Journal of the Korea Academia-Industrial cooperation Society
Vol. 23, No. 11 pp. 352-358, 2022

https://doi.org/10.5762/KAIS.2022.23.11.352
ISSN 1975-4701 / eISSN 2288-4688

시험평가 유형별 신뢰도 향상을 위한 통계적 시험횟수 판단

정승원[1*], 전일국[2]
[1]육군 시험평가단, [2]방위사업청

Statistical Approach for Determining the Number of Tests to Enhance the Reliability of Test and Evaluation Items

Sueng-Won Chung[1*], Il-Kuk Jeon[2]
[1]ROKA Test & Evaluation Group
[2]Defense Acquisition Program Administration

요 약 새로운 무기체계를 도입하는 과정에서 시험평가는 전력 발휘를 보장하는 핵심적인 절차이다. 시험평가를 수행함에 있어서 가장 중요한 요소 중 하나는 통계적으로 신뢰할 수 있는 적정 시험 횟수판단이라고 할 수 있다. 그동안 시험횟수를 판단함에 있어서 통계적 이론에 근거한 시험횟수 판단보다는 시간적, 경제적 상황을 주로 고려하거나, 과거 유사무기체계의 선례를 따르는 등의 방법으로 횟수를 판단하였다. 그 결과 시험평가에 대한 통계적 신뢰성을 확보하지 못한 경우가 일부 발생했다. 이러한 문제점을 개선하기 위해 통계 이론을 활용하여 시험횟수를 판단하는 과학적 시험평가 방법들이 수행되고 있지만, t-검정이라는 특정 이론에 국한된 시험횟수 위주로 판단되고 있어, 개별 시험 항목에 적절한 통계적 방법을 적용하지 못해 신뢰성을 확보하지 못한 경우가 있었다. 본 논문에서는 시험평가 계획수립 시 다양한 시험평가 항목별로 적용 가능한 통계적 방법을 제시하였고 그에 따른 적절한 시험횟수를 제안하였다. 이를 통해 국방 분야의 시험평가에 더욱 과학적 신뢰성을 확보할 수 있고, 더 나아가 국방력의 신뢰성을 보장할 수 있는 다양한 과학적 방법론의 연구에 기여할 수 있을 것으로 기대된다.

Abstract The test and evaluation (T&E) procedure is an essential legal requirement for ensuring fitness for purpose before any new weapon system is introduced. Determining the number of tests that should be performed is an important aspect of this procedure. However, the number of tests required has been determined based on experience of similar weapons and/or time and cost considerations, and as a result, the statistical reliability of the T&E procedure has not been achieved. To solve this problem, scientific T&E methods have been proposed, and several statistical tests have been used. Although there are examples of their proper use, in too many cases, these methods have not been appropriately implemented. In particular, reliability was often not achieved because appropriate statistical methods were not applied to specific test items. This paper presents statistical methods for T&E and proposes required test numbers. We believe that implementing our findings will improve reliability in the T&E defense domain and contribute to research in scientific methodologies that improve the reliability of defense systems.

Keywords : Test and Evaluation, Number of Test, Statistic Methods, T-Test, Proportion Test, ANOVA

*Corresponding Author : Sueng-Won Chung(ROKA Test & Evaluation Group)
email: seanlisa15@gmail.com
Received August 19, 2022 Revised September 23, 2022
Accepted November 4, 2022 Published November 30, 2022

1. 서론

새로운 무기체계를 군에 도입하기 위해서는 시험평가 기준과 항목, 방법, 시기 등이 포함된 시험평가 계획을 수립하고 일정 기간 시험평가를 수행한 후 해당 기준을 충족한 장비가 군에 전력화된다[1]. 시험평가를 수행함에 있어 시험평가 계획수립이 무엇보다도 중요한데, 시험평가 계획수립에 중요한 부분 중 하나는 시험횟수의 판단이다. 우리 군은 시험횟수를 판단함에 있어 다음의 세 가지 요소를 우선 고려해왔다.

