STAT346: Statistical Data Science I

Final: Thursday, Dec 16 2021, 05:00–06:15 p.m.

Yoon Minseo / 2021320322

0

```
rm(list = ls())
ls()
```

```
## character(0)
```

1

```
library(tidyverse)
library(dslabs)
data(heights)
```

(a)

```
mean_sd <- heights %>%
  group_by(sex) %>%
  summarize(sample_mean = mean(height), sample_sd = sd(height))

mean_sd
```

```
## # A tibble: 2 x 3
##   sex    sample_mean sample_sd
##   <fct>        <dbl>     <dbl>
## 1 Female        64.9      3.76
## 2 Male          69.3      3.61
```

(b)

```r
x = heights %>% filter(sex == 'Female') %>% pull(height)
F <- function(a) mean(x <= a)
1 - F(64.5)
```

```
## [1] 0.5294118
```

c

```r
B <- 1000
set.seed(100)
tallest <- replicate(B, {
  simulated_data <- rnorm(500, mean_sd$sample_mean, mean_sd$sample_sd)
  max(simulated_data)
})

mean(tallest > 80)
```

```
## [1] 0.289
```

2

```r
library(dslabs)
library(lubridate)
data('polls_us_election_2016')
polls = polls_us_election_2016 %>% filter(enddate >= '2016-10-31' & state == 'U.S.')
```

a

```r
N <- polls$samplesize[2]
x_hat <- polls$rawpoll_trump[2] / 100
se_hat <- sqrt(x_hat * (1 - x_hat) / N)
cat('The 99% confidence interval for the election night proportion is [',
    x_hat + qnorm(0.005) * se_hat, ',', x_hat + qnorm(0.995) * se_hat, ']')
```

```
## The 99% confidence interval for the election night proportion is [ 0.3493299 , 0.3644701 ]
```

b

```
library(dplyr)
trump <- polls %>%
  mutate(N = samplesize, x_hat = rawpoll_trump / 100,
         se_hat = sqrt(x_hat * (1 - x_hat) / N),
         lower = x_hat + qnorm(0.005) * se_hat,
         upper = x_hat + qnorm(0.995) * se_hat) %>%
  select(x_hat, lower, upper, grade)
head(trump)
```

```
##    x_hat     lower     upper grade
## 1 0.4300 0.4029347 0.4570653    A+
## 2 0.3569 0.3493299 0.3644701     B
## 3 0.3900 0.3631838 0.4168162    A-
## 4 0.4100 0.3891076 0.4308924     B
## 5 0.4300 0.4201139 0.4398861    B-
## 6 0.4400 0.4044694 0.4755306     A
```

```
p = 0.461
trump2 = trump %>%
  mutate(hit = case_when(lower <= p & upper >= p ~ "Include", TRUE ~ "Fail")) %>%
  select(lower, upper, hit, grade)
trump2 %>% head()
```

```
##       lower     upper     hit grade
## 1 0.4029347 0.4570653    Fail    A+
## 2 0.3493299 0.3644701    Fail     B
## 3 0.3631838 0.4168162    Fail    A-
## 4 0.3891076 0.4308924    Fail     B
## 5 0.4201139 0.4398861    Fail    B-
## 6 0.4044694 0.4755306 Include     A
```

c

c1

```
trump2 %>%
  summarize(missing_value = sum(is.na(grade)))
```

```
##   missing_value
## 1            18
```

c2

```
trump3 <- trump2 %>%
  filter(!is.na(grade)) %>%
  mutate(grade2 = case_when(
    grade %in% c('A+', 'A', 'A-') ~ 'A',
    grade %in% c('B+', 'B', 'B-') ~ 'B',
    TRUE ~ 'Others'
  ))
head(trump3)
```

```
##       lower     upper     hit grade grade2
## 1 0.4029347 0.4570653    Fail    A+      A
## 2 0.3493299 0.3644701    Fail     B      B
## 3 0.3631838 0.4168162    Fail    A-      A
## 4 0.3891076 0.4308924    Fail     B      B
## 5 0.4201139 0.4398861    Fail    B-      B
## 6 0.4044694 0.4755306 Include     A      A
```

c3

```
# tab <- trump3 %>%
#   summarize(include_A = sum(hit == 'Include' & grade2 == 'A'),
#             fail_A = sum(hit == 'Fail' & grade2 == 'A'),
#             include_B = sum(hit == 'Include' & grade2 == 'B'),
#             fail_B = sum(hit == 'Fail' & grade2 == 'B')) %>%
#   pivot_longer(c('include_A', 'fail_A', 'include_B', 'fail_B'),
#                names_to = 'name', values_to = 'num') %>%
#   separate(name, c('hit', 'grade')) %>%
#   pivot_wider(names_from = grade, values_from = num)
# tab

tab <- table(trump3$hit[trump3$grade2 != 'Others'],
             trump3$grade2[trump3$grade2 != 'Others'])
tab
```

```
##
##             A  B
##   Fail     11  6
##   Include 19  4
```

c4

```
# tab %>%
#   select(-hit) %>%
```

```
#   chisq.test()

chisq.test(tab)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab
## X-squared = 0.85251, df = 1, p-value = 0.3558
```

c5

Answer: 0.3558, larger, do not reject

d

```
polls_us_election_2016 %>%
  summarize(proportion =
              sum(year(startdate) == 2016 & month(startdate) == 9)
            / length(startdate))
```

```
##   proportion
## 1  0.1896388
```

