Existe diferença significante de atrasos de voo entre duas companhia aéreas?

Rodolfo R. Terra | Cientista de Dados

Sumário

1.Coletando os Dados
2.Exploração de Dados
3.Construindo o Dataset
4.Criação de Amostragem
5.Intervalo de Confiança
6.Gráfico dos Intervalos de Confiança
7.Criação do Teste de Hipótese
7.1.Teste t
7.2.Valor p
8.Conclusão
Definição do Problema de Negócio

Para analisar se existe diferença de atraso de vôo entre a companhia utilizaremos teste de hipósete de um conjunto de dados que possui 336,776 observações e 19 variáveis, chamado flights", que demonstram dados pontuais de todos os voos que partiram de Nova York em 2013. Dentre estas companhia escolheremos duas companhias: Delta Airlines (DL) e a United Airlines (UA).

Etapa 1 - Coletando os Dados

Aqui está a coleta de dados.

```
# Coletando dados
library('ggplot2')
library('dplyr')

##
## Attaching package: 'dplyr'

##   The   following   objects   are   masked   from   'package:stats':
##
##     filter, lag
```

```
##    The    following    objects    are    masked    from    'package:base':
##
##     intersect, setdiff, setequal, union

library('nycflights13')
head(flights)

##          #          A          tibble:          6          x          19
##       year  month     day  dep_time  sched_dep_time  dep_delay  arr_time
sched_arr_time
##    <int> <int> <int>     <int>             <int>      <dbl>     <int>
<int>
## 1   2013      1      1       517               515          2       830
819
## 2   2013      1      1       533               529          4       850
830
## 3   2013      1      1       542               540          2       923
850
## 4   2013      1      1       544               545         -1      1004
1022
## 5   2013      1      1       554               600         -6       812
837
## 6   2013      1      1       554               558         -4       740
728
## # ... with 11 more variables: arr_delay <dbl>, carrier <chr>, flight
<int>,
## #   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance
<dbl>,
## #   hour <dbl>, minute <dbl>, time_hour <dttm>
```

Etapa 2 - Exploração dos Dados

```
head(flights)

##          #          A          tibble:          6          x          19
##       year  month     day  dep_time  sched_dep_time  dep_delay  arr_time
sched_arr_time
##    <int> <int> <int>     <int>             <int>      <dbl>     <int>
<int>
## 1   2013      1      1       517               515          2       830
819
## 2   2013      1      1       533               529          4       850
830
## 3   2013      1      1       542               540          2       923
850
## 4   2013      1      1       544               545         -1      1004
1022
## 5   2013      1      1       554               600         -6       812
```

