

# Power Analysis for $t$ -Test with Non-normal Data and Unequal Variances

Han Du, Zhiyong Zhang, and Ke-Hai Yuan

**Abstract** A Monte Carlo-based power analysis is proposed for  $t$ -test to deal with non-normality and heterogeneity in real data. The step-by-step procedure of the proposed method is introduced in the paper. For comparing the performance of the Monte Carlo-based power analysis to that of conventional pooled-variance  $t$ -test, a simulation study was conducted. The results indicate the Monte Carlo-based power analysis provided well-controlled empirical Type I error rate, whereas the conventional pooled-variance  $t$ -test failed to yield nominal-level Type I error rate. Both an R package and its corresponding online interface are provided to implement the proposed method.

**Keywords** Power analysis • Monte Carlo simulation • Non-normality • Heterogeneity

Power analysis is widely used for sample size determination (e.g., Cohen 1988). With appropriate power analysis, an adequate but not “too large” sample size is determined to detect an existing effect. The conventional method for power analysis for the  $t$ -test is limited by two strict assumptions: normality and homogeneity (two-sample pooled-variance  $t$ -test). The two-sample separated-variance  $t$ -test (also known as the Welch’s  $t$ -test; Welch 1947) tolerates heterogeneity but still assumes normally distributed data. Thus, the corresponding exact power solution for the separated - variance  $t$ -test assumes normality with either numerical integration of noncentral density function or approximation (Moser et al. 1989; Disantostefano and Muller 1995).

Practical data in social, behavioral, and education research are rarely normal or homogeneous (Blanca et al. 2013; Micceri 1989). This poses challenges on statistical power analysis for the  $t$ -test (Cain et al. in press). To deal with the

---

H. Du (✉) • K.-H. Yuan

Department of Psychology, University of Notre Dame, 123A Hagger Hall, Notre Dame, IN 46556, USA

e-mail: [hdu1@nd.edu](mailto:hdu1@nd.edu)

Z. Zhang

University of Notre Dame, 118 Hagggar Hall, Notre Dame, IN 46556, USA

problems, we develop a general method to conduct power analysis for  $t$ -test through Monte Carlo simulation. The method can flexibly take into account non-normality in one-sample  $t$ -test, two-sample  $t$ -test, and paired  $t$ -test and unequal variances in two-sample  $t$ -test. We provide an R package as well as an online interface for implementing the proposed Monte Carlo-based power analysis procedure.

## 1 One-Sample $t$ -Test

The one-sample  $t$ -test concerns whether the population mean  $\mu$  is different from a specific target value  $\mu_0$  (usually  $\mu_0 = 0$ ). Thus, the null hypothesis is

$$H_0 : \mu = \mu_0.$$

The alternative hypothesis can be either two-sided ( $H_{a1}$ ) or one-sided ( $H_{a2}$  or  $H_{a3}$ ):

$$H_{a1} : \mu \neq \mu_0,$$

$$H_{a2} : \mu > \mu_0,$$

$$\text{or } H_{a3} : \mu < \mu_0.$$

The statistic given sample size  $n$ ,  $t = \frac{\bar{y} - \mu_0}{s \sqrt{\frac{1}{n}}}$ , follows a  $t$  distribution with degrees of freedom  $n - 1$  under the normality assumption, where  $s$  is the sample standard deviation. When the normality assumption is violated, the  $t$  statistic does not follow a  $t$  distribution any more. When sample size increases, the statistic approximately follows a normal distribution. However, power analysis is less meaningful with a huge sample size because the power would be always 1.

Non-normality can take many forms. In this study, we focus on continuous variables with skewness and kurtosis different from a normal distribution (e.g., Cain et al. [in press](#)). With such non-normal data, it is extremely difficult to use an analytical formula to calculate power as in traditional power analysis. Instead, a Monte Carlo simulation method can be conveniently used (e.g., Muthén and Muthén 2002; Zhang 2014). The basic procedure of the Monte Carlo method is to first simulate the empirical null distribution of a chosen test statistic with the first four moments under the null distribution to get the critical value for null hypothesis testing and then simulate the distribution of the test statistic under the alternative hypothesis. Finally, the power can be estimated using the empirical distribution under the alternative hypothesis and the empirical critical value.

