Evaluating the One-Sample t-Test Under Non-Normality

Junjie Zeng

Github Repository: https://github.com/JunjieZeng1/STATS-506.git

Introduction

The t-test is one of the most important tools in statistics. One-sample t-test, like many other tests, assumes that the data have been drawn from a normally distributed population. However, this assumption is relaxed by the presence of the Central Limit Theorem(CLT), which guarantees an approximately normally distributed mean of the sampling distribution for large samples. This paper strives to investigate the extent to which non-normality impacts the Type I error rate of the one-sample t-test, the effectiveness of the CLT in mitigating these issues, and the sample sizes at which these effects are minimized.

Methodology

To answer the research questions, the performance of the one-sample t-test is evaluated through following steps:

1. **Define distributions:**
    - Various distributions were carefully designed to have mean 0 for convenience: uniform distribution; normal distribution; t-distribution with degrees of freedom 1, 2, 5, and 10; exponential distribution with lambda 0.5, 1, 2, and 5; bimodal normal distribution.
2. **Sample sizes:**
    - Sample sizes of 5, 10, 20, 50, 100, 200, 500, 1000 and 10,000 were chosen.
3. **Simulation:**
    - Generate data according to distribution and sample sizes.
    - Conduct t-tests on each sample, testing the null hypothesis that the mean equals zero at significance level $\alpha = 0.05$.
    - Calculate the type I error rates for each sample.
4. **Visualization and Analysis**

- Create table and line chart of type I errors for different distributions
- Analyze based on tables and charts

Results and Analysis

Uniform distribution

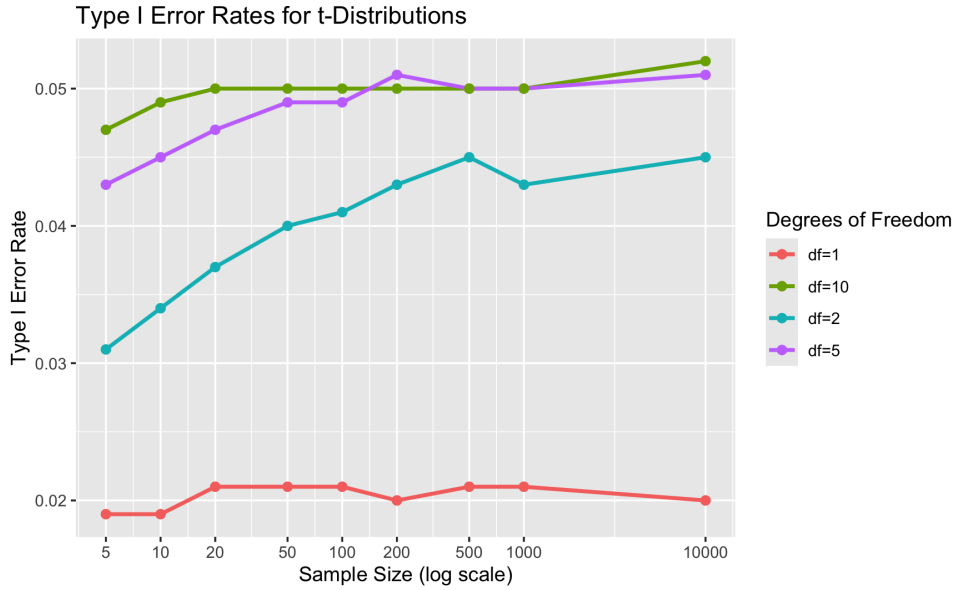
From the graph (see Appendix), the CLT overcomes non-normality for small samples for uniform distribution. The type I error rate for sample size 10 is 0.055, which is fairly close to 0.05.

Normal distribution

For normal distribution, the type I error is close to 0.05 for all sample sizes (see Appendix). This aligns with our knowledge since the one-sample t-test assumes normality in the population. Since the population distribution aligns perfectly with this assumption, the test performs optimally, regardless of sample size.

