The Effect of Non-Normality on T-Test and Central Limit Theorem

Department of Statistics, University of Michigan, Ann Arbor, MI, USA

Zhao Shengchun

Introduction

Normality is a vital property in statistics. Many statistical tests, such as the t-test, tests for regression coefficients, and analysis of variance, all have a fundamental assumption that the sampled data comes from a normal distribution. Among these tests, the t-test must be counted among the best-known statistical procedures in current use; this test has received a surge of interest from statistical researchers over the years given its familiarity and utility. However, most datasets do not follow a normal distribution in practice; thus, it is necessary to investigate the robustness of the t-test when applied to non-normally distributed data.

Also, the central limit theorem plays an important role in statistical inference. In the meanwhile, determining the minimum sample size required to ensure the sample mean approximates a normal distribution is essential. In real research, Incorrect collection of samples makes us waste unnecessary manpower, material, and financial resources. Therefore, as for a data set that is not normally distributed, it is significant to determine how large the sample size needs to be for the central limit theorem to overcome non-normality.

Methodology

Analyzing the impact of non-normality on the t-test:

In the one-sample t-test analysis, multiple commonly seen population distributions were applied, such as Normal, Gamma, Beta, and Weibull (see Figures 1), with different parameters to simulate varying types of data distribution (symmetric, unsymmetric, monotone, etc.) in the real world. Then, we use simple sampling methods to draw the sample with different sample sizes n = 10, 50, 100, and 500, respectively. Moreover, three more complex multimodal normal distributions (see Figure 2 for their density plots) were proposed for analysis:

$$Model\ 1:\ Y1 \sim \frac{1}{2} * N(0,1) + \frac{1}{2} * N(3,1)$$

$$Model2:\ Y2 \sim \frac{3}{4} * N(-1,1) + \frac{1}{4} * N(3,\frac{1}{2})$$

$$Model3:\ Y3 \sim \frac{1}{2} * N(0,\frac{3}{4}) + \frac{3}{20} * N(3,\frac{1}{2}) + \frac{7}{20} * N(-4,\frac{3}{2})$$

As for these three models, it is too hard to find the quantile function to do simple sampling, therefore, in this case, the Metropolis-Hastings (MH) algorithm was used with the proposal distribution:

$$A = min(1, \frac{f(x^*)q(x_t|x^*)}{f(x_t)q(x^*|x_t)})\ ,\ q(x^*|x_t) = \begin{cases} x^* \sim N(x_t, 0.5),\ prob = 0.8 \\ x^* \sim Uniform(-10,10),\ prob = 0.2 \end{cases}$$

Here, we do not use simple random walk or other basic distributions as our proposal distribution since they do not work efficiently for multimodal distributions. By using the proposed distribution above, as shown in Figure 3, the MH method could do the sampling very well.

In the two-sample t-test, we referenced the generalized lambda distribution proposed by Ramberg, Dudewicz, Tadikamalla, and Mykytka (1979) in their paper *A Probability Distribution and Its Uses in Fitting Data,* the distribution is defined as follows:

$$Percentile\ Function\ R(p) = \lambda_1 + [p^{\lambda_3} - (1-p)^{\lambda_4}]/\lambda_2\ (0 \le p \le 1)$$

$$PDF\colon f(R(p)) = \lambda_2[\lambda_3 p^{\lambda_3-1} + \lambda_4(1-p)^{\lambda_4-1}]^{-1} \quad (0 \le p \le 1)$$

two extreme non-normal distributions were generated with specified λ values (see Figure 2) to test the robustness of the independent two-sample t-test by controlling the sample size (n = 10, 50, 100, 500) and whether the two sample sizes are equal or not respectively.

After that, the Monte Carlo simulation was put into use on both two types of t-tests to calculate the type I error for each sample size and compare them with the significant level 0.05, and then draw the plot.

