Readership Modeling v1.1 Comments

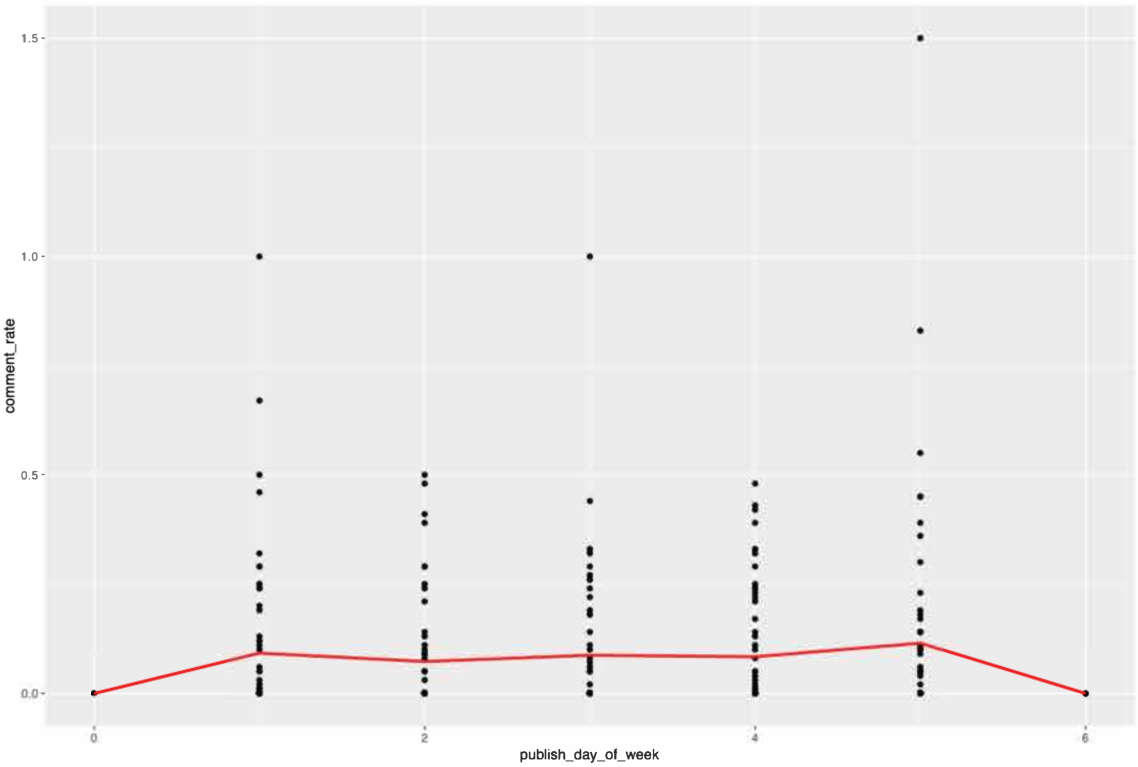
Paula Lee

11/26/2018

In this document, we look at how average comment rates depend on categorial variables (days of the week). In particular, we want to know *how to identify* if any difference that we see in the comment rates (depending on the frequency of the categorical variables) are "real" or not.

We will be focusing on the case regarding the relationship between comment rate and publish day of week.

Comment Rates from Customer1 and Customer2 by publish day of week:



We want to find out if this difference means anything.

A commonway to do this is to conduct a one-way ANOVA through an F-test:

To understand what ANOVA(analysis of variance) and an F-test is, we will cover some basics. ANOVA is used to test variability within and between three or more groups. In our case, we are using a one-way ANOVA, which compares three or more independent groups to one variable. Our independent groups are the publish day of week and our one variable is the comment rate. To find if there is any statistically significant differences of the means within and between groups, ANOVA uses F-tests to test those means.

$$F - statistic = \frac{\text{variation between groups}}{\text{variation within groups}}$$

