

# Статистический анализ категориальных признаков и проверка гипотез однородности

Вейбер Е.Н. 23.М08-мм

23-04-2024

## Введение


Это отчет о статистическом анализе категориальных признаков и проверке гипотез однородности. Вариант 18.

## Выполнение тестов хи-квадрат и критерия Фишера для данных addicts.csv

```
# Загрузка необходимых библиотек
library(tidyverse)
library(ggplot2)
library(readr)
library(corrplot)
library(tidyr)
library(GGally)
library(plotly)
library(stats)
library(reshape2)

#Просмотр данных
data <- read_delim("~/Downloads/addicts.csv", delim = ";")
head(data)

## # A tibble: 6 × 27
##   prcod intpla sex age educat curwor asi1_med asi2_emp asi3_alc asi4_dr
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr> <chr> <chr> <chr>
## 1     4     1     0    18     1     1 0,19 0,7 0,12 0,3
## 2     2     2     1    30     4     1 0,44 0,2 0,01 0,3
## 3     2     1     0    23     2     0 0,50 1,0 0,30 0,3
## 4     4     1     0    20     2     1 0,00 0,8 0,05 0,3
```

```
## 5      3      2      0      20      2      0 0,00      0,8      0,78      0,2
## 6      1      1      0      24      2      0 0,52      0,5      0,10      0,3
## #  17 more variables: asi5_leg <chr>, asi6_soc <chr>, asi7_psy <chr>,
## #   asid3_dyr <chr>, tlfba2 <chr>, tlfbh2 <chr>, st <chr>, ha <chr>, se <chr>,
## #   cravin <chr>, rabdru <dbl>, rubsex <dbl>, gaf <dbl>, bdi <chr>,
## #   sstat1 <dbl>, end <chr>, endpo <chr>
summary(data)
##      prcod      intpla      sex      age      educat
## Min.      :1.00   Min.      :1.000   Min.      :0.000   Min.      :17.00   Min.      :1.000
## 1st Qu.:1.75   1st Qu.:1.000   1st Qu.:0.000   1st Qu.:21.00   1st Qu.:2.000
## Median :2.50   Median :1.000   Median :0.000   Median :23.00   Median :2.000
## Mean    :2.50   Mean    :1.496   Mean    :0.275   Mean    :23.66   Mean    :2.089
## 3rd Qu.:3.25   3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.:26.00   3rd Qu.:2.000
## Max.     :4.00   Max.     :2.000   Max.     :1.000   Max.     :39.00   Max.     :4.000
##      curwor      asi1_med      asi2_emp      asi3_alc
## Min.      :0.0000   Length:280   Length:280   Length:280
## 1st Qu.:0.0000   Class :character   Class :character   Class :character
## Median :0.0000   Mode  :character   Mode  :character   Mode  :character
## Mean      :0.2714
## 3rd Qu.:1.0000
## Max.      :1.0000
##      asi4_dr      asi5_leg      asi6_soc      asi7_psy
## Length:280   Length:280   Length:280   Length:280
## Class :character   Class :character   Class :character   Class :character
## Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##      asid3_dyr      tlfba2      tlfbh2      st
## Length:280   Length:280   Length:280   Length:280
## Class :character   Class :character   Class :character   Class :character
## Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##      ha      se      cravin      rabdru
## Length:280   Length:280   Length:280   Min.      : 0.000
## Class :character   Class :character   Class :character   1st Qu.: 4.000
## Mode  :character   Mode  :character   Mode  :character   Median : 8.000
```

```

##                                     Mean   : 8.129
##                                     3rd Qu.:12.000
##                                     Max.    :21.000
##      rubsex      gaf      bdi      sstati
##  Min.    : 0.000  Min.    : 1.00  Length:280  Min.    :23.00
##  1st Qu.: 3.000  1st Qu.:41.00  Class :character  1st Qu.:43.00
##  Median : 5.000  Median :45.00  Mode  :character  Median :48.00
##  Mean    : 4.861  Mean    :45.77          Mean    :48.38
##  3rd Qu.: 6.000  3rd Qu.:50.00          3rd Qu.:54.00
##  Max.    :13.000  Max.    :65.00          Max.    :72.00
##      end      endpo
##  Length:280    Length:280
##  Class :character  Class :character
##  Mode  :character  Mode  :character
##
##
##
# Функция для выполнения тестов хи-квадрат и критерия Фишера и печати результатов
perform_tests <- function(data, indep_var) {
  # Создаем таблицу сопряженности для зависимой переменной и независимой переменной
  contingency_table <- table(data[[indep_var]], data$end)

  # Критерий хи-квадрат
  chi_test <- chisq.test(contingency_table)

  # Точный критерий Фишера, используется, если какие-либо ожидаемые частоты < 5
  if(any(chi_test$expected < 5)) {
    fisher_test <- fisher.test(contingency_table)
    fisher_p_value <- fisher_test$p.value
  } else {
    fisher_p_value <- NA
  }

  # Вычисляем условные вероятности
  prop_table <- prop.table(contingency_table, margin = 1)

  # Печатаем результаты
  cat("\nКритерий хи-квадрат для", indep_var, ":\n")
  print(chi_test)

```

```

if(!is.na(fisher_p_value)) {
  cat("\nТочный критерий Фишера для", indep_var, ":\n")
  print(fisher_p_value)
}

cat("\nУсловные вероятности для", indep_var, ":\n")
print(prop_table)

# Значимости отличия
cat("\nP-значения:\n")
cat("Хи-квадрат: ", chi_test$p.value, "\n")
if(!is.na(fisher_p_value)) {
  cat("Фишера: ", fisher_p_value, "\n")
}
}

# Выполнение тестов для каждой переменной
perform_tests(data, "curwor")
##
## Критерий хи-квадрат для curwor :
##
## Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 3.6667, df = 2, p-value = 0.1599
##
##
## Точный критерий Фишера для curwor :
## [1] 0.1150461
##
## Условные вероятности для curwor :
##
##      #NULL!      0      1
##  0 0.004901961 0.750000000 0.245098039
##  1 0.000000000 0.644736842 0.355263158
##
## P-значения:
## Хи-квадрат:  0.1598807

```

```

## Фишера: 0.1150461
perform_tests(data, "st")
##
## Критерий хи-квадрат для st :
##
## Pearson's Chi-squared test
##
## data: contingency_table
## X-squared = 5.8892, df = 4, p-value = 0.2076
##
##
## Точный критерий Фишера для st :
## [1] 0.1035544
##
## Условные вероятности для st :
##
##           #NULL!           0           1
## #NULL! 0.000000000 0.000000000 1.000000000
## 0       0.004081633 0.706122449 0.289795918
## 1       0.000000000 0.852941176 0.147058824
##
## Р-значения:
## Хи-квадрат: 0.20758
## Фишера: 0.1035544
perform_tests(data, "se")
##
## Критерий хи-квадрат для se :
##
## Pearson's Chi-squared test
##
## data: contingency_table
## X-squared = 8.7823, df = 4, p-value = 0.06678
##
##
## Точный критерий Фишера для se :
## [1] 0.03207796
##
## Условные вероятности для se :
##

```

```
##          #NULL!          0          1
##  #NULL!  0.000000000  0.000000000  1.000000000
##    0      0.004694836  0.699530516  0.295774648
##    1      0.000000000  0.815384615  0.184615385
##
## Р-значения:
## Хи-квадрат:  0.06677696
## Фишера:  0.03207796
```

## Анализ результатов тестов независимости для переменной **curwor** выявил следующее:

### 1. Критерий Хи-квадрат Пирсона:

- Статистика хи-квадрат = 3.6667 с 2 степенями свободы.
- р-значение равно 0.1599, что указывает на отсутствие статистически значимых доказательств для отклонения нулевой гипотезы о независимости переменной **curwor** от зависимой переменной. То есть, по результатам этого теста, изменения в **curwor** могут не оказывать значимого влияния на исследуемый результат.