To use the Monte Carlo method, information regarding the first four moments is needed. Specifically, we need the population mean ( $\mu$ ) and standard deviation ( $\sigma$ ). In addition, we need the population skewness  $\gamma_1 = E\left[\left(\frac{x-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3}$  and kurtosis  $\gamma_2 = E\left[\left(\frac{x-\mu}{\sigma}\right)^4\right] = \frac{\mu_4}{\sigma^4}$ . For testing the population mean, the means under the null

and alternative hypotheses should be different, denoted by  $\mu_0$  and  $\mu_1$ , respectively. However, we assume that the shapes of distributions under the null and alternative are the same with the same standard deviation, skewness, and kurtosis in this study although they can be different. In practice, the population statistics are unknown, but they can be decided based on meta-analysis or literature review (e.g., Schmidt and Hunter 2014).

For the one-sample test, the following step-by-step procedure can be used to obtain the power for a given sample size  $n$  for testing:

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu = \mu_1.$$

1. Given the mean ( $\mu_0$ ), standard deviation ( $\sigma$ ), skewness ( $\gamma_1$ ), and kurtosis ( $\gamma_2$ ), generate  $R_0$  sets of non-normal data, each with the sample size  $n$ .  $R_0$  should be sufficiently large, and we recommend a minimum value 100,000.
2. Calculate the mean and variance for each of the  $R_0$  datasets denoted as  $\bar{y}_{0j}$  and  $s_{0j}^2, j = 1, \dots, R_0$ . Calculate the statistics  $t_{0j}^* = \frac{\bar{y}_{0j} - \mu_0}{s_{0j} \sqrt{\frac{1}{n}}}$ . Obtain the critical value  $c_\alpha$  according to the prespecified Type I error rate  $\alpha$ , typically 0.05, and the alternative hypothesis. For example, if the alternative hypothesis is  $H_{a2}$ ,  $c_\alpha$  is the  $100(1 - \alpha)$ th percentile of  $t_{0j}^*$ 's.
3. Generate  $R_1$  sets of non-normal data, each with the sample size ( $n$ ), mean ( $\mu_1$ ), standard deviation ( $\sigma$ ), skewness ( $\gamma_1$ ), and kurtosis ( $\gamma_2$ ). We recommend a minimum value 1000 for  $R_1$ .
4. Calculate the mean and variance for each dataset in Step (3) and denote them as  $\bar{y}_{ai}$  and  $s_{ai}^2, i = 1, \dots, R_1$ , and calculate the corresponding statistic  $t_{ai}^* = \frac{\bar{y}_{ai} - \mu_0}{s_{ai} \sqrt{\frac{1}{n}}}$  statistic.
5. The power is estimated as the proportion that  $t_{ai}^*$  is greater than the critical value  $c_\alpha$ :  $\pi = \#(t_{ai}^* > c_\alpha) / R_1$ .

The Monte Carlo procedure works equally for the normal data, in which the data in Steps (1) and (3) can be generated from normal distributions. The procedure above also works for the paired samples where the population mean, standard deviation, skewness, and kurtosis of the difference scores are used.

## 2 Two-Sample *t*-Test

The two-sample *t*-test is used to test whether two independent population means are equal. The null hypothesis is

$$H_0 : \mu_1 - \mu_2 = 0.$$

The alternative hypothesis can be either two-sided or one-sided:

$$H_{a1} : \mu_1 - \mu_2 \neq 0,$$

$$H_{a2} : \mu_1 - \mu_2 > 0,$$

$$\text{or } H_{a3} : \mu_1 - \mu_2 < 0.$$

The pooled-variance  $t$ -test where the statistic  $t_{pooled} = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2} \cdot \frac{1}{n_1} + \frac{1}{n_2}}}$  follows a  $t$  distribution with degrees of freedom  $n_1 + n_2 - 2$ , where  $n_1$  and  $n_2$  are sample sizes for the two independent samples.  $\bar{y}_1$  and  $\bar{y}_2$  are the sample means and  $s_1^2$  and  $s_2^2$  are the sample variances of the two groups, respectively. The pooled  $t$ -test assumes homogeneity and normality. When the variances of the two groups are not the same, the separated-variance  $t$ -test should be used where the test statistic  $t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$  follows a  $t$  distribution with the degrees of freedom  $\frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}$ . As for the one-sample  $t$ -test, when the normality assumption is violated, the distribution of the statistic is not a  $t$  distribution. Therefore, the Monte Carlo-based method could be used for power analysis.

