

# RNA-seq course- week1

Serhiy Naumenko

2023-09-30

## Contents

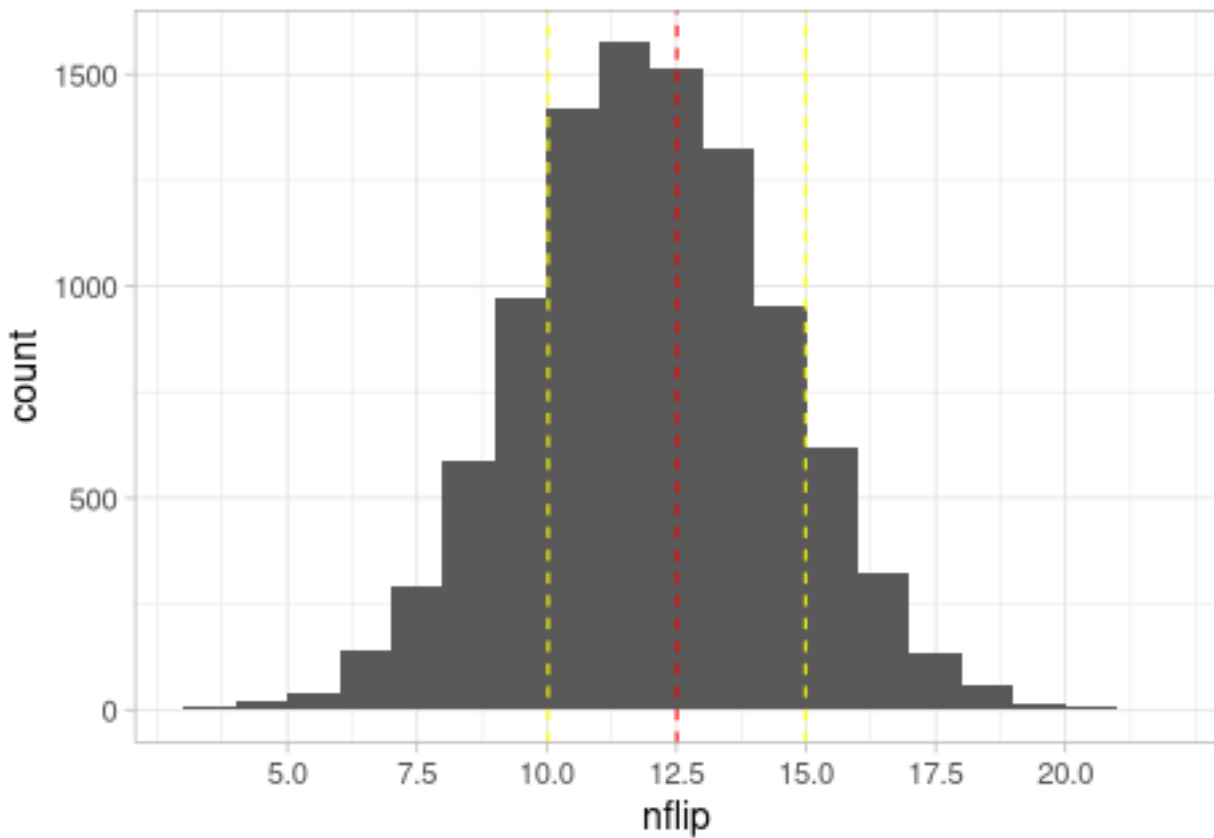
<b>1</b>	<b>Overview</b>	<b>1</b>
<b>2</b>	<b>Біноміальний розподіл</b>	<b>2</b>
<b>3</b>	<b>Спостереження</b>	<b>3</b>
<b>4</b>	<b>P-значення</b>	<b>4</b>
<b>5</b>	<b>nflir &gt;= 15</b>	<b>5</b>
<b>6</b>	<b>Кластер хвороби</b>	<b>6</b>
<b>7</b>	<b>sessionInfo()</b>	<b>8</b>

## 1 Overview

- P-values and multiple testing correction
- <https://bookdown.org/jgscott/DSGI/p-values.html>
- [https://uk.wikipedia.org/wiki/Метод\\_Монте-Карло](https://uk.wikipedia.org/wiki/Метод_Монте-Карло)
- [https://uk.wikipedia.org/wiki/Біноміальний\\_розподіл](https://uk.wikipedia.org/wiki/Біноміальний_розподіл)