ⅰ. 과거 유사무기체계의 시험횟수 적용사례
ⅱ. 시험평가 기간과 전력화 시기에 대한 시간적 고려
ⅲ. 사업예산과 시제 수에 대한 경제적인 고려

과거 유사 무기체계 시험평가 시 시험횟수 적용사례를 이용한 경우 어떠한 통계적 근거 없이 과거의 시험횟수를 그대로 가져와 시험을 진행하게 되고 이는 다음 시험평가에도 똑같이 영향을 주어 결과적으로 과학적 시험평가를 수행할 수 없도록 한다. 예를 들어, 신형 무전기의 운용시간에 관한 성능에 대해 시험평가를 진행하는 경우를 생각해 보자. 과거 무전기의 운용시간에 대한 시험평가를 위해 5회의 시험을 진행하였다는 것을 근거로, 이번에도 비슷하게 5회의 시험을 시행한다면 이는 통계적 근거를 기반으로 한 것이 아니므로 과학적 시험평가가 이루어졌다고 할 수 없다. 결과적으로 이러한 시험횟수의 판단은 적절한 시험평가를 수행하지 못하게 만든다.

사업예산과 시제 수에 대한 시간적, 경제적 측면에만 초점을 맞춰 진행한 시험평가의 경우 통계적 유의성을 확보하지 못할 정도의 시험횟수만을 실시하였다. 과거 우리 군에 전력화된 유도무기의 경우 그 당시 전문가 의견에 따르면 최소 20여 발 이상의 시험평가 후 그 결과에 따라 전력화가 필요했으나, 그 당시 사업예산과 기간 등의 부족으로 4발밖에 시험하지 못하고 전력화가 된 사례가 있다[2].

또한, 최근 과학적 시험평가를 적용한 시험평가 계획수립에 있어서 시험횟수를 산정하는 경우, 대부분의 시험평가에서 ROC와의 비교를 목적으로 특정 집단의 평균과 실수(real number), 혹은 특정 집단의 평균과 다른 집단의 평균을 비교하는 방법인 t-검정을 주로 사용하여 27회의 시험을 진행하였다[2]. 하지만 이는 몇 가지 시험평가 항목에 있어서 적절하지 못한 방법이다. 예를 들어, 어떤 3개 이상의 집단들의 평균을 비교하고 싶은 경우에, 통계적으로는 분산분석(ANOVA: Analysis of variance, 이하 ANOVA) 혹은 Kruskal-Wallis 이론을 적용하여 시험횟수를 산출하는 것이 보다 적절하다[3]. 이렇게 일괄적으로 t-검정을 적용하여 시험횟수를 구하는 것과 같이 적절하지 못한 방법을 이용하여 시험횟수를 산출하는 것은 유의수준, 신뢰수준 및 검정력 등을 원하는 정도로 보장하지 못하게 되고 적절한 시험평가가 이루어지지 못할 가능성이 발생하여 무기체계 및 전력지원 체계가 실전에 배치되었을 때 예상하지 못한 문제점들이 발견될 수 있음을 의미한다.

앞서 강석원 외 2인은 '효과 크기를 이용해 적절한 시험횟수 및 시제 수량 결정방법제안(2021)'[4]에서 효과크기 d를 활용하여 t-검정과 비율 비교에 관한 연구를 진행하였다. 이를 통해 통계적으로 유의한 시제 수량 및 시험횟수를 산출하는 방법을 설명하였다. 하지만 한가지의 효과 크기에 대한 설명만 포함되어 있고 비율검정에 대한 정규 근사 방법만을 사용하여 시험횟수 및 시제 수량을 판단하였다는 한계가 있다.

또 장한얼 외 1인은 '다양한 통계적 검정을 활용한 국방체계 시험횟수 판단에 관한 연구(2022)'[2]에서 t-검정 이외에도 비모수 방법, 다변량 방법 등 다양한 검정 방법과 이에 따른 시험횟수 산출과 다양한 효과 크기를 소개하였다. 그러나 특정한 시험 항목에 대한 검정법에 관한 논의가 이루어지지 않았고 그에 따른 횟수를 제안하지 않았기 때문에, 각각의 항목에 대한 검정법을 선택하고 선택된 검정법을 이용하여 적절한 시험횟수를 산출하는 것에 어려움이 있다.

이러한 문제점들의 해결을 위해 본 논문에서는 다음의 3가지 시험평가 항목에 대해 각각의 항목에 대한 적절한 검정법과 이를 통해 구할 수 있는 시험횟수를 제안할 것이다.

