Отчет с описательными статистиками

Alisa Seleznyova

2022-11-10

#Посмотреть на данные  
df <- read\_excel("data\_excel.xlsx")  
#нужно что-то сделать со страницей визитов, возможно сделать отдельный столбец и обращаться  
#к нему при подсчете показателей крови  
#table(df$Группа)  
#table(df$Возраст)  
summary(df)

## Группа Возраст Пол Рост   
## Length:100 Min. :21.00 Length:100 Min. :155.0   
## Class :character 1st Qu.:28.00 Class :character 1st Qu.:164.0   
## Mode :character Median :30.50 Mode :character Median :168.0   
## Mean :30.25 Mean :167.7   
## 3rd Qu.:33.00 3rd Qu.:171.2   
## Max. :42.00 Max. :181.0   
## Группа крови Базофилы\_E1 Эозинофилы\_E1 Гемоглобин\_E1   
## Length:100 Length:100 Length:100 Length:100   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## Эритроциты\_E1 Базофилы\_E2 Эозинофилы\_E2 Гемоглобин\_E2   
## Length:100 Length:100 Length:100 Length:100   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## Эритроциты\_E2   
## Length:100   
## Class :character   
## Mode :character   
##   
##   
##

df <- df %>%  
 mutate\_at(vars(`Базофилы\_E1`:`Эритроциты\_E2`), ~ gsub(",", ".", .) %>% as.numeric()) %>%  
 mutate\_at(vars(`Группа`, `Пол`), ~ factor(.))   
summary(df)

## Группа Возраст Пол Рост Группа крови   
## Группа 1:50 Min. :21.00 Женский:53 Min. :155.0 Length:100   
## Группа 2:50 1st Qu.:28.00 Мужской:47 1st Qu.:164.0 Class :character   
## Median :30.50 Median :168.0 Mode :character   
## Mean :30.25 Mean :167.7   
## 3rd Qu.:33.00 3rd Qu.:171.2   
## Max. :42.00 Max. :181.0   
## Базофилы\_E1 Эозинофилы\_E1 Гемоглобин\_E1 Эритроциты\_E1   
## Min. :-0.2188 Min. :-1.227 Min. : 5.352 Min. :2.821   
## 1st Qu.: 0.4020 1st Qu.: 2.325 1st Qu.:10.681 1st Qu.:3.605   
## Median : 0.6509 Median : 3.728 Median :11.711 Median :4.082   
## Mean : 0.6509 Mean : 3.707 Mean :11.860 Mean :4.104   
## 3rd Qu.: 0.8644 3rd Qu.: 5.083 3rd Qu.:13.175 3rd Qu.:4.599   
## Max. : 1.7186 Max. : 8.434 Max. :16.232 Max. :5.728   
## Базофилы\_E2 Эозинофилы\_E2 Гемоглобин\_E2 Эритроциты\_E2   
## Min. :0.1854 Min. :-0.2124 Min. : 6.073 Min. :5.136   
## 1st Qu.:0.8062 1st Qu.: 3.3403 1st Qu.:11.402 1st Qu.:5.920   
## Median :1.0551 Median : 4.7428 Median :12.432 Median :6.398   
## Mean :1.0551 Mean : 4.7216 Mean :12.581 Mean :6.420   
## 3rd Qu.:1.2686 3rd Qu.: 6.0976 3rd Qu.:13.896 3rd Qu.:6.914   
## Max. :2.1228 Max. : 9.4492 Max. :16.952 Max. :8.044

После обзора данных, стало понятно, что некоторые показатели крови имеют отрицательные значения. Я посчитала, что следует заменить их на NA. Помимо этого, в столбец с группой крови вместо пропущенных столбцов закралась Северная Америка, её я тоже заменила на NA.

df <- df %>%  
 mutate\_at(vars(`Базофилы\_E1`:`Эритроциты\_E2`), ~ ifelse(. <= 0, NA, .)) %>%  
 mutate(`Группа крови` = `Группа крови` %>% na\_if("NA"))  
str(df)