T-distribution

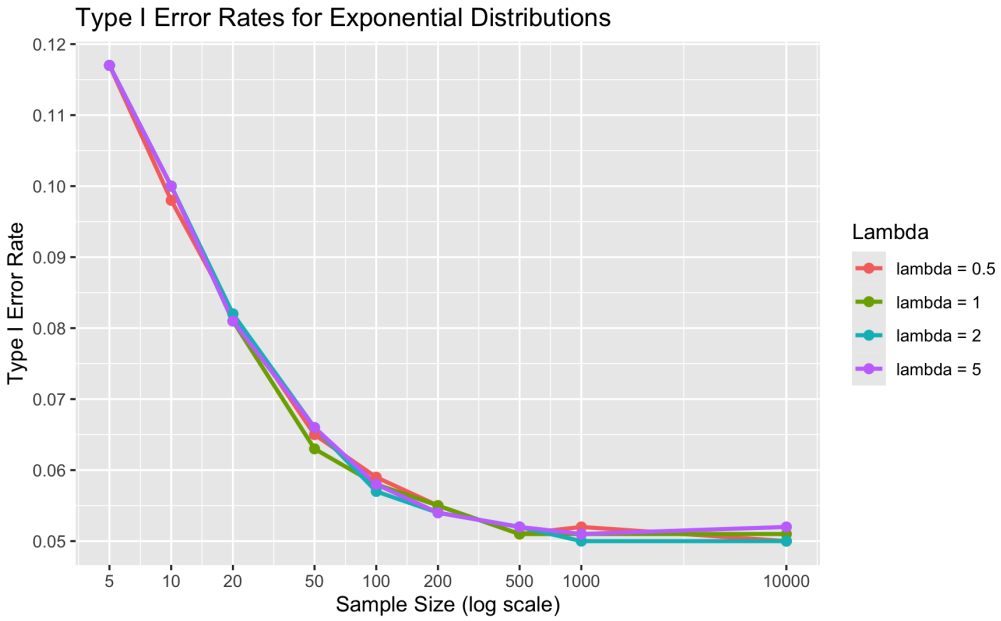
From the line chart, we see that for heavy-tailed distributions (df=1, 2), type I error rates are significantly below 0.05 for small samples (n ≤ 20), stabilizing near 0.05 only for n > 200. For 1 degree of freedom, the CLT won't work even for very large samples (10,000). For heavy-tailed distributions, the CLT worked for less sample sizes.



Exponential Distribution

The exponential distribution is highly skewed, violating the t-test's assumption of approximate normality. At small sample sizes (n=5, 10, 20, 50, 100), Type I error rates are significantly higher than the nominal 0.05. The CLT starts to overcome

non-normality when the sample size is larger than 200. When sample size is about 500, the type I error rate is very close to 0.05.



Bimodal Normal Distribution

For bimodal normal distribution (see Appendix), type I error rates are inflated (~0.07) for small sample sizes due to the t-test's susceptibility to non-normality but rapidly decline and quickly stabilize near the nominal level by n = 50, where the CLT started to play a role. Eventually, for larger sample sizes (n > 500), minor fluctuations around 0.05 are observed but negligible. The t-test is robust for n ≥ 50.

