

Разрыв в доходах между мужчинами и женщинами

1. Загрузка и подготовка данных.

Загрузка данных с официального сайта RLMS HSE <https://www.hse.ru/rlms>.

```
library(haven)

data_individuals_2023 <- read_dta('r32i_os_73.dta')
```

Данные будут взяты за 2023 год, по Санкт-Петербургу и сфере образования.

```
filtered_data <- subset(data_individuals_2023, #данные скачены за 2023 год
                        psu == 1 & # Выбор города (Санкт-Петербург)
                        bbj4_1 == 10) # Выбор отрасли (образование)
```

#функция subset() фильтрует указанный датафрейм и оставляет те строки, которые удовлетворяют критериям, в данном случае, где город - СПб и отрасль - образование

На сайте можно найти расшифровку и описание всех переменных (в том числе с вариантами значений, которые может принимать переменная). Для этого анализа потребуется уровень зарплаты и пол респондента.

```
wages <- filtered_data[, c("bbj13_2", "bbh5")] #bbj13_2 - уровень зарплаты, bbh5
- пол респондента
```

#сохраним в переменную wages датафрейм, содержащий только уровень зарплаты и пол респондента, фильтруем по столбцам, поэтому часть ДО запятой пустая (по строка нет фильтрации)

#Чистим те данные, где респондент не дал точного ответа

```
wages$bbj13_2[wages$bbj13_2 %in% c(99999997, 99999998, 99999999)] <- NA
dim(wages)
```

```
## [1] 26 2
```

3. Построение доверительных интервалов.

Смотрим на количество объектов в каждой из выборок.

```
n_men <- length(wages$bbj13_2[wages$bbh5==1])
n_women <- length(wages$bbj13_2[wages$bbh5==2])

print(sprintf('Количество опрошенных мужчин: %d', n_men))
## [1] "Количество опрошенных мужчин: 7"

print(sprintf('Количество опрошенных женщин: %d', n_women))
## [1] "Количество опрошенных женщин: 19"
```

Так как количество объектов в каждой выборке мало (меньше 30), будем использовать распределение Стьюдента для построения доверительных интервалов (концы распределения будут более “тяжелыми” по сравнению с нормальным распределением.

#Доверительные интервалы для мужчин

```
men_10 <- t.test(wages$bbj13_2[wages$bbh5==1], conf.level = 0.90)$conf.int #  $\alpha = 10\%$ 
men_5 <- t.test(wages$bbj13_2[wages$bbh5==1], conf.level = 0.95)$conf.int #  $\alpha = 5\%$ 
men_1 <- t.test(wages$bbj13_2[wages$bbh5==1], conf.level = 0.99)$conf.int #  $\alpha = 1\%$ 

print(sprintf('90-ый %% доверительный интервал для мужчин: (%.3f, %.3f)',
men_10[1], men_10[2]))

print(sprintf('95-ый %% доверительный интервал для мужчин: (%.3f, %.3f)',
men_5[1], men_5[2]))

print(sprintf('99-ый %% доверительный интервал для мужчин: (%.3f, %.3f)',
men_1[1], men_1[2]))

## [1] "95-ый % доверительный интервал для мужчин: (29628.339, 63514.518)"
## [2] "90-ый % доверительный интервал для мужчин: (33116.315, 60026.542)"
## [3] "99-ый % доверительный интервал для мужчин: (20900.179, 72242.678)"
```

#Доверительные интервалы для женщин

```
women_10 <- t.test(wages$bbj13_2[wages$bbh5==2], conf.level = 0.90)$conf.int #  $\alpha = 10\%$ 
women_5 <- t.test(wages$bbj13_2[wages$bbh5==2], conf.level = 0.95)$conf.int #  $\alpha = 5\%$ 
women_1 <- t.test(wages$bbj13_2[wages$bbh5==2], conf.level = 0.99)$conf.int #  $\alpha = 1\%$ 

print(sprintf('90-ый %% доверительный интервал для женщин: (%.3f, %.3f)',
women_10[1], women_10[2]))

print(sprintf('95-ый %% доверительный интервал для женщин: (%.3f, %.3f)',
women_5[1], women_5[2]))

print(sprintf('99-ый %% доверительный интервал для женщин: (%.3f, %.3f)',
women_1[1], women_1[2]))

## [1] "95-ый % доверительный интервал для женщин: (32613.400, 52916.012)"
## [2] "90-ый % доверительный интервал для женщин: (34404.427, 51124.985)"
## [3] "99-ый % доверительный интервал для женщин: (28778.351, 56751.061)"

print(sprintf('Средняя зп для мужчин: %.2f', mean(wages$bbj13_2[wages$bbh5 == 1], na.rm = TRUE)))
## [1] "Средняя зп для мужчин: 46571.43"

print(sprintf('Средняя зп для женщин: %.2f', mean(wages$bbj13_2[wages$bbh5 == 2], na.rm = TRUE)))
## [1] "Средняя зп для женщин: 42764.71"
```

Мужчины в данной выборке в среднем зарабатывают больше, однако их доходы имеют большую вариативность по сравнению с доходами женщин. Из-за того, что данных по зарплатам женщин чуть больше, доверительные интервалы для зарплат женщин получаются более узкими, чем для зарплат мужчин.

