Статистический отчет

Оценка показателей крови в двух группах для взитов 1, 2

Liliya

2022-11-19

* Cтатистический отчёт с таблицами описательной статистики по всем имеющимся визитам по всем группам для каждой переменной
* Сравнение групп

### Загрузка библиотек

#install.packages(c("dplyr", "tidyr", "readxl", "purrr", "tibble", "ggplot2", "ggpubr","tidyverse", "flextable"))  
#install.packages("readr")  
#install.packages("stringr")  
#install.packages("split")  
  
# dplyr - базовая грамматика манипуляции данных  
# tidyr - приведение данных к аккуратному виду   
# ggplot2, ggpubr - визуализация данных  
# purrr - дополнительные функции для функционального программирования  
# tidyverse - метапакет, содержащий в себе основные пакеты для автоматизации анализа данных  
# tibble - улучшенный аналог data.frame  
# readxl - чтение данных  
# flextable - статистические пакеты для печати  
# stringr - извлечение подстроки  
# DescTools, psych - пакеты со статистическими финкциями  
  
#detach(package:dplyr)

### Чтение данных

library('readxl')  
library('readr')  
data <- readxl::read\_excel('data\_excel.xlsx')  
data <- readr::read\_tsv("data\_excel.txt")

## Rows: 100 Columns: 13  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: "\t"  
## chr (3): Группа, Пол, Группа крови  
## dbl (10): Возраст, Рост, Базофилы\_E1, Эозинофилы\_E1, Гемоглобин\_E1, Эритроци...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

### Сводная статистика

Базовая функция summary, дающая нам сводку по всем типам переменных:

* для количественных переменных возвращает: минимум, первый квартиль, медиану, среднее, третий квартиль, максимум;
* для категориальных/строковых переменных возвращает просто количество каждой категории.

summary(data)

## 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 Min. :-0.2188 Min. :-1.227 Min. : 5.352   
## Class :character 1st Qu.: 0.4020 1st Qu.: 2.325 1st Qu.:10.681   
## Mode :character Median : 0.6509 Median : 3.728 Median :11.711   
## Mean : 0.6509 Mean : 3.707 Mean :11.860   
## 3rd Qu.: 0.8644 3rd Qu.: 5.083 3rd Qu.:13.175   
## Max. : 1.7186 Max. : 8.434 Max. :16.232   
## Эритроциты\_E1 Базофилы\_E2 Эозинофилы\_E2 Гемоглобин\_E2   
## Min. :2.821 Min. :0.1854 Min. :-0.2124 Min. : 6.073   
## 1st Qu.:3.605 1st Qu.:0.8062 1st Qu.: 3.3403 1st Qu.:11.402   
## Median :4.082 Median :1.0551 Median : 4.7428 Median :12.432   
## Mean :4.104 Mean :1.0551 Mean : 4.7216 Mean :12.581   
## 3rd Qu.:4.599 3rd Qu.:1.2686 3rd Qu.: 6.0976 3rd Qu.:13.896   
## Max. :5.728 Max. :2.1228 Max. : 9.4492 Max. :16.952   
## Эритроциты\_E2   
## Min. :5.136   
## 1st Qu.:5.920   
## Median :6.398   
## Mean :6.420   
## 3rd Qu.:6.914   
## Max. :8.044

library('psych')  
psych::describe(data, na.rm = TRUE, skew = FALSE, ranges = TRUE)

## vars n mean sd min max range se  
## Группа\* 1 100 1.50 0.50 1.00 2.00 1.00 0.05  
## Возраст 2 100 30.25 3.98 21.00 42.00 21.00 0.40  
## Пол\* 3 100 1.47 0.50 1.00 2.00 1.00 0.05  
## Рост 4 100 167.70 5.77 155.00 181.00 26.00 0.58  
## Группа крови\* 5 83 2.39 1.30 1.00 4.00 3.00 0.14  
## Базофилы\_E1 6 100 0.65 0.38 -0.22 1.72 1.94 0.04  
## Эозинофилы\_E1 7 100 3.71 2.15 -1.23 8.43 9.66 0.21  
## Гемоглобин\_E1 8 100 11.86 1.78 5.35 16.23 10.88 0.18  
## Эритроциты\_E1 9 100 4.10 0.67 2.82 5.73 2.91 0.07  
## Базофилы\_E2 10 100 1.06 0.38 0.19 2.12 1.94 0.04  
## Эозинофилы\_E2 11 100 4.72 2.15 -0.21 9.45 9.66 0.21  
## Гемоглобин\_E2 12 100 12.58 1.78 6.07 16.95 10.88 0.18  
## Эритроциты\_E2 13 100 6.42 0.67 5.14 8.04 2.91 0.07