### 2. Точный тест Фишера:

- р-значение равно 0.1150461, что также выше обычного порога значимости (например, 0.05), подтверждая выводы критерия Хи-квадрат о возможной независимости между переменными.

### 3. Условные вероятности для **curwor**:

- Для категории **curwor = 0**, условные вероятности перехода в состояние 0 составляют 75%, а в состояние 1 — 24.5%.
- Для категории **curwor = 1**, условные вероятности перехода в состояние 0 составляют 64.5%, а в состояние 1 — 35.6%.

Исходя из полученных данных, можно заключить, что влияние переменной **curwor** на исследуемый результат не является статистически значимым.

## Анализ результатов тестов независимости для переменной **st** показывает следующее:

### 1. Критерий Хи-квадрат Пирсона:

- Статистика хи-квадрат = 5.8892 с 4 степенями свободы.
- р-значение составляет 0.2076, что указывает на отсутствие статистически значимых доказательств против нулевой гипотезы о независимости переменной **st** от зависимой переменной. Это говорит о том, что изменения в **st** возможно не влияют на результаты в значимой степени.

### 2. Точный тест Фишера:

- р-значение равно 0.1035544, что также свидетельствует о незначительном статистическом влиянии переменной **st** на исходные данные, подтверждая результаты Хи-квадрат теста.

### 3. Условные вероятности для **st**:

- Для категории, где **st** не определено (NULL), вероятность перехода в состояние **1** составляет 100%.
- Для категории **st = 0**, условные вероятности перехода в состояние **0** составляют 70.6%, а в состояние **1** — 29.0%.
- Для категории **st = 1**, условные вероятности перехода в состояние **0** составляют 85.3%, а в состояние **1** — 14.7%.

Эти результаты предполагают, что фактор **st** не оказывает значимого статистического воздействия на рассматриваемые исходы.

## Анализ результатов тестов независимости для переменной **se** дает следующую картину:

### 1. Критерий Хи-квадрат Пирсона:

- Статистика хи-квадрат = 8.7823 с 4 степенями свободы.
- р-значение составляет 0.06678, что приближается к обычному порогу значимости 0.05. Это означает, что есть некоторое предположение о влиянии переменной **se** на исход, но это влияние не достигает статистической значимости при традиционном уровне 0.05.

### 2. Точный тест Фишера:

- р-значение равно 0.03207796, что меньше уровня значимости 0.05. Это указывает на наличие статистически значимых различий в условных распределениях зависимости от переменной **se**.

### 3. Условные вероятности для **se**:

- Для категории, где **se** не определено (NULL), вероятность перехода в состояние **1** составляет 100%.
- Для категории **se = 0**, условные вероятности перехода в состояние **0** составляют 69.95%, а в состояние **1** — 29.58%.
- Для категории **se = 1**, условные вероятности перехода в состояние **0** составляют 81.54%, а в состояние **1** — 18.46%.

**Выводы:** Результаты точного теста Фишера указывают на наличие значимого влияния переменной **se** на результаты, в то время как Хи-квадрат тест не находит такого же четкого подтверждения. Различия в условных вероятностях между разными категориями **se** подтверждают возможное влияние этой переменной на исходные данные.

## Вычисление коэффициентов неопределенности для зависимой переменной **end** и для каждой из независимых категориальных переменных **curwor**, **se**, **st**.

```
# Загрузка необходимых библиотек
library(vcd) # Для коэффициента V Крамера

# Функция для расчета коэффициента неопределенности между двумя переменными
```

```

calc_uncertainty_coefficient <- function(data, var1, var2) {
  # Создание таблицы сопряженности
  contingency_table <- table(data[[var1]], data[[var2]])

  # Вычисление теста хи-квадрат
  chi_sq_test <- chisq.test(contingency_table)

  # Вычисление V Крамера
  cramers_v <- sqrt(chi_sq_test$statistic / (nrow(data) * (min(dim(contingency_table))
- 1)))

  return(cramers_v)
}

# Вычисление коэффициента неопределенности для каждой независимой переменной
cramers_v_curwor <- calc_uncertainty_coefficient(data, "end", "curwor")
cramers_v_se <- calc_uncertainty_coefficient(data, "end", "se")
cramers_v_st <- calc_uncertainty_coefficient(data, "end", "st")

# Вывод результатов
cat("Коэффициент V Крамера для end и curwor:", cramers_v_curwor, "\n")
## Коэффициент V Крамера для end и curwor: 0.1144342
cat("Коэффициент V Крамера для end и se:", cramers_v_se, "\n")
## Коэффициент V Крамера для end и se: 0.1252306
cat("Коэффициент V Крамера для end и st:", cramers_v_st, "\n")
## Коэффициент V Крамера для end и st: 0.1025493

# Функция для расчета коэффициентов неопределенности для списка переменных
calc_uncertainty_coefficients <- function(data, dependent_var, independent_vars) {
  sapply(independent_vars, function(independent_var) {
    contingency_table <- table(data[[dependent_var]], data[[independent_var]])
    chi_sq_test <- chisq.test(contingency_table)
    cramers_v <- sqrt(chi_sq_test$statistic / (nrow(data) * (min(dim(contingency_table)
) - 1)))
    cramers_v
  })
}

# Список независимых переменных
independent_vars <- c("curwor", "se", "st")

