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 sampling distribution normality 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, and 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. $skewness \leq 1$) 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. Our work 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, 3], 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 500 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 [4].

Our new heuristic takes into account the effects of skewness on the sample size requirement of inferential procedures relying on the Central Limit Theorem. We provide a simple, robust

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

check that can be considered a drop-in replacement for the classic validity check of $n \geq 30$.

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, and finish with our conclusions and future work in section 5.

2. Problems with the current guideline. Whist $n \geq 30$ 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 perfectly good normality tests such as D'Agostino or Anderson-Darling, these are 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.

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, and lesser than -1.96 the lower tail. We looked at right-skewed distributions, so for a distribution to be "approximately normal" we wanted both tail probabilities to be within 20% of 0.025l; i.e. the lower tail is above 0.020 and the upper tail is below 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 1 we can see that the sum of the tail probabilities is 0.048, 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, which is ultimately our goal.

Let us look at the results of these criteria. 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 plotted in Figure 1.

A visual inspection may classify Figure 1 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 0.014 and the upper tail probability is 0.034. Herein 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. This new number is the result of our proposed heuristic, being $n \geq 36 \times skewness^2$, which we will elaborate upon in section 4. With a skewness of 2, our guideline proposes a minimum sample size of roughly 144 needed for an approximately normal sampling distribution of means for samples taken from an exponential population.

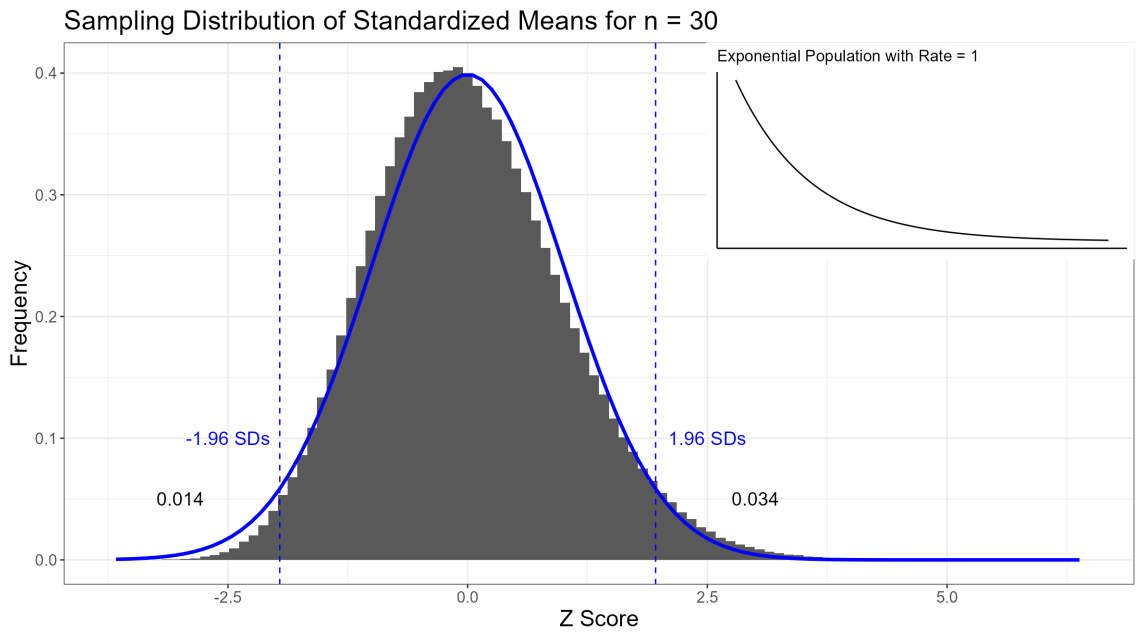

Figure 1. *Sampling distribution of means using samples of size 30 taken from an exponential distribution.*

