Лабораторная работа №4

Выполнил: Сайфуллин Динислам

Вариант 2

В данной работе проведён статистический анализ результатов экзаменов студентов (N=1000) по трём предметам: математика, чтение и письменная часть. Цели:

1. Проверить равномерность распределения участников по признакам «пол» и «этническая группа».
2. Сравнить зависимые выборки: средние оценки по математике и по письменной части.
3. Оценить влияние прохождения подготовительного курса на успеваемость.

Данные представлены в csv файле:

- **Файл:** `exams_dataset.csv`
- **Параметры:**
  - `gender` — пол (male/female)
  - `race/ethnicity` — этническая группа (group A…E)
  - `math score`, `reading score`, `writing score` — баллы по трём предметам
  - `test preparation course` — курс подготовки (completed/none)

Сформулируем гипотезы:

1. Распределение по полу
   - H_0: доли мужчин и женщин равны ($p_{male} = p_{female} = 0.5$)
   - H_1: доли отличаются ($p_{male} \neq p_{female}$)
2. Распределение по этническим группам
   - H_0: доли по группам A–E равны ($p_i = 1/5$)
   - H_1: хотя бы одна доля отличается
3. Однородность результатов
   - H_0: $\mu_{math} - \mu_{writing} = 0$
   - H_1: $\mu_{math} - \mu_{writing} \neq 0$
4. Влияние прохождения курсов (completed vs none)
   Для каждого предмета:
   - H_0: $\mu_{completed} = \mu_{none}$
   - H_1: $\mu_{completed} \neq \mu_{none}$

In [ ]:
```
import pandas as pd
from scipy.stats import chisquare, ttest_rel, ttest_ind, chi2, binomtest, t
import numpy as np

df = pd.read_csv('exams_dataset.csv')

obs_gender = df['gender'].value_counts().sort_index()  # female, male
n = obs_gender.sum()

exp_gender = pd.Series([n/2, n/2], index=obs_gender.index)
chi2_stat = ((obs_gender - exp_gender)**2 / exp_gender).sum()
p_chi2_gender = chi2.sf(chi2_stat, df=1)

res_binom = binomtest(obs_gender['female'], n=n, p=0.5)

print("Распределение по полу")
print("Наблюдаемые:", obs_gender.to_dict())
print(f"1a) X² = {chi2_stat:.3f}, p = {p_chi2_gender:.3f}")
print(f"1b) binomtest: p = {res_binom.pvalue:.3f}\n")

obs_eth = df['race/ethnicity'].value_counts().sort_index()
k = len(obs_eth)

exp_eth = np.full(k, n/k)
chi2_eth_stat = ((obs_eth.values - exp_eth)**2 / exp_eth).sum()
p_chi2_eth = chi2.sf(chi2_eth_stat, df=k-1)

chi2_scipy_eth, p_scipy_eth = chisquare(obs_eth.values, f_exp=exp_eth)

print("Распределение по этническим группам")
print(obs_eth.to_string())
print(f"2a) X² = {chi2_eth_stat:.3f}, p = {p_chi2_eth:.3f}")
```

```
print(f"2b) SciPy X² = {chi2_scipy_eth:.3f}, p = {p_scipy_eth:.3f}\n")

diff = df['math score'] - df['writing score']
n_pairs = len(diff)

mean_diff = diff.mean()
sd_diff = diff.std(ddof=1)
t_stat_manual = mean_diff / (sd_diff/np.sqrt(n_pairs))
p_manual = 2 * t.sf(abs(t_stat_manual), df=n_pairs-1)

t_stat_scipy, p_scipy = ttest_rel(df['math score'], df['writing score'])

print("Парный t-тест (Math/Writing)")
print(f"3a) manual t = {t_stat_manual:.3f}, p = {p_manual:.3e}")
print(f"3b) SciPy t = {t_stat_scipy:.3f}, p = {p_scipy:.3e}\n")

grp_c = df[df['test preparation course']=='completed']
grp_n = df[df['test preparation course']=='none']

print("Влияние подготовительного курса")
for subj in ['math score','reading score','writing score']:
    x1 = grp_c[subj]
    x2 = grp_n[subj]
    n1, n2 = len(x1), len(x2)
    m1, m2 = x1.mean(), x2.mean()
    s1, s2 = x1.std(ddof=1), x2.std(ddof=1)

    se = np.sqrt(s1**2/n1 + s2**2/n2)
    t_manual = (m1 - m2) / se
    df_welch = (s1**2/n1 + s2**2/n2)**2 / ((s1**2/n1)**2/(n1-1) + (s2**2/n2)**2/(n2-1))
    p_manual = 2 * t.sf(abs(t_manual), df=df_welch)

    t_scipy, p_scipy = ttest_ind(x1, x2, equal_var=False)
    subj_name = subj.replace(' score','').title()
    print(f"{subj_name}:")
    print(f"4a) manual t = {t_manual:.3f}, p = {p_manual:.3e}")
    print(f"4b) SciPy t = {t_scipy:.3f}, p = {p_scipy:.3e}\n")
```

```
Распределение по полу
Наблюдаемые: {'female': 494, 'male': 506}
1a) X² = 0.144, p = 0.704
1b) binomtest: p = 0.728

Распределение по этническим группам
race/ethnicity
group A     77
group B    204
group C    324
group D    261
group E    134
2a) X² = 192.990, p = 0.000
2b) SciPy X² = 192.990, p = 0.000

Парный t-тест (Math vs Writing)
3a) manual t = -6.653, p = 4.737e-11
3b) SciPy t = -6.653, p = 4.737e-11

Влияние подготовительного курса
Math:
4a) manual t = 5.577, p = 3.354e-08
4b) SciPy t = 5.577, p = 3.354e-08

Reading:
4a) manual t = 8.136, p = 1.503e-15
4b) SciPy t = 8.136, p = 1.503e-15

Writing:
4a) manual t = 11.063, p = 1.311e-26
4b) SciPy t = 11.063, p = 1.311e-26
```

