MAT-335 Project 2

How to tell Normal or Laplace?

Ziwen Chen & Hongyuan Zhang

December 15, 2018

1 Introduction

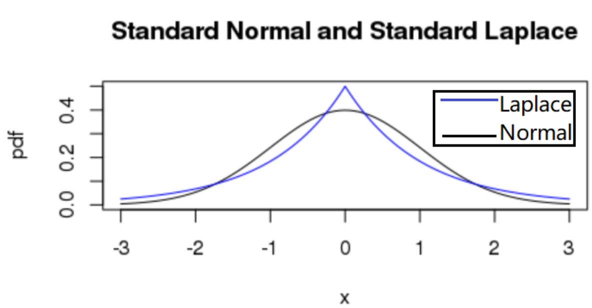
Historically, there has been a long-term problem in stock market that the Normal distribution is not a good predictor for large up and down days in the market. Using the Normal distribution constructed by historical stock price changes, the model predicts that the probability of getting a 5% change in S&P 500 index is 0.0001 in 67 years. However, in reality this event happened twice in 67 years (Harwood, 2018). This aberration indicates that the tail of the distribution of data set is much heavier than the Normal distribution model. People have found that Laplace might be a better model to fit the data (Harwood, 2018). Intrigued by this problem, in this project, we explore methods to facilitate researchers to decide whether a given sample of data comes from a Normal distribution or a Laplace distribution, given that the sample is from one of these two distributions. The results of our projects can help researchers to avoid the mistake similar to the one in stock market mentioned above.

In this report, we will introduce two indicators that can be used to determine whether the data is generated from a Normal or Laplace distribution. The first one is Kurtosis and the second one is Residual Standard Deviation. The first indicator is accurate for samples of 100 or more observations, whereas the second indicator is accurate for samples of 3000 or more observations.

2 Research Results

2.1 Kurtosis

The kurtosis of a Normal distribution is 3, whereas that of a Laplace distribution is 6 (Aryal, 2006). Looking at the following figure, we see that the Laplace distribution has a fatter tail, thus confirming its larger kurtosis.



In fact, given a fixed sample size n, the approximate probability that the kurtosis of a Laplace sample is smaller than or equal to that of a Normal sample of the same size is shown below for various n:

	n=50	n=100	n=200	n=500
$P(K_{Laplace} \leq K_{Normal})$	9%	3%	0.4%	0

Thus, given a sample, we can potentially tell whether it is drawn from a Normal population or a Laplace population by calculating the sample kurtosis. We developed a test for kurtosis which works as follows:

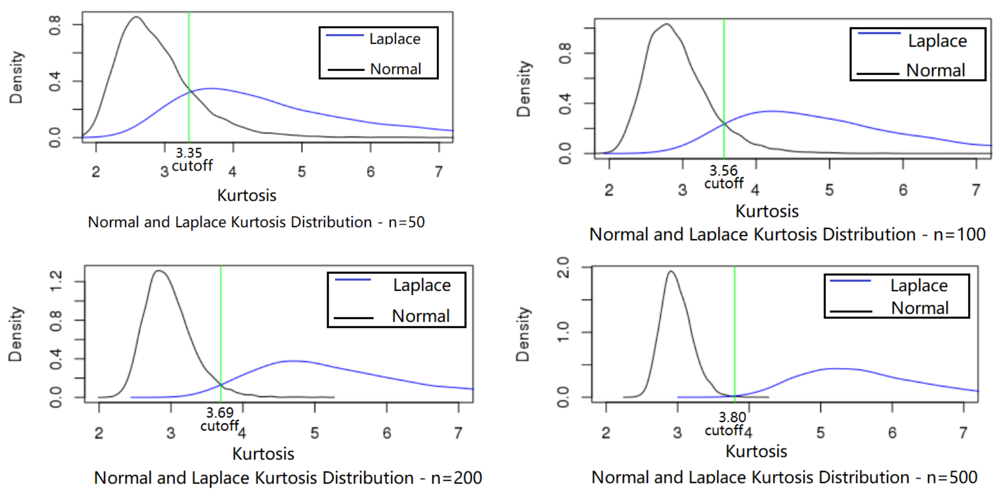
- Since the probability that the kurtosis of a Laplace sample of size 50 is below 2.85 is around 5% or lower for a larger sample size, if the sample kurtosis is smaller or equal to 2.85, then we conclude that it is from a Normal distribution.

