Toward a robust and simple guideline for checking the Central Limit Theorem

Visruth Srimath Kandali[†]

Project advisor: Beth Chance[†]

Abstract. In statistical practice, many introductory statistical procedures require the sampling distribution of means to be approximately normal. Most students learn a simplified check of this condition as "$n \geq 30$", which often becomes a black-and-white mantra replacing visual inspection of the data. A slightly more detailed version might be "$n \geq 30$ as long as the population distribution is not too skewed." Our research seeks to clarify a guideline that incorporates measures of skewness along with sample size. We used simulation to explore the consequences of skewed populations with different sample sizes. We hope to provide students and practitioners with a slightly more refined rule that allows a way to operationalize the degree of skewness in statistical analysis.

1. Introduction. Ensuring the validity conditions of a statistical procedure are met is a critical component of valid statistical practice. One example of a common check is in inference for population means, wherein we often need the sampling distribution of the sample means to be approximately normal. One way of checking this is by relying on the Central Limit Theorem (CLT), which asserts that the sampling distribution of means converges to a normal distribution as sample size (n) increases. Following from this, a common check for the normality of a sampling distribution of sample means is "$n \geq 30$" with the idea being that a sample size of 30 suffices for the CLT to "kick in" and ensure that the sampling distribution of the means is sufficiently normal for the procedure.

It should be noted that 30 is not entirely an arbitrary sample size. Though we will go on to show that a rigid guideline of 30 isn't robust, with distributions that aren't too skewed (e.g. skewness ≤ 1), the current guideline typically suffices to produce an approximately normal sampling distribution. However the current guideline also appears to be overly stringent when the skewness is modest. At larger skewnesses, the current guideline no longer holds up. Our exploration mainly focuses on what happens beyond small skewnesses and that is where this guideline no longer applies.

It has been shown that the convergence rate of the (strong) CLT is directly related to skewness, as measured by the third absolute normalized moment [1, 4], known as the Berry-Esseen Theorem. We seek to codify a relationship that captures this relationship in a manner which is easier to teach and simple to use.

To observe the effects of skewness on the shape of the sampling distribution at various sample sizes, we took a simulation approach. We generated samples of sizes ranging from 5 to 6000 from numerous levels of skewed continuous population distributions. We looked at exponential, log-normal, and gamma distributions with varying parameters to achieve population skewnesses from around 0.5 to around 3. We used 1,000,000 repetitions to mitigate any potential effects of random sampling, thus ensuring that our empirical sampling distributions are accurate [5].

Our main focus is pedagogical. Whilst skewness isn't a great tool (sample skewness from

[†]Department of Statistics, California Polytechnic State University San Luis Obispo, California.

non-normal populations is a biased estimator of population skewness [6]), it is better than nothing as it does have an effect on the sampling distribution of single mean's convergence rate [1, 4]. The classical guideline does not use this information; our guideline extends the classical guideline to use skewness in a simple manner, encouraging students to look at their data and consider various aspects of it (such as skewness) when checking validity conditions.

We begin with further analysis of the problem in section 2, followed by a discussion of our data generation procedure in section 3, the results of our analysis in section 4, extensions of our results in section 5, and finish with our conclusions and future work in section 6.

2. Problems with the current guideline.

Whilst $n \geq 30$ roughly holds at smaller skewness (i.e. populations with skewness lesser than 1), and can even be conservative, at larger skewness this guideline no longer holds–the sampling distribution of means ceases to be approximately normal. Let us first clarify what exactly we mean when referring to an "approximately normal" sampling distribution. Whilst there are many generally appropriate normality tests such as D'Agostino or Anderson-Darling, these can be overly sensitive at very large sample sizes such as 1,000,000, which is the number of samples we used to create our sampling distributions. Since these classic tests won't work for our scenario, we decided upon an alternative form of checking for normality by looking at tail probabilities. Additionally, the context for invoking the CLT tends to be with inferential procedures where we care about a p-value; looking at tail probabilities is a natural extension of that idea.

In a standard normal distribution, the probability of observing a z-score greater than 1.96 (or lesser than -1.96) is 0.025. Let us call this probability a tail probability. Further, let us call the probability of observing z-score greater than 1.96 the upper tail probability, and lesser than -1.96 the lower tail probability. We looked at right-skewed distributions, so for a distribution to be "approximately normal" we wanted both tail probabilities to be within 20% of 0.025; i.e. the lower tail probability is greater than 0.020 and the upper tail tail probability is lesser than 0.030 (vice-versa for a left-skewed distribution). While it might seem natural to instead look at the sum of the tails instead of individually, doing so would fail to recognize skewness. In Figure 2 we can see that the sum of the tail probabilities is around 0.0478, which is quite close to the 0.05 we would expect from a normal distribution. However there clearly is skewness as the upper tail and lower tail individually aren't very close to 0.025. Due to cases like these, we opted to look at tail probabilities individually. Having the lower tail greater than 0.02 and the upper tail lesser than 0.03 ensures that, in our opinion, inferential procedures could still rely on this distribution being approximately normal, e.g. to calculate p-values, which is ultimately our goal.

We decided upon using tail probabilities as a measure of normality, given the pertinence of tail probabilities in inference. Additionally, since our focus is more on introductory statistics and pedagogy, we didn't want to use a more complicated method like Kolmogorov-Smirnov distance. Percent error in tail probabilities is simple and easily interpretable; additionally, its relevance is quite clear when couched in the context of p-values and should be easy to explain to students.

