки**.

Тип порога. Порог можно указать как числовое значение или как процентное отношение к среднему. Если в раскрывающемся списке **Тип порога** выбрано Процентиль, введите значение в текстовом поле Порог как значение. Если в раскрывающемся списке **Тип порога** выбрано Процент, введите значение в текстовом поле Порог как процент.

Количество наблюдений для выборки. Указывает количество наблюдений для использования при автоматической подгонке распределений для имитированных входных данных в соответствии с активным набором данных. Если ваш набор данных очень большой, можно ограничить количество наблюдений, которые используются для подгонки распределений. Если выбрать **Ограничить до N наблюдений**, то будут использованы первые N наблюдений.

Критерий статистики согласия (количественный). Для количественных входных данных можно использовать тест согласия статистики Андерсона-Дарлинга или тест Колмогорова-Смирнова для ранжирования распределений при их подгонке для имитированных входных значений в соответствии с активным набором данных. Тест Андерсона-Дарлинга выбирается по умолчанию и в особенности рекомендуется, когда необходимо обеспечить наилучшую возможную подгонку в областях хвоста.

Эмпирическое распределение. Для количественных входных данных эмпирическое распределение является кумулятивной функцией распределения хронологических данных. Можно указать количество интервалов, которые используются для расчета эмпирического распределения для количественных входных данных. По умолчанию задано значение 100, максимальное значение - 1000.

Воспроизвести результаты. Задание стартового числа генератора псевдослучайных чисел позволяет воспроизвести имитацию. Задайте целое число или щелкните по **Генерировать**, чтобы сгенерировать псевдослучайное целое число в диапазоне между 1 и 2147483647 включительно. Значение по умолчанию - 629111597.

Примечание: Для определенного случайного стартового числа результаты воспроизводятся, если число потоков не изменено. На одно и том же компьютере число потоков не меняется, если его не изменили командой SET THREADS. Число потоков может измениться, если вы запускаете имитацию на другом компьютере, так как для определения числа потоков на каждом компьютере используется внутренний алгоритм.

Пользовательские значения отсутствия как входная информация с категориальным распределением. Эти управляющие элементы задают, будут ли пользовательские значения отсутствия с категориальным распределением рассматриваться как допустимые. Системные и пользовательские значения отсутствия для всех прочих типов входных полей всегда рассматриваются как недопустимые. Все входные поля должны иметь допустимые значения, чтобы наблюдение было включено в подгонку распределения, вычисление корреляций и вычисление необязательной таблицы сопряженности.

Функции плотности

Эти настройки позволяют настроить вывод для функций плотности вероятности и кумулятивных функций распределения для количественных целей, а также столбчатые диаграммы прогнозных значений для категориальных целей.

Функция плотности вероятности (Probability Density Function, PDF). Эта функция показывает распределение целевых значений. Для количественных целевых значений она позволяет определять вероятность того, что они находятся в данной области. Для категориальных целевых значений (целевые значения с количественной или порядковой шкалой измерения) создается столбчатая диаграмма, в которой показан процент наблюдений, которые относятся к каждой из категорий целевого значения. Для категориальных значений доступны дополнительные параметры категориальных целей моделей PMML для описанной далее настройки отчета.

При использовании двухэтапного кластерного анализа и кластерного анализа методом k-средних создается столбчатая диаграмма принадлежности к кластеру.

Кумулятивная функции распределения (CDF). Кумулятивная функция распределения показывает вероятность того, что целевое значение меньше указанного значения либо равно ему. Она доступна только для количественных целевых значений.

Положения ползунка. Вы можете задать начальные положения подвижных опорных линий на диаграммах PDF и CDF. Задаваемые значения для нижней и верхней линий относятся к положениям по горизонтальной оси, а не к процентилям. Можно удалить нижнюю линию, выбрав **-Infinity**, или верхнюю линию, выбрав **Infinity**. По умолчанию эти линии располагаются на 5-й и 95-й процентилях. Если на одной диаграмме показаны несколько функций распределения (из-за нескольких целевых значений или результатов из итераций анализа чувствительности), значения по умолчанию относятся к функции распределения для первой итерации или первого назначения.

Опорные линии (количественные). Для функции плотности вероятности и кумулятивных функций распределения для количественных целевых значений можно добавить различные вертикальные опорные линии.

- **Сигмы.** Можно добавить опорные линии с амплитудой указанного количества стандартных отклонений от среднего целевого значения.
- **Процентили.** Можно добавить опорные линии в одном или двух значениях процентилей распределения для целевого значения в текстовых полях Нижняя и Верхняя. Например, значение 95 в текстовом поле Верхняя представляет 95-ый процентиль, который является значением, ниже которого попадают 95 % наблюдений. Точно так же, значение 5 в текстовом поле Нижняя представляет 5-ый процентиль, который является значением, ниже которого попадают 5% наблюдений.
- **Настраиваемые опорные линии.** Можно добавить опорные линии в указанных значениях цели.

Примечание: Если на одной диаграмме показаны несколько функций распределения (из-за нескольких целевых значений или результатов из итераций анализа чувствительности), опорные линии применяются только к функции распределения для первой итерации или первого назначения. Вы можете добавить опорные линии к другим распределениям в диалоговом окне Параметры диаграмм, к которому можно обратиться с диаграммы PDF или CDF.

Перекрыть результаты из отдельных количественных целевых значений. При наличии нескольких количественных целевых значений определяет вывод на экран функций распределения для всех таких целевых значений на одной диаграмме: одна диаграмма для функций плотности вероятности, другая - для функций кумулятивного распределения. Если этот параметр не выбран, результаты для каждого целевого значения будут показаны на отдельной диаграмме.

Значения категории для отчета. Для моделей PMML с категориальными целевыми значениями результатом модели является набор прогнозных вероятностей (по одной для каждой категории) того, что целевое значение попадает в каждую из категорий. Категория с наивысшей вероятностью выбирается в качестве предсказанной и используется при генерировании столбчатой диаграммы, описанной для настройки **Функция плотности вероятности** выше. Если выбрано **Предсказанная категория**, будет создана столбчатая диаграмма. Если выбрать **Предсказанные вероятности**, для каждой из категорий назначения будут сгенерированы гистограммы распределения.

Группирование для анализа чувствительности. Имитации, которые включают анализ чувствительности, создают независимый набор предсказанных целевых значений для каждой итерации, определенной анализом (варьируется одна итерация для каждого значения входных данных). При наличии итераций столбчатая диаграмма предсказанной категории для категориального целевого значения показывается в качестве кластеризованной столбчатой диаграммы, которая включает результаты для всех итераций. Категории или итерации можно сгруппировать.

Вывод

Диаграммы торнадо. Диаграммы торнадо - это столбчатые диаграммы, показывающие отношения между целевыми и имитированными входящими значениями при помощи множества метрик.

- **Корреляция целевых данных с входными.** Позволяет создать диаграммы торнадо для коэффициентов корреляции между данной целью и каждым из ее имитированных значений. Этот тип диаграмм торнадо не поддерживает целевые значения с номинальной или порядковой шкалой измерений и имитированные входные значения с категориальным распределением.
- **Вклад в дисперсию.** Позволяет создать диаграммы торнадо, которые показывают вклад в дисперсию каждого целевого значения из его имитированных входных значений, позволяя оценить степень, в которой каждое входное значение имеет вклад в общую неопределенность цели. Этот тип диаграмм торнадо не поддерживает целевые значения с порядковой или номинальной шкалой измерений и имитированные входные значения с любым из следующих распределений: категориальным, Бернулли, биномиальным, Пуассона или отрицательным биномиальным.
- **Чувствительность целевого значения к изменению.** Позволяет создать диаграммы торнадо, которые показывают влияние на целевое значение модулирования каждого имитированного входного значения с амплитудой указанного количества стандартных отклонений распределения, связанного с входными данными. Этот тип диаграмм торнадо не поддерживает целевые значения с порядковой или номинальной шкалой измерений и имитированные входные значения с любым из следующих распределений: категориальным, Бернулли, биномиальным, Пуассона или отрицательным биномиальным.

