Assignment 3a: Understanding Hypotheses

Adrianne Avila-Mangual
Purdue University
SCLA 51000-001: Data and AI Storytelling
Sorin Adam Matei, PhD
November 9, 2025

Assignment 3a: Understanding Hypotheses

1. Hypotheses

The following are my null hypothesis and my alternative hypothesis:

- H_0: There is no relationship between humor and the number of likes per post on our company's social media.
- H_A: There is a relationship between humor and the number of likes per post on our company's social media.

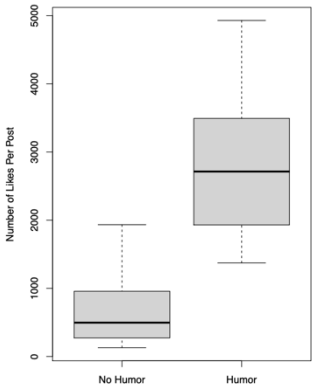

Figure 1:Boxplot of Likes Per Post by Humor. This was given in the assignment document.

I chose the null hypothesis to further explore the relationship between the two variables: humor and number of likes. When studying Figure 1, it seems as though the number of likes increases with humor. This is a surface level investigation, and the hypothesis would need to be tested to see if we accept or reject the null hypothesis.

2. Why are you seeing (or not seeing) relationships between these two variables?

The two variables we are studying are humor (categorical) and likes (numerical). After reviewing the dataset and Figure 1 (both given in the assignment), it appears that posts with humor receive more likes implying that humor increases online engagement. This insinuates a positive relationship between humor and likes. Furthermore, there could potentially be a relationship between humor and engagement. First steps would include testing for correlation and/or a causal relationship. The potential of third variable confounding the relationship would need to be investigated. If it is found that the variables are correlated and indeed causally related, then more research on "why" humor receives more likes therefore showing

increased engagement to find other ways for our company to increase social media engagement.

3. **How do we know whether any relationship is strong and should be trusted (in other words, what evidence can you provide that your hypothesis is correct)? For example, can you obtain a correlation coefficient, create a plot, or provide some other kind of analysis?**

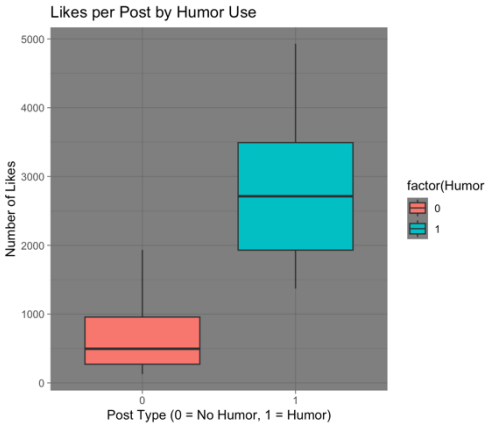

Figure 3: Recreated boxplot to check for accuracy after cleaning data. Data included an additional line.

I am an advocate for interacting with the data before drawing and conclusions; therefore, I started by recreating a boxplot with the given data although a boxplot was given as seen in Figure 3. The next step was to examine the relationship through paired histograms. As illustrated in Figure 2, it was found that the humorous posts cluster at higher engagement levels, whereas the posts lacking humor cluster at much lower counts. In fact, there is very little crossover between the two histograms indicating that the difference between the two (humor/no humor) is both strong and consistent.

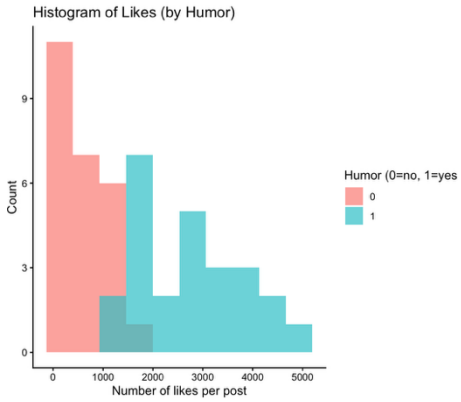

Figure 2: The Histograms show the distribution of likes for posts with humor and without humor.

After having the opportunity to visually explore the data, it was time to calculate the measures of central tendency. It was noted that not only were there more likes for humorous posts, but also the mean for humor post likes was more than four times the mean for none humor post likes. The measures are illustrated in Figure 4 and Figure 5. The next step was to look at the correlation

through the Pearson correlation as well as a scatterplot of the data. The graph showed two distinct clusters. The lower cluster (left) represents the posts without humor and the higher cluster (right) represents the post with humor. The upward slope signifies a positive relationship between humor and likes. The thinner shaded confidence band implies that we are fairly confident in this positive relationship. The correlation was found to be very strong positive correlation (r = 0.8107). Figure 6 visualizes the correlation between humor and likes.

Humor (0 = No humor, 1 = …[1]	count	mean_likes	median_likes	sd_likes	var_likes
<dbl>	<int>	<dbl>	<dbl>	<dbl>	<dbl>
0	25	628.	496	464.	215013.
1	25	2782.	2713	1023.	1045815.