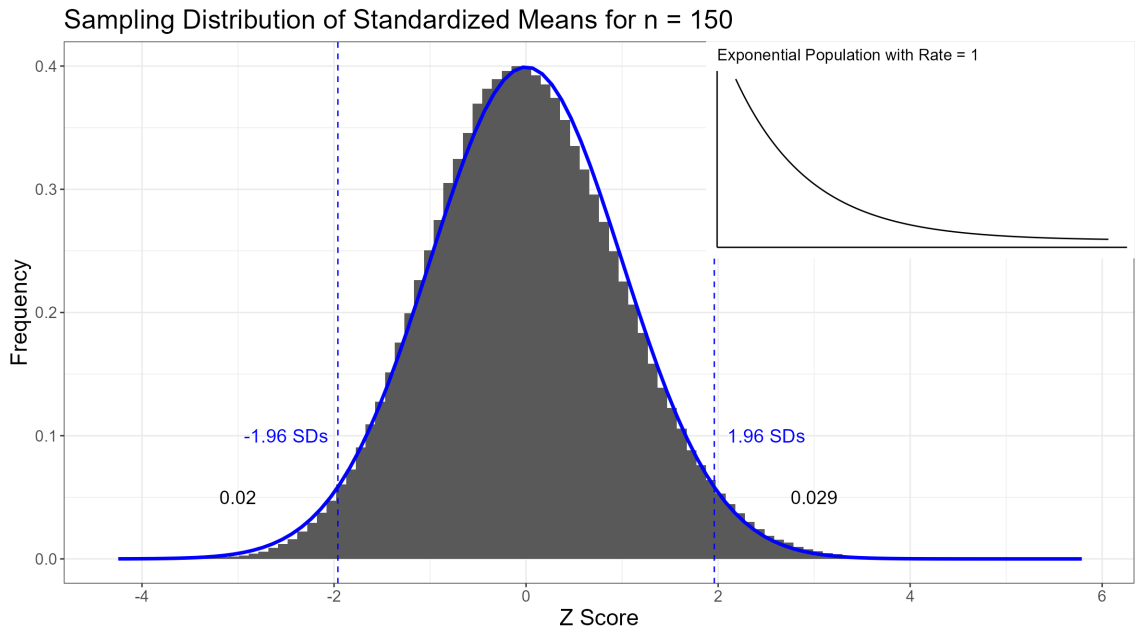

Figure 2. *Sampling distribution of means using samples of size 150 taken from an exponential distribution.*

At this much larger sample size, we can see in Figure 2 that there is less disparity in tail probabilities and the histogram looks more normal and symmetric. This distribution has a lower tail probability of 0.02 and an upper tail probability of 0.029, thus fitting our definition 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.

Figures 1 and 2 show that whilst the current guideline would claim that the sampling distribution of sample means is "approximately normal" with samples of size 30, in reality these sampling distributions could be skewed and non-normal as we have defined. 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. There are a number of ways to measure skewness, both for samples and populations [2, 5]. 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 [6, 2]. Here, m_3 is the third central moment and m_2 is the second central moment [2]. For sample skewness, there is the adjusted Fisher-Pearson coefficient. It can be defined as $G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1$ [2, 5].

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 distributions. We didn't see any particular differences across distributions, so the distribution itself matters less than its skewness. See Table 1 for the distributions we selected along with their population skewnesses. Our Gamma distribution is parameterized by shape and scale.

Table 1

The population distributions we chose along with their skewnesses.

Distribution	Skewness
Gamma ($\alpha = 16, \theta = 1$)	0.5
Log-normal ($\mu = 0, \sigma = 0.25$)	0.778
Gamma ($\alpha = 4, \theta = 1$)	1.000
Gamma ($\alpha = 2, \theta = 1$)	1.414
Log-normal ($\mu = 0, \sigma = 0.5$)	1.750
Exponential ($\theta = 1$)	2
Gamma ($\alpha = 1, \theta = 1$)	2
Gamma ($\alpha = 0.64, \theta = 1$)	2.5
Log-normal ($\mu = 0, \sigma = 0.75$)	3.263

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 1 and 2 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 [4].