## tibble [100 x 13] (S3: tbl\_df/tbl/data.frame)  
## $ Группа : Factor w/ 2 levels "Группа 1","Группа 2": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Возраст : num [1:100] 31 28 33 26 33 28 27 31 23 29 ...  
## $ Пол : Factor w/ 2 levels "Женский","Мужской": 1 1 1 1 1 2 2 2 1 1 ...  
## $ Рост : num [1:100] 174 157 166 168 170 172 157 174 175 172 ...  
## $ Группа крови : chr [1:100] "A (II)" "A (II)" NA "O (I)" ...  
## $ Базофилы\_E1 : num [1:100] 0.4222 0.327 0.7994 0.0237 0.6636 ...  
## $ Эозинофилы\_E1: num [1:100] 0.646 4.974 3.388 4.54 3.316 ...  
## $ Гемоглобин\_E1: num [1:100] 10.68 9.62 10.16 10.64 12.26 ...  
## $ Эритроциты\_E1: num [1:100] 4.26 3.88 5.06 3.81 3.03 ...  
## $ Базофилы\_E2 : num [1:100] 0.826 0.731 1.204 0.428 1.068 ...  
## $ Эозинофилы\_E2: num [1:100] 1.66 5.99 4.4 5.56 4.33 ...  
## $ Гемоглобин\_E2: num [1:100] 11.4 10.3 10.9 11.4 13 ...  
## $ Эритроциты\_E2: num [1:100] 6.57 6.2 7.38 6.12 5.35 ...

Посчитаю описательные статистики для количественных переменных.

#Cтатистики для каждой количественной переменной  
statistics\_1 <- list(  
 `\_Количество субъектов` = ~length(.x) %>% as.character(),  
 `\_Есть данные` = ~sum(!is.na(.x)) %>% as.character(),  
 `\_Нет данных` = ~sum(is.na(.x)) %>% as.character(),  
 `\_Среднее значение` = ~ifelse(sum(!is.na(.x)) == 0, "н/п", mean(.x, na.rm = TRUE) %>% round(2) %>% as.character()),  
 `\_Стандартное отклонение` = ~ifelse(sum(!is.na(.x)) == 0, "н/п", sd(.x, na.rm = TRUE) %>% round(2) %>% as.character()),  
 `\_Размах` = ~ifelse(sum(!is.na(.x)) == 0, "н/п", (max(.x, na.rm = TRUE) - min(.x, na.rm = TRUE)) %>% round(2)) %>% as.character(),  
 `\_95% ДИ для среднего` = ~sd(.x, na.rm = TRUE) %>% round(2) %>% as.character(),  
 `\_мин. - макс.` = ~ifelse(sum(!is.na(.x)) == 0, "Н/П\*", paste0(min(.x, na.rm = TRUE) %>% round(2), " - ", max(.x, na.rm = TRUE) %>% round(2))),  
 `\_Медиана` = ~ifelse(sum(!is.na(.x)) == 0, "Н/П\*", median(.x, na.rm = TRUE) %>% round(2) %>% as.character()),  
 `\_Q1 - Q3` = ~ifelse(sum(!is.na(.x)) == 0, "Н/П\*", paste0(quantile(.x, 0.25, na.rm = TRUE) %>% round(2), " - ", quantile(.x, 0.75, na.rm = TRUE) %>% round(2)))  
)  
  
#Статистики по данным анализов крови  
df %>%   
 select(`Группа`, where(is.numeric)) %>%  
 group\_by(`Группа`) %>%  
 summarise(across(where(is.numeric), statistics\_1)) %>%  
 pivot\_longer(!`Группа`) %>%  
 separate(name, into = c("Переменная", "Статистика"), sep = "\_\_") %>%  
 rename(`Значение` = value) %>%  
 flextable() %>%  
 theme\_zebra() %>%  
 merge\_v(c("Группа", "Переменная")) %>%  
 set\_table\_properties(width = 1, align = "center", layout = "autofit")