### Статистика по группам

table(data$Группа, data$Возраст)

##   
## 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 42  
## Группа 1 1 0 3 1 4 3 5 7 2 3 6 4 5 2 2 1 0 1 0  
## Группа 2 1 1 0 0 2 1 2 4 5 5 2 6 7 6 4 0 2 1 1

table(data$Группа, data$Пол)

##   
## Женский Мужской  
## Группа 1 32 18  
## Группа 2 21 29

table(data$Группа, data$Рост)

##   
## 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171  
## Группа 1 1 2 2 1 0 0 0 3 1 4 3 2 2 2 3 4 6  
## Группа 2 0 0 0 0 4 2 4 1 1 5 3 2 4 4 3 3 3  
##   
## 172 173 174 175 176 177 179 181  
## Группа 1 3 2 2 3 1 1 1 1  
## Группа 2 1 1 2 1 4 2 0 0

table(data$Группа, data$`Группа крови`)

##   
## A (II) AB (IV) B (III) O (I)  
## Группа 1 16 4 11 10  
## Группа 2 18 4 5 15

### Статистика по визитам

library('dplyr')

##   
## Присоединяю пакет: 'dplyr'

## Следующие объекты скрыты от 'package:stats':  
##   
## filter, lag

## Следующие объекты скрыты от 'package:base':  
##   
## intersect, setdiff, setequal, union

data %>%  
 rowwise() %>%  
 mutate('Среднее по базофилам за два визита' = mean(c(`Базофилы\_E1`, `Базофилы\_E2`))) %>%  
 ungroup() %>%  
 select(`Базофилы\_E1`, `Базофилы\_E2`, `Среднее по базофилам за два визита`)

## # A tibble: 100 × 3  
## Базофилы\_E1 Базофилы\_E2 `Среднее по базофилам за два визита`  
## <dbl> <dbl> <dbl>  
## 1 0.422 0.826 0.624  
## 2 0.327 0.731 0.529  
## 3 0.799 1.20 1.00   
## 4 0.0237 0.428 0.226  
## 5 0.664 1.07 0.866  
## 6 0.481 0.885 0.683  
## 7 0.890 1.29 1.09   
## 8 0.858 1.26 1.06   
## 9 0.383 0.787 0.585  
## 10 0.281 0.685 0.483  
## # … with 90 more rows

data %>%  
 rowwise() %>%  
 mutate('Среднее по эозинофилам за два визита' = mean(c(`Эозинофилы\_E1`, `Эозинофилы\_E2`))) %>%  
 ungroup() %>%  
 select(`Эозинофилы\_E1`, `Эозинофилы\_E2`, `Среднее по эозинофилам за два визита`)

## # A tibble: 100 × 3  
## Эозинофилы\_E1 Эозинофилы\_E2 `Среднее по эозинофилам за два визита`  
## <dbl> <dbl> <dbl>  
## 1 0.646 1.66 1.15  
## 2 4.97 5.99 5.48  
## 3 3.39 4.40 3.89  
## 4 4.54 5.56 5.05  
## 5 3.32 4.33 3.82  
## 6 2.79 3.80 3.29  
## 7 2.34 3.36 2.85  
## 8 3.98 4.99 4.49  
## 9 3.39 4.40 3.90  
## 10 7.94 8.95 8.44  
## # … with 90 more rows

data %>%  
 rowwise() %>%  
 mutate('Среднее по гемоглобинам за два визита' = mean(c(`Гемоглобин\_E1`, `Гемоглобин\_E2`))) %>%  
 ungroup() %>%  
 select(`Гемоглобин\_E1`, `Гемоглобин\_E2`, `Среднее по гемоглобинам за два визита`)

## # A tibble: 100 × 3  
## Гемоглобин\_E1 Гемоглобин\_E2 `Среднее по гемоглобинам за два визита`  
## <dbl> <dbl> <dbl>  
## 1 10.7 11.4 11.0   
## 2 9.62 10.3 9.98  
## 3 10.2 10.9 10.5   
## 4 10.6 11.4 11.0   
## 5 12.3 13.0 12.6   
## 6 12.4 13.1 12.7   
## 7 11.3 12.0 11.6   
## 8 14.9 15.6 15.3   
## 9 15.7 16.4 16.0   
## 10 12.9 13.6 13.3   
## # … with 90 more rows

data %>%  
 rowwise() %>%  
 mutate('Среднее по эритроцитам за два визита' = mean(c(`Эритроциты\_E1`, `Эритроциты\_E2`))) %>%  
 ungroup() %>%  
 select(`Эритроциты\_E1`, `Эритроциты\_E2`, `Среднее по эритроцитам за два визита`)

