Crunch! Potato Chips!

Summary

This project aims to investigate several count datasets of potato chip consumptions in the U.S. The data is supplied by Simmons Research LLC, whose national survey continuously samples 25,000 U.S. adults for lifestyle preferences. While the initial motivator was to impose and analyze NBD models on several count datasets, including the combined chip consumption distribution and individual brand consumption distribution, the paper also discussed the shortcomings of having binned datasets and the implications of such data.

Introduction & Background

As popular and common as they are today, potato chips have a mysterious origin. Some says that it was cook George Crum who invented them back in the 1850s, when one of his customers complained that his fries were "too thick" and wanted them to be thinly sliced potato. Regardless of how it started, in 1920s, Herman Lay took this rising American potato addiction to a national level, eventually "laying" the foundation of *Lay's* that we love today. Their slogan, "betcha can't just eat one!" perfectly demonstrates the company's confidence in their chip production, as well as their dominance in the snack industry, accounting for 60% of the salty snack market in the U.S.

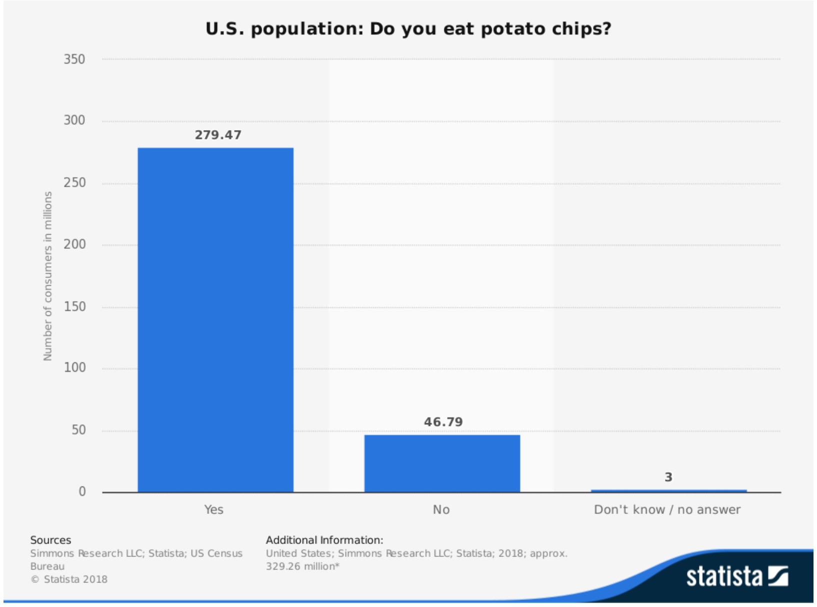
(Source: https://www.thoughtco.com/history-of-potato-chips-1991777)

Just how much chips do Americans consume? How long would it take for over 80% of the U.S. population to consume more than 8 bags of chips? Is Frito-Lay a much more preferred chip brand in comparison to others? Motivated by these questions, this project aims to use the NBD to model consumption distribution of potato chips in the U.S., as well as establish cross comparisons between different brands of potato chips.

Data Selection

Using *Statista* as my data platform, I found five sets of data for further investigation. All of the data used were provided by Simmons Research LLC and the US Census Bureau, recorded in 2018.

The first dataset is very straightforward: ***Do you eat chips?*** The data is used to calculate the percentage of Americans who never consume chips (14.21%), which then serves as a good proxy for the probability of having a Hard Core Never Buyer (HCNB) of chips used in later NBD models. (URL: https://www.statista.com/statistics/277158/us-households-consumption-of-potato-chips/)