| **Группа** | **Переменная** | **Статистика** | **Значение** |
| --- | --- | --- | --- |
| Группа 1 | Возраст | Количество субъектов | 50 |
| Есть данные | 50 |
| Нет данных | 0 |
| Среднее значение | 29.24 |
| Стандартное отклонение | 3.8 |
| Размах | 17 |
| 95% ДИ для среднего | 3.8 |
| мин. - макс. | 21 - 38 |
| Медиана | 29 |
| Q1 - Q3 | 27 - 32 |
| Рост | Количество субъектов | 50 |
| Есть данные | 50 |
| Нет данных | 0 |
| Среднее значение | 168.04 |
| Стандартное отклонение | 6.16 |
| Размах | 26 |
| 95% ДИ для среднего | 6.16 |
| мин. - макс. | 155 - 181 |
| Медиана | 169 |
| Q1 - Q3 | 164 - 172 |
| Базофилы\_E1 | Количество субъектов | 50 |
| Есть данные | 47 |
| Нет данных | 3 |
| Среднее значение | 0.61 |
| Стандартное отклонение | 0.34 |
| Размах | 1.57 |
| 95% ДИ для среднего | 0.34 |
| мин. - макс. | 0.02 - 1.59 |
| Медиана | 0.56 |
| Q1 - Q3 | 0.34 - 0.84 |
| Эозинофилы\_E1 | Количество субъектов | 50 |
| Есть данные | 49 |
| Нет данных | 1 |
| Среднее значение | 3.24 |
| Стандартное отклонение | 1.7 |
| Размах | 7.86 |
| 95% ДИ для среднего | 1.7 |
| мин. - макс. | 0.27 - 8.13 |
| Медиана | 3.16 |
| Q1 - Q3 | 2.29 - 4.25 |
| Гемоглобин\_E1 | Количество субъектов | 50 |
| Есть данные | 50 |
| Нет данных | 0 |
| Среднее значение | 11.95 |
| Стандартное отклонение | 2.1 |
| Размах | 10.88 |
| 95% ДИ для среднего | 2.1 |
| мин. - макс. | 5.35 - 16.23 |
| Медиана | 11.73 |
| Q1 - Q3 | 10.65 - 13.27 |
| Эритроциты\_E1 | Количество субъектов | 50 |
| Есть данные | 50 |
| Нет данных | 0 |
| Среднее значение | 4.17 |
| Стандартное отклонение | 0.67 |
| Размах | 2.91 |
| 95% ДИ для среднего | 0.67 |
| мин. - макс. | 2.82 - 5.73 |
| Медиана | 4.23 |
| Q1 - Q3 | 3.74 - 4.63 |
| Базофилы\_E2 | Количество субъектов | 50 |
| Есть данные | 50 |
| Нет данных | 0 |
| Среднее значение | 0.97 |
| Стандартное отклонение | 0.38 |
| Размах | 1.81 |
| 95% ДИ для среднего | 0.38 |
| мин. - макс. | 0.19 - 2 |
| Медиана | 0.91 |
| Q1 - Q3 | 0.73 - 1.24 |
| Эозинофилы\_E2 | Количество субъектов | 50 |
| Есть данные | 50 |
| Нет данных | 0 |
| Среднее значение | 4.19 |
| Стандартное отклонение | 1.75 |
| Размах | 8.43 |
| 95% ДИ для среднего | 1.75 |
| мин. - макс. | 0.71 - 9.14 |
| Медиана | 4.11 |
| Q1 - Q3 | 3.2 - 5.2 |
| Гемоглобин\_E2 | Количество субъектов | 50 |
| Есть данные | 50 |
| Нет данных | 0 |
| Среднее значение | 12.67 |
| Стандартное отклонение | 2.1 |
| Размах | 10.88 |
| 95% ДИ для среднего | 2.1 |
| мин. - макс. | 6.07 - 16.95 |
| Медиана | 12.45 |
| Q1 - Q3 | 11.37 - 13.99 |
| Эритроциты\_E2 | Количество субъектов | 50 |
| Есть данные | 50 |
| Нет данных | 0 |
| Среднее значение | 6.48 |
| Стандартное отклонение | 0.67 |
| Размах | 2.91 |
| 95% ДИ для среднего | 0.67 |
| мин. - макс. | 5.14 - 8.04 |