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 prior defined, we simulated samples of varying sizes. We used roughly 20 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 2.

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 our Github.

Table 2

The population distributions we chose along with their skewnesses and empirical minimum sample size to pass our criteria.

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	30	0.143
Gamma ($\alpha = 4, \theta = 1$)	1.000	40	0.154
Gamma ($\alpha = 2, \theta = 1$)	1.414	80	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	250	0.162
Log-normal ($\mu = 0, \sigma = 0.75$)	3.263	400	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 $\widehat{skewness} = 0.165 \times \sqrt{n} - 0.0619$. This can be roughly reduced to $\widehat{skewness} = \frac{1}{6}\sqrt{n}$. Solving for n gets us to what we proposed earlier:

$$n \geq 36 \times skewness^2$$

Our model has an R^2 value of 99.78%, and as Figure 3 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

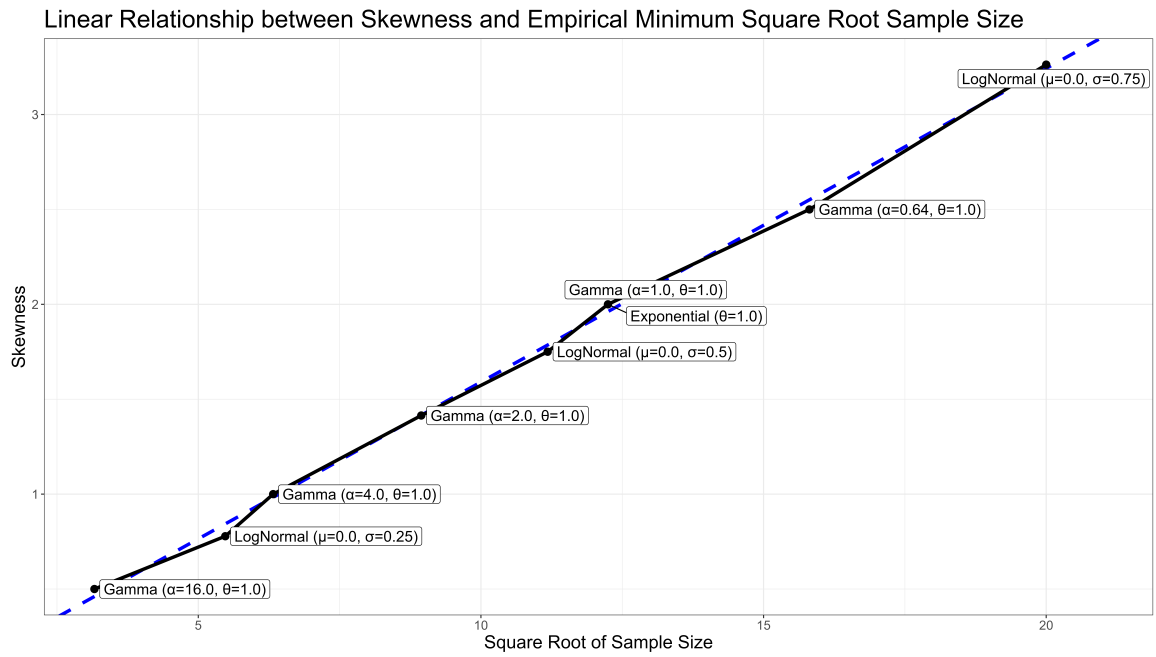

Figure 3. *Modeling the relationship between $\sqrt{n}$ and population skewness.*

a sample size less than 36, which seems to be agreeing with the classic guideline. The main difference between our heuristic and the old one is that our guideline can be extended past skewnesses of 1 as that extreme skewness is taken into account. This also means that we can rely on the CLT with smaller, less conservative sample sizes if our population distribution isn't too skewed.

