Reproduction of Schucany and Ng's Paper

Ming Zhang

Yifan Lu

October 25, 2019

Abstract

One of the most basic topics in statistics is the inference for the population mean. Although t-test is robust on departure from non-normal distribution, there are many software packages to test the normality before doing t-test, and graphical assessments are also used to detect, such as quantile-quantile plot or scatter plot. We agree with those recommendations. To well understand the effect of preliminary goodness-of-fit test and further t-test, we did a simulation study using Shapiro-Wilk statistics, W as the convincing evidence to test the normality. We analyze the results of systematically screening all samples from standard normal and uniform. This pretest at large significance level will not help do the t-test, and we recommend that use very low level such as 0.1% or, in practice, use graphical diagnostics to substitute the pretest.

1 Background

The Gaussian assumption was a good idea to be tested before making the statistical inference about the population mean by using a sample $X_1, X_2, ..., X_n$. There are many software packages to check the normality before doing the t test, such as PROC UNIVARIATE in SAS or shapiro.test and ks.test in R. The assumption can be also judged by informal graphical diagnostics. For example,Schafer (2002) claim that formal preliminary test is not necessary.

Suppose the sample of size n is from distribution F with population mean μ. The goodness-of-fit (GOF) test at significance level α_g will be:

$$\begin{aligned} H_0^* &: \text{The true } F \text{ is normal} \\ \text{against } H_1^* &: \text{F is not normal} \end{aligned} \tag{1}$$

If one do not reject the null hypothesis, then treat the sample as from a normal distribution and do the following Student t-test:

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}, \text{ where } \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \text{ and } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

and corresponding hypothesis at the significant level α_t:

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ \text{against } H_1^* &: \mu \neq \mu_0 \end{aligned} \tag{2}$$

For assessing the necessity of the preliminary goodness-of-fit test for normality, we compare the true Type-I error rate after passing the pretest:

$$\alpha = \Pr(\text{Reject } H_0 | \text{ do not reject } H_0^* \text{ and } H_0 \text{ is true}) \tag{3}$$

to the pre-defined nominal Type-I error rate α_t.

There are many ways to test the normality for the first preliminary goodness-of-fit test, and, according to Schucany and Ng (2006), for instance, many of them based on moments, probability plots, or other empirical distribution tests, which lead people not recommending Kolmogorov-Smirnov. Also, Tukey cited Michael (1983) for transformation better suited to the sup norm. Therefore, we decide to use Shapiro-Wilk statistic, W Shaphiro and Wilk (1965) for normality, which has also roughly same results with Anderson-Darling statistics, A^2. So, we only report W for following simulation process.

2 Methods

According to Schucany and Ng (2006), the algorithm for getting the estimated Type-I error rate in (3) follows:

- Step1. Simulate a random sample of size n from a distribution F
- Step2. Use the Shapiro-Wilk statistic, W, to test the normal assumption at α_g significant level.
- Step3. Continue t test at α_t significant level if null hypothesis is not rejected in Step2. If H_0 is rejected in Step2, then go back Step1.

In this simulation, the sample size n = 10,20,30 and 50 from two underlying distribution, (i) uniform (0,1), μ_0=0.5, (ii) standard normal,μ_0=0. Five fixed levels of significance will be set for preliminary goodness-of-fit test α_g=10% ,5% ,1% ,0.5% ,0.1% and crossed with four levels of significance for t-test α_t=10% ,5% ,1% ,0.5% . For each combination of $n, \alpha_g, \alpha_t$and F, we need independently repeat all steps until the null hypothesis is not rejected in Step2 M = 100,000 times, which means that we need have M times t-test in Step3. Then, the Type-I error rate in (3) can for t-test can be estimated by:

$$\hat{\alpha} = \frac{\text{number of times } H_0 \text{ (2) is rejected}}{M} \tag{4}$$

3 Result

Table 1-2 in the Appendix summarize the estimated Type-I error rates under five pretest levels and without any pretest, four different sample size, four conventional levels of significance α_t, and under two different distributions of uniform and standard normal. These estimated α_t are estimated by 100,000 replications of the Student t-statistic. The nominal SE have been show at head of each column by equation $(\alpha_t * (1 - \alpha_t)/\sqrt{M})$. The number of simulations to get final 100,000 replications of t-test are achieved by the power list in the last column. For example, for Table 2, α_g = 0.1 and n=50, power=88%, implying 100,000/(1-0.88) = 833,333 samples to produce 100,000 cases passing goodness-of-fit pretest.

We are using the standard normal distribution as our benchmark to compare whether there are any difference from the uniform distribution. From Fig.2, we can see the estimated α_t values all fluctuate around 5% under all sample sizes and α_g settings, which makes sense because our underlying distribution for that is standard normal and nominal pre-defined α_t is equal 5% for this graph and pretest do not hurt distribution even the significance level is extremely small (0.1 %). From Table 2, all values from the third to sixth column are very close to the α_t we have already defined. For the power of Shapiro-Wilk test, the values are also very close to α_g we have set. In the last row of Table 2, w/o pretest, the values are very still close to α_t because the distribution is normal and pretest is not hurting it.

