

Статистический анализ гипотез на данных вакансий и резюме

- Провел: Мамонтов Олег Дмитриевич, Б23-215

Методология

Предобработка данных

Использовались методы описательной статистики, t-тесты, доверительные интервалы и биномиальные тесты. Уровень значимости $\alpha = 0.05$ установлен для всех тестов.

```
# Удаление колонок с более чем 90% пропусков
df.dropna(axis=1, thresh=len(df) - 0.90 * len(df), inplace=True)
```

Гипотеза 1: Влияние высшего образования на ожидаемую зарплату

Формулировка гипотез

- **H₀**: Медианная ожидаемая зарплата не зависит от наличия высшего образования
- **H₁**: Медианная ожидаемая зарплата зависит от наличия высшего образования

Методика проверки

Для проверки использовался двухвыборочный t-тест с расчетом доверительного интервала.

```
# Расчет ожидаемой зарплаты как среднего между min и max
df_edu["expected_salary"] = (df_edu["salaryMin_cv"] + df_edu["salaryMax_cv"]) / 2

# Разделение на группы
has_edu = df_edu["education"].str.contains('высшее', case=False, na=False)
group_edu = df_edu[has_edu]['expected_salary'].values
group_no_edu = df_edu[~has_edu]['expected_salary'].values
```

Результаты

- Кандидатов с высшим образованием: 69,610
- Кандидатов без высшего образования: 66,671
- Средняя ожидаемая зарплата (с высшим): 55,267.91 руб.
- Средняя ожидаемая зарплата (без высшего): 40,854.34 руб.

Статистические показатели:

- t-статистика: 76.5383
- Разница средних: 14,413.57 руб.

- 95% ДИ для разницы: [14,044.47, 14,782.67] руб.
- P-value: < 0.0001

Вывод

Гипотеза H_0 отвергается. Обнаружена статистически значимая разница в ожидаемых зарплатах: кандидаты с высшим образованием ожидают зарплату в среднем на 14,413.57 руб. выше.

Гипотеза 2: Влияние опыта работы на приглашение на собеседование

Формулировка гипотез

- H_0 : Статус отклика не зависит от опыта работы
- H_1 : Существует зависимость между опытом работы и статусом отклика

Методика проверки

Использовался метод сравнения доверительных интервалов для средних значений опыта работы в двух группах.

```
def confidence_interval(data, confidence=0.95):  
    n = len(data)  
    mean = np.mean(data)  
    std_err = np.std(data, ddof=1) / np.sqrt(n)  
    t_critical = stats.t.ppf(0.975, n - 1)  
    return (mean - t_critical * std_err, mean + t_critical * std_err)
```

Результаты

Приглашенные кандидаты:

- Средний опыт: 6.46 лет
- 95% ДИ: [6.36, 6.56] лет

Неприглашенные кандидаты:

- Средний опыт: 7.06 лет
- 95% ДИ: [7.02, 7.10] лет

Доверительные интервалы не пересекаются, что указывает на статистически значимую разницу.

Вывод

Гипотеза H_0 отвергается. Обнаружена статистически значимая разница: неприглашенные кандидаты имеют больший опыт работы (7.06 лет) по сравнению с приглашенными (6.46 лет).

Гипотеза 3: Соответствие ожидаемых и предлагаемых зарплат

Формулировка гипотез

- H_0 : Медианная разница между желаемой и предлагаемой зарплатой равна нулю
- H_1 : Медианная разница между желаемой и предлагаемой зарплатой не равна нулю

Методика проверки

Использовался биномиальный тест для проверки симметричности распределения разниц.

```
def manual_binomial_test(n_success, n_trials, p_h0=0.5):  
    p_obs = n_success / n_trials  
    prob_observed = stats.binom.pmf(n_success, n_trials, p_h0)  
    all_outcomes = range(0, n_trials + 1)  
    all_probs = stats.binom.pmf(all_outcomes, n_trials, p_h0)  
    extreme_probs = all_probs[all_probs <= prob_observed]  
    p_value = np.sum(extreme_probs)  
    return p_obs, p_value
```

Результаты

- Всего пар: 150,492
- Медиана разниц: 0.00 руб.
- Среднее разниц: 923.48 руб.
- Разниц больше 0: 67,946 (45.1%)
- Разниц меньше 0: 73,376 (48.8%)
- Разниц равных 0: 9,170 (6.1%)

Биномиальный тест:

- Наблюдаемая доля "ожидают больше": 0.481 (48.1%)
- P-value: 0.0000

Вывод

Несмотря на то, что медиана разниц равна нулю, биномиальный тест показывает статистически значимое отклонение от симметричного распределения ($p < 0.0001$). Кандидаты значимо чаще ожидают зарплату ниже предлагаемой (48.8% против 45.1%).

Общие выводы

1. **Образование существенно влияет** на зарплатные ожидания - кандидаты с высшим образованием ожидают на 35% большую зарплату.
2. **Опыт работы показывает парадоксальную зависимость** - кандидаты с меньшим опытом чаще получают приглашения на собеседование.
3. **Зарплатные ожидания асимметричны** - хотя медианная разница равна нулю, существует статистически значимый перекося в сторону занижения ожиданий.