19. Обзор процедур начального анализа данных в SPSS

Автор: Александр
26.08.2014 13:13

SPSS является программным продуктом, предназначенным для выполнения всех этапов статистического анализа: от просмотра данных, создания таблиц и вычисления дескриптивных статистик до применения сложных статистических методов.

Смысл термина ***«анализ данных»*** неодинаково трактуется разными специалистами, в зависимости от различных областей применения. Некоторые считают, что анализ данных заканчивается с выводом дескриптивных статистик, графика или результата статистического вычисления. Для других он представляет собой последовательность шагов, каждый из которых может предполагать дальнейший анализ и появление новых задач для исследования. SPSS является универсальной статистической системой программ, поддерживающей процесс анализа данных на любом уровне и предназначенной для реализации полной последовательности шагов анализа данных: от просмотра данных, создания таблиц и вычисления дескриптивных статистик до сложного статистического анализа. Графические средства, встроенные в статистические процедуры, облегчают понимание данных и интерпретацию результатов анализа; они неоценимы для представления результатов анализа.

SPSS позволяет читать много различных типов файлов или вводить данные непосредственно в *Редакторе Данных*. Какой бы ни была структура вашего исходного файла данных, в *Редакторе Данных* он будет представлен в прямоугольном виде - так принято не только в SPSS, но и в большинстве других систем анализа данных, причем строки соответствуют наблюдениям, а столбцы - переменным. Наблюдение содержит информацию об одной единице анализа. Переменные содержат информацию, собранную об одном наблюдении. В данных часто встречаются так называемые
пропущенные значения -
они возникают из-за отсутствия ответов в некоторых наблюдениях, ошибок при измерениях или в результате неправильных вычислений. Каждое такое значение заменяется в SPSS специальным кодом - системным кодом пропущенного значения.

Результаты проведенного анализа появляются в навигаторе вывода SPSS. Большинство процедур Базового модуля представляют результаты в виде мобильных таблиц, которые можно редактировать различными способами с целью выделения наиболее важных результатов анализа. SPSS предоставляет пользователю большой набор возможностей для преобразования, отбора и сортировки данных. Термин «
преобразование»
охватывает очень большой набор функций, арифметических и логических операций, которые могут быть применены к данным. Чтобы отобрать подмножество наблюдений

для анализа, можно использовать значения переменных, функции и операции сравнения. Диалоговое окно
SelectCases
(Отбор наблюдений) в
меню Данные
Редактора Данных позволяет также отобрать случайную подвыборку или диапазон наблюдений для просмотра или анализа. Это может пригодится, например, в случае, когда нужно провести анализ для каждого из значений переменной отдельно.

Реальные данные редко удается собрать без проблем. Первым шагом после ввода данных является выявление ошибок при их записи и вводе, а также проверка соответствия данных предположениям, лежащим в основе планируемых методов анализа. В больших исследованиях проверка данных отнимает чрезвычайно много времени и сил.

Первый шаг при проверке данных обычно состоит в поиске значений, выходящих за разумные пределы значений переменной, - необходимо выяснить, действительно ли это выбросы или это ошибки.

Необходимо использовать процедуру *Частоты* для подсчета появления каждого отдельного значения . Так можно обнаружить опечатки и неожиданные значения следует искать также пропущенные значения, которые представлены как валидные.

Для количественных переменных используются гистограммы в процедурах *Частоты* или *Исследовать,*
а также ящичковые диаграммы и диаграммы "ствол-лист" в процедуре *Исследовать.*
Необходимо обращать внимание на выбросы, которые показывают диаграммы.

Зачастую выбросы легче обнаружить, если исследовать две или более переменных одновременно. Для дискретных данных неправдоподобные или нежелательные комбинации значений могут быть выявлены с помощью таблиц сопряженности.

Автор: Александр
26.08.2014 13:13

Распределение данных может оказаться не таким, как предполагалось - не похожим на нормальное и даже несимметричным. Если распределения переменных сильно асимметричны, использование процедуры *Регрессия* для предсказания одной переменной с помощью набора других переменных может привести к неадекватным результатам. Иногда эту проблему можно преодолеть, используя логарифмическое преобразование.