Analyzing the sample size needs to be for the CLT to overcome extreme non-normality:

In this analysis, only some representative function distributions referenced in the last part were picked, the first one is the Beta distribution, beta(0.5, 0.5), beta(5, 1), and beta(1, 3) (see Figure 1) with sample size n = 5, 10, 25, 50, 75, 100; In the Gamma distribution, gamma(2, 0.5), gamma(9, 2), and gamma(0.5, 1) (see Figure 1) with the sample size n = 5, 10, 50, 100, 200, 500 were chosen, furthermore, we also adopted the Metropolis-Hastings (MH) algorithm to these three Gamma models with the random walks as proposal distribution, compare the results with those obtained by simple sampling methods to find out whether different sampling methods will affect the sample size that needs to overcome the non-normality issue. In addition, those three multimodal normal distributions in the previous part were also employed to do further CLT investigation with sample size n = 5, 10, 50, 100, 500, 1000.

Beyond these, we also constructed two mixture distributions by using the Gibbs sampling method with sample size n = 5, 10, 50, 100, 500, 1000:

Distribution 1:

$$X \sim Binomial(n_0, \theta), \pi(\theta) \sim Beta(a, b), n_0 = 16, a = 1, b = 2$$

The full conditional distribution in Gibbs sampling:

$$x_i^{(t+1)} | p^{(t)} \sim Binomial(n_0, p^{(t)}), i = 1, 2, 3, \dots, n$$

$$p^{(t+1)} | x_i^{(t+1)} \sim Beta(\sum x_i^{(t+1)} + a,\ n * n_0 - \sum x_i^{(t+1)} + b)$$

Distribution 2:

$$Y \sim N(\mu, \sigma^2), \mu \sim N(\mu_0, \tau^2), \sigma^2 \sim InverseGamma(\alpha, \beta), \mu_0 = 0, \tau^2 = 10, \alpha = 2, \beta = 1$$

The full conditional distributions in Gibbs sampling:

$$\mu^{(t+1)} | y_i^{(t)}, \sigma^{2(t)} \sim N(\frac{\sum y_i^{(t)} / \sigma^{2(t)} + \mu_0/\tau^2}{n/\sigma^{2(t)} + 1/\tau^2}, \frac{1}{n/\sigma^{2(t)} + 1/\tau^2})$$

$$\sigma^{2(t+1)} | y_i^{(t)}, \mu^{(t+1)} \sim Inverse\ Gamma(\alpha + \frac{n}{2}, \beta + \frac{\sum(y_i^{(t)} - \mu^{(t+1)})^2}{2})$$

$$y_i^{(t+1)} | \mu^{(t+1)}, \sigma^{2(t+1)} \sim N(\mu^{(t+1)}, \sigma^{2(t+1)})$$

Results

For the one-sample t-test, the result is satisfactory when the data is indeed normally distributed. As for the Gamma and Weibull distributions, the smaller shape values result in highly skewed distributions, causing inaccuracies in the t-test's standard error estimation for small sample sizes and then inflating Type I error values. This effect is reduced by increasing the sample size. The beta distribution's shape parameters control its skewness and kurtosis. Extreme skewness (e.g., shape1 = 1, shape2 = 3) accelerates deviations from normality at small sample sizes. This occurs because, under such parameters, much of its density is concentrated at the boundaries, leading to asymmetric and heavily tailed data, and resulting in the increasing Type I error rate (see Figure 4).

In the multi-modal model distributions, we can see that Model 1 and Model 3 are roughly symmetric, which causes the value of the Type I error of these two models to be smaller than that of the asymmetric Model 2. Moreover, the non-normality leads to the expansion of the value of Type I error. However, with the increase in sample size, the asymmetry and non-normality exert less influence on the Type I error values (see Figure 5).

For the two-sample t-test, when sample sizes are unequal, the Type I error of GLD (λ1 = 0, λ2 = 1, λ3 =2, λ4 = 3) always remains close to the significant level 0.05, but in equal sample sizes, The smaller the sample size, the greater the deviation produced by Type I error. However, as for the GLD (λ1 = 0.166, λ2 = 0.5901, λ3 = 1.7680, λ4 = 1.1773), the result is the exact opposite (see Figure 6).

