Analysing Twitter for Ubisoft

Ryan Greenup

April 27, 2020

Contents

8.1 Analysing the Relationship Between Friends and Followers for Twitter Users

8.1.1 Retrieve the posts from Twitter

relevant posts can be retrieved from twitter by utilising the `rtweet` package, packages can be loaded for use in **R** thusly:

The `rtweet` API will search for tweets that contain all the words of a query regardless of uppercase or lowercase usage [4].

In order to leverage the *Twitter* API it is necessary to use tokens provided through a *Twitter* developer account:

and hence all tweets containing a mention of *Ubisoft* can be returned and saved to disk as shown in listing 3:

8.2.2 Count of Followers and Friends

In order to identify the number of users that are contained in the *tweets* the `unique()` function can be used to return a vector of names which can then be passed as an index to the vector of counts as shown in listing 4, this provides that 81.7% of the tweets are by unique users.

```r
# Load Packages
 ↪  ------------------------------------------------------------
setwd("~/Dropbox/Notes/DataSci/Social_Web_Analytics/SWA-Project/scripts
 ↪  /")

if (require("pacman")) {
  library(pacman)
} else{
  install.packages("pacman")
  library(pacman)
}

pacman::p_load(xts, sp, gstat, ggplot2, rmarkdown, reshape2,
               ggmap, parallel, dplyr, plotly, tidyverse,
               reticulate, UsingR, Rmpfr, swirl, corrplot,
               gridExtra, mise, latex2exp, tree, rpart,
               lattice, coin, primes, epitools, maps, clipr,
               ggmap, twitteR, ROAuth, tm, rtweet, base64enc,
               httpuv, SnowballC, RColorBrewer, wordcloud,
               ggwordcloud, tidyverse, boot)
```

Listing 1: Load the Packages for ***R***

8.1.3 Summary Statistics

The average number of friends and followers from users who posted tweets mentioning *Ubisoft* can be returned using the `mean()` as shown in listing 5 this provides that on average each user has 586 friends and 63,620 followers.

8.1.4 Above Average Followers

Each user can be compared to the average number of followers, by using a logical operator on the vector (e.g. `y > ybar`), this will return an output of logical values. ***R*** will coerce logicals into 1/0 values meaning that the mean value will return the proportion of `TRUE` responses as shown in listing 6. This provides that 20.6% of the users identified have above average friend counts, while only 2.4% have an above average numbmer of followers.

```r
# Set up Tokens ==========================================================

options(RCurlOptions = list(
  verbose = FALSE,
  capath = system.file("CurlSSL", "cacert.pem", package = "RCurl"),
  ssl.verifypeer = FALSE
))

setup_twitter_oauth(
  consumer_key = "************************",
  consumer_secret = "**************************************************",
  access_token = "**************************************************",
  access_secret = "******************************************"
)

# rtweet ================================================================
tk <-    rtweet::create_token(
  app = "SWA",
  consumer_key    = "************************",
  consumer_secret = "**************************************************",
  access_token    = "**************************************************",
  access_secret   = "******************************************",
  set_renv        = FALSE
```

Listing 2: Import the twitter tokens (redacted)

```r
 n <- 1000
 tweets.company <- search_tweets(q = 'ubisoft', n = n, token = tk,
                                 include_rts = FALSE)
 save(tweets.company[,], file = "resources/Download_1.Rdata")
```

Listing 3: Save the Tweets to the HDD as an `rdata` file

```
(users <- unique(tweets.company$name)) %>% length()
x <- tweets.company$followers_count[duplicated(tweets.company$name)]
y <- tweets.company$friends_count[duplicated(tweets.company$name)]

## > [1] 817
```

Listing 4: Return follower count of twitter posts

```
x<- rnorm(090)
y<- rnorm(090)
(xbar <- mean(x))
(ybar <- mean(y))

## > [1] 4295.195
## > [1] 435.9449
```

Listing 5: Determine the average number of friends and followers

```
(px_hat <- mean(x>xbar))
(py_hat <- mean(y>ybar))

## > [1] 0.0244798
## > [1] 0.2729498
```

Listing 6: Calculate the proportion of users with above average follower counts

8.1.5 Bootstrap confidence intervals

a.) Generate a bootsrap distribution

A bootstrap assumes that the population is an infinitely large repetition of the sample, a bootstrap of the follower counts can be produced by resampling with replacement/repetition and plotted using the `ggplot2` library as shown in listing 7 and figure 1.