In Figure 4 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 our selected sample sizes, then smoothed. In Figure 4 we highlighted the region of tail probabilities we used to deem a distribution "approximately normal" 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. Conclusions. Our research provides a guideline to help practitioners determine an appropriate sample size to invoke the CLT which encourages consideration of population skewness as well as part of a more thorough, visual inspection of the data. In lieu of the classical guideline of $n \geq 30$ we propose $n \geq 36 \times skewness^2$. 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 practitioners look at their data.

Clearly, there is a relationship between skewness and the sample size needed to get an approximately normal sampling distribution of the means. Our heuristic finds such a relationship, providing researchers with a simple, drop-in replacement for the current method which scales quite well with skewnesses.

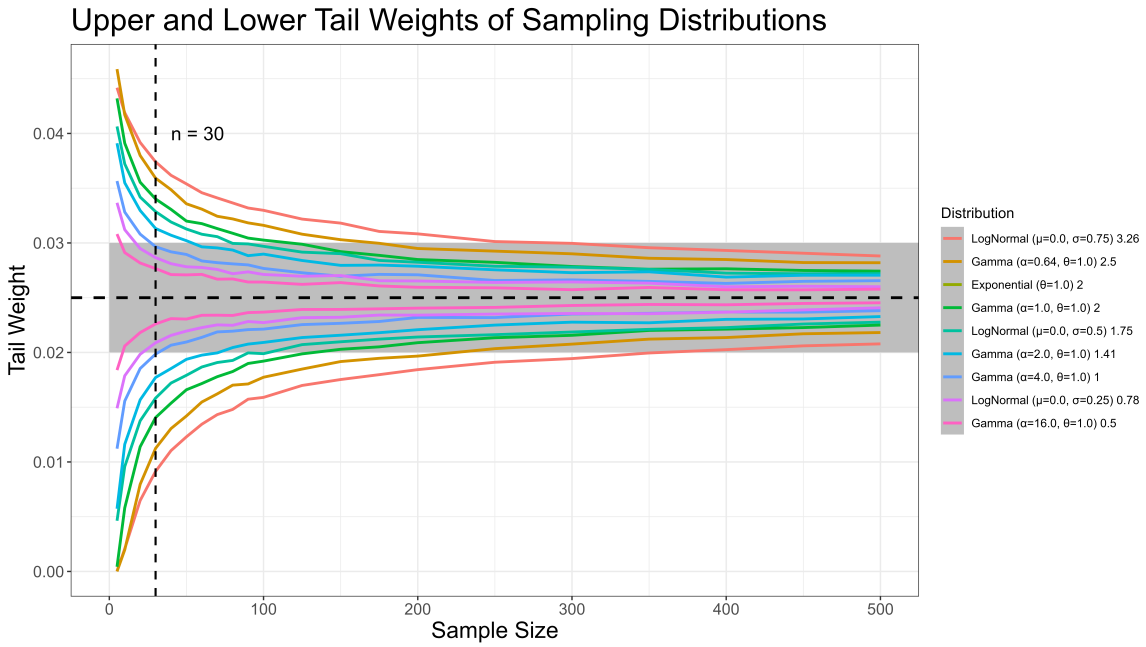

Figure 4. *Tail probabilities converge to 0.025 as sample size increases.*

This is in accord with the Berry-Esseen theorem, but is geared more towards pedagogy and practitioners in that our guideline is far more applied and most applicable to the normality check many inferential procedures require. Our guideline is not much more difficult to use 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.

One of the main things we would like to turn our focus to in the future is the bias of sample skewness. Our preliminary research suggests that something like $n \geq 36 \times (1.3 \times G_1)^2$ might suffice to account for this bias in most cases whilst still remaining a simple guideline; we would still like to look further into this however.

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; we think that 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 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, $\frac{q-p}{\sqrt{npq}}$.

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] 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.

[3] 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.

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

[5] 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).

[6] StatsBase.jl, *Statsbase.jl (version 0.34.3)*. Github, https://github.com/JuliaStats/StatsBase.jl/blob/60fb5cd400c31d75efd5cdb7e4edd5088d4b1229/src/moments.jl#L275-L280 (accessed 2024-06-15).