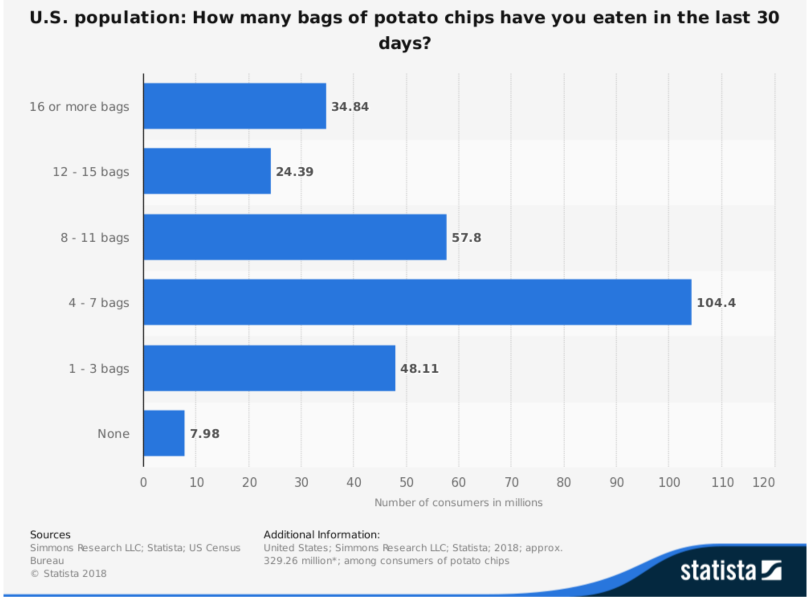

U.S. population: Do you eat potato chips?

Sources
Simmons Research LLC; Statista; US Census Bureau
© Statista 2018

Additional Information:
United States; Simmons Research LLC; Statista; 2018; approx. 329.26 million*

The second data is a distribution of potato chip consumption over the course of one month. The unit of measurement is by bags of chips consumed per person. A standard bag of chip used is 8 ounce in weight.

(URL: https://www.statista.com/statistics/277190/us-households-amount-of-potato-chips-eaten-within-30-days/)

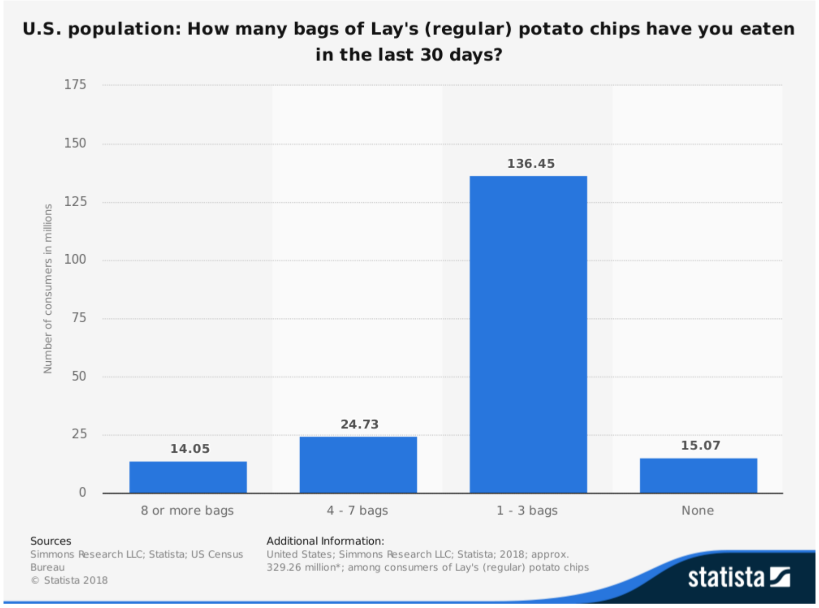

U.S. population: How many bags of potato chips have you eaten in the last 30 days?

Sources
Simmons Research LLC; Statista; US Census Bureau
© Statista 2018

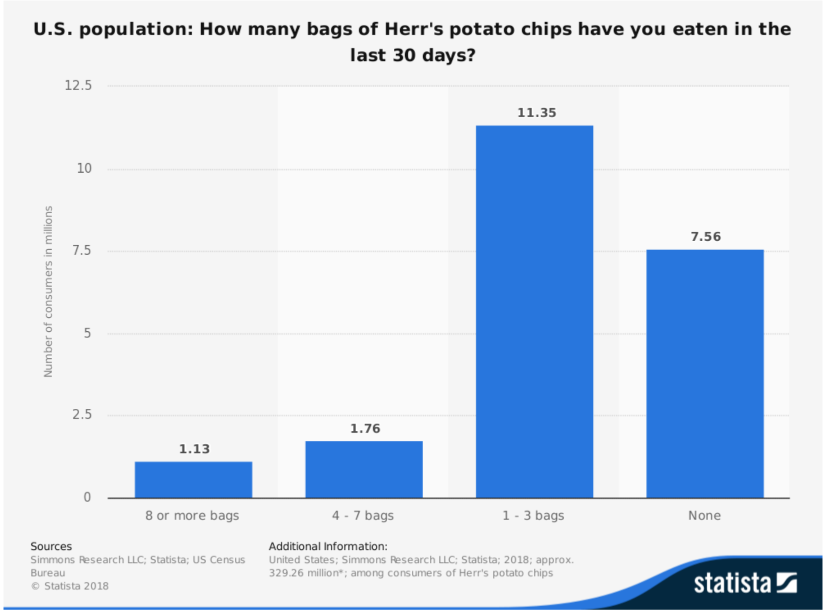
Additional Information:
United States; Simmons Research LLC; Statista; 2018; approx. 329.26 million*; among consumers of potato chips

Then we have a series of data accounting for monthly potato chip consumption by specific brands: Lay’s (regular), Herr’s, and Kettle.