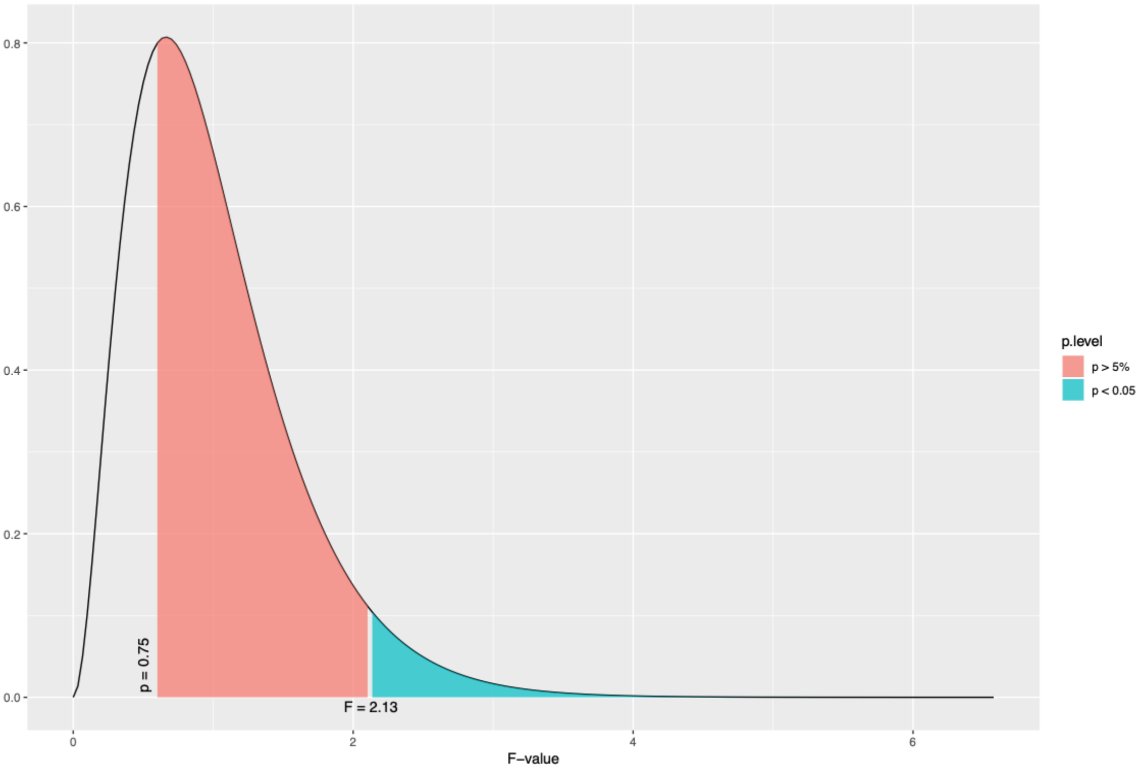
As in all Null Hypothesis Significance Testing (NHST), we establish a Null Hypothesis (H0), and an Alternative Hypothesis (H1).

In this case, our H0 is: the means are equal (there is no difference between groups), and our H1 is: the means are NOT equal.

We use the F-statistic to decide whether or not: 1. to reject the null hypothesis H0 2. to favor the alternative hypothesis H1

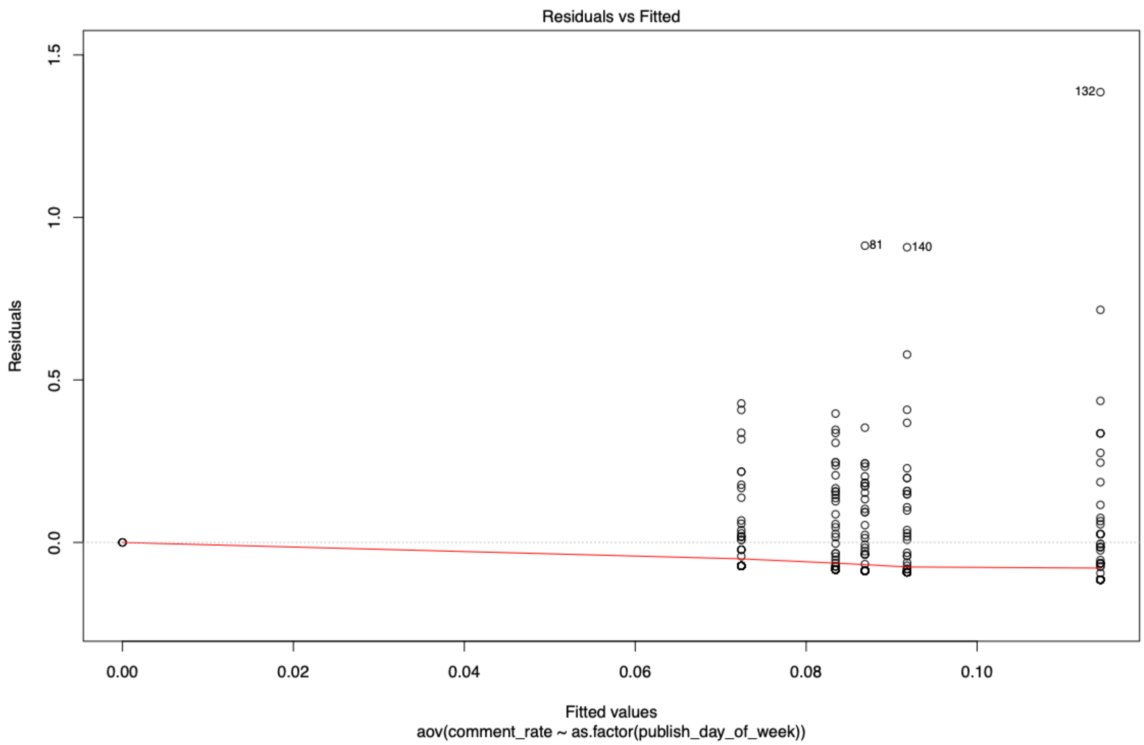
However, if we fail to reject the null hypothesis H0, that does not mean the null hypothesis is true. The F-statistic only assesses whether enough evidence exists to reject the null hypothesis.

