lab3-ckp4

Fernando Lordao

28 de maio de 2018

```
running_mode = "finalWeb"
#running_mode = "reducedLocal"
#running_mode = "finalLocal"

if(running_mode == "finalWeb") {
    path_to_read = "https://github.com/wikimedia-research/Discovery-Hiring-Analyst-2016/raw/master/event
    lines_to_read = -1
} else if(running_mode == "finalLocal") {
    path_to_read = "data/events_log.csv"
    lines_to_read = -1
} else if(running_mode == "reducedLocal") {
    path_to_read = "data/events_log.csv"
    lines_to_read = 5000
} else {
    message("Running mode not specified.")
    break() #encontrar a funÃ§Ã£o certa para finalizar o script.
}

set.seed(20180528)
```

```
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 2.2.1     v purrr   0.2.4
## v tibble  1.4.2     v dplyr   0.7.4
## v tidyr   0.8.0     v stringr 1.3.0
## v readr   1.1.1     v forcats 0.3.0
```

```
## -- Conflicts -------------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
library(chron)
```

```
##
## Attaching package: 'chron'
```

```
## The following objects are masked from 'package:lubridate':
##
##     days, hours, minutes, seconds, years
```

```
library(resample)
theme_set(theme_bw())
```

Descrição geral

Neste documento vamos explorar os dados disponibilizados pela Wikimedia e tentar responder às questões levantadas fazendo inferências sobre a população através de *testes de hipóteses*.

Para esse estudo iremos revisitar as questões 1 e 3 ajustando-as para o cenário de testes de hipóteses onde compararemos o comportamento de dois grupos.

Carga do log de eventos

```
events = read_csv(path_to_read) %>%
    head(lines_to_read)
```

```
## Parsed with column specification:
## cols(
##   uuid = col_character(),
##   timestamp = col_double(),
##   session_id = col_character(),
##   group = col_character(),
##   action = col_character(),
##   checkin = col_integer(),
##   page_id = col_character(),
##   n_results = col_integer(),
##   result_position = col_integer()
## )
```

```
events = events %>%
    mutate(
        date = round_date(ymd_hms(timestamp), unit = "day"),
        date_week = paste(date, weekdays(date, abbreviate = TRUE))
    )
```

PERGUNTA 1:

Existe diferença significativa entre a taxa geral de cliques dos grupos?

Conforme descrição do problema pela Wikimedia, a taxa de cliques é definida da seguinte maneira:

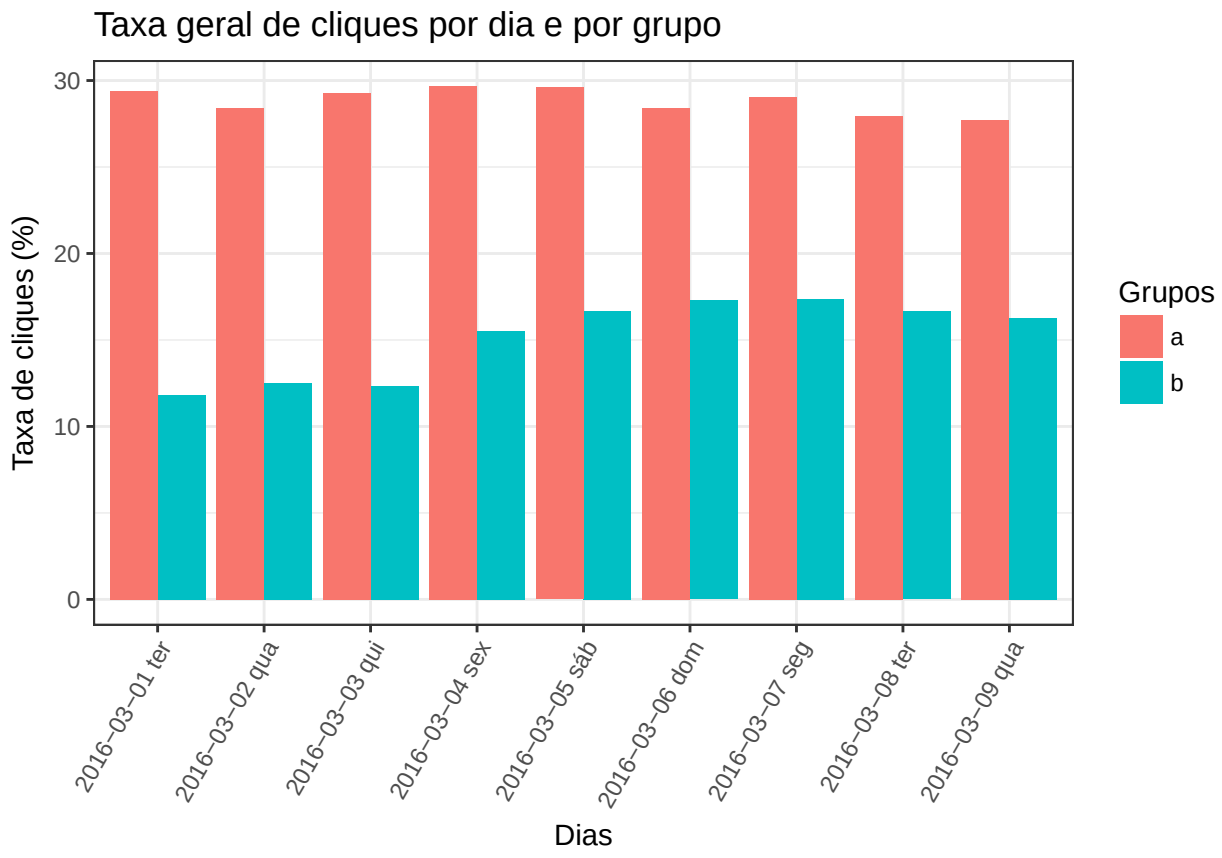
clickthrough rate: the proportion of search sessions where the user clicked on one of the results displayed