4. Проверка гипотезы о равенстве средних зарплат у мужчин и женщин.

Видим, что средняя зарплата мужчин оказалась чуть выше средней зарплаты женщин в Санкт-Петербурге и в сфере образования.

Проведем двухсторонний тест, чтобы убедиться, действительно ли мужчины в этой категории зарабатывают больше, или различие статистически незначимо.

1. Формулировка гипотез:

$H_0 : \mu_{\text{муж}} = \mu_{\text{жен}}$ vs. $H_1 : \mu_{\text{муж}} \neq \mu_{\text{жен}}$

Нулевая гипотеза(H_0): средняя зарплата женщин равна средней зарплате мужчин.

Альтернативная: средние зарплаты не равны.

2. Выбор тестовой статистики и её распределения.

Рассчитываем необходимые параметры.

```
# Средняя зарплата мужчин и женщин
mean_men <- mean(wages$bbj13_2[wages$bbh5==1], na.rm = TRUE)
mean_women <- mean(wages$bbj13_2[wages$bbh5==2], na.rm = TRUE)

# Размеры выборок (исключая NA)
n_men <- sum(!is.na(wages$bbj13_2[wages$bbh5 == 1]))
n_women <- sum(!is.na(wages$bbj13_2[wages$bbh5 == 2]))

# Стандартные отклонения
SE_women <- sd(wages$bbj13_2[wages$bbh5==2], na.rm = TRUE)
SE_men <- sd(wages$bbj13_2[wages$bbh5==1], na.rm = TRUE)
SE_total <- ((n_men-1)*SE_men^2+(n_women-1)*SE_women^2)/(n_men+n_women-2)
```

Выбираем t-тест для **независимых** выборок с **разными** дисперсиями, так как дисперсии в выборках мужчин и женщин могут отличаться. Кроме того, они неизвестны. В виду того, что выборка мала, используем t-тест Уэлча с скорректированными степенями свободы.

```
# можем посчитать вручную

se_welch <- sqrt((SE_men^2 / n_men) + (SE_women^2 / n_women)) #расчет общей
стандартной ошибки разности

t_stat_welch <- (mean_men - mean_women) / se_welch

# Степени свободы для теста
df_welch <- (SE_men^2/n_men + SE_women^2/n_women)^2 /
  ((SE_men^4/(n_men^2*(n_men-1))) + (SE_women^4/(n_women^2*(n_women-1))))
```

#а можем воспользоваться готовой функцией

```
t_test <- t.test(wages$bbj13_2[wages$bbh5 == 1],
               wages$bbj13_2[wages$bbh5 == 2],
               var.equal = FALSE) #параметр FALSE говорит о том, что
используем тест Уэлча
cat("Результаты сравнения зарплат\n")

## Результаты сравнения зарплат

cat(sprintf("t-статистика: %.4f\n", t_test$statistic))

## t-статистика: 0.4522

cat(sprintf("Степени свободы (df): %.2f\n", t_test$parameter))

## Степени свободы (df): 12.08

cat(sprintf("p-value: %.5f\n", t_test$p.value))

## p-value: 0.65916
```

Видим, что p-value достаточно большое, и какой бы уровень значимости мы не взяли (1, 5 или 10), нулевую гипотезу о равенстве зарплат отвергнуть не можем.

Посмотрим то же самое через критическую область.

3. Критическая область .

Критические значения для разных α (гипотеза H1: z_1 не равны, поэтому делим α пополам)

```
alpha <- c(0.01, 0.05, 0.10)
critical_values <- qt(1 - alpha/2, df = df_welch)
print(critical_values)

## [1] 3.051057 2.177299 1.781357
```

Статистика 0.4522 по модулю меньше критических значений, это так же говорит о том, что на любом уровне значимости гипотеза о равенстве не может быть отвергнута.

4. Статистическое решение.

Значит, что зарплаты женщин и мужчин в Санкт-Петербурге в сфере образования не отличаются статистически.

4.1. Альтернативная гипотеза $H_1 : \mu_{\text{муж}} > \mu_{\text{жен}}$

1. Формулировка гипотез

H_0 : средняя зарплата женщин равна средней зарплате мужчин

H_1 : средняя зарплата женщин больше средней зарплате мужчин

2. Проводим односторонний тест

t-статистика осталась такой же. Поменялись лишь критические значения, так как тест теперь односторонний.

Критические значения для одностороннего теста

```
critical_values_welch_one <- qt(1 - alpha, df_welch)
```

```
print(critical_values_welch_one)
```

```
## [1] 2.678474 1.781357 1.355723
```

Снова статистика 0.4522 по модулю оказывается меньше критических значений при любом уровне значимости.

3. Вывод

Это значит, что зарплаты женщин и мужчин не отличаются статистически.

Можно сделать вывод, что статистически значимых различий между зарплатами работников сферы образования в Санкт-Петербурге мужского и женского пола нет. Это означает отсутствие явной необходимости проведения специальной зарплатной политики.