DAAD RNA-seq course- lesson4

Serhiy Naumenko

2024-08-08

Contents

1 Overview

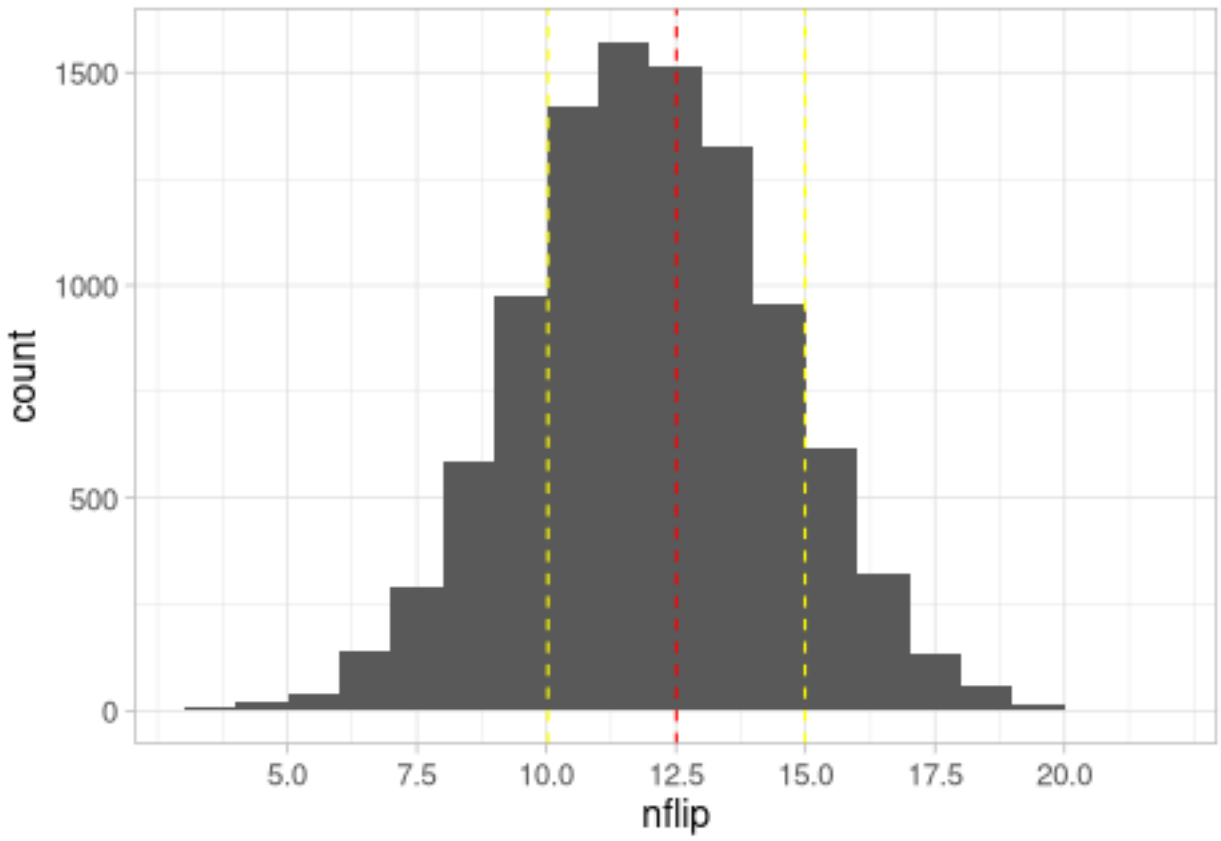
- P-values and multiple testing correction
- https://bookdown.org/jgscott/DSGI/p-values.html
- https://uk.wikipedia.org/wiki/Метод_Монте-Карло
- https://uk.wikipedia.org/wiki/Біноміальний_розподіл

2 Біноміальний розподіл

```
# heads/tails    -
p <- 0.5
n <- 25
bootstrap_n <- 10000

binomial_sim <- mosaic::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$nflip), linetype = "dashed", col
    geom_vline(xintercept = mean(binomial_sim$nflip) - sd(binomial_sim$nflip), linetype = "dashed", col
    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
```