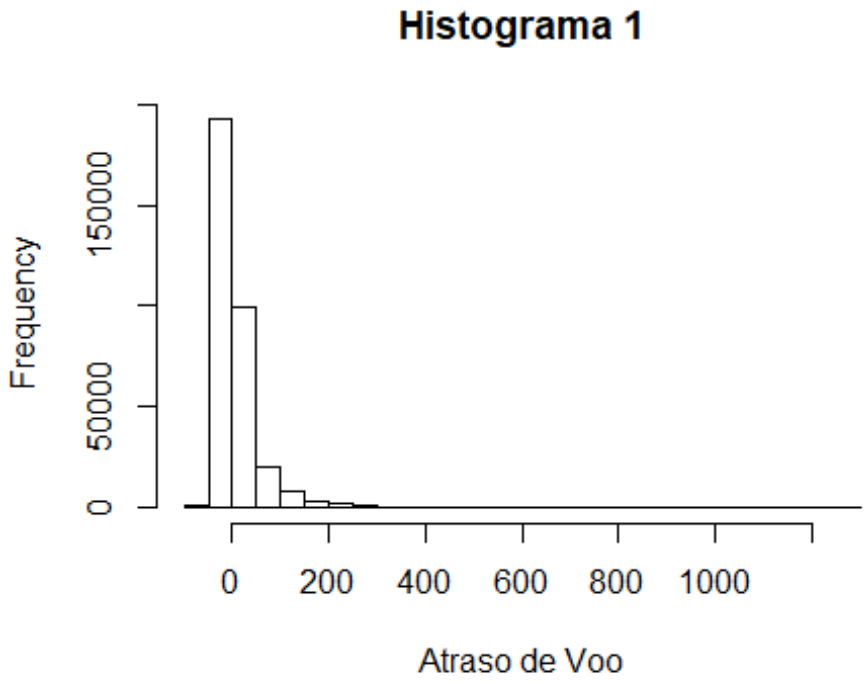
```
837
## 6  2013     1     1       554             558         -4       740
728
## # ... with 11 more variables: arr_delay <dbl>, carrier <chr>, flight
<int>,
## #   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance
<dbl>,
## #   hour <dbl>, minute <dbl>, time_hour <dttm>

#                Visualizando                as                variáveis
str(flights)

## Classes 'tbl_df', 'tbl' and 'data.frame':      336776 obs. of   19
variables:
##  $ year           : int  2013 2013 2013 2013 2013 2013 2013 2013 2013
2013                                                                  ...
##   $ month                 :  int    1  1  1  1  1  1  1  1  1  1  ...
##   $ day                    :  int    1  1  1  1  1  1  1  1  1  1  ...
##  $ dep_time       : int  517 533 542 544 554 554 555 557 557 558 ...
##  $ sched_dep_time: int  515 529 540 545 600 558 600 600 600 600 ...
##   $ dep_delay         :  num    2  4  2  -1  -6  -4  -5  -3  -3  -2  ...
##  $ arr_time       : int  830 850 923 1004 812 740 913 709 838 753 ...
##  $ sched_arr_time: int  819 830 850 1022 837 728 854 723 846 745 ...
##   $ arr_delay        :  num   11  20  33  -18  -25  12  19  -14  -8  8  ...
##    $  carrier                :  chr    "UA"  "UA"  "AA"  "B6"  ...
##  $ flight         : int  1545 1714 1141 725 461 1696 507 5708 79 301
...
##   $ tailnum          :  chr   "N14228"  "N24211"  "N619AA"  "N804JB"  ...
##    $  origin                :  chr    "EWR"  "LGA"  "JFK"  "JFK"  ...
##    $  dest                   :  chr    "IAH"  "IAH"  "MIA"  "BQN"  ...
##  $ air_time       : num  227 227 160 183 116 150 158 53 140 138 ...
##    $  distance             :  num    1400  1416  1089  1576  762  ...
##    $  hour                  :  num    5  5  5  5  6  5  6  6  6  6  ...
##    $  minute              :  num    15  29  40  45  0  58  0  0  0  0  ...
##  $ time_hour     : POSIXct, format: "2013-01-01 05:00:00" "2013-01-01
05:00:00" ...
```

Medidas de Tendência Central da Variável Numéricas

```
summary(flights$arr_delay)

##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
##  -86.000  -17.000   -5.000    6.895   14.000 1272.000     9430

#      Tabela      de      contigêcia      das      linhas      aéreas
table(flights$carrier)

##
##    9E    AA    AS    B6    DL    EV    F9    FL    HA    MQ    OO
UA                                                                  US
```

```
## 18460 32729   714 54635 48110 54173   685  3260   342 26397    32
58665                                                            20536
##                    VX                      WN                    YV
##  5162 12275   601
```

```
#Histograma
hist(flights$arr_delay, main = 'Histograma 1', xlab = 'Atraso de Voo')
```



Etapa 3 - Construindo o Dataset

Construção do dataset pop_data com os dados de voos das companhias aéreas UA (United Airlines) e DL (Delta Airlines). Contruiremos um dataset irá conter apenas duas colunas, nome da companhia e atraso nos voos de chegada. Será conssiderado este dataset como sendo nossa população de voos:

Metodologia da formula:

- 1º eliminação de todos os dados na (Valores missing, não disponivel);
- 2º Filtro pela companhia aérea UA (United Airlines) e DL (Delta Airlines);
- 3º Filtro para retornar apenas os valores que forem igual ou maior a 'Zero'. Os valores negativos serão descontiderádos;

