Лабораторная работа №3

Описание данных. Статистический вывод

Цель: Знакомство с этапом понимание данных стандарта CRISP-DM.

Задача этапа – найти, описать основные закономерности, которые содержатся в данных и статистически их подтвердить, попытаться выявить связи между этими данными, фактами.

Задание

Получив первоначальное представление о данных, рассмотрите закономерности, присущие данным. В этой лабораторной работе - категориальным.

Сформулируйте гипотезы о связи категориальных переменных, которые помогут в решении выбранной вами в предыдущей работе задачи. Проследите правильность формулировок гипотез H_0 и H_1 . Проведите испытание на независимость.

Зависимости категориальных переменных. Для проверки статистической связи между двумя категориальными переменными проводится испытание хи-квадрат ($\chi 2$) или используются ранговые тесты.

Теория

СТАТИСТИЧЕСКИЕ ИСПЫТАНИЯ НА НЕЗАВИСИМОСТЬ

Если Вы проводите, например, маркетинговое исследование среди читателей газеты, Вам придется рассмотреть одновременно множество переменных. Читательская аудитория газеты сегментируется по возрасту, полу, образованию, социальному классу и т.д.

Собрав данные, Вы захотите определить, нет ли связи между переменными, и если такая связь есть, то насколько она велика. В результате можно определить целевую группу читателей газеты. Полученная информация станет основой маркетинговых решений.

В самом начале этого курса были определены два типа данных: категориальные и количественные. В зависимости от типа данных применяются два различных подхода для определения того, существует или нет взаимосвязь между переменными. Если Вы должны определить, существует ли связь между двумя количественными переменными, Вы можете вычислить величину известную как *коэффициент корреляции*. Он будет рассмотрен позднее. Если же Вам необходимо установить наличие связи между категориальными данными, то Вы можете использовать *критерий хи-квадрат* и провести *испытание на независимость*.

Пример: .1

Университетская администрация желает определить, существует ли различие между предпочтениями студентов (мужчин и женщин), когда они выбирают учебные курсы, не входящие в число строго обязательных.

Использованы данные за последний год.

Курсы	Женщины	Мужчины	Всего

Бизнес	160	200	360
Естественные науки	80	70	150
Проектирование	50	20	70
Бухгалтерия	70	85	155
Маркетинг	60	90	150
Языки	50	65	115
Всего	470	530	1000

Чтобы определить, имеет ли какой-либо курс большую пропорцию мужчин- или женщин-претендентов, можно вычислить ожидаемые значения и сравнить их с наблюдаемыми значениями для каждой категории.

Если бы у мужчин и женщин не было никаких предпочтений, то в идеальном случае число ожидаемых заявлений на курс от мужчин было бы пропорционально как общему числу заявлений на этот курс, так и общему количеству мужчин. То же самое для женщин.

Ожидаемые значения могут быть рассчитаны по формуле:

$$\frac{\text{строка «всего»} \times \text{колонка «всего»}}{\text{итоговое «всего»}}$$

Например, для первой клетки

$$\frac{360 \times 470}{1000} = 169.2$$

Ожидаемые значения для каждого курса показаны в следующей таблице. Она называется *таблицей сопряженности*. Обратите внимание, что колонка и строка «Всего» не изменились

Курсы	Женщины	Мужчины	Всего
Бизнес	169.20с	190.80	360
Естественные науки	70.50	79.50	150
Проектирование	32.90	37.10	70
Бухгалтерия	72.85	82.15	155
Маркетинг	70.50	79.50	150
Языки	54.05	60.95	115
Всего	470	530	1000

Далее мы могли бы просто посчитать в процентах различия между наблюдаемыми и ожидаемыми значениями. Но этого недостаточно, мы должны доказать, что различия неслучайны. Необходимо провести формальное статистическое испытание. Это испытание называется испытанием хи-квадрат на независимость.

ИСПЫТАНИЕ ХИ-КВАДРАТ НА НЕЗАВИСИМОСТЬ

Испытание хи-квадрат (χ^2) позволяет нам определять значимость статистической связи между двумя категориальными переменными. Испытание выполняется в следующей последовательности.

- Определите нулевую (H_0) и альтернативную (H_1) гипотезы. В общем случае, эти гипотезы имеют форму:

H_0: не существует никакой связи между категориальными переменными
H_1: существует связь между категориальными переменными

Определяя H_0 и H_1 таким образом, Вы предполагаете, что никакой связи между двумя категориальными переменными не существует, пока не доказано иное.
Обратите внимание. В формулировке гипотез не присутствует никакой параметр генеральной совокупности, такой тип испытания называется *непараметрическим*.
Для нашего примера
H_0: не существует никакой связи между полом студентов и предпочитаемыми ими курсами
H_1: существует связь между полом студентов и предпочитаемыми ими курсами

- Выберите уровень значимости (для коммерческих решений обычно 1% или 5%). Как и ранее, уровень значимости указывает вероятность ошибки Типа I, то есть вероятность отклонения правильной нулевой гипотезы. Для нашего примера возьмем α=0,005.
- Вычислите проверочную статистику по формуле.

$$\chi^2 = \sum \left| \frac{(\text{наблюдаемое} - \text{ожидаемое})^2}{\text{ожидаемое}} \right|,$$

где:

наблюдаемое – фактическое число претендентов на каждый курс в описанном примере.

ожидаемое – значение из таблицы сопряженности.

Важное замечание. Единственное, за чем Вы должны следить, – это чтобы ни одно ожидаемое значение не было меньше числа 5. Если Вы сталкиваетесь именно с такой ситуацией, то придется объединить маленькие курсы в категорию «Другие», так, чтобы ожидаемое значение получилось больше 5. В нашем примере самое маленькое значение – 20 (мужчины/проектирование), нам объединять строки не требуется.

