-МИНОБРНАУКИ РОССИИ

ФГБОУ ВО Череповецкий государственный университет

Институт информационных технологий

Кафедра: МПО ЭВМ

Дисциплина: Основы Data Science

ЛАБОРАТОРНАЯ РАБОТА № 3

Тема: **Описание данных. Статистический вывод**

Выполнил: студент гр. 1ПИб-00-1оп-21 Макаренков В. С.

Проверил: доцент, Юдина Ольга Вадимовна

Череповец, 2023 г.

Задание

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.

Ход работы

Гипотезы:

Существование связи между полом сотрудника и уровнем работы;

Существование связи между наличием переработок и полом сотрудника;

Существование связи между увольнением и специальностью сотрудника.

Гипотеза связи пола сотрудника и уровня работы

H0: не существует никакой связи между полом сотрудников и уровнем работы

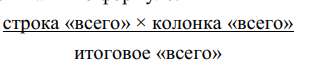
Н1: существует связь между полом сотрудников и уровнем работы

Таблица 1

Распределение по полу

|  |  |  |  |
| --- | --- | --- | --- |
| Названия строк | Female | Male | Общий итог |
| 1 | 199 | 344 | 543 |
| 2 | 220 | 314 | 534 |
| 3 | 94 | 124 | 218 |
| 4 | 51 | 55 | 106 |
| 5 | 24 | 45 | 69 |
| Общий итог | 588 | 882 | 1470 |

Формула расчета сопряженности:



Проведем испытание хи-квадрат на 5% значимости:

Таблица 2

Таблица сопряженности

|  |  |  |  |
| --- | --- | --- | --- |
| Названия строк | Female | Male | Общий итог |
| 1 | 217,2 | 325,8 | 543 |
| 2 | 213,6 | 320,4 | 534 |
| 3 | 87,2 | 130,8 | 218 |
| 4 | 42,4 | 63,6 | 106 |
| 5 | 27,6 | 41,4 | 69 |
| Общий итог | 588 | 882 | 1470 |

Таблица сопряженности отражает то, как распределялись бы рабочие, если бы у сотрудников и у работодателя не было бы предпочтений в выборе должности, а число рабочих конкретного пола на конкретной должности было бы пропорционально как общему числу рабочих на данной должности, так и числу рабочих конкретного пола в данной компании.

Проверочная статистика:

= 7.43

DF = 4

Критическое значение:

9,48

Так как значение проверочной статистики не превышает критическое значение, гипотеза H0 принимается. Это значит, что в компанию нанимают специалистов по их опытности, невзирая на пол.

Гипотеза связи между наличием переработок и полом сотрудника

H0: не существует никакой связи между наличием переработок и полом сотрудника

Н1: существует связь между наличием переработок и полом сотрудника

Таблица 3

Распределение по переработкам

|  |  |  |  |
| --- | --- | --- | --- |
| Названия строк | No | Yes | Общий итог |
| Female | 408 | 180 | 588 |
| Male | 646 | 236 | 882 |
| Общий итог | 1054 | 416 | 1470 |

Проведем испытание хи-квадрат на 5% значимости:

Таблица 4

Таблица сопряженности

|  |  |  |  |
| --- | --- | --- | --- |
| Названия строк | Female | Male | Общий итог |
| No | 408 | 646 | 1054 |
| Yes | 180 | 236 | 416 |
| Общий итог | 588 | 882 | 1470 |

Проверочная статистика:

= 2,58

DF = 1

Критическое значение:

3,84

Так как значение проверочной статистики не превышает критическое значение, гипотеза H0 принимается. Это значит, что в компании сотрудники перерабатывают/ не перерабатывают вне зависимости от пола.

Гипотеза связи между увольнением и специальностью сотрудника

H0: не существует никакой связи между увольнением и специальностью сотрудника

Н1: существует связь между увольнением и специальностью сотрудника

Таблица 5

Распределение по увольнениям

|  |  |  |  |
| --- | --- | --- | --- |
| Названия строк | No | Yes | Общий итог |
| Human Resources | 20 | 7 | 27 |
| Life Sciences | 517 | 89 | 606 |
| Marketing | 124 | 35 | 159 |
| Medical | 401 | 63 | 464 |
| Other | 71 | 11 | 82 |
| Technical Degree | 100 | 32 | 132 |
| Общий итог | 1233 | 237 | 1470 |

Проведем испытание хи-квадрат на 5% значимости:

Таблица 6

Таблица сопряженности

|  |  |  |  |
| --- | --- | --- | --- |
| Названия строк | No | Yes | Общий итог |
| Human Resources | 22,64693878 | 4,353061224 | 27 |
| Life Sciences | 508,2979592 | 97,70204082 | 606 |
| Marketing | 133,3653061 | 25,63469388 | 159 |
| Medical | 389,1918367 | 74,80816327 | 464 |
| Other | 68,77959184 | 13,22040816 | 82 |
| Technical Degree | 110,7183673 | 21,28163265 | 132 |
| Общий итог | 1233 | 237 | 1470 |

Проверочная статистика:

= 16,02

DF = 4

Критическое значение:

9,48

Так как значение проверочной статистики превышает критическое значение, гипотеза H0 отклоняется. Это значит, что в компании существует связь между количеством увольнений и специальностью .

Контрольные вопросы:

1. Что такое нулевая гипотеза в испытаниях на независимость?

Нулевая гипотеза в испытаниях на независимость – утверждение, что значения категориальных переменных не связаны.

1. Как формулируется альтернативная гипотеза?

Альтернативная гипотеза – утверждение о зависимости категориальных переменных.

1. Для чего нужно критическое значение?

Вычисленная 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.

1. Где можно найти проверочную статистику?

Проверочную статистику можно найти в таблице критерия Пирсона.