Ящичная диаграмма распределения целевых значений. Ящичные диаграммы доступны для количественных целевых значений. Выберите **Перекрыть результаты из отдельных целевых значений**, если у прогнозной модели несколько количественных целевых значений и вы хотите выводить ящичные диаграммы для всех целевых значений на одной диаграмме.

Сравнение диаграмм рассеяния целевых и входящих значений. Диаграммы рассеяния против имитированных входных данных доступны как для количественных, так и для категориальных целевых значений, и включают рассеяния целевых значений как с количественными, так и с категориальными входными данными. Диаграммы рассеяния, включающие категориальные целевые значения или категориальные входные данные, показаны в виде тепловой карты.

Создать таблицу значений процентилей. Для количественных целевых значений можно получить таблицу указанных процентилей целевых распределений. Квартили - это 25%-е, 50%-е и 75%-е процентили, которые разделяют наблюдения на четыре группы одинакового объема. Если вы хотите получить разбивку на иное число равных групп, выберите **Интервалы** и задайте число. Выберите **Настраиваемые процентили**, чтобы указать отдельные процентили, например, 99-й процентиль.

Описательные статистики целевых распределений. Этот параметр позволяет создать таблицы описательных статистик для количественных и категориальных целевых значений, а также для количественных входных данных. Для количественных целевых значений таблица включает среднее, стандартное отклонение, медиану, минимум и максимум, доверительный интервал среднего на указанном уровне, а также 5-ый и 95-ый процентили целевого распределения. Для категориальных целевых значений в таблицу входит процент наблюдений, которые попадают в каждую из категорий целевого значения. Для категориальных целевых значений моделей PMML таблица также включает среднюю вероятность каждой категории целевого значения. Для количественных входных данных в таблицу входят среднее, стандартное отклонение, минимум и максимум.

Корреляции и таблица сопряженности как входная информация. Эта опция выводит таблицу коэффициентов корреляции между имитированными входными полями. Когда входные поля с категорийным распределением генерируются из таблицы сопряженности, выводится также таблица сопряженности данных, сгенерированный для этих входных полей.

Имитированные входные данные для включения в вывод. По умолчанию все имитированные входные данные включены в вывод. Выбранные входные имитированные данные можно исключить из вывода. Это также исключит их из диаграмм торнадо, диаграмм рассеяния и табличного вывода.

Ограничить диапазоны для непрерывных полей назначения. Вы можете задать диапазон допустимых значений для одного или нескольких последовательных назначений. Значения вне заданного диапазона исключаются из всего вывода и анализа, связанных с этим назначением. Чтобы задать нижний предел, выберите **Нижний** в столбце Предел и введите значение в столбце Минимум. Чтобы задать верхний предел, выберите **Верхний** в столбце Предел и введите значение в столбце Максимум. Чтобы задать и нижний, и верхний предел, выберите **Оба** в столбце Предел и введите значения в столбцах Минимум и Максимум.

Форматы вывода на экран. Можно задать формат, который используется при выводе на экран значений целевых значений и входных данных (как для фиксированных, так и для имитированных входных данных).

Сохранение

Сохранение плана этой симуляции. Текущие характеристики симуляции можно сохранить в файл плана симуляции. Расширение файлов планов имитации - *.splan*. План имитации можно открыть заново в мастере имитаций, внести изменения (при необходимости) и выполнить имитацию. Можно поделиться планом имитации с другими пользователями, которые затем могут выполнить его в диалоговом окне Выполнение имитации . Планы имитации включают в себя все спецификации, кроме следующих: настройки для функций плотности, настройки вывода для диаграмм и таблиц, расширенные параметры для соответствия, эмпирического распределения и случайного значения.

Сохранение имитированных данных в новый файл данных. Можно сохранить имитированные входные данные, фиксированные входные данные и предсказанные целевые значения в файл данных SPSS Statistics, новый набор данных в текущем сеансе или файле Excel. Каждое наблюдение (или строка) файла данных состоят из предсказанных значений целей вместе с имитированными входными данными и фиксированными входными данными, которые генерируют целевые значения. Если анализ чувствительности указан, то при каждой итерации создается последовательный набор наблюдений, которые отмечены номером итерации.

Диалоговое окно Выполнение имитации

Диалоговое окно Выполнение имитации разработано для пользователей, которые имеют план имитации и хотят только выполнить ее. Также в нем предоставлены функции, необходимые для выполнения имитации при различных условиях. Он позволяет выполнить общие задачи, которые перечислены ниже.

- Настройка или изменение анализа чувствительности для изучения эффекта вариации фиксированных значений или вариации параметра распределения для имитированных входных данных.
- Изменение распределений вероятности для неопределенных входных данных (и корреляции между этими входными данными) в соответствии с новыми данными.
- Изменение распределения для имитированных входных данных.
- Настройка вывода.
- Выполнение имитации.

Вкладка Имитация

Вкладка Имитация позволяет определять анализ чувствительности, изменять распределение вероятности для имитированных входных данных и корреляции между новыми имитированными входными данными, а также изменять распределение вероятности, связанное с имитированными входными данными.

Сетка Имитированные входные данные содержит запись для каждого входящего значения, определенного в плане имитации. В каждой записи выводится имя входных данных и связанный с ними тип распределения вероятностей с образцом диаграммы соответствующей кривой распределения. Каждый набор входных данных имеет значок состояния (цветной круг с переключателем), который полезен при изменении распределений в соответствии с новыми данными. Кроме того, входные данные могут иметь значок блокировки, который указывает, что они заблокированы и не могут быть изменены в диалоговом окне Выполнение имитации . Чтобы изменить заблокированные входные данные, необходимо открыть план имитации в мастере имитации.

Каждое входное значение является имитированным либо фиксированным. Имитированные входные значения являются неопределенными и создаются на основе указанного распределения вероятностей. Фиксированные входные значения известны и остаются постоянными при каждом генерировании имитации. Чтобы обработать те или иные входящие данные, выберите соответствующую запись в сетке Имитированные входные данные.

Определение анализа чувствительности

Анализ чувствительности позволяет изучить влияние изменения систематических изменений фиксированных входных данных или параметра распределения для имитированных входных данных посредством формирования независимого набора имитированных наблюдений (то есть фактически отдельной имитации) для каждого указанного значения. Чтобы определить анализ чувствительности, выберите фиксированные или имитированные входные данные и нажмите кнопку **Анализ чувствительности**. Анализ чувствительности ограничен единым фиксированным входным параметром или единым параметром распределения для имитированного входного параметра. Дополнительную информацию смотрите в разделе “Анализ чувствительности” на стр. 188.

Изменение распределений в соответствии с новыми данными

Порядок автоматического изменения распределения вероятностей для имитированных входных данных (и корреляций между ними) в соответствии с новым активным набором данных.

1. Проверьте, что каждая модель входных данных соответствует корректному полю в активном наборе данных. Каждое имитированное входное значение соответствует полю в активном наборе данных, указанному в связанном с этим значением выпадающем списке **Поле**. Несоответствующие входные значение легко определить - на значке состояния будет указан вопросительный знак.