As in one-sample  $t$ -test, we assume that the shapes of the data distribution for each group under the null and alternative are the same with the same standard deviation, skewness, and kurtosis, which can be estimated from meta-analysis or based on literature review. The step-by-step procedure for the two-sample  $t$ -test power calculation with given sample sizes  $n_1$  and  $n_2$  for the two groups is given below:

1. Let  $\mu_{10}$  and  $\mu_{20}$  be the means of the two groups under the null hypothesis, typically,  $\mu_{10} - \mu_{20} = 0$ . Given the population means ( $\mu_{10}$  and  $\mu_{20}$ ), standard deviations ( $\sigma_1$  and  $\sigma_2$ ), skewness values ( $\gamma_{11}$  and  $\gamma_{12}$ ), and kurtosis values for two groups ( $\gamma_{21}$  and  $\gamma_{22}$ ), generate  $R_0$  sets of non-normal data, one with sample size  $n_1$  and another with sample size  $n_2$ . We recommend a minimum value 100,000 for  $R_1$ .
2. For the  $R_0$  sets of data from previously simulated data pool, calculate the mean and variance of each group for each dataset denoted as  $\bar{y}_{01j}$ ,  $\bar{y}_{02j}$ ,  $s_{01j}^2$ , and  $s_{02j}^2$ ,  $j = 1, \dots, R_0$ . Calculate the separated-variance test statistics  $t_{0j}^* = \frac{\bar{y}_{01j} - \bar{y}_{02j}}{\sqrt{\frac{s_{01j}^2}{n_1} + \frac{s_{02j}^2}{n_2}}}$ .

Obtain the critical value  $c_\alpha$  according to the prespecified Type I error rate  $\alpha$  and the alternative hypothesis.

3. Let  $\mu_{11}$  and  $\mu_{21}$  be the means of the two groups under the alternative hypothesis. Generate  $R_1$  sets of non-normal data, each with the sample sizes ( $n_1$  and  $n_2$ ), means ( $\mu_{11}$  and  $\mu_{21}$ ), standard deviations ( $\sigma_1$  and  $\sigma_2$ ), skewness values ( $\gamma_{11}$  and  $\gamma_{12}$ ), and kurtosis values ( $\gamma_{21}$  and  $\gamma_{22}$ ) for the two groups separately. We recommend a minimum value 1000 for  $R_1$ .

4. Calculate the means and variances for each group in each dataset in Step (3) and denote them as  $\bar{y}_{a1j}$ ,  $\bar{y}_{a2j}$ ,  $s_{a1j}^2$ , and  $s_{a2j}^2$ ,  $i = 1, \dots, R_1$ , and calculate the corresponding  $t_{ai}^* = \frac{\bar{y}_{a1j} - \bar{y}_{a2j}}{\sqrt{\frac{s_{a1j}^2}{n_1} + \frac{s_{a2j}^2}{n_2}}}$  statistic.
5. The power is estimated as the proportion that  $t_{ai}^*$  is greater than the critical value  $c_\alpha$ :  $\pi = \#(t_{ai}^* > c_\alpha) / R_1$ .

### 3 Implementation

The Monte Carlo procedure for power analysis for the one-sample, paired-sample, and two-sample analysis is implemented in an R package WebPower. Specifically, the function `wp.mc.t()` is utilized. The basic usage of the function `wp.mc.t()` has the following form:

```
wp.mc.t(n, R0, R1, mu0, mu1, sd, skewness, kurtosis,
alpha, type, alternative).
```

In the function, `n` is the sample size; `mu0`, `mu1`, `sd`, `skewness`, and `kurtosis` are the mean under the null hypothesis, mean under the alternative hypothesis, standard deviation, skewness, and kurtosis, with the default values 0, 0, 1, 0, and 3, respectively. `R0` and `R1` specify the total number of replications under null and alternative hypotheses with the default value 100,000 and 1000, respectively. `alpha` is the significance level with the default value 0.05. `type` specifies the type of analysis such as one-sample test or two-sample test, and `alternative` specifies the direction of the alternative hypothesis.

