Анализ качественных переменных

Структура лекции

1. Таблицы сопряженности
2. Критерий Хи-квадрат
3. Логлинейный анализ таблиц сопряженности

Объекты исследования обладают несколькими признаками.

Вопрос: насколько эти признаки связаны между собой?

Можно ли по степени выраженности одного признака судить о выраженности другого, либо все-таки следует считать эти признаки проявляющимися независимо (в вероятностном смысле)?

Сначала решается более простая задача: проверить, существует ли вообще какая-либо связь между этими признаками, или же они ведут себя независимо друг от друга?

Статистический способ ответа основан на изучении выборки.

Таблицы сопряженности служат для описания связи двух или более номинальных (категориальных переменных).

Анализ таблиц сопряженности:

1. Составление таблиц сопряженности признаков (перекрестных таблиц);

2. Проверка гипотезы независимости переменных.

Таблицы сопряженности
Кросстабуляция (Crosstabulations)

Для описания двухвходовых (многомерных) таблиц используемые **термины**:

Факторы (признаки) – переменные, табулированные в таблицы;

Уровни – значения факторов.

A\ B	B_1	B_2	B_j	B_s	
A_1	n_{11}	n_{12}	n_{1j}	n_{1s}	$\sum_{j=1}^{s} n_{1j}$
A_2	n_{21}	n_{22}	n_{2j}	n_{2s}	$\sum_{j=1}^{s} n_{2j}$
A_i	n_{i1}	n_{i2}	n_{ij}	n_{is}	$\sum_{j=1}^{s} n_{ij}$
A_r	n_{r1}	n_{r2}	n_{rj}	n_{rs}	$\sum_{j=1}^{s} n_{rj}$
	$\sum_{i=1}^{r} n_{i1}$	$\sum_{i=1}^{r} n_{i2}$	$\sum_{i=1}^{r} n_{ij}$	$\sum_{i=1}^{r} n_{is}$	n

Для проверки гипотез о зависимости качественных переменных, измеряемых по номинальной шкале, используется тест Хи-квадрат.

Для применения метода требуется выполнение двух условий:

1. Набор данных представляет случайную выборку из рассматриваемой генеральной совокупности;
2. Для каждой комбинации категорий ожидаемое количество наблюдений в ячейке не меньше 5. Если это условие нарушается, надо перекодировать переменные, объединяя категории так, чтобы условие начало выполняться. Поскольку при всяком объединении теряется информация, желательно сделать изменения минимальными.

Основная гипотеза: переменные независимы

Альтернативная гипотеза: переменные зависимы

Идея метода

Основана на теореме (К. Пирсон, Р. Фишер).

Если верна модель, по которой рассчитаны теоретические частоты T , то при неограниченном росте числа наблюдений распределение случайной величины X^2 стремится к распределению хи-квадрат. Число степеней свободы этого распределения определяется как разность между числом событий и числом связей, налагаемых моделью.

В этой теореме

T- ожидаемые (теоретические) частоты,

H – наблюдаемые частоты,

$$X^2 = \sum \frac{(H-T)^2}{T}$$

Если модель правильно описывает действительность, числа H и T должны быть близки друг к другу.

Логлинейный анализ таблиц сопряженности

1. Понятие логлинейной модели

2. Логлинейный метод подбора модели

Понятие логлинейной модели

Логлинейная модель – множественная регрессионная модель, в которой категориальные переменные и их взаимодействия выступают в качестве предикторов, а роль зависимой переменной играет натуральный логарифм частот категорий. Использование логарифмической меры обусловливает линейность модели.

В этом уравнении частота – это частота текущей ячейки частотной таблицы, *λ*- воздействие со стороны одной или более независимых переменных, *μ*- общее среднее воздействия, *A*, *C*, *Y* – переменные агрессия, условия, симпатия:

$$\ln(\textit{частота}) = \mu + \lambda A + \lambda C + \lambda \mathrm{Y} + \lambda \mathrm{A} \cdot \mathrm{C} + \lambda \mathrm{A} \cdot \mathrm{Y} + \lambda \mathrm{C} \cdot \mathrm{Y} + \lambda \mathrm{A} \cdot \mathrm{C} \cdot \mathrm{Y}$$

Модель называется насыщенной, если она содержит все предикторы и их возможные взаимодействия.