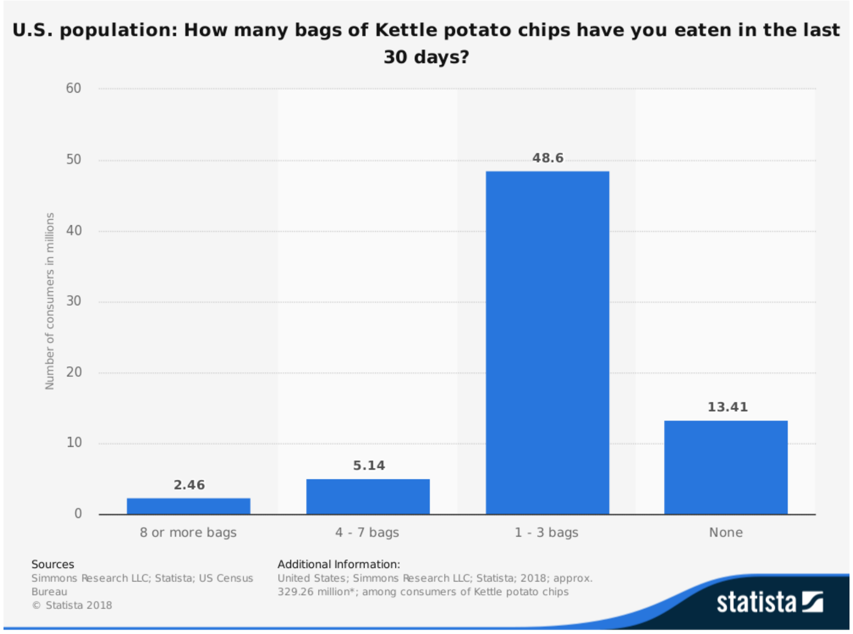
(URL: https://www.statista.com/statistics/289264/bags-of-lay-s-regular-potato-chips-eaten-in-the-us/)



(URL: https://www.statista.com/statistics/289259/bags-of-herr-s-potato-chips-eaten-in-the-us/)



(URL: https://www.statista.com/statistics/933533/bags-of-kettle-potato-chips-eaten-usa/)


U.S. population: How many bags of Kettle potato chips have you eaten in the last 30 days?

Sources
Simmons Research LLC; Statista; US Census Bureau
© Statista 2018

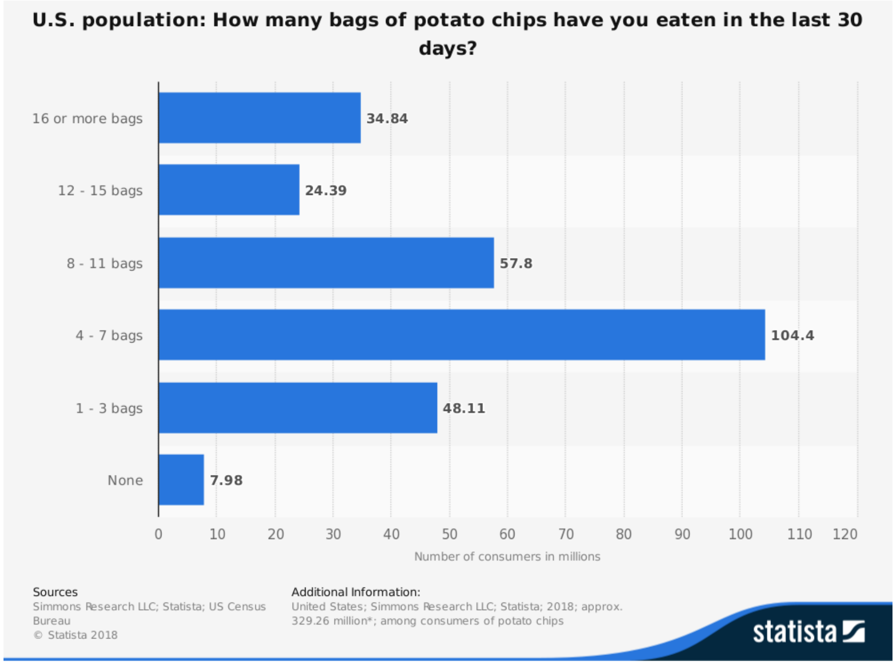
Additional Information:
United States; Simmons Research LLC; Statista; 2018; approx. 329.26 million*; among consumers of Kettle potato chips