2. Измените все необходимые соответствия полям; для этого выберите **Подогнать по полю в наборе данных**, а затем выберите нужное поле из списка.
3. Нажмите кнопку **Подогнать все**.

Для каждого соответствующего входного значения наиболее точно соответствующее распределение показано рядом с диаграммой распределения, которая наложена на гистограмму (или столбчатую диаграмму) хронологических данных. При невозможности найти приемлемое соответствие используется эмпирическое распределение. Для входящих значений, которые соответствуют эмпирическому распределению, вы увидите только гистограмму хронологических данных, поскольку эмпирическое распределение фактически представлено данной диаграммой.

Примечание: Полный список значков состояния смотрите в теме “Имитированные поля” на стр. 184.

Изменение вероятности распределений

Невозможно изменить вероятность распределений для имитированных данных и дополнительно изменить имитированные данные в фиксированные и наоборот.

1. Выберите нужные входные данные и нажмите кнопку **Ручное распределение**.
2. Выберите тип распределения и задайте его параметры. Чтобы изменить имитированные входные данные на фиксированные входные данные, выберите Фиксированные в выпадающем списке **Тип**.

После ввода параметров для распределения, его образец (показанный в записи входных данных) будет обновлен в соответствии с изменениями. Дополнительную информацию о задании распределений вероятности вручную смотрите в теме “Имитированные поля” на стр. 184.

Включить пользовательские пропущенные значения категориальных входных полей при подгонке Задает, будут ли пользовательские значения отсутствия с категориальным распределением рассматриваться как допустимые при переподгонке по данным активного набора данных. Системные и пользовательские значения отсутствия для всех прочих типов входных полей всегда рассматриваются как недопустимые. Все входные поля должны иметь допустимые значения, чтобы наблюдение было включено в подгонку распределения и вычисление корреляций.

Вкладка Вывод

Вкладка Вывод позволяет настроить вывод, созданный имитацией.

Функции плотности. Функции плотности являются основными средствами проверки набора результатов имитации.

- **Функция плотности вероятности.** Функция плотности вероятности показывает целевые значения распределения, позволяя пользователю определить вероятность нахождения целевого значения в нужной области. Для целевых значений с фиксированным набором результатов, например, неудовлетворительное обслуживание, удовлетворительное обслуживание, хорошее обслуживание и отличное обслуживание, создается столбчатая диаграмма, на которой выводятся процентные показатели наблюдений, которые соответствуют каждой из категорий целевого значения.
- **Кумулятивная функция распределения.** Кумулятивная функция распределения показывает вероятность того, что целевое значение меньше указанного значения либо равно ему.

Диаграммы торнадо. Диаграммы торнадо - это столбчатые диаграммы, показывающие отношения между целевыми и имитированными входящими значениями при помощи множества метрик.

- **Корреляция целевых данных с входными.** Позволяет создать диаграммы торнадо для коэффициентов корреляции между данной целью и каждым из ее имитированных значений.
- **Вклад в дисперсию.** Позволяет создать диаграммы торнадо, которые показывают вклад в дисперсию каждого целевого значения из его имитированных входных значений, позволяя оценить степень, в которой каждое входное значение имеет вклад в общую неопределенность цели.
- **Чувствительность целевого значения к изменению.** Позволяет создать диаграммы торнадо, которые показывают влияние на цель модулирования каждого имитированного входного значения с амплитудой в одно стандартное отклонение распределения, связанного с входными данными.

Сравнение диаграмм рассеяния целевых и входящих значений. Позволяет создать диаграммы рассеяния целевых значений против имитированных входящих значений.

Ящичная диаграмма распределения целевых значений. Позволяет создать ящичные диаграммы распределения целевых значений.

Таблица квартилей. Этот параметр позволяет создать таблицу квартилей целевых распределений. Квартили распределения - это 25-ый, 50-ый и 75-ый процентили распределения, которые разделяют наблюдения на четыре группы одинакового объема.

Корреляции и таблицы сопряженности для входных полей. Эта опция выводит таблицу коэффициентов корреляции между имитированными входными полями. Таблица сопряженности связей между входными полями с категориальным распределением выводится, когда план имитации задает генерирование категорийных данных из таблицы сопряженности.

Перекрыть результаты из отдельных целевых значений. Если имитируемая прогнозная модель содержит несколько целевых значений, можно задать вывод на экран на одной диаграмме результатов из отдельных целей. Эта настройка применяется к диаграммам функций плотности вероятности, кумулятивным функциям распределения и ящичным диаграммам. Например, если выбрать этот параметр, то функции плотности вероятности для всех целей будут показаны на одной диаграмме.

Сохранение плана этой имитации. Любые изменения имитации можно сохранить в файл плана имитации. Расширение файлов планов симуляции - *.splan*. План можно повторно открыть в диалоговом окне Выполнение имитации или в мастере имитаций. В планы имитации включены все характеристики за настроек вывода.

Сохранение имитированных данных в новый файл данных. Можно сохранить имитированные входные данные, фиксированные входные данные и предсказанные целевые значения в файл данных SPSS Statistics, новый набор данных в текущем сеансе или файле Excel. Каждое наблюдение (или строка) файла данных состоят из предсказанных значений целей вместе с имитированными входными данными и фиксированными входными данными, которые генерируют целевые значения. Если анализ чувствительности указан, то при каждой итерации создается последовательный набор наблюдений, которые отмечены номером итерации.

Если необходима более глубокая настройка вывода, выполните имитацию при помощи мастера имитаций. Дополнительную информацию смотрите в разделе “Порядок выполнения имитации из плана” на стр. 181.

Работа с выводом диаграммы из имитации

Ряд диаграмм, созданных на основе имитации, имеют интерактивные функции, которые позволяют настроить вывод на экран. Для использования интерактивных функций активируйте объект диаграммы (двойным щелчком мыши) в окне вывода средства просмотра. Все диаграммы имитаций являются визуализациями графической панели.

Диаграммы функций плотности вероятности для непрерывных целевых переменных. Эта диаграмма имеет две скользящих вертикальных опорных линии, которые разделяют ее на отдельные области. В таблице ниже на диаграмме показана вероятность того, что целевое значение находится в каждой из областей. Если на одной диаграмме показаны несколько функций плотности, то таблица имеет отдельную строку для вероятностей, связанных с каждой функцией плотности. Каждая из этих опорных линий имеет ползунок (перевернутый треугольник), который позволяет легко переместить ее. Ряд дополнительных функций доступны при нажатии кнопки **Параметры диаграмм** на диаграмме. В частности, вы сможете явно задать позиции ползунков, добавить фиксированные опорные линии и изменить вид диаграммы с непрерывной кривой на гистограмму и наоборот. Дополнительную информацию смотрите в разделе “Опции диаграмм”.

Кумулятивная функция плотности для непрерывных целевых переменных. Эта диаграмма имеет такие же две перемещаемые вертикальные опорные линии и связанную таблицу, описанную для функции плотности вероятности на диаграмме выше. На ней также предоставлен доступ к диалоговому окну Параметры диаграмм, которое позволяет явно задать положения ползунков, добавлять фиксированные опорные линии и указывать порядок вывода на экран кумулятивной функции распределения: восходящий (по умолчанию) или нисходящий. Дополнительную информацию смотрите в разделе “Опции диаграмм”.