## 2 Біноміальний розподіл

```
# heads/tails -  
p <- 0.5  
n <- 25  
bootstrap_n <- 10000  
binomial_sim <- do(bootstrap_n) * nflip(p = p, n = n)  
ggplot(binomial_sim) +  
  geom_histogram(aes(x=nflip), binwidth = 1, boundary = 5) +  
  geom_vline(xintercept = mean(binomial_sim$nflip), linetype = "dashed", color = "red") +  
  geom_vline(xintercept = mean(binomial_sim$nflip) + sd(binomial_sim$nflip), linetype = "dashed", color = "yellow") +  
  geom_vline(xintercept = mean(binomial_sim$nflip) - sd(binomial_sim$nflip), linetype = "dashed", color = "yellow") +  
  scale_x_continuous(breaks = c(5, 7.5, 10, 12.5, 15, 17.5, 20))
```



- Mean= 12.5101
- Var = 6.1551135
- SD = 2.4809501

### 3 Спостереження

```
#      19
rare_event <- 19
sum(binomial_sim >= rare_event) / bootstrap_n

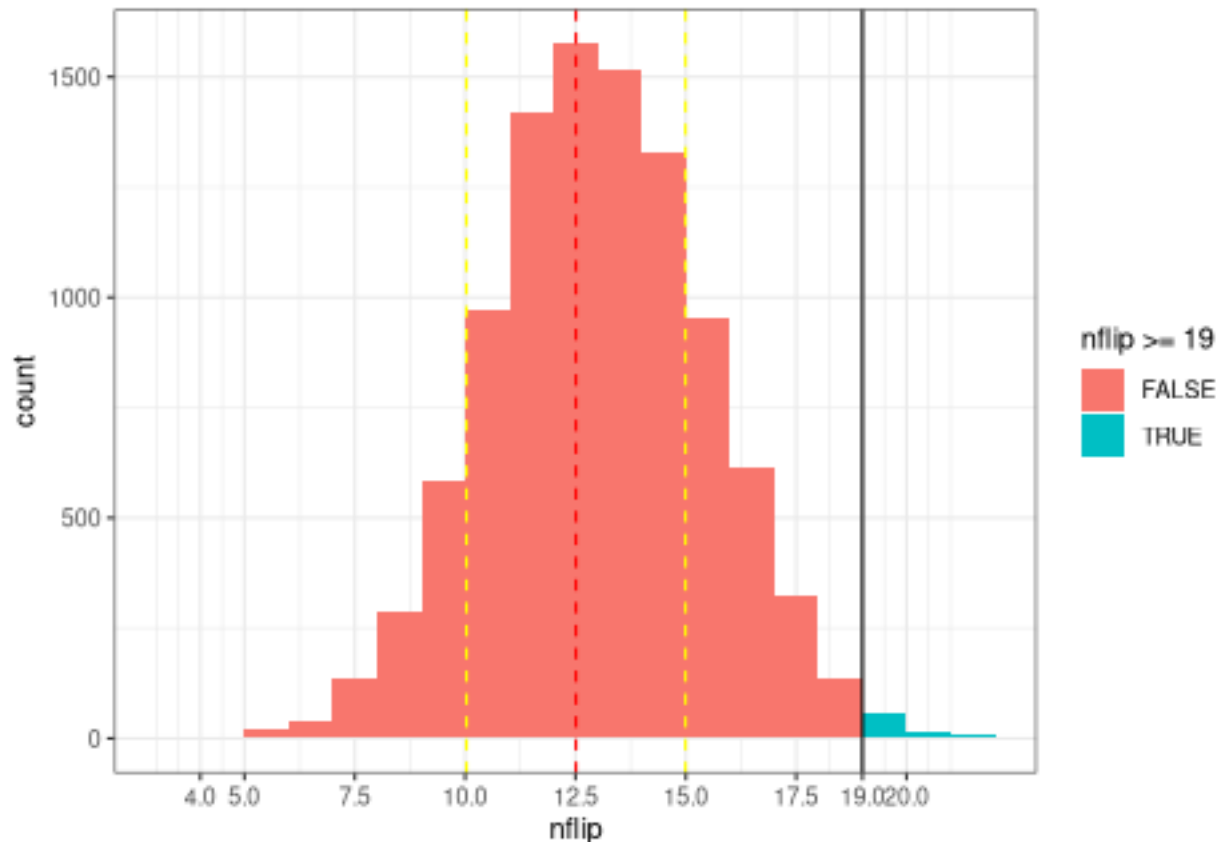
## [1] 0.008

binomial_sim$nflip %>% table()

## .
##    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18
##    1    5   22   41  138  289  585  972 1419 1574 1517 1327  955  617  325  133
##   19   20   21   22
##   60   14    5    1
```