| Медиана | 6.55 |
| Q1 - Q3 | 6.06 - 6.95 |
| Группа 2 | Возраст | Количество субъектов | 50 |
| Есть данные | 50 |
| Нет данных | 0 |
| Среднее значение | 31.26 |
| Стандартное отклонение | 3.94 |
| Размах | 21 |
| 95% ДИ для среднего | 3.94 |
| мин. - макс. | 21 - 42 |
| Медиана | 32 |
| Q1 - Q3 | 29 - 34 |
| Рост | Количество субъектов | 50 |
| Есть данные | 50 |
| Нет данных | 0 |
| Среднее значение | 167.36 |
| Стандартное отклонение | 5.4 |
| Размах | 18 |
| 95% ДИ для среднего | 5.4 |
| мин. - макс. | 159 - 177 |
| Медиана | 167 |
| Q1 - Q3 | 164 - 171 |
| Базофилы\_E1 | Количество субъектов | 50 |
| Есть данные | 49 |
| Нет данных | 1 |
| Среднее значение | 0.75 |
| Стандартное отклонение | 0.34 |
| Размах | 1.56 |
| 95% ДИ для среднего | 0.34 |
| мин. - макс. | 0.15 - 1.72 |
| Медиана | 0.72 |
| Q1 - Q3 | 0.49 - 0.95 |
| Эозинофилы\_E1 | Количество субъектов | 50 |
| Есть данные | 47 |
| Нет данных | 3 |
| Среднее значение | 4.57 |
| Стандартное отклонение | 2.03 |
| Размах | 8.3 |
| 95% ДИ для среднего | 2.03 |
| мин. - макс. | 0.14 - 8.43 |
| Медиана | 4.76 |
| Q1 - Q3 | 3.49 - 6.11 |
| Гемоглобин\_E1 | Количество субъектов | 50 |
| Есть данные | 50 |
| Нет данных | 0 |
| Среднее значение | 11.77 |
| Стандартное отклонение | 1.4 |
| Размах | 6.28 |
| 95% ДИ для среднего | 1.4 |
| мин. - макс. | 8.85 - 15.13 |
| Медиана | 11.71 |
| Q1 - Q3 | 10.77 - 13.01 |
| Эритроциты\_E1 | Количество субъектов | 50 |
| Есть данные | 50 |
| Нет данных | 0 |
| Среднее значение | 4.04 |
| Стандартное отклонение | 0.66 |
| Размах | 2.42 |
| 95% ДИ для среднего | 0.66 |
| мин. - макс. | 2.84 - 5.26 |
| Медиана | 4.03 |
| Q1 - Q3 | 3.53 - 4.53 |
| Базофилы\_E2 | Количество субъектов | 50 |
| Есть данные | 50 |
| Нет данных | 0 |
| Среднее значение | 1.14 |
| Стандартное отклонение | 0.36 |
| Размах | 1.75 |
| 95% ДИ для среднего | 0.36 |
| мин. - макс. | 0.38 - 2.12 |
| Медиана | 1.11 |
| Q1 - Q3 | 0.89 - 1.35 |
| Эозинофилы\_E2 | Количество субъектов | 50 |
| Есть данные | 48 |
| Нет данных | 2 |
| Среднее значение | 5.48 |
| Стандартное отклонение | 2.15 |
| Размах | 9.09 |
| 95% ДИ для среднего | 2.15 |
| мин. - макс. | 0.36 - 9.45 |
| Медиана | 5.75 |
| Q1 - Q3 | 4.25 - 7.08 |
| Гемоглобин\_E2 | Количество субъектов | 50 |
| Есть данные | 50 |
| Нет данных | 0 |
| Среднее значение | 12.49 |
| Стандартное отклонение | 1.4 |
| Размах | 6.28 |
| 95% ДИ для среднего | 1.4 |
| мин. - макс. | 9.57 - 15.85 |
| Медиана | 12.43 |
| Q1 - Q3 | 11.49 - 13.73 |
| Эритроциты\_E2 | Количество субъектов | 50 |
| Есть данные | 50 |
| Нет данных | 0 |
| Среднее значение | 6.36 |
| Стандартное отклонение | 0.66 |
| Размах | 2.42 |
| 95% ДИ для среднего | 0.66 |
| мин. - макс. | 5.16 - 7.58 |
| Медиана | 6.34 |
| Q1 - Q3 | 5.84 - 6.85 |

