The power analysis technique in determining sample size for military equipment test and evaluation

Yun Bo¹, Zhiyuan Du¹ and Xuejun Liao²

¹ Graduate School, Space Engineering University, Huairou, Beijing, 101416, China
² Space Engineering University, Huairou, Beijing, 101416, China

Corresponding author’s e-mail: yunmills@163.com

Abstract. Test and evaluation is a statutory behaviour to verify that certain design of military equipment has met established operational requirements. The determination of its sample size directly affects the scientific rigor and impartiality of test identification. This paper summarizes the basic theory of power analysis, discusses the algorithm of performing power analysis, and some suggestions for parameter setting of power analysis in different applications are given.

1. Introduction
In statistics, power refers to the probability of being able to reject the null hypothesis correctly under the significance of α, usually in the form of (1-β), when the alternative hypothesis is true [1]. Where α and β are the probabilities of the errors of type I & II. Simply put, α, β and (1-β) are the probabilities of "wrong judgment", "missed judgment" and "right judgment" respectively.

The so-called power analysis is a statistical technique to study the relationship between α, β, sample size N, and effect size ES [2]. The effect size ES refers to the actual difference between the true value of the tested index and the comparison value.

Before the power analysis, though impossible to get the true value, ES does exist, and can be estimated by historical or simulation data, so it can be regarded as fixed. Therefore, efficacy analysis generally has three applications [3]: one is to specify α and β before the test to get the required N; the other is to specify α and N before the test to get the actual value of β; the third is to determine N before the test and give the relative importance of α and β, which is generally expressed in the form of ratio q=β/α, and then determine the appropriate value of α and β.

Because the power analysis provides a theoretical basis for determining the sample size, it effectively enhances the scientific rigor and impartiality of test & evaluation. In addition, test and evaluation, as an acceptance behaviour of weapon equipment design results, is bound to contain a large number of hypothesis testing contents for indicators. Therefore, the application of research efficacy analysis in test and evaluation is of great theoretical and practical significance. In this paper, the single-tailed t-test [4] on the right side is taken as an example to discuss the algorithm implementation of power analysis, study the relationship between various parameters of power analysis, and then give relevant suggestions in practice.

2. Algorithm implementation
Take the single-tailed t test on the right side as an example, when the null hypothesis is true, the distribution curve is the t-distribution [5] probability density curve on the left side of FIG. 1 (when the
statistic $t$ equals to the comparison value); otherwise, it is the non-central $t$-distribution [6] probability density curve on the right side. If $\alpha$ and $N$ are determined, a critical value $t_c$ can be determined by using the $t$ distribution in figure 1, and then the integral of the probability density function of the non-central $t$ distribution less than $t_c$ can be used to calculate $\beta$.

![Figure 1. Example of hypothesis testing statistical power](image)

The algorithm implementation of these calculations is shown in the three flow charts in FIG. 2-4. Note that the meaning of each symbol in the flowcharts is: $\mu_0$-mean of null hypothesis, $\mu_1$-mean of alternative hypothesis, $\sigma$-standard deviation of sample, $N$-sample size, $q$-the ratio of $\beta$&$\alpha$, tol-error allowed by the program, step-step size of $\alpha$, pdf-probability density function; and the meaning of "change sign" in the flowchart is that the tested value changes from negative $t$ to positive or from positive to negative.

3. Power analysis parameter setting study

3.1. Sample size $N$ and statistical power $(1-\beta)$

Using Cohen $d$ [7] as the measure of $ES$, i.e. $ES = \mu_1 - \mu_0 / \alpha$. When $ES$ and $\alpha$ remain unchanged, the statistical power $(1-\beta)$ is consistent with the increasing trend of sample size $N$ by observing the calculation results in table 1($ES=0.5, \alpha=0.05$). For example, reducing the requirement for $(1-\beta)$ from 0.9855 to 0.8483 can reduce the sample size from 60 to 30.

This is mainly because the smaller the $N$ is, the smaller the non-central parameter [6] is, that is, the closer the distance between the $t$-distribution and the non-central $t$-distribution is, and the smaller the $N$ is, the larger the sample standard deviation is, that is, the wider the probability density curve is, which results in the larger overlap coverage of the two distribution curves. If $\alpha$ remains constant, it inevitably leads to an increase in $\beta$, i.e., a reduction in statistical efficacy $(1-\beta)$.
However, statistical power (1-β) cannot be reduced indefinitely, as shown in table 1. When N is 15 and 10, (1-β) is only 0.5781 and 0.4273. Take military equipment test & evaluation as an example, if a certain performance of a certain equipment does meet the requirements of the index, but under such a poor power, the probability of "passing" the correct evaluation is close to 0.5, which is almost the same as the decision based on coin toss.