## 4 Р-значення

```
ggplot(binomial_sim) +  
  geom_histogram(aes(x = nflip, fill = nflip >= 19), binwidth = 1, boundary = 5, closed = "left") +  
  geom_vline(xintercept = mean(binomial_sim$nflip), linetype = "dashed", color = "red") +  
  geom_vline(xintercept = mean(binomial_sim$nflip) + sd(binomial_sim$nflip), linetype = "dashed", color = "yellow") +  
  geom_vline(xintercept = mean(binomial_sim$nflip) - sd(binomial_sim$nflip), linetype = "dashed", color = "yellow") +  
  geom_vline(xintercept = 19, linetype = "solid", color = "black") +  
  scale_x_continuous(breaks = c(4, 5, 7.5, 10, 12.5, 15, 17.5, 19, 20),  
    limits = c(min(binomial_sim$nflip), max(binomial_sim$nflip))) + theme_bw()
```

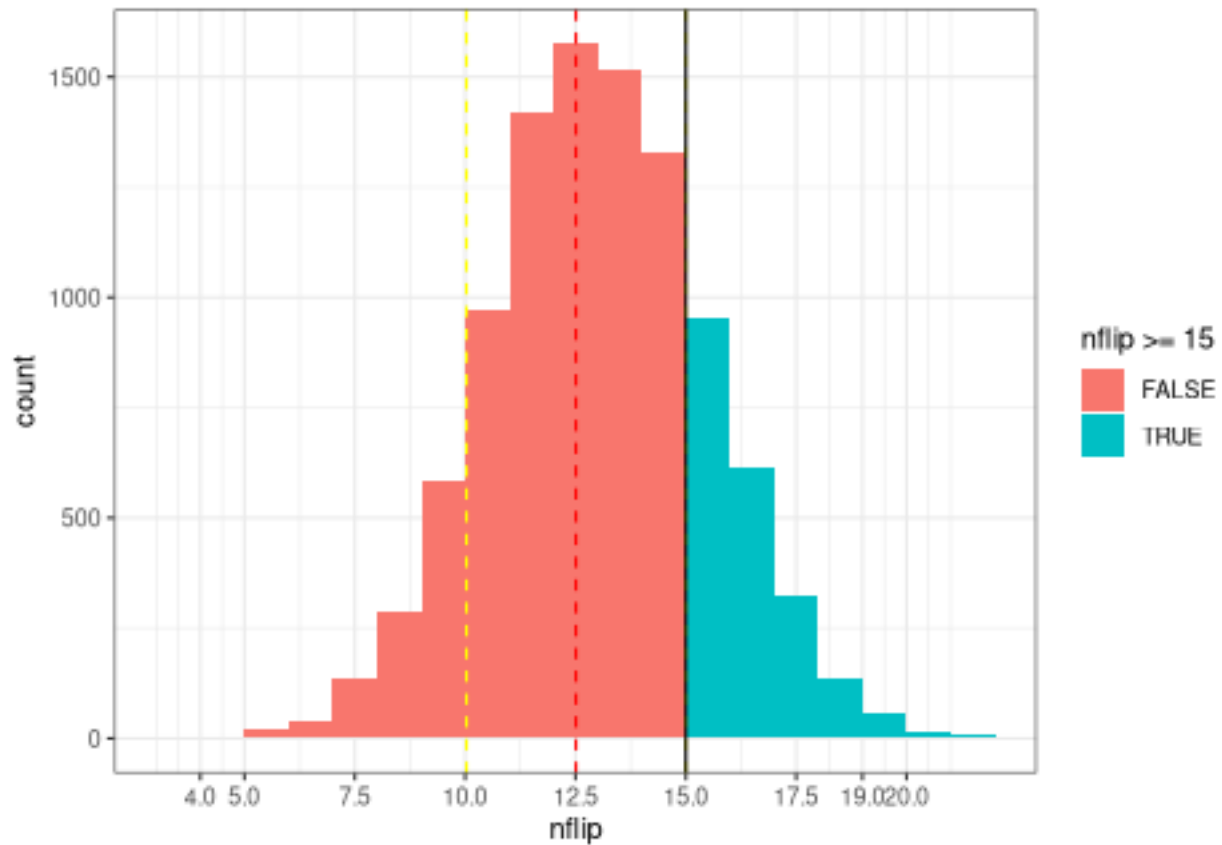


```
#scale_y_log10()
```

- $H_0$ : монетка випадкова ( $p = 0.5$ ) (нема біологічного сигналу)
- тестова статистика: скільки успіхів (heads) із 25 експериментів
- порахували розподіл ймовірності за методом Монте-Карло на 10,000 повторів