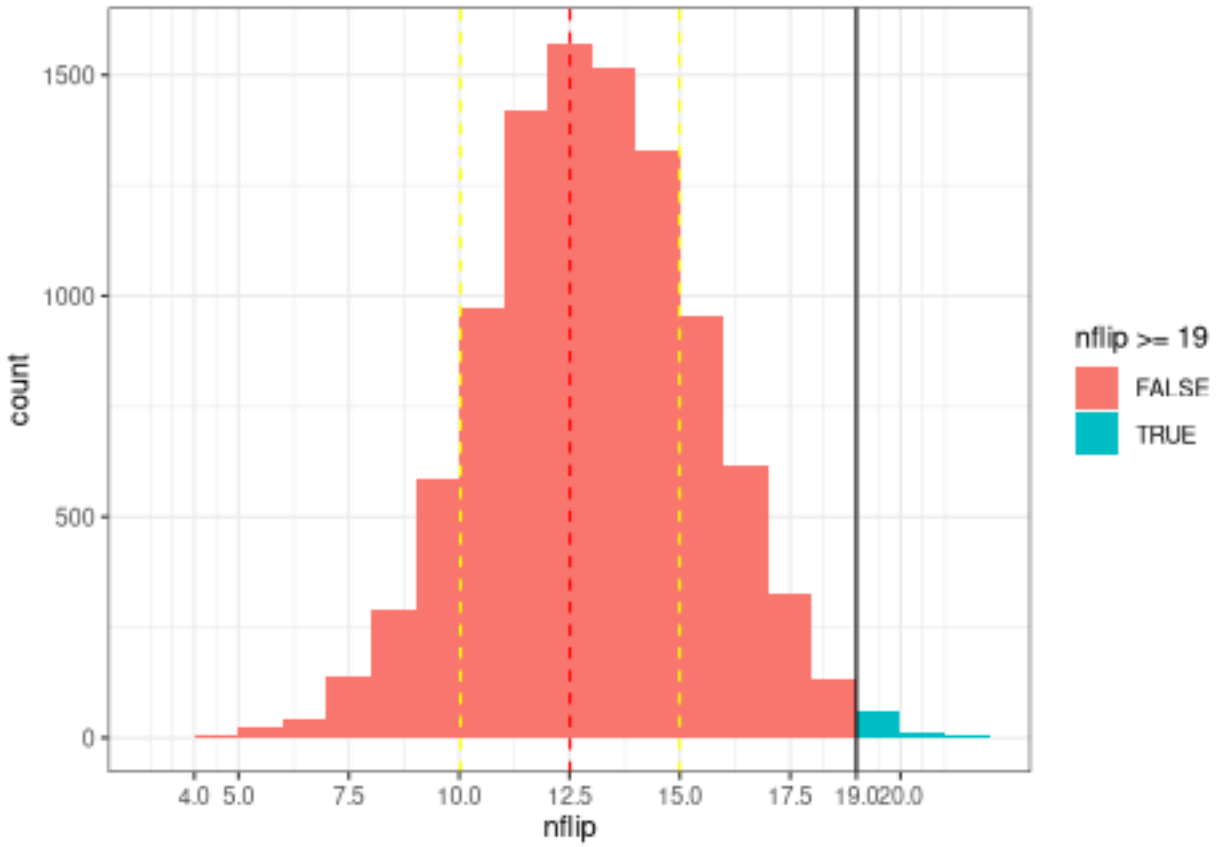
```
## [1] 0.008
```

```
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 P-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$nflip), linetype = "dashed", col
    geom_vline(xintercept = mean(binomial_sim$nflip) - sd(binomial_sim$nflip), linetype = "dashed", col
    geom_vline(xintercept = 19, linetype = "solid", color = "black") +
    scale_x_continuous(breaks = c(4, 5, 7.5, 10, 12.5, 15, 17.5, 19, 20),
                       limits = c(min(binomial_sim$nflip), max(binomial_sim$nflip))) + theme_bw()
```