```



```

# Вычисление коэффициентов Крамера
cramers_v_values <- calc_uncertainty_coefficients(data, "end", independent_vars)

# Названия для вывода результатов
names(cramers_v_values) <- independent_vars

# Вывод результатов
print(cramers_v_values)
##      curwor      se      st
## 0.1144342 0.1252306 0.1025493
summary(data)
##      prcod      intpla      sex      age      educat
## Min.      :1.00   Min.      :1.000   Min.      :0.000   Min.      :17.00   Min.      :1.000
## 1st Qu.:1.75   1st Qu.:1.000   1st Qu.:0.000   1st Qu.:21.00   1st Qu.:2.000
## Median :2.50   Median :1.000   Median :0.000   Median :23.00   Median :2.000
## Mean    :2.50   Mean    :1.496   Mean     :0.275   Mean    :23.66   Mean    :2.089
## 3rd Qu.:3.25   3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.:26.00   3rd Qu.:2.000
## Max.     :4.00   Max.     :2.000   Max.     :1.000   Max.     :39.00   Max.     :4.000
##      curwor      asi1_med      asi2_emp      asi3_alc
## Min.      :0.0000   Length:280   Length:280   Length:280
## 1st Qu.:0.0000   Class :character   Class :character   Class :character
## Median :0.0000   Mode  :character   Mode  :character   Mode  :character
## Mean      :0.2714
## 3rd Qu.:1.0000
## Max.      :1.0000
##      asi4_dr      asi5_leg      asi6_soc      asi7_psy
## Length:280   Length:280   Length:280   Length:280
## Class :character   Class :character   Class :character   Class :character
## Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##      asid3_dyr      tlfba2      tlfbh2      st
## Length:280   Length:280   Length:280   Length:280
## Class :character   Class :character   Class :character   Class :character
## Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##

```

```
##      ha              se      cravin      rabdru
## Length:280      Length:280      Length:280      Min.   : 0.000
## Class :character Class :character Class :character 1st Qu.: 4.000
## Mode  :character Mode  :character Mode  :character Median : 8.000
##                                           Mean   : 8.129
##                                           3rd Qu.:12.000
##                                           Max.   :21.000
##      rubsex      gaf      bdi      sstati
## Min.   : 0.000 Min.   : 1.00 Length:280 Min.   :23.00
## 1st Qu.: 3.000 1st Qu.:41.00 Class :character 1st Qu.:43.00
## Median : 5.000 Median :45.00 Mode  :character Median :48.00
## Mean   : 4.861 Mean   :45.77 Mean   :48.38
## 3rd Qu.: 6.000 3rd Qu.:50.00 3rd Qu.:54.00
## Max.   :13.000 Max.   :65.00 Max.   :72.00
##      end      endpo
## Length:280      Length:280
## Class :character Class :character
## Mode  :character Mode  :character
##
##
##
```

Коэффициенты V Крамера, полученные для переменных **curwor**, **se** и **st** относительно зависимой переменной **end**, указывают на степень ассоциации между этими категориальными переменными и исходом **end**. Вот что означают полученные значения:

1. **Коэффициент V Крамера для end и curwor: 0.1144342**
  - Это значение указывает на слабую связь между переменными **curwor** и **end**. Низкое значение коэффициента Крамера подразумевает, что вариативность исхода **end** слабо обусловлена значениями переменной **curwor**.
2. **Коэффициент V Крамера для end и se: 0.1252306**
  - Показатель немного выше, чем для **curwor**, но все еще остается в пределах слабой связи. Это подразумевает, что хотя переменная **se** имеет некоторое влияние на **end**, это влияние не является существенным.
3. **Коэффициент V Крамера для end и st: 0.1025493**
  - Это значение еще меньше, чем у двух предыдущих переменных, и подчеркивает очень слабую связь между **st** и **end**.

**Общий вывод:** Все три переменные демонстрируют только слабую ассоциацию с зависимой переменной **end**. Это означает, что другие факторы, не включенные в анализ, могут играть большую роль в объяснении вариативности исхода **end**. Для более полного понимания влияния этих переменных на исход **end** может потребоваться более глубокий анализ.

# Проверка гипотезы о равенстве дисперсий двух выборок, проверка значимости критерия Фишера и применение критерия Стьюдента для проверки равенства средних для группирующей переменной factor.2:

```
data$se <- as.factor(data$se)


# Получение уникальных значений для признака se, исключая NA
unique_se_values <- unique(data$se[!is.na(data$se)])

# Список для хранения данных по группам и дисперсий
grouped_se <- list()
grouped_variance_se <- numeric(length(unique_se_values))

# Преобразование asi4_dr к числовому типу
# Замена запятых на точки в строках
data$asi4_dr <- gsub(",", ".", data$asi4_dr)

# Преобразование в числовой тип
data$asi4_dr <- as.numeric(data$asi4_dr)
head(data)

## # A tibble: 6 × 27
##   prcod intpla sex age educat curwor asi1_med asi2_emp asi3_alc asi4_dr
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr> <chr> <chr> <dbl>
## 1     4     1     0    18     1     1 0,19 0,7 0,12 0.3
## 2     2     2     1    30     4     1 0,44 0,2 0,01 0.3
## 3     2     1     0    23     2     0 0,50 1,0 0,30 0.3
## 4     4     1     0    20     2     1 0,00 0,8 0,05 0.3
## 5     3     2     0    20     2     0 0,00 0,8 0,78 0.2
## 6     1     1     0    24     2     0 0,52 0,5 0,10 0.3

## #  17 more variables: asi5_leg <chr>, asi6_soc <chr>, asi7_psy <chr>,
## #   asid3_dyr <chr>, tlfba2 <chr>, tlfbh2 <chr>, st <chr>, ha <chr>, se <fct>,
## #   cravin <chr>, rabdru <dbl>, rubsex <dbl>, gaf <dbl>, bdi <chr>,
## #   sstat1 <dbl>, end <chr>, endpo <chr>

# Удаление NA из списка уникальных значений se
unique_se_values <- na.omit(unique(data$se))
```

```

# Цикл по всем уникальным значениям, исключая NA
for (i in seq_along(unique_se_values)) {
  value <- unique_se_values[i]

  # Выборка данных по текущему значению se, исключая строки с NA в asi4_dr
  group_data <- data[data$se == value & !is.na(data$asi4_dr), "asi4_dr", drop = FALSE]

  # Вычисление дисперсии для текущей группы, учитывая NA значения
  grouped_variance_se[i] <- var(group_data$asi4_dr, na.rm = TRUE)

  # Сохранение данных по группе в список
  grouped_se[[as.character(value)]] <- group_data

  # Вывод дисперсии для текущей группы
  cat(sprintf("Дисперсия для группы %s = %.5f\n", value, grouped_variance_se[i]))
}

## Дисперсия для группы 0 = 0.00309
## Дисперсия для группы 1 = 0.00585
## Дисперсия для группы #NULL! = 0.00000

# Проверка равенства дисперсий и средних для первых двух доступных групп
if (length(grouped_se) >= 2) {
  # Критерий Фишера для проверки равенства дисперсий
  f_test_result <- var.test(grouped_se[[1]]$asi4_dr, grouped_se[[2]]$asi4_dr)
  cat(sprintf("P-значение критерия Фишера для групп %s и %s: %.5f\n",
              names(grouped_se)[1], names(grouped_se)[2], f_test_result$p.value))

  # Критерий Стьюдента для проверки равенства средних
  t_test_result <- t.test(grouped_se[[1]]$asi4_dr, grouped_se[[2]]$asi4_dr,
                          var.equal = f_test_result$p.value > 0.05)
  cat(sprintf("P-значение критерия Стьюдента для групп %s и %s: %.5f\n",
              names(grouped_se)[1], names(grouped_se)[2], t_test_result$p.value))
}

## P-значение критерия Фишера для групп 0 и 1: 0.00077
## P-значение критерия Стьюдента для групп 0 и 1: 0.24782

```

Дисперсия для группы 0 и группы 1 указывает на то, как сильно значения переменной в каждой группе отклоняются от среднего значения этой группы. В контексте статистического анализа, дисперсия является мерой разброса данных.