We can see that the present guideline is too conservative in some scenarios. Figure 1 shows the normalized sampling distribution of means for samples of size 10 taken from a normal population. We can see that this sampling distribution is very close to being normal, as

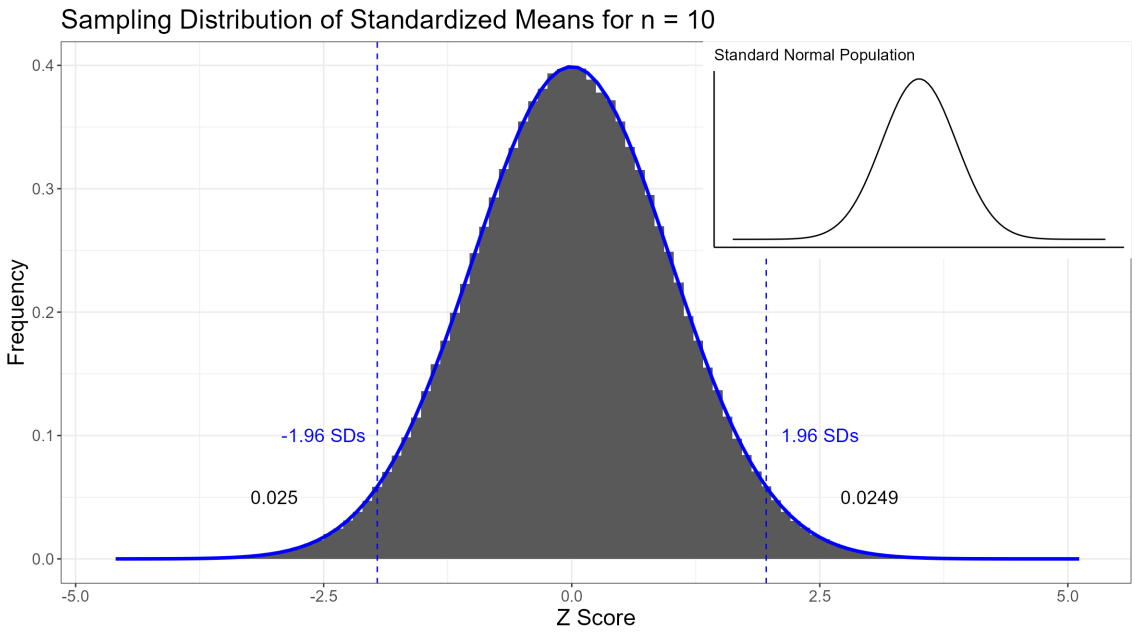

Figure 1. *Sampling distribution of standardized means using samples of size 10 taken from a normal population. The blue curve is a standard normal distribution. The two probabilities are the lower and upper tail probabilities as mentioned above.*

defined by our tail probability criteria. This holds even though our sample size is much smaller than what the current current guideline asserts would be necessary to get an approximately normal sampling distribution. This is, of course, entirely expected and addressed by the CLT.

Suppose we take samples of size 30 from an exponential population, which has a population skewness of 2 (regardless of the value of the rate parameter), and look at the standardized sampling distribution of means. We standardized using the population parameters μ and σ (though standardizing with the sample data produced similar results). Given that the sample size is 30, we might expect an approximately normal distribution, and this sampling distribution is shown in Figure 2.

A visual inspection may classify Figure 2 as being "approximately normal" but looking at the tail probabilities like we are recommending reveals that the distribution differs from a normal distribution; the lower tail probability is just 0.0142 and the upper tail probability is 0.0336. Here we can see that a sample size of 30 would not suffice, and that applying an inference procedure under the assumption of normality may not be appropriate here. Let us look then at another sampling distribution, this time with samples of size 150.

At this much larger sample size, we can see in Figure 3 that there is less disparity in tail probabilities, and the histogram looks more normal and symmetric. This distribution has a lower tail probability of 0.0204 and an upper tail probability of 0.0292, thus fitting our idea of normality. We think that such a distribution can be considered reasonably normal, enough for procedures that rely on the normality of this distribution at least. We think that this distribution would give reasonably accurate tail probabilities, which is what we're concerned