We briefly illustrate the application of the `wp.mc.t` function via three examples. First, in a one-sample  $t$ -test, we are interested in whether the population mean is equal to 0 with a two-sided alternative hypothesis. The population distribution follows a normal distribution with mean equal to 0.5 and standard deviation equal to 1. To calculate the power with sample size equal to 20, the R input is as follows:

```
wp.mc.t(n=20, mu0=0, mu1=0.5, sd=1, skewness=0,
kurtosis=3, type = c("one.sample"), alternative =
c("two.sided")).
```

The power is 0.557 in this example.

Second, in a paired  $t$ -test, we plan to test whether the matched pairs have equal means with one-sided alternative hypothesis ( $H_a: \mu_D > 0$ ). The mean, standard deviation, skewness, and kurtosis of the difference scores are 0.3, 1, 1, and 6, respectively. To calculate the power with sample size equal to 40, the specification of the R function is as follows:

```
wp.mc.t(n=40, mu0=0, mu1=0.3, sd=1, skewness=1,
kurtosis=6, type = c("paired"), alternative =
c("larger")).
```

The power is 0.657 in this example.

Third, in a two-sample independent  $t$ -test, we plan to examine whether two independent population means are equal with one-sided alternative hypothesis ( $H_a: \mu_1 - \mu_2 < 0$ ). The means for two groups are 0.2 and 0.5, standard deviations for two groups are 0.2 and 0.5, skewnesses for two groups are 1 and 2, and kurtoses for two groups are 4 and 6, respectively. To calculate the power with sample size equal to 15 per group, the specification of the R function is as follows:

```
wp.mc.t(n=c(15, 15), mu1=c(0.2, 0.5), sd=c(0.2, 0.5),
skewness=c(1, 2), kurtosis=c(4, 6), type = c("two.
sample"), alternative = c("less"))
```

The power is 0.879 in this example.

For those who are not familiar with R, an online application is also created to conduct the same power analysis using a simple interface on this webpage: <http://w.psychstat.org/tnonnormal>.

## 4 A Simulation Study

We conducted a simulation study to examine the performance of the Monte Carlo-based power analysis for the two-sample analysis under the null hypothesis  $H_0: \mu_1 - \mu_2 = 0$ . This is to investigate whether the Type I error can be well controlled. The performance of the Monte Carlo method (MC) is also compared with conventional pooled-variance  $t$ -test (CP).

We varied the following four factors in the simulation: normality of data (either normal or non-normal), ratio of variance of group 1 to that of group 2 with  $\sigma_2^2 = 50$  ( $\frac{\sigma_1^2}{\sigma_2^2} = 0.2, 1, 2, \text{ and } 5$ ), ratio of sample size of group 1 to that of group 2 ( $\frac{n_1}{n_2} = 0.2, 1, \text{ and } 2$ ), and sample size of group 1 ( $n_1 = 10, 50, \text{ and } 100$ ). The non-normal data are generated from a Gamma distribution. Overall, a total of 72 conditions ( $2 \times 4 \times 3 \times 3$ ) are evaluated.

The empirical Type I error rates are listed in Table 1. Clearly, the Monte Carlo-based power analysis controlled the Type I error rates well around the nominal level ( $\alpha = 0.05$ ) regardless of the shape of distribution, the level of heterogeneity ( $\frac{\sigma_1^2}{\sigma_2^2}$ ), the ratio of sample size of group 1 to that of group 2 ( $\frac{n_1}{n_2}$ ), and the sample size of group 1 ( $n_1$ ). The conventional pooled-variance  $t$ -test only controlled the Type I error rates at the nominal level under homogeneity and/or equal-sample-size situations as expected. When two groups have different variance and sample sizes, the conventional pooled-variance  $t$ -test yielded either too small rejection rate (e.g., 0.002) or too large rejection rate (e.g., 0.242). Given that practical data are often non-normal and heterogeneous, the Monte Carlo-based power analysis is therefore recommended.