Для проверки распределения можно построить гистограммы с наложенными нормальными кривыми, используя процедуру *Частоты* или графики из меню *Графики;* а также процедуру
Исследовать
или P-P-графики (P-P plots) из меню
Графики
для построения графиков на вероятностной бумаге. Такие графики можно использовать для сравнения эмпирического распределения не только с нормальным, но и с несколькими другими стандартными видами распределений; при большом объеме данных сравнить величины среднего, 5%-го усеченного среднего и медианы. Если они сильно различаются, распределение асимметрично в качестве формального теста нормальности можно использовать критерии Колмогорова-Смирнова или Шапиро-Уилка процедуры
Исследовать.

Если сравнивать групповые средние, проблем может возникнуть еще больше. Например, при проведении дисперсионного анализа уровни значимости могут оказаться искаженными, в тех случаях когда распределения в сравниваемых группах значительно отклоняются от нормального или их разбросы сильно различаются (то есть нарушается предположение о равенстве дисперсий). Для сравнения эмпирических распределений с нормальным и для сопоставления разбросов распределений внутри групп используют ящичковые диаграммы.

Дескриптивные статистики могут полностью исчерпать потребности текущего исследования, но могут и оказаться первым шагом в изучении и понимании нового набора данных. Перед тем как начать описание данных (положение центра распределения, его разброс и т.п.), следует определить *типы имеющихся переменных*.

Подавляющее большинство статистических показателей разработаны для количественных переменных. В частности, вычисление среднего и стандартного отклонения допустимо для количественных переменных с нормальным распределением.

Однако для реальных данных предположение о нормальности часто не выполняется.

Для проверки гипотез о средних значениях количественных переменных предназначены t-критерий и дисперсионный анализ (ANOVA). С их помощью можно сделать выводы о характеристиках популяции по статистикам, описывающим выборочные данные. Эти критерии выбираются в меню *Сравнение средних* и *Общая линейная модель.*

Для данных с распределениями, значительно отклоняющимися от нормального, более подходящими могут оказаться непараметрические критерии. Некоторые из критериев ориентированы на обработку ранговых данных (во время вычисления статистик таких критериев SPSS преобразует данные в ранги). Применять непараметрические тесты для спасения данных нужно осторожно. Если данные не удовлетворяют предположениям, необходимым *t*-критерию или дисперсионному анализу, перед обращением к непараметрической статистике следует попробовать преобразование данных. Хотя непараметрические критерии не требуют нормальности, они, как и их параметрические аналоги, все же основываются на некоторых предположениях. Например, критерий Манна-Уитни предполагает, что формы сравниваемых распределений сходны. Кроме того, если *на самом деле* популяции различаются, для доказательства этого различия с помощью непараметрической процедуры может потребоваться большая выборка, чем для критерия, основанного на предположении о нормальности распределения.

SPSS предлагает три типа *t*-критериев. Выбирать нужный следует в зависимости от того, что именно сравнивает пользователь.

Дисперсионный анализ применяется для тех же целей, что и двухвыборочный *t*-критерий, но для большего числа выборок. Этот метод позволяет сравнить вариабельность выборочных средних с разбросом наблюдений в каждой из групп. Нулевая гипотеза заключается в том, что выборки составлены из популяций с *равными* средними.

Для однофакторного дисперсионного анализа (One-Way ANOVA) группы или ячейки, определяются уровнями одного группирующего фактора с двумя или более уровней. В многофакторной процедуре ANOVA группы определяются уровнями двух или более факторов. Например, если объекты исследования группируются по *полу* (мужской,

женский) и *мест у проживания* (Москва, Орел, Смоленск), получается шесть групп: мужчины из Москвы, женщины из Москвы, мужчины из Орла, женщины из Орла и т.д. Полная вариация зависимой переменной делится на составляющие - для *пола,* для *места проживания* и для их взаимодействия. Базовый модуль SPSS обеспечивает три процедуры дисперсионного анализа: средние, однофакторный дисперсионный анализ, многофакторный дисперсионный анализ.

В некоторых ситуациях ковариата (или, на языке регрессионного анализа, *независимая* переменная) может вносить дополнительный вклад в изменчивость зависимой переменной. При анализе ковариаций изменчивость зависимой переменной корректируется по вкладу ковариаты.

