До и после: здоровье и пандемия

Исследовательский вопрос: Изменилась ли самооценка здоровья (или другое состояние) после пандемии COVID-19?

1. Выбор показателя.

За переменную, связанную со здоровьем, возьмем количество дней, пропущенных по болезни за последние 12 месяцев.

2. Выбор волн RLMS.

Выберем для волны, за 2017 год и за 2022.

```
library(haven)

#волна за 2017 год
data_individuals_2017 <- read_spss('r26i_os_74.sav')

#волна за 2022 год уже сохранена в переменную data_2022, но перезагрузим их в но
вую переменную, чтобы не было накладок
data_individuals_2022 <- read_dta('r31i_os_75.dta')
```

Важно, чтобы люди участвовали в обоих волнах, поэтому отфильтруем наблюдения по этому критерию. Сохраним полезные для исследования переменные.

```
#Создаем 2 датафрейма, которые содержат в себе информацию о поле, возрасте, коли
честве дней, пропущенных по болезни, и id респондентов в 2017 и 2022

data_2017 <- data.frame(days_ill_2017 = data_individuals_2017$vl90, id_2017=data
_individuals_2017$idind, age_2017 = data_individuals_2017$v_age, sex_2017 = data
_individuals_2017$vh5)
data_2022 <- data.frame(days_ill_2022 = data_individuals_2022$aal90, id_2022=dat
a_individuals_2022$idind, age_2022 = data_individuals_2022$aa_age, sex_2022 = da
ta_individuals_2022$aah5)

data_2017$days_ill_2017<-as.character(data_2017$days_ill_2017)
data_2022$days_ill_2022<-as.character(data_2022$days_ill_2022)
```

NA - ответ людей, которые в предыдущем вопросе “Пропускали ли вы рабочие дни по болезни?” ответили “Нет”, поэтому NA в данном случае важно, заменяем это значение на 0.

```
#Заменяем NA на 0 (NA обозначены те, кто не пропускал дни по болезни)
data_2017$days_ill_2017[is.na(data_2017$days_ill_2017)==TRUE]<-0
data_2022$days_ill_2022[is.na(data_2022$days_ill_2022)==TRUE]<-0
data_2017$days_ill_2017<-as.double(data_2017$days_ill_2017)
data_2022$days_ill_2022<-as.double(data_2022$days_ill_2022)
#Очищаем датафреймы от респондентов, которые не ответили на вопрос

data_2017 <- data_2017[data_2017$days_ill_2017 < 99999997,]
data_2022 <- data_2022[data_2022$days_ill_2022 < 99999997,]

#Пересекаем датафреймы так, чтобы остались только те люди, которые участвовали в
обоих опросах
data_2017 <- data_2017[data_2017$id_2017 %in% data_2022$id_2022,]
data_2022 <- data_2022[data_2022$id_2022 %in% data_2017$id_2017,]

#Объединяем датафреймы в один
data <- merge(data_2017, data_2022, by.x='id_2017', by.y='id_2022')
```

3. Создание переменной разности.

Для каждого индивида рассчитаем разность значений выбранного показателя:* $d_i = x_{i,\text{после}} - x_{i,\text{до}}$*. Тогда задача превратится в анализ зависимых (paired) выборок.

```
differ <- data$days_ill_2022-data$days_ill_2017
mean_of_differ <- mean(differ)
print(mean_of_differ)

## [1] 0.3074293

std <- sd(differ) #в R по умолчанию считается выборочное стандартное отклонение
n <- length(differ)
```

4. Построение доверительного интервала.

• *Построим доверительный интервал для среднего изменения d для уровней значимости 90%, 95%, 99%.*

Выбираем t-статистику, так как дисперсии в обеих группах неизвестны. Значит, нужно пользоваться оценкой дисперсии - выборочной дисперсией.Тогда лучше использовать распределение Стьюдента.

```r
confint_10 <- t.test(differ, conf.level = 0.90)$conf.int  # α = 10%
confint_5 <- t.test(differ, conf.level = 0.95)$conf.int  # α = 5%
confint_1 <- t.test(differ, conf.level = 0.99)$conf.int  # α = 1%

print(sprintf('90-ый %% доверительный интервал для разности: (%.3f, %.3f)', confint_10[1],
              confint_10[2]))
```

```
## [1] "90-ый % доверительный интервал для разности: (0.094, 0.521)"
```

```r
print(sprintf('95-ый %% доверительный интервал для разности: (%.3f, %.3f)', confint_5[1], confint_5[2]))
```

```
## [1] "95-ый % доверительный интервал для разности: (0.053, 0.562)"
```

```r
print(sprintf('99-ый %% доверительный интервал для разности: (%.3f, %.3f)', confint_1[1], confint_1[2]))
```

```
## [1] "99-ый % доверительный интервал для разности: (-0.027, 0.642)"
```

• *Интерпретация*

На уровне 5% и 10% можно сказать, что с вероятностью 95% и 90% соответственно доверительный интервал НЕ содержит 0, то есть различия ЕСТЬ, здоровье ухудшилось (но совсем не значительно, так как разница небольшая). Но при этом видим, что 0 содержится в 99% доверительном интервале. Значит, на уровне значимости 1% ухудшение состояния здоровья статистически не значимо.