Столбчатые диаграммы для категориальных целевых значений с итерациями анализа чувствительности. Для категориальных целевых значений с итерациями анализа чувствительности результаты для прогнозной категории целевых значений показаны в виде кластеризованной столбчатой диаграммы, которая включает результаты всех итераций. Диаграмма включает раскрывающийся список, который позволяет выполнить кластеризацию по категории или по итерации. При использовании двухэтапного кластерного анализа и кластерного анализа методом k-средних можно выбрать кластеризацию по номеру кластера или итерации.

Ящичные диаграммы для нескольких целевых значений с итерациями анализа чувствительности. В случае прогнозных моделей с несколькими количественными целевыми значениями и итерациями анализа чувствительности если выбрать вывод на экран ящичных диаграмм для всех целевых значений на одной диаграмме, то создастся кластеризованная ящичная диаграмма.. Диаграмма включает раскрывающийся список, который позволяет выполнить кластеризацию по целевому значению или по итерации.

Опции диаграмм

Диалоговое окно Опции диаграмм позволяет настроить вывод на экран активированных диаграмм функций плотности вероятности и кумулятивных функций распределения, созданных из имитации.

Вид. Выпадающий список **Вид** применяется только к диаграмме функции плотности вероятности. Оно позволяет изменить форму вида диаграммы с непрерывной кривой на гистограмму. Эта функция недоступна, если на одной диаграмме показано несколько функций плотности. В этом случае функции плотности можно просмотреть только как непрерывные кривые.

Порядок. Выпадающий список **Вид** применяется только к диаграмме кумулятивной функции распределения. Оно указывает порядок вывода на экран функции: восходящий (по умолчанию) или убывающий. При выводе на экран в убывающем порядке значение функции в данной точке на горизонтальной оси является вероятностью того, что целевое значение находится справа от этой точки.

Положения ползунка. Позиции опорных линий ползунка можно задать явно. Для этого нужно ввести значения в текстовые поля Нижняя и Верхняя . Можно удалить левую линию и задать отрицательную бесконечность, выбрав **-Бесконечность**, а также удалить правую линию и задать положительную бесконечность, выбрав **Бесконечность**.

Опорные линии. Вы можете добавлять различные неподвижные вертикальные опорные линии для функций плотности вероятности и кумулятивных функций распределения. Если на одной диаграмме показаны несколько функций (из-за нескольких целевых значений или результатов из итераций анализа чувствительности), можно указать конкретные функции, к которым эти линии применяются.

- **Сигмы.** Можно добавить опорные линии с амплитудой указанного количества стандартных отклонений от среднего целевого значения.
- **Процентили.** Можно добавить опорные линии в одном или двух значениях процентилей распределения для целевого значения в текстовых полях Нижняя и Верхняя. Например, значение 95 в текстовом поле Верхняя представляет 95-ый процентиль, который является значением, ниже которого попадают 95 % наблюдений. Точно так же, значение 5 в текстовом поле Нижняя представляет 5-ый процентиль, который является значением, ниже которого попадают 5% наблюдений.
- **Настраиваемые позиции.** Можно добавить опорные линии в указанных значениях по горизонтальной оси.

Опорные линии меток. Этот параметр определяет, применяются ли метки к выбранным опорным линиям.

Чтобы удалить опорную линию, очистите соответствующий выбор в диалоговом окне Параметры диаграмм и нажмите кнопку **Продолжить**.

Глава 35. Геопространственное моделирование

Методы геопространственного моделирования предназначены для обнаружения шаблонов в данных, в которых содержится геопространственный компонент (карта). Мастер по геопространственному моделированию предоставляет методы анализа геопространственных данных, как с временным компонентом, так и без.

Найдите связи на основе событий и геопространственных данных (геопространственные правила связывания)
Пользуясь геопространственными правилами, можно искать шаблоны в данных с учетом как пространственных, так и не пространственных свойств. Например, иногда можно найти шаблоны в данных о преступлениях по атрибутам положения и демографическим атрибутам. По этим шаблонам можно построить правила, предсказывающие вероятные места определенных типов преступлений.

Сделайте прогнозы по временным рядам и геопространственным данным (пространственно-временное предсказание)
В пространственно-временном предсказании используются данные, содержащие информацию о положении, входные поля для прогноза (предикторы), одно или несколько полей времени и целевое поле. В этих данных для каждого положения в каждом интервале времени по каждому предиктору и назначению есть значительный ряд значений.

Использование Мастера по геопространственному моделированию

1. Выберите в меню:
   Анализ > Пространственное моделирование и моделирование во времени > Пространственное моделирование
2. Выполните указания в мастере.

Примеры

В системе справки доступны подробные примеры.

- Геопространственные правила связывания: **Справка > Темы > Примеры анализа > База статистики > Пространственные правила связывания**
- Пространственно-временное предсказание: **Справка > Темы > Примеры анализа > База статистики > Пространственно-временное предсказание**

Выбор карт

В геопространственном моделировании может использоваться один или несколько источников данных карты. Источники данных карты содержат информацию, определяющую географические области и другие географические объекты, например, дороги и реки. Многие источники карт содержат также демографические и иные описательные данные и данные о событии, например, отчеты о преступности или уровень безработицы. Можно использовать ранее определенный файл спецификации карты или определить спецификации карты здесь и сохранить эти спецификации для использования в дальнейшем.

Загрузите спецификацию карты
Загружает ранее определенный файл спецификаций карты (.mplan). Источники данных карты, определяемые здесь, можно сохранить в файле спецификаций. Для пространственно-временного предсказания, если выбрать файл спецификаций карты, в котором указывается несколько карт, вам предлагается выбрать одну карту из файла.

Добавить файл карты
Добавьте файл начертаний ESRI (файл .shp) или архив .zip, содержащий файл начертаний ESRI.

- Соответствующий файл .dbf должен находиться в том же положении, что файл .shp, и корневое имя этого файла должно быть таким же, как у файла .shp.

- Если файл - архив .zip, корневые имена файлов .shp и .dbf должны быть те же, что у файла архива .zip.
- Если нет соответствующего файла проекции (.prj), выводится приглашение выбрать систему проекции.

Взаимосвязь
Для геопространственных правил связывания в этом столбце определяется, как события соотносятся с объектами на карте. Этот параметр недоступен для пространственно-временного предсказания.

Переместить вверх, переместить вниз
Порядок слоев элементов карты определяется их порядком в этом списке. Первая карта в списке - это нижний слой.

Выбор карты

Для пространственно-временного предсказания, если выбрать файл спецификаций карты, в котором указывается несколько карт, вам предлагается выбрать одну карту из файла. Использование нескольких карт при пространственно-временном предсказании не поддерживается.

Геопространственная взаимосвязь

Для геопространственных правил связывания в диалоговом окне Геопространственная взаимосвязь определяется, как события соотносятся с объектами на карте.

- Этот параметр применим только к геопространственным правилам связывания.
- Этот параметр воздействует только на источники данных, связанные с картами, заданными как данные контекста на шаге выбора источников данных.

Взаимосвязь

Близко
Событие происходит близко к указанной точке или области на карте.

В Событие происходит в указанной области на карте.

Содержит
Область события содержит объект контекста карты.

Пересекает
Положения, в которых линии или области из разных карт пересекают друг друга.

Пересечение
Для нескольких карт - положения, где линии (для дорог, рек, железнодорожных путей) из нескольких карт пересекаются друг с другом.

К северу от, к югу от, к востоку от, к западу от
Событие происходит в области к северу, югу, востоку или западу от указанной точки на карте.

Задание системы координат

Если при карте нет файла проекции (.prj) или в качестве набора координат определены два поля из источника данных, нужно задать систему координат.

Географические данные по умолчанию (долгота и широта)
Система координат - долгота и широта.