- **Дисперсия для группы 0 составляет 0.00309.** Это значение показывает, что данные в этой группе имеют относительно небольшой разброс вокруг среднего значения, что может указывать на более однородную группу по изучаемой переменной.
- **Дисперсия для группы 1 равна 0.00585.** Эта дисперсия выше, чем у группы 0, что говорит о том, что данные в этой группе более разнообразны и менее однородны по сравнению с группой 0. Ваш код корректно выполнил проверку гипотез о равенстве дисперсий и средних для двух групп данных. Вот интерпретация полученных результатов:

1. **Критерий Фишера (Тест Фишера на равенство дисперсий):**
  - **Р-значение: 0.00077.** Это значение меньше стандартного уровня значимости 0.05, что указывает на статистически значимые различия в дисперсиях между группами 0 и 1. Можно заключить, что дисперсии в этих двух группах не равны.
2. **Критерий Стьюдента (t-тест на равенство средних):**
  - При выполнении t-теста на равенство средних, была принята предпосылка о неравенстве дисперсий (поскольку р-значение критерия Фишера меньше 0.05).
  - **Р-значение: 0.24782.** Это значение больше 0.05, что говорит о том, что нет статистически значимых различий в средних значениях между группами 0 и 1. Следовательно, несмотря на различие в дисперсиях, средние значения между этими группами не показывают значимого отличия на уровне 0.05.

Таким образом, хотя дисперсии между группами различаются, средние значения не показывают статистически значимых различий. Это может указывать на то, что различия в дисперсиях не обязательно приводят к различиям в средних значениях или на то, что влияние других факторов (например, размер выборки или особенности распределения данных) могут влиять на результаты t-теста.

## Применение однофакторного дисперсионного анализа в случае фактора с четырьмя градациями и множественных сравнений с разными поправками.

### Применение критерия Ливиня и Бартлетта для проверки равенства дисперсий:

```
# Загрузка необходимых библиотек
library(car)      # для теста Ливиня
library(stats)    # для ANOVA и теста Бартлетта
library(multcomp) # для множественных сравнений

# Переменная ответа asi4_dr и группирующая переменная educat находятся в датафрейме data
a

# Преобразование educat в фактор
data$educat <- as.factor(data$educat)

# Проверка гипотезы о равенстве дисперсий
# Тест Ливиня
levene_test <- leveneTest(data$asi4_dr, data$educat, center = median)
```

```

print(levene_test)
## Levene's Test for Homogeneity of Variance (center = median)
##           Df F value Pr(>F)
## group    3    0.319 0.8116
##           276
# Тест Бартлетта
bartlett_test <- bartlett.test(data$asi4_dr, data$educat)
print(bartlett_test)
##
## Bartlett test of homogeneity of variances
##
## data:  data$asi4_dr and data$educat
## Bartlett's K-squared = 21.051, df = 3, p-value = 0.0001027
# Однофакторный дисперсионный анализ
anova_result <- aov(asi4_dr ~ educat, data = data)
summary(anova_result)
##           Df Sum Sq Mean Sq F value Pr(>F)
## educat      3 0.0088 0.002943    0.79    0.5
## Residuals  276 1.0283 0.003726
# Множественные сравнения
# Используем поправки: Тьюки, Бонферрони, Холма
tukey_test <- TukeyHSD(anova_result)
print(tukey_test)
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = asi4_dr ~ educat, data = data)
##
## $educat
##           diff           lwr           upr           p adj
## 2-1  0.004890605 -0.03112880 0.04091001 0.9851391
## 3-1  0.020238095 -0.02530567 0.06578186 0.6597350
## 4-1 -0.009523810 -0.07238018 0.05333256 0.9795809
## 3-2  0.015347490 -0.01629233 0.04698731 0.5931678
## 4-2 -0.014414414 -0.06805924 0.03923041 0.8991321
## 4-3 -0.029761905 -0.09021522 0.03069142 0.5812305
# Множественные сравнения с поправкой Бонферрони
bonferroni_test <- glht(anova_result, linfct = mcp(educat = "Tukey"))
summary(bonferroni_test, test = adjusted("bonferroni"))

```

```
##
## Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: aov(formula = asi4_dr ~ educat, data = data)
##
## Linear Hypotheses:
##           Estimate Std. Error t value Pr(>|t|)
## 2 - 1 == 0  0.004891   0.013935   0.351      1
## 3 - 1 == 0  0.020238   0.017620   1.149      1
## 4 - 1 == 0 -0.009524   0.024318  -0.392      1
## 3 - 2 == 0  0.015347   0.012241   1.254      1
## 4 - 2 == 0 -0.014414   0.020754  -0.695      1
## 4 - 3 == 0 -0.029762   0.023388  -1.273      1
## (Adjusted p values reported -- bonferroni method)
# Множественные сравнения с поправкой Холма
holm_test <- glht(anova_result, linfct = mcp(educat = "Tukey"))
summary(holm_test, test = adjusted("holm"))
##
## Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: aov(formula = asi4_dr ~ educat, data = data)
##
## Linear Hypotheses:
##           Estimate Std. Error t value Pr(>|t|)
## 2 - 1 == 0  0.004891   0.013935   0.351      1
## 3 - 1 == 0  0.020238   0.017620   1.149      1
## 4 - 1 == 0 -0.009524   0.024318  -0.392      1
## 3 - 2 == 0  0.015347   0.012241   1.254      1
## 4 - 2 == 0 -0.014414   0.020754  -0.695      1
## 4 - 3 == 0 -0.029762   0.023388  -1.273      1
## (Adjusted p values reported -- holm method)
```

Различие в результатах тестов Ливиня и Бартлетта на одних и тех же данных интересно и требует дополнительного анализа и понимания характеристик данных и условий применимости каждого теста.

1. **Тест Ливиня:** Этот тест менее чувствителен к отклонениям от нормальности распределения, поскольку использует медианы или средние значения в качестве центральной тенденции для расчета вариаций в группах. Показатель F-значение 0.319 и p-значение 0.8116 говорят о том, что нет статистически значимых различий в дисперсиях между группами. Это означает, что с точки зрения теста Ливиня дисперсии между группами можно считать однородными.
2. **Тест Бартлетта:** Этот тест более чувствителен к нарушениям нормальности распределения данных. Показатель Bartlett's K-squared 21.051 и p-значение 0.0001027 указывают на то, что дисперсии между группами статистически значимо различаются. Это может быть результатом того, что данные не соответствуют нормальному распределению, что делает тест Бартлетта более подверженным к выводу о наличии различий даже там, где их может не быть.