Посчитаю статистики для категориальных переменных.

df %>%   
 select(`Группа`, where(is.character)) %>%  
 mutate(`Группа крови` = `Группа крови` %>% replace\_na("Нет данных") %>% as.factor()) %>%  
 count(`Группа`, `Группа крови`) %>%  
 group\_by(`Группа`) %>%  
 mutate(`Процент по группе` = (n / sum(n)) %>% round(4) %>% `\*`(100) %>% str\_c("%")) %>%  
 ungroup() %>%  
 mutate(`Процент по выборке` = (n / sum(n)) %>% round(4) %>% `\*`(100) %>% str\_c("%")) %>%  
 flextable() %>%  
 theme\_zebra() %>%  
 merge\_v("Группа") %>%  
 set\_table\_properties(width = 1, align = "center", layout = "autofit")

| **Группа** | **Группа крови** | **n** | **Процент по группе** | **Процент по выборке** |
| --- | --- | --- | --- | --- |
| Группа 1 | A (II) | 16 | 32% | 16% |
| AB (IV) | 4 | 8% | 4% |
| B (III) | 11 | 22% | 11% |
| O (I) | 10 | 20% | 10% |
| Нет данных | 9 | 18% | 9% |
| Группа 2 | A (II) | 18 | 36% | 18% |
| AB (IV) | 4 | 8% | 4% |
| B (III) | 5 | 10% | 5% |
| O (I) | 15 | 30% | 15% |
| Нет данных | 8 | 16% | 8% |

Таблица со сравнением двух групп.

df %>%  
 tbl\_summary(by = "Группа", missing\_text = "Нет данных") %>%   
 add\_p() %>%  
 as\_flex\_table()

| Characteristic | Группа 1, N = 501 | Группа 2, N = 501 | p-value2 |
| --- | --- | --- | --- |
| Возраст | 29.0 (27.0, 32.0) | 32.0 (29.0, 34.0) | 0.008 |
| Пол |  |  | 0.028 |
| Женский | 32 (64%) | 21 (42%) |  |
| Мужской | 18 (36%) | 29 (58%) |  |
| Рост | 169.0 (164.0, 172.0) | 167.0 (164.0, 171.0) | 0.4 |
| Группа крови |  |  | 0.3 |
| A (II) | 16 (39%) | 18 (43%) |  |
| AB (IV) | 4 (9.8%) | 4 (9.5%) |  |
| B (III) | 11 (27%) | 5 (12%) |  |
| O (I) | 10 (24%) | 15 (36%) |  |
| Нет данных | 9 | 8 |  |
| Базофилы\_E1 | 0.56 (0.34, 0.84) | 0.72 (0.49, 0.95) | 0.043 |
| Нет данных | 3 | 1 |  |
| Эозинофилы\_E1 | 3.16 (2.29, 4.25) | 4.76 (3.49, 6.11) | <0.001 |
| Нет данных | 1 | 3 |  |
| Гемоглобин\_E1 | 11.73 (10.65, 13.27) | 11.71 (10.77, 13.01) | 0.6 |
| Эритроциты\_E1 | 4.23 (3.74, 4.63) | 4.03 (3.53, 4.53) | 0.3 |
| Базофилы\_E2 | 0.91 (0.73, 1.24) | 1.11 (0.89, 1.35) | 0.024 |
| Эозинофилы\_E2 | 4.11 (3.20, 5.20) | 5.75 (4.25, 7.08) | <0.001 |
| Нет данных | 0 | 2 |  |
| Гемоглобин\_E2 | 12.45 (11.37, 13.99) | 12.43 (11.49, 13.73) | 0.6 |
| Эритроциты\_E2 | 6.55 (6.06, 6.95) | 6.34 (5.84, 6.85) | 0.3 |
| 1Median (IQR); n (%) | | | |
| 2Wilcoxon rank sum test; Pearson's Chi-squared test; Fisher's exact test; Wilcoxon rank sum exact test | | | |

Сколько красивых табличек можно делать с помощью gtsummary… <https://www.danieldsjoberg.com/gtsummary/articles/tbl_summary.html> <https://www.pipinghotdata.com/posts/2021-07-14-polished-summary-tables-in-r-with-gtsummary/>