**Table 1** The empirical Type I error in Monte Carlo-based power analysis (MC) and conventional pooled-variance *t*-test (CP) under the null hypothesis

$\frac{n_1}{n_2}$	$n_1$	$\frac{\sigma_1^2}{\sigma_2^2} = 0.2$		$\frac{\sigma_1^2}{\sigma_2^2} = 1$		$\frac{\sigma_1^2}{\sigma_2^2} = 2$		$\frac{\sigma_1^2}{\sigma_2^2} = 5$	
		MC	CP	MC	CP	MC	CP	MC	CP
Normal data									
0.2	10	0.048	0.003	0.051	0.049	0.049	0.117	0.049	0.227
0.2	50	0.050	0.001	0.047	0.048	0.056	0.120	0.047	0.219
0.2	100	0.052	0.002	0.047	0.048	0.051	0.116	0.050	0.225
1	10	0.053	0.057	0.052	0.051	0.048	0.050	0.048	0.055
1	50	0.049	0.051	0.052	0.050	0.051	0.051	0.047	0.050
1	100	0.053	0.054	0.052	0.053	0.048	0.049	0.048	0.048
2	10	0.050	0.131	0.052	0.050	0.047	0.028	0.052	0.020
2	50	0.046	0.116	0.051	0.051	0.048	0.028	0.048	0.015
2	100	0.049	0.121	0.052	0.054	0.052	0.029	0.050	0.015
Non-normal data									
0.2	10	0.050	0.005	0.047	0.048	0.049	0.109	0.050	0.234
0.2	50	0.051	0.003	0.046	0.046	0.053	0.119	0.051	0.242
0.2	100	0.050	0.002	0.047	0.050	0.049	0.119	0.047	0.224
1	10	0.050	0.065	0.052	0.047	0.053	0.056	0.052	0.103
1	50	0.047	0.055	0.049	0.049	0.049	0.049	0.049	0.067
1	100	0.052	0.052	0.048	0.048	0.044	0.048	0.048	0.062
2	10	0.047	0.131	0.053	0.047	0.048	0.038	0.045	0.072
2	50	0.049	0.122	0.049	0.048	0.051	0.032	0.050	0.034
2	100	0.050	0.120	0.050	0.050	0.046	0.029	0.050	0.027

## 5 Conclusion

To flexibly deal with non-normality and unequal variances in the real data, we proposed a Monte Carlo-based power analysis procedure for one-sample *t*-test, two-sample *t*-test, and paired *t*-test. Simulation results showed that the Monte Carlo-based method achieved well-controlled Type I rate even when the assumptions for the conventional power analysis do not hold. In contrast, when homogeneity assumption does not hold and/or two groups have unequal sample size, the conventional pooled-variance *t*-test could be either too liberal or too conservative. Both an R package WebPower and an online application are provided for researchers to easily carry out the Monte Carlo-based power analysis. The Monte Carlo-based method can be generalized to power analysis for ANOVA, regression, structural equation modeling, and multilevel modeling to handle non-normal data. Missing data can also be considered in the Monte Carlo method.

## References

- J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn. (Lawrence Erlbaum, Hillsdale, NJ, 1988)
- B.L. Welch, The generalization of student's problem when several different population variances are involved. *Biometrika* **34**(1/2), 28–35 (1947)
- B.K. Moser, G.R. Stevens, C.L. Watts, The two-sample t test versus Satterthwaite's approximate F test. *Commun. Stat. Theory Methods* **18**(11), 3963–3975 (1989)
- R.L. Disantostefano, K.E. Muller, A comparison of power approximations for Satterthwaite's test. *Commun. Stat. Simul. Comput.* **24**(3), 583–593 (1995)
- M.J. Blanca, J. Arnau, D. López-Montiel, R. Bono, R. Bendayan, Skewness and kurtosis in real data samples. *Methodology* **9**(2), 78–84 (2013)
- T. Micceri, The unicorn, the normal curve, and other improbable creatures. *Psychol. Bull.* **105**(1), 156 (1989)
- M. Cain, Z. Zhang, K. Yuan, Univariate and multivariate skewness and kurtosis for measuring nonnormality: prevalence, Influence and estimation. *Behav. Res. Methods* (in press)
- L.K. Muthén, B.O. Muthén, How to use a Monte Carlo study to decide on sample size and determine power. *Struct. Equation Model.* **9**(4), 599–620 (2002)
- Z. Zhang, Monte Carlo based statistical power analysis for mediation models: methods and software. *Behav. Res. Methods* **46**(4), 1184–1198 (2014)
- F.L. Schmidt, J.E. Hunter, *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings* (Sage, Newbury Park, CA, 2014)