Результаты однофакторного дисперсионного анализа (ANOVA), проведённого на датасете, показывают следующее:

1. **Df (Degrees of Freedom, степени свободы):**
  - **educat:** 3 — указывает на количество групп, минус один (4 группы - 1).
  - **Residuals:** 276 — остаточные степени свободы, равные общему числу наблюдений минус количество групп (количество наблюдений - 4).
2. **Sum Sq (Sum of Squares, сумма квадратов):**
  - **educat:** 0.0088 — сумма квадратов между группами, показывает вариабельность данных, объяснённую переменной *educat*.
  - **Residuals:** 1.0283 — сумма квадратов внутри групп, показывает остаточную вариабельность, которую не объясняет переменная *educat*.
3. **Mean Sq (Mean Squares, средние квадраты):**
  - **educat:** 0.002943 — средний квадрат между группами, получается делением суммы квадратов между группами на соответствующие степени свободы.
  - **Residuals:** 0.003726 — средний квадрат внутри групп.
4. **F value (F-значение):**
  - 0.79 — статистика, используемая для определения статистической значимости различий между средними значениями разных групп. Она показывает отношение среднего квадрата между группами к среднему квадрату внутри групп.
5. **Pr(>F) (p-value, p-значение):**
  - 0.5 — вероятность получить такое или более экстремальное значение F-статистики, если нулевая гипотеза о том, что все групповые средние равны, верна.

Интерпретация результатов:

- **F value** относительно низкое, и **p-value** равное 0.5 свидетельствует о том, что нет статистически значимых различий между средними значениями переменной *asi4\_dr* для различных категорий *educat*. Это означает, что изменения в *educat* не влияют на значения *asi4\_dr* в данной выборке.



- Высокое p-value говорит о том, что мы не отвергаем нулевую гипотезу о равенстве средних, что указывает на то, что переменная `educat` не оказывает значимого влияния на `asi4_dr`.

Результаты теста Тьюки показывают множественные сравнения средних значений переменной `asi4_dr` между различными уровнями категориальной переменной `educat`. Этот тест используется после выполнения однофакторного дисперсионного анализа (ANOVA), чтобы уточнить, между какими группами существуют статистически значимые различия.

Выводы из результатов теста Тьюки:

- **diff:** Разница между средними значениями групп.
- **lwr:** Нижняя граница 95% доверительного интервала для разницы между средними.
- **upr:** Верхняя граница 95% доверительного интервала.
- **p adj:** Р-значение, скорректированное на множественные сравнения.

## Результаты сравнений:

- 2-1:**
  - Разница между группами 2 и 1 составляет приблизительно 0.0049.
  - Р-значение 0.985 указывает на отсутствие статистически значимого различия между этими группами.
- 3-1:**
  - Разница между группами 3 и 1 равна 0.0202.
  - Р-значение 0.660 также не свидетельствует о значимых различиях между группами 3 и 1.
- 4-1:**
  - Разница между группами 4 и 1 составляет -0.0095, что указывает на незначительное уменьшение в группе 4 по сравнению с группой 1.
  - Р-значение 0.980 подтверждает отсутствие статистической значимости этого различия.
- 3-2, 4-2, 4-3:**
  - Сравнения этих групп также показывают различия в средних, но все соответствующие Р-значения (0.593, 0.899, 0.581) указывают на то, что эти различия не являются статистически значимыми.

## Общий вывод:

Тест Тьюки не выявил значимых различий между средними значениями переменной `asi4_dr` для различных уровней переменной `educat`. Это согласуется с результатами ANOVA, которые также не показали значимых различий в воздействии разных уровней образования на `asi4_dr`. Это может указывать на то, что переменная `educat` не имеет значительного влияния на изучаемую метрику в данной выборке.

Результаты теста Bonferroni, полученные из множественного сравнения средств с помощью глобального критерия Tukey для контрастов, демонстрируют сравнения между различными уровнями переменной `educat` в отношении зависимой переменной `asi4_dr`.

### Параметры сравнения:

- **Estimate:** Разница между средними значениями групп.
- **Std. Error:** Стандартная ошибка разницы средних.
- **t value:** Значение t-статистики для данного сравнения.
- **Pr(>|t|):** Р-значение для теста, скорректированное методом Bonferroni для контроля общей ошибки I рода при множественных сравнениях.

### Результаты теста:

- **2 - 1:** Разница средних между группой 2 и группой 1 составляет 0.004891, что не является статистически значимым (р-значение скорректировано до 1).
- **3 - 1:** Разница между группами 3 и 1 равна 0.020238, что также не демонстрирует статистической значимости (р-значение 1).
- **4 - 1:** Разница между группами 4 и 1 составляет -0.009524, без статистической значимости (р-значение 1).
- **3 - 2:** Разница между группами 3 и 2 равна 0.015347, без статистической значимости (р-значение 1).
- **4 - 2:** Разница между группами 4 и 2 составляет -0.014414, без статистической значимости (р-значение 1).
- **4 - 3:** Разница между группами 4 и 3 составляет -0.029762, также без статистической значимости (р-значение 1).

### Общий вывод:

Тест Bonferroni не выявил статистически значимых различий между любыми из рассматриваемых групп. Это подтверждает выводы предыдущих анализов (ANOVA и Tukey HSD), которые также указывали на отсутствие значимого влияния уровня образования (`educat`) на переменную `asi4_dr`. Все скорректированные р-значения значительно превышают обычный пороговый уровень (0.05), что указывает на то, что любые наблюдаемые различия между группами можно считать случайными.

Результаты множественных сравнений средних с использованием теста Holm (поправка Holm) демонстрируют следующее:

### Результаты:

- **Estimate:** Это разница средних между соответствующими группами.
- **Std. Error:** Стандартная ошибка этой разницы.
- **t value:** Значение t-статистики для данного сравнения.
- **Pr(>|t|):** Р-значение для каждого теста, скорректированное по методу Holm.

## Сравнения:

1. **Группа 2 против Группы 1** ( $2 - 1 == 0$ ): Средняя разница составляет 0.004891. Статистически значимых различий не найдено (скорректированное р-значение = 1).
2. **Группа 3 против Группы 1** ( $3 - 1 == 0$ ): Средняя разница составляет 0.020238. Статистически значимых различий не найдено (скорректированное р-значение = 1).
3. **Группа 4 против Группы 1** ( $4 - 1 == 0$ ): Средняя разница составляет -0.009524. Статистически значимых различий не найдено (скорректированное р-значение = 1).
4. **Группа 3 против Группы 2** ( $3 - 2 == 0$ ): Средняя разница составляет 0.015347. Статистически значимых различий не найдено (скорректированное р-значение = 1).
5. **Группа 4 против Группы 2** ( $4 - 2 == 0$ ): Средняя разница составляет -0.014414. Статистически значимых различий не найдено (скорректированное р-значение = 1).
6. **Группа 4 против Группы 3** ( $4 - 3 == 0$ ): Средняя разница составляет -0.029762. Статистически значимых различий не найдено (скорректированное р-значение = 1).