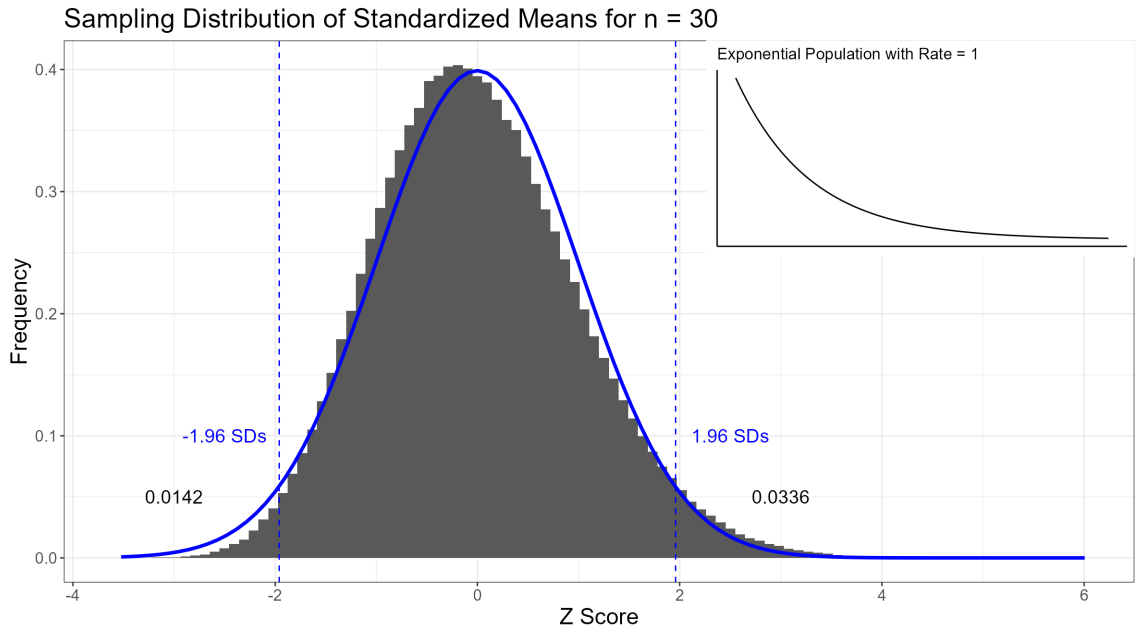

Figure 2. *Sampling distribution of standardized means using samples of size 30 taken from an exponential population. The blue curve is a standard normal distribution. The two probabilities are the lower and upper tail probabilities as mentioned above.*

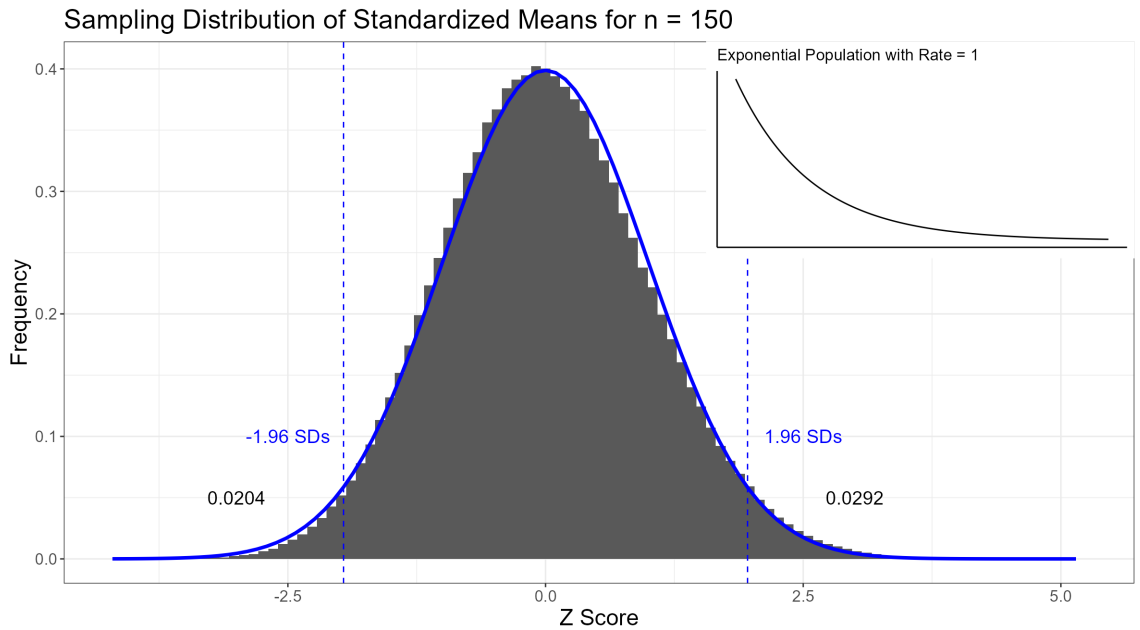

Figure 3. *Sampling distribution of standardized means using samples of size 150 taken from an exponential population. The blue curve is a standard normal distribution. The two probabilities are the lower and upper tail probabilities as mentioned above.*

about.

As mentioned, we decided to take a simulation approach to determine a better guideline which would take into account the skewness of the population distribution to develop a new, robust validity check.

3. Computational work. Before we move on to our computing methodology, it is worthwhile to take a moment to properly define some measures of skewness as there are a number of ways to measure skewness, both for samples and populations [3, 6]. In Julia, the formula used for skewness in StatsBase.jl is the Fisher-Pearson coefficient of population skewness,

$$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2]^{3/2}}$$

as we can see from [2, 3]. Here, m_3 is the third central moment and m_2 is the second central moment [3]. For sample skewness, there is the adjusted Fisher-Pearson coefficient. It can be defined as $G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1$ [3, 6].

To look at the effects of skewness, we selected a number of continuous distributions from a few different families to discern the effects of skewness within and between population distributions. See Table 1 for the distributions we selected along with their population skewnesses. Our Gamma distribution is parameterized by shape and scale.

We simulated random samples of various sizes from each of these distributions, ranging from $n = 5$ to $n = 500$. We took 1,000,000 samples of each sample size from each distribution to create sampling distributions of means; we used such a large repetition size (r) to ensure that our results would not be swayed by random variability. Figures 2 and 3 are graphical representations of sampling distributions created by this method. Random variability can have a material impact on results if sufficient repetitions aren't taken, and we found that in our research that smaller values for r wouldn't suffice as there was noticeable variation in our sampling distributions of means [5].

Given the computational intensity required for our task, we turned to Julia and relied on parallelism to cut down on run times. To ascertain the minimum sample size required to generate a sampling distribution which is "approximately normal" as defined in section 2, we simulated samples of varying sizes. We used roughly 50 different sample sizes and generated sampling distributions for each combination of sample size and distribution. We then looked at their tail probabilities. We know there to be a monotonic relationship between tail probabilities and sample size, with the lower tail increasing to 0.025 and the upper tail decreasing similarly to 0.025 (assuming a right-skewed population distribution). As such, we simply selected the smallest sample size in our set for each distribution which met our criteria to be "approximately normal." The resulting empirical minimum sample sizes along with the resulting skewness measure for the generated sampling distributions can be seen in Table 1.