Figure 4: Measures of Central Tendency

Humor (0 = No humor, 1 =…[1]	min_likes	lower_quartile	upper_quartile	max_likes
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
0	129	272	957	1933
1	1372	1930	3492	4929

Figure 5: Data (by humor/no humor) in Quartiles

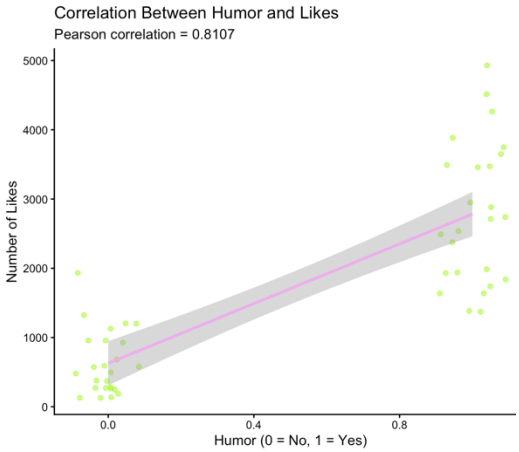

Figure 6: This scatterplot shows a strong positive relationship (r = 0.81) between the use of humor in social media posts and the number of likes each post received.

Finally, I conducted a two sample t-test to compare the means of the two groups as the t-test is used to in scenarios where the measurements of two groups have a suspected link to each other (Wadhwa & Marappa-Ganeshan, 2023). The t-value is used to measure the difference between humor and no humor relative to the variation within the dataset. The results of the Welch Two Sample t-test are shown in Figure 7. The data shows t = -9.5935. When looking at the absolute value of -9.5935, it is plain to see that the absolute value has a large magnitude. In fact, the value of 9 is exceptionally strong and evidence that the post with humor and without humor differ far beyond what

```
        Welch Two Sample t-test

data:  Likes by Humor
t = -9.5935, df = 33.468, p-value = 3.923e-11
alternative hypothesis: true difference in means between
group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -2611.095 -1697.785
sample estimates:
mean in group 0 mean in group 1
         627.92         2782.36
```

Figure 7: Humorous posts (M = 2782.36) received significantly more likes than non-humorous posts (M = 627.92), t(33.47) = –9.59, p < .001, 95% CI [–2611.10, –1697.79].

would occur by a random chance/variation. The p-value is much smaller than the rule of thumb ($p < 0.05$) which makes this finding statistically significant.

4. **Can you draw on course concepts to better understand any relationships you see?**

In conclusion, it was found that there is a statistically significant, strong positive relationship between humor and the amount of like, hence an increase in social media engagement. However, it would be remiss of me not to state that this is a correlation and "correlation does not imply causation". We only studied one dimension of the posts. We did not consider, however, who was posting, what time the posts were made, were there photos/visuals, what are the mean age groups of the social media consumers, what is the mean age of our consumer, or what were the analytics of who saw the posts?

One article in this unit discussed how Presidents Trumps social media was analyzed and it was shown how different the tone was from one tweet author to the other (Robinson, 2016; Wadhwa & Marappa-Ganeshan, 2023). The scenario with our social media could be similar in that the humorous author has a more positive overtone than our non-humorous author. As the assignment does not discuss who makes the post, I cannot rule out that there is a third variable at play. I could look at the counterfactual of if the social media post wasn't humorous, it would not have been liked this could be tested against to test causality. As a future data scientist, I would recommend using humor as a part of a robust, data-driven strategy to boost engagement; however, I would want more data so see if there are any more potential *whats* as well as the *why* behind the humor relationship to engagement.

References

Kelleher, A. (2018, January 5). *If correlation doesn't imply causation, then what does?* Medium. https://medium.com/causal-data-science/if-correlation-doesnt-imply-causation-then-what-does-c74f20d26438

Levy, J. S. (2015). Counterfactuals, causal inference, and historical analysis. *Security Studies*, *24*(3), 378–402. https://doi.org/10.1080/09636412.2015.1070602

Robinson, D. (2016, August 9). *Text analysis of Trump's tweets confirms he writes only the (angrier) Android half*. Variance Explained. http://varianceexplained.org/r/trump-tweets/

Silver, N. (2013). *The signal and the noise: why most predictions fail, but some succeed*. Penguin Books.

Team Geckoboard. (2021, July 12). *How to analyze data: A basic guide*. Geckoboard Blog. https://www.geckoboard.com/blog/how-to-analyze-data/

Wadhwa, R. R., & Marappa-Ganeshan, R. (2023). *T test*. PubMed; StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK553048/

Wheelan, C. (2014). *Naked statistics: stripping the dread from the data*. W.W. Norton & Company Ltd.

Yu, V. (2020, December 14). *How to tell a story with data | towards data science*. Towards Data Science. https://towardsdatascience.com/how-to-a-tell-story-with-data-3200bfadce6d/