Существуют более предпочтительные альтернативы в виде ненасыщенных моделей, которые отражают лишь статистически значимые главные эффекты и взаимодействия переменных.

Подменю Логлинейный анализ содержит три команды.

1. Общий — эта команда допускает вхождение в модель любых факторов и их взаимодействий и предполагает, что исследователь перед проведением анализа уже имеет гипотезы о составе модели.
2. Логит — применение этой команды позволяет рассматривать дихотомические переменные как зависимые, а одну (или более) категориальную переменную как независимую. При этом зависимая дихотомическая переменная используется не для прогнозирования частот категорий, а для разделения всех категорий на две группы. ^ 3. Подбор модели — эта команда позволяет из всех возможных ненасыщенных моделей подобрать ту, которая в наибольшей степени соответствует исходным данным. Подбор осуществляется, как правило, автоматически. В результате выявляется совокупность значимых связей между категориальными переменными и вычисляются параметры μ и λ логлинейной модели.

Логлинейный метод подбора модели

Теоретически из насыщенной модели можно удалить любые элементы, получив произвольную ненасыщенную модель.
Далее можно проверить состоятельность этой модели и в случае несоответствия ее исходным данным перейти к анализу другой ненасыщенной модели.
Предпочтение отдается иерархическим логлинейным моделям, которые позволяют упорядочить процесс подбора окончательной состоятельной модели.
Основной особенностью иерархических моделей является то, что присутствие какого-либо взаимодействия переменных означает присутствие всех взаимодействий, имеющих более низкий порядок, и главных эффектов этих переменных. Например, если в модели присутствует взаимодействие агрессия х симпатия, то в ней присутствуют главные эффекты переменных агрессия и симпатия;
если в модели присутствует взаимодействие
агрессия х симпатия х условия,
то в ней также присутствуют взаимодействия агрессия х симпатия,
агрессия х условия и симпатия х условия, и т. д.

Существуют три вспомогательных метода, которые предназначены для подбора адекватной модели. Все три метода оказываются полезными и приводят к сходным результатам

Метод *исследования оценок параметров* предназначен для вычисления оценок параметров для насыщенной модели. SPSS вычисляет также стандартизованные оценки. Если значения последних невелики, то они не оказывают значимого влияния на модель и обычно исключаются.

Метод *вычисления частичного критерия хи-квадрат* в дополнение к оценкам параметров модели SPSS вычисляет критерий *хи-квадрат,* характеризующий степень соответствия модели исходным данным. При помощи этого критерия проверяется, являются ли все однофакторные эффекты, а также эффекты более высоких порядков статистически значимыми. При этом отсутствие общей значимости эффектов второго порядка вовсе не означает, что все эффекты первого порядка не являются значимыми. Аналогично, из отсутствия общей значимости эффектов любого порядка не следует отсутствие значимости отдельных взаимодействий этого порядка. Вследствие этих двух особенностей в SPSS предусмотрена возможность раздельной проверки главных эффектов и эффектов взаимодействий.

Суть метода *пошагового исключения* состоит в автоматической «подгонке» модели и сходна с методом исключения предикторов из уравнения регрессии: из насыщенной модели постепенно исключаются те элементы (переменные и их взаимодействия), которые не оказывают значимого воздействия. Данный метод построения модели относится к иерархическому логлинейному моделированию. Если обнаружено статистически значимое взаимодействие четырех переменных, не проверяется (на предмет исключения из модели) взаимодействие трех из этих переменных, иначе модель не являлась бы иерархической по определению. Окончательный результат «подгонки» модели наиболее приемлем, если все оставшиеся в ней элементы оказываются статистически достоверными.