Decision-making Risks of Operational Tests from a Statistical Point of View

Yun Bo1,2*, Yu Bai1,3 and Xuejun Liao4*
1 Graduate School, Space Engineering University, Huairou, Beijing, 101416, China
2 Baicheng Ordnance Test Center, Baicheng, Jilin, 137001, China
3 Joint Service College, National Defence University, Haidian, Beijing, 100039, China
4 Department of Space Support, Space Engineering University, Changping, Beijing, 102206, China
*Corresponding author’s e-mail: yunmails@163.com

Abstract. The operational test, along with the evaluation based on it, is the most important decision-making basis for the Department of Defence in acquiring weapons and equipment. However, there has been relatively little attention to the credibility of its conclusions. In order to evaluate the decision-making risks quantitatively, the relationship between the two types of errors in the operational tests is studied. By analysing the interaction between the factors affecting decision risk errors, it is pointed out that it is of great practical significance to ensure that the tests have sufficient statistical power, which is usually overlooked before in operational tests.

1. Introduction
The operational test is a process of observing the completion degree of the combat task of the weapon and equipment under the control of professional fighters and the degree of adaptation to the given combat task by simulating the actual combat. There are many factors affecting the reliability of its conclusions, such as capital consumption, schedule, resource allocation, and the credibility of test conclusions. However, among these factors, the most fundamental factor is the reliability of the test results, because other factors are directly or indirectly evolved from it. Therefore, from the perspective of statistics, this paper investigates the decision risks of extrapolating relevant assessment indicators of weapons and equipment based on test data.

2. Basic concepts

2.1. Statistical hypothesis test
Hypothesis testing is the process of making a hypothesis about the parameter(s) of the population, and then deducing whether the hypothesis is valid from the information of the sample. This process is called statistical test of hypothesis[1].

The process of arriving at a conclusion by hypothesis testing is analogous to a proof by contradiction. It calls the conclusion or theoretical result which wants to support the alternative hypothesis, and it's denoted as $H_a$. The conclusion or theory that is opposed to the alternative hypothesis is called the null hypothesis and is denoted as $H_0$. Then, come up with a function that is completely determined by the
sample data, that is the test statistic, and construct the probability distribution of the test statistic if $H_0$ is true. Finally, the sample data obtained from the test is substituted into the test statistics to calculate the probability $p$ of the occurrence of the sample and more extreme cases than the sample under this distribution. If the $p$ value is very small, $H_0$ is considered very suspicious when the sample is used as evidence and $H_a$ should be supported. On the other hand, if the $p$ value is big, then it is considered that we cannot deny $H_0$, when the sample is taken as evidence. Therefore $H_a$ could not be supported.

2.2. Two types of errors, significance levels, statistical power and effect size

In the hypothesis test, when $H_0$ is true, the case of rejecting $H_0$ is called the type I error, the case where $H_a$ is true and $H_0$ is not rejected is called a type II error. They are two types of risks that researchers face when extrapolating population information based on sample data.

The statistical tool for controlling the risk of making the type I error is the significance level. It’s the acceptable level of probability of making the type I error, denoted as $\alpha$. When the $p$ value calculated based on the sample data is less than $\alpha$, it is called "reject $H_0$ at the $\alpha$ level", otherwise, it is called "$H_0$ cannot be rejected at the $\alpha$ level".

The statistical tool for controlling the risk of making the type II error is the statistical power. It is the probability of not making the type II error, denoted as $(1 - \beta)$, where $\beta$ is the probability of making the type II error.

Effect Size ($ES$) measures provide an index of how much impact treatments actually have on the dependent variable.[2] The most common $ES$ measures are proposed by Cohen[3].

