

# Задание 3. Семья и работа

Тема: Влияет ли семейный статус на вероятность трудоустройства?

## 1. Загрузка и подготовка данных.

Загрузите RLMS-HSE с сайта <https://www.hse.ru/rlms> (<https://www.hse.ru/rlms>). Выберите один конкретный год (отличающийся от выбранного в задании 1).

```
library(haven)
data_2022 <- read_dta('C:\\Users\\arina\\Downloads\\r3li_os_75.dta') #Читаем данные
```

## 2. Фильтрация данных:

- Возраст: выберите только людей от 25 до 45 лет.

```
data_2022 <- data_2022[data_2022$aa_age >= 25 & data_2022$aa_age <= 45,]
```

- Переменная занятости: выберите переменную, отражающую, работает ли человек (например, по статусу занятости).

```
data_2022$job_status[data_2022$aa_j1 == 1] <- 1 #работают
```

```
## Warning: Unknown or uninitialised column: `job_status`.
```

```
data_2022$job_status[data_2022$aa_j1 != 1 & data_2022$aa_j1 != 99999997 & data_2022$aa_j1 != 99999998 & data_2022$aa_j1 != 99999999] <- 0 #на данный момент без работы, в декрете или отпуске
```

- Переменная семейного положения: выделите две группы:

– Женатые/замужние (официально зарегистрированный брак);

– Неженатые/незамужние (включая вдовых, разведённых, никогда не состоявших в браке).

```
data_2022$marriage[data_2022$aa_marst == 2 | data_2022$aa_marst == 6] <- 1 # в зарегистрированном браке
```

```
## Warning: Unknown or uninitialised column: `marriage`.
```

```
data_2022$marriage[data_2022$aa_marst == 1 | data_2022$aa_marst == 3 | data_2022$aa_marst == 4 | data_2022$aa_marst == 5] <- 0 # не в браке на данный момент
```

- Перенесите все данные в R (четыре переменные: пол; возраст; переменная, отражающая наличие работы, переменная, отражающая семейный статус).

```
job <- data_2022$job_status #создаем переменную со статусом работы (1 или 0)
marriage <- data_2022$marriage #создаем переменную со статусом замужества (1 или 0)
age <- data_2022$aa_age #создаем переменную с возрастом
gender <- data_2022$aa_h5 #создаем переменную с полом
```

## 3 Расчет долей.

- Рассчитайте долю трудоустроенных среди женатых/замужних p1;

```
#расчет отношения трудоустроенных среди замужних/женатых ко всему количеству женатых
```

```
#выбираем те строки, где объект и в браке (статус 1), и работает (статус 1)
married_employed <- data_2022[data_2022$marriage == 1 & data_2022$job_status == 1,]
```

```
total_married <- data_2022[data_2022$marriage == 1,] #выбираем те строки, где объект в браке (1)
```

```
#считаем количество соответствующих строк (объектов)
```

```
r_1 <- nrow(married_employed)
```

```
n_1 <- nrow(total_married)
```

```
p_1 <- r_1 / n_1
```

```
paste('Доля трудоустроенных среди женатых/замужних:', p_1)
```

```
## [1] "Доля трудоустроенных среди женатых/замужних: 0.796514896008994"
```

- Рассчитайте долю трудоустроенных среди неженатых/незамужних p2.

```
#расчет отношения трудоустроенных среди незамужних/неженатых ко всему количеству людей не в браке
```

```
#выбираем те строки, где объект не в браке(статус 0), но при этом работает (статус 1)
single_married <- data_2022[data_2022$marriage == 0 & data_2022$job_status == 1,]
total_single <- data_2022[data_2022$marriage == 0,] #выбираем строки, где объект не состоит в браке

#считаем количество соответствующих строк (объектов)
r_2 <- nrow(single_married)
n_2 <- nrow(total_single)
p_2 <- r_2 / n_2
paste('Доля трудоустроенных среди неженатых/незамужних:', p_2)
```

```
## [1] "Доля трудоустроенных среди неженатых/незамужних: 0.798279906176701"
```

