МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное

образовательное учреждение высшего образования

«ЧЕРЕПОВЕЦКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Институт информационных технологий

*наименование института(факультета)*

Кафедра математического и программного обеспечения ЭВМ

*наименование кафедры*

Основы Data Science

*наименование дисциплины в соответствии с учебным планом*

ЛАБОРАТОРНАЯ РАБОТА №2

“Описание данных. Статистический вывод”

|  |  |
| --- | --- |
| Исполнитель |  |
| Студент | 1ПИб-02-1оп-22 |
|  | *группа* |
|  | Харламов Д.А. |
|  | *ФИО* |
| Руководитель | Юдина О.В. |
|  | *ФИО преподавателя* |
| Оценка |  |
| Подпись |  |

2024 год

**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, знаниями, получить что-то новое, понять, что они могут дать.

**Задание:**

Получив первоначальное представление о данных, рассмотрите закономерности, присущие данным. Для задачи, которую вы сформулировали в предыдущей работе: Подумайте, какие статистические гипотезы могут подтвердить ил опровергнуть ваши предположения. Сформулируйте гипотезы о равенстве выборочных средних и долей категориальных, разделяя переменные. Используйте закономерности, особенности, которые вы смогли обнаружить в предыдущей работе.

Проследите правильность формулировок гипотез H0 и H1 . Проведите испытание, используя соответствующие критерии. Для того, чтобы сформулировать полезные гипотезы: Подумайте, что вы хотите выяснить, проверить. Можно использовать фильтры разного уровня. Например, чтобы сравнить учебные успехи студентов, имеющих двоих и более друзей и тех, у кого друзей меньше – отфильтруйте по этому признаку показатели успеваемости, посчитайте среднее и доверительные интервалы. Если доверительные интервалы пересекаются – сформулируйте гипотезы. Результат позволит сделать заключение – отличается ли успеваемость студентов, которые общаются мало и их более общительных товарищей.

Ход работы

Гипотезы:

1. Существование связи между должностью сотрудника (JobRole) и полом сотрудника (Gender).

2. Существование связи между частотой командировок (BusinessTravel) и уровнем образования сотрудника (EducationField(Life Sciences,Technical Degree)).

3. Существование связи между присутствием в компании (Attrition) и удовлетворённостью отношениями в рабочем коллективе (JobSatisfaction).

Существование связи между должностью сотрудника (JobRole) и полом сотрудника (Gender).

H0: не существует никакой связи между полом сотрудников и занимаемой должностью.

Н1: существует связь между полом сотрудников и занимаемой должностью.

Уровень значимости — 5%

Наблюдаемые данные:

|  |  |  |  |
| --- | --- | --- | --- |
| Количество по полю Gender | Названия столбцов |  |  |
| Названия строк | Female | Male | Общий итог |
| Healthcare Representative | 51 | 80 | 131 |
| Human Resources | 16 | 36 | 52 |
| Laboratory Technician | 85 | 174 | 259 |
| Manager | 47 | 55 | 102 |
| Manufacturing Director | 72 | 73 | 145 |
| Research Director | 33 | 47 | 80 |
| Research Scientist | 114 | 178 | 292 |
| Sales Executive | 132 | 194 | 326 |
| Sales Representative | 38 | 45 | 83 |
| Общий итог | 588 | 882 | 1470 |

Ожидаемые данные:

|  |  |  |  |
| --- | --- | --- | --- |
| Названия строк | Female | Male | Общий итог |
| Healthcare Representative | 52,4 | 78,6 | 131 |
| Human Resources | 20,8 | 31,2 | 52 |
| Laboratory Technician | 103,6 | 155,4 | 259 |
| Manager | 40,8 | 61,2 | 102 |
| Manufacturing Director | 58 | 87 | 145 |
| Research Director | 32 | 48 | 80 |
| Research Scientist | 116,8 | 175,2 | 292 |
| Sales Executive | 130,4 | 195,6 | 326 |
| Sales Representative | 33,2 | 49,8 | 83 |
| Общий итог | 588 | 882 | 1470 |

ХИ2-ТЕСТ (фактический интервал; ожидаемый интервал) = 0,041954442.

0,041954442< 0,05.

Вывод: р-значение меньше уровня значимости, значит нулевая гипотеза отклоняется, это означает, что существует связь между полом сотрудников и занимаемой должностью.

Существование связи между частотой командировок (BusinessTravel) и уровнем образования сотрудника (EducationField(Life Sciences, Technical Degree)).

H0: не существует никакой связи между частотой командировок и уровнем образования сотрудников.

Н1: существует связь между частотой командировок и уровнем образования сотрудников.

