САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
ФАКУЛЬТЕТ ИНФОКОММУНИКАЦИОННЫХ ТЕХНОЛОГИЙ

Отчет по лабораторной работе №2

по курсу «Современные инструменты анализа данных»

Выполнили:
Гусейнова М. Э.,
Евдокимова У. В.,
Кадникова Е. М.,
Платонова А. С.
Проверила: Максимова Т. Г.

Санкт-Петербург
2024 г.

Задание 2.1

Проверить гипотезу о статистической значимости различия между доходами двух групп работающих и получающих доход граждан Петербурга:
1 группа - имеющие образование среднее и ниже,
2 группа - имеющие среднее специальное или высшее образование
Представить в отчете ход проверки гипотезы (рисунки, таблицы) и выводы.
Для формирования выборки использовать признаки:

- Регион
- ЗАКОНЧЕННОЕ ОБРАЗОВАНИЕ (ГРУППА)
- Ваше основное занятие в настоящее время? = 1 | Вы сейчас работаете
- Сколько денег в течение последних 30 дней Вы получили по основному месту работы после вычета налогов и отчислений? Если все или часть денег Вы получили в иностранной валюте, переведите все в рубли и назовите общую сумму

Для проверки гипотезы использовать jamovi (можно Python), ДА/DA, однофакторный дисперсионный анализ, Т-тесты.

Для проверки этой и последующих гипотез был выбран язык Python. Было рассмотрено два варианта проверки гипотезы: однофакторный дисперсионный анализ и Т-тест для независимых выборок. В первом случае проверка осуществлялась по критерию Фишера, во втором - по критерию Стьюдента.

Были выдвинуты следующие гипотезы:

- Нулевая гипотеза (H_0): Средние доходы двух групп не различаются или различия между доходами групп статистически незначимы.
- Альтернативная гипотеза (H_1): Средние доходы двух групп статистически значимо различаются.

<u>Вариант 1</u>

Описательная статистика для 1 группы (среднее и ниже):

```
count       61.000000
mean     29785.245902
```

```
std      13170.457295
min       4300.000000
25%      20000.000000
50%      28000.000000
75%      36000.000000
max      70000.000000
```

Описательная статистика для 2 группы (среднее специальное и выше):

```
count        88.000000
mean      35835.227273
std       14646.399579
min        8000.000000
25%       25000.000000
50%       34000.000000
75%       44250.000000
max      100000.000000
```

Результаты однофакторного дисперсионного анализа (ANOVA):

```
F-статистика: 6.668023048918451
p-значение: 0.010792595536295 53
```

Отклоняем нулевую гипотезу. Доходы двух групп статистически значимо различаются.

На основе результатов однофакторного дисперсионного анализа (ANOVA) можно сделать следующие выводы:

Значения статистики:

- **F-статистика**: 6.68. Это значение указывает на соотношение между группами и вариацией внутри групп. Более высокое значение F-статистики говорит о большем различии между группами по сравнению с изменчивостью внутри групп.
- **p-значение**: 0.0108. Это значение указывает на вероятность получения таких же или более экстремальных результатов, если нулевая гипотеза верна.

Заключение

Таким образом, результаты анализа ANOVA подтверждают, что уровень образования имеет значительное влияние на доходы работающих граждан Петербурга.

Вариант 2

Описательная статистика для 1 группы (среднее и ниже):

```
count       61.000000
mean     29785.245902
std      13170.457295
min       4300.000000
25%      20000.000000
50%      28000.000000
75%      36000.000000
max      70000.000000
```

Описательная статистика для 2 группы (среднее специальное и выше):

```
count        88.000000
mean      35835.227273
std       14646.399579
min        8000.000000
25%       25000.000000
50%       34000.000000
75%       44250.000000
max      100000.000000
```

Результаты T-теста:

```
T-статистика: -2.632588514983112
p-значение: 0.009444010854348454
```

Отклоняем нулевую гипотезу. Доходы двух групп статистически значимо различаются, что видно и на графике на рисунке 1.