In the central limit theorem analysis, for the unimodal distribution (Beta, Gamma), unsymmetric distributions, and some distributions with heavy tails, such as Gamma (2, 0.5) or Beta (5, 1), demand larger sample sizes because the heavy tails which contain extreme values (high or low) more frequently. These extreme values could significantly affect the sample mean, the CLT only can smooth these effects by consuming larger sample sizes (see Figure 7-8). Furthermore, comparing the sampling method for three Gamma distribution models from Figure 7 (simple sampling) and Figure 8 (MH sampling), the result is obvious, the MH sampling with random walk proposal distribution does a much worse job than the simple sampling.

Regarding multimodal distribution, for the symmetric and homogeneous distribution (model 1), the QQ plots aligned closely with the theoretical normal distribution when n increased to 100, indicating that the CLT could quickly eliminate the effect of bimodality. However, for unsymmetric multimodal and heterogeneous distributions (model 2, model 3), Only at extremely massive sample sizes (n=500,1000) do the effects of multimodality and heterogeneity diminish, resulting in near-normal sample means. This slower convergence is primarily due to the high variance of some components and the influence of the extreme tail (see Figure 10).

In regard to the mixture models, both of the models reveal a slow convergence to the normal distribution of their sample mean when we increase the sample size, this may be due to the heavy tail nature of beta and inverse gamma distributions caused these substantial deviations, from the QQ-plots, we can find the 1000 sample size is far from enough to remove that effect, maybe a new advanced sampling method should be needed to use or use larger sample size (see Figure 11).

Discussion

The performance of the t-test under non-normal conditions is influenced by a combination of factors, including skewness, kurtosis, and sample size. Extreme skewness and kurtosis can lead to biased estimates of the mean and variance, disrupting the t-test's performance. However, increasing the sample size can highly reduce this effect and ensure the robustness of the test. Besides, in the CLT analysis, while the CLT ensures eventual convergence to normality for the models in most cases, the rate of convergence depends on the distribution's complexity. Symmetric and balanced distributions converge rapidly, while asymmetric, variance-heterogeneous, and hierarchical distributions require larger sample sizes for reliable normal approximations. Besides, if we can get the quantile function of the data, it is better to use the simple sampling method which is more efficient to draw the sample, instead of other complex sampling methods.

Therefore, when applying both t-test and CLT analysis to practical problems, researchers should assess the characteristics of the original data distributions and may adopt some beneficial methods, including efficient data sampling, transformations, or larger sample size.

GitHub URL

https://github.com/ZHAOShengchun67734538/STAT-506-Final-Project

Reference

Ramberg, J. S., Dudewicz, E. J., Tadikamalla, P. R., & Mykytka, E. F. (1979). A Probability Distribution and its Uses in Fitting Data. *Technometrics*, 21(2), 201–214. https://doi.org/10.1080/00401706.1979.10489750

