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

Тема: Описание данных. Статистический вывод.

|  |  |
| --- | --- |
| Руководитель | Юдина О.В. |
|  | *ФИО преподавателя* |
| Исполнитель |  |
| студент | 1ПИб-02-1оп-22 |
|  | *группа* |
|  | Маслов В.А. |
|  | *Фамилия, имя, отчество* |
| Оценка |  |
| Подпись |  |

2025 год

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

Задача: 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.

Ход работы

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

Уровень значимости 5% = 0,05.

1. Проверка гипотез

Гипотеза 1. Связь между успеваемостью и проблемами с учебой.

H0 – Не существует, H1 – Существует.

Сводная таблица:

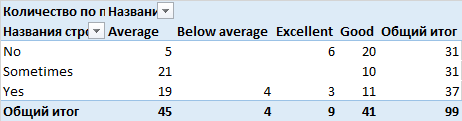
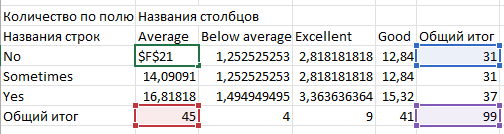
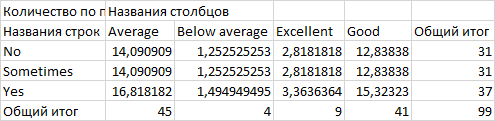
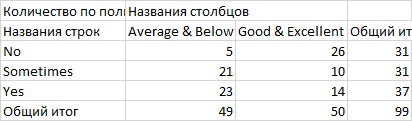


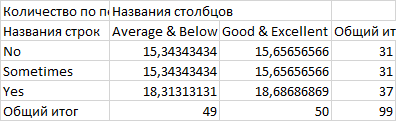
Таблица сопряжённости:

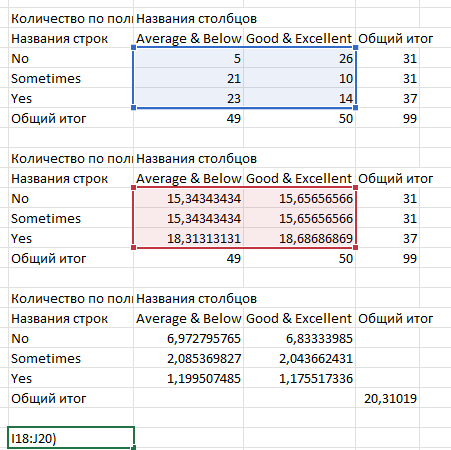




Необходимо объединить столбцы:







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

Гипотеза 2. Связь между успеваемостью и наличием депрессии

H0 – Не существует, H1 – Существует.

Сводная таблица:

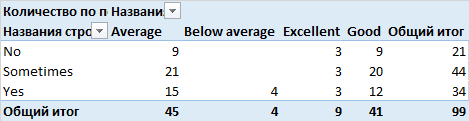
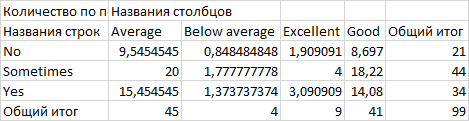
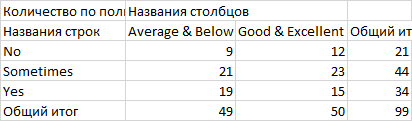
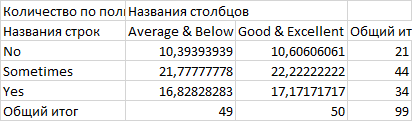


Таблица сопряжённости:



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.

Гипотеза 3: H0 – Не существует связь между успеваемостью и полом.

H1 – Существует связь между успеваемостью и полом.

Сводная таблица:

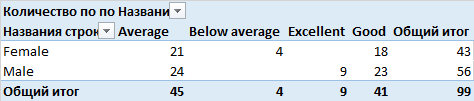
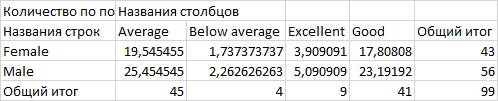
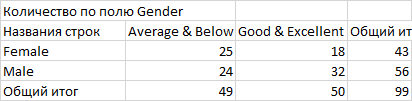
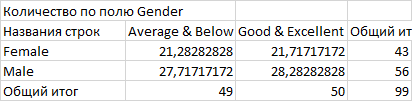


Таблица сопряжённости:



Необходимо объединить столбцы:





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

Гипотеза 4. Связь между успеваемостью и ведением конспектов

H0 – Не существует, H1 – Существует.

Сводная таблица:

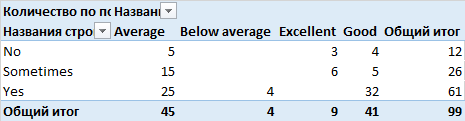
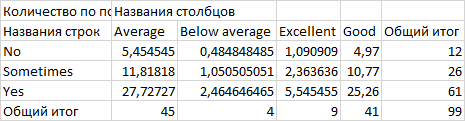
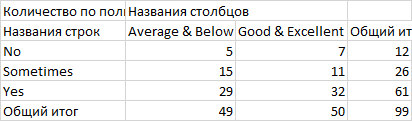
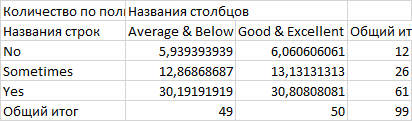


Таблица сопряжённости:



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.

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).

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: