The Law of Large Numbers – Product Reviews & e-Commerce

(Bonus Project 1)

The law of large numbers says that the relative frequency of a certain event occurring in an experiment converges to the value of its theoretical probability as the experiment is repeated a large number of times. This idea can be applied to many real world examples wherever probability is involved, whether it is a coin toss or Russian Roulette. Specifically, we will be looking at the way the law of large numbers applies to product reviews on e-commerce sites like Amazon.com.

If you are a savvy online shopper, you will most likely hesitate before buying a product that has few customer reviews, or none at all. You may not realize it, but you are looking for products where the law of large numbers has already played out with respect to the reviews before you will consider making a purchase. Everyone knows that a product with no reviews carries a risk of being a disappointment. They also know that a few customer reviews of a product here and there don't accurately represent the quality of the product or customer service as a whole, whether the reviews paint the product or company in a positive or a negative light. The experience of a few people cannot accurately determine the experience that most other customers will have with a particular product. This is where the law of large numbers comes into play, where the distribution of 1-5 star product reviews begins to more accurately represent the product in question as more and more reviews are added. We can use our computer simulation to model this phenomenon.

The simulation written for this project uses a set of theoretical probabilities for 1-5 star product reviews, which are given either by the user or by a default set of values. The default values given for these probabilities represent what could be considered a typical product review distribution, with 5 star reviews having the highest frequency, followed by 4, then 3, then 2, then 1 star reviews. If the user wanted to use their own review distribution, these probabilities could be based off of any real world product example (or an imaginary one) and entered into the simulation.

The number of reviews to be generated for the simulation is determined in the same way: the default number of reviews to be generated is specified by the simulation, but the user can enter any other valid number in its place. When the simulation is run, weighted, random ratings are generated based on the given set of theoretical probabilities. The results – the relative frequencies of 1-5 star reviews for each added review, and the accumulated total for each star review – are written to files that can be imported into Excel and graphed. The code for the simulation is attached at the end of this report.

To generate the graphs and charts you will see later, the computer simulation was run four times – once with 12 reviews generated, once with 100 reviews generated, once with 1000 reviews generated, and once with 5000 reviews generated. Because the sample size was so small for the 12 review run, the graph of relative frequencies was all over the place and in no way resembled the given theoretical probabilities for 1-5 star reviews. The 100 review run had somewhat better results, with each star review's relative frequency smoothing out as the sample size became larger and each added review had less of an impact on the relative frequency. The 1000 review run saw the relative frequencies converging fairly close to their theoretical probabilities, and the 5000 review run brought those frequencies even closer to their theoretical values. These results are shown in the charts and graphs below.

1st Run

2nd Run

3rd Run

4th Run

Screenshot of a Sample Run