## 4. Построение доверительного интервала для разности долей.

• Постройте доверительный интервал для разности долей трудоустроенных между женатыми/замужними и нет для уровней значимости 90%, 95% и 99%.

```
print(n_1)
```

```
## [1] 1779
```

```
print(n_2)
```

```
## [1] 1279
```

```
#так как обе выборки большие >1000 объектов, то критические значения будем брать, пользуясь графиком нормального
распределения
z_90 <- 1.64 #критические значения для разных уровней значимости
z_95 <- 1.96
z_99 <- 2.57
```

```
#разница между двумя долями
difference = p_2 - p_1
```

```
#построение доверительных интервалов для каждого уровня значимости (ci - confidence interval)
#рассчитаем по отдельности левые и правые границы интервалов
```

```
ci_90_left <- difference - z_90 * sqrt(p_1*(1 - p_1)/n_1 + p_2*(1 - p_2)/n_2)
ci_90_right <- difference + z_90 * sqrt(p_1*(1 - p_1)/n_1 + p_2*(1 - p_2)/n_2)
ci_95_left <- difference - z_95 * sqrt(p_1*(1 - p_1)/n_1 + p_2*(1 - p_2)/n_2)
ci_95_right <- difference + z_95 * sqrt(p_1*(1 - p_1)/n_1 + p_2*(1 - p_2)/n_2)
ci_99_left <- difference - z_99 * sqrt(p_1*(1 - p_1)/n_1 + p_2*(1 - p_2)/n_2)
ci_99_right <- difference + z_99 * sqrt(p_1*(1 - p_1)/n_1 + p_2*(1 - p_2)/n_2)

ci_90 <- sprintf("Confidence interval (%.3f, %.3f)", ci_90_left, ci_90_right)
print(ci_90)
```

```
## [1] "Confidence interval (-0.022, 0.026)"
```

```
ci_95 <- sprintf("Confidence interval (%.3f, %.3f)", ci_95_left, ci_95_right)
print(ci_95)
```

```
## [1] "Confidence interval (-0.027, 0.031)"
```

```
ci_99 <- sprintf("Confidence interval (%.3f, %.3f)", ci_99_left, ci_99_right)
print(ci_99)
```

```
## [1] "Confidence interval (-0.036, 0.040)"
```

• Интерпретируйте результат: каков возможный разрыв между группами?

По результатам построения доверительных интервалов для каждого из 3 уровней значимости получилось, что 0 содержится в каждом из них. Иными словами, если брать другие выборки, то в 90%, 95% и 99% из них, соответственно, доверительный интервал "поймает" 0. Таким образом, можно заключить, что разрыва между долями трудоустроенных среди людей в браке и трудоустроенных среди незамужних и неженатых нет.

## 5. Проверка гипотезы о равенстве долей.

Выполните анализ по всем 5 этапам:

### 1. Формулировка гипотез:

**Нулевая гипотеза:** доли трудоустроенных среди замужних/женатых и среди незамужних/неженатых равны (не различаются) ( $H_0: \pi_1 = \pi_2$ )

**Альтернативная гипотеза:** эти доли различаются ( $H_1: \pi_1 \neq \pi_2$ )

### 2. Выбор и расчет тестовой статистики:

Альтернативная гипотеза подразумевает неравенство, значит критерий будет *двусторонним*. Выборки большие (количество указано выше), значит используем z-значение (то есть смотрим на график нормального распределения).

Воспользуемся формулой для **тестовой статистики** для сравнения двух долей генеральных совокупностей: 
$$z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{(P * (1 - P) * (\frac{1}{n_1} + \frac{1}{n_2}))}}$$

Статистика является стандартной нормальной величиной.

```
#оцениваем совокупную долю
```

```
P <- (r_1 + r_2) / (n_1 + n_2)
```

```
#так как в H_0 полагаем равенство долей генеральных совокупностей, то \pi_1 - \pi_2 = 0
```

```
Z <- difference / sqrt(P*(1 - P)*(1 / n_1 + 1/n_2))
```

```
paste('Значение тестовой статистики равно:', Z)
```

```
## [1] "Значение тестовой статистики равно: 0.119750270473002"
```

### 3. Критическая область или p-value: выберите уровни значимости $\alpha = 0.10, 0.05, 0.01$ .

В случае двустороннего критерия критические значения будут точно такими же, как и в пункте с доверительными интервалами.

