Институт информационных технологий

Кафедра: Математическое и программное обеспечение ЭВМ

Дисциплина: Основы Data Science

ЛАБОРАТОРНАЯ РАБОТА № 2

Тема: Структура и описание данных. Описательные характеристики

Выполнил:

студент гр. 1ПИб-02-2оп-23

Кринкин Олег Алексеевич

Проверил:

Юдина Ольга Вадимовна

# 1. Задание

**Цель**. На основе результатов предыдущей работы составить представление о структуре и распределении данных.

Задача этапа – понять, что именно содержится в этих данных, какие факты они описывают, понять, что они могут дать.

**Задания**

Вспомните формулировки задач (исследований), которые вы сделали на прошлом занятии. Запишите предположения, которые нужно проверить. Подумайте, что может их подтвердить или опровергнуть.

Например, мы предполагаем, что в игры «с головой» чаще уходят дети, у которых нет интереса к учебе, а родители тоже в учебе не были заинтересованы.

Что может подтверждать эти предположения?

*Например, определим чрезмерную увлеченность. Как это можно сделать? Посмотрим* *на распределение – можно ли выделить группу «игроков» по времени? А другие группы?* *Можем в этих группах посчитать среднее время в игре, среднюю частоту…Посчитать* *доверительные интервалы – сопоставить*. Нужна проверка гипотез о среднем?

Используя результаты прошлой работы по структурированию данных. а также с учетомпостроенных описаний:

Для выделенных для сравнения групп, вычислите (количественных) подходящие средние значения, доверительные интервалы меры разброса

Для интересующих категорий – доли и их доверительные интервалы.

*Например, совпадают или нет доли самых активных игроков среди лучших и не* *лучших? С начала посчитаем доли и доверительные интервалы, а затем примем решение,* *нужна ли проверка статистических гипотез*

**Помните, что в исследовании должны быть задействованы максимум переменных**

Заключение: какие выводы можно сделать по вашей работе

# 2. Ход работы

Сформулированы следующие гипотезы о долях:

1. Среди часто играющих доля из бедных и богатых семей одинакова;
2. Среди редко играющих доля учеников разных полов одинакова;
3. В семьях с малообразованными родителями доля часто играющих и мало играющих различны;
4. Доля много играющих и мало играющих в разных школах различна.

Проверка гипотез:

Гипотезы проверялись при помощи двух выборочного z-теста, вычисляемого при помощи функции пакета анализа «Двухвыборочный z-тест для средних».

Дисперсия для функции вычислена по формуле:

1. Среди часто играющих доля из бедных и богатых семей одинакова

*H0*: Среди часто играющих доля из бедных и богатых семей одинакова  
*H1*: Среди часто играющих доля из бедных и богатых семей различна  
Уровень значимости – 5%.



|-0.629| < |1.959| – гипотеза *H0* не отклоняется

1. Среди редко играющих доля учеников разных полов одинакова

*H0*: Среди редко играющих доля учеников разных полов одинакова  
*H1*: Среди редко играющих доля учеников разных полов различна  
Уровень значимости – 5%.



|-7.657| > |1.959| – гипотеза *H0* отклоняется

1. В семьях с малообразованными родителями доля часто играющих и мало играющих различны  
   *H0*: В семьях с малообразованными родителями доля часто играющих и мало играющих одинаковы  
   *H1*: В семьях с малообразованными родителями доля часто играющих и мало играющих различны

Уровень значимости – 5%.



|-24.109| > |1.959| - гипотеза *H0* для отцов отклоняется



|-29.928| > |1.959| - гипотеза *H0* для матерей отклоняется

1. Доля много играющих и мало играющих в разных школах различна

*H0*: Доля много играющих и мало играющих в разных школах одинакова  
*H1*: Доля много играющих и мало играющих в школах различна

Уровень значимости – 5%.

Гипотеза была проверена для нескольких школ:

Школа 1:



|-21.249| > |1.959| - гипотеза *H0* для школы 1 отклоняется

Школа 2:



|-18.876| > |1.959| - гипотеза *H0* для школы 2 отклоняется

Школа 4:



|-25.938| > |1.959| - гипотеза *H0* для школы 4 отклоняется

Школа 8:



|-25.583| > |1.959| - гипотеза *H0* для школы 8 отклоняется

Для других школ z-тест не может быть применён, так как размеры выборки меньше, чем 30. Значения z-теста были довольно высоки во всех школах, что может значить что разница в долях частоты игры в разных школах одинаково высока.

Выводы:

* Среди часто играющих доля из бедных и богатых семей одинакова;
* Среди редко играющих доля учеников разных полов различна;
* В семьях с малообразованными родителями доля часто играющих и мало играющих различны;
* Доля много играющих и мало играющих в разных школах одинаково сильно различна.