В следующей таблице представлены слагаемые для χ^2.

Курсы	Женщины	Мужчины
Бизнес	0.5002	0.4436
Естественные науки	1.2801	1.1352
Проектирование	8.8878	7.8817
Бухгалтерия	0.1115	0.0989
Маркетинг	1.5638	1.3868
Языки	0.3035	0.2691
Всего		

Первое слагаемое (0.5002) было рассчитано как

$$(160\text{-}169.20)^2/169.20$$

Сумма всех чисел в таблице χ^2= 23.8623. Эта проверочная статистика всегда имеет положительную значение.

- Чтобы определить критическое значение статистики по таблице распределения χ^2, нужно кроме уровня значимости, знать число степеней свободы, которое для этого испытания рассчитывается как:

$$DF = (r - 1) \times (c - 1),$$

где **r** и **c** – число строк и столбцов соответственно в таблице сопряженности. В этом примере r = 6 (число курсов) и c = 2 (пол претендентов), поэтому имеем 5 степеней свободы. По таблице хи-квадрат находим

$$\chi^2_{0.05;5} = 11.07$$

- Следующий шаг – сравнение проверочной статистики с критическим значением. Если различия статистически значимы, то значение окажется в критической зоне, т.е. в области хвоста хи-квадрат распределения. Иначе говоря

Если проверочная статистика > критического значения, отклоняем нулевую гипотезу H_0

Если проверочная статистика< критического значения, принимаем нулевую гипотезу H_0

Для хи-квадрат распределения всегда выполняется одностороннее испытание, т.к. проверочная статистика всегда положительна.

В этом примере значение проверочной статистики 23.8623 больше, чем критическое значение 11.07, следовательно, значение проверочной статистики расположено в хвосте χ^2-распределения. Поэтому на уровне значимости 5% мы должны отклонить нулевую гипотезу H_0.

- В заключительной части испытания гипотезы необходимо прокомментировать результат.

В нашем примере, Вы можете сказать, что на уровне значимости 5% имеются очевидные свидетельства, чтобы предполагать существование связи между полом студента и выбираемыми курсами.

EXCEL КОМАНДЫ

В EXCEL есть встроенная функция ХИ2ТЕСТ( ). Чтобы воспользоваться этой функцией нужно заранее подготовить таблицу наблюдаемых значений и таблицу сопряженности.

Если собранные выборочные данные представлены не в форме таблицы наблюдаемых значений, то привести их к нужной форме таблицы можно, используя команду меню EXCEL: Данные – Сводная таблица.

Далее строится таблица сопряженности.

При использовании функции =ХИ2ТЕСТ(фактический интервал; ожидаемый интервал) будьте осторожны. Функция ХИ2ТЕСТ( ) вычисляет значение уровня значимости на котором нулевая гипотеза должна быть отклонена.

Если это значение меньше требуемого Вами уровня значимости, то Вы отклоняете нулевую гипотезу.

До сих пор, испытывая гипотезу, мы сравнивали *проверочную статистику* с ее *критическим значением*. Этот подход сложился исторически, так как для вычисления вручную он проще. Но современные компьютерные статистические программы для большинства тестов вычисляют пороговый уровень значимости, на котором нулевую гипотезу следует отклонить. Обычно эта величина называется **p-значением**. Если этот уровень меньше того, который Вы установили, то это достаточное свидетельство против нулевой гипотезы.

Обратите внимание. Правило для p-значения протвоположно правилу для проверочной статистики.

Если проверочная статистика > критического значения, отклоняем нулевую гипотезу H_0

Если проверочная статистика< критического значения, принимаем нулевую гипотезу H_0
НО:
Если р-значение < заданного уровня значимости α, отклоняем нулевую гипотезу H_0
Если р-значение> заданного уровня значимости α, принимаем нулевую гипотезу H_0

Для нашего примера
функция ХИ2ТЕСТ(фактический интервал; ожидаемый интервал)= 0,000231
Это значение намного меньше принятого нами уровня значимости 5% (или в долях от единицы α=0,05). То есть H_0 отвергается.

Для нашего примера целесообразно проверить связи между такими характеристиками ка образование, область деятельности, подразделение, вовлеченность сотрудника, уровень, на котором он находится, должность, семейное положение. Важно понимать, что некоторые переменные, хотя и обозначены цифрами, по факту являются качественными.

Иногда переменные у нас в описаны числовыми признаками. Для этого испытания нужно перейти к качественным: вы можете использовать плохо, удовлетворительно, хорошо, отлично, либо упростить до перехода через заданную вами границу. Границы вы можете задать исходя из распределения значений.

Таблицы данных можно построить, используя сводные таблицы:

Например

Должность	Длительность нахождения			Всего
	Недавно *(например, менее 3 лет)*	Средне *(от 3 до 5)*	…..	
A				
V				
Всего				1000

Таким образом должно быть проведено не менее 5 испытаний, отражающих закономерности. В отчете должны быть приведены формулировки гипотез, таблицы с данными, таблицы сопряженности,

значения статистик, проверочных статистик, вывод. Для ранговых тестов должны быть приведены условия присвоения рангов.

Полученные результаты желательно визуализировать.

Оценка от 3 до 5.

Дополнительные баллы: логика, построение плана испытаний. В выводах должны быть отражены результаты

Контрольные вопросы.

1. Что такое нулевая гипотеза в испытаниях на независимость
2. Как формулируется альтернативная гипотеза
3. Для чего нужно критическое значение
4. Где можно найти проверочную статистику