Before imposing any models, it is essential to address several major concerns and limitations on using these sets of data. As seen through histograms above, the data supplied are NOT raw sampled data, but rather adjusted proportional data that reflect a certain population (e.g. U.S. population, all consumers of chips, all consumers of a specific brand of chips). The true sample, in accordance with Simmons Research LLC methodology website, is a random sample of U.S. adults with a size of 25,000 individuals. Since the original count distributions are not available, in order to avoid possible erroneous inferences made with numerical count values, all inferences in this project will be drawn from proportions.

A second concern is the usage of binned data. Since the survey calls for the number of bags of chips consumed over the past 30 days, it is hard to record such number in precise increments. The models use the combined probability of falling into the range of values assorted by bins. The risk is that the model might not fit as well with binned data than un-binned data, since there are essentially fewer rows of data when values are binned.

A minor concern, though not particularly influential to how one treats the actual data, is the validity of the original data. As discussed in class, it is hard to collect certain counting data over a monthly period, since people tend to lose track and/or do not have enough incentive to do so. As Simmons Research LLC's methodology uses home surveys, a 30-day chip consumption recall might be highly inaccurate. It is suggested that the data be measured in a weekly fashion for higher accuracy.

Analysis 1: potato chip consumption of all brands



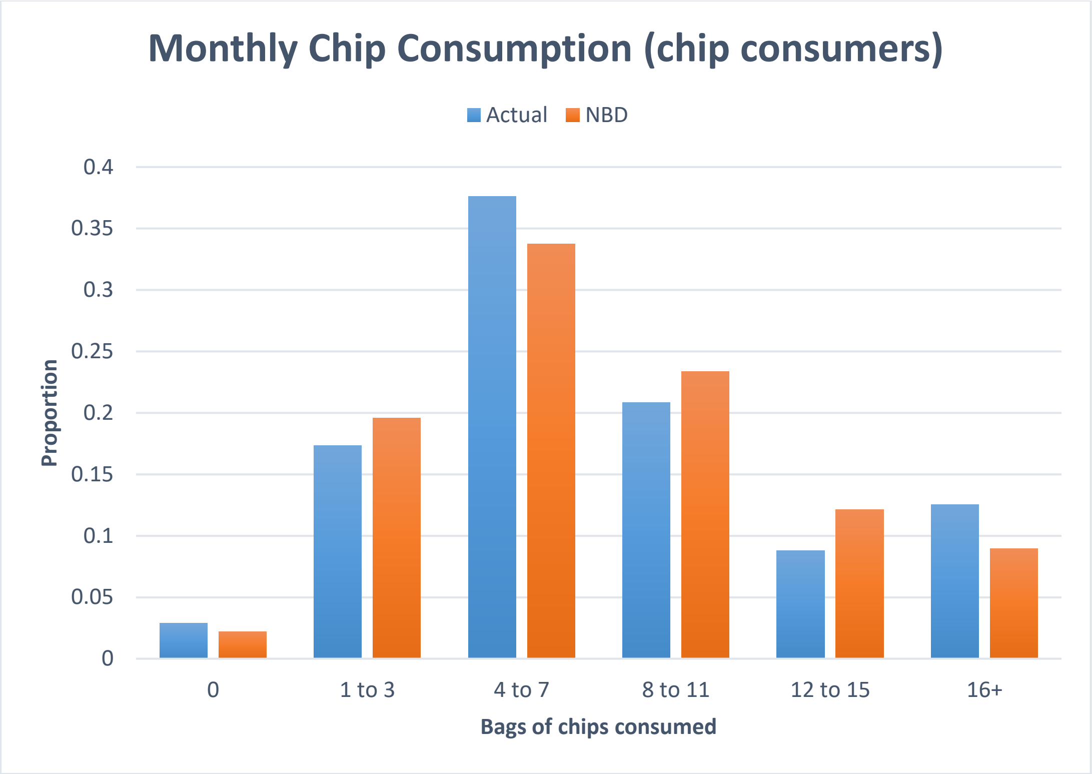
Sources
Simmons Research LLC; Statista; US Census Bureau
© Statista 2018