3

```
library(Lahman)
library(broom)
data(Teams)
data(Batting)
data(Salaries)
```

a

```
fit3 = Teams %>%
  filter(yearID %in% 1961:2015) %>%
  mutate(BB = BB/G, singles = (H-X2B-X3B-HR)/G, doubles = X2B/G,
         triples = X3B/G, HR = HR/G, R = R/G) %>%
  lm(R ~ BB + singles + doubles + triples + HR, data = .)
tidy(fit3, conf.int = T)
```

```
## # A tibble: 6 x 7
##   term        estimate std.error statistic   p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>
## 1 (Intercept)   -2.85    0.0709      -40.2 9.62e-238   -2.99     -2.71
## 2 BB             0.368   0.00968      38.0 2.47e-219    0.349     0.387
## 3 singles        0.541   0.0102       52.8 0            0.521     0.561
## 4 doubles        0.727   0.0180       40.3 2.76e-238    0.691     0.762
## 5 triples        1.29    0.0661       19.5 2.38e- 75    1.16      1.42
## 6 HR             1.46    0.0211       69.0 0            1.42      1.50
```

b

```
Teams %>%
  filter(yearID == 2016) %>%
  mutate(BB = BB/G, singles = (H-X2B-X3B-HR)/G, doubles = X2B/G,
         triples = X3B/G, HR = HR/G, R = R/G) %>%
  mutate(R_hat = predict(fit3, newdata = .)) %>%
  summarize(correlation = cor(R, R_hat))
```

```
##   correlation
## 1   0.9225805
```

c

```
Batting1 <- Batting %>% filter(yearID == 2015)
Salaries1 <- Salaries %>% filter(yearID == 2015)
Salaries1 %>%
  left_join(Batting1, by = 'playerID') %>%
  head(6)
```

```
##   yearID.x teamID.x lgID.x  playerID  salary yearID.y stint teamID.y lgID.y   G
## 1     2015      ARI     NL ahmedni01  508500     2015     1      ARI     NL 134
## 2     2015      ARI     NL anderch01  512500     2015     1      ARI     NL  28
## 3     2015      ARI     NL chafian01  507500     2015     1      ARI     NL  66
## 4     2015      ARI     NL collmjo01 1400000     2015     1      ARI     NL  44
## 5     2015      ARI     NL corbipa01  524000     2015     1      ARI     NL  16
## 6     2015      ARI     NL delarru01  516000     2015     1      ARI     NL  32
##    AB  R  H X2B X3B HR RBI SB CS BB SO IBB HBP SH SF GIDP
## 1 421 49 95  17   6  9  34  4  5 29 81   1   1  5  3    4
## 2  48  0  5   0   0  0   3  0  0  1 23   0   0  8  0    2
## 3   3  0  0   0   0  0   0  0  0  0  1   0   0  0  0    0
## 4  27  2  5   0   0  0   1  0  0  3  9   0   0  2  0    1
## 5  25  1  3   0   0  0   3  0  0  3 11   0   0  1  0    1
## 6  64  3  6   0   0  0   2  0  0  0 25   0   0  4  0    0
```

4

```
library(dslabs)
data(admissions)
dat <- admissions %>% dplyr::select(-applicants)
```

a

```
dat %>%
  pivot_wider(names_from = gender, values_from = admitted)
```

```
## # A tibble: 6 x 3
##   major   men women
##   <chr> <dbl> <dbl>
## 1 A        62    82
## 2 B        63    68
## 3 C        37    34
## 4 D        33    35
## 5 E        28    24
## 6 F         6     7
```

b

```
tmp <- admissions %>%
  pivot_longer(c('admitted', 'applicants'), names_to = 'key', values_to = 'value')
head(tmp)
```

```
## # A tibble: 6 x 4
##   major gender key        value
##   <chr> <chr>  <chr>      <dbl>
## 1 A     men    admitted      62
## 2 A     men    applicants   825
## 3 B     men    admitted      63
## 4 B     men    applicants   560
## 5 C     men    admitted      37
## 6 C     men    applicants   325
```