Простые декартовы координаты (X и Y)
Система координат - простые координаты X и Y.

Использовать как известный ID (Well Known ID, WKID)
"Известный ID" для общепринятых проекций.

Использовать имя системы координат

Система координат основана на именованной проекции. Имя заключается в скобки.

Задание проекции

Если систему проекции нельзя определить из информации, предоставленной с картой, нужно ее указать. Наиболее общая причина этого условия - отсутствие файла проекции (.prj), связанного с картой, или невозможность использования существующего файла проекции.

- **Город, район или страна (Меркатор)**
- **Большая страна, несколько стран или континенты (Тройная Винкеля)**
- **Область, близкая к экватору (Меркатор)**
- **Область, близкая к одному из полюсов (Стереографическая)**

Проекция Меркатора - это общая проекция, используемая на многих картах. Эта проекция рассматривает земной шар как цилиндр, раскатанный по плоской поверхности. Проекция Меркатора искажает размер и форму больших объектов. Это искажение увеличивается от экватора к полюсам. В тройной проекции Винкеля и в стереографической проекции делаются корректировки, учитывающие тот факт, что карта представляет собой трехмерную сферу, показанную в двух измерениях.

Система проекции и координат

Если выбрать несколько карт, и эти карты различаются по системам проекции и координат, нужно выбрать карту с той системой проекции, которую вы хотите использовать. Эти система проекции будут использоваться для всех карт, которые будут объединяться с этой картой в объекте вывода.

Источники данных

Источником данных может быть файл dBase, предоставленном вместе с файлом начертаний, файл данных IBM SPSS Statistics или открытый набор данных в текущем сеансе.

Данные контекста. В данных контекста указываются объекты на карте. Данные контекста могут содержать также поля для использования в качестве входных данных модели. Чтобы использовать файл dBase (.dbf) контекста, связанный с файлом форм (.shp) карты, файл dBase контекста должен располагаться там же, где файл форм, и у него должно быть такое же корневое имя. Например, если файл форм - `geodata.shp`, файл dBase должен называться `geodata.dbf`

Данные события. Данные события содержат информацию о происходящих событиях, например, о преступлениях или авариях. Опция доступна только для геопространственных правил связывания.

Плотность точек. Интервал времени и данные о координатах для ядерных оценок плотности. Опция доступна только для пространственно-временного предсказания.

Добавить. Открывает диалоговое окно для добавления источников данных. Источником данных может быть файл dBase, предоставленном вместе с файлом начертаний, файл данных IBM SPSS Statistics или открытый набор данных в текущем сеансе.

Связать. Открывает диалоговое окно для указания идентификаторов (координат или ключей), используемых для связывания данных с картами. Каждый источник данных должен содержать один или несколько идентификаторов, связывающих данные с картой. Файлы dBase, поступающие вместе с файлом начертаний, обычно содержат поле, которое автоматически используется как идентификатор по умолчанию. Для других источников данных нужно указать поля, используемые как идентификаторы.

Проверить ключи. Открывает диалоговое окно для проверки соответствия между ключами карты и источника данных.

Геопространственные правила связывания

- Хотя бы один источник данных должен быть источником данных о событии.
- Все источники данных о событии пользуются одними и теми же идентификаторами связи с картой: значениями координат или ключей.
- Если источники данных о событии связываются с картами при помощи значений ключей, то все источники событий должны пользоваться тем же типом объектов карты (например, многоугольниками, точками, линиями).

Пространственно-временное предсказание

- Должен существовать источник данных контекста.
- Если есть только один источник данных (файл данных, не связанный с картой), он должен включать в себя значения координат.
- Если у вас два источника данных, один источник данных должен представлять собой данные контекста, а другой - данные плотности точек.
- Нельзя включать более двух источников данных.

Добавить источник данных

Источником данных может быть файл dBase, предоставленном вместе с файлом начертаний и файлом контекста, файл данных IBM SPSS Statistics или открытый набор данных в текущем сеансе.

Можно добавить один и тот же источник данных несколько раз, если с ним нужно использовать различные пространственные ассоциации.

Связывание данных и карт

Каждый источник данных должен содержать один или несколько идентификаторов, связывающих данные с картой.

Координаты

Источник данных содержит поля, представляющие декартовы координаты; выберите поля, представляющие координаты X и Y. Для правил геопространственного связывания можно также выбрать координату Z.

Значения ключей

Значения ключей в полях в источнике данных соответствует выбранным ключам карты. Например, на карте областей каждая область может быть помечена идентификатором имени (ключом карты). Этот идентификатор соответствует полю в данных, которое также содержит имена областей (ключ данных). Поля сопоставляются ключам карты с учетом порядка в каждом из двух списков.

Проверка ключей

Диалоговое окно Проверка ключей содержит сводку сопоставления записей карты и источника данных с учетом выбранных ключей-идентификаторов. Если некоторые значения ключей данных не сопоставлены значениям ключей, вы можете сделать это вручную.

Геопространственные правила связывания

Для геопространственных правил связывания после определения карт и источников данных в мастере остается выполнить следующие действия:

- При наличии нескольких источников данных событий определите способ их слияния.
- Выберите поля для использования в качестве условий и предсказаний в анализе.

Дополнительно можно выполнить следующие действия.

- Выберите другие опции вывода.
- Сохраните файл оценки модели.

- Создайте новые поля для предсказанных значений и правил в используемых в модели источниках данных.
- Настройте параметры для построения правил связывания.
- Настройте параметры категоризации и агрегации.

Определить поля данных о событии

Для геопространственных правил связывания, если есть несколько источников данных о событии, такие источники объединяются.

- По умолчанию это только поля, общие для всех источников данных о событии.
- Можно вывести список общих полей, полей для конкретного источника данных или полей из всех источников данных и выбрать нужные поля.
- Для общих полей **Тип** и **Измерение** должны быть одинаковы для всех источников данных. В случае конфликтов можно указать нужные тип и уровень измерения для каждого общего поля.

Выбрать поля

Список допустимых полей включает в себя поля из источников данных о событии и поля из источников данных контекста.

- Списком выводимых полей можно управлять, выбрав источник данных в списке **Источники данных**.
- Нужно выбрать хотя бы два поля. Хотя бы одно должно быть условием и хотя бы одно должно быть прогнозом. Есть ряд способов выполнить это требование, включая выбор двух полей в списке **Оба типа (условие и прогноз)**.
- Правилами связывания предсказываются значения полей прогноза с учетом значений полей условия. Например, в правиле "Если x=1 и y=2, то z=3" значения x и y - это условия, а значение z - это прогноз.

Объект вывода

Таблицы правил

В каждой таблице правил показаны лучшие правила и такие значения, как достоверность, поддержка правила, прирост, поддержка условия и внедряемость. Каждая таблица сортируется по значениям выбранного критерия. Можно вывести все правила или лучшие **N** правил с учетом выбранного критерия.

Сортируемое облако слов

Список лучших правил с учетом значений выбранного критерия. Размер текста показывает относительную важность правила. Интерактивный объект вывода содержит лучшие правила по таким показателям, как достоверность, поддержка правила, прирост, поддержка условия и внедряемость. Выбранным критерием задается тот список правил, который выводится по умолчанию. Можно выбрать другой критерий интерактивно в выводе. **Максимум выводимых правил** задает число правил в объекте вывода.

Карты Интерактивная полосчатая диаграмма и карта лучших правил с учетом выбранного критерия. В каждом интерактивном объекте вывода содержатся лучшие правила по таким показателям, как достоверность, поддержка правила, прирост, поддержка условия и внедряемость. Выбранным критерием задается тот список правил, который выводится по умолчанию. Можно выбрать другой критерий интерактивно в выводе. **Максимум выводимых правил** задает число правил в объекте вывода.