Additional Information:
United States; Simmons Research LLC; Statista; 2018; approx. 329.26 million*; among consumers of potato chips

Treating the binned X values with a combined probability, I imposed an NBD model on the monthly consumption data for chips. The underlying story behind using this model is to have every chip consumer spin their random wheel with a specific lambda at the beginning of each month, from which they decide how many bags of chips to consume over the course of one month. In this model, no spike was introduced. This is because the data only accounts for the population of chip consumers. Hence, there are no “Hard Core Never Buyers”. When using solver, instead of using LL to maximize likelihood, I chose to minimize the sum of squared errors between each pair of expected probability and actual probability instead. As explained previously, due to the nature of the dataset being a proportional data adjusted to the population rather than a raw data, using LL may not be appropriate since it relies on the raw counts. Using square errors, however, one may utilize only the proportions, which remains constant despite the scaling of the original data.

Looking at the original data distribution, I predict that the value of r would be greater than 1 due to the data’s relatively low heterogeneity, having a spike in the middle of the distribution. While the value of alpha is often not deemed critical, it should be below 1 since the mean of the distribution appears to be near 7.

The parameters of the model, as well as a visualization of the model, are as follows:



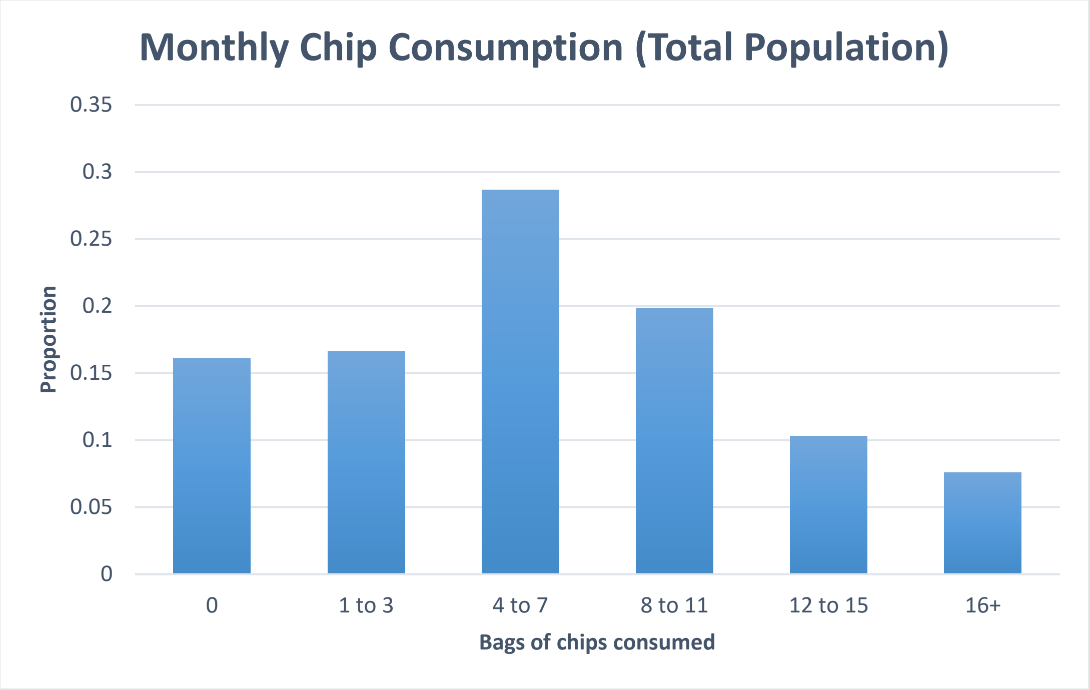
r	α	sq error	mean
2.927	0.374	0.005	7.822

As predicted, the model returns a high r, a low alpha, and a mean (7.822) near 7.

Here I am graphing the probability of occurrence, rather than the frequency, on the y-axis. The same reason as before: the data is an adjusted data, whose counts do not contain physical meaning besides the percentage they occupy in the population.

By observation, the NBD model is an "okay" fit of the actual data. If one performs a x^2goodness-of-fit test using the given data and expected counts, p > 0.001. However, due to the lack of physical significance of the adjusted data, x^2 test statistic should not be utilized as a method to evaluate how good the model is.