Результаты

1. Распределение по полу

- $X^2 = 0.144, p_{value} = 0.704$
- $p_{value} = 0.728$
- Оба $p \geq 0.05$ $\rightarrow$ нет оснований отвергать нулевую гипотезу о равных долях мужчин и женщин.

2. Распределение по этническим группам

- $X^2 = 192.990, p_{value} < 0.001$

- $X^2 = 192.990$, $p_{value} < 0.001$
- $p < 0.05 \rightarrow$ распределение по этническим группам существенно отличается от равномерного.

3. Сравнение математики и письма

- t = –6.653, p_{value} = 4.737 × 10^{-11}
- t = –6.653, p_{value} = 4.737 × 10^{-11}
  p < 0.05 \rightarrow средние оценки по математике и по письменной части статистически различаются. Письменная часть в среднем выше.

4. Влияние прохождения подготовительного курса

- Math:
  - p_{value} = 3.354 × 10^{-8}
  - p_{value} = 3.354 × 10^{-8}
- Reading:
  - p_{value} = 1.503 × 10^{-8}
  - p_{value} = 1.503 × 10^{-8}
- Writing:
  - p_{value} = 1.311 × 10^{-8}
  - p_{value} = 1.311 × 10^{-8}
- Для всех трёх предметов p < 0.05 \rightarrow прохождение подготовительного курса статистически значимо повышает результаты экзаменов.

Выводы

1. **Распределение по полу**
   Половая структура выборки близка к равномерной (нет значимого отклонения от пропорции 50/50).

2. **Распределение по этническим группам**
   Наблюдается значимая неравномерность: группы C и D представлены сильнее, чем ожидалось при равномерном распределении.

3. **Сравнение математики и письменной части**
   Средние оценки по письменной части существенно выше, чем по математике (t = –6.653, p < 0.05).

4. **Эффект подготовительного курса**
   Студенты, прошедшие курс, демонстрируют статистически значимо более высокие баллы по всем трём предметам, что подтверждает эффективность подготовки.