Are Android Phones More Negative?

David Gerard

2017-09-18

Learning Objectives

- Introduction to hypothesis testing.
- Section 1.8 of DBC.

Trump's Tweets

```
library(tidyverse)
read_csv("../../data/trump.csv") %>%
  select(source, text, hour,
         quote, picture, positive, negative) %>%
  filter(quote == "no_quote") ->
  trump
glimpse(trump)
```

```
Observations: 1,208
Variables: 7
$ source   <chr> "Android", "iPhone", "iPhone", "Androi...
$ text     <chr> "My economic policy speech will be car...
$ hour     <int> 10, 8, 19, 18, 16, 8, 21, 21, 20, 15, ...
$ quote    <chr> "no_quote", "no_quote", "no_quote", "n...
$ picture  <chr> "no_picture", "picture", "picture", "n...
$ positive <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, FA...
$ negative <lgl> FALSE, FALSE, FALSE, TRUE, TRUE, TRUE,...
```

Why are we interested in this?

8/4/2017 Todd Vaziri on Twitter: "Every non-hyperbolic tweet is from iPhone (his staff). Every hyperbolic tweet is from Android (from him). https://t.co/GWr6D8h5ed"

Todd Vaziri
@tvaziri

Follow

Every non-hyperbolic tweet is from iPhone (his staff).

Every hyperbolic tweet is from Android (from him).

12:20 PM - 6 Aug 2016

10,380 Retweets **14,521** Likes

266 10K 15K

https://twitter.com/tvaziri/status/762005541388378112/photo/1?ref_src=twsrc%5Etfw&ref_url=http%3A%2F%2Fvarianceexplained.org%2Fr%2Ftrump-tweets%2F 1/1

Example

- Tweet from android: “The dishonest media didn’t mention that Bernie Sanders was very angry looking during Crooked’s speech. He wishes he didn’t make that deal!”
- Tweet from iPhone: “Join me in Fayetteville, North Carolina tomorrow evening at 6pm. Tickets now available at:”
- Let’s see if these differences are actually statistically meaningful.

We use sentiment analysis to evaluate this statement.

- According to one annotation, each word can consist of one/none of two sentiments (positive or negative) and some/all/none of eight primary emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust).
- See `http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm`.
- Examples:
  - abandon has the **negative** sentiment and the **fear** and **sadness** emotions.
  - trump has no sentiment and the **surprise** emotion.
  - maroon has the **negative** sentiment and no emotions.

Caveat: Sentiment analysis is not perfect:

- Tweet: "Michael Morell, the lightweight former Acting Director of C.I.A., and a man who has made serious bad calls, is a total Clinton flunky!"
- Seems negative.
- bad has sentiments "disgust", "fear", "negative", and "sadness"
- calls has sentiments "anticipation", "negative", "trust"
- director has sentiments "positive" and "trust".
- So we would say it has elements of disgust, fear, negative, positive, sadness, anticipation, and trust? This seems a little too complicated for a negative tweet.

two-way table

```
tabdat <- table(trump$negative, trump$source)
rownames(tabdat) <- c("Non-negative", "Negative")
tabdat
```

```
##
##                Android iPhone
##   Non-negative     245    456
##   Negative         341    165
```

If we want to see an association between phone-source and negative sentiments, what conditional distribution should we look at?

two-way table

```
proptab <- prop.table(tabdat, margin = 2)
proptab
```

```
              Android iPhone
  Non-negative  0.4181 0.7343
  Negative      0.5819 0.2657
```

- 58% of tweets from Androids contain some negative words.
- 27% of tweets from iPhones contain some negative words.
- Seems like a large difference = 32%. But couldn't we have just seen this by chance?
- E.g. if President Trump uses a new phone at random, but by chance he happened to use the Android phone for more negative tweets.

Hypotheses

- We label these hypotheses H_0 and H_A.
- H_0: The variables `source` and `negative` are independent. They have no relationship, and the observed difference in negative proportions was due to chance.
- H_A: The variables `source` and `negative` are not independent (they are associated). The observed difference in negative proportions is not due to chance.

Observed/Expected counts under H_0

Observed:

	Android	iPhone	Total
Non-negative	245	456	701
Non-negative	341	165	506
Total	586	621	1207

Expected:

	Android	iPhone	Total
Non-negative	$586\frac{701}{1207} = 340$	$621\frac{701}{1207} = 361$	701
Non-negative	$586\frac{506}{1207} = 246$	$621\frac{506}{1207} = 260$	506
Total	586	621	1207

Expected $=$ sample size $\times$ observed overall rate.

Do we expect exactly this result?

- If H_0 were true, would we expect the difference in proportions of tweets that are negative to be *exactly* zero?
- NO! Just by chance, we would expect one phone to send out a few more negative tweets than the other phone.
- If you flip a fair coin, do you always expect *exactly* 50% of the flips to be tails?
- But what constitutes "a few"?

How are tweets generated under H_0?