The function we wrote to generate a sampling distribution is quite flexible, and could easily be extended beyond how we used it in this project–viz. beyond a sampling distribution of means and the particular distributions and sample sizes we selected. Please find our code on the first author's GitHub.

Table 1

The population distributions we chose along with their skewnesses and empirical minimum sample size to pass our criteria. The column titled "Skewness" details the population skewness for each distribution, and is exact. The "Sample Distribution Skewness" column lists the skewness of the normalized sampling distribution of sample means for samples from the given population at the associated "Empirical Minimum Sample Size" and is measured by g_1.

Distribution	Skewness	Empirical Minimum Sample Size	Sampling Distribution Skewness
Gamma ($\alpha = 16, \theta = 1$)	0.500	10	0.158
Log-normal ($\mu = 0, \sigma = 0.25$)	0.778	20	0.143
Gamma ($\alpha = 4, \theta = 1$)	1.000	40	0.154
Gamma ($\alpha = 2, \theta = 1$)	1.414	70	0.159
Log-normal ($\mu = 0, \sigma = 0.5$)	1.750	125	0.157
Exponential ($\theta = 1$)	2.000	150	0.162
Gamma ($\alpha = 1, \theta = 1$)	2.000	150	0.162
Gamma ($\alpha = 0.64, \theta = 1$)	2.500	200	0.162
Log-normal ($\mu = 0, \sigma = 0.75$)	3.263	350	0.163

4. Empirical results. Given the relationship posited by the Berry-Esseen theorem, we decided to look at skewness and square root empirical minimum sample size. We found that there exists a linear relationship between the two, with our regression equation being $\widehat{\text{skewness}} = 0.175 \times \sqrt{n} - 0.08$ This can be roughly reduced to $\widehat{\text{skewness}} = \frac{1}{6}\sqrt{n}$. Solving for n gets us to our proposed guideline:

$$n \geq 36 \times \text{skewness}^2$$

Our model has an R^2 value of 99.22%, and as Figure 4 shows there is a strong, positive, linear relationship between root sample size and skewness.

It is interesting to note the 36 in this expression, lending some credence to the typical guideline of $n \geq 30$. Using our heuristic, anything with a skewness lesser than 1 would need a sample size less than 36. The classic guideline can be too conservative in many cases with small skewnesses, but too lax with large population skewnesses. This means that we can rely on the CLT with smaller, less conservative sample sizes if our population distribution isn't too skewed, as seen in Figure 1.

In Figure 5 we can clearly once again see how both the lower and upper tail probabilities steadily converge to 0.025, showing the effects of the CLT as sample size tends to infinity. This graph was created by using our empirical tail weights for each distribution at various sample sizes, then smoothed. In Figure 5 we highlighted the region of tail probabilities we used to deem a distribution "approximately normal" at 20% error, as well as provide a line to demarcate the classic $n \geq 30$ guideline. We can see that our new, robust guideline and the classic guideline roughly agree for distributions with minimal skewnesses.

5. Extending our Results.

5.1. Error Rate. As a reminder, we considered a sampling distribution to be "approximately normal" if the tail probabilities were both within 20% of 0.025. We can abstract away the 20% easily. To do so, we looked at error bounds ranging from 6% to 30%, and proceeded with our modelling analysis at each error rate. This resulted in a number of models, each with

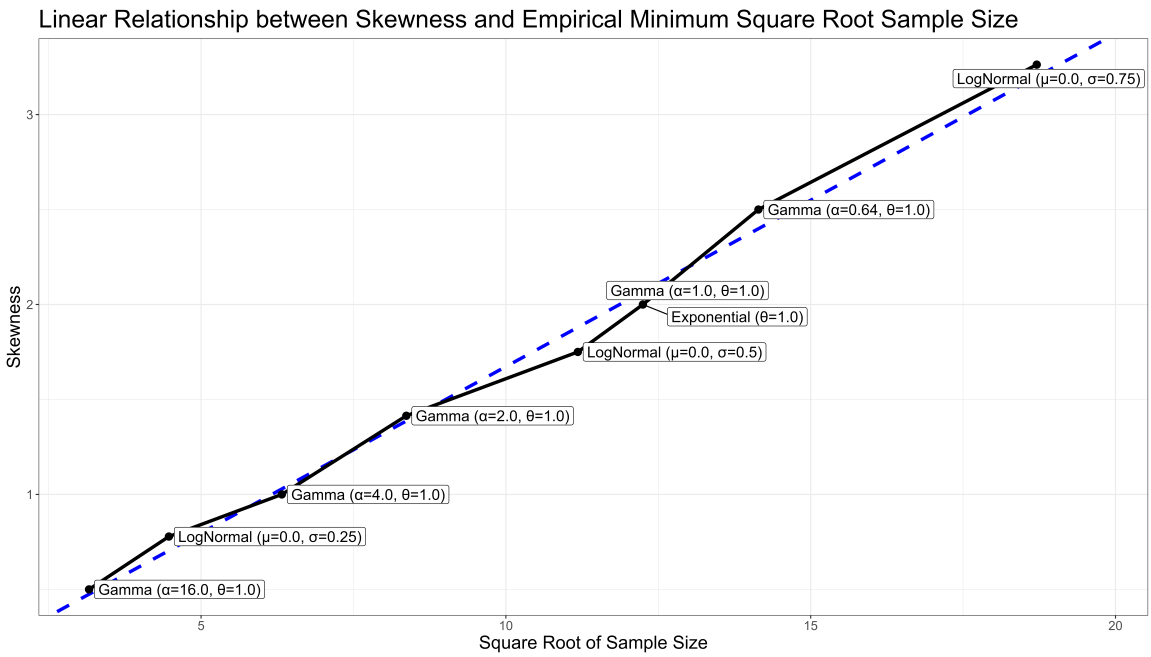

Figure 4. *Modeling the relationship between $\sqrt{n}$ and population skewness.*