Taking figure 1 as an example, suppose the population is normally distributed and the variance is known, and $H_0$ is that the population mean is less than or equal to 10, namely $\mu \leq 10$, $H_a$ is that the population mean is greater than or equal to 10, namely $\mu \geq 10$. While the sample size is $n$, and $H_0$ is true, the sample mean $\mu$ follows the normal distribution of parameters 10 and 1. If $H_a$ is true, the sample mean follows a normal distribution with parameters 12 and 1. After the significance level $\alpha$ is determined, a threshold value of $z_{critical}$ can be determined in the distribution of $H_0$, which is the $(1 - \alpha)$ quantile of the distribution of $H_0$. Then by definition, $H_0$ can only be rejected if the sample mean is greater than $z_{critical}$. Therefore, the area covered by the red region on the right of the $H_0$ distribution graph is the significance level $\alpha$, the area covered by the blue area to the right of $H_a$ is the statistical power $(1 - \beta)$. Then, the relationship between the two types of errors is shown on the right side of figure 1. If $H_0$ or $H_a$ is true, the probability of making type I or type II error is $\alpha$ or $\beta$. Where, $d$ in the figure 1 is the $ES$(Cohen $d = (\mu_a - \mu_0)/\sigma$ [3]), $\sigma$ is the standard deviation of both of the distributions.

![Figure 1. Null hypothesis, alternative hypothesis, significance level, and statistical power.](image)

3. Influence of sample size on statistical power

Since $ES$ is objective, and the significance level $\alpha$ adopts the accepted standard (generally 0.05), the adjustment of sample size $n$ will mainly affect the statistical power $(1 - \beta)$.

Taking the normal distribution as an example, if the sample size is $n$, $H_0$ and $H_a$ obeys $N(10,1)$ and $N(12,1)$ respectively, so first of all, from the equation of

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{(2\pi)}} e^{-\frac{(x-10)^2}{2}} dx = 0.95$$
it can be derived that \( z_{\text{critical}_1} = 11.6449 \).

Then, the statistical power was calculated as follows:

\[
(1 - \beta_1) = 1 - \int_{-\infty}^{11.6449} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma_x^2}} dx
\]

\[= 0.6387 \]

If the sample size is \( 2n \), then \( H_0 \) and \( H_\alpha \) obey \( N(10,1/2) \) and \( N(12,1/2) \) respectively, that is, the standard deviation of the sample mean decreases to \( \sigma_x = 1/\sqrt{2} \). As above, from the equation of

\[
\int_{-\infty}^{12} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-10)^2}{2(1/2)^2}} dx = 0.95
\]

it can be derive that \( z_{\text{critical}_2} = 11.1631 \).

Then, the statistical power was calculated as follows:

\[
(1 - \beta_2) = 1 - \int_{-\infty}^{11.1631} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma_x^2}} dx
\]

\[= 0.8817 \]

If the sample size is \( n/2 \), then \( H_0 \) and \( H_\alpha \) obey \( N(10,2) \) and \( N(12,2) \) respectively, that is, the standard deviation of the sample mean increases to \( \sigma_x = \sqrt{2} \), as above, from the equation of

\[
\int_{-\infty}^{12} \frac{1}{2\sqrt{\pi}} e^{-\frac{(x-10)^2}{2\sqrt{2}}} dx = 0.95
\]

it can be derived that \( z_{\text{critical}_3} = 12.3262 \).

Then, the statistical power was calculated as follows:

\[
(1 - \beta_3) = 1 - \int_{-\infty}^{12.3262} \frac{1}{2\sqrt{\pi}} e^{-\frac{x^2}{2\sigma_x^2}} dx
\]

\[= 0.4089 \]

That is, when the \( ES \) and significance level are fixed, the statistical power increases with the increase of sample size \( n \), as shown in figure 2.

4. Influence of lack of statistical power on test results

Statistical power has a decisive influence on the conclusion of test, but it is often ignored in current researches [4][5][6][7]. Because of research funding constraints, the statistical power is very low in many studies, of which the results reported are type I or type II errors, or the so-called false positives or false negatives. This is so common in the academic community that many of the results published in the top journals cannot be avoided[4][5][6][7]. Literature [5][6][7] discussed the subjective and objective causes which result in this situation, while literature[4][8] focused on how the ability to find effects through tests was severely weakened in the absence of statistical power. Take figure 3 as an example, which reflects the fact that the tests, with different statistical power, will obtain the positive predictive value (ppv), in the case that the results with effect account for different proportions of all the results.
Figure 3. The variation trend of PPV with statistical power.