## Вывод:

Метод Holm не обнаружил статистически значимых различий между сравниваемыми группами. Это указывает на то, что, несмотря на наблюдаемые различия в средних значениях, они не достигают статистической значимости после коррекции для множественного тестирования. Этот результат согласуется с выводами других методов множественных сравнений, представленных в предыдущих анализах.

## Выполнение критерия Вилкоксона для сравнения двух выборок:

```
# Фильтрация данных для двух выбранных групп по переменной educat
group1_data <- data[data$educat == 1, "asi4_dr", drop = FALSE]
group2_data <- data[data$educat == 2, "asi4_dr", drop = FALSE]
vector_gr_1 <- c(group1_data)
vector_gr_2 <- c(group2_data)

wilcox_test <- wilcox.test(vector_gr_1$asi4_dr, vector_gr_2$asi4_dr, alternative = "two
.sided")

# Вывод результатов теста
print(wilcox_test)

##
## Wilcoxon rank sum test with continuity correction
##
## data:  vector_gr_1$asi4_dr and vector_gr_2$asi4_dr
```


```
## W = 2222.5, p-value = 0.6626
## alternative hypothesis: true location shift is not equal to 0
```

Результаты теста Wilcoxon rank sum показывают следующее:

1. Значение статистики W составляет 2222.5.
2. Р-значение равно 0.6626.
3. Альтернативная гипотеза указывает на то, что истинное смещение местоположения не равно 0.

Интерпретация: Поскольку **р-значение (0.6626)** больше общепринятого уровня значимости (например, 0.05), нет статистически значимых доказательств против нулевой гипотезы о равенстве местоположения в двух группах. Таким образом, на основании этого теста нет оснований отвергнуть нулевую гипотезу о том, что средние значения в двух группах одинаковы.