5. Проверка гипотезы о средних изменениях здоровья по выбранному показателю.

1. Формулировка гипотез:

Нулевая гипотеза (H_0): среднее разности =0 или $\mu_d = 0$ Альтернативная гипотеза (H_1): среднее разности!=0 $\mu_d \neq 0$

2. Выбор и расчет тестовой статистики.

Выбираем t-статистику, так как дисперсии в обеих группах неизвестны. Значит, нужно пользоваться оценкой дисперсии - выборочной дисперсией.Тогда лучше использовать распределение Стьюдента.

```
t_stat <- mean_of_differ/(std/sqrt(n))
print(t_stat)

## [1] 2.366438
```

3. Для уровней значимости = 0,10, 0,05, 0,01 рассчитайте критическое значение.

```
t_value_1 <- qt(0.995, n-1)
t_value_5 <- qt(0.975, n-1)
t_value_10 <-qt(0.95, n-1)

t_values <- c(t_value_10, t_value_5, t_value_1)
print(t_values)

## [1] 1.645054 1.960276 2.576476
```

4. Результат

Так как t_stat > t_value_5, то гипотеза H_0 отвергается на уровне значимости 5% (и на 10% в том числе), но не может быть отвергнута на уровне значимости 1%.

5. Интерпретация.

Таким образом мы можем сделать вывод, что пандемия ухудшила состояние здоровья населения, так как на уровне значимости 5% среднее изменилось.

6. Повторим анализ отдельно для мужчин и женщин.

```
#создаем отдельные датафреймы для мужчин и для женщин
data_2017$sex_2017<-as.character(data_2017$sex_2017)
data_2022$sex_2022<-as.character(data_2022$sex_2022)

female <- data[data$sex_2022 == 2,]
male <- data[data$sex_2022 == 1,]

#считаем разность между количеством дней, которые пропустили по болезни респонде
нты мужского пола в 2022 и в 2017
male_differ <- male$days_ill_2022 - male$days_ill_2017

#Аналогично для респондентов женского пола
female_differ <- female$days_ill_2022 - female$days_ill_2017
```

• *Сравним, кто сильнее испытал негативные изменения.*

```
#вычисляем среднее, стандартное отклонение и количество респондентов
male_mean <- mean(male_differ)
male_std <- sd(male_differ)
male_n <- length(male_differ)
```

```r
t_stat_male <- male_mean/(male_std/sqrt(male_n))
t_value_1_male <- qt(0.995, male_n-1)
t_value_5_male <- qt(0.975, male_n-1)
t_value_10_male <-qt(0.95, male_n-1)
t_value_male_overall <- c(t_value_10_male, t_value_5_male, t_value_1_male)
print(t_stat_male)
```

```
## [1] 3.761783
```

```r
print(t_value_male_overall)
```

```
## [1] 1.645330 1.960705 2.577366
```

```r
female_mean <- mean(female_differ)
female_std <- sd(female_differ)
female_n <- length(female_differ)

t_stat_female <- female_mean/(female_std/sqrt(female_n))
t_value_1_female <- qt(0.995, female_n-1)
t_value_5_female <- qt(0.975, female_n-1)
t_value_10_female <-qt(0.95, female_n-1)
t_value_female_overall <- c(t_value_10_female, t_value_5_female, t_value_1_femal
e)
print(t_stat_female)
```

```
## [1] 0.3439006
```

```r
print(t_value_female_overall)
```

```
## [1] 1.645200 1.960503 2.576947
```

Для мужчин: t_stat_male > t_value_11_male, значит мы нашли эффект для мужчин на уровне значимости 1% (то есть и на 5% и на 10% тоже). Таким образом пандемия повлияла на количество дней, которые мужчины проводят на больничном.

Для женщин: -t_value_10_female < t_stat_female < t_value_10_female, значит гипотеза H_0 не может быть отвергнута на уровне значимости 10% (а следовательно, и на 5%, и на 1%). Для женщин эффект не находится, статистически значимой разности нет.