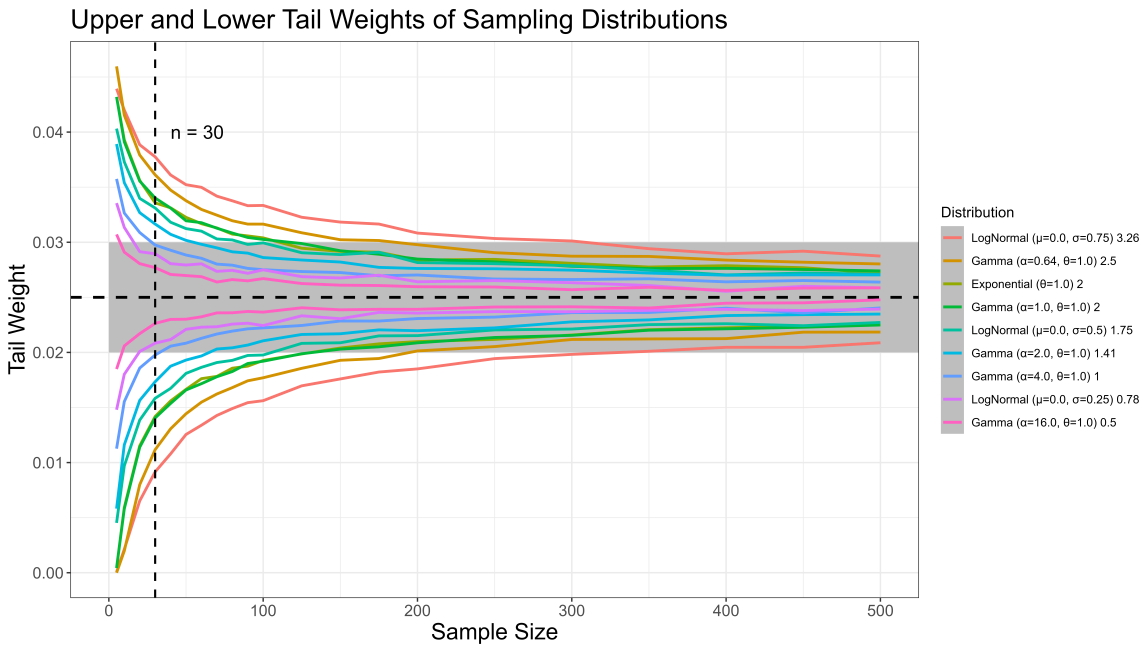

Figure 5. *Tail probabilities converge to 0.025 as sample size increases.*

an associated coefficient that can be manipulated and used in our formula as a coefficient for skewness (e.g., 20% error corresponds to 36.)

We selected 54 sample sizes ranging from 5 to 6000. We ran our full analysis with error rates ranging from 30% to 6%. Six percent was the smallest error rate we could get with these data to still find an empirical minimum sample size for all distributions. We could not get an approximately normal sampling distribution using samples of size 6000 from a our Log-normal ($\mu = 0, \sigma = 0.75$) distribution (with a skewness of approximately 3.263) at error rates more stringent than 6%.

With these data, we then fit a linear regression to predict the coefficient of each individual regression at its own error rate, from the error rate itself. We can see another strong, positive, linear association in Figure 6. This is confirmed by the model, which has an R^2 value of 99.05%.