After fitting the model for chip consumers, a natural question I had was to construct a model that accounts for the entire U.S. population (including those who do not eat chips). This is equivalent of manually adding a spike at zero. To do this, I utilized a second set of data – ***Do You Eat Chips?*** This gives me a proxy for the proportion of chip consumers as well as the proportion of Hard Core Never Buyers (p). Rescaling the probability distribution from the previous model using these two proportions, adding a spike at zero to account for HCNB, the monthly chip consumption of all U.S. population can be visualized as the following:



With the complete data that reflects the entire U.S. population, we can now answer a lot of questions that might interest us.

Question 1:

How long would it take for 80% of the U.S. population to consume at least 8 bags of chips?

Solution:

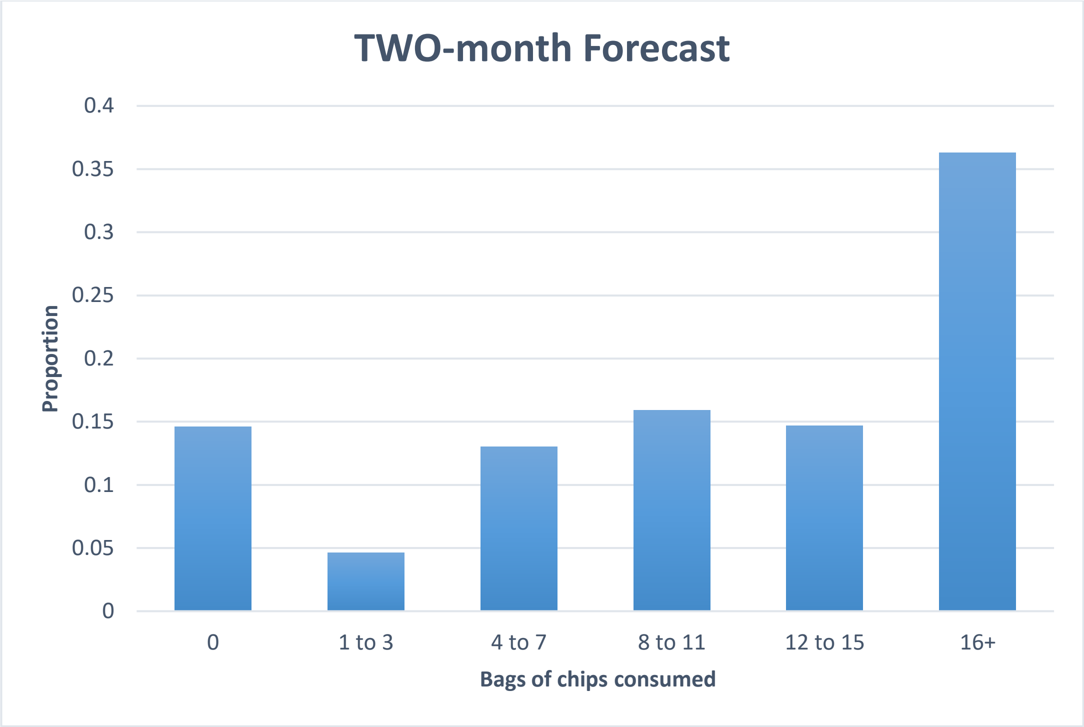
Summing up the total probability of the right tail (X > 7) and using solver to solve for time t such that P (X > 7) = 80%, it takes about 3.85 months (117 days).

Questions 2:

What would the chip consumption distribution look like for a two-month period?

Solution:

By changing the time duration to t=2, where t is measured in months, the distribution is as follows.



