Exercise: Facebook Statistics

James M. Murray, Ph.D.
University of Wisconsin - La Crosse

Updated: September 13, 2017

PDF file location: http://www.murraylax.org/rtutorials/facebook.pdf

HTML file location: http://www.murraylax.org/rtutorials/facebook.html

---

Note on required packages: The following code requires the packages in the `tidyverse`. The `tidyverse` actually contains many packages that allow you to organize, summarize, and plot data. The package `psych` is used to perform statistics related to the median. If you have not already done so, download and install the libraries (needed only once per computer), and load the libraries (need to do every time you start R) with the following code:

```
install.packages("psych") # This only needs to be executed once for your machine
install.packages("tidyverse") # This only needs to be executed once for your machine
library("psych") # This needs to be executed every time you load R
library("tidyverse") # This needs to be executed every time you load R
```

Data Set

The following data set comes from the following study on Facebook marketing and performance metrics:

Moro, S., Rita, P. and Vala, B., (2016) "Predicting Social Media Performance Metrics and Evaluation of the Impact on Brand Building: A Data Mining Approach" *Journal of Business Research*, Vol. 68, pp. 3341-3351. Available at http://www.sciencedirect.com/science/article/pii/S0148296316000813

Download and load into memory the data set (see other formats at end of document):

```
load(url("http://murraylax.org/datasets/facebook.RData"))
```

The data set includes statistics from 500 Facebook posts in 2014 related to the marketing of a globally known cosmetic brand. Facebook marketing is an important part of many businesses marketing strategy. Facebook interaction can help businesses build their brand and market new products. Marketing executives such statistics to better understand the effectiveness of their Facebook marketing.

The data set includes the following variables:

1. **Type**: Scale / Class: Nominal / Factor. Type of post. Possible outcomes are "Link", "Photo", "Status", and "Video"
2. **Month**: Scale / Class: Ordinal / Ordered factor. Month of the year for the post.
3. **Weekday**: Scale / Class: Ordinal / Ordered factor. Day of the week for the post.
4. **Hour**: Scale / Class: Ratio / Integer. Hour of the day - between 0 (12:00AM) and 23 (11:00PM)
5. **Paid**: Scale / Class: Binary / Integer. Dummy variable equal to 1 if a paid post, 0 if a free or unsolicited post.

6. **Reach**: Scale / Class: Ratio / Integer. Number of unique individuals who saw the post appear on their news feeds.
7. **Impressions**: Scale / Class: Ratio / Integer. Number of times the post appeared on people's news feeds (some individuals may have had the post appear more than once)
8. **EngagedUsers**: Scale / Class: Ratio / Integer. Number of unique individuals that clicked anywhere in the post.
9. **Comments**: Scale / Class: Ratio / Integer. Number of comments on the post.
10. **Likes**: Scale / Class: Ratio / Integer. Number of likes for the post
11. **Shares**: Scale / Class: Ratio / Integer. Number of shares for the post
12. **Interactions**: Scale / Class: Ratio / Integer. The sum, Comments + Likes + Shares.
13. **Weekday.Int**: Scale / Class: Ordinal / Integer: Number associated with day of the week in **Weekday**
14. **Month.Int**: Scale / Class: Ordinal / Integer: Number associated with month in **Month**

Exercises

1. Descriptive Statistics

A. Compute the mean and standard deviation for the number of engaged users per post.

B. Compute the median and interpolated median for the number of comments per post.

C. Compute the frequencies of Facebook posts by day of the week. What two days have the most number of Facebook posts?

D. Compute the median number of comments by day of week. What day results in the most number of comments? What three days result in the least number of comments?

E. Compute the mean number of engaged users by post *type*. What type of post leads to the most engaged users? What type of post leads to least number of engaged users? Is this what you expected?

2. Inferential Statistics

A. Compute and interpret a 95% confidence interval for the mean number of shares of a Facebook post

B. Compute the mean and interpret a 95% confidence interval for the mean for the number of shares generated by a Facebook post created on a Monday

C. Compare Facebook posts of photos versus videos. Is there statistical evidence that one results in a higher mean number of interactions than the other?

D. Answer the following questions as they related to the difference in the mean number of engaged users from paid posts versus unpaid posts.

D.1. Compute and interpret a 95% confidence interval for the difference in the mean number of engaged users between paid posts versus unpaid posts.

D.2. Is there statistical evidence that paid posts result in *more engaged users* than unpaid posts?

E. Report the medians and interpolated medians for the number of shares and comments for Facebook posts. Is there statistical evidence that the medians for shares and comments are different?

Answers

1. Descriptive Statistics

A. Compute the mean and standard deviation for the number of engaged users per post.

```
mean(df$EngagedUsers)
```

```
## [1] 920.344
```

```
sqrt( var(df$EngagedUsers) )
```

```
## [1] 985.0166
```

B. Compute the median and interpolated median for the number of comments per post.

```
median(df$Comments)
```

```
## [1] 3
```

```
interp.median(df$Comments)
```

```
## [1] 2.805556
```

C. Compute the frequencies of Facebook posts by day of the week. What two days have the most number of Facebook posts?

```
summary(df$Weekday)
```

```
##    Monday   Tuesday Wednesday  Thursday    Friday  Saturday    Sunday
##        68        66        64        72        67        81        82
```

Saturday and Sunday have the most number of Facebook posts.