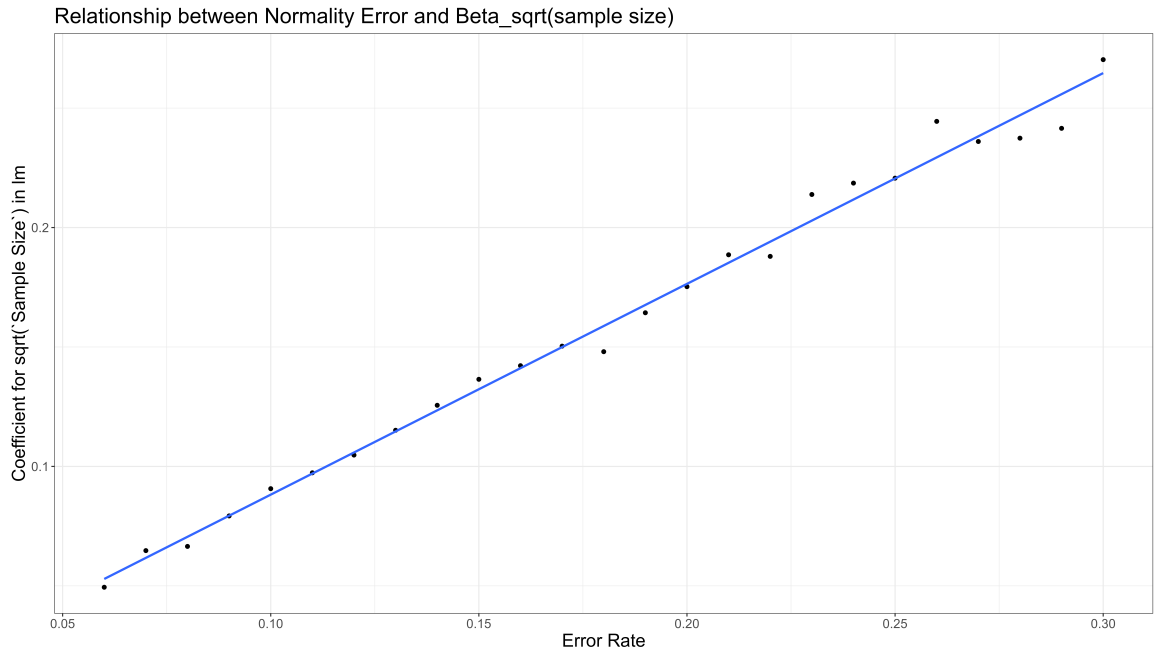

Figure 6. *Minimum requisite sample size for normal sampling distribution increases as error tolerance increases.*

Now, to predict the appropriate coefficient of skewness[2] for a fixed error, we can just use the model:

$$\widehat{\text{coefficient}} = 0.8824 \times \text{error}$$

This allows us to approximate the coefficient for our analysis done with a given error rate. To use this, we then need to square and reciprocate the fitted value to receive a number that can replace the 36 in $n \geq 36 \times \text{skewness}^2$, just as we did before with a fixed error rate. Combining these two results then gets us a rough formula for requisite sample size parametrized by error ($0 < \text{error} < 1$) and population skewness.

$$n \geq 1.284 \times \text{error}^{-2} \times \text{skewness}^2$$

Setting error = 20% as we did in our study gets us a coefficient of around 32, which is similar to our 36 from earlier. This combined formula is less precise than re-running our simulation at a desired error rate, but is of course much easier to use. Since we are estimating the coefficient for skewness in our formula through another regression, we expect to see deviations from the coefficients we would get if we reran the simulation at each error rate. This is why we get 32 instead of 36 as expected; we are paying a penalty in terms of accuracy for the ease gained by not needing to run the entire simulation setup again. Regardless, this combined formula gracefully dismisses the fixed error rate seen before.

5.2. Sample Skewness. Our work has found a relationship between population skewness and required sample size. Whilst interesting, this isn't directly applicable to a sample and its skewness. Sample skewness is biased for non-normal populations [6], so we developed an empirical heuristic to side-step that and allow our rule to be used by roughly estimating population skewness from a sample.

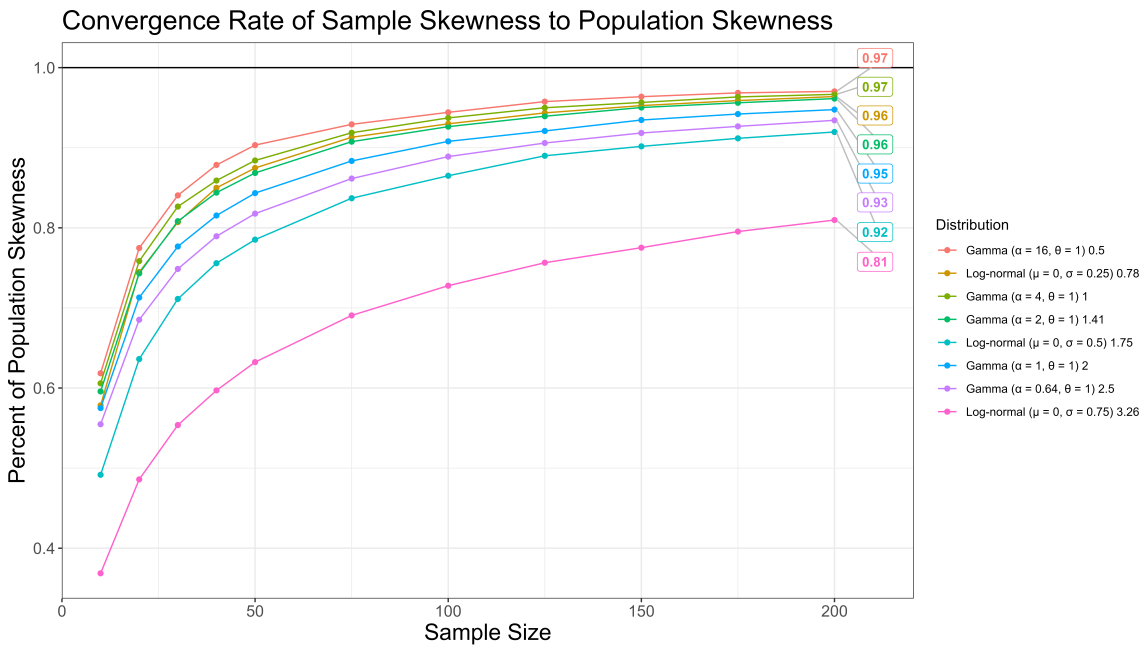

Figure 7. *Sample skewness converges to population skewness at different rates, depending on population skewness.*

We can see in Figure 7 that sample skewnesses converge to population skewness at different rates, parametrized by population skewness. Due to that, we opted to not aim for a general solution which would account for all population skewnesses. Figure 7 shows how sample skewnesses are naturally biased downwards; sample skewnesses tend to underestimate population skewness. We generated Figure 7 by looking at $\frac{\text{Mean Sampling Skewness}}{\text{Population Skewness}}$ at the listed sample sizes and our typical distributions. If we invert that fraction and look at $\frac{\text{Population Skewness}}{\text{Mean Sampling Skewness}}$, we can get a correction factor–i.e. a number we can multiply our sample skewness by to get a better estimate of population skewness.

