МИНОБРАНУКИ РОССИИ

федеральное государственное бюджетное

образовательное учреждение высшего образования

«ЧЕРЕПОВЕЦКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Институт информационных технологий

*наименование института (факультета)*

Математическое и программное обеспечение ЭВМ

*наименование кафедры*

Основы Data Science

*наименование дисциплины в соответствии с учебным планом*

ЛАБОРАТОРНАЯ РАБОТА №3

Описание данных. Статистический вывод

|  |  |
| --- | --- |
| Руководитель | Юдина О. В. |
|  | ФИО преподавателя |
| Исполнитель |  |
| студент | 1ПИб-02-1оп-22 |
|  | группа |
|  | Тихомиров Владислав |
|  | Валерьевич |
|  | Фамилия, имя, отчество |
| Оценка |  |
| Подпись |  |

Череповец, 2024 г.

**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. 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 H0 и H1. Проведите испытание на независимость.

**Зависимости категориальных переменных.** Для проверки статистической связи между двумя категориальными переменными проводится испытания хи-квадрат или используются ранговые тесты.

Ход работы

Тест 1

Существует ли зависимость уровня оценок от пола

|  |
| --- |
| H0- не существует |
| H1-существует |

Сводная таблица

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Female** | **Male** | **Общий итог** |
| Средне и ниже среднего | 25 | 24 | 49 |
| Хорошо и отлично | 18 | 32 | 50 |
| **Общий итог** | **43** | **56** | **99** |

Таблица сопряжённости

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Female** | **Male** | **Общий итог** |
| Средне и ниже среднего | 21,28282828 | 27,71717172 | 49 |
| Хорошо и отлично | 21,71717172 | 28,28282828 | 50 |
| **Общий итог** | **43** | **56** | **99** |

Таблица Хи-квадрат

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Female** | **Male** | **Общий итог** |
| Средне и ниже среднего | 0,649226005 | 0,498512825 | 1,14773883 |
| Хорошо и отлично | 0,636241485 | 0,488542569 | 1,124784053 |
| **Общий итог** | **43** | **56** | **99** |

Для последующих расчётов было принято взять уровень значимости, равным 5% (0,05)

Расчётная таблица

|  |  |  |
| --- | --- | --- |
| Уровень значимости | 5% | 0,05 |
| Степень свободы DF | (2-1)\*(2-1) | 1 |
| **Проверочная статистка** | 2,272522883 | |
| Критическое значение | 3,84 | |

|2,27|<|3,84| Принимает гипотезу H0

Тест 2

Существует ли зависимость между успеваемостью и депрессией

|  |  |
| --- | --- |
| H0- не существует |  |
| H1-существует |  |
|  |  |

Сводная таблица

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Количество по полю AcademicPerformance** | **Названия столбцов** |  |  |  |
| **Названия строк** | **Average** | **Below average** | **Good** | **Общий итог** |
| No | 9 |  | 9 | 18 |
| Yes | 15 | 4 | 12 | 31 |
| **Общий итог** | **24** | **4** | **21** | **49** |

Таблица Сопряжённости

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Названия строк** | **Average** | **Below average** | **Good** | **Общий итог** |
| No | 8,816326531 | 1,469387755 | 7,714285714 | 18 |
| Yes | 15,18367347 | 2,530612245 | 13,28571429 | 31 |
| **Общий итог** | **24** | **4** | **21** | **49** |

Таблица Хи-квадрат

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Названия строк** | **Average** | **Below average** | **Good** | **Общий итог** |
| No | 0,003826531 | 1,469387755 | 0,214285714 | 18 |
| Yes | 0,002221856 | 0,85319289 | 0,124423963 | 31 |
| **Общий итог** | **24** | **4** | **21** | **49** |

Расчётная таблица

|  |  |  |
| --- | --- | --- |
| **Проверочная статистика** | 2,66733871 | |
| **Уровень значимости** | 5% | 0,05 |
| **Степень свободы** | df= (2-1)\*(3-1) | 2 |
| **t-крит** | 5,99 | |

| 2,67|<|5,99|, принимаем гипотезу H0.