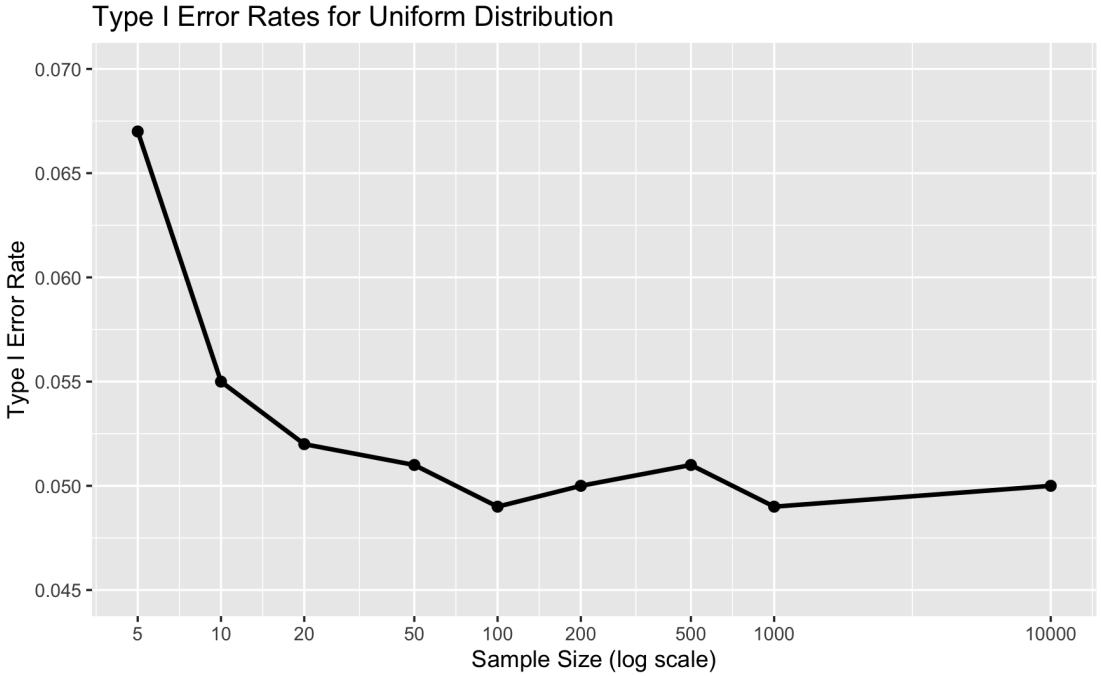
Conclusion

This study investigated the performance of one-sample t-test under non-normality. The one-sample t-test performs strongly for both normally distributed and uniformly distributed data, and small sample size (n ≥ 10) did not affect such performance. For both t-distributed and exponential distribution, Type I error rates stabilize towards 0.05 for larger sample sizes of n > 200. The t-test is robust for the bimodal distribution at n ≥ 50. Such findings again underlined an impact of the proper choice of sample size and were pointing towards further research on other methods for the cases of small samples with highly non-normal population.