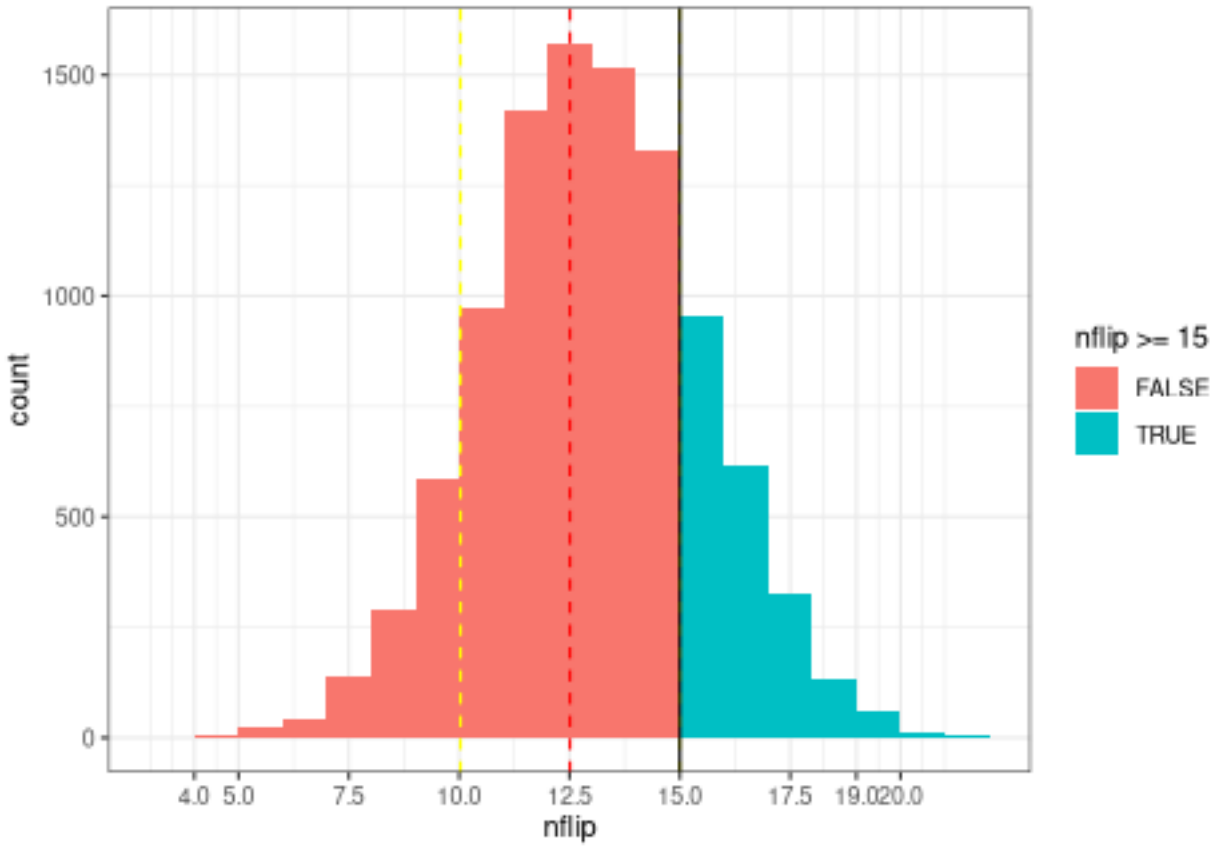



```
#scale_y_log10()
```

- H0: монетка випадкова (p = 0.5) (нема біологічного сигнала)
- тестова статистика: скільки успіхів (heads) із 25 експериментів
- порахували розподіл ймовірності за методом Монте-Карло на 10,000 повторів
- оцінили, наскільки ймовірно, що H0 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$nflip), linetype = "dashed", col
    geom_vline(xintercept = mean(binomial_sim$nflip) - sd(binomial_sim$nflip), linetype = "dashed", col
    geom_vline(xintercept = 15, linetype = "solid", color = "black") +
    scale_x_continuous(breaks = c(4, 5, 7.5, 10, 12.5, 15, 17.5, 19, 20),
