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-02-1оп-22 |
|  | *группа* |
|  | Харламов Д.А. |
|  | *ФИО* |
| Руководитель | Юдина О.В. |
|  | *ФИО преподавателя* |
| Оценка |  |
| Подпись |  |

2024 год

**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 предыдущей работе: 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 предыдущей работе.

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, 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.

Ход работы

Проверка гипотезы по доле

Известно, что доля студентов сталкивающихся с трудностями в учебе в равна 0,28. Доля перерабатывающих среди работников с низким уровнем работы (1) равна 0,38. Можно ли сделать вывод, что доля перерабатывающих работников с низким уровнем работы (1) больше всех?

Нулевая гипотеза: доля перерабатывающих работников с низким уровнем работы (1) равна доле перерабатывающих в компании:

H0: p = 0,28

Альтернативная гипотеза: доля перерабатывающих работников с низким уровнем работы (1) выше:

H1: p > 0,28

Испытание одностороннее, проведём его на 5%-ном уровне значимости.

Проверочная статистика

Критическое значение по таблице Стьюдента при n = 415 и α = 0,05 равно 1,965696725

|8,5| > |1,965696725|

Вывод: значение проверочной статистики меньше критического значения, то H0 отклоняется. С 95% вероятностью доля перерабатывающих среди работников с низким уровнем работы больше среднего значения.

Контрольные вопросы:

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

Нулевая гипотеза — принимаемое по умолчанию предположение о том, что не существует связи между двумя наблюдаемыми событиями, феноменами. Так, нулевая гипотеза считается верной, пока нельзя доказать обратное. Для численных переменных нулевая гипотеза это предположение о связи выборки и генеральной совокупности

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

Альтернативная гипотеза заключается в том, что распределения не идентичны.

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

Отличие t-теста от рангового теста заключается в том, что в t-тестах используются значения единиц наблюдения, в ранговых тестах – ранги (за ранг берётся номер элемента в упорядоченной). Пример рангового теста - коэффициент корреляции Спирмена.

1. Где можно найти проверочную статистику?

Проверочную статистику можно найти в статистических таблицах, например, таблица Стьюдента.