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 №2

«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

Задание на лабораторную работу

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, 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 предыдущей работе:

* Подумайте, какие статистические гипотезы могут подтвердить или опровергнуть ваши предположения.
* Сформулируйте гипотезы о равенстве выборочных средних и долей категориальных, разделяя переменные. Используйте закономерности, особенности, которые вы смогли обнаружить в предыдущей работе. Проследите правильность формулировок гипотез Н0 и Н1. 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, имеющих двоих и более друзей и тех, у кого друзей меньше – отфильтруйте по этому признаку показатели успеваемости, посчитайте среднее и доверительные интервалы. Если доверительные интервалы пересекаются – сформулируйте гипотезы. Результат позволит сделать заключение – отличается ли успеваемость студентов, которые общаются мало и их более общительных товарищей

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

Гипотезы:

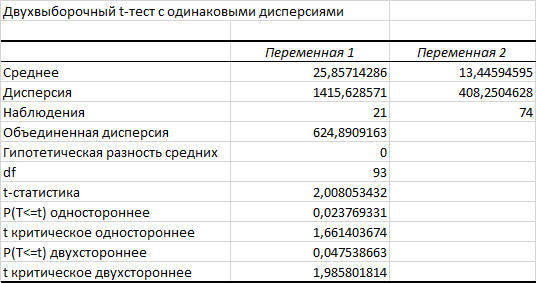
1. У здоровых людей среднее количество друзей такое же, как и у людей, страдающих депрессией
2. У здоровых людей среднее количество часов сна такое же, как и у людей, страдающих депрессией
3. У людей без проблем по учебе среднее количество часов сна такое же, как и у людей с проблемами
4. У людей, которые хорошо учатся, среднее количество часов сна такое же, как и у людей, которые учатся плохо
5. Доля мужчин и женщин, страдающих депрессией, одинакова

Уровень значимости 5% = 0,05

Проверка гипотез

1.

H0 – У здоровых людей среднее количество друзей такое же, как и у людей, страдающих депрессией

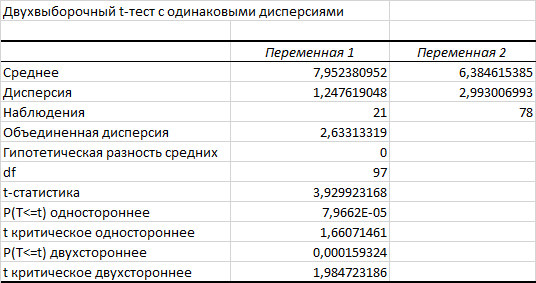
H1 – У здоровых людей среднее количество друзей больше, чем у людей, страдающих депрессией 

|2| > |1,66|, следовательно, гипотеза H0 отклоняется

2.

H0 – У здоровых людей среднее количество часов сна такое же, как и у людей, страдающих депрессией

H1 – У здоровых людей среднее количество часов сна больше, чем у людей, страдающих депрессией



|3,93| > |1,66|, следовательно, гипотеза H0 отклоняется

3.

H0 – У людей без проблем по учебе среднее количество часов сна такое же, как и у людей с проблемами

H1 – У людей без проблем по учебе среднее количество часов сна больше, чем у людей с проблемами



|0,84| < |1,66|, следовательно, гипотеза H0 принимается

4.

H0 – У людей, которые хорошо учатся, среднее количество часов сна такое же, как и у людей, которые учатся плохо

H1 – У людей, которые хорошо учатся, среднее количество часов сна больше, чем у людей, которые учатся плохо



|-1,14| < |1,66|, следовательно, гипотеза H0 принимается

5.

H0 – Доля мужчин и женщин, страдающих депрессией, одинакова

H1 – Доля мужчин и женщин, страдающих депрессией, отличается

= 25/43 = 0,58

= 9/56 = 0,16

= 34/99 = 0,34

4,37 > 1,96, следовательно, гипотеза H0 отклоняется

Контрольные вопросы:

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

Нулевая гипотеза в тестах для численных переменных обычно формулируется так, чтобы предположить отсутствие различий между группами.

Например, если при проведении теста на сравнение средних значений двух групп, нулевая гипотеза будет утверждать, что средние значения в этих группах равны. Если результаты теста позволяют отвергнуть нулевую гипотезу, это означает, что между группами есть статистически значимые различия.

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

U-критерий Манна-Уитни – непараметрический статистический критерий, используемый для сравнения 2 независимых выборок по уровню какого-либо признака, измеренного количественно. Как и t-тест, он пытается определить, есть ли отличие первой выборки от второй. Он меряет однородность этих выборок.

Метод Манна-Уитни основан на определении того, достаточно ли мала зона перекрещивающихся значений между 2 вариационными рядами (ранжированным рядом значений параметра в первой выборке и таким же во второй). Чем меньше значение критерия, тем вероятнее, что различия между значением параметра достоверны.

Для теста Манна-Уитни альтернативная гипотеза обычно формулируется так, чтобы утверждать наличие статистически значимых различий между двумя независимыми выборками.

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

Главное различие между t-тестами и ранговыми тестами заключается в том, как они обрабатывают данные и какие предположения о распределении данных они делают.

t-тесты предназначены для сравнения средних значений между двумя группами. Они предполагают, что данные распределены нормально и имеют одинаковые дисперсии. t-тесты более чувствительны к выбросам и требуют более строгих условий для применения.

Ранговые тесты не основаны на предположениях о нормальном распределении данных и могут использоваться, когда данные не соответствуют этому распределению или когда у вас маленький объем выборки. Вместо того чтобы использовать сами значения, ранговые тесты работают с рангами данных. Они менее чувствительны к выбросам и могут быть более устойчивыми к нарушению условий.

Примеры ранговых тестов:

* Анализ статистических связей
* Критерий Манна-Уитни
* Дисперсионный анализ
* Метод Краскела-Уоллиса

1. Где можно найти проверочную статистику?

Проверочную статистику можно найти в результате применения статистических тестов или анализа данных. Это числовая величина, которая используется для принятия решений о статистической значимости гипотезы на основе данных из выборки.

Вывод:

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

В результате проведения тестов были подтверждены гипотезы:

* У людей без проблемам по учебе среднее количество часов сна такое же, как и у людей с проблемами
* У людей, которые хорошо учатся, среднее количество часов сна такое же, как и у людей, которые учатся плохо

Остальные гипотезы были опровергнуты:

* У здоровых людей среднее количество друзей такое же, как и у людей, страдающих депрессией
* У здоровых людей среднее количество часов сна такое же, как и у людей, страдающих депрессией
* Доля мужчин и женщин, страдающих депрессией, одинакова