Таблицы информации модели

Преобразования полей.
Описывает преобразования, применяемые к полям, которые используются в анализе.

Сводка записей.
Число и процент включенных и исключенных записей.

Статистика правил.
Сводная статистика для таких показателей, как поддержка условия, достоверность,

поддержка правила, прирост и внедряемость. Статистические показатели включают в себя среднее, минимум, максимум и стандартное отклонение.

Часто встречаемые элементы.
Элементы, которые встречаются чаще всего. Элемент включается в условие или предсказание в правиле. Например, возраст < 18 или пол=женский.

Часто встречаемые поля.
Поля, которые встречаются в правилах чаще всего.

Исключенные входные данные.
Поля, исключенные из анализа, и причина исключить правило.

Критерий для таблиц правил, облака слов и карт

Достоверность.
Процент верных предсказаний правила.

Поддержка правила.
Процент наблюдений, для которых значение правила равно true. Например, пусть правило таково: "Если x=1 и y=2, то z=3". Поддержка правила - это фактически процент наблюдений в данных, где x=1, y=2 и z=3.

Подъем.
Прирост - это мера того, насколько сильно правило улучшает предсказание по сравнению со случайным выбором. Он вычисляется как отношение числа верных прогнозов к общему числу наблюдений предсказываемого значения. Это значение должно быть больше 1. Например, если предсказываемое значение наблюдалось в 20% случаев, а достоверность в прогнозе составила 80%, то значение прироста равно 4.

Поддержка условия.
Процент наблюдений, для которых имеет место условие правила. Например, пусть правило таково: "Если x=1 и y=2, то z=3". Поддержка условия - это доля наблюдений в данных, для которых x=1 и y=2.

Внедряемость.
Процентная доля неправильных предсказаний, когда значения условий - true. Значение внедряемости равно результату умножения значения (1-достоверность) на значение поддержки условия (вторым множителем может быть поддержка условия минус поддержка правила).

Сохранение

Сохранить карту и данные контекста как спецификацию карты
Сохраните спецификации карты во внешний файл (.mplan). Этот файл спецификации карты можно загрузить в мастер для последующего анализа. Кроме того, файл спецификации карты можно использовать при помощи команды `SPATIAL ASSOCIATION RULES`.

Скопировать все карты и файлы данных в спецификацию
Данные из файлов начертаний карты, файлов внешних данных и наборов данных, используемых в спецификации карты, сохраняются в файле спецификации.

Скоринг
Сохраняет лучшие значения правил, значения достоверности правил и значения числовых ID для правил как новые поля в указанном источнике данных.

Источник данных для оценки
Источник или источники данных, где создаются новые поля. Если источник данных не открыт в текущем сеансе, он будет открыт в текущем сеансе. Чтобы сохранить новые поля, нужно явным образом сохранить измененный файл.

Значения назначения
Создать новые поля для выбранных целевых полей (полей прогноза).

- Два новых поля создается для каждого целевого поля: предсказанное значение и значение достоверности.
- Для непрерывных (количественных) целевых полей предсказанное значение - это строка, описывающая диапазон значений. Значение в формате "(значение1, значение2]" значит "больше значение1 и меньше или равно значение2."

Число лучших правил

Создайте новые поля для указанного числа лучших правил. Для каждого правила создается три новых поля: значение правила, значение достоверности и значение числового ID для правила.

Префикс имени

Префикс для имен новых полей.

Построение правил

Параметры построения правил задают критерии для сгенерированных правил связывания.

Элементов на правило

Число значений полей, которые можно включить в условия и предсказания правила. Общее число элементов не может превышать 10. Например, в правиле "Если x=1 и y=2, то z=3" есть два элемента условия и один элемент прогноза.

Максимальное число предсказаний.

Максимальное число значений полей, которое может войти в предсказания правила.

Максимальное число условий.

Максимальное число значений полей в условиях для одного правила.

Исключить пару

Исключает вхождение указанной пары полей в одно и то же правило.

Критерии правил

Достоверность.

Минимальная достоверность правила, при которой оно может быть включено в объект вывода. Достоверность - это процент верных предсказаний.

Поддержка правила.

Минимальная поддержка правила, при которой оно может быть включено в объект вывода. Это значение представляет процент наблюдений, при которых значение правила равно true в данных наблюдения. Например, пусть правило таково: "Если x=1 и y=2, то z=3". Поддержка правила - это фактически процент наблюдений в данных, где x=1, y=2 и z=3.

Поддержка условия.

Минимальная поддержка условия правила, при которой правило может быть включено в объект вывода. Это значение представляет процент наблюдений, при которых имеет место условие правила. Например, пусть правило таково: "Если x=1 и y=2, то z=3". Поддержка условия - это процент наблюдений в данных, для которых x=1 и y=2.

Подъем.

Минимальный подъем правила, при которой оно может быть включено в объект вывода. Подъем - это мера того, насколько сильно правило улучшает предсказание по сравнению со случайным выбором. Он вычисляется как отношение числа верных прогнозов к общему числу наблюдений предсказываемого значения. Например, если предсказываемое значение наблюдалось в 20% случаев, а достоверность в прогнозе составила 80%, значение подъема - 4.

Рассматривать как одинаковые

Указывает пары полей, которые нужно рассматривать как одно и то же поле.

Разбивка по интервалам и агрегация

- Агрегирование необходимо, когда число записей в данных больше, чем число объектов на карте. Например, у вас есть записи данных для отдельных графств, а на карте представлены штаты.
- Можно задать метод вычисления сводной меры для непрерывных и порядковых полей. Номинальные поля агрегируются с учетом модального значения.

Количественный
Для непрерывных (количественных) полей сводная мера может быть средним, медианой или суммой.

Порядковый
Для порядковых полей суммарная мера может быть медианой, модой, наибольшим или наименьшим.

Число интервалов
Задает максимальное число интервалов для непрерывных (количественных) полей. Непрерывные поля всегда группируются или "категоризируются" по диапазонам значений. Например: меньше или равно 5, больше 5 и меньше или равно 10, больше 10.

Агрегировать карту
Применить агрегирование и к данным, и к картам.

Пользовательские параметры для конкретных полей
Вы можете переопределить сводную меру по умолчанию и количество интервалов для конкретных полей.

- Щелкните по значку, чтобы открыть диалоговое окно **Средство выбора полей**, и выберите поле, добавляемое в список.
- В столбце **Агрегирование** выберите сводную меру.
- В случае непрерывных полей нажмите кнопку в столбце **Интервалы** и задайте свое число интервалов для поля в диалоговом окне **Интервалы**.

Пространственно-временное предсказание

Для пространственно-временного предсказания после определения карт и источников данных остальные действия в мастере следующие:

- Задайте поле назначение, поля времени и необязательные предикторы.
- Определите интервалы времени или циклические периоды для полей времени.

Дополнительно можно выполнить следующие действия.

- Выберите другие опции вывода.
- Настройте параметры построения моделей.
- Настройте параметры агрегации.
- Сохраните предсказанные значения в наборе данных текущего сеанса или в файле данных формата IBM SPSS Statistics.

Выбрать поля

Список допустимых полей включает в себя поля из выбранных источников данных. Списком выводимых полей можно управлять, выбрав источник данных в списке **Источники данных**.

Назначение
Целевое поле - обязательное. Цель - это то поле, значения которого предсказываются.