Para isso é preciso criar novas variáveis que permitam identificar as buscas e também as visitas dentre de cada busca. Faremos isso usando o código a seguir.

```
events = events %>%
    group_by(session_id) %>%
    arrange(timestamp) %>%
    mutate(
        search_index = cumsum(action == "searchResultPage") # Sequenciador de buscas realizadas dentro
    )
```

```r
events = events %>%
    group_by(session_id, search_index) %>%
    arrange(timestamp) %>%
    mutate(
        visit_index = cumsum(action == "visitPage") # Sequenciador de visitas realizadas dentro de cada
    )
```

A seguir visualizamos a *clickthrough rate* ao longo dos dias para cada grupo, para ter uma noção de como os dados se comportam.



Analisando a diferença da média de taxa de cliques entre o grupo A e o grupo B

Agora sim, de fato, faremos uma comparação entre os grupos usando teste de hipóteses através de **permutações aleatórias dos grupos A e B**.

```r
events_q1dif = events %>%
    group_by(group, session_id, search_index) %>%
    mutate(click_count = max(visit_index)) %>%
    subset(action == "searchResultPage", select = c("group", "click_count"))

permutationTest2(events_q1dif, mean((sum(click_count>0) / sum(click_count>=0)) * 100), treatment = grou
```

```
## Call:
## permutationTest2(data = events_q1dif, statistic = mean((sum(click_count >
##     0)/sum(click_count >= 0)) * 100), treatment = group)
```

```
## Replications: 9999
## Two samples, sample sizes are 92056 44178
##
## Summary Statistics for the difference between samples 1 and 2:
##                                                                    Observed
## mean((sum(click_count > 0)/sum(click_count >= 0)) * 100): a-b  13.6344
##                                                                          Mean
## mean((sum(click_count > 0)/sum(click_count >= 0)) * 100): a-b 9.936611e-05
##                                                                    Alternative
## mean((sum(click_count > 0)/sum(click_count >= 0)) * 100): a-b   two.sided
##                                                                    PValue
## mean((sum(click_count > 0)/sum(click_count >= 0)) * 100): a-b  2e-04
```

Aplicando o teste de permutações observa-se que há uma diferença significativa em torno de 13,63 pontos percentuais quando comparamos a média taxa de cliques do grupo A contra a média do grupo B. Isso se confirma observando o p-value que resultou em 0,0002 e representa, portanto, uma possibilidade muito baixa (0,02%) de que a diferença das médias entre o grupo A e o grupo B seja nula.