This shows that the population follower counts is a non-normal skew-right distribution, which is expected because the number of friends is an integer value bound by zero [5].

```
## Resample the Data
kt_pop <- sample(x, size = 10^6, replace = TRUE)

## Make the Population
bt_pop_data <- tibble("Followers" = bt_pop)
ggplot(data = bt_pop_data, aes(x = Followers)) +
  geom_histogram(aes(y = ..density..), fill = "lightblue", bins = 35,
  ↪  col = "pink") +
  geom_density(col = "violetred2") +
  scale_x_continuous(limits = c(1, 800)) +
  theme_bw() +
  labs(x = "Number of Followers", y = "Density",
       title = "Bootstrapped population of Follower Numbers")
```

Listing 7: Bootstrapping a population from the sample.

b.) Estimate a Confidence Interval for Follower Counts

- The normal t value bootstrap offers now advantage over using a t distribution (other than being illustrative of bootstrapping generally) [3, Section 4.1]
- It is more appropriate to use a percentile interval for skewed data such as this, but it is still not very accurate [3, Section 4.1]

  The `boot` package implements confidence intervals consistent with work by Davison and Hinkley [6] in there texbook *Bootstrap Methods and their Application* wherein the consensus is that BC_a methods will be [2, Ch. 5] superior to mere percentile methods in terms of accuracy. [1, 2, Ch. 5]

  Wait, hold on, the population is non-normal, sure, but the sampling distribution will obviously be normal, have I gotten confused?

references

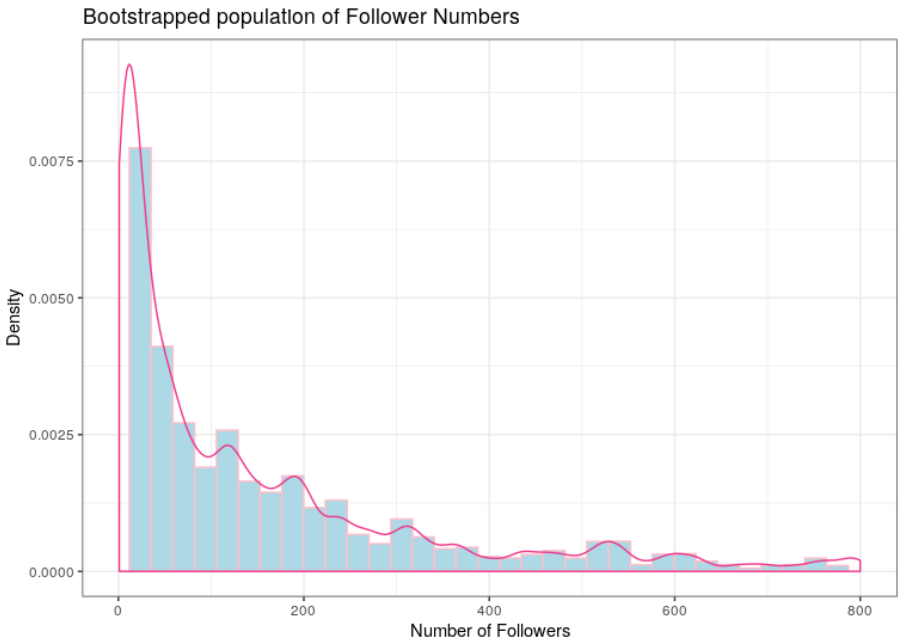

Figure 1: Histogram of the bootrapped population of follower counts

References

[1] James Carpenter and John Bithell. "Bootstrap Confidence Intervals: When, Which, What? A Practical Guide for Medical Statisticians". en. In: *Statistics in Medicine* 19.9 (2000), pp. 1141–1164. ISSN: 1097-0258. DOI: 10.1002/(SICI)1097-0258(20000515)19:9<1141::AID-SIM479>3.0.CO;2-F. URL: https://doi-org.ezproxy.uws.edu.au/10.1002/(SICI)1097-0258(20000515)19:9%3C1141::AID-SIM479%3E3.0.CO;2-F (visited on 04/27/2020) (cit. on p. 5).

[2] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and Their Application*. Cambridge ; New York, NY, USA: Cambridge University Press, 1997. ISBN: 978-0-521-57391-7 978-0-521-57471-6 (cit. on p. 5).

[3] Tim C. Hesterberg. "What Teachers Should Know About the Bootstrap: Resampling in the Undergraduate Statistics Curriculum". In: *The American Statistician* 69.4 (Oct. 2015), pp. 371–386. ISSN: 0003-1305. DOI: 10.1080/00031305.2015.1089789. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4784504/ (visited on 04/26/2020) (cit. on p. 5).

[4] Michael Kearney. *Get Tweets Data on Statuses Identified via Search Query. Search_tweets*. en. Manual. 2019. URL: https://rtweet.info/reference/search_tweets.html (visited on 04/26/2020) (cit. on p. 1).

[5] NIST. *1.3.3.14.6. Histogram Interpretation: Skewed (Non-Normal) Right*. Oct. 2013. URL: https://www.itl.nist.gov/div898/handbook/eda/section3/histogr6.htm (visited on 04/26/2020) (cit. on p. 2).

[6] Brian Ripley. *Boot.Ci Function | R Documentation*. Apr. 2020. URL: https://www.rdocumentation.org/packages/boot/versions/1.3-25/topics/boot.ci (visited on 04/27/2020) (cit. on p. 5).