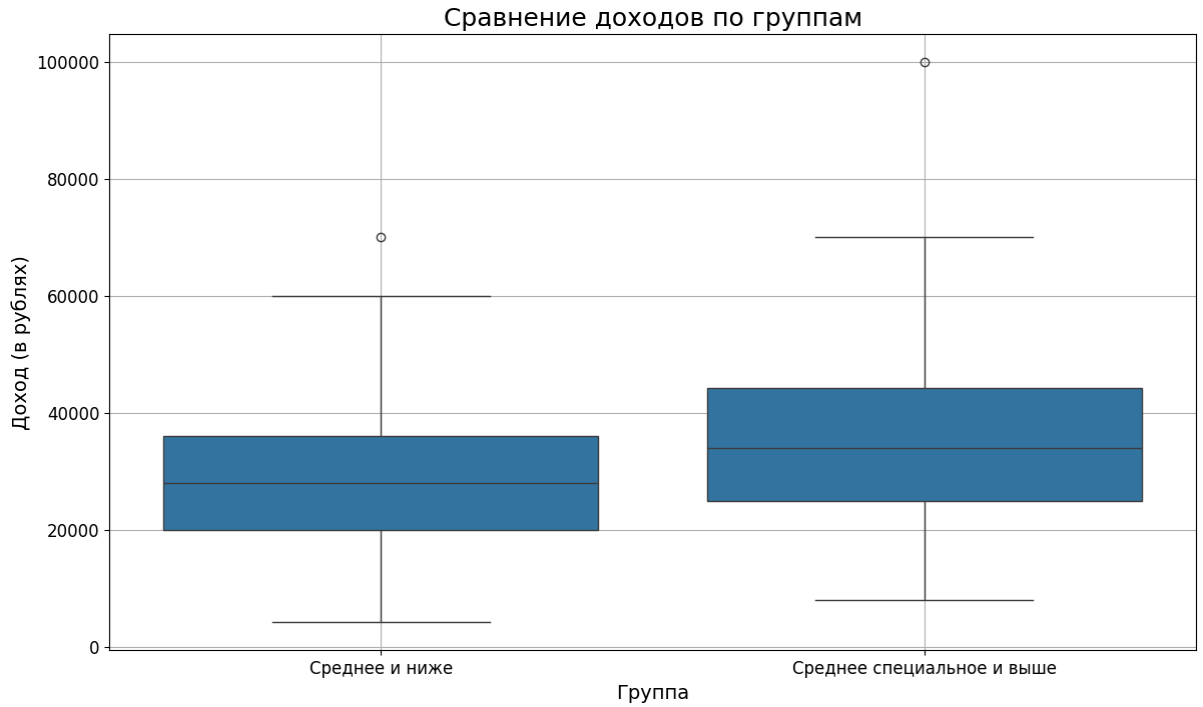

Рисунок 1 - График к первой гипотезе

На основе результатов Т-теста можно сделать следующие выводы:

Значения статистики:

- **Т-статистика**: -2.63. Это значение показывает, насколько стандартные ошибки отличаются от нуля в сравнении между группами. Отрицательное значение Т-статистики указывает на то, что средний доход первой группы (среднее и ниже) ниже, чем у второй группы (среднее специальное и выше).
- **p-значение**: 0.0094. Это значение показывает вероятность того, что наблюдаемое различие в доходах между группами произошло случайно, если нулевая гипотеза верна.

Результаты анализа:

- Поскольку p-значение (0.0094) меньше уровня значимости $\alpha = 0.05$, мы отклоняем нулевую гипотезу. Это свидетельствует о том, что существует статистически значимое различие в доходах между двумя группами.

Аналогичные результаты видим на боксплоте, который показывает, что и среднее значение, и максимальное значение, и минимальное значение зарплаты во второй группе выше, чем в первой.

Заключение

Таким образом, результаты Т-теста подтверждают наличие статистически значимого различия в доходах между группами работающих граждан Петербурга в зависимости от уровня образования.

Задание 2.2

Для выделенных ранее групп проверить гипотезу о равенстве средней продолжительности работы в неделю.
Выполнять только в jamovi (можно Python)
Представить в отчете ход проверки гипотезы (рисунки, таблицы) и выводы.

Были выдвинуты следующие гипотезы:

- Нулевая гипотеза (H0): Средние значения продолжительности рабочей недели для двух групп не отличаются.
- Альтернативная гипотеза (H1): Средние значения продолжительности рабочей недели отличаются.