Appendix

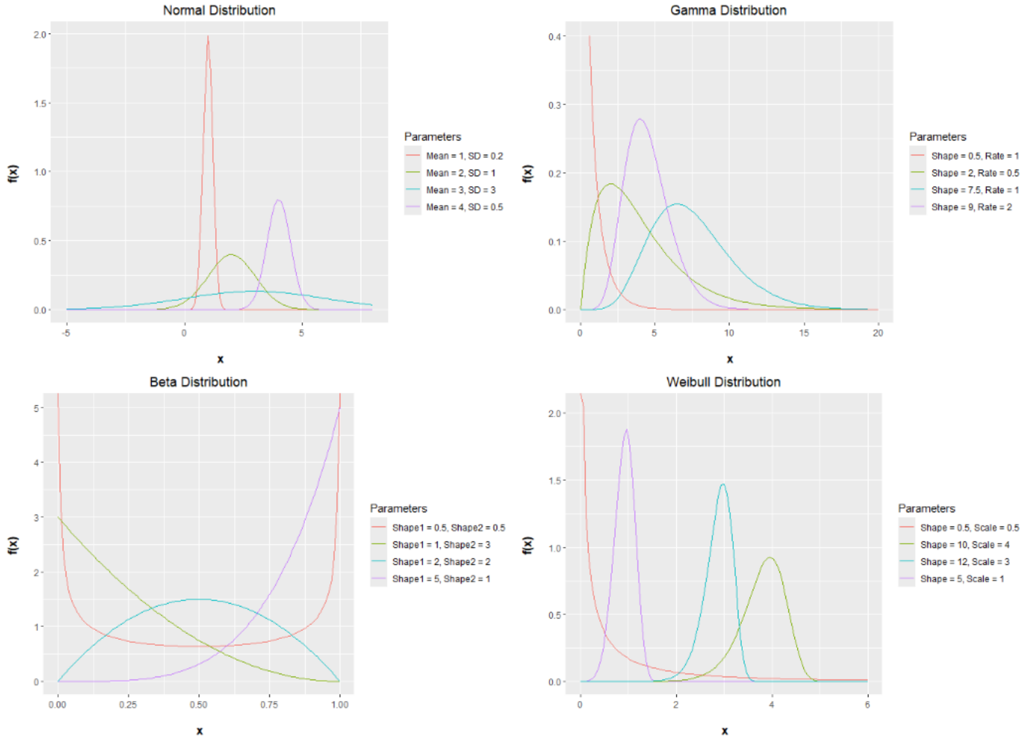

Figure 1

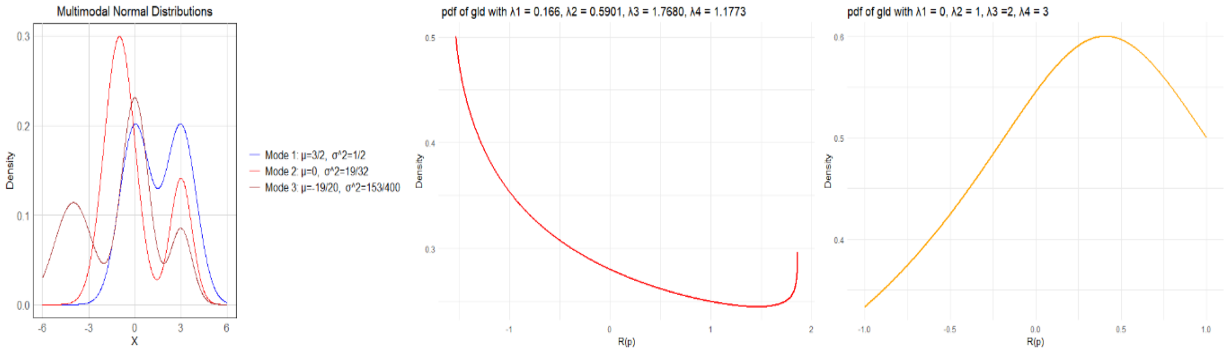

Figure 2

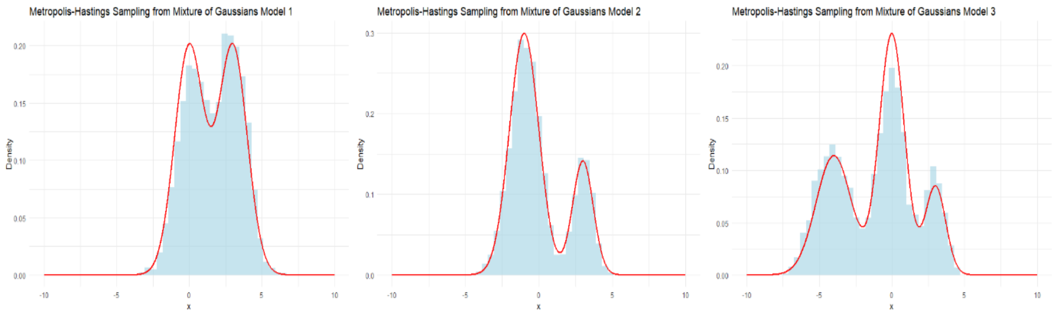

