Институт информационных технологий

Кафедра: Математическое и программное обеспечение ЭВМ

Дисциплина: Основы Data Science

ЛАБОРАТОРНАЯ РАБОТА № 3

Тема: Описание данных. Статистический вывод

Выполнил:

студент гр. 1ПИб-02-2оп-23

Кринкин Олег Алексеевич

Проверил:

Юдина Ольга Вадимовна

# 1. 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, 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 предыдущих работах: 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 предыдущих работах. 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, проводящих разное время за играми– отфильтруйте по этому признаку показатели успеваемости, посчитайте среднее и доверетельные интервалы. Если доверительные интервалы пересекаются – сформулируйте гипотезы. Результат позволит сделать заключение – отличается ли успеваемость игроков. Аналогично используйте возможность разделить на доли. В отчете должны быть приведены: вид используемого теста (одновыборочный , двухвыборочный, ранговый и т.д.), формулировки нулевой и альтернативной гипотез, вычисленные и взятые из таблиц статистики.

Должны быть приведены 2-3 варианта гипотез о среднем и о доле. Не забудьте о контрольных вопросах

# 2. Ход работы

На основе полученных ранее данных можно выдвинуть и проверить следующие гипотезы:

1. Часто и редко играющие дети имеют одинаковый средний балл;
2. Игровой стаж у разных полов различен;
3. Игроки со стажем в день проводят отличное от новичков время за компьютерными играми;
4. Дети из бедных и обеспеченных семей учатся по-разному.

Проверка гипотез

1. Часто и редко играющие дети имеют одинаковый средний балл

*H0* – Часто и редко играющие дети имеют одинаковый средний балл

*H1* – Часто и редко играющие дети имеют одинаковый средний балл

Уровень значимости – 5%.



|0.272| < |1.647|, следовательно, гипотеза *H0* не отклоняется

1. Один из полов имеет больший игровой стаж, чем другой

*H0* – Разные пола имеют одинаковый средний игровой стаж

*H1* – Разные пола имеют различный средний игровой стаж

Уровень значимости – 5%.



|-4.621| > |1.647|, следовательно, гипотеза *H0* отклоняется

1. Игроки со стажем в день проводят больше времени за компьютерными играми;

*H0* – Игроки с малым стажем и большим в среднем играют одинаково долго

*H1* – Игроки с малым стажем и большим в среднем играют по разному долго

Уровень значимости – 5%.



|-22.996| > |1.647|, следовательно, гипотеза *H0* отклоняется

1. Дети из бедных и обеспеченных семей учатся по-разному

*H0* – Дети из бедных и обеспеченных семей имеют одинаковый средний балл

*H1* – Дети из бедных и обеспеченных семей имеют разный средний балл

Уровень значимости – 5%.



|-1.953| > |1.648|, следовательно, гипотеза *H0* отклоняется

Выводы:

* У детей, часто играющих в игры и редко играющих в игры одинаковый средний балл.
* Ученики разных полов имеют различный в играх;
* Игроки, что начали играть недавно играют не одинаковое время в день с теми, кто начал играть давно;
* Успеваемость детей из обеспеченных и бедных семей различна.

Контрольные вопросы:

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

*H0* – гипотеза о равном среднем значении в двух выборках

1. Как формируется альтернативная гипотеза для теста Манна-Уитни?

*H1* – гипотеза о неоднородности двух выборок

1. В чём различие в применении t-тестов и ранговых тестов? Приведите примеры ранговых тестов.

Ранговые тесты используют ранги вместо фактических значений. Тест Спирмена, U-критерий Манна-Уитни.

1. Где можно найти проверочную статистику?

В таблице