## # A tibble: 100 × 3  
## Эритроциты\_E1 Эритроциты\_E2 `Среднее по эритроцитам за два визита`  
## <dbl> <dbl> <dbl>  
## 1 4.26 6.57 5.42  
## 2 3.88 6.20 5.04  
## 3 5.06 7.38 6.22  
## 4 3.81 6.12 4.97  
## 5 3.03 5.35 4.19  
## 6 4.79 7.11 5.95  
## 7 3.36 5.68 4.52  
## 8 4.46 6.78 5.62  
## 9 4.43 6.75 5.59  
## 10 4.33 6.64 5.49  
## # … with 90 more rows

### Таблицы со сравнением групп

data %>%  
 split(~ `Группа`)

## $`Группа 1`  
## # A tibble: 50 × 13  
## Группа Возраст Пол Рост Групп…¹ Базоф…² Эозин…³ Гемог…⁴ Эритр…⁵ Базоф…⁶  
## <chr> <dbl> <chr> <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Группа 1 31 Женск… 174 A (II) 0.422 0.646 10.7 4.26 0.826  
## 2 Группа 1 28 Женск… 157 A (II) 0.327 4.97 9.62 3.88 0.731  
## 3 Группа 1 33 Женск… 166 <NA> 0.799 3.39 10.2 5.06 1.20   
## 4 Группа 1 26 Женск… 168 O (I) 0.0237 4.54 10.6 3.81 0.428  
## 5 Группа 1 33 Женск… 170 A (II) 0.664 3.32 12.3 3.03 1.07   
## 6 Группа 1 28 Мужск… 172 B (III) 0.481 2.79 12.4 4.79 0.885  
## 7 Группа 1 27 Мужск… 157 A (II) 0.890 2.34 11.3 3.36 1.29   
## 8 Группа 1 31 Мужск… 174 <NA> 0.858 3.98 14.9 4.46 1.26   
## 9 Группа 1 23 Женск… 175 A (II) 0.383 3.39 15.7 4.43 0.787  
## 10 Группа 1 29 Женск… 172 A (II) 0.281 7.94 12.9 4.33 0.685  
## # … with 40 more rows, 3 more variables: Эозинофилы\_E2 <dbl>,  
## # Гемоглобин\_E2 <dbl>, Эритроциты\_E2 <dbl>, and abbreviated variable names  
## # ¹​`Группа крови`, ²​Базофилы\_E1, ³​Эозинофилы\_E1, ⁴​Гемоглобин\_E1,  
## # ⁵​Эритроциты\_E1, ⁶​Базофилы\_E2  
##   
## $`Группа 2`  
## # A tibble: 50 × 13  
## Группа Возраст Пол Рост Групп…¹ Базоф…² Эозин…³ Гемог…⁴ Эритр…⁵ Базоф…⁶  
## <chr> <dbl> <chr> <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Группа 2 37 Мужск… 165 AB (IV) 0.895 5.13 13.5 2.93 1.30   
## 2 Группа 2 25 Женск… 168 O (I) 1.33 6.90 13.3 4.02 1.73   
## 3 Группа 2 32 Женск… 159 A (II) -0.0283 4.03 8.85 3.50 0.376  
## 4 Группа 2 35 Мужск… 164 AB (IV) 0.828 7.10 10.2 3.95 1.23   
## 5 Группа 2 29 Мужск… 170 O (I) 0.412 6.02 9.66 4.09 0.816  
## 6 Группа 2 34 Мужск… 169 B (III) 0.337 3.68 12.0 5.26 0.741  
## 7 Группа 2 34 Мужск… 161 A (II) 0.774 4.40 11.9 3.25 1.18   
## 8 Группа 2 27 Женск… 167 <NA> 0.191 3.83 11.3 5.08 0.595  
## 9 Группа 2 34 Женск… 175 AB (IV) 0.414 -1.05 10.6 3.95 0.818  
## 10 Группа 2 30 Мужск… 167 A (II) 0.841 1.30 15.1 3.75 1.24   
## # … with 40 more rows, 3 more variables: Эозинофилы\_E2 <dbl>,  
## # Гемоглобин\_E2 <dbl>, Эритроциты\_E2 <dbl>, and abbreviated variable names  
## # ¹​`Группа крови`, ²​Базофилы\_E1, ³​Эозинофилы\_E1, ⁴​Гемоглобин\_E1,  
## # ⁵​Эритроциты\_E1, ⁶​Базофилы\_E2