- Since the probability that the kurtosis of a Normal sample of size 50 is above 4 is around 5% or lower for a larger sample size, if the sample kurtosis is larger or equal to 4, then we conclude that it is from a Laplace distribution.
- If the sample kurtosis is between 2.85 and 4, we look at how likely a random sample of the same size as the sample from a Normal distribution (a Laplace distribution) has a kurtosis larger than (smaller than) the sample kurtosis. Specifically, depending on a sample size n, we determine a cutoff value. If the sample kurtosis is larger than this cutoff value, then we conclude that it is from a Laplace distribution; if it is smaller than this cutoff, then it is from a Normal distribution.

In the third case, the cutoff value is a value that makes the probability that a Normal random sample of the same size as the sample has a kurtosis larger than or equal to this cutoff equal to the probability that a Laplace random sample of the same size has a kurtosis smaller than or equal to this cutoff. The cutoff is determined in a way that makes the probability that we incorrectly conclude that a sample is from a Normal distribution equal to the probability that we incorrectly conclude that a sample is from a Laplace distribution.

Looking at the table and figures below, it is clear that for larger n, our test is more accurate, with a smaller error rate. The error rate is defined as the probability that we make an incorrect conclusion. We suggest a sample of at least 100 observations to guarantee the accuracy of our test. We approximated error rates and cutoff values by using 10,000 samples for each distribution.

	n=50	n=100	n=200	n=500
Error Rate	17%	9%	3%	0.2%
Cutoff Value	3.35	3.56	3.69	3.80

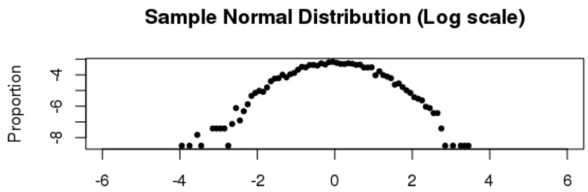


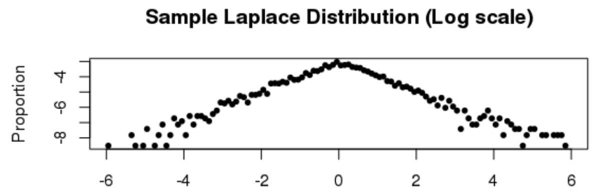
2.2 Residual Standard Deviation

The indicator we use in this section is inspired by pdfs of the Normal and Laplace distributions. The pdf of a Normal distribution with mean μ and standard deviation σ is $f_n(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$, and the pdf of a Laplace distribution with mean μ and standard deviation $\sqrt{2}b$ is $f_l(x) = \frac{1}{2b} \exp(-\frac{|x-\mu|}{b})$. Taking log of both pdfs, we get $\ln(f_n(x)) = \ln(\frac{1}{\sqrt{2\pi\sigma^2}}) - \frac{(x-\mu)^2}{2\sigma^2}$, and $\ln(f_l(x)) = \ln(\frac{1}{2b}) - \frac{|x-\mu|}{b}$. Thus, $\ln(f_n(x))$ is a quadratic function of x, while $\ln(f_l(x))$ is a linear function of x.

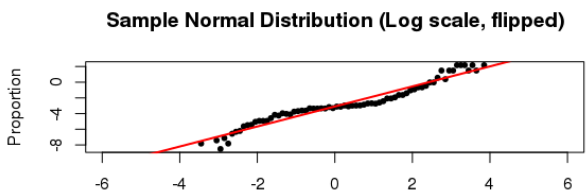
We thus can reasonably speculate that the estimated probabilities of each value occurring of a Laplace sample of data, after being taken log, will display a larger "linearity" than those of a Normal sample of data. We measure this "linearity" by fitting a straight line to the logged frequencies using least-square regression, and then looking at the standard deviation of the residuals.

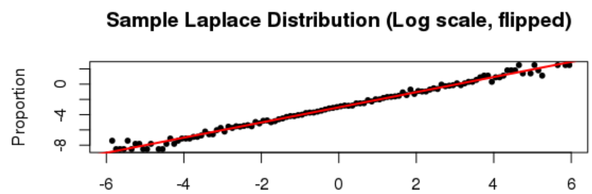
Here are some graphs of samples of size 5000 from Normal and Laplace distributions:





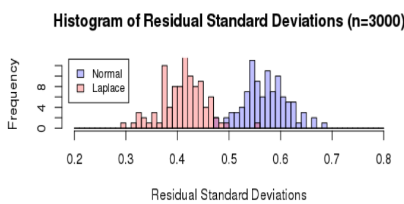
In order to fit a straight line to the data more conveniently, we flip the right half of the data horizontally (the red line is the fitted regression line):