Figure 3

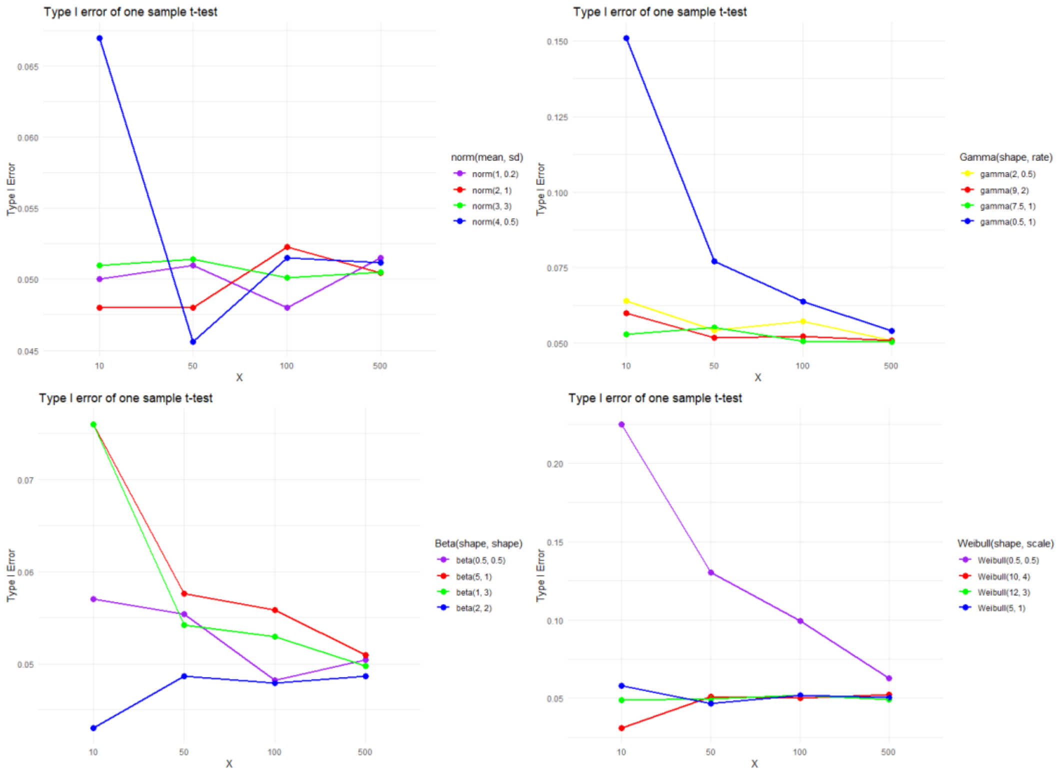

Figure 4

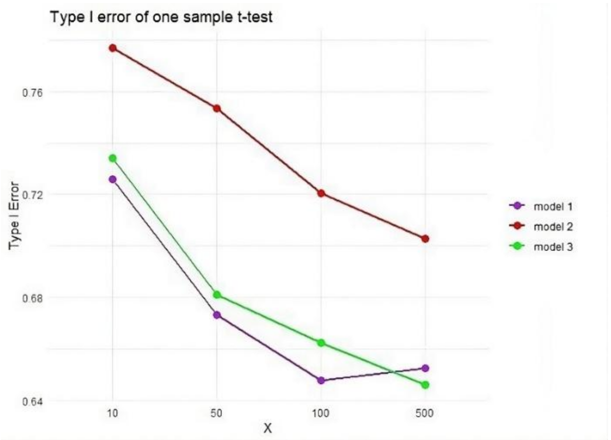

Figure 5