                       limits = c(min(binomial_sim$nflip), max(binomial_sim$nflip))) + theme_bw()
```



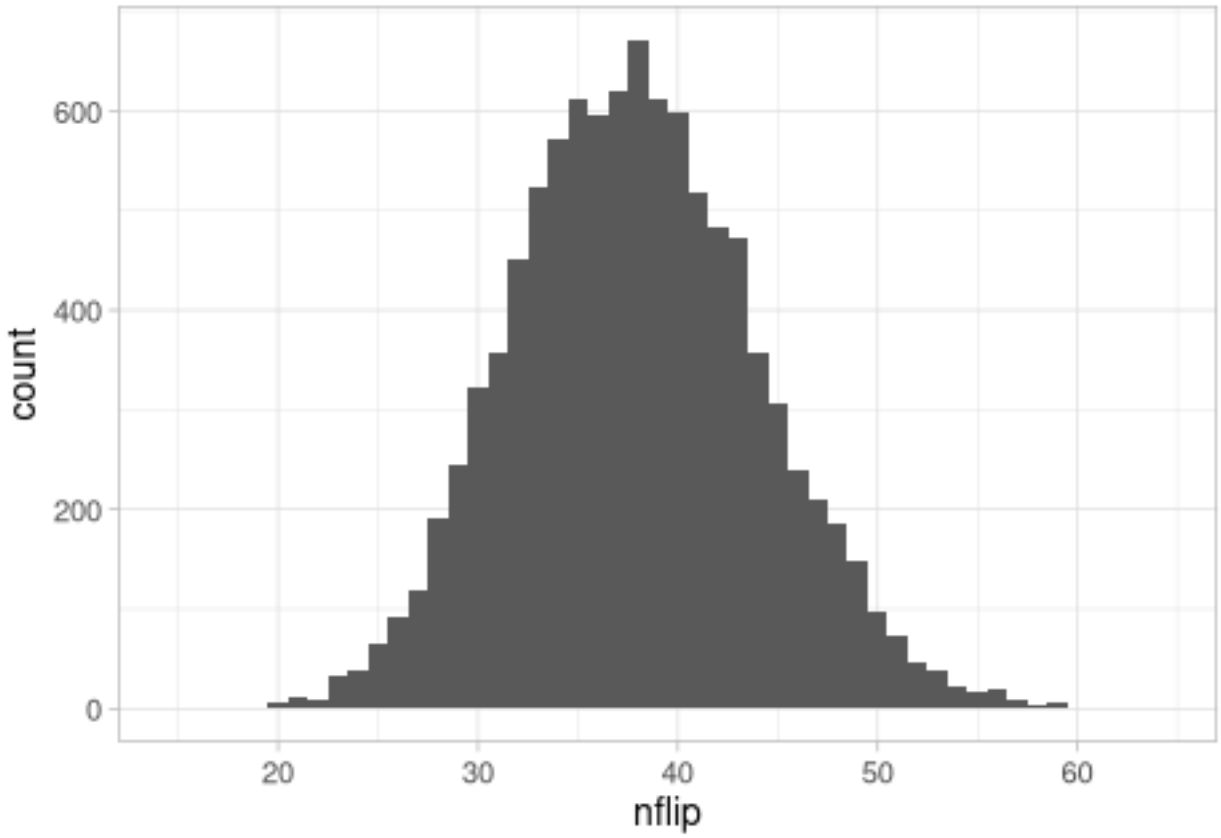

```
rare_event <- 15
sum(binomial_sim >= rare_event) / bootstrap_n
```

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

6 Кластер хвороби

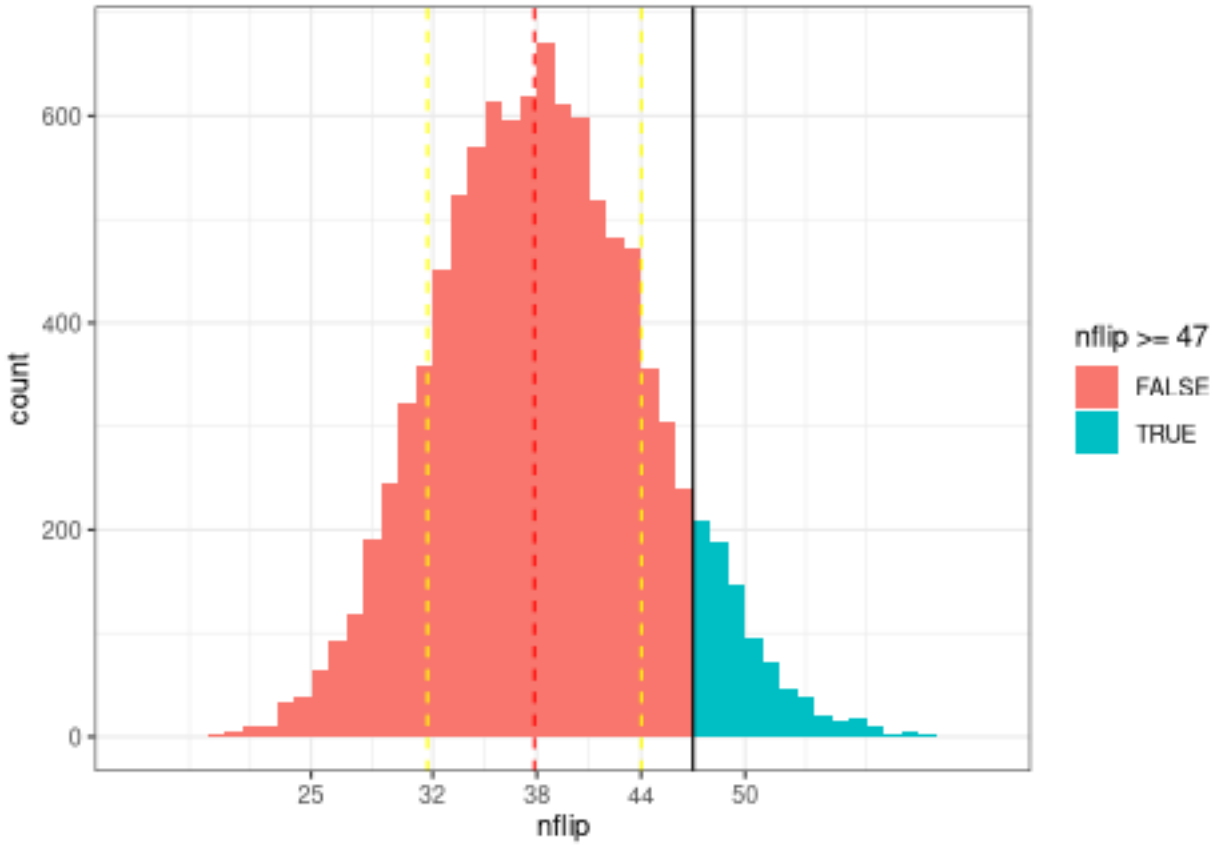
- < 10 км від атомної станції: 5.8 випадків на 10,000: 47 / 80,515
- > 30 км від атомної станції: 4.7 випадків на 10,000: 851 / 1,819,636
- incidence ratio: 5.8/4.7 = 1.23
- H0: IR = 4.7
- тестова статистика: кількість захворювань %
- розподіл тестової статистики за припущення H0