Appendix

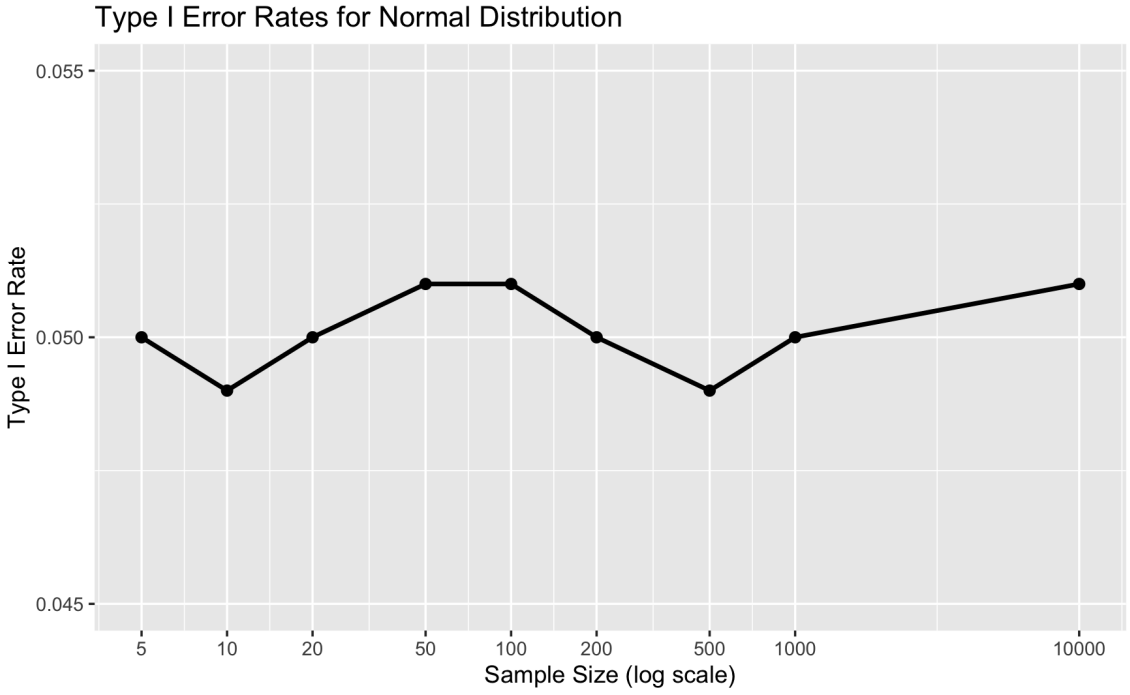
Uniform distribution

SampleSize <dbl>	TypeIError <dbl>
5	0.067
10	0.055
20	0.052
50	0.051
100	0.049
200	0.050
500	0.051
1000	0.049
10000	0.050



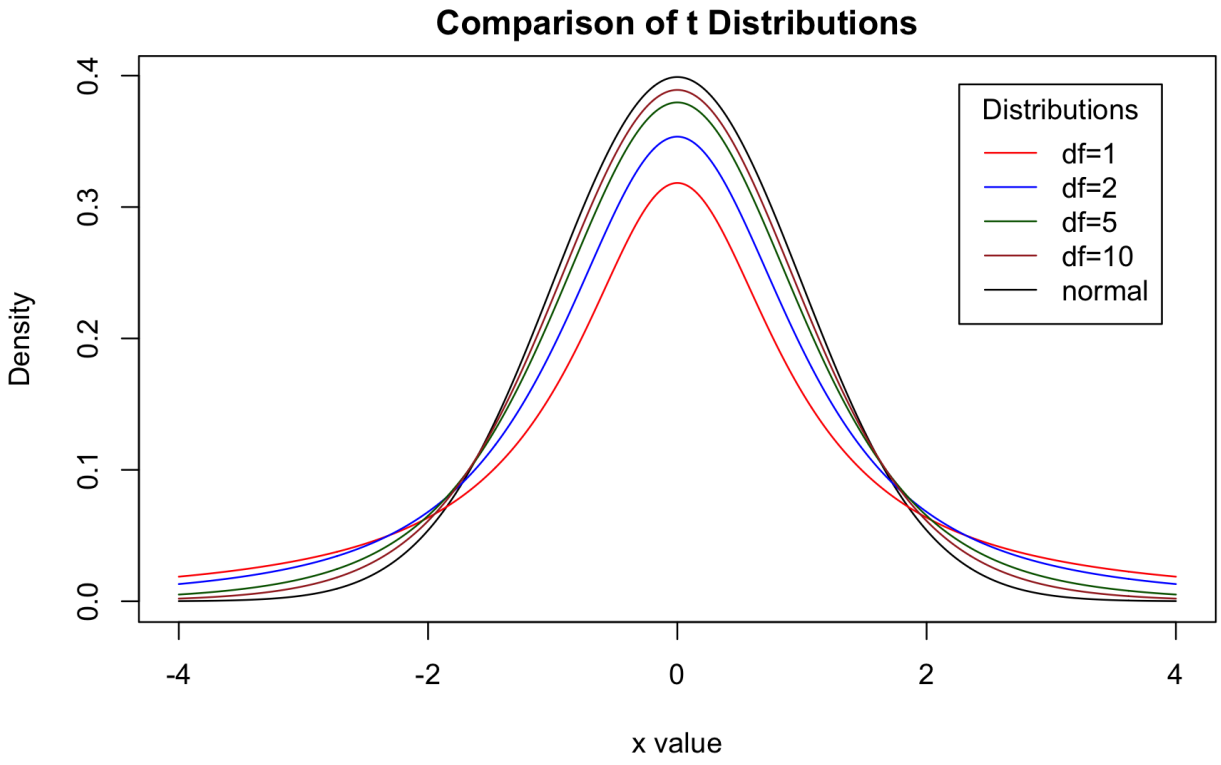
Normal distribution

SampleSize <dbl>	TypeIError <dbl>
5	0.050
10	0.049
20	0.050
50	0.051
100	0.051
200	0.050
500	0.049
1000	0.050
10000	0.051