#### Расчёт средних значений по группам

library(`dplyr`)  
data %>%   
 select(`Группа`, where(is.numeric)) %>%  
 group\_by(`Группа`) %>%  
 dplyr::summarise(across(where(is.numeric), function(x) mean(x, na.rm = TRUE)))

## # A tibble: 2 × 11  
## Группа Возраст Рост Базоф…¹ Эозин…² Гемог…³ Эритр…⁴ Базоф…⁵ Эозин…⁶ Гемог…⁷  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Группа 1 29.2 168. 0.564 3.17 11.9 4.17 0.968 4.19 12.7  
## 2 Группа 2 31.3 167. 0.738 4.24 11.8 4.04 1.14 5.26 12.5  
## # … with 1 more variable: Эритроциты\_E2 <dbl>, and abbreviated variable names  
## # ¹​Базофилы\_E1, ²​Эозинофилы\_E1, ³​Гемоглобин\_E1, ⁴​Эритроциты\_E1, ⁵​Базофилы\_E2,  
## # ⁶​Эозинофилы\_E2, ⁷​Гемоглобин\_E2

#### Расчёт статистической таблицы для количественных и категориальных переменных по группам

library('tidyr')  
library('stringr')  
statistics <- list(  
 `Количество субъектов` = ~length(.x) %>% as.character(),  
 `Количество (есть данные)` = ~sum(!is.na(.x)) %>% as.character(),  
 `Нет данных` = ~sum(is.na(.x)) %>% as.character(),  
 `Ср. знач.` = ~ifelse(sum(!is.na(.x)) == 0, "Н/П\*", mean(.x, na.rm = TRUE) %>% round(2) %>% as.character()),  
 `Станд. отклон.` = ~ifelse(sum(!is.na(.x)) < 3, "Н/П\*", sd(.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)))  
)  
  
# для количественных  
data %>%  
 select(`Группа`, where(is.numeric)) %>%  
 group\_by(`Группа`) %>%  
 summarise(across(where(is.numeric), statistics)) %>%  
 pivot\_longer(!`Группа`) %>%  
 separate(name, into = c("Переменная", "Статистика"), sep = "\_\_") %>%  
 rename(`Значение` = value)

## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 140 rows [1, 2,  
## 3, 4, 5, 6, 7, 10, 11, 12, 13, 14, 15, 16, 19, 20, 21, 22, 23, 24, ...].

## # A tibble: 180 × 4  
## Группа Переменная Статистика Значение  
## <chr> <chr> <chr> <chr>   
## 1 Группа 1 Возраст\_Количество субъектов <NA> 50   
## 2 Группа 1 Возраст\_Количество (есть данные) <NA> 50   
## 3 Группа 1 Возраст\_Нет данных <NA> 0   
## 4 Группа 1 Возраст\_Ср. знач. <NA> 29.24   
## 5 Группа 1 Возраст\_Станд. отклон. <NA> 3.8   
## 6 Группа 1 Возраст\_95% ДИ для среднего <NA> 3.8   
## 7 Группа 1 Возраст\_мин. - макс. <NA> 21 - 38   
## 8 Группа 1 Возраст Медиана 29   
## 9 Группа 1 Возраст Q1 - Q3 27 - 32   
## 10 Группа 1 Рост\_Количество субъектов <NA> 50   
## # … with 170 more rows

data$Пол <- as.factor(data$Пол)  
data$`Группа крови` <- as.factor(data$`Группа крови`)  
  
# для категориальных  
data %>%  
 select(`Группа`, where(is.factor)) %>%  
 mutate(`Группа крови` = `Группа крови` %>% as.character() %>% 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("%"))

## # A tibble: 10 × 5  
## Группа `Группа крови` n `Процент по группе` `Процент по выборке`  
## <chr> <fct> <int> <chr> <chr>   
## 1 Группа 1 A (II) 16 32% 16%   
## 2 Группа 1 AB (IV) 4 8% 4%   
## 3 Группа 1 B (III) 11 22% 11%   
## 4 Группа 1 O (I) 10 20% 10%   
## 5 Группа 1 Нет данных 9 18% 9%   
## 6 Группа 2 A (II) 18 36% 18%   
## 7 Группа 2 AB (IV) 4 8% 4%   
## 8 Группа 2 B (III) 5 10% 5%   
## 9 Группа 2 O (I) 15 30% 15%   
## 10 Группа 2 Нет данных 8 16% 8%