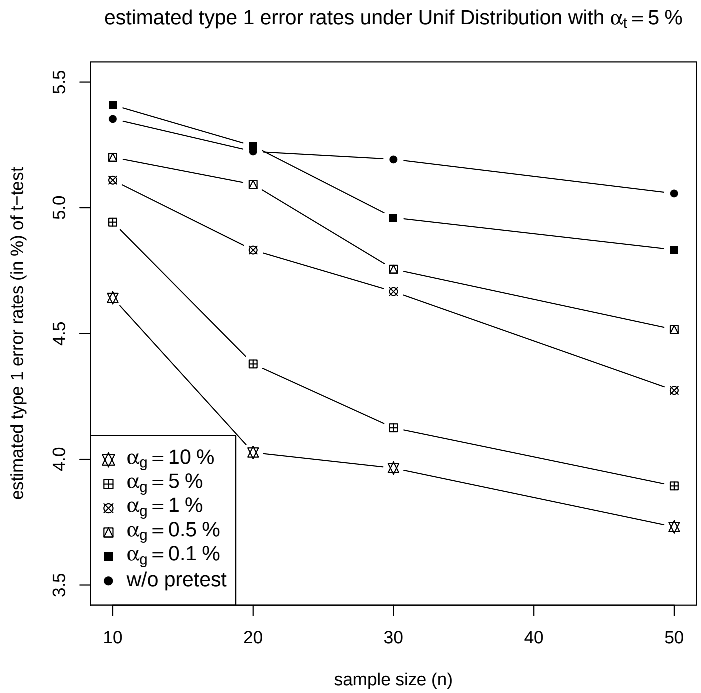

Figure 1

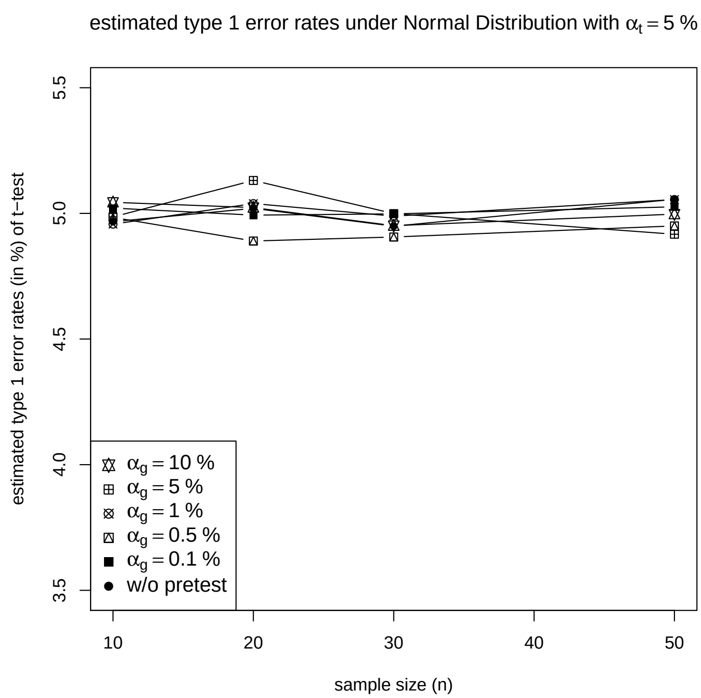

Figure 2

When F is uniform distribution, however, from Table 1, the estimated Type-I error rates are departing from the nominal α_t and it get slightly better for small α_g when the sample size increase but it will get worse when sample size is too large. We can find that under all α_g settings, the estimated Type-I errors tend to decrease with the increase of sample sizes. And when α_g is large, the estimated Type-I error rates get worse when sample size become larger. For instance, when α_g= 10 % and α_t=10 %, estimated Type-I error rate decreases from 8.71 % to 7.81 %. The power gets bigger when the size become larger for five levels of α_g. However, when $\alpha_g = 0.1\%$, most of values perform much better than the bigger significant levels. For the section of without pretest, if sample size is small, the differences between nominal α_t and estimated Type-I error rates are bigger than pretest with lower significant level, such as 0.1%, but it will get better when sample size become bigger.

Therefore, same pattern can also be seen in Fig 1, which represent the change for estimated Type-I error with the change of sample size under α_t=5% and use different symbols to denote different α_g levels. It's clear that when sample size is small, 10 and 20, all α_g levels work well, but with the increasing of the sample size, only α_g=0.1% and w/o pretest is still close to 5%, which is good.