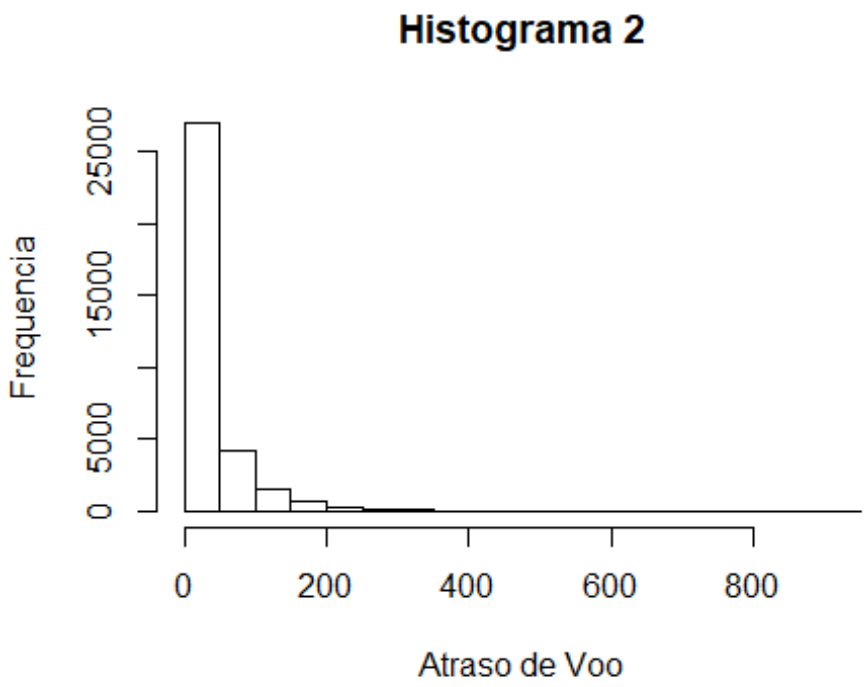
```r
pop_data = na.omit(flights) %>%
  filter(carrier == 'UA' | carrier == 'DL', arr_delay >=0) %>%
  select(carrier, arr_delay) %>%
  group_by(carrier) %>%
  sample_n(17000) %>%
  ungroup()


head(pop_data)
```

```
## # A tibble: 6 x 2
##   carrier arr_delay
##   <chr>       <dbl>
## 1 DL             34
## 2 DL             56
## 3 DL              9
## 4 DL              9
## 5 DL            374
## 6 DL              4
```

```r
tail(pop_data)
```

```
## # A tibble: 6 x 2
##   carrier arr_delay
##   <chr>       <dbl>
## 1 UA            159
## 2 UA             10
## 3 UA            102
## 4 UA             24
## 5 UA              9
## 6 UA             15
```

```r
hist(pop_data$arr_delay, main = 'Histograma 2', xlab = 'Atraso de Voo',
ylab="Frequencia")
```



É importante observar que neste histograma removemos os valores negativos, pois temos como principal objetivo observar somente os atrasos de voos. Os valores no conjunto de dados que representam menores que zero seguinifica que o voo chegou a sei destido antes da data prevista, então não houve atraso.

Etapa 4 - Criação de Amostragem

Criação de duas amostras de 1000 observações cada uma a partir do dataset pop_data apenas com dados da companhia DL para amostra 1 e apenas dados da companhia UA na amostra 2.

```
amostra1 <- na.omit(pop_data) %>%
  select(carrier, arr_delay) %>%
  filter(carrier == 'DL') %>%
  mutate(sample_id = '1') %>%
  sample_n(1000)

head(amostra1)
```

```
## # A tibble: 6 x 3
##   carrier arr_delay sample_id
##   <chr>       <dbl> <chr>
## 1 DL             29 1
## 2 DL             92 1
## 3 DL              8 1
```

```
## 4 DL              4 1
## 5 DL              0 1
## 6 DL             14 1
```

```
amostra2 <- na.omit(pop_data) %>%
  select(carrier, arr_delay) %>%
  filter(carrier == "UA") %>%
  mutate(sample_id = "2") %>%
  sample_n(1000)
```

```
head(amostra2)
```

```
## # A tibble: 6 x 3
##   carrier arr_delay sample_id
##   <chr>       <dbl> <chr>
## 1 UA              8 2
## 2 UA              3 2
## 3 UA              7 2
## 4 UA              1 2
## 5 UA             84 2
## 6 UA             51 2
```

```
# Criação de um dataset contendo os dados das 2 amostras criadas no item
anterior.
```

```
samples = rbind(amostra1, amostra2)
head(samples)
```

```
## # A tibble: 6 x 3
##   carrier arr_delay sample_id
##   <chr>       <dbl> <chr>
## 1 DL             29 1
## 2 DL             92 1
## 3 DL              8 1
## 4 DL              4 1
## 5 DL              0 1
## 6 DL             14 1
```

```
tail(samples)
```

```
## # A tibble: 6 x 3
##   carrier arr_delay sample_id
##   <chr>       <dbl> <chr>
## 1 UA              2 2
## 2 UA              8 2
## 3 UA             60 2
## 4 UA            113 2
## 5 UA             57 2
## 6 UA             80 2
```

Etapa 5 - Intervalo de Confiança

Calculo do intervalo de confiança (95%) da amostra1.

Fórmula de erro padrão: erro_padrao = sd(amostra$arr_delay) / sqrt(nrow(amostra))

Esta fórmula é usada para calcular o desvio padrão de uma distribuição da média amostral (de um grande número de amostras de uma população). Em outras palavras, só é aplicável quando você está procurando o desvio padrão de médias calculadas a partir de uma amostra de tamanho nn, tirada de uma população.