• *Возможные причины.*

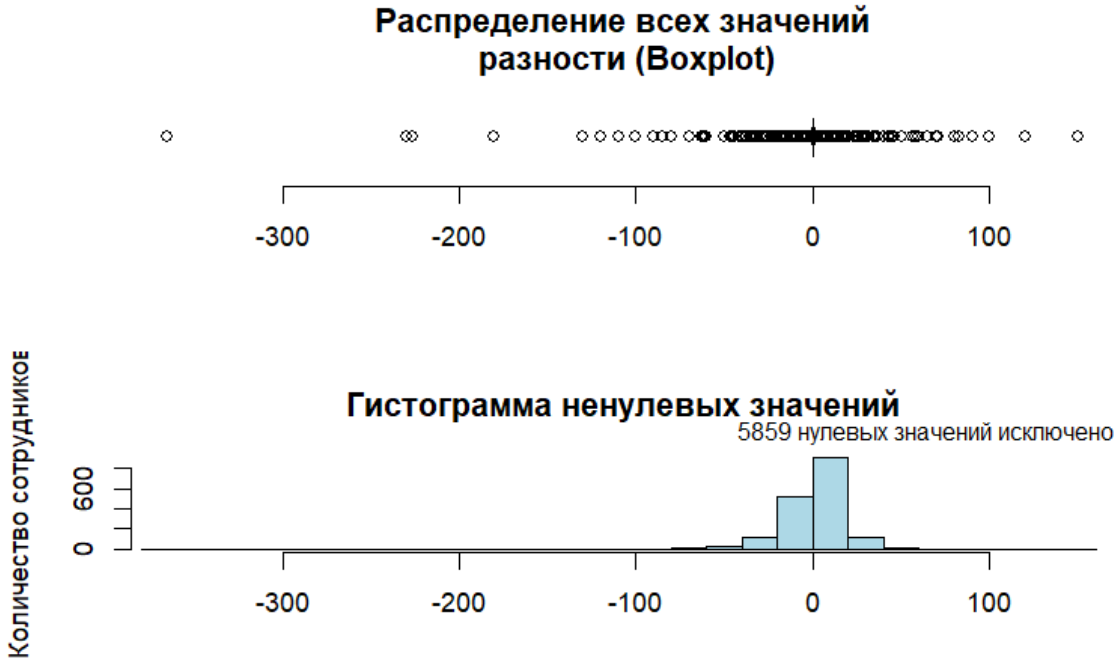
Одно из возможных объяснений заключается в различиях в восприятии болезни, толерантности к дискомфорту и готовности сообщать о днях нетрудоспособности между мужчинами и женщинами. Существует гипотеза, что женщины, в силу социокультурных ролей и ожиданий, могут демонстрировать более высокую склонность продолжать профессиональную деятельность даже при наличии симптомов заболевания, которые мужчины могли бы счесть достаточным основанием для пропуска работы.

Эти поведенческие различия могут усугубляться экономическими и социальными факторами. Например, в полных семьях женщины чаще берут на себя основные обязанности по ведению домашнего хозяйства и уходу за детьми, что может создавать дополнительный стимул избегать оформления больничных. Для матерей-одиночек, составляющих значительную часть неполных семей, необходимость обеспечения семьи может делать выход на работу приоритетом, несмотря на состояние здоровья, что приводит к недооценке или нерегистрации дней болезни в самоотчетах.

7. Визуализация распределение изменений.

```r
# Установка параметров графики
par(mfrow = c(2, 1),  # 2 строки, 1 столбец
    mar = c(5, 4, 3, 2) + 0.1,  # Увеличиваем нижний отступ
    oma = c(0, 0, 0, 0))  # Внешние отступы

# 1. Boxplot (все значения)
boxplot(differ,
        horizontal = TRUE,
        main = "Распределение всех значений разности (Boxplot)",
        xlab = "",
        ylab = "",
        col = "lightgreen",
        frame.plot = FALSE,
        notch = TRUE)

# 2. Гистограмма (без нулей)
hist(differ[differ != 0],
     main = "Гистограмма ненулевых значений",
     xlab = "",  # Временно убираем подпись оси X
     ylab = "Количество сотрудников",
     col = "lightblue",
     border = 'black',
     breaks = 20,
     xlim = range(differ[differ != 0]))

# Добавляем общую подпись оси X внизу
title(xlab = "Разница пропущенных дней (2022 - 2017)",
      outer = TRUE,  # Размещаем внизу внешней области
      line = 3)  # Отступ от края

num_zeros <- sum(differ == 0)
# Добавляем примечание в правый верхний угол
mtext(text = paste(num_zeros, "нулевых значений исключено"),
      side = 3,       # Сторона 3 - верхнее поле графика
      line = 0.25,     # Отступ в строках от края области построения (нужно подо
брать)
      adj = 1,        # Выравнивание: 1 - вправо, 0 - влево, 0.5 - по центру
```

```
      cex = 0.9,       # Размер шрифта (чуть больше, чем 0.8)
      col = "black")
```



В ящике с усами большинство разностей между количеством пропущенных дней по болезни в 2017 и 2022 годах находятся вблизи нуля. Видна высокая кучность в центре и присутствуют выраженные выбросы влево и вправо, что говорит о том, что у немногих людей были значительные колебания (в обе стороны). При этом правый хвост длиннее, что иллюстрирует отклонение гипотезы H_0 о равенстве средних. Несмотря на то, что для большинства респондентов индивидуальные изменения были минимальны (медиана разности близка к нулю), анализ совокупных данных показывает небольшое, но статистически значимое увеличение среднего числа дней болезни за этот период. Если исключить нулевые значения (при построении гистограммы) для более наглядного анализа ситуации, то видно, что самый высокий столбец находится по правую сторону от 0.