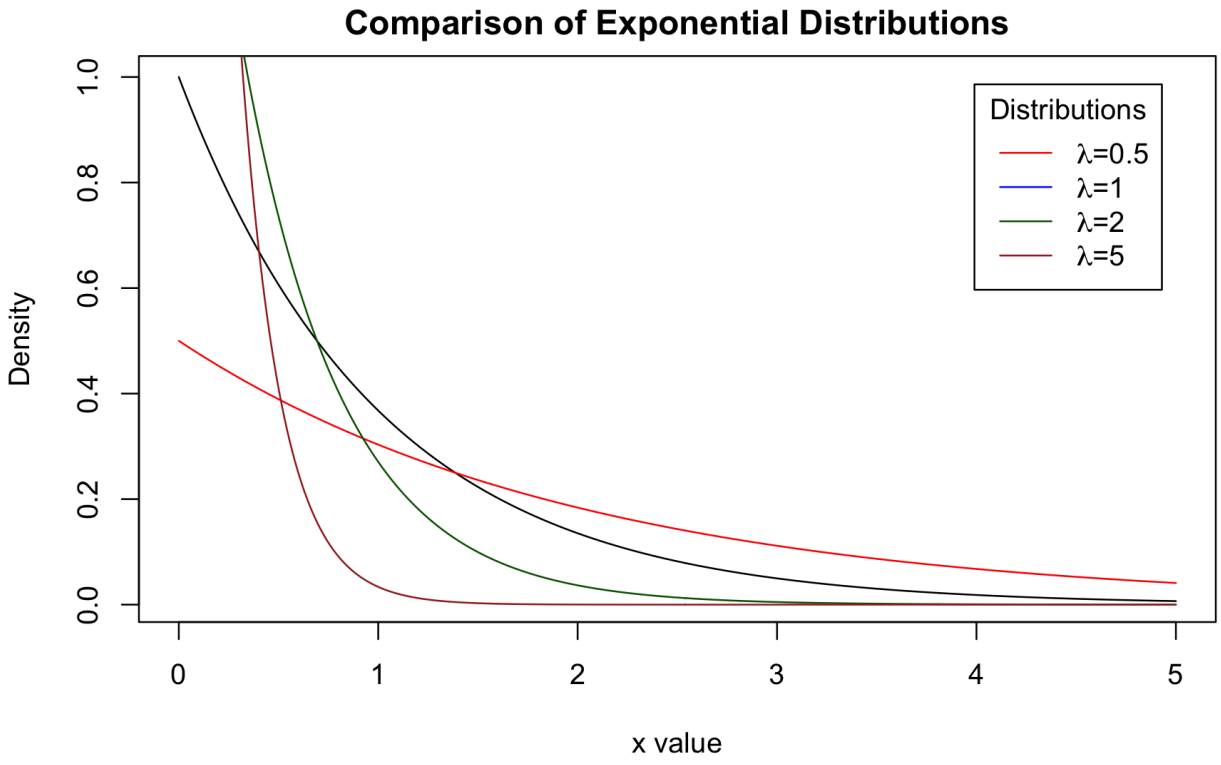
T-distribution

SampleSize <dbl>	df=1 <dbl>	df=2 <dbl>	df=5 <dbl>	df=10 <dbl>
5	0.019	0.031	0.043	0.047
10	0.019	0.034	0.045	0.049
20	0.021	0.037	0.047	0.050
50	0.021	0.040	0.049	0.050
100	0.021	0.041	0.049	0.050
200	0.020	0.043	0.051	0.050
500	0.021	0.045	0.050	0.050
1000	0.021	0.043	0.050	0.050
10000	0.020	0.045	0.051	0.052