D. Compute the median number of comments by day of week. What day results in the most number of comments? What three days result in the least number of comments?

```
df %>%
  group_by(Weekday) %>%
  summarize(MedianComments = median(Comments))
```

```
## # A tibble: 7 x 2
##     Weekday MedianComments
##       <ord>          <dbl>
## 1    Monday              2
## 2   Tuesday              3
## 3 Wednesday              3
## 4  Thursday              5
## 5    Friday              3
## 6  Saturday              2
## 7    Sunday              2
```

Saturday, Sunday, and Monday results in the least number of comments, each generating a median number of comments equal 2.

E. Compute the mean number of engaged users by post *type*. What type of post leads to the most engaged users? What type of post leads to least number of engaged users? Is this what you expected?

```
df %>%
  group_by(Type) %>%
  summarize(MeanEngagedUsers = mean(EngagedUsers))
```

```
## # A tibble: 4 x 2
##     Type MeanEngagedUsers
##   <fctr>            <dbl>
## 1   Link         342.8182
## 2  Photo         818.9460
## 3 Status        2040.2222
## 4  Video        1707.0000
```

Statuses result in the most number of engaged users. Links result in the least number of engaged users.

2. Inferential Statistics

A. Compute and interpret a 95% confidence interval for the mean number of shares of a Facebook post

```
shares.test <- t.test(df$Shares)
shares.test$conf.int
```

```
## [1] 23.50676 31.02550
## attr(,"conf.level")
## [1] 0.95
```

We are 95% confident that the interval, (23.5 - 31.0), contains the true mean number of shares per Facebook post.

B. Compute the mean and interpret a 95% confidence interval for the mean for the number of shares generated by a Facebook post created on a Monday

```
df %>%
  filter(Weekday=="Monday") ->
  testdata

mean(testdata$Shares, na.rm=T)
```

```
## [1] 24.76471
```

```
shares.test <- t.test(testdata$Shares)
shares.test$conf.int
```

```
## [1] 17.79351 31.73590
## attr(,"conf.level")
## [1] 0.95
```

The mean number of shares of posts created on Mondays is equal to 27.57.

We are 95% confident the interval (23.3, 31.8) includes the true number of shares from Facebook posts created on Mondays.

C. Compare Facebook posts of photos versus videos. Is there statistical evidence that one results in more interactions than the other?

H_0: $\mu_{Photo} - \mu_{Video} = 0$

H_1: $\mu_{Photo} - \mu_{Video} \neq 0$

```
df %>%
  filter(Type=="Photo" | Type=="Video") %>%
  droplevels() ->
  testdata
```

```r
t.test(Interactions ~ Type, testdata, paired=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  Interactions by Type
## t = -1.097, df = 6.9964, p-value = 0.309
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -250.17989   91.62523
## sample estimates:
## mean in group Photo mean in group Video
##            216.5798            295.8571
```

With a p-value equal to 0.309, we fail to reject the null hypothesis.

We fail to find statistical evidence that photo posts result in more interactions than videos.

D. Answer the following questions as they related to the difference in the mean number of engaged users from paid posts versus unpaid posts.

D.1. Compute and interpret a 95% confidence interval for the difference in the mean number of engaged users between paid posts versus unpaid posts.

```r
t.test(EngagedUsers ~ Paid, data=df)
```

```
##
##  Welch Two Sample t-test
##
## data:  EngagedUsers by Paid
## t = -2.0857, df = 173.66, p-value = 0.03846
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -500.34855  -13.80736
## sample estimates:
## mean in group 0 mean in group 1
##        849.4472       1106.5252
```

We are 95% confidence the interval (13.81, 500.34) includes the mean difference in engaged users of paid posts versus unpaid posts.

D.2. Is there statistical evidence that paid posts result in *more engaged users* than unpaid posts?

H_0: $\mu_0 - \mu_1 = 0$

H_1: $\mu_0 - \mu_1 < 0$

```r
t.test(EngagedUsers ~ Paid, alternative='less', data=df)
```

```
##
##  Welch Two Sample t-test
##
## data:  EngagedUsers by Paid
## t = -2.0857, df = 173.66, p-value = 0.01923
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -53.25427
## sample estimates:
```

```
## mean in group 0 mean in group 1
##        849.4472       1106.5252
```

With a p-value = 0.01923 which is less than 0.05, we reject the null hypothesis.

We find statistical evidence that paid posts result in more engaged users than unpaid posts.

E. Report the medians and interpolated medians for the number of shares and comments for Facebook posts. Is there statistical evidence that the medians for shares and comments are different?

```
median(df$Shares, na.rm=T)
```

```
## [1] 19
```

```
median(df$Comments, na.rm=T)
```

```
## [1] 3
```

```
interp.median(df$Shares, na.rm=T)
```

```
## [1] 18.75
```

```
interp.median(df$Comments, na.rm=T)
```

```
## [1] 2.805556
```

```
wilcox.test(df$Shares, df$Comments, paired = T)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  df$Shares and df$Comments
## V = 111330, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

The median and interpolated median for number of shares is 19 and 18.75, respectively.

The median and interpolated median for number of comments is 3 and 2.8, respectively.

H_0 : Median Shares = Median Comments

H_A : Median Shares $\neq$ Median Comments

The p-value is less than 2.2^{-16}, which is below a significance level of 0.05. Reject the null hypothesis.

There is statistical evidence that the median number of shares is different than the median number of comments.

Other Formats for Data Download

You can download the data in alternative formats using the following links. Always cite the above source if using the data.

RData: http://murraylax.org/datasets/facebook.RData

Excel: http://murraylax.org/datasets/facebook.xlsx

CSV: http://murraylax.org/datasets/facebook.csv

SPSS: http://murraylax.org/datasets/facebook.sav

Open Document Spreadsheet (LibreOffice, OpenOffice): http://murraylax.org/datasets/facebook.ods