- Under H_0, Trump chooses a tweet, then randomly chooses a phone to send out the tweet, regardless of it being negative or not.
- We can actually perform this randomization!
- I.e., randomly assign 586 of the tweets (whose negativity we know) to be sent from the Android phone and the rest (622) to be sent from the iPhone.
- Why these numbers?

```
table(trump$source)


Android  iPhone
    586     622
```

Resampling

The idea of resampling is to

- use only the observed data (not a statistical model)
- resample (sample from the sample)
- with or without replacement
- I create different realizations of possible experimental results (if the null hypothesis were actually true).
- I compare many, many resampled experimental results with the observed experimental results I decide if observed result is common or rare to occur by chance

- If observed data are rare compared to resampled results: the data may point to something interesting (an effect)
- If observed data are common within resampled results: maybe result just occurred by chance (no evidence of an effect)

Applet Simulation:

`http://www.rossmanchance.com/applets/ChiSqShuffle.html?yawning=1`

One such simulation

```r
tabdat <- table(trump$negative, sample(trump$source))
propdat <- prop.table(tabdat, margin = 2)
propdat
```

```
       Android iPhone
  FALSE  0.5573 0.6029
  TRUE   0.4427 0.3971
```

So in this case, 0.5573 of the Android tweets are negative and 0.6029 of the iPhone tweets are negative.
This difference -0.0456 is much smaller than in the original dataset.

Wait, what was that code?

```
new_dat <- data_frame(negative = trump$negative,
                      source = sample(trump$source))
print(new_dat, n = 7)
```

```
# A tibble: 1,208 x 2
  negative  source
     <lgl>   <chr>
1    FALSE Android
2    FALSE Android
3    FALSE  iPhone
4     TRUE Android
5     TRUE  iPhone
6     TRUE Android
7    FALSE Android
# ... with 1,201 more rows
```

Wait, what was that code?

```
new_dat <- data_frame(negative = trump$negative,
                      source = sample(trump$source))
print(new_dat, n = 7)
```

```
# A tibble: 1,208 x 2
  negative  source
     <lgl>   <chr>
1    FALSE Android
2    FALSE Android
3    FALSE Android
4     TRUE  iPhone
5     TRUE  iPhone
6     TRUE  iPhone
7    FALSE Android
# ... with 1,201 more rows
```

Wait, what was that code?

```
new_dat <- data_frame(negative = trump$negative,
                      source = sample(trump$source))
print(new_dat, n = 7)

# A tibble: 1,208 x 2
  negative  source
     <lgl>   <chr>
1    FALSE  iPhone
2    FALSE Android
3    FALSE  iPhone
4     TRUE Android
5     TRUE Android
6     TRUE  iPhone
7    FALSE Android
# ... with 1,201 more rows
```

Wait, what was that code?

I am keeping `negative` fixed while shuffling the ordering of `source`.
Then I create the contingency table.

```
table(new_dat$negative, new_dat$source)
```

```
        Android iPhone
  FALSE     336    365
  TRUE      249    257
```

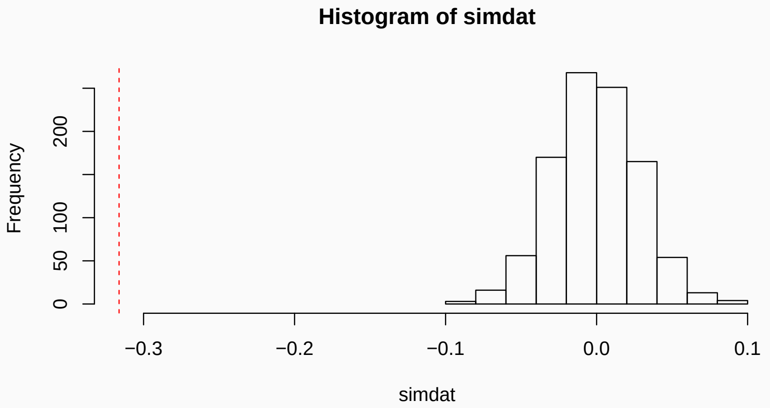
We can repeat this

Repeating this many times will tell us what the “likely” values of the difference are under H_0.

```
simdat <- rep(NA, length = 1000)
for (index in 1:1000) {
  tabdat <- table(trump$negative, sample(trump$source))
  propdat <- prop.table(tabdat, margin = 2)
  simdat[index] <- propdat[1, 1] - propdat[1, 2]
}
realtab <- prop.table(table(trump$negative, trump$source),
                      margin = 2)
realstat <- realtab[1, 1] - realtab[1, 2]
```

Plot the simulations

```
hist(simdat, xlim = c(realstat, max(simdat)))
abline(v = realstat, col = 2, lty = 2)
```



Possible conclusions

- H_0: `source` and `negative` are not associated, what we observed was just do to random chance, even though the probability of observing the data we saw (given that this was just due to random chance) is remarkably small.
- H_A: `source` and `negative` are associated.
- Since the data we observe is incredibly unlikely under H_0, we reject H_0 and conclude H_A.
- This idea of rejecting a hypothesis when the data are rare under said hypothesis is the foundation of much of statistical inference.

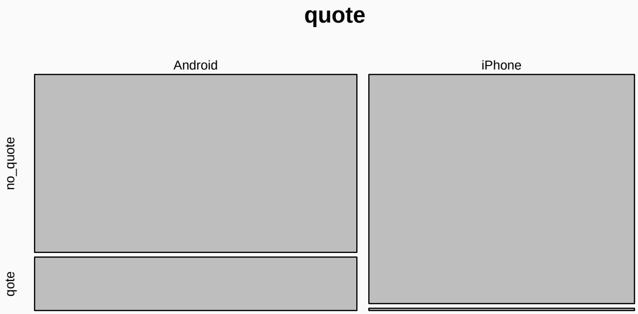
Some Fun

```
read_csv("../../data/trump.csv") %>%
  select(source, text, hour,
         quote, picture) ->
  trump
```

Some Fun

Weird copy and pasting:

```
plot(prop.table(table(trump$source, trump$quote)),
     main = "quote")
```



Some Fun

Pictures for advertising events:

```
plot(prop.table(table(trump$source, trump$picture)),
     main = "picture")
```