### 4. Примите решение о принятии/отклонении нулевой гипотезы.

$z = 0.11975$ , Z не попала в критическую область на уровне значимости 10%. Значит, Z не попала и в другие критические области (при уровнях значимости 5% и 1%). Таким образом,  $H_0$  не может быть отвергнута при всех  $\alpha$ .

### 5. Интерпретация: сделайте вывод, влияет ли семейный статус на занятость. Что это может означать с социальной точки зрения?

Исходя из того, что  $H_0$  не отвергается, можно подтвердить, что различия между долями нет, то есть семейный статус не влияет на занятость.

С социальной точки зрения, это может означать следующее:

- **Экономическая независимость:** Как замужние/женатые, так и незамужние/неженатые люди достигают уровня экономической независимости, позволяющей им активно участвовать в трудовом рынке. В каждой группе могут быть люди, стремящиеся к карьерным достижениям и финансовой обеспеченности независимо от их семейного положения.
- **Изменение семейных норм:** Параллельная трудовая активность может свидетельствовать о смене традиционных представлений о семье и браке. В современном обществе все чаще происходит осознание того, что брак не является необходимым условием для обеспечения материального благополучия.

## 6 Сравнение по полу.

Повторите весь анализ отдельно для мужчин и для женщин:

Мужчины:

- Разделите выборку на мужчин и женщин.

```

#фильтруем строки, в которых объект - женатый мужчина с работой
man_married_employed <- data_2022[data_2022$marriage == 1 & data_2022$aah5 == 1 & data_2022$job_status == 1,]

#фильтруем строки, в которых объект - женатый мужчина
man_total_married <- data_2022[data_2022$marriage == 1 & data_2022$aah5 == 1,]

#фильтруем строки, в которых объект - неженатый мужчина с работой
man_single_employed <- data_2022[data_2022$marriage == 0 & data_2022$aah5 == 1 & data_2022$job_status == 1,]

#фильтруем строки, в которых объект - неженатый мужчина
man_total_single <- data_2022[data_2022$marriage == 0 & data_2022$aah5 == 1,]

#фильтруем строки, в которых объект - замужняя женщина с работой
woman_married_employed <- data_2022[data_2022$marriage == 1 & data_2022$aah5 == 2 & data_2022$job_status == 1,]

#фильтруем строки, в которых объект - замужняя женщина
woman_total_married <- data_2022[data_2022$marriage == 1 & data_2022$aah5 == 2,]

#фильтруем строки, в которых объект - незамужняя женщина с работой
woman_single_employed <- data_2022[data_2022$marriage == 0 & data_2022$aah5 == 2 & data_2022$job_status == 1,]

#фильтруем строки, в которых объект - незамужняя женщина
woman_total_single <- data_2022[data_2022$marriage == 0 & data_2022$aah5 == 2,]

```

- Для каждой подгруппы (мужчины, женщины) повторите всю работу по п.5, т.е. проверьте гипотезу о равенстве долей трудоустроенных между женатыми и неженатыми мужчинами, замужними и незамужними женщинами.