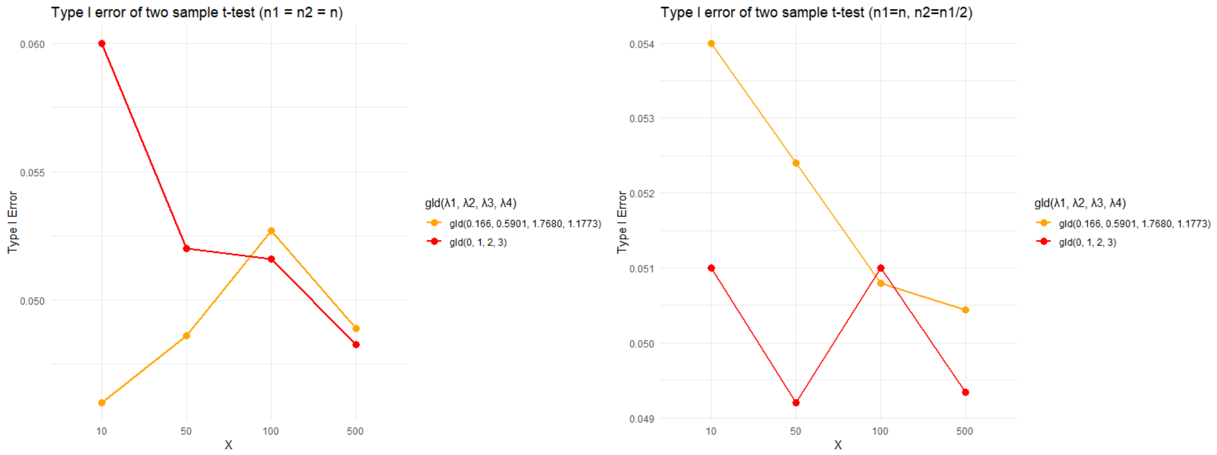

Figure 6

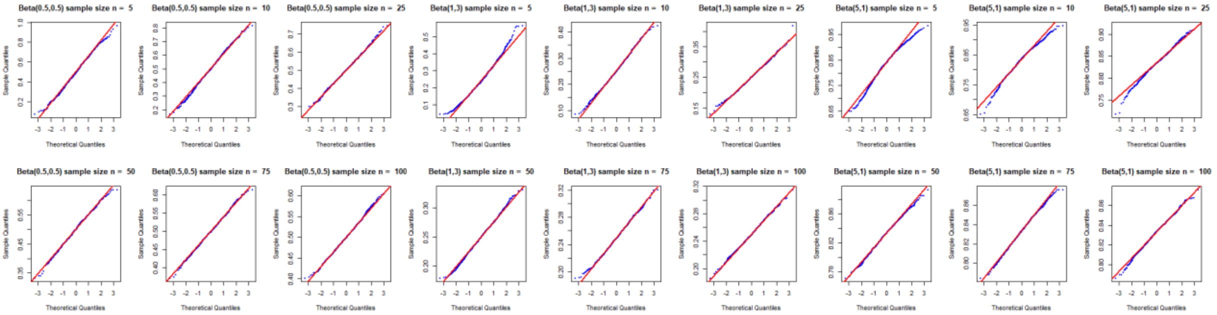

Figure 7

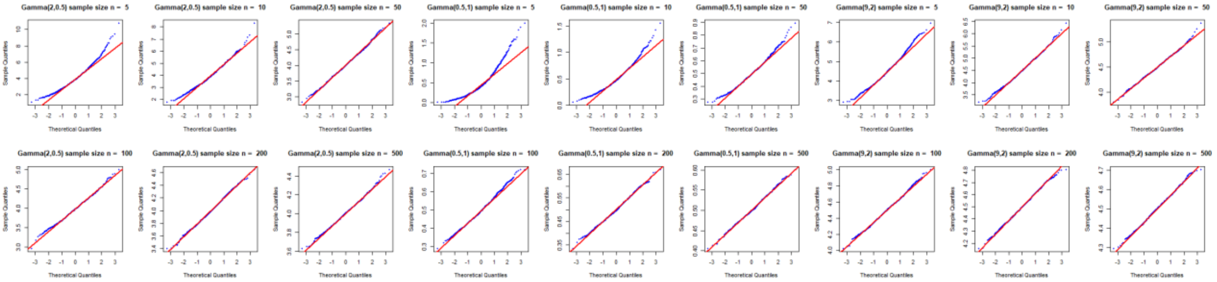

Figure 8