After calculating the F-statistic, we want to see if our test case has any statistical significance. If the p-value < 0.05, then we can reject the null hypothesis and we have enough evidence to support the alternative hypothesis. If the p-value > 0.05, then we failed to reject the null hypothesis



```
##                                 Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(publish_day_of_week)   6  0.104 0.01726   0.576  0.749
## Residuals                      315  9.439 0.02996
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = comment_rate ~ as.factor(publish_day_of_week), data = pub_comment_rates_by_pub_Real)
##
## $`as.factor(publish_day_of_week)`
##              diff         lwr        upr     p adj
## 1-0  9.180328e-02 -0.21197836 0.39558491 0.9728790
## 2-0  7.241379e-02 -0.23173589 0.37656348 0.9921542
## 3-0  8.687500e-02 -0.21657271 0.39032271 0.9793672
## 4-0  8.343750e-02 -0.22001021 0.38688521 0.9832317
## 5-0  1.144286e-01 -0.18843627 0.41729341 0.9213083
## 6-0 -2.359224e-15 -0.46892831 0.46892831 1.0000000
## 2-1 -1.938949e-02 -0.11359832 0.07481935 0.9964532
## 3-1 -4.928279e-03 -0.09684556 0.08698900 0.9999986
## 4-1 -8.365779e-03 -0.10028306 0.08355150 0.9999678
## 5-1  2.262529e-02 -0.06734907 0.11259966 0.9894973
## 6-1 -9.180328e-02 -0.46094016 0.27733360 0.9900962
## 3-2  1.446121e-02 -0.07866523 0.10758765 0.9992747
## 4-2  1.102371e-02 -0.08210273 0.10415015 0.9998493
## 5-2  4.201478e-02 -0.04919450 0.13322406 0.8189768
## 6-2 -7.241379e-02 -0.44185361 0.29702603 0.9972944
## 4-3 -3.437500e-03 -0.09424508 0.08737008 0.9999998
## 5-3  2.755357e-02 -0.06128682 0.11639396 0.9691351
## 6-3 -8.687500e-02 -0.45573712 0.28198712 0.9925980
## 5-4  3.099107e-02 -0.05784932 0.11983146 0.9455075
## 6-4 -8.343750e-02 -0.45229962 0.28542462 0.9940494
## 6-5 -1.144286e-01 -0.48281134 0.25395420 0.9688990
```