We will proceed under the assumption that the most skewed population samples could come from is an exponential distribution. This means that the largest population skewness is 2. Since our focus is mainly pedagogical, we think this is an okay assumption; also, a similar analysis could be done to any desired skewness. We are essentially biasing our sample skewnesses upwards, correcting to a population skewness of 2. Under this assumption, we can look at a correction factor to better estimate population skewness given a sample skewness.

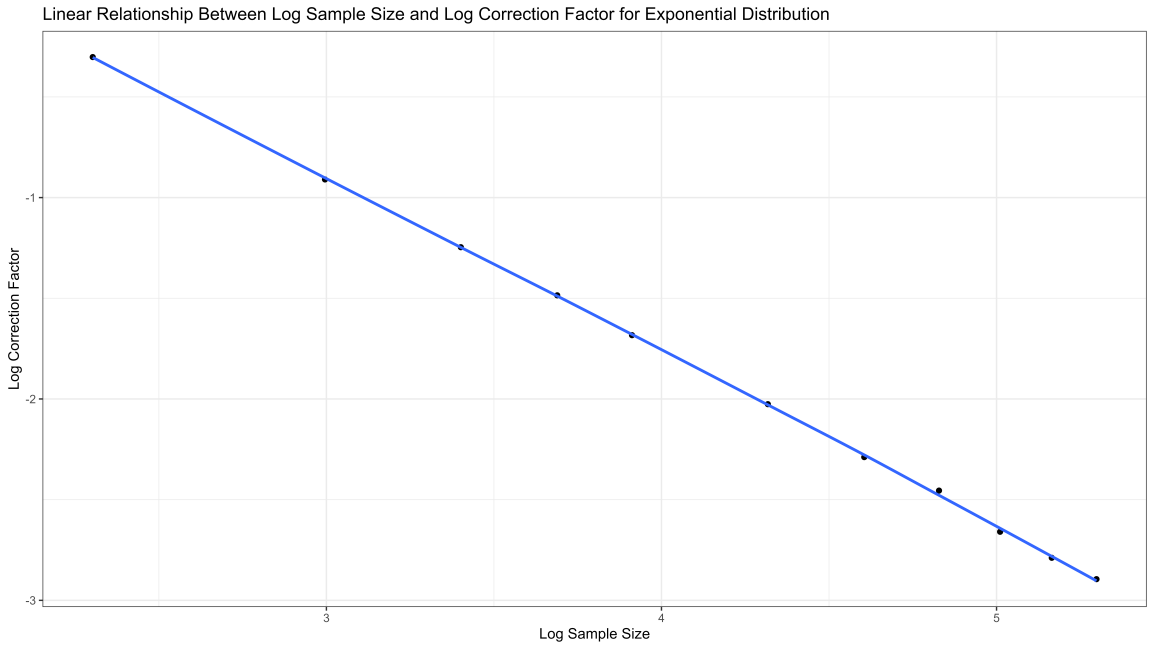

Figure 8. *Log correction factor decreases linearly with log sample size. Log correction factor is derived by* $\ln(\frac{\textit{Population Skewness}}{\textit{Mean Sampling Skewness}} - 1)$.

As we can see in Figure 8, there is a linear relationship on the log-scale between correction factor and sample size. We fit a model to these data which had an R^2 value of 99.97%. The regression equation is $\log \widehat{\text{Correction Factor}} = -0.865 \times \log n + 1.696$. We can invert this formula to get a correction formula which better approximate population skewness. Corrected Skewness $\approx$ Sample Skewness $\times$ $(e^{-0.865 \times \log n + 1.696} + 1)$ This relationship allows us to use sampling skewness in our heuristic by applying an appropriate correction factor to approximate population skewness, then plugging that estimate in to our guideline.

Using this correction, we plotted adjusted skewnesses for mean sampling skewnesses in Figure 9. We can see that we do a much better job of estimating population skewness in populations with skewnesses lesser than 2–comparing these skewnesses estimates to Figure 7 shows that we've done a lot better than the unadjusted estimates. This guideline is a touch conservative for samples taken from population skewnesses which are much smaller than exponential, but we think this is better than underestimating skewness in this context.

6. Conclusions. Our research provides a guideline to help students determine an appropriate sample size to invoke the CLT which encourages consideration of population skewness