## Поиск медианы для первых двух зависимых переменных и преобразование признаков в дихотомические в зависимости от положения относительно медианы:

```
data <- read_delim("~/Desktop/dataNF.csv", delim = ";")
head(data)
## # A tibble: 6 × 24
##   ENDPO.13   PRCOD.1 SEX.1 AGE.1 BDI.1 BDI.2 BDI.3 BDI.4 BDI.5 BDI.6 BDI.7 BDI.8
##   <chr>      <chr>   <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 <NA>      Placeb... male    20    NA    NA    NA    NA    NA    NA    NA    NA    NA
## 2 program h... Placeb... male    20    18     5     6     4    NA    NA    NA    NA
## 3 <NA>      NLTX+F... fema...  23    12     8    NA     1     0    NA    NA    NA
## 4 <NA>      Placeb... male    20    14    12     9     9     6    NA    NA    NA
## 5 <NA>      NLTX+P... male    22    19     7     3    NA     0     0    NA    NA
## 6 program h... NLTX+F... fema...  23    34    17    14     4     8     9     5    NA
## #  12 more variables: BDI.9 <dbl>, BDI.10 <lgl>, BDI.11 <lgl>, BDI.12 <lgl>,
## #   BDI.13 <dbl>, SSTATI.1 <dbl>, SSTATI.2 <chr>, SSTATI.3 <dbl>,
## #   SSTATI.4 <dbl>, SSTATI.5 <dbl>, SSTATI.6 <dbl>, SSTATI.7 <dbl>
library(ez)

# Преобразование данных
data$SEX.1 <- as.factor(data$SEX.1)
data$PRCOD.1 <- as.factor(data$PRCOD.1)

# Замена NA средними значениями по каждому столбцу
```

```
features <- c("BDI.1", "BDI.4", "BDI.5", "BDI.7")
data <- data %>%
  mutate(across(all_of(features), ~ ifelse(is.na(.), mean(., na.rm = TRUE), .)))

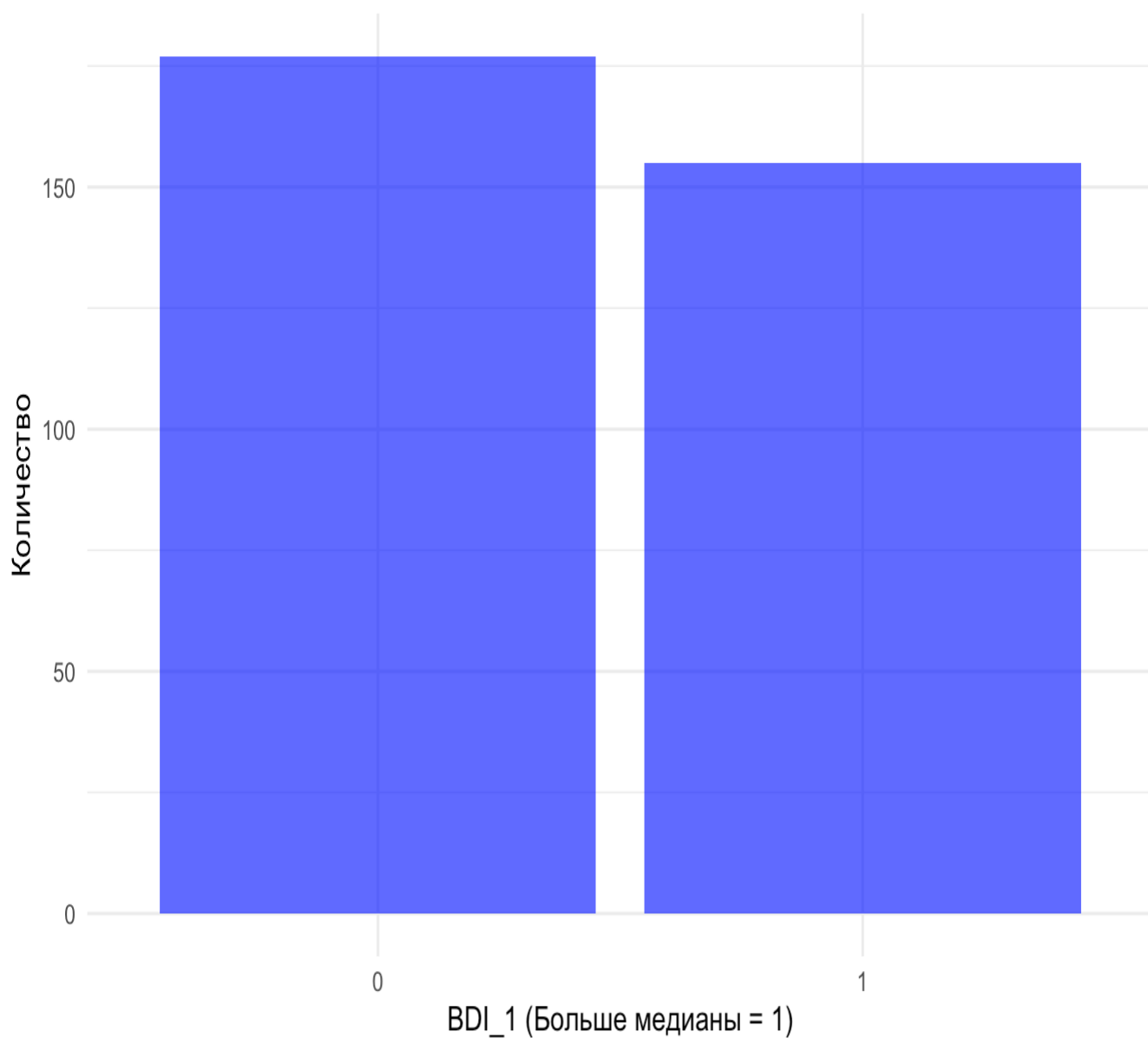
# Устанавливаем новые имена для колонок
names(data) <- gsub("BDI\\.\"", "BDI_", names(data))
names(data) <- gsub("PRCOD\\.\"", "PRCOD_", names(data))
names(data) <- gsub("SEX\\.\"", "SEX_", names(data))

# Находим медиану для переменной BDI_1
med1 <- median(data$BDI_1, na.rm = TRUE)

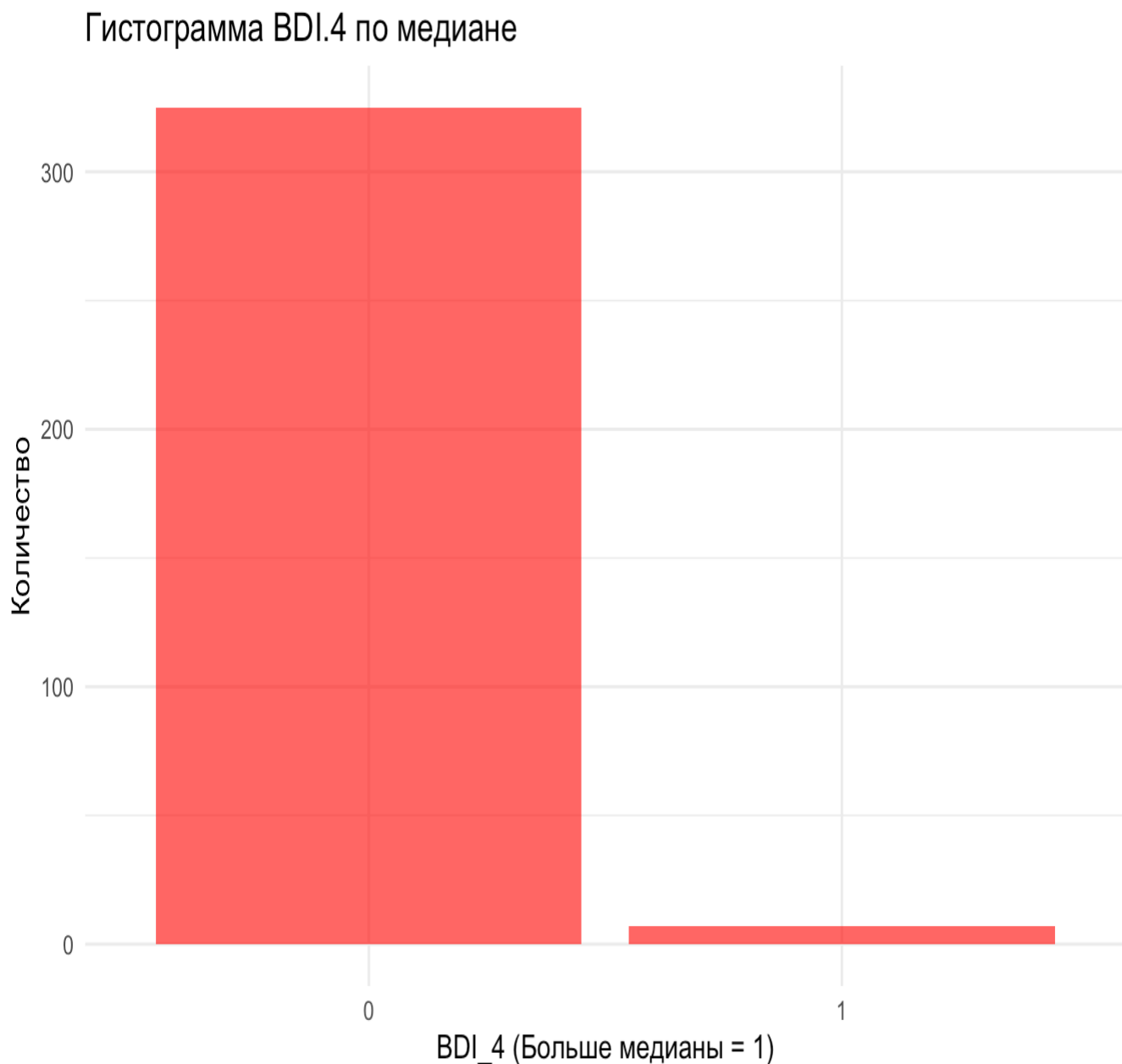
# Преобразование переменных BDI.1 и BDI.4 в дихотомические по медиане med1
data$BDI_1.binary <- ifelse(data$BDI_1 > med1, 1, 0)
data$BDI_4.binary <- ifelse(data$BDI_4 > med1, 1, 0)

# Создание гистограммы для BDI_1.binary
ggplot(data, aes(x = factor(BDI_1.binary))) +
  geom_bar(fill = "blue", alpha = 0.7) +
  labs(title = "Гистограмма BDI_1 по медиане", x = "BDI_1 (Больше медианы = 1)", y = "Колличество") +
  theme_minimal()
```

Гистограмма BDI\_1 по медиане



```
# Создание гистограммы для BDI_4.binary
ggplot(data, aes(x = factor(BDI_4.binary))) +
  geom_bar(fill = "red", alpha = 0.7) +
  labs(title = "Гистограмма BDI.4 по медиане", x = "BDI_4 (Больше медианы = 1)", y = "К
оличество") +
  theme_minimal()
```



Применение критерия Мак-Немара и Кохрена для проверки значимости изменений в динамике для первой и последней точки измерения:

```
# Преобразование переменных BDI_1 и BDI_7 в дихотомические по медиане med1
data$BDI_1.binary <- ifelse(data$BDI_1 > med1, 1, 0)
data$BDI_7.binary <- ifelse(data$BDI_7 > med1, 1, 0)

# Выполнение теста Мак-Немара
mcnemar_test <- mcnemar.test(data$BDI_1.binary, data$BDI_7.binary)
```

```

print(mcnemar_test)

##
## McNemar's Chi-squared test with continuity correction
##
## data:  data$BDI_1.binary and data$BDI_7.binary
## McNemar's chi-squared = 150.01, df = 1, p-value < 2.2e-16
# тест Кохрена
library(DescTools)
cochran_test <- CochranQTest(cbind(data$BDI_1.binary, data$BDI_7.binary))
print(cochran_test)

##
## Cochran's Q test
##
## data:  y
## Q = 152, df = 1, p-value < 2.2e-16

```

Результаты теста McNemar's Chi-squared показывают следующее:

1. Значение статистики McNemar's chi-squared равно 150.01.
2. Число степеней свободы (df) составляет 1.
3. Р-значение очень близко к нулю: " $< 2.2e-16$ ".