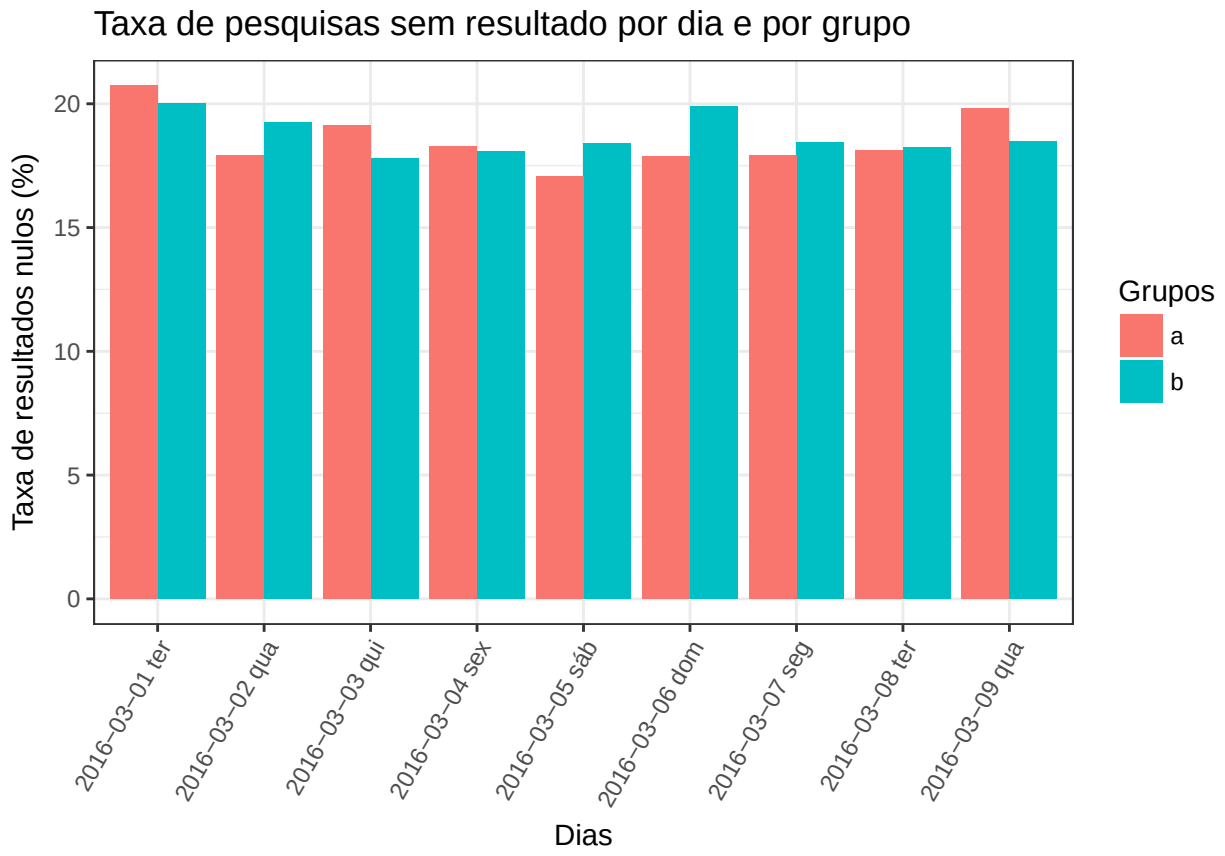
PERGUNTA 3:

Existe diferença significativa entre a taxa geral de pesquisas sem resultados dos grupos?

Para responder esta questão precisamos contar a quantidade eventos do tipo "searchResultPage" que possuem a coluna *n_result* zerada e dividir esse resultado pela quantidade total de eventos desse mesmo tipo.

O gráfico que segue é bem semelhante ao gráfico plotado na questão 1, só muda a informação que estamos medindo, que aqui é a taxa de resultados nulos, por assim dizer.

De fato, este gráfico é apenas para termos uma noção geral do comportamento dos dados. A análise por teste de hipóteses será realizada no tópico seguinte.



Analisando a diferença da média de taxa de pesquisas sem resultado entre o grupo A e o grupo B

Agora sim, segue o código para realizar a análise da diferença entre os grupos utilizando teste de hipóteses por **permutações aleatórias dos grupos A e B**.

```
events_q3dif = events %>%
    group_by(group, session_id, search_index) %>%
    subset(action == "searchResultPage", select = c("group", "n_results"))

permutationTest2(events_q3dif, statistic = mean((sum(n_results==0) / sum(n_results>=0)) * 100), treatme
```

```
## Call:
## permutationTest2(data = events_q3dif, statistic = mean((sum(n_results ==
##     0)/sum(n_results >= 0)) * 100), treatment = group)
## Replications: 9999
## Two samples, sample sizes are 92056 44178
##
## Summary Statistics for the difference between samples 1 and 2:
##                                                          Observed
## mean((sum(n_results == 0)/sum(n_results >= 0)) * 100): a-b -0.257301
##                                                               Mean
## mean((sum(n_results == 0)/sum(n_results >= 0)) * 100): a-b 0.0003817068
##                                                         Alternative
## mean((sum(n_results == 0)/sum(n_results >= 0)) * 100): a-b   two.sided
```

```
##                                                          PValue
## mean((sum(n_results == 0)/sum(n_results >= 0)) * 100): a-b 0.1248
```

Aplicando o teste de permutações observa-se que **NÃO HÁ** uma diferença significativa quando comparamos a média da taxa de pesquisas sem resultados no grupo A contra a média no grupo B. Isso se confirma observando o p-value que resultou em 0,1248 e representa, portanto, possibilidade significativa (12,48%) de que a diferença das médias entre o grupo A e o grupo B seja nula.

Comentários: Teste de Hipóteses vs Intervalos de Confiança

Após a compreensão e aplicação do teste de hipóteses (TH), tive a impressão que esta abordagem tem aplicação mais restrita que os intervalos de confiança (IC). Isso se deve ao fato de que para aplicar TH precisamos sempre comparar grupos distintos de amostras e normalmente fazemos isso para confirmar ou descartar associação entre tais grupos.

Já os ICs permitem realizar inferências mais diversificadas sobre a população, por exemplos estimar o valor de determinada medida para um único grupo, ou até mesmo quando não existe alguma seguimentação evidente.

Além disso os ICs nos trazem mais informações além simplesmente do percentual de significância existente nos THs.