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 №2

Описание данных. Статистический вывод

|  |  |
| --- | --- |
| Руководитель | Юдина О. В. |
|  | ФИО преподавателя |
| Исполнитель |  |
| студент | 1ПИб-02-2оп-22 |
|  | группа |
|  | Тихомиров Владислав |
|  | Валерьевич |
|  | Фамилия, имя, отчество |
| Оценка |  |
| Подпись |  |

Череповец, 2024 г.

**Цель:**

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.

**Задания:**

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 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

В отчете должны быть приведены: вид используемого теста (одно выборочный, двух выборочный, ранговый и т.д.), формулировки нулевой и альтернативной гипотез, вычисленные и взятые из таблиц статистики.

Для подтверждения и опровержения гипотез, будем использовать Двух выборочный t-тест с одинаковыми дисперсиями. Гипотеза 1 будет наименована, как H1, гипотеза 2 - H2.

Ход работы:

Тест 1.

Гипотезы:

H0 – Девушки спят столько же, сколько и парни

H1 – Девушки спят больше/меньше парней



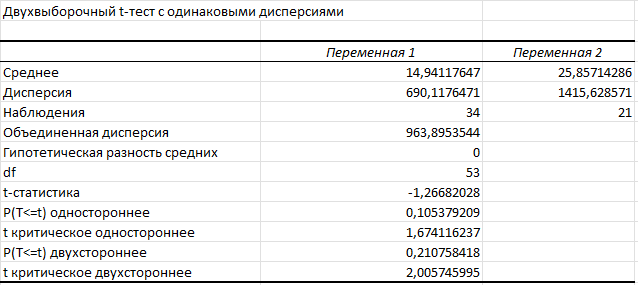
|3,69|>|1,98|-Нулевая гипотеза отклоняется.

Девочки спят больше/меньше мальчиков

Тест 2.

H0 – У людей с депрессией столько же друзей, сколько и у людей без неё

H1 – У людей без депрессии друзей больше/меньше



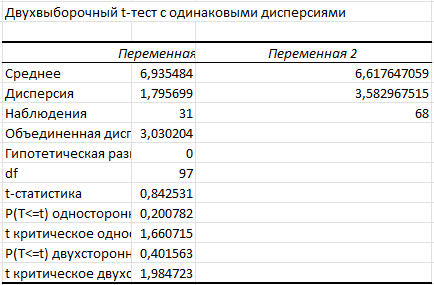
|-1,2|<|2|-Нулевая гипотеза принимается.

У людей с депрессией столько же друзей, сколько у людей без неё.

Тест 3.

H0 – Те, кто справляется с трудностями спят больше/меньше чем те, кто не справляется с трудностями

H1 – Те, кто справляется с трудностями спят одинаково как и те, кто не справляется с трудностями.



|0,8|<|1,98| Нулевая гипотеза принимается.

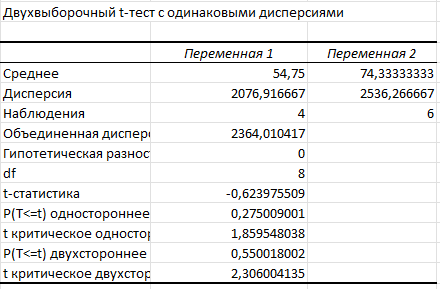
Те, кто справляются с трудностями спят больше/меньше нежели те, кто не справляется с трудностями

Тест 4.

Н0 – Люди, которым нравятся презентации, спят столько же сколько и люди которым не нравятся презентации

H1 – Люди, которые любят презентации спят меньше

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Сумма по полю SleepPerDayHours** | **Названия столбцов** |  |  |  |  |  |  |  |
| **Названия строк** | **4** | **5** | **6** | **7** | **8** | **10** | **12** | **Общий итог** |
| No |  | 45 |  | 14 | 120 | 40 |  | 219 |
| Yes | 44 | 55 | 48 | 147 | 128 |  | 24 | 446 |
| **Общий итог** | **44** | **100** | **48** | **161** | **248** | **40** | **24** | **665** |



|-0,6|<|2,3| Нулевая гипотеза принимается.

Люди, которым нравятся презентации, спят столько же сколько и люди которым не нравятся презентации

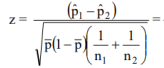
Тест 5.

H0 – Людей с депрессией готовых выполнять задачи столько же сколько людей без депрессии

H1 – Доля людей с депрессией готовых выполнять задачи отличается от людей без депрессии

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Количество по полю FaceChallangesToCompleteAcademicTask** | **Названия столбцов** |  |  |  |
| **Названия строк** | **No** | **Sometimes** | **Yes** | **Общий итог** |
| No | 13 | 4 | 4 | 21 |
| Yes | 7 | 12 | 15 | 34 |
| **Общий итог** | **20** | **16** | **19** | **55** |

|  |  |
| --- | --- |
| p1= | 0,19047619 |
| p2= | 0,441176471 |
| p | 0,345454545 |
| t-крит | 1,985251004 |
| z | -14,39327485 |
| |-14,4| > 1,98 | Нулевая гипотеза отклоняется |
| Доля людей с депрессией готовых выполнять задачи отличается от людей без депрессии | |

-14,39327485

Контрольные вопросы

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

В статистике, нулевая гипотеза (H0) обычно формулируется в контексте проверки статистических гипотез и предполагает отсутствие значимого эффекта или различий между группами или переменными. 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?

Для теста Манна-Уитни, который используется для сравнения средних значений двух независимых выборок, альтернативная гипотеза обычно формулируется с учетом того, что между выборками существует статистически значимая разница. Это означает, что если мы отвергаем нулевую гипотезу (которая утверждает, что средние значения выборок равны), то мы принимаем альтернативную гипотезу, которая говорит о наличии различий между выборками.

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

Проверочная статистика — это значение, которое рассчитывается на основе данных и используется для проверки нулевой гипотезы. Проверочную статистику можно найти в статистических таблицах, например, таблица Стьюдента.