Описательная статистика для 1 группы (среднее и ниже)

```
count    55.000000
mean     41.000000
std       9.972184
min       8.000000
25%      40.000000
50%      40.000000
75%      48.000000
max      60.000000
```

Описательная статистика для 2 группы (среднее специальное и выше)

```
count    78.000000
mean     41.448718
std       6.872961
min      24.000000
25%      40.000000
50%      40.000000
75%      43.750000
max      60.000000
```

Результаты T-теста:
```
T-статистика: -0.475884714777291
p-значение: 0.6348629818901632
```

Не удалось отклонить нулевую гипотезу. Статистически значимых различий в средней продолжительности рабочей недели между группами нет.

Выводы:

T-статистика: -0.476. Значение T-статистики говорит о том, что средние значения рабочих часов для обеих групп находятся близко друг к другу. Положительные и отрицательные значения T-статистики указывают на то, в какую сторону (в большую или меньшую сторону) смещено одно среднее значение относительно другого. В вашем случае оно близко к нулю, что подтверждает, что различия незначительны.

p-значение: 0.635. Это значение значительно больше стандартного уровня значимости 0.05, что означает, что у нас недостаточно статистических оснований для отклонения нулевой гипотезы. Это подтверждают и графики на рисунке 2.

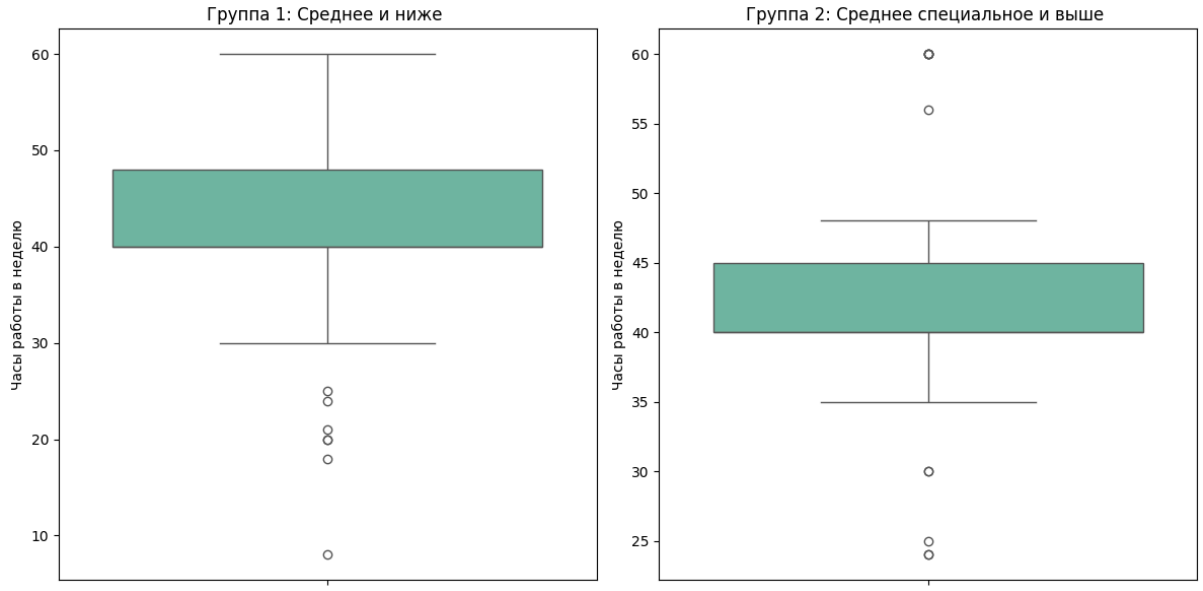

Рисунок 2 - Графики ко второй гипотезе

Задание 2.3

Для работающих, указавших продолжительность работы и получающих доход граждан (проживающих в любом населенном пункте) исследовать взаимосвязь двух признаков: курение и употребление алкоголя за последние 30 дней.
Выполнять только в jamovi (можно Python)
Использовать Частотный анализ, Таблицы сопряженности парных выборок.
Представить в отчете ход проверки гипотезы (рисунки, таблицы) и выводы.