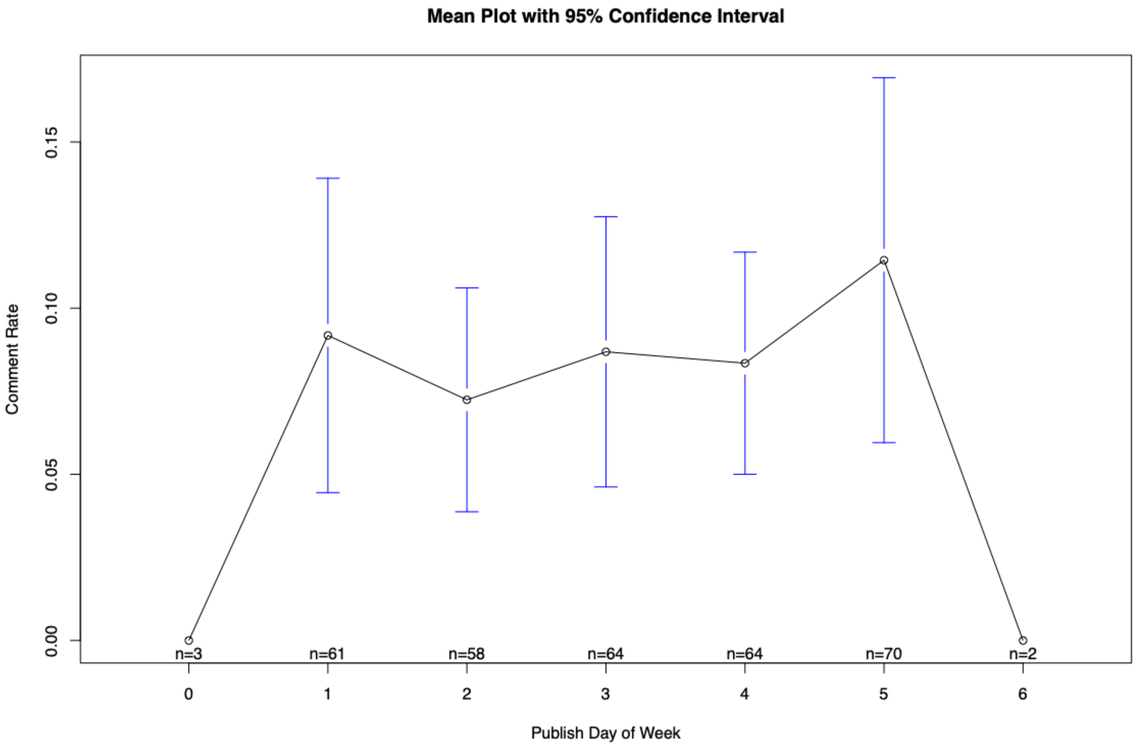


The test failed to reject the null hypothesis, because p is 0.759 > 0.05. Thus we have not established that the means are significantly different. (Again, this is not proof that they are equal (H0).)

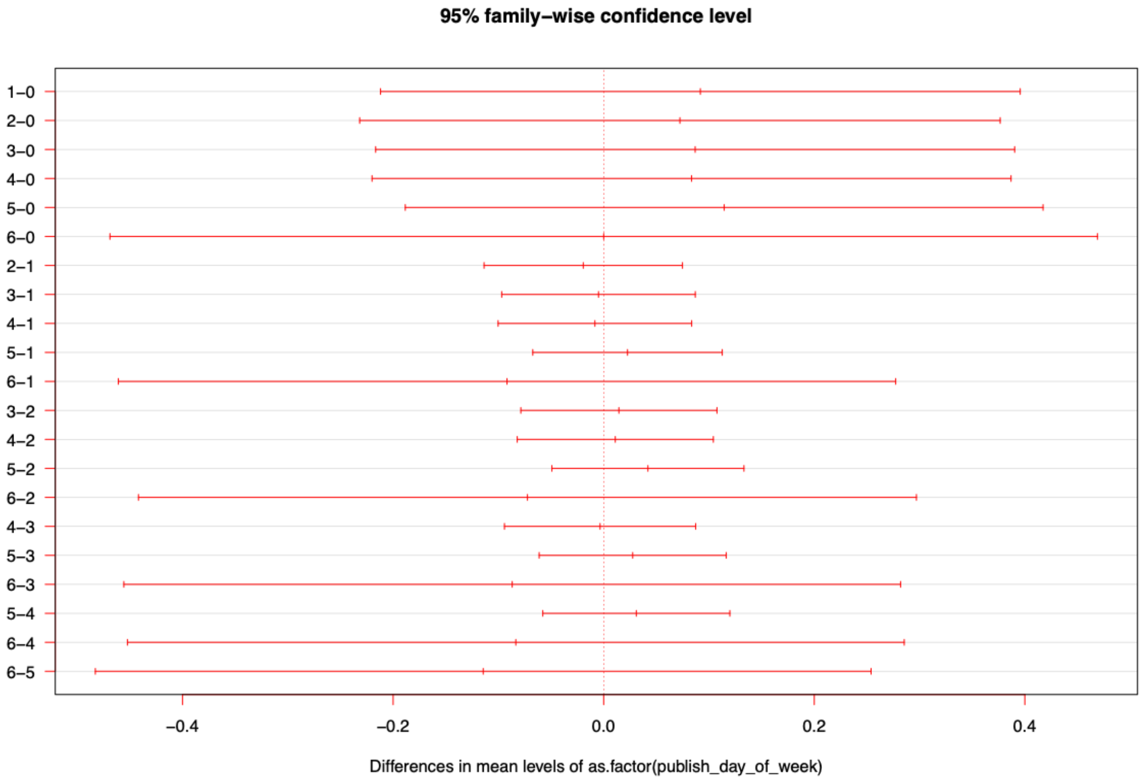
Here are summary statistics of comment_rate by publish_day_of_week - count, mean, standard deviation(sd), and median. This gives us a quick quantitative visual that on Sunday(0) and Saturday(6) have the lowest comment rates.

```
## # A tibble: 7 x 5
##   publish_day_of_week count   mean    sd median
##   <fct>               <int>  <dbl> <dbl>  <dbl>
## 1 0                       3 0      0       0
## 2 1                      61 0.0918 0.185   0
## 3 2                      58 0.0724 0.128   0
## 4 3                      64 0.0869 0.163   0
## 5 4                      64 0.0834 0.134   0
## 6 5                      70 0.114  0.230   0.03
## 7 6                       2 0      0       0
```