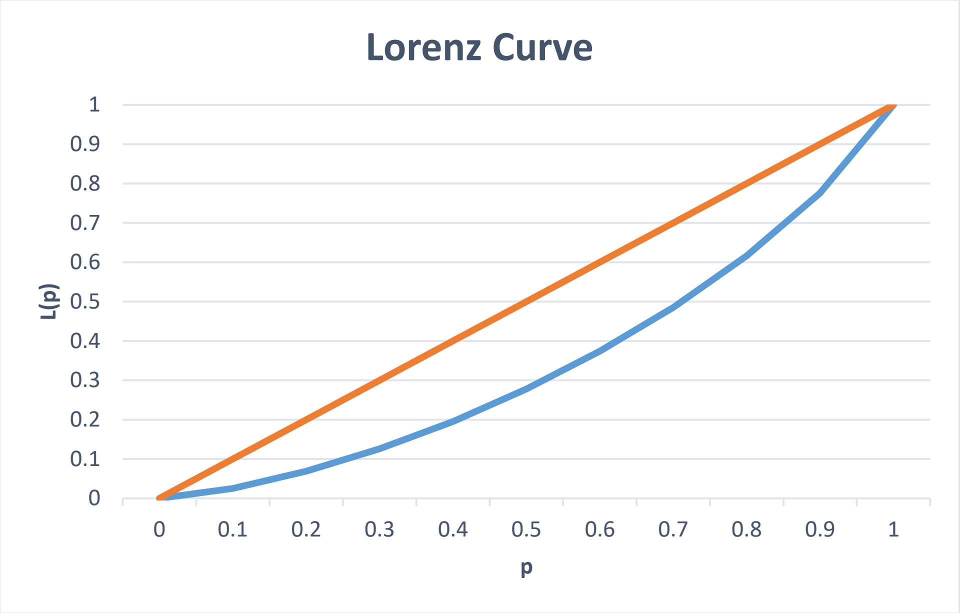
Please notice that as the right tail becomes more concentrated (i.e. more people consume 16+ bags), it becomes increasing less meaningful to produce forecasts by increasing the time duration. While it is possible to use the NBD to predict the probability distribution where X is beyond the current maximum value, it is difficult to produce accurate inferences for this dataset. The extrapolation power of this model is low mainly due to the original dataset being already binned with limited data points. It is suggested to use this NBD model to only produce forecasts within 3 months.

Question 3:

Does the consumption of potato chips follow the 80:20 rule?

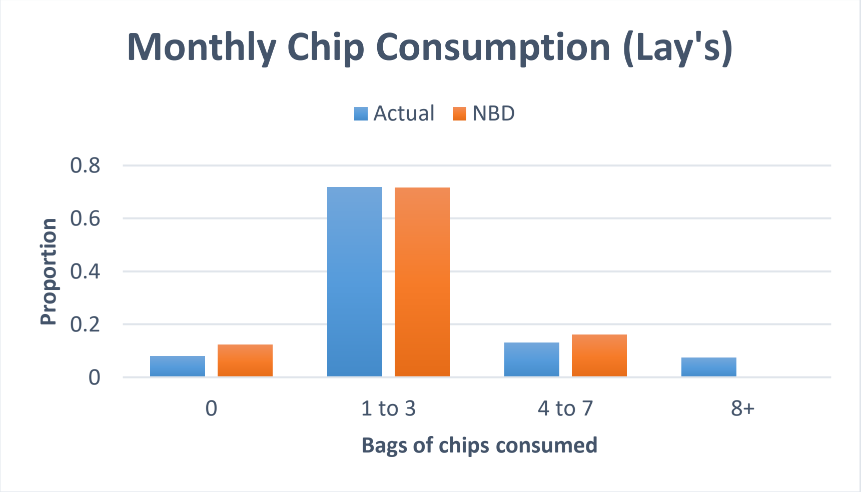
Solution:

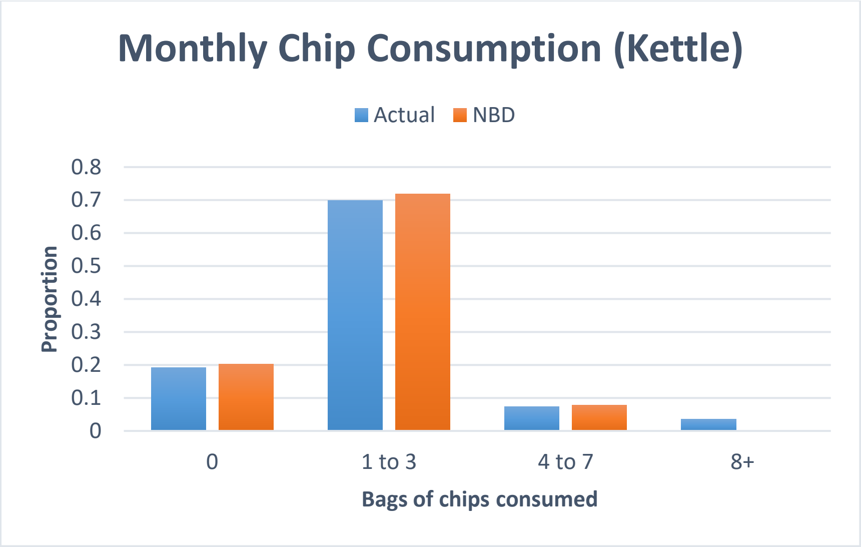
Computing and graphing the Lorenz Curve, one sees that the distribution does not follow the 80:20 rule. In fact, the top 20% only accounts for about 40% of consumption.