4 Conclusion

Based on the results here, we recommend: (1). If the population follows normal distribution, pretest is not hurting the test but not needed to use. (2)(i). If we do not know what distribution but suppose to be normal, and then if sample size is small, use pretest significance level at a very low level, say $\alpha_g = 0.1\%$. (2)(ii). If sample size is fairly large, use either pretest at 0.01% significance level or just not use any pretest, which can substitute to the graphical assessments, such as quantile-quantile plot. In practice, according to Schucany and Ng (2006), graphical diagnostics are better than a formal pretest. Furthermore, rank or permutation methods are recommended for exact validity in the symmetric case.

References

Michael, J. R. (1983). The stabilized probability plot. *Biometrika*, 70(1):11–17.

Schafer, D. W. (2002). *The statistical sleuth: a course in methods of data analysis*. Duxbury/Thomson Learning.

Schucany, W. R. and Ng, H. K. T. (2006). Preliminary goodness-of-fit tests for normality do not validate the one-sample student t. *Communications in Statistics*, 35(12):2275–2286.

Shaphiro, S. and Wilk, M. (1965). An analysis of variance test for normality. *Biometrika*, 52(3):591–611.

5 Appendix

Tables following are the simulation results of estimated Type-I errors under Uniform (0,1), Standard Normal and Exponential distribution with μ=1.

Table 1

Simulated Type-I error rates (in %) of the t-test after acceptable Shapiro–Wilk tests of normality under Uniform Distribution

α_g (in %)	n	$\alpha_t = 10$ SE = 0.10	$\alpha_t = 5$ SE = 0.07	$\alpha_t = 1$ SE = 0.03	$\alpha_t = 0.5$ SE = 0.02	Power of Shapiro–Wilk test
10	10	8.711	4.642	1.260	0.755	17.255
	20	8.290	4.027	0.877	0.483	36.015
	30	8.214	3.965	0.762	0.371	57.652
	50	7.806	3.731	0.654	0.323	88.024
5	10	9.127	4.943	1.345	0.768	8.038
	20	8.841	4.379	0.951	0.510	19.892
	30	8.438	4.125	0.847	0.431	38.282
	50	8.147	3.894	0.731	0.367	74.842
1	10	9.647	5.110	1.330	0.743	1.158
	20	9.615	4.832	1.081	0.556	3.022
	30	9.347	4.667	0.989	0.521	9.057
	50	8.758	4.274	0.840	0.456	35.496
0.5	10	9.854	5.201	1.365	0.812	0.460
	20	9.833	5.093	1.101	0.589	1.102
	30	9.492	4.756	1.024	0.530	3.884
	50	9.208	4.516	0.910	0.457	21.537
0.1	10	10.006	5.409	1.375	0.783	0.040
	20	9.952	5.246	1.224	0.674	0.061
	30	9.835	4.961	1.049	0.560	0.321
	50	9.758	4.833	1.013	0.513	4.006
w/o pretest	10	10.016	5.354	1.407	0.814	0.000
	20	10.080	5.224	1.188	0.663	0.000
	30	10.175	5.193	1.135	0.583	0.000
	50	10.044	5.057	1.044	0.533	0.000

Table 2

Simulated Type-I error rates (in %) of the t-test after acceptable Shapiro–Wilk tests of normality under Normal Distribution

α_g (in %)	n	$\alpha_t = 10$ SE = 0.10	$\alpha_t = 5$ SE = 0.07	$\alpha_t = 1$ SE = 0.03	$\alpha_t = 0.5$ SE = 0.02	Power of Shapiro–Wilk test
10	10	10.047	5.044	1.056	0.524	9.830
	20	9.966	5.023	1.014	0.520	10.180
	30	10.044	4.951	0.979	0.480	10.070
	50	9.986	4.997	1.029	0.509	9.922
5	10	9.859	4.986	0.984	0.489	5.046
	20	10.107	5.131	1.001	0.506	5.075
	30	9.976	4.999	0.955	0.510	5.037
	50	9.975	4.917	0.985	0.506	5.063
1	10	9.980	4.957	0.928	0.450	1.063
	20	10.076	5.039	0.915	0.470	0.942
	30	10.070	4.989	0.999	0.497	0.890
	50	9.922	5.055	1.029	0.493	0.976
0.5	10	9.966	4.984	1.010	0.466	0.530
	20	9.807	4.890	0.974	0.505	0.462
	30	10.024	4.906	0.976	0.483	0.508
	50	9.943	4.950	0.967	0.477	0.532
0.1	10.005	10.005	5.021	1.036	0.536	0.118
	10.071	10.071	4.993	1.019	0.530	0.083
	10.014	10.014	4.998	1.023	0.502	0.093
	10.048	10.048	5.026	1.034	0.533	0.121
w/o pretest	10	9.897	4.967	0.968	0.486	0.000
	20	9.938	5.020	0.983	0.482	0.000
	30	9.952	4.949	0.996	0.508	0.000
	50	10.086	5.056	0.997	0.491	0.000