Тест 3

Существует зависимость записи конспектов от пола

H0 – не существует

H1 – существует

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Количество по полю TakingNoteInClass** | **Названия столбцов** |  |  |  |
| **Названия строк** | **No** | **Sometimes** | **Yes** | **Общий итог** |
| Female |  | 4 | 39 | 43 |
| Male | 12 | 22 | 22 | 56 |
| **Общий итог** | **12** | **26** | **61** | **99** |

Таблица Сопряжённости

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Названия строк** | **No** | **Sometimes** | **Yes** | **Общий итог** |
| Female | 5,212121212 | 11,2929293 | 26,49 | 43 |
| Male | 6,787878788 | 14,7070707 | 34,51 | 56 |
| **Общий итог** | **12** | **26** | **61** | **99** |

Таблица Хи-квадрат

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Названия строк** | **No** | **Sometimes** | **Yes** | **Общий итог** |
| Female | 5,212121212 | 4,70974504 | 5,902 | 43 |
| Male | 4,002164502 | 3,61641137 | 4,532 | 56 |
| **Общий итог** | **12** | **26** | **61** | **99** |

Расчётная таблица

|  |  |  |
| --- | --- | --- |
| **Проверочная статистика** | 27,97454157 | |
| **Уровень значимости** | 5% | 0,05 |
| **Степень свободы** | df= (2-1)\*(3-1) | 2 |
| **t-крит** | 5,99 | |

| 27,97|>|5,99|, гипотеза H1 однозначно является верной

Тест 4

Существует зависимость статуса депрессии от пола

Н0 – Не существует

Н1 – Существует

Сводная таблица

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Количество по полю DepressionStatus** | **Названия столбцов** |  |  |  |
| **Названия строк** | **No** | **Sometimes** | **Yes** | **Общий итог** |
| Female | 8 | 10 | 25 | 43 |
| Male | 13 | 34 | 9 | 56 |
| **Общий итог** | **21** | **44** | **34** | **99** |

Таблица сопряжённости

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Названия строк** | **No** | **Sometimes** | **Yes** | **Общий итог** |
| Female | 9,121212121 | 19,1111111 | 14,7677 | 43 |
| Male | 11,87878788 | 24,8888889 | 19,2323 | 56 |
| **Общий итог** | **21** | **44** | **34** | **99** |

Таблица Хи-квадрат

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Названия строк** | **No** | **Sometimes** | **Yes** | **Общий итог** |
| Female | 0,137823417 | 4,34366925 | 7,08984 | 43 |
| Male | 0,105828695 | 3,33531746 | 5,44398 | 56 |
| **Общий итог** | **21** | **44** | **34** | **99** |

Расчёты

|  |  |  |
| --- | --- | --- |
| **Проверочная статистика** | 20,45645991 | |
| **Уровень значимости** | 5% | 0,05 |
| **Степень свободы** | df= (2-1)\*(3-1) | 2 |
| **t-крит** | 5,99 | |

| 20,45|>|5,99|, гипотеза H1 однозначно является верной

Тест 5

Существует зависимость между полом и любовью к презентациям

Н0 – Не существует

Н1 – Существует

Сводная таблица

|  |  |  |  |
| --- | --- | --- | --- |
| **Количество по полю LikeNewThings** | **Названия столбцов** |  |  |
| **Названия строк** | **No** | **Yes** | **Общий итог** |
| Female | 2 | 41 | 43 |
| Male | 8 | 48 | 56 |
| **Общий итог** | **10** | **89** | **99** |

Таблица сопряжённости

|  |  |  |  |
| --- | --- | --- | --- |
| **Названия строк** | **No** | **Yes** | **Общий итог** |
| Female | 4,343434343 | 38,657 | 43 |
| Male | 5,656565657 | 50,343 | 56 |
| **Общий итог** | **10** | **89** | **99** |

Таблица Хи-квадрат

|  |  |  |  |
| --- | --- | --- | --- |
| **Названия строк** | **No** | **Yes** | **Общий итог** |
| Female | 1,264364576 | 0,1421 | 43 |
| Male | 0,970851371 | 0,1091 | 56 |
| **Общий итог** | **10** | **89** | **99** |

Расчёты

|  |  |  |
| --- | --- | --- |
| **Проверочная статистика** | 2,486363806 | |
| **Уровень значимости** | 5% | 0,05 |
| **Степень свободы** | df= (2-1)\*(3-1) | 2 |
| **t-крит** | 5,99 | |

| 2,48|<|5,99|, гипотеза H0 является верной

# Контрольные вопросы

1. Что такое нулевая гипотеза в испытаниях на независимость?

Нулевая гипотеза в испытаниях на независимость утверждает, что нет связи между двумя переменными. Другими словами, она предполагает, что две изучаемые переменные независимы друг от друга, и их значения не влияют друг на друга.

1. Как формулируется альтернативная гипотеза?

Альтернативная гипотеза — это противоположность нулевой гипотезы. Она утверждает, что существует связь между двумя переменными.

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?

Проверочная статистика — это значение, которое рассчитывается на основе данных и используется для проверки нулевой гипотезы. Проверочную статистику можно найти в статистических таблицах, например, таблица X^2.