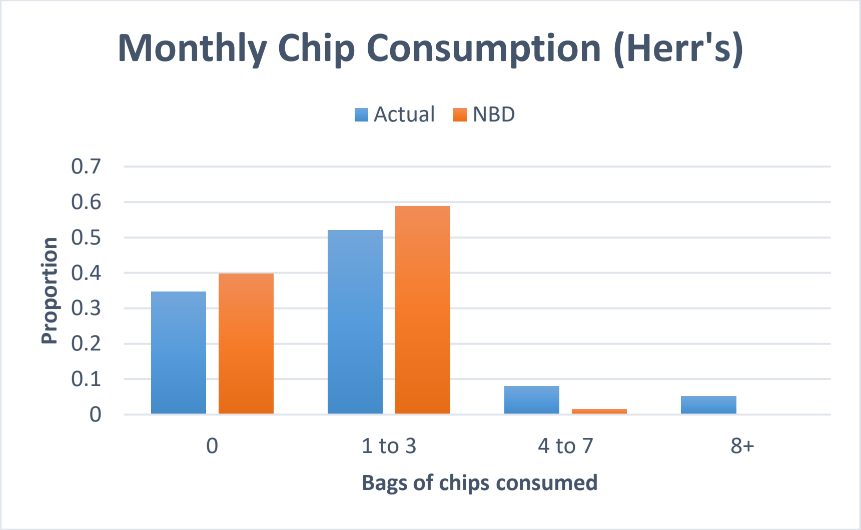


Analysis 2: Lay’s vs. Kettle vs. Herr’s – A Major Shortcoming of Binned Data

Now we have modeled the combined chip consumption distribution, we might be curious about the distributions of specific chip brands. Using the same method as the previous model (i.e. attaching no spikes, minimizing the least squared error to find the parameters, and graphing proportions rather than frequency), here are the graphs of the three brands:







While the NBD models imposed on these three datasets appear to fit well, one could immediately spot a (huge) concern once they compare the parameters of these models.

brand	r	alpha	sq error
Lay's	15365.51	7308.97	0.008004
Kettle	13150.76	8227.946	0.001783
Herr's	8511.006	9218.626	0.014153

When the data is clustered at the center, and contain only a few numbers of bins, the NBD model returns extremely high values for r and alpha. This is a major shortcoming of having binned data, as well as having a low number of rows of data.

Look at the chip consumption distribution for Lay's. Because the data is binned, the most probable values for X (e.g. 1, 2, 3) are grouped together into a single spike that accounts for more than 70% of the data. Recall that the parameter r measures heterogeneity, and a high value of r means that the data is more homogeneous. In this distribution, the combined "spike" is so significantly larger than the rest that the distribution becomes almost homogeneous centered at the spike, causing the r value to go up drastically. Similarly, the NBD model for Kettle and Herr's also returned large values of r. In order to maintain a constant mean, as r goes up, so does alpha, explaining why both r and alpha have extreme values.

Going from a narrative perspective, one can also see why consumption among individual brands tends to be homogeneous. Speaking only from personal experience, since the tastes of chips do not variate much among different brands, one might expect to see a lower brand preference. The chance that an individual only purchases chips from a specific brand, rather than picking what's cheap or what's available on the shelf, is relatively low. This results in less heterogeneity among a specific brand's consumers. Due to homogeneity, the Lorenz curves of the models are close to straight lines passing through the origin at a 45-degree angle.

Notice that even though the binned data resulted in extreme parameters among specific brand models, it worked fairly well for the combined chip consumption distribution (analysis 1). This is because the combined data has more rows of data (more bins) as well as more spread-out proportions among each bin, which makes the data more heterogeneous. From a narrative perspective, while there might not be a specific brand preference, those who enjoy chips will consume a greater amount of chips than those who do not enjoy chips as much. This contributes to heterogeneity as seen in the combined data.

Future Remarks

While the NBD model can be used to analyze a variety of counting datasets, one should be careful when treating datasets with either too few rows of data or datasets that consist of binned values. As the NBD seeks to capture heterogeneity, having bins that increase the data concentration over a specific range of values may defy this purpose. On another note, should the model be accurate, it only takes less than 4 months for 80% of the U.S. population (that's 263 million!) to consume at least 8 bags of potato chips each! That's a lot of junk food for anyone.