Уровень значимости – 5%

Построим сводную таблицу распределения по частоте командировок:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Количество по полю BusinessTravel | Названия столбцов |  |  |  |
| Названия строк | Non-Travel | Travel\_Frequently | Travel\_Rarely | Общий итог |
| Life Sciences | 63 | 123 | 420 | 606 |
| Technical Degree | 14 | 21 | 97 | 132 |
| Общий итог | 77 | 144 | 517 | 738 |

Далее проведем испытание хи-квадрат на 5% значимости, построим для этого таблицу сопряженности:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Количество по полю BusinessTravel | Названия столбцов |  |  |  |
| Названия строк | Non-Travel | Travel\_Frequently | Travel\_Rarely | Общий итог |
| Life Sciences | 63,22764228 | 118,2439024 | 424,5284553 | 606 |
| Technical Degree | 13,77235772 | 25,75609756 | 92,47154472 | 132 |
| Общий итог | 77 | 144 | 517 | 738 |

Посчитаем проверочную статистику.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Названия строк | Non-Travel | Travel\_Frequently | Travel\_Rarely | Общий итог |
| Life Sciences | 0,000819594 | 0,191303429 | 0,048305142 | 0,240428165 |
| Technical Degree | 0,003762682 | 0,878256652 | 0,221764515 | 1,103783849 |
|  | Х2 | 1,344212014 |  |  |

Х2= 1,344212014

DF = 2

ХИ2РАСП(Х2; степень свободы) = 0,51063205

0,51063205> 0,05

Вывод: р-значение больше уровня значимости, значит нулевая гипотеза не отклоняется, это означает, что не существует связи между частотой командировок и уровнем образования сотрудников.

Существование связи между присутствием в компании (Attrition) и удовлетворённостью отношениями в рабочем коллективе (JobSatisfaction).

H0: не существует никакой связи между тем, что работник ушёл из компании и его взаимоотношениями в коллективе.

Н1: существует связь между тем, что работник ушёл из компании и его взаимоотношениями в коллективе.

Построим сводную таблицу распределения по удовлетворённостью отношениями:

|  |  |  |  |
| --- | --- | --- | --- |
| Количество по полю Attrition | Названия столбцов |  |  |
| Названия строк | No | Yes | Общий итог |
| 1 | 223 | 66 | 289 |
| 2 | 234 | 46 | 280 |
| 3 | 369 | 73 | 442 |
| 4 | 407 | 52 | 459 |
| Общий итог | 1233 | 237 | 1470 |

Далее проведем испытание хи-квадрат на 5% значимости, построим для этого таблицу сопряженности:

|  |  |  |  |
| --- | --- | --- | --- |
| Количество по полю Attrition | Названия столбцов |  |  |
| Названия строк | No | Yes | Общий итог |
| 1 | 242,4061224 | 46,59 | 289 |
| 2 | 234,8571429 | 45,14 | 280 |
| 3 | 370,7387755 | 71,26 | 442 |
| 4 | 384,9979592 | 74 | 459 |
| Общий итог | 1233 | 237 | 1470 |

Посчитаем проверочную статистику.

|  |  |  |  |
| --- | --- | --- | --- |
| Названия строк | No | Yes | Общий итог |
| 1 | 1,553581175 | 8,083 | 9,6361364 |
| 2 | 0,003128259 | 0,016 | 0,01940312 |
| 3 | 0,008154907 | 0,042 | 0,05058107 |
| 4 | 1,257382769 | 6,542 | 7,79895642 |
| Х2 | 17,50507701 |  |  |

Х2= 17,50507701

DF = 3

Критическое значение: 7,814727903

|17,50507701| > |7,814727903|

Вывод: так как проверочная статистика больше, чем критическое значение, то мы отказываемся от нулевой гипотезы , а это значит, что существует связь между тем, что работник ушёл из компании и его взаимоотношениями в коллективе.

Контрольные вопросы.

1. Что такое нулевая гипотеза в тестах для численных переменных?

Нулевая гипотеза – это предположение о том, что две совокупности X и Y, которые сравниваются по одному или нескольким признакам, не различаются между собой. Нулевая гипотеза всегда формируется для утверждения того, что выборочная статистика согласуется с принятым параметром генеральной совокупности, мы никогда не можем на 100% доказать или опровергнуть нулевую гипотезу.

2. Как формулируется альтернативная гипотеза?

Утверждение, которое проверяется на соответствие нулевой гипотезе, является альтернативной гипотезой. При статистической проверке гипотезы, чтобы доказать, что альтернативная гипотеза верна, следует показать, что данные противоречат нулевой гипотезе. А именно, имеется достаточно доказательств против нулевой гипотезы, чтобы продемонстрировать, что альтернативная гипотеза верна.

3. Для чего нужно критическое значение?

Проверочная статистика сравнивается с критическим значением. В результате сравнения делается выбор в пользу одной из гипотез: если проверочная статистика меньше критического значения, принимается гипотеза H0, иначе принимается гипотеза H1.

4. Где можно найти проверочную статистику?

Её можно вычислить с помощью excel команды, например ХИ2.ОБР.