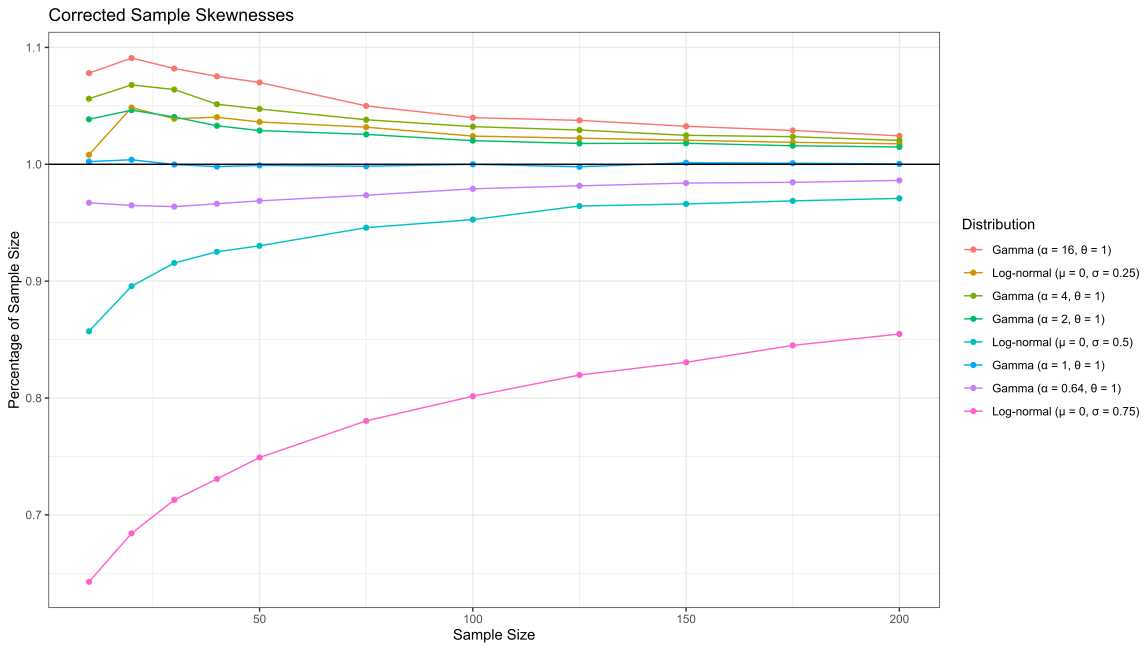

Figure 9. *Ratio of adjusted sample skewness to population skewness for mean sampling skewness of distributions*

as part of a more thorough, visual inspection of the data. In lieu of the classical guideline of $n \geq 30$ we propose $n \geq 36 \times \text{skewness}^2$. Sample skewness is a relatively easy statistic to find and use, and so we recommend the trivial step of reporting sample skewness with sample size when students look at their data.

Our results are in accord with the Berry-Esseen theorem, but are geared more towards students and practitioners in that our guideline is far more applied and directly applicable to the normality check many inferential procedures require. Our guideline is more robust and better than the classic guideline, and it encourages some exploratory data analysis, thus ensuring that procedures aren't applied without any care for how the data themselves look. Our guideline clearly shows how factors of the data which may not be easy to discern at a first glance, such as skewness above 1, can have an important impact on the validity of a procedure. Thus it is important to ensure that one does all the validity checks needed for a method, and also looks at their data to see whether there are any noticeable visual inconsistencies.

In the future we would also like to expand our work to include more types of distributions of varying skewnesses and such. We would also like to look at how the sampling distribution of difference in means would look like and how important skewness is there, especially if both samples have dissimilar skewnesses. We would like to measure and codify the effects of these differences in skewnesses; empirical evidence suggests if both samples have similar skewnesses and sample sizes, the sampling distribution might converge quicker to a normal distribution but dissimilar skewness may lead to problems. We would like to measure and detail this result so it may be used in a pedagogical context as well.

We also did some preliminary exploration on the effects of kurtosis, and would like to do

more research to see what role kurtosis plays in our research. We would also like to see whether t-statistics converge to a t-distribution, and if they do at what rate. A similar analysis as we have performed on means may reveal the robust nature of t-statistics. Additionally, we would like to work more with discrete distributions; we began working with the Poisson distribution early on but ran into issues and decided to focus on continuous univariate distributions. Furthermore we would like to look at the Binomial distribution and further explore another common sample size criteria, "$np \geq 10$ and $n(1-p) \geq 10$" as well, perhaps by considering Binomial skewness, $\frac{q-p}{\sqrt{npq}}$.

Acknowledgments. We would like to acknowledge two anonymous reviewers for their advice, especially on suggesting that we extend our work to cover arbitrary error rates and sample skewness. We would also like to acknowledge the editors for their suggestions and flexibility. The first author would also like to thank their advisor, Dr. Chance, for her invaluable help and unwavering support. This work was supported in part by the Frost Fund.

REFERENCES

[1] A. C. Berry, *The accuracy of the gaussian approximation to the sum of independent variates*, Transactions of the American Mathematical Society, 49 (1941), pp. 122–136, http://www.jstor.org/stable/1990053 (accessed 2024-06-17).

[2] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, *Julia: A fresh approach to numerical computing*, SIAM Review, 59 (2017), pp. 65–98, https://doi.org/10.1137/141000671, https://epubs.siam.org/doi/10.1137/141000671.

[3] D. P. Doane and L. E. Seward, *Measuring skewness: A forgotten statistic?*, Journal of Statistics Education, 19 (2011), https://doi.org/10.1080/10691898.2011.11889611, https://doi.org/10.1080/10691898.2011.11889611, https://arxiv.org/abs/https://doi.org/10.1080/10691898.2011.11889611.

[4] C. Esseen, *On the Liapounoff Limit of Error in the Theory of Probability*, Arkiv för matematik, astronomi och fysik, Almqvist & Wiksell, 1942, https://books.google.com/books?id=VjXgPgAACAAJ.

[5] T. Hesterberg, *What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum*, 2014, https://arxiv.org/abs/1411.5279.

[6] D. N. Joanes and C. A. Gill, *Comparing measures of sample skewness and kurtosis*, Journal of the Royal Statistical Society. Series D (The Statistician), 47 (1998), pp. 183–189, http://www.jstor.org/stable/2988433 (accessed 2024-06-17).