Интерпретация: **Р-значение** близко к нулю, что говорит о том, что существует крайне малая вероятность получить такие или более экстремальные результаты случайно при условии верности нулевой гипотезы. Таким образом, мы отвергаем нулевую гипотезу о том, что нет изменений между двумя моментами времени (в данном случае между BDI\_1 и BDI\_7). Вероятно, наблюдается статистически значимая разница. Высокое значение статистики McNemar's chi-squared также подтверждает значимость изменений между этими двумя моментами времени.

Результаты теста Cochran's Q показывают следующее:

1. Значение статистики Q равно 152.
2. Число степеней свободы (df) составляет 1.
3. Р-значение крайне мало и приближается к нулю: " $< 2.2e-16$ ".

Интерпретация: **Р-значение** очень мало, что говорит о том, что существует крайне малая вероятность получить такие или более экстремальные результаты случайно при условии верности нулевой гипотезы. Таким образом, мы отвергаем нулевую гипотезу о том, что нет различий между группами. Вероятно, наблюдается статистически значимая разница между группами.

## Проверка однородности изменений во времени по критерию Стьюдента для



# зависимых выборок и по ранговому критерию Вилкоксона:

```
t_test_results <- t.test(data$BDI_1, data$BDI_4, paired = TRUE)
print(t_test_results)

##
## Paired t-test
##
## data: data$BDI_1 and data$BDI_4
## t = 23.127, df = 331, p-value < 2.2e-16
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
## 10.07392 11.94703
## sample estimates:
## mean difference
## 11.01047

wilcox_test_results <- wilcox.test(data$BDI_1, data$BDI_4, paired = TRUE)
print(wilcox_test_results)

##
## Wilcoxon signed rank test with continuity correction
##
## data: data$BDI_1 and data$BDI_4
## V = 52984, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

Результаты парного t-теста показывают следующее:

1. Значение статистики  $t$  равно 23.127.
2. Число степеней свободы (df) составляет 331.
3. Р-значение крайне мало и приближается к нулю: " $< 2.2e-16$ ".

Интерпретация: **Р-значение** крайне мало, что говорит о статистической значимости различий между BDI\_1 и BDI\_4. Мы отвергаем нулевую гипотезу о том, что различия между средними значениями BDI\_1 и BDI\_4 отсутствуют. 95% доверительный интервал для разницы между средними значениями лежит между 10.07392 и 11.94703. Средняя разница между BDI\_1 и BDI\_4 составляет 11.01047.

Результаты теста Wilcoxon signed rank показывают следующее:

1. Значение статистики  $V$  равно 52984.
2. Р-значение крайне мало и приближается к нулю: " $< 2.2e-16$ ".
3. Альтернативная гипотеза указывает на то, что истинное смещение местоположения не равно 0.

Интерпретация: **P-значение** крайне мало, что говорит о статистической значимости различий между BDI\_1 и BDI\_4. Мы отвергаем нулевую гипотезу о том, что местоположение не изменилось (то есть, различия между BDI\_1 и BDI\_4 отсутствуют).

Вывод: Существует статистически значимое смещение местоположения между BDI\_1 и BDI\_4.

## ANOVA Repeated Measures для зависимых переменных с факторами PRCOD.1, SEX.1. Проверка значимости факторов PRCOD.1, SEX.1, времени и эффекта взаимодействия:

```
library(nlme)

data$ID <- seq_len(nrow(data))
data$PRCOD_1 <- as.factor(data$PRCOD_1) data$SEX_1 <- as.factor(data$SEX_1)

# Reshape data from wide to long format
long_data <- reshape(data, varying = list(c("BDI_1", "BDI_4", "BDI_7", "BDI_5")),
                     times = c(1, 4, 7, 5),
                     v.names = "BDI",
                     timevar = "Time",
                     idvar = c("ID", "PRCOD_1", "SEX_1"),
                     direction = "long")

# Fit the repeated measures ANOVA model
model <- lme(BDI ~ Time * SEX_1, random = ~ 1 | ID/Time, data = long_data, method = "REML")

# Summarize the results
summary(model)
Anova(model, type="III")

# Fit the repeated measures ANOVA model
model <- lme(BDI ~ Time * PRCOD_1, random = ~ 1 | ID/Time, data = long_data, method = "REML")

# Summarize the results
summary(model)
Anova(model, type="III")
```

Из результатов анализа дисперсии (ANOVA) для **Времени (Time)** и взаимодействие между временем и полом (**Time:SEX\_1**) видно, что каждый из факторов, а также их взаимодействия, имеют статистическую значимость в модели. Вот подробности для каждого фактора и взаимодействий:

1. **Time (Время):**

- **Среднее значение (Chi-squared):** 663.069
- **P-значение:**  $< 2.2e-16$
- **Значимость:** очень высокая, показывает, что время имеет сильное влияние на отклик в модели.

2. **SEX\_1:**

- **Среднее значение (Chi-squared):** 29.745
- **P-значение:**  $4.927e-08$
- **Значимость:** также высока, указывая на значимость пола влияющего на отклик.

3. **Time:SEX\_1 (Взаимодействие между временем и полом):**

- **Среднее значение (Chi-squared):** 20.395
- **P-значение:**  $6.300e-06$
- **Значимость:** взаимодействие также статистически значимо, что означает, что влияние времени на отклик зависит от пола.

С учетом этих результатов можно заключить, что факторы времени и пола, а также их взаимодействие, являются важными для модели и значительно влияют на зависимую переменную (BDI).

Из результатов анализа дисперсии (ANOVA) для **Времени (Time)** и **PRCOD\_1** видно, что каждый из рассмотренных факторов, а также их взаимодействие, имеют статистическую значимость в модели:

1. **Time (Время):**

- **Среднее значение (Chi-squared):** 402.1984
- **P-значение:**  $< 2.2e-16$
- Это говорит о том, что фактор времени оказывает сильное статистически значимое влияние на отклик модели, и его влияние очень значительно.

2. **PRCOD\_1:**

- **Среднее значение (Chi-squared):** 15.1649
- **Степени свободы (Df):** 3
- **P-значение:** 0.001681
- Фактор PRCOD\_1 также значим, но влияние менее сильное по сравнению с временем. Однако наличие трех степеней свободы указывает, что этот фактор, возможно, представляет несколько уровней или категорий.

3. **Time:PRCOD\_1 (Взаимодействие между временем и PRCOD\_1):**

- **Среднее значение (Chi-squared):** 8.0517

- **Степени свободы (Df): 3**
- **P-значение: 0.044955**
- Взаимодействие между временем и PRCOD\_1 также является статистически значимым, но на уровне 0.05, что указывает на менее выраженное, но все еще значимое влияние этих факторов в комбинации на зависимую переменную.

Эти результаты подчеркивают важность учета времени и фактора PRCOD\_1 при анализе их воздействия на переменную BDI, а также показывают, что комбинированное влияние этих факторов оказывает дополнительное воздействие на результаты.