This Mean Plot is just to show a simple dispersion of data with 95% confidence interval for each day of the week.

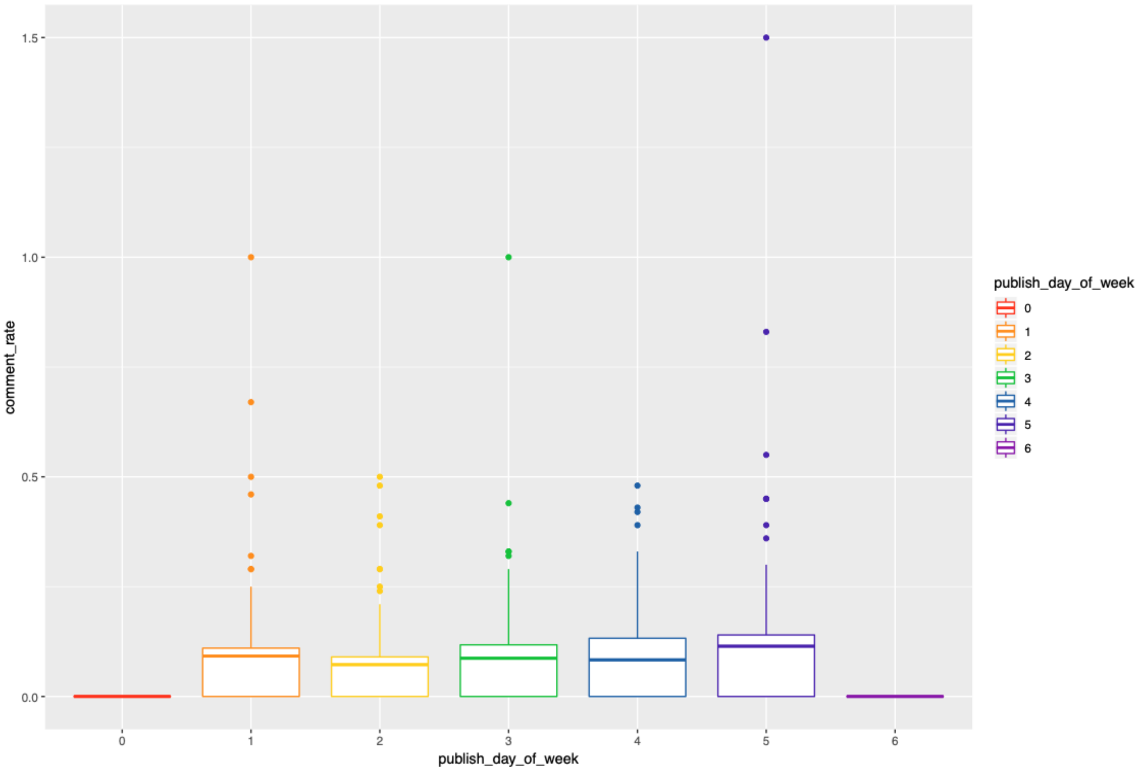


This display is of the Tukey Test, or Tukey's Honest Significance Difference test. The ANOVA with the F-statistic only told us whether or not our results were significant, but does not tell you exactly what the differences were. Since we have our results from the F-statistic we can use Tukey's HSD to find which groups(days) are different by comparing all possible pairs of means.



This box plot shows us the overall patterns of the comment rate by publish day of week. Here is a breakdown of the graph:

- Lower quartile (bottom line of box) is zero for all x-values
  - This means that 25% of the data is less than or equal to this value
- By default the line is the median, but was changed to be the mean to show more value
- Upper quartile (top line of box) varies between the days of the week
  - This means that 25% of data is greater than or equal to this value
- Outliers(dots) more than 3/2 times the upper quartile


comment_rate
publish_day_of_week
publish_day_of_week