The 6 colored blocks in figure 3 respectively represent the 6 different situations constituted by the proportion of all results. In each set of colored blocks, if the research results do not have an effect, the correctly identified proportion, \(prop_{pn}\), is represented by green colored blocks, and the incorrectly identified proportion (i.e., type ii error), \(prop_{nn}\), is represented by red colored blocks. If the results of the study do have an effect, the proportion correctly identified, \(prop_{pa}\), is represented by blue colored blocks, and the incorrectly identified proportion (i.e., type ii error), \(prop_{na}\), is presented by grey colored blocks. Let the ratio of the results with effect to all the results be

\[
p_{peff} = \frac{prop_{pa} + prop_{na}}{prop_{pn} + prop_{pa} + prop_{pa} + prop_{na}}
\]

and,

\[
PPV = \frac{prop_{pa}}{prop_{pn} + prop_{pa}}
\]

It can be seen that when \(p_{peff} = 0.5\) (as shown in the above 3 sub-graphs of figure 3), if \((1 - \beta)\) is 0.2, 0.5 and 0.8, PPV is 0.80, 0.91 and 0.94, respectively. When \(p_{peff} = 0.1\) (in the three cases above), and when \((1 - \beta)\) is 0.2, 0.5 and 0.8, \(PPV\) is 0.31, 0.53 and 0.64, respectively. That is, the accuracy of inferred conclusions from experiments decreases with the decrease of statistical power, and this situation will rapidly deteriorate with the decrease of the proportion of correct conclusions in all conclusions, as shown in figure 3.

Therefore, the literature [8] points out that it is because most studies expect to get seismic conclusions, but the seismic conclusions account for a small proportion of all conclusions, plus the statistical power of tests cannot be guaranteed to be sufficient, therefore, many of the results claimed to have effects are actually false positive results, and the proportion is very high, that is, the \(PPV\) is very small.

5. Conclusion

So from what has been discussed above, it is of great practical significance to improve the correctness of operational test and appraisal conclusion. That is because whether the weapons or equipment of the DoD have the claimed operational effectiveness and operational suitability is related to the life safety of the warfighter and the victory of the wars. If, as shown in figure 3, when \(p_{peff} = 0.1\), \((1 - \beta) = 0.2\), the \(PPV\) is only 0.31, that is, only about 3 pieces have the claimed operational effectiveness and operational suitability among the 10 pieces of SUT. The consequences is obviously catastrophic. Therefore, it is necessary to seriously study the statistical power of OT, and its sample size.

References

[1] W.M. Mendenhall, T. Sincich. (2016) Statistics for engineering and the sciences, Sixth edition, CRC Press, Taylor & Francis Group, Boca Raton.
[2] P.D. Ellis. (2010) The essential guide to effect sizes: statistical power, meta-analysis, and the interpretation of research results, Cambridge University Press, Cambridge, New York.
[3] J. Cohen. (1988) Statistical power analysis for the behavioral sciences, 2nd ed, L. Erlbaum Associates, Hillsdale, N.J.
[4] R. Nuzzo. (2014) Scientific method: Statistical errors, Nature., 506: 150–152.
[5] J.P.A. Ioannidis. (2005) Why Most Published Research Findings Are False, PLoS Medicine., 2: 6.

[6] L.R. Jager, J.T. Leek. (2014) An estimate of the science-wise false discovery rate and application to the top medical literature, Biostatistics., 15: 1–12.

[7] J.P.A. Ioannidis. (2014) Discussion: Why “An estimate of the science-wise false discovery rate and application to the top medical literature” is false, Biostatistics., 15: 28–36.

[8] M. Krzywinski, N. Altman. (2013) Power and sample size, Nat Methods., 10: 1139–1140.