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 №3

«Описание данных. Статистический вывод»

|  |  |
| --- | --- |
| Исполнитель |  |
| Студент | 1ПИб-02-3оп-22 |
|  | *группа* |
|  | Маркелов С. А. |
|  | *ФИО* |
| Руководитель | Юдина О.В. |
|  | *ФИО преподавателя* |
| Оценка |  |
| Подпись |  |

2024 год

Цель: Знакомство с этапом понимания 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 H0 и H1. Проведите испытание на независимость.

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

Гипотезы:

1. Не существует связь между успеваемостью и проблемами с учебой
2. Не существует связь между успеваемостью и наличием депрессии
3. Не существует связь между успеваемостью и полом
4. Не существует связь между успеваемостью и ведением конспектов

Уровень значимости 5% = 0,05

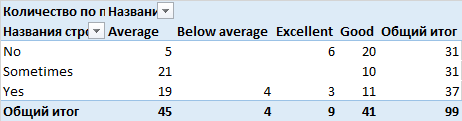
Проверка гипотез

1.

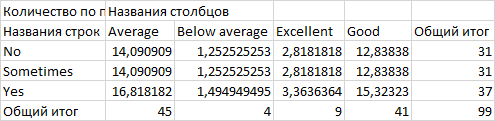
H0 – Не существует связь между успеваемостью и проблемами с учебой

H1 – Существует связь между успеваемостью и проблемами с учебой

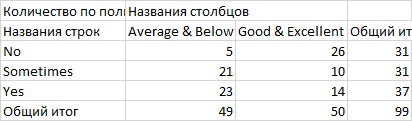
Наблюдаемые значения:

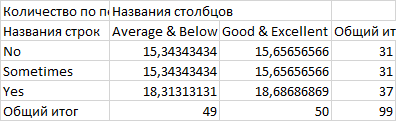


Ожидаемые значения:



Т. к. часть значений менее 5, необходимо объединить столбцы:





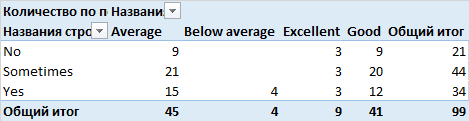
С помощью функции =ХИ2ТЕСТ находим p = |0,00004| < |0,05|, следовательно, гипотеза H0 отклоняется. Значит, зависимость между успеваемостью и проблемами с учебой есть

2.

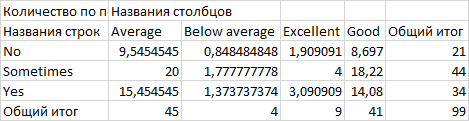
H0 – Не существует связь между успеваемостью и наличием депрессии

H1 – Существует связь между успеваемостью и наличием депрессии

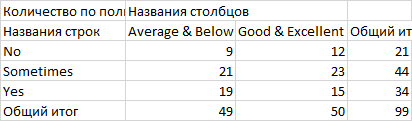
Наблюдаемые значения:

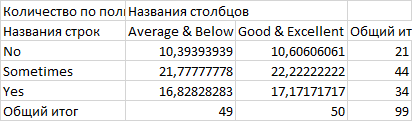


Ожидаемые значения:



Т. к. часть значений менее 5, необходимо объединить столбцы:





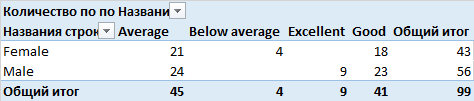
С помощью функции =ХИ2ТЕСТ находим p = |0,61| > |0,05|, следовательно, гипотеза H0 принимается. Значит, зависимости между успеваемостью и наличием депрессии нет

3.

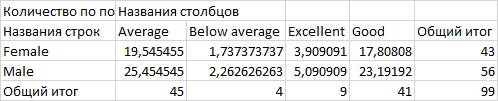
H0 – Не существует связь между успеваемостью и полом

H1 – Существует связь между успеваемостью и полом

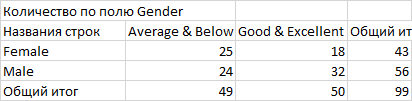
Наблюдаемые значения:

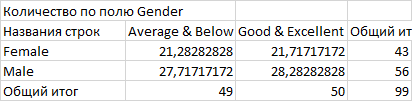


Ожидаемые значения:



Т. к. часть значений менее 5, необходимо объединить столбцы:





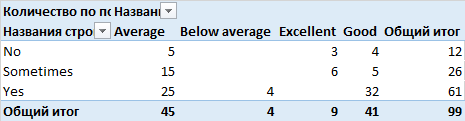
С помощью функции =ХИ2ТЕСТ находим p = |0,13| > |0,05|, следовательно, гипотеза H0 принимается. Значит, зависимости между полом и успеваемостью нет

4.

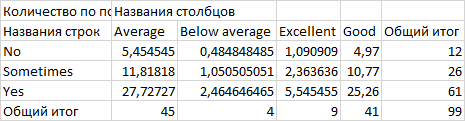
H0 – Не существует связь между успеваемостью и ведением конспектов

H1 – Существует связь между успеваемостью и ведением конспектов

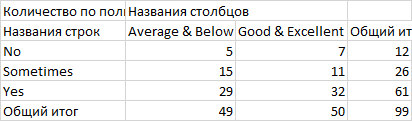
Наблюдаемые значения:

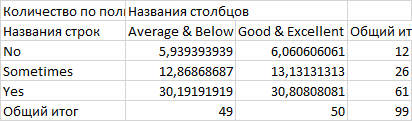


Ожидаемые значения:



Т. к. часть значений менее 5, необходимо объединить столбцы:





С помощью функции =ХИ2ТЕСТ находим p = |0,58| > |0,05|, следовательно, гипотеза H0 принимается. Значит, зависимости между успеваемостью и ведением конспектов нет

Контрольные вопросы:

1. Что такое нулевая гипотеза в испытаниях на независимость?

В испытаниях на независимость нулевая гипотеза обычно утверждает, что два набора данных независимы друг от друга. То есть, нулевая гипотеза предполагает отсутствие взаимосвязи между двумя переменными.

1. Как формулируется альтернативная гипотеза?

Альтернативная гипотеза – это утверждение, являющееся логическим отрицанием нулевой гипотезы. В случае, когда нулевая гипотеза утверждает, что два набора данных независимы друг от друга, альтернативная гипотеза соответственно утверждает, что наборы данных зависимы.

1. Для чего нужно критическое значение?

Критическое значение – это значение χ2-статистики, при котором нулевая гипотеза отвергается в пользу альтернативной.

Если значение χ2-статистики превышает критическое значение, то это означает, что наблюдаемые данные находятся в критической области, что позволяет отвергнуть нулевую гипотезу в пользу альтернативной с заданным уровнем значимости. Если значение статистики находится вне критической области, то нулевая гипотеза не отвергается.

1. Где можно найти проверочную статистику?

Проверочная статистика вычисляется на основе данных, которые анализируются в рамках статистического теста. Её можно найти путем применения соответствующей формулы, которая зависит от выбранного метода анализа и статистического теста. В хи-квадрат тесте:

Вывод:

В ходе выполнения лабораторной работы были выдвинуты несколько гипотез по набору данных о студентах. Для проверки выдвинутых гипотез были проведены хи-тесты.

В результате проведения тестов были подтверждены гипотезы:

* Не существует связь между успеваемостью и наличием депрессии
* Не существует связь между успеваемостью и полом
* Не существует связь между успеваемостью и ведением конспектов

Остальные гипотезы были опровергнуты:

* Не существует связь между успеваемостью и проблемами с учебой (т. е. связь есть)