При выборе индикаторов, измеряющих зависимости между переменными, необходимо принимать во внимание типы исследуемых переменных. Если переменные дискретны, найти соответствующие меры можно в процедуре *Таблицы сопряженности.* Если переменные количественные, причем распределение их значений можно считать нормальным, можно использовать линейную модель в процедуре *Регрессия* или корреляцию Пирсона в процедуре *Парные корреляции.* Если предположение о нормальности распределении не кажется правдоподобным, следует использовать корреляцию Спирмена.

Для двумерных частотных таблиц наблюдений, соответствующих сочетанию значений двух дискретных переменных, процедура *Таблицы сопряженности* предлагает 22 критерия значимости и мер связи. Каждый критерий относится к определенным типам таблиц (с определенным числом строк и столбцов); несколько критериев подходят для упорядоченных категорий.

Коэффициент корреляции является мерой линейной связи между двумя количественными переменными. Простая регрессия представляет собой другой подход к той же проблеме. Корреляционная матрица отображает статистики для множества переменных попарно, а многомерная регрессия характеризует линейную связь между

одной переменной и подмножеством других переменных.

Корреляция по Приснону доступна в процедурах *Парные корреляции, Частные корреляции, Регрессия* и *Таблицы сопряженности.*
Данные должны иметь нормальное распределение. В тех случаях, когда это не так, в процедурах
Парные корреляции
и
Таблицы сопряженности
используют корреляцию по Спирмену. При вычислении этой статистики каждое значение переменной заменяется на его ранг в совокупности всех значений (с поправками, если встречаются одинаковые значения).

Регрессия дает количественное выражение линейной зависимости между переменными, когда изменение значений одной переменной зависит от изменения значений нескольких других переменных. Наиболее простым видом линейной зависимости является уравнение прямой: Y=A+BX.

Для оценки того, насколько хорошо прямая линия описывает имеющуюся зависимость, полезна диаграмма рассеяния. Линия представляет собой линию наилучшего соответствия, оцененной с помощью регрессионной процедуры.

Решая прикладную задачу, исследователь может не знать, какое именно множество из переменных следует включить в многомерную регрессионную модель, и, возможно, захочет отделить важные переменные от тех, которые несущественны для предсказания. В процедуре *Регрессия* пользователь может выбрать одну из нескольких стратегий включения и исключения переменных по одной в каждый момент времени в пошаговом режиме.

Графическое представление результатов полезно на всех стадиях анализа. После того как подбрана регрессионная модель для данных, следует изучить остатки, предсказанные значения и диагностические индикаторы. Последние полезны для определения выбросов и отклонений от предположений, лежащих в основе анализа.

Автор: Александр
26.08.2014 13:13

В Базовый модуль SPSS входят кластерный, дискриминантный и факторный анализы. Эти процедуры полезны для выявления групп.

Кластерный анализ является многофакторной процедурой для обнаружения группировок в данных. При использовании процедуры *k*-средних и иерархической процедуры кластеры образуются группами наблюдений. Иерархическая процедура может быть использована также для формирования групп переменных. Кластеризация является хорошим методом, если необходимо разбить данные на классы или когда данные неоднородны, и надо увидеть, существуют ли явные группы.

Для классификации наблюдений может быть использован также дискриминантный анализ. В нем идет работа с выборкой наблюдений, принадлежность которых классам уже известна. Процедура анализа позволяет найти линейные комбинации переменных, наилучшим образом характеризующие различия между группами (эти комбинации далее могут быть использованы для классификации новых наблюдений). Для того чтобы определить переменные, которые максимально полно описывают различие между группами, их можно вводить в функцию в пошаговом режиме.

Факторный анализ подходит для выявления групп коррелированных числовых переменных. Можно изучать корреляцию большого числа переменных, группируя переменные в факторы. Переменные в пределах каждого фактора коррелированы друг с другом сильнее, чем с переменными из других факторов. Возможно также интерпретировать каждый фактор в соответствии со смыслом переменных и свести большое количество переменных к небольшому числу факторов. Факторные нагрузки могут быть использованы в качестве данных для *t*-критерия, регрессии и т.д.