- оцінили, наскільки ймовірно, що  $H_0$  пояснює спостереження (більше 19)

## 5 nflip >= 15

```
ggplot(binomial_sim) +  
  geom_histogram(aes(x = nflip, fill = nflip >= 15), binwidth = 1, boundary = 5, closed = "left") +  
  geom_vline(xintercept = mean(binomial_sim$nflip), linetype = "dashed", color = "red") +  
  geom_vline(xintercept = mean(binomial_sim$nflip) + sd(binomial_sim$nflip), linetype = "dashed", color = "red") +  
  geom_vline(xintercept = mean(binomial_sim$nflip) - sd(binomial_sim$nflip), linetype = "dashed", color = "red") +  
  geom_vline(xintercept = 15, linetype = "solid", color = "black") +  
  scale_x_continuous(breaks = c(4, 5, 7.5, 10, 12.5, 15, 17.5, 19, 20),  
    limits = c(min(binomial_sim$nflip), max(binomial_sim$nflip))) + theme_bw()
```



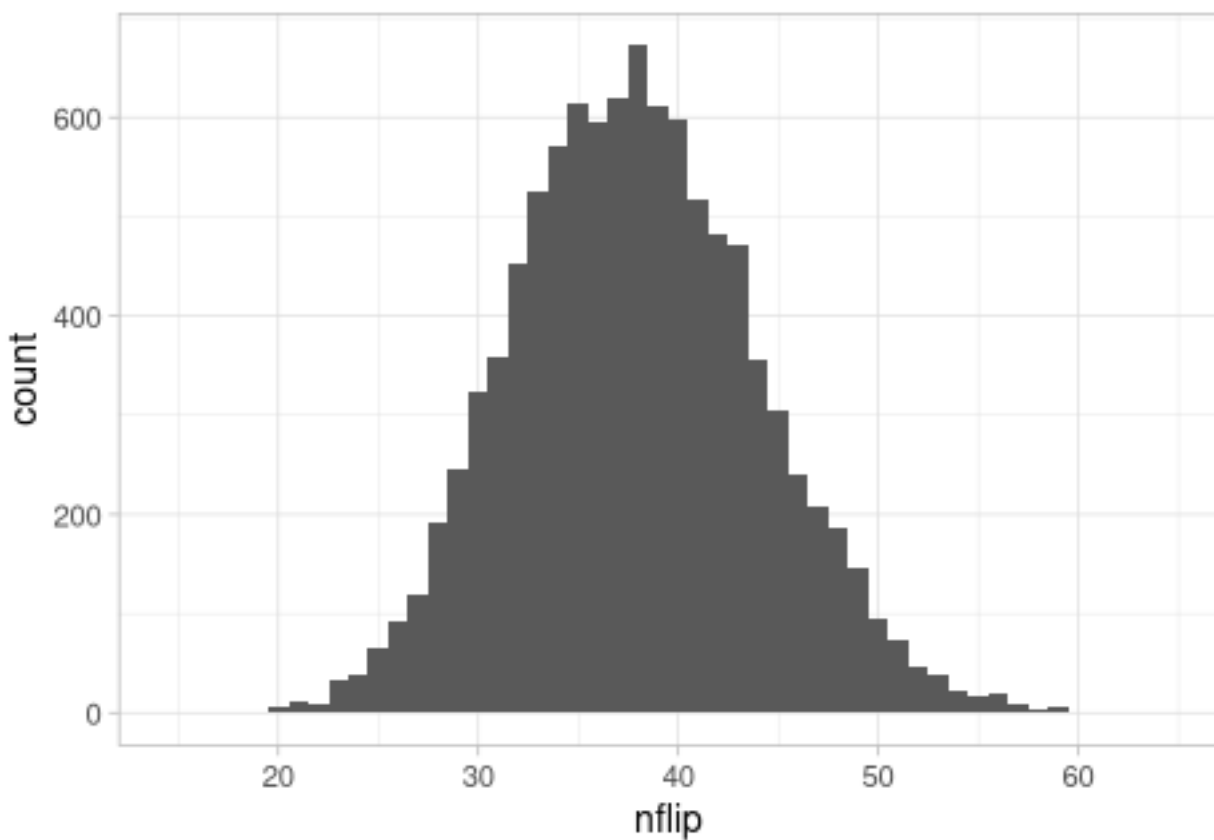
```
rare_event <- 15  
sum(binomial_sim >= rare_event) / bootstrap_n
```

```
## [1] 0.211
```