Digamos que você obtenha 10000 amostras de uma população qualquer com um tamanho de amostra de n = 2. Então calculamos as médias de cada uma dessas amostras (teremos 10000 médias calculadas). A equação acima informa que, com um número de amostras grande o suficiente, o desvio padrão das médias da amostra pode ser aproximado usando esta fórmula: sd(amostra) / sqrt(nrow(amostra))

Deve ser intuitivo que o seu desvio padrão das médias da amostra será muito pequeno, ou em outras palavras, as médias de cada amostra terão muito pouca variação.

Com determinadas condições de inferência (nossa amostra é aleatória, normal, independente), podemos realmente usar esse cálculo de desvio padrão para estimar o desvio padrão de nossa população. Como isso é apenas uma estimativa, é chamado de erro padrão. A condição para usar isso como uma estimativa é que o tamanho da amostra n é maior que 30 (dado pelo teorema do limite central) e atende a condição de independência n <= 10% do tamanho da população.

amostra 1

Erro Padrão

```
erro_padrao_amostra1 = sd(amostra1$arr_delay) / sqrt(nrow(amostra1))
```

Limites Inferior e Superior

1.96 é o valor de z score para 95% de confiança

```
lower1 = mean(amostra1$arr_delay) - 1.96 * erro_padrao_amostra1
upper1 = mean(amostra1$arr_delay) + 1.96 * erro_padrao_amostra1
```

Intervalo de Confiança

```
ic_1 = c(lower1, upper1)
mean(amostra1$arr_delay)

## [1] 38.57

ic_1

## [1] 34.69659 42.44341
```

amostra 2

Erro Padrão

```
erro_padrao_amostra2 = sd(amostra1$arr_delay) / sqrt(nrow(amostra2))
```

Limites Inferior e Superior

1.96 é o valor de z score para 95% de confiança

```
lower2 = mean(amostra2$arr_delay) - 1.96 * erro_padrao_amostra2
upper2 = mean(amostra2$arr_delay) + 1.96 * erro_padrao_amostra2
```

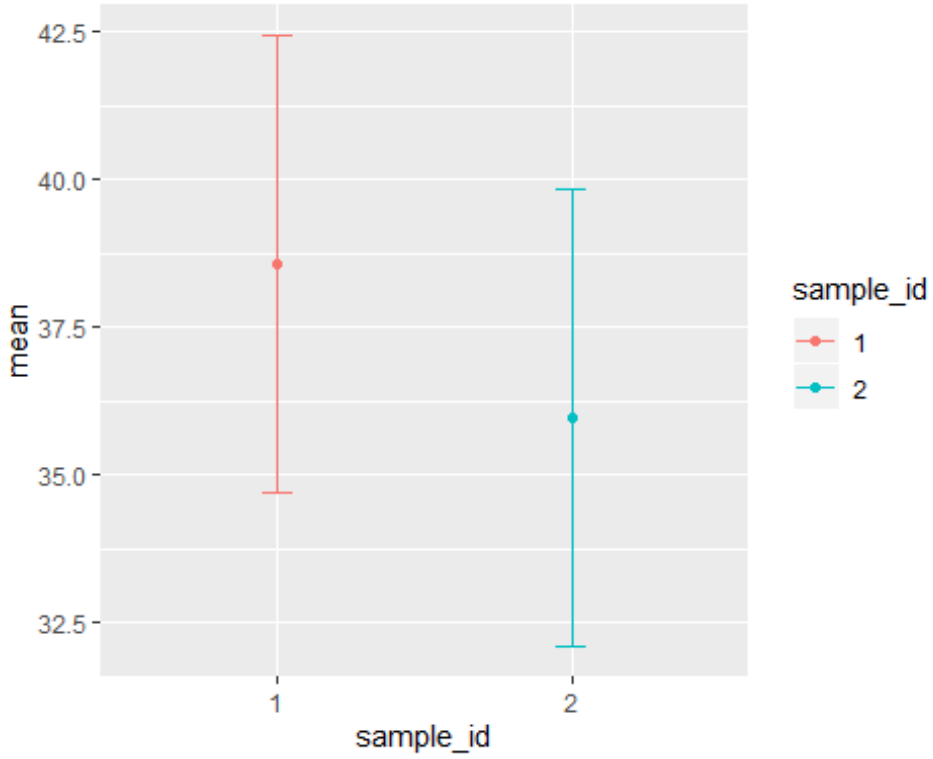
Intervalo de Confiança

```
ic_2 = c(lower2, upper2)
mean(amostra1$arr_delay)

## [1] 38.57

ic_2

## [1] 32.10559 39.85241
```