According to statistical practice, 0.8 is generally set as (1-β). By observing table 1, it can be seen that a value between 15 and 30 is appropriate for the sample size N at this time (the accurate value is N=27).

| 1-β  | N    | Critical t |
|------|------|------------|
| 0.9996 | 100  | 1.6604     |
| 0.9855 | 60   | 1.6711     |
| 0.8483 | 30   | 1.6991     |
| 0.5781 | 15   | 1.7613     |
| 0.4273 | 10   | 1.8331     |

3.2. Sample size N and significance level α

When ES is fixed (ES=0.35 in the table), it can be seen from the observation in FIG. 5 that if you want to ensure statistical power as high as possible (in this case, (1-β) =0.8), you can loose the significance level α, that is, increase its value, so as to reduce the sample size N. For example, raising the requirement for α from 0.05 to 0.1 would reduce the sample size from 52 to 38.

The reason for this is the same as for 2.2, except that the (1-β) values are fixed and the values of alpha are calculated.

However, α could not also be loosen indefinitely. As shown in figure 5, when N<38, α would be greater than 0.1, which means that if the design under test does not achieve the desired purpose, in 10 tests, there could probably be more than one “passed”. In the field of procurement of military equipment, this risk is often unbearable, because they have a bearing on the success or failure of front-line operations and the life and death of soldiers.
In order to improve the comprehensive benefits of test and evaluation, it is usually expected to reduce the sample size $N$ of the tested equipment as much as possible. However, based on the above analysis,
when reducing the sample size $N$ in test, the preconditions shown in table 2 should be considered and appropriate adjustment measures should be taken.

| The focus of the test | Approach | Application scenarios |
|-----------------------|----------|-----------------------|
| To avoid “wrong judgment” | Fix $\alpha$, increase $\beta$ | In order to strictly control the quality of the projects with mature technology or wide application in the front line of the army, you would rather miss the design which has met the demand. |
| To guarantee the “right judgment” | Fix $\beta$, increase $\alpha$ | For major scientific and technological innovation projects or projects with forward guidance, if the design really meets the needs, and the risk of wrong judgment is tolerable, try to select it. |
| Both | Fix both | For projects with relatively mature technology or extensive front-line military applications, under the premise of strict quality control, there is evidence to prove that the degree of performance better than the index requirements is greater than previously expected, so the sample size requirements can be reduced without sacrificing the probability of correct selection. |

5. Conclusion
Taking t test as an example, this paper expounds the basic principle, calculation method and basic application of statistical power analysis in the test of military equipment design, the basic way of adjusting test sample size by using power analysis and the precondition that should be paid attention to are given, which can better serve the related work of equipment design and acceptance.

When analyzing the power of F test, chi-square test and other hypothesis test methods, the basic principle is consistent with the content of this paper, and the solution method, application premise and applicable scenarios are basically similar, therefore, in practical work, the content elaborated in this paper can be adjusted appropriately according to the type of hypothesis test to meet the actual needs.

References
[1] Gravetter, F.J., Wallnau, L.B. (2014). Essentials of Statistics for the Behavioural Sciences, 8e. Cengage Learning Asia Pte. Ltd, Singapore.
[2] Wikipedia (2018). Power (statistics). https://en.wikipedia.org/wiki/Power_(statistics).
[3] Mayr, S., Erdfelder, E., Buchner, A., Faul, F. (2007) A short tutorial of GPower. Tutorials in Quantitative Methods for Psychology, Vol.3(2), p.51-59.
[4] Wikipedia (2019). Student's t-test. https://en.wikipedia.org/wiki/Student%27s_t-test.
[5] Wikipedia (2019). t-distribution. https://en.wikipedia.org/wiki/Student%27s_t-distribution.
[6] Wikipedia (2019). Noncentral t-distribution. https://en.wikipedia.org/wiki/Noncentral_t-distribution.
[7] Cohen, J. (1988). Statistical Power Analysis for the Behavioural Sciences (2nd ed.). Lawrence Erlbaum Associates, New York.