## 6 Кластер хвороби

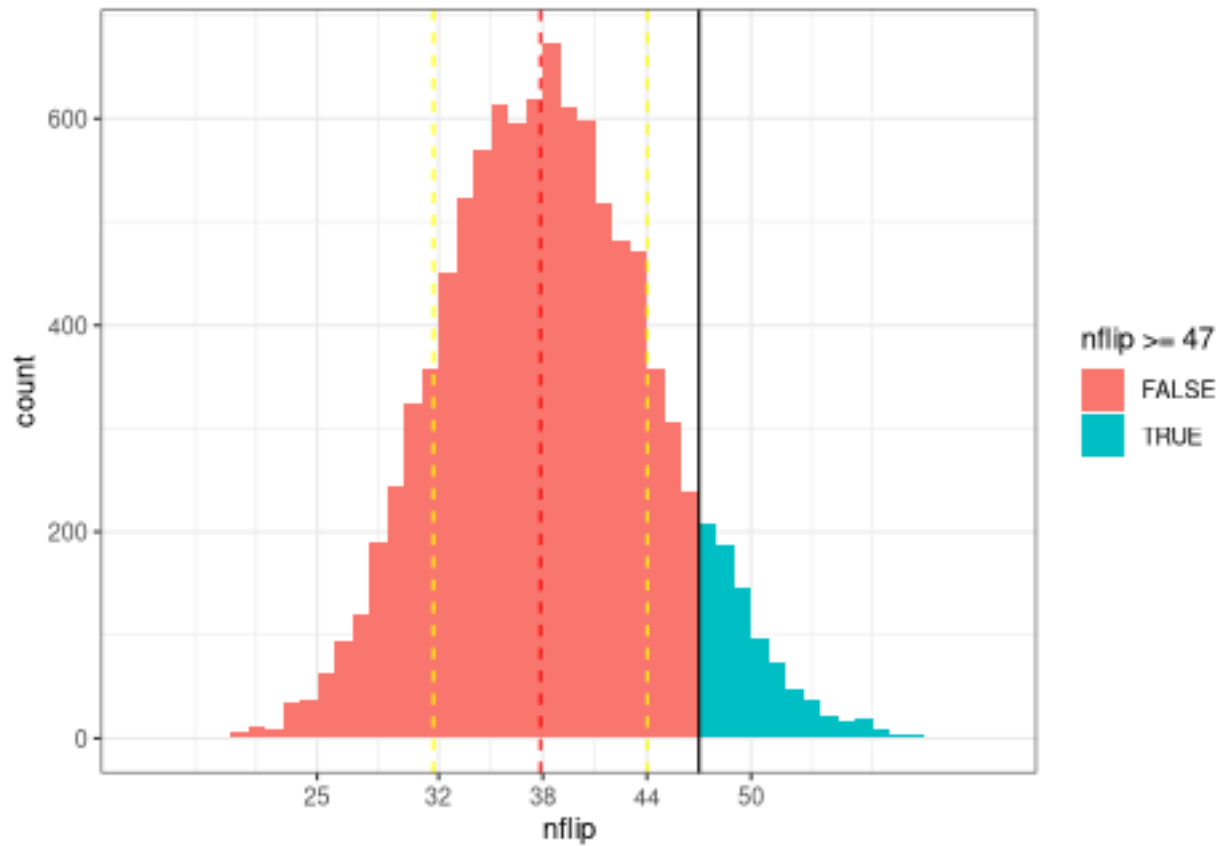
- < 10 км від атомної станції: 5.8 випадків на 10,000: 47 / 80,515
- > 30 км від атомної станції: 4.7 випадків на 10,000: 851 / 1,819,636
- incidence ratio:  $5.8/4.7 = 1.23$
- $H_0$ : IR = 4.7
- тестова статистика: кількість захворювань %
- розподіл тестової статистики за припущення  $H_0$

```
sim_cancer = do(10000)*nflip(n = 80515, prob = 0.00047)
ggplot(sim_cancer) +
  geom_histogram(aes(x = nflip), binwidth = 1)
```