Были выдвинуты следующие гипотезы:

- Нулевая гипотеза (H0): Нет статистически значимой связи между курением и употреблением алкоголя.
- Альтернативная гипотеза (H1): Существует статистически значимая связь между курением и употреблением алкоголя.

Таблица сопряженности парных выборок:

Употребляли алкоголь за последние 30 дней	**Да**	**Нет**
Курят		
Да	954	253
Нет	1245	749

Результаты Хи-квадрат теста:

```
Хи-квадрат: 95.59611320423234
p-значение: 1.408827425755191e-22
```

p-значение меньше 0.05, следовательно **отклоняем** нулевую гипотезу.

Вывод:
Существует статистически значимая связь между курением и употреблением алкоголя, что подтверждают и графики на рисунках 3 и 4.

Рисунок 3 - Тепловая карта к третьей гипотезе (1 - Да, 2 - Нет)

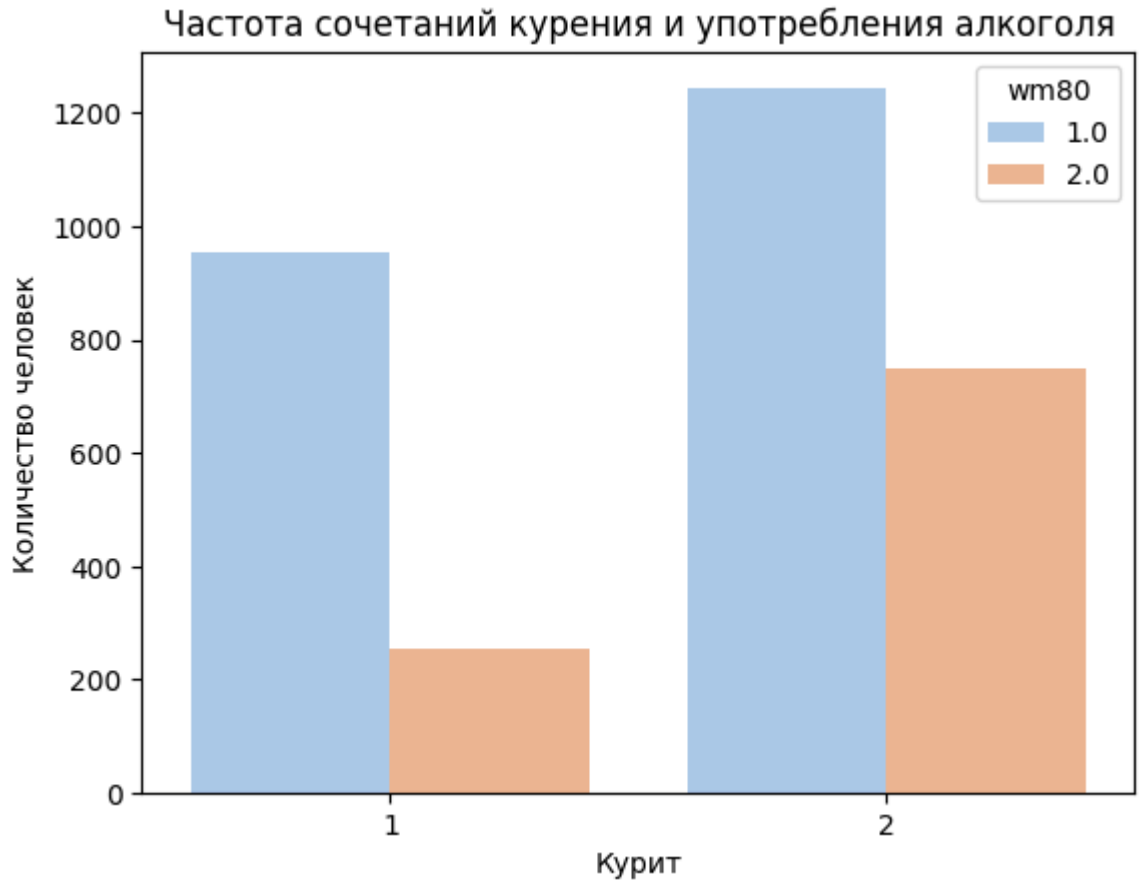

Рисунок 4 - Столбчатая диаграмма к третьей гипотезе (wm80 - Употребление алкоголя, 1 - Да, 2 - Нет)