Etapa 6 - Gráfico dos Intervalos de Confianças

Criação de um plot Visualizando os intervalos de confiança criados nos itens anteriores.

```
toPlot = summarise(group_by(samples, sample_id), mean = mean(arr_delay))
toPlot = mutate(toPlot, lower = ifelse(toPlot$sample_id ==
1,ic_1[1],ic_2[1]))
toPlot = mutate(toPlot, upper = ifelse(toPlot$sample_id ==
1,ic_1[2],ic_2[2]))
ggplot(toPlot, aes(x = sample_id, y=mean, colour = sample_id)) +
  geom_point() +
  geom_errorbar(aes(ymin=lower, ymax=upper), width=.1)
```



A maior parte dos dados reside no mesmo intervalo de confiança nas duas amostras. Motimo pelo qual podemos dizer que muito provavelmente, as amostras vieram da mesma população.

Etapa 7 - Criação do Teste de Hipótese

Crição de um teste de hipótese para verificar se os voos da Delta Airlines (DL) atrasam mais do que os voos da UA (United Airlines).

H0 e H1 devem ser mutuamente exclusivas.

- H0 = Não há diferença significativa entre os atrasos da DL e UA (diff da média de atrasos = 0).
- H1 = Delta atrasa mais (diff das médias > 0).

Criação das Amostras

```
dl <- sample_n(filter(pop_data, carrier == "DL", arr_delay > 0), 1000)
ua <- sample_n(filter(pop_data, carrier == "UA", arr_delay > 0), 1000)
```

Calculo do Erro Padrão Médio

Amostra 1

```
se1 = sd(dl$arr_delay) / sqrt(nrow(dl))
mean(dl$arr_delay)

## [1] 37.459
```

Limites Inferior e Superior

```
lower11 = mean(dl$arr_delay) - 1.96 * se1
upper11 = mean(dl$arr_delay) + 1.96 * se1
ic_dl = c(lower11,upper11)
ic_dl

## [1] 33.98665 40.93135
```

Calcula erro padrão e média Amostra 2

```
se2 = sd(ua$arr_delay) / sqrt(nrow(ua))
mean(ua$arr_delay)

## [1] 37.051
```

Limites inferior e superior

```
lower22 = mean(ua$arr_delay) - 1.96 * se2
upper22 = mean(ua$arr_delay) + 1.96 * se2
ic_ua = c(lower22,upper22)
ic_ua

## [1] 34.13043 39.97157
```

Etapa 7.1 - Teste T

O teste t (de Student) foi desenvolvido por Willian Sealy Gosset em 1908 que usou o pseudônimo "Student" em função da confidencialidade requerida por seu empregador (cervejaria Guiness) que considerava o uso de estatística na manutenção da qualidade como uma vantagem competitiva. O teste t de Student tem diversas variações de aplicação, e pode ser usado na comparação de duas (e somente duas) médias e as variações dizem respeito às hipóteses que são testadas.

```
t.test(dl$arr_delay, ua$arr_delay, alternative="greater")

##
##  Welch Two Sample t-test
##
## data:  dl$arr_delay and ua$arr_delay
## t = 0.17625, df = 1941, p-value = 0.4301
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -3.40156      Inf
```

```
## sample estimates:
## mean of x mean of y
##    37.459    37.051
```

Etapa 7.2 - Valor p

O valor-p é uma quantificação da probabilidade de se errar ao rejeitar H0 e a mesma decorre da distribuição estatística adotada. Se o valor-p é menor que o nível de significância, conclui-se que o correto é rejeitar a hipótese de nulidade.

Valor p é a probabiblidade de que a estatística do teste assuma um valor extremo em relação ao valor observado quando H0 é verdadeira.

Estamos trabalhando com alfa igual a 0.05 (95% de confiança)

Regra

- Baixo valor p: forte evidência empírica contra h0
- Alto valor p: pouca ou nenhuma evidência empírica contra h0

Etapa 8 - Conclusão

- Falhamos em rejeitar a hipótese nula, pois p-valor é maior que o nível de significância.
- Isso que dizer que há uma probabilidade alta de não haver diferença significativa entre os atrasos.
- Para os nossos dados, não há evidência estatística de que a DL atrase mais que a UA.

Dados Pessoais

Site www.rodolfoterra.com
Linkedin rodolffoterra
Repertório no GitHub: Teste de Hipotese
E-mail consultoriaterra@hotmail.com