May 9, 2020

Alexander Davies and Max Nadeau

Math 23c

A Glass Half Full: Analyzing a Dataset of Wine Reviews

For our final project, we used the techniques learned in Math 23c to perform statistical analyses on a dataset of about 130,000 reviews of wine. The data came from Kaggle, where a user named zackthoutt had posted the results of scraping the website WineEnthusiast. Each row in the data frame has a points rating, ostensibly out of 100 but actually confined to between 80 and 100, a price in dollars, a region of origin, and the type of wine, among other columns. The points and price acted as our quantitative variables, while region of origin and variety functioned as categorical variables.

We began with exploratory data analysis and visualizations. First, we made a barplot of the top 10 regions of origin for the wines in the dataset. California, with about 35,000 wines, was far ahead of all the rest. Washington State, in second place, had less than 10,000 wines. Outside of the United States, Bordeaux, Tuscany, and Northern Spain produced the most wine. Next, we generated a violin plot with the ggplot2 package to see how points rating varied by region of origin. The distributions were very similar, but the median rating for Bordeaux wines was, surprisingly, less than the other high-production regions. The overall distribution of wine ratings was roughly symmetrical, slightly skewed right and centered on around 88 points.

We then performed a permutation test to see if the difference in the average rating of wines from Southwest France, around 88.6 points, was statistically significant from the population mean, 88.4. Somewhat surprisingly, the permutation test produced a p-value of about 0.01, which made the difference statistically significant. As the CLT implies that the sampling distribution should be approximately normal, we confirmed this p-value with pnorm, the normal cdf, which also produced a p-value of around 0.01. Finally, we generated a 95% confidence interval for the average rating of wines from SW France, 88.5 to 88.8, which does not include the population mean.

Next, we analyzed the contingency table between two binary variables, whether the wine was Californian and whether it was a Chardonnay, to assess our hypothesis that Chardonnays

were disproportionately Californian. A permutation test and a Fisher exact test both produced p-values of essentially zero, confirming our hunch.

Finally, we regressed point ratings against prices to see if more expensive ratings were given higher prices. Using a linear model, the price of wine is positively correlated with the rating but only explains 17% of the variation in price, indicating that many factors besides price impact the rating of a wine.