- Целевое поле должно быть непрерывным (количественным), числовым полем.
- Если есть два источника данных, целью будут оценки ядерной плотности, а именем цели будет "Плотность". Этот выбор нельзя изменить.

Предикторы

Можно задать одно или несколько полей предикторов. Это необязательный параметр.

Поля времени

Нужно выбрать одно или несколько полей, которые представляют периоды, или выбрать **Циклические периоды**.

- Если есть два источника данных, нужно выбрать поля времени из обоих источников данных. Оба поля времени должны представлять один и тот же интервал.
- Для циклических периодов нужно задать поля, определяющие циклы периодичности на панели Интервал времени мастера.

Интервалы времени

Опции на этой панели учитывают, что на шаге выбора полей были выбраны **Поля времени** или **Циклический период**.

Поля времени

Выбранные поля времени. Если на шаге выбора полей выбрать одно или несколько полей времени, эти поля выводятся в данном списке.

Интервал времени. Выберите нужный интервал времени в списке. В зависимости от интервала времени можно задать также другие параметры, такие как интервал между наблюдениями (инкремент) или начальное значение. Этот интервал времени используется для всех выбранных полей времени.

- Процедура предполагает, что все наблюдения (записи) представляют интервалы с одинаковыми промежутками.
- С учетом выбранного интервала времени процедура может обнаружить пропущенные наблюдения или несколько наблюдений в одном интервале времени, которые нужно объединить. Например, если интервал времени - это Дни, и после даты 2014-10-27 следует 2014-10-29, пропущенным считается наблюдение за дату 2014-10-28. Например, если интервал времени - это Месяцы, то несколько дат одного месяца агрегируются.
- Для некоторых интервалов времени дополнительным параметром могут определяться перерывы в интервалах с одинаковыми промежутками. Например, если интервал времени - это Дни, но допускаются только рабочие дни, можно указать, что в неделе есть только пять дней и она начинается в понедельник.
- Если выбранное поле времени не в формате даты или времени, для интервала времени автоматически задается значение **Периоды**, и оно не может быть изменено.

Циклические поля

Если на шаге выбора полей выбрать **Циклический период**, нужно указать поля, которыми определяются циклические периоды. Циклический период показывает повторяющиеся периодические изменения, например, число месяцев в году или число дней в неделе.

- Можно задать до трех полей, определяющих циклические периоды.
- Первое поле циклического периода представляет высший уровень цикла. Например, если учитываются циклические изменения в течение года, квартала и месяца, в первом циклическом поле представлен год.
- Длина цикла для первого и второго поля - это кратность периода относительно следующего уровня. Например, если циклические поля - это год, квартал и месяц, то длина первого цикла 4, а второго - 3.
- Начальное значение для второго и третьего циклического поля - это первое значение в каждом из этих циклических периодов.
- Длины циклов и начальные значения должны быть положительными целыми.

Агрегирование

- Если выбрать любые **Предикторы** на шаге выбора полей, можно выбрать метод агрегирования сводных предикторов.

- Агрегирование необходимо, если есть несколько записей в определенном временном интервале. Например, если временной интервал - месяц, то несколько дат в одном месяце агрегируются вместе.
- Можно задать метод вычисления суммарной меры агрегирования для непрерывных и порядковых полей. Номинальные поля агрегируются с учетом модального значения.

Количественный
Для непрерывных (количественных) полей суммарная мера может быть средним, медианой или суммой.

Порядковый
Для порядковых полей суммарная мера может быть медианой, модой, наибольшим или наименьшим.

Пользовательские параметры для конкретных полей
Вы можете переопределить суммарную меру агрегирования для конкретных предикторов.

- Щелкните по значку, чтобы открыть диалоговое окно **Средство выбора полей**, и выберите поле, добавляемое в список.
- В столбце **Агрегирование** выберите суммарную меру.

Объект вывода

Карты

Значения назначения.
Карта значений для выбранного целевого поля.

Корреляция
Карта корреляции.

Кластеры
Карта, на которой выделены кластеры положений, аналогичных друг другу. Карты кластеров доступны только для эмпирических моделей.

Порог сходства положения.
Сходство, требуемое для создания кластеров. Значение должно быть числом больше нуля и меньше 1.

Укажите максимальное количество кластеров.
Максимальное число кластеров для вывода.

Таблицы оценки моделей

Спецификации моделей.
Сводка спецификаций, используемых при выполнении анализа, включая целевые поля, входные поля и поля положения.

Сводка временной информации.
Показывает поля времени и интервалы времени, используемые в модели.

Критерий эффектов в структуре средних.
Выходной объект включает в себя значение статистики критерия, степени свободы и уровень значимости для модели и для каждого эффекта.

Структура средних для коэффициентов модели.
Объект вывода включает в себя значение коэффициента, стандартную ошибку, значение статистики критерия, уровень значимости и доверительные интервалы для каждого члена модели.

Коэффициенты авторегрессии.
Объект вывода включает в себя значение коэффициента, стандартную ошибку, значение статистики критерия, уровень значимости и доверительные интервалы для каждой задержки.

Критерии пространственной ковариации.
Для параметрических моделей на основе вариограммы показывает результаты критерия согласия для структуры пространственной ковариации. По результатам этого критерия можно определить, моделировать ли структуру пространственной ковариации параметрически, или использовать непараметрическую модель.

Параметрическая пространственная ковариация.
Для параметрических моделей на основе вариограммы показывает оценки параметра для параметрической пространственной ковариации.

Опции модели

Параметры модели

Автоматически включать свободный член
Включить в модель свободный член.

Максимальная задержка авторегрессии
Максимальная задержка авторегрессии. Номер должен представлять собой целое число от 1 до 5.

Пространственная ковариация
Задает метод оценки для пространственной ковариации.

Параметрическое
Метод оценки - параметрический. Метод может быть **Гауссов**, **Экспоненциальный** или **Показатель степени**. Для метода показатель степени можно задать значение **Степень**.

Непараметрическое
Метод оценки - не параметрический.

Сохранение

Сохранить карту и данные контекста как спецификацию карты
Сохраните спецификации карты во внешний файл (.mplan). Этот файл спецификации карты можно загрузить в мастер для последующего анализа. Кроме того, файл спецификации карты можно использовать при помощи команды `SPATIAL TEMPORAL PREDICTION`.

Скопировать все карты и файлы данных в спецификацию
Данные из файлов начертаний карты, файлов внешних данных и наборов данных, используемых в спецификации карты, сохраняются в файле спецификации.

Скоринг
Сохраняет предсказанные значения, дисперсию и верхнюю и нижнюю границу доверительного интервала для целевого поля в выбранном файле данных.

- Можно сохранить предсказанные значения в открытый набор данных в текущем сеансе или в файл данных формата IBM SPSS Statistics.
- Этот файл данных должен отличаться от источника данных, используемого в модели.
- Файл данных должен содержать все поля времени и предикторы, используемые в модели.
- Значения времени должны быть больше, чем используемые в модели значения.

Дополнительные параметры

Максимальное число наблюдений с пропущенными значениями (%)
Максимальный процент наблюдений с отсутствующими значениями.

Уровень значимости
Уровень значимости для выяснения, уместна ли параметрическая модель на основе вариограммы. Это значение должно быть больше 0 и меньше 1. Значение по умолчанию - 0,05. Уровень значимости

используется в критерии согласия для структуры пространственной ковариации. Определить, какую модель использовать, параметрическую или непараметрическую, можно с помощью статистики согласия.