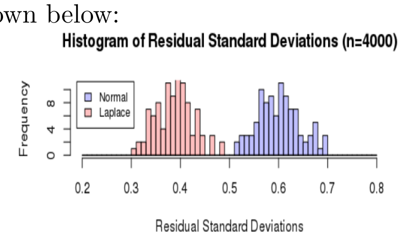


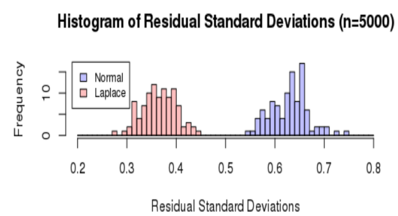


From the graphs we can see that the residuals in the case of the Normal distribution apparently have a larger variance than the residuals in the case of the Laplace distribution.

We simulated the sampling process 100 times for each distribution, and the sampling distributions of residual standard deviation are shown below:



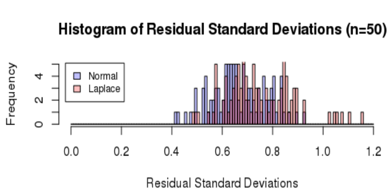


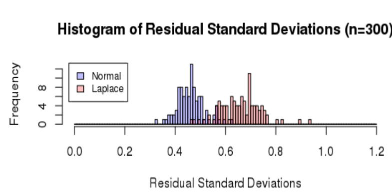


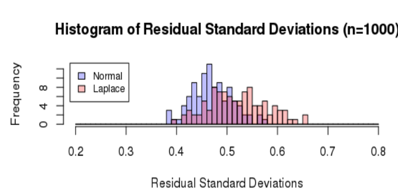
As in the Kurtosis section, we again developed a test, determined error rate and cutoff value for each sample size, as shown in the table below. If the residual standard deviation of the sample whose size is n is smaller than the corresponding cutoff value, then we conclude that it is from a Laplace distribution; if it is larger, then we conclude that it is from a Normal distribution. The error rate and cutoff value are determined in the same way as in the Kurtosis section. Notice that just as the kurtosis test, this test is more accurate with a larger sample size.

	n=3000	n=4000	n=5000
Error Rate	2.5%	0.2%	0
Cutoff Value	0.49	0.49	0.48

Note that however, this test on residual standard deviation σ is valid only for large size sample data. The sampling distribution of σ displays a curious pattern of movement as the n grows. According to our experiment, when $n = 50$, the mean of Laplace σ is 0.75, larger than the mean of Normal σ which is 0.66. When n grows to 300, Laplace σ is even significantly greater than Normal σ (with cutoff value 0.54, the error rate is only 5.8%). When n grows to 1000, Laplace σ and Normal σ merge back together. And as n continues to grow, Laplace σ finally gets significantly smaller than Normal σ.







3 Future Work

In this project we learned what are some of the potential indicators that can be used to differentiate between two similar distributions. Looking at the figures in the Kurtosis section, we observe that the distribution of kurtosis for the Lapalce distribution seems to be more spread out than that for the Normal distribution. We would like to investigate why this is the case and what specific distributions the sampling distributions of kurtosis follow. Also, as the sample size increases, the kurtosis distribution for both distributions are less skewed. We would like to know whether this can be explained by the Central Limit Theorem. Further, we would like to study the curious movement pattern of the sampling distribution of residual standard deviation mentioned in the second section.

We would like to systematically investigate more indicators using the relationship: If $X_1 \ldots X_4 \sim Normal(0,1)$, then $X_1X_2 - X_3X_4 \sim Laplace(0,1)$ (Aryal, 2006). Also, we would like to learn another indicator proposed by Dr. Debasis Kundu, which is based on maximum likelihood estimation (Kundu, 2003).

References

Gokarna Raj Aryal. Study of laplace and related probability distributions and their applications. *Graduate Theses and Dissertations*, 2006. URL `https://scholarcommons.usf.edu/etd/2443/`.

Vance Harwood. Predicting stock market returns—lose the normal and switch to laplace. *6 Figure Investing*, 2018. URL `https://sixfigureinvesting.com/2016/03/modeling-stock-market-returns-with-laplace-distribution-instead-of-normal/`.

Debasis Kundu. Discriminating between the normal and the laplace distributions. *Advances in Ranking and Selection, Multiple Comparisons, and Reliability*, 2003. URL `http://home.iitk.ac.in/ kundu/paper98.pdf`.