CS100 Homework 2 (three-day late pass)

Mariel Pacada

10/31/2019

Part 1: Deflategate

```
avg_psi <- football %>%
             separate(Football, c("Team", "Ball")) %>%
             group_by(Team) %>%
             summarize(Average_B = mean(Blakeman), Average_P = mean(Prioleau))

avg_psi
```

```
## # A tibble: 2 x 3
##   Team     Average_B Average_P
##   <chr>        <dbl>     <dbl>
## 1 Colts         12.6      12.4
## 2 Patriots      11.1      11.5
```

For both of the given measurements, the Patriots had a lower average psi than the Colts.

```
pats_vs_colts <- football %>%
                   separate(Football, c("Team", "Ball")) %>%
                   mutate(Average = (Blakeman + Prioleau)/2)

Drop <- c()
for (i in pats_vs_colts$Team) {
  switch(i,
         "Patriots" = Drop <- 12.5 - pats_vs_colts$Average[pats_vs_colts$Team == i],
         "Colts" = Drop <- c(Drop, 13 - pats_vs_colts$Average[pats_vs_colts$Team == i]),
         break)
} # there's a bug here i don't know how to fix
  # returning a vector with 27 elements == (11 patriots + 4 Colts) +
  # all the Colts 3 extranenous times

Drop <- Drop[1:15] #working around the bug

avg_drop <- pats_vs_colts %>%
              mutate(Drop = Drop) %>%
              group_by(Team) %>%
              summarize(Average_Drop = mean(Drop))

avg_drop
```

```
## # A tibble: 2 x 2
##   Team     Average_Drop
##   <chr>           <dbl>
## 1 Colts           0.469
## 2 Patriots        1.20
```

The Patriots' average drop was greater than that of the Colts.

```
test_stat <- avg_drop$Average_Drop[avg_drop$Team == "Patriots"] -
  avg_drop$Average_Drop[avg_drop$Team == "Colts"]
test_stat
```

```
## [1] 0.7335227
```

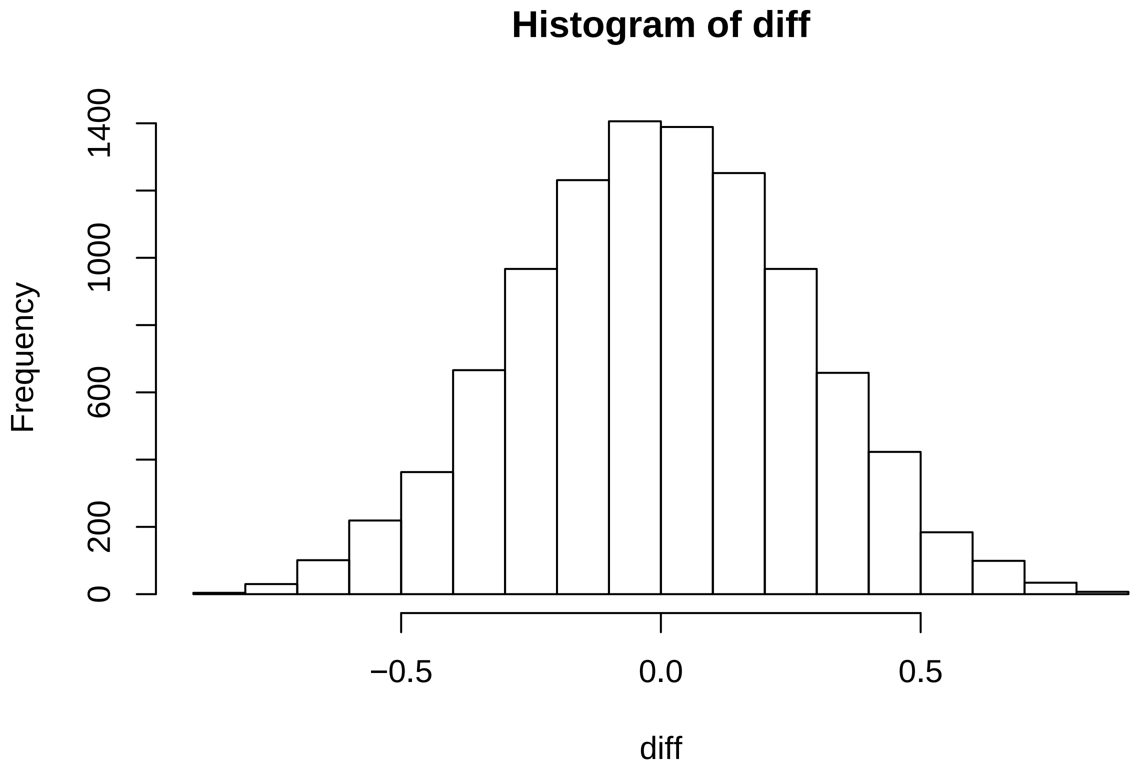
The difference between the means is about 0.7 psi.

```
diff_sim <- function(data) {
  m <- sample(data, replace = TRUE)
  mean(m[1:11]) - mean(m[12:15])
}

diff = replicate(10000, diff_sim(Drop))
head(diff, 10)
```

```
##  [1]  0.49431818  0.41420455 -0.04488636  0.19943182 -0.25397727
##  [6] -0.30284091  0.18522727 -0.24545455  0.66761364  0.11534091
```

```
hist(diff)
```



```r
sum <- 0
for (i in diff > 0.7335) {
  if (i == TRUE) {
    sum <- sum + 1
  }
}

sum / 10000 # == 0.0029
```

```
## [1] 0.0021
```