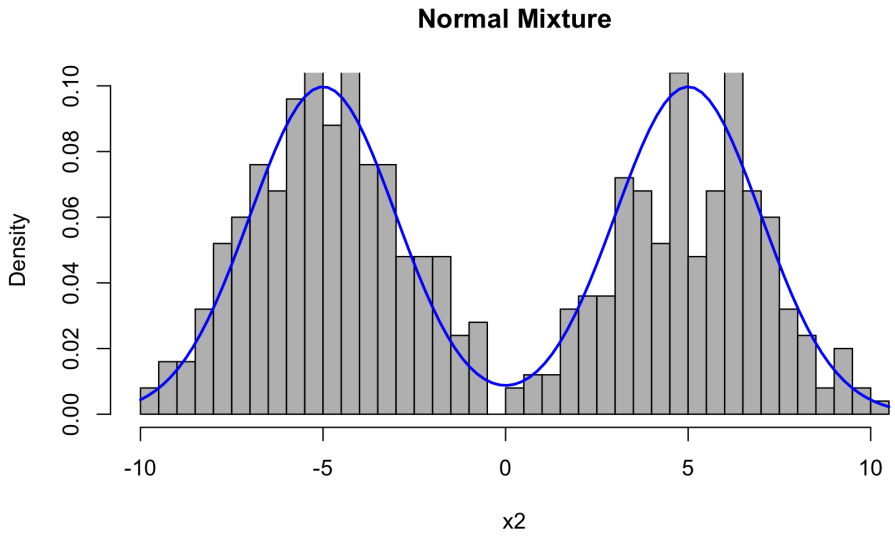


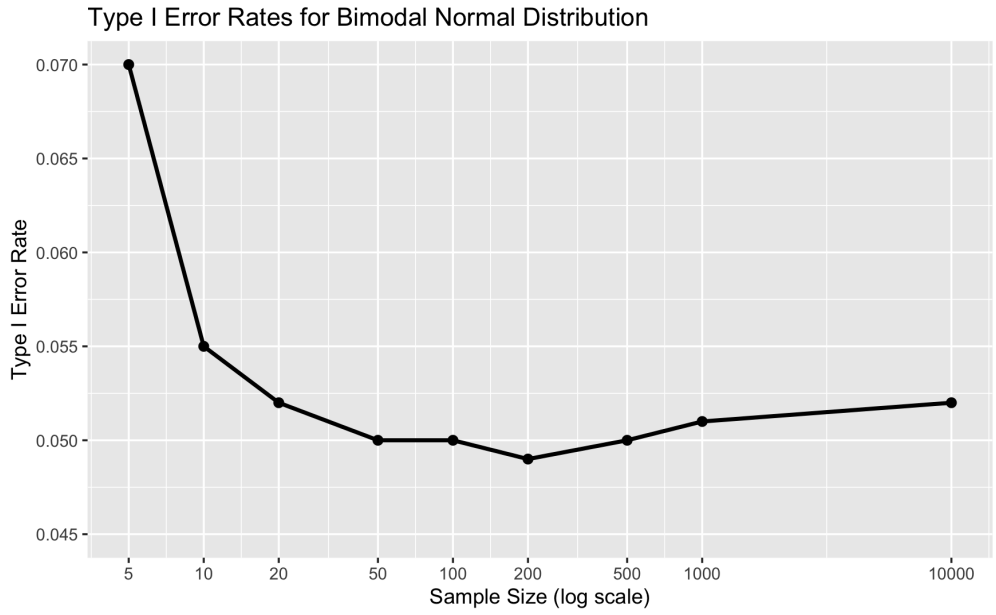
Exponential distribution

SampleSize <dbl>	lambda = 0.5 <dbl>	lambda = 1 <dbl>	lambda = 2 <dbl>	lambda = 5 <dbl>
5	0.117	0.117	0.117	0.117
10	0.098	0.100	0.100	0.100
20	0.082	0.081	0.082	0.081
50	0.065	0.063	0.066	0.066
100	0.059	0.058	0.057	0.058
200	0.055	0.055	0.054	0.054
500	0.051	0.051	0.052	0.052
1000	0.052	0.051	0.050	0.051
10000	0.050	0.051	0.050	0.052



Bimodal Normal Distribution





SampleSize <dbl>	TypeIError <dbl>
5	0.070
10	0.055
20	0.052
50	0.050
100	0.050
200	0.049
500	0.050