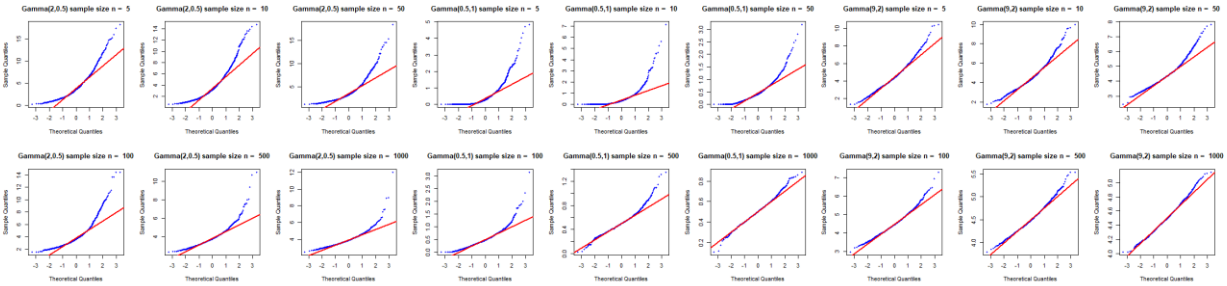

Figure 9

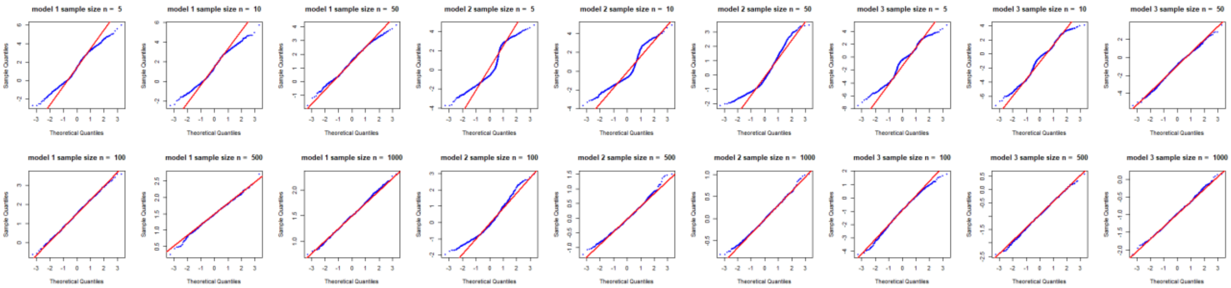

Figure 10

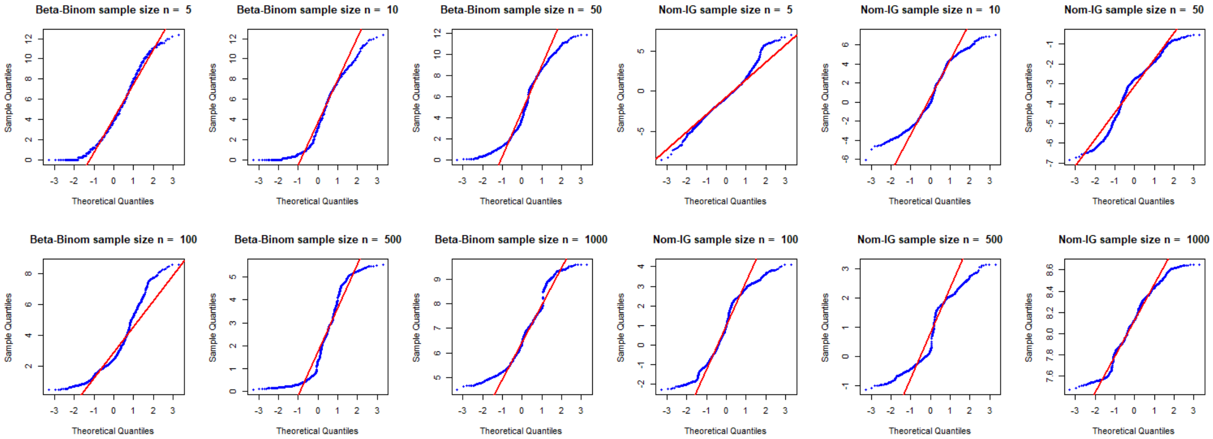

Figure 11