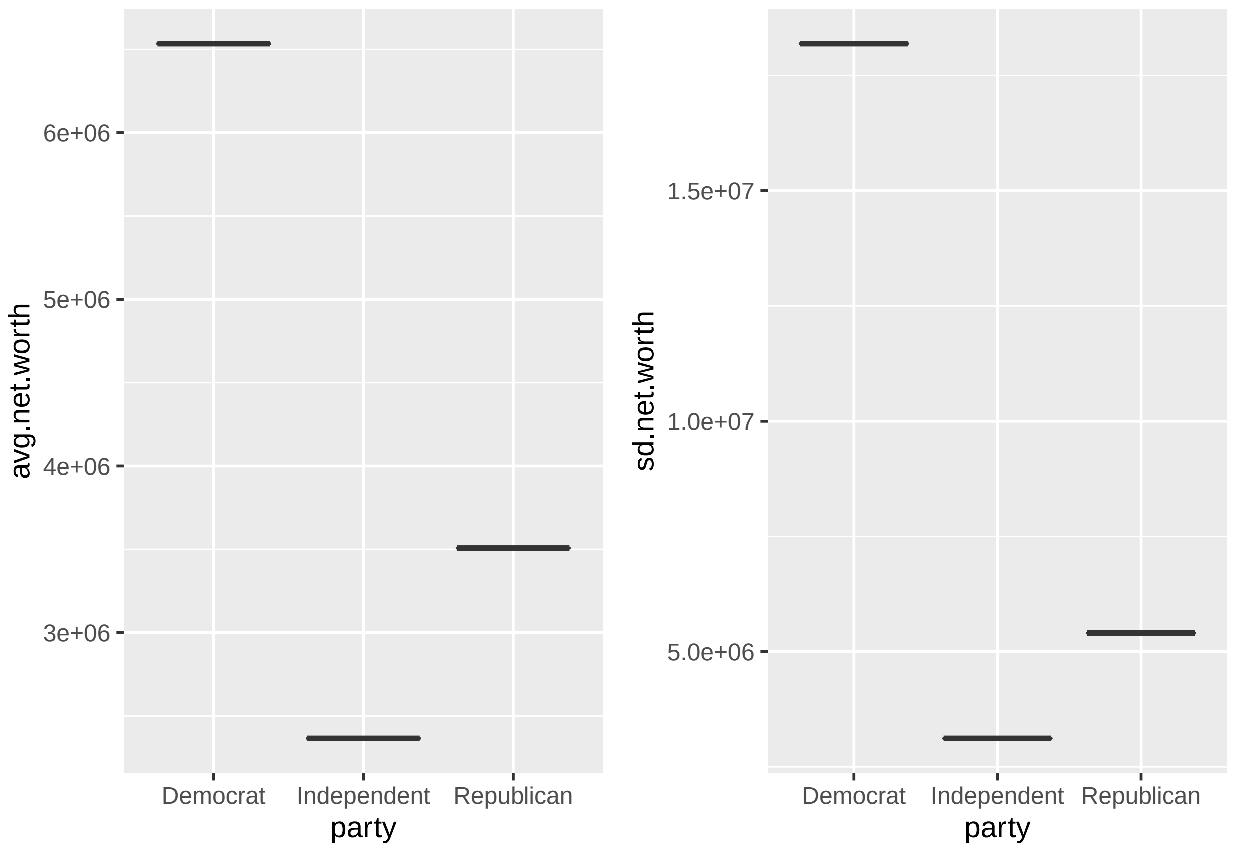
Just by looking at the histogram, it seems that achieving a 0.7 psi difference in mean is low-probability. We confirm this my calculating the empirical probability, which is 0.0029. This means that there had been a 0.0029 probability of getting the mean difference of 0.7, which could suggest that it had not happened only by chance.

(Note: Because the replication differs each time the code is run, the empirical probability might be slightly different.)

Part 2: Deflategate

```r
senate_income <- senate %>%
                   select(party, net.worth) %>%
                   group_by(party) %>%
                   summarize(avg.net.worth = mean(net.worth), sd.net.worth = sd(net.worth))

p1 <- ggplot(senate_income, aes(x = party, y = avg.net.worth)) + geom_boxplot()
p2 <- ggplot(senate_income, aes(x = party, y= sd.net.worth)) + geom_boxplot()
grid.arrange(p1, p2, nrow = 1, ncol = 2)
```



From this visualization, we can clearly see that Democrats have a higher average net worth then Republicans. They also have a higher standard deviation, which means the data is more spread out for Democrats. Now, we perform a significance test to determine if there is a difference between the mean net worth of Democrats and the mean net worth of Republicans.

```
# average net worth in variables
d_avg <- senate_income$avg.net.worth[senate_income$party == "Democrat"]
r_avg <- senate_income$avg.net.worth[senate_income$party == "Republican"]

# count democrats and republicans
senate_tally <- senate %>%
                  select(party) %>%
                  add_count(party) # bug i don't know how to fix, again :(
                                   # returns all 100 observations instead of groups
                                   # (group_by doesn't fix it)

# individual standard error = sample sd / square root of sample size
d_se <- senate_income$sd.net.worth[senate_income$party == "Democrat"] / 44
r_se <- senate_income$sd.net.worth[senate_income$party == "Republican"] / 54

# standard error for difference in means
diff_mean_se <- (d_se ** 2 + r_se ** 2) ** (1/2)

# z-statistic
z <- (d_avg - r_avg) / diff_mean_se
```

```
# significance
pnorm(z, 0, 1, lower.tail = FALSE)
```

```
## [1] 5.4708e-13
```

The p-value of this test turned out to be 5 times 10 to the negative thirteenth power (5.4708e-13). This is clearly lower than the our value of alpha, which is 0.05. Thus, we reject the null hypothesis and conclude that the Democrats are significantly richer than the Republicans.