Tests for Normality and Hypothesis Testing

Brayden Yates

University of the Cumberlands

MSDS-530: Fundamentals of Data Science

Dr. Jason Turner

10 February 2024

Tests for Normality and Hypothesis Testing

Univariate tests, as implied by the prefix uni- identify whether a single variable follows a normal distribution. This is critical, because parametric tests are founded in the assumption that the data is normalized. Parametric tests also assume that the dataset is homogenious. These types of tests are used to compare proportions between various groups. Hypothesis testing is used to make inferences about population parameters based on the data that is provided. In order to this, you must use the sample data to determine whether the evidence is strong enough to reject the *null hypothesis*—the assumption we begin with that the hypothesis is false.

For this dataset, I chose the Fast Food Nutrition Dataset, which is a collection of nutritional facts regarding various meals at a variety of fast food locations, such as McDonald's, Burger King, Wendy's, KFC, Taco Bell, and Pizza Hut. I selected this dataset because I believe it is of interest and relevance to the average consumer.

Here is the code:

```
import pandas as pd
from scipy.stats import f_oneway
df = pd.read_csv('FastFoodNutritionMenuV3.csv')
df['Calories'] = pd.to_numeric(df['Calories'], errors='coerce')
df.dropna(subset=['Calories'], inplace=True)
anova_result = f_oneway(
    df[df['Company'] == 'McDonald's']['Calories'],
    df[df['Company'] == 'KFC']['Calories'],
```

```
    df[df['Company'] == 'Burger King']['Calories'],
    df[df['Company'] == 'Taco Bell']['Calories'],
    df[df['Company'] == 'Wendy's']['Calories']
)
print("ANOVA F-Statistic:", anova_result.statistic)
print("ANOVA p-value:", anova_result.pvalue)
```

```
ANOVA F-Statistic: 11.180960013457812
ANOVA p-value: 6.918084671576456e-09
```

For my Univariate Analysis, I used the Shapiro-Wilk test. Here is the code I used:

```
shapiro_test_statistic, shapiro_p_value = shapiro(df['Calories'])
print("Test Statistic:", shapiro_test_statistic)
print("p-value:", shapiro_p_value)
```

```
Test Statistic: 0.9165760656989711
p-value: 1.71061953890427e-24
```

After running these analyses, we can gather some information regarding this dataset. With the ANOVA Test Results, our null hypothesis was that all of the fast food chains in the dataset offer a similar number of calories per item on the menu. With an F-statistic of 11.18 and a p-value of 6.92*10^-9, there is a significant difference in calorie content among all of these fast food chains. Because of our hypothesis, our null hypothesis was that the calories are all equal regardless of fast food chain choice. We are able to reject the null hypothesis based on this analysis.

With our univariate analysis, we used a Shapiro-Wilk test. This test produced a test statistic of 0.91 and a p-value of 1.71 * 10^-24. With these values, we are able to reject the null hypothesis—the data is *not* normally distributed.

When we synthesize these results, we can determine that, while the ANOVA test confirmed significant differences among fast food companies' caloric density, the data itself is not normally distributed. This means that ANOVA's assumption—the dataset is normally distributed—is not fully met. This could have a major impact on the ANOVA results, so the statistic provided by the dataset may in fact be inaccurate.

References

Joakim Arvidsson. (n.d.). *Fast Food Nutrition*. Kaggle.

https://www.kaggle.com/datasets/joebeachcapital/fast-food