#### Мужчины:

```

#считаем количество подходящих строк (объектов)
n_3 <- nrow(man_total_married)
r_3 <- nrow(man_married_employed)
p_3 <- r_3 / n_3

paste('Доля трудоустроенных среди женатых мужчин составляет:', p_3)

```

```
## [1] "Доля трудоустроенных среди женатых мужчин составляет: 0.915441176470588"
```

```

n_4 <- nrow(man_total_single)
r_4 <- nrow(man_single_employed)
p_4 <- r_4 / n_4
paste('Доля трудоустроенных среди неженатых мужчин составляет:', p_4)

```

```
## [1] "Доля трудоустроенных среди неженатых мужчин составляет: 0.819905213270142"
```

Сформулируем гипотезы. **H0:  $P_3 = P_4$**  (между долями трудоустроенных женатых мужчин и трудоустроенных неженатых мужчин *разницы нет*) и **H1:  $P_3 \neq P_4$**  (разница есть).

```
print(n_3)
```

```
## [1] 816
```

```
print(n_4)
```

```
## [1] 633
```

Так как выборки большие, используем график нормального распределения для определения критического значения.

```

P_m <- (r_3 + r_4) / (n_3 + n_4)

Z_m <- (p_3 - p_4) / sqrt(P_m*(1 - P_m)*(1/n_3 + 1/n_4))

paste('Значение тестовой статистики равно:', Z_m)

```

```
## [1] "Значение тестовой статистики равно: 5.4300723477812"
```

#### Женщины:

```
#считаем количество подходящих строк (объектов)
```

```
n_5 <- nrow(woman_total_married)
r_5 <- nrow(woman_married_employed)
p_5 <- r_5 / n_5
paste('Доля трудоустроенных среди замужних женщин составляет:', p_5)
```

```
## [1] "Доля трудоустроенных среди замужних женщин составляет: 0.695742471443406"
```

```
n_6 <- nrow(woman_total_single)
r_6 <- nrow(woman_single_employed)
p_6 <- r_6 / n_6
paste('Доля трудоустроенных среди незамужних женщин составляет:', p_6)
```

```
## [1] "Доля трудоустроенных среди незамужних женщин составляет: 0.777089783281734"
```

Сформулируем гипотезы. **H0:  $\Pi_5 = \Pi_6$**  (между долями трудоустроенных замужних женщин и трудоустроенных незамужних женщин *разницы нет*) и **H1:  $\Pi_5 \neq \Pi_6$**  (*разница есть*).

```
print(n_5)
```

```
## [1] 963
```

```
print(n_6)
```

```
## [1] 646
```

Так как выборки большие, используем график нормального распределения для определения критического значения.

```
P_w <- (r_5 + r_6) / (n_5 + n_6)

Z_w <- (p_5 - p_6) / sqrt(P_w*(1 - P_w)*(1/n_5 + 1/n_6))

paste('Значение тестовой статистики равно:', Z_w)
```

```
## [1] "Значение тестовой статистики равно: -3.5962187937727"
```

• Сравните результаты: различаются ли эффекты для мужчин и женщин?

При уровне значимости 1% H0 отвергается в пользу альтернативной гипотезы H1, то есть в подгруппах наблюдается разница между долями трудоустроенных с разным семейным положением. При этом если посмотреть на односторонние критерии, то для мужчин принимается альтернативная гипотеза, что доля занятых среди женатых выше, чем доля среди неженатых, а для женщин - наоборот. Доля среди замужних ниже, чем доля среди незамужних.

• Обсудите возможные причины различий: экономические, культурные, гендерные роли, ожидания со стороны работодателей и семьи.

В культуре существует стереотип, что женатые мужчины являются главами семейства и имеют большую ответственность за финансовое обеспечение. Это может приводить к тому, что более высокие доли женатых мужчин показывают более высокую занятость. При этом, замужние женщины, наоборот, могут сталкиваться с ожиданиями, связанными с домашними обязанностями, что ограничивает возможности занятости.

Замужние женщины могут нести больше ответственности за уход за детьми или другими родственниками, что сказывается на их трудовой занятости. Это может приводить к более низкой доле замужних женщин среди трудоустроенных. В то же время женатые мужчины могут восприниматься как имеющие меньше таких обязанностей, что поддерживает их присутствие на рынке труда.

Таких же ожиданий могут придерживаться работодатели, полагая, что замужняя женщина вскоре уйдет в декрет и на время будет удалена от работы. За мужчинами же декретный отпуск является исключением, что может приводить к тому, что работодатели более

охотно их нанимают