- Mean= 37.8623
- Var = 37.9789366
- SD = 6.1627053
- P = 0.0874

```
ggplot(sim_cancer) +
  geom_histogram(aes(x = nflip, fill = nflip >= 47), binwidth = 1, boundary = 5, closed = "left") +
  geom_vline(xintercept = mean(sim_cancer$nflip), linetype = "dashed", color = "red") +
  geom_vline(xintercept = mean(sim_cancer$nflip) + sd(sim_cancer$nflip), linetype = "dashed", color = "red") +
  geom_vline(xintercept = mean(sim_cancer$nflip) - sd(sim_cancer$nflip), linetype = "dashed", color = "red") +
  geom_vline(xintercept = 47, linetype = "solid", color = "black") +
  scale_x_continuous(breaks = c(25, 32, 38, 44, 50),
    limits = c(min(sim_cancer$nflip), max(sim_cancer$nflip))) + theme_bw()
```



## 7 sessionInfo()

```
sessionInfo()
```

```
## R version 4.2.2 (2022-10-31)
## Platform: x86_64-redhat-linux-gnu (64-bit)
## Running under: Fedora Linux 37 (Workstation Edition)
##
## Matrix products: default
## BLAS/LAPACK: /usr/lib64/libflexiblas.so.3.3
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8      LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8      LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8  LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8     LC_NAME=C
##  [9] LC_ADDRESS=C             LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats      graphics  grDevices  utils      datasets  methods    base
##
## other attached packages:
##  [1] knitr_1.41      mosaic_1.8.4.2  mosaicData_0.20.3 ggformula_0.10.4
##  [5] Matrix_1.6-1    lattice_0.20-45 forcats_0.5.2    stringr_1.5.0
##  [9] dplyr_1.0.10    purrr_1.0.0     readr_2.1.3      tidyr_1.2.1
## [13] tibble_3.1.8    ggplot2_3.4.0   tidyverse_1.3.2
##
## loaded via a namespace (and not attached):
##  [1] httr_1.4.4      jsonlite_1.8.4   modelr_0.1.10
##  [4] assertthat_0.2.1 highr_0.10        googlesheets4_1.0.1
##  [7] ggstance_0.3.6  cellranger_1.1.0 yaml_2.3.6
## [10] pillar_1.8.1    backports_1.4.1  glue_1.6.2
## [13] digest_0.6.31   polyclip_1.10-4  rvest_1.0.3
## [16] colorspace_2.0-3 htmltools_0.5.4  pkgconfig_2.0.3
## [19] broom_1.0.2     labelled_2.12.0  haven_2.5.1
## [22] scales_1.2.1    tweenr_2.0.2     tzdb_0.3.0
## [25] ggforce_0.4.1   timechange_0.1.1 googledrive_2.0.0
## [28] generics_0.1.3  farver_2.1.1     ellipsis_0.3.2
## [31] withr_2.5.0     cli_3.5.0        magrittr_2.0.3
## [34] crayon_1.5.2    readxl_1.4.1     evaluate_0.19
## [37] fs_1.5.2        fansi_1.0.3      MASS_7.3-58.1
## [40] xml2_1.3.3      tools_4.2.2      hms_1.1.2
## [43] gargle_1.2.0    lifecycle_1.0.3  munsell_0.5.0
## [46] reprex_2.0.2    compiler_4.2.2   rlang_1.0.6
## [49] grid_4.2.2      ggridges_0.5.4   rstudioapi_0.14
## [52] mosaicCore_0.9.2.1 labeling_0.4.2    rmarkdown_2.19
## [55] gtable_0.3.1    DBI_1.1.3        R6_2.5.1
## [58] lubridate_1.9.0 fastmap_1.1.0     utf8_1.2.2
## [61] stringi_1.7.8   Rcpp_1.0.9        vctrs_0.5.1
## [64] dbplyr_2.2.1    tidyselect_1.2.0 xfun_0.36
```