```
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

```r
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$nflip), linetype = "dashed", color =
    geom_vline(xintercept = mean(sim_cancer$nflip) - sd(sim_cancer$nflip), linetype = "dashed", color =
    geom_vline(xintercept = 47, linetype = "solid", color = "black") +
    scale_x_continuous(breaks = c(25, 32, 38, 44, 50),
                       limits = c(min(sim_cancer$nflip), max(sim_cancer$nflip))) + theme_bw()
```



7 sessionInfo()

```
sessionInfo()
```

```
## R version 4.4.1 (2024-06-14)
## Platform: x86_64-redhat-linux-gnu
## Running under: Fedora Linux 40 (Workstation Edition)
##
## Matrix products: default
## BLAS/LAPACK: FlexiBLAS OPENBLAS-OPENMP;  LAPACK version 3.11.0
##
## 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
##
## time zone: America/Toronto
## tzcode source: system (glibc)
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] knitr_1.48        mosaic_1.9.1      mosaicData_0.20.4 ggformula_0.12.0
##  [5] Matrix_1.7-0      lattice_0.22-6    lubridate_1.9.3   forcats_1.0.0
##  [9] stringr_1.5.1     dplyr_1.1.4       purrr_1.0.2       readr_2.1.5
## [13] tidyr_1.3.1       tibble_3.2.1      ggplot2_3.5.1     tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] utf8_1.2.4         generics_0.1.3     stringi_1.8.4      hms_1.1.3
##  [5] digest_0.6.36      magrittr_2.0.3     evaluate_0.24.0    grid_4.4.1
##  [9] timechange_0.3.0   fastmap_1.2.0      fansi_1.0.6        scales_1.3.0
## [13] cli_3.6.3          labelled_2.13.0    rlang_1.1.4        munsell_0.5.1
## [17] withr_3.0.0        yaml_2.3.9         tools_4.4.1        tzdb_0.4.0
## [21] colorspace_2.1-0   mosaicCore_0.9.4.0 vctrs_0.6.5        R6_2.5.1
## [25] ggridges_0.5.6     lifecycle_1.0.4    MASS_7.3-60.2      pkgconfig_2.0.3
## [29] pillar_1.9.0       gtable_0.3.5       glue_1.7.0         highr_0.11
## [33] haven_2.5.4        xfun_0.45          tidyselect_1.2.1   rstudioapi_0.16.0
## [37] farver_2.1.2       htmltools_0.5.8.1  labeling_0.4.3     rmarkdown_2.27
## [41] compiler_4.4.1
```