Фактор неопределенности (%)

Фактор неопределенности - это значение в процентах, представляющее рост неопределенности для будущих предсказаний. Верхний и нижний пределы неопределенности прогноза увеличиваются на эту процентную долю при всяком шаге в будущее.

Готово

На последнем шаге работы с Мастером можно запустить модель или вставить сгенерированный командный синтаксис в окно синтаксиса. Сгенерированный синтаксис можно отредактировать и сохранить для последующего использования.

Замечания

Эта публикация разрабатывалась для продуктов и услуг, предлагаемых в США. Этот материал может быть доступен от IBM на других языках. Однако для его получения может понадобиться приобрести продукт или версию продукта на нужном языке.

IBM может не предоставлять в других странах продукты, услуги и аппаратные средства, описанные в данном документе. За информацией о продуктах и услугах, предоставляемых в вашей стране, обращайтесь к местному представителю IBM. Ссылки на продукты, программы или услуги IBM не означают и не предполагают, что можно использовать только указанные продукты, программы или услуги IBM. Разрешается использовать любые функционально эквивалентные продукты, программы или услуги, если при этом не нарушаются права IBM на интеллектуальную собственность. Однако ответственность за оценку и проверку работы любого продукта, программы или сервиса, не произведенного корпорацией IBM, лежит на пользователе.

IBM может располагать патентами или рассматриваемыми заявками на патенты, относящимися к предмету данного документа. Предъявление данного документа не предоставляет какую-либо лицензию на эти патенты. Вы можете послать письменный запрос о лицензии по адресу:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

По поводу лицензий, связанных с использованием наборов двухбайтных символов (DBCS), обращайтесь в отдел интеллектуальной собственности IBM в вашей стране или направьте запрос в письменной форме по адресу:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

КОРПОРАЦИЯ INTERNATIONAL BUSINESS MACHINES ПРЕДОСТАВЛЯЕТ ДАННУЮ ПУБЛИКАЦИЮ "КАК ЕСТЬ", БЕЗ КАКИХ-ЛИБО ЯВНЫХ ИЛИ ПОДРАЗУМЕВАЕМЫХ ГАРАНТИЙ, ВКЛЮЧАЯ, НО НЕ ОГРАНИЧИВАЯСЬ ТАКОВЫМИ, ПОДРАЗУМЕВАЕМЫЕ ГАРАНТИИ ОТСУТСТВИЯ НАРУШЕНИЙ, КОММЕРЧЕСКОЙ ПРИГОДНОСТИ ИЛИ СООТВЕТСТВИЯ КАКОЙ-ЛИБО КОНКРЕТНОЙ ЦЕЛИ. В некоторых странах для ряда сделок не допускается отказ от явных или предполагаемых гарантий; в таком случае данное положение к вам не относится.

Эта информация может содержать технические неточности и типографские ошибки. В представленную здесь информацию периодически вносятся изменения; эти изменения будут включаться в новые издания данной публикации. Фирма IBM может в любое время без уведомления вносить изменения и усовершенствования в продукты и программы, описанные в этой публикации.

Любые ссылки в данной информации на сайты, не принадлежащие IBM, приводятся только для удобства и никоим образом не означают поддержки этих сайтов. Материалы на этих сайтах не входят в число материалов по данному продукту IBM, и весь риск пользования этими сайтами несете вы сами.

IBM может использовать или распространять предоставленную вами информацию любым способом, как фирма сочтет нужным, без каких-либо обязательств перед вами.

Если обладателю лицензии на данную программу понадобится информация о возможности: (i) обмена данными между независимо разработанными программами и другими программами (включая данную) и (ii) совместного использования таких данных, он может обратиться по адресу:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

Такая информация может быть доступна при соответствующих условиях и соглашениях, включая в некоторых случаях взимание платы.

Описанную в данном документе лицензионную программу и все прилагаемые к ней лицензированные материалы IBM предоставляет на основе положений Соглашения между IBM и Заказчиком, Международного Соглашения о Лицензиях на Программы IBM или любого эквивалентного соглашения между IBM и заказчиком.

Упоминаемые данные о производительности и примеры клиентов представлены только для иллюстративных целей. Фактические результаты производительности могут быть иными в зависимости от определенных конфигураций и конкретных условий.

Информация, касающаяся продуктов других компаний (не IBM) была получена от поставщиков этих продуктов, из опубликованных ими заявлений или из прочих общедоступных источников. IBM не проводила тестирования этой продукции и не может подтвердить или опровергнуть информацию о точности ее работы и совместимости, а также другие заявления относительно продуктов других производителей (не IBM). Вопросы относительно возможностей продуктов других компаний (не IBM) следует адресовать поставщикам этих продуктов.

Утверждения, касающиеся намерений и планов IBM, могут быть изменены без предварительного предупреждения; они приведены здесь только для обозначения целей и задач IBM.

Эти сведения содержат примеры данных и отчетов, используемых в повседневных деловых операциях. Чтобы проиллюстрировать их настолько полно, насколько это возможно, данные примеры включают имена индивидуумов, названия компаний, брендов и продуктов. Все эти имена и названия вымышлены и любое их сходство с реальными именами и названиями компаний полностью случайно.

ЛИЦЕНЗИЯ НА КОПИРОВАНИЕ:

Эта информация содержит примеры исходных текстов прикладных программ, которые иллюстрируют приемы программирования на различных платформах. Разрешается копировать, изменять и распространять эти примеры программ в любой форме без оплаты фирме IBM для целей разработки, использования, сбыта или распространения прикладных программ, соответствующих интерфейсу прикладного программирования операционных платформ, для которых эти примера программ написаны. Эти примеры не были всесторонне проверены во всех возможных условиях. Поэтому IBM не может гарантировать их надежность, пригодность и функционирование. Примеры программ предоставляются "КАК ЕСТЬ", без каких-либо гарантий. IBM не несет никакой ответственности за какой либо ущерб, причиненный в результате использования этих программ.

Каждая копия или каждая часть этих примеров программ или работы, основанной на них, должна содержать следующее замечание об авторских правах:

© (название вашей компании) (год). Части этого кода получены из примеров программ IBM Corp.

Товарные знаки

IBM, логотип IBM, и ibm.com являются товарными знаками или зарегистрированными товарными знаками компании International Business Machines Corp., зарегистрированными во многих странах мира. Прочие наименования продуктов и услуг могут быть товарными знаками, принадлежащими IBM или другим компаниям. Текущий список товарных знаков IBM можно найти в Интернете в разделе "Copyright and trademark information" ("Информация об авторских правах и товарных знаках") по адресу www.ibm.com/legal/copytrade.shtml.

Adobe, логотип Adobe, PostScript и логотип PostScript являются либо зарегистрированными товарными знаками, либо товарными знаками корпорации Adobe Systems в Соединенных Штатах и/или других странах.

Intel, логотип Intel, Intel Inside, логотип Intel Inside, Intel Centrino, логотип Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium и Pentium являются товарными знаками или зарегистрированными товарными знаками компании Intel или ее дочерних компаний в Соединенных Штатах и других странах.

Linux является зарегистрированным товарным знаком Linus Torvalds в Соединенных Штатах и других странах.

Microsoft, Windows, Windows NT и логотип Windows являются товарными знаками корпорации Microsoft в Соединенных Штатах и других странах.

UNIX является зарегистрированным товарным знаком The Open Group в Соединенных Штатах и других странах.

Java и все основанные на Java товарные знаки и логотипы являются товарными знаками Oracle и/или его филиалов.

Индекс

Д

Е

З

И

К

Л

M

H

O

П

Р

С

T

У

Ф

Х

Ч

Ш

Э

Я

Напечатано в Дании