ⅰ. 집단의 평균과 특정 기준을 비교하는 경우
ⅱ. 체계의 성공률을 검증하는 경우
ⅲ. 3개 이상 집단의 평균을 비교하는 경우

적절하게 제안된 시험횟수를 기반으로 시험평가에 더욱 통계적 신뢰성을 확보할 수 있을 것이다.

이 논문에서는 횟수판단에 필요한 검정력 분석(power analysis)의 여러 요소를 설명하고, 이를 토대로 시험평가 항목별 적절한 시험 횟수판단에 대해 알아볼 것이다.

2. 본론

2.1 검정력 분석

검정력 분석이란 가설검정(hypothesis test), 유의수준(significance level), 검정력(power), 효과 크기(effect size) 등을 기반으로 시험횟수를 판단하는 분석기법을 의미한다[5].

2.1.1 가설검정, 유의수준, 검정력

가설검정이란 추론하고자 하는 모집단의 성질과 관련된 가설을 검정하는 과정이다. 여기에서 가설이란 통계적으로 검증하고 싶은 주장 및 진술을 의미한다. 가설검정을 위해서는 귀무가설(null hypothesis, H_0)과 대립가설(alternative hypothesis, H_a) 두 가지 가설을 설정해야 하며, 일반적으로 대립가설은 통계적으로 검증하고 싶은 새로운 주장이고 귀무가설은 이러한 주장이 받아들여지지 않았을 경우 선택하게 되는 가설이다. 예를 들어 '지뢰탐지기의 검출률이 90% 이상'이라는 시험 항목이 있다고 해보자. 이를 아래와 같은 Eq. (1)으로 표현할 수 있다.

$$H_0 : p = 0.9 (\leq 0.9)$$
$$H_a : p > 0.9 \quad (1)$$

위와 같은 형식의 대립가설을 단측 대립가설이라고 한다. 반면, 대립가설의 형태가 다음 Eq. (2)과 같은 경우 이를 양측 대립가설이라고 한다.

$$H_0 : p = 0.9$$
$$H_a : p \neq 0.9 \quad (2)$$

단측 대립가설을 검정하는 과정을 단측 검정, 양측 대립가설을 검정하는 과정을 양측검정이라고 하며, 본 논문에서는 시험평가에 있어서 대부분의 대립가설이 '특정 기준보다 크다.' 혹은 '특정 기준보다 작다.'라는 것과 같은 형식이므로 3개 이상의 집단의 평균을 비교하는 경우를 제외하고 단측 검정에 대해서만 다루도록 하겠다.

가설검정을 하는 과정에서 두 가지 위험이 존재하게 된다. 하나는 귀무가설이 참인 경우에도 귀무가설을 기각하는 경우이고 또 다른 하나는 대립가설이 참인 경우에도 대립가설을 기각하는 경우이다. 전자를 통계학에서는 제1종 오류(Type 1 error), 후자를 제2종 오류(Type 2 error)라고 한다. 이를 표로 정리하면 다음 Table 1과 같다.

Table 1. Type of error

Decision \ Reality	H_0 is true	H_a is true
reject H_0	Type 1 error	Correct decision
reject H_a	Correct decision	Type 2 error

제1종 오류를 범할 확률의 최대 허용한계를 유의수준이라고 하고 일반적으로 α로 나타낸다. 제2종 오류를 범할 확률은 일반적으로 β로 나타내며 1-β를 검정력이라고 한다. 즉 검정력이란 대립가설이 사실일 때 귀무가설을 기각할 확률을 의미한다. 다시 말해 대립가설이 참일 때 적절한 판단을 내릴 확률이다. 일반적으로 유의수준은 낮을수록, 검정력은 클수록 적절한 검정이지만 두 가지를 모두 동시에 달성하는 것에는 통계적으로 어려움이 있다. 이러한 한계에 의해 사회의 다양한 분야에서 유의수준은 0.05, 검정력은 0.8을 기준으로 하고 있지만, 시험평가의 횟수판단에 있어서 다양한 상황이 존재하므로 이 논문에서는 유의수준 0.01, 0.05, 0.1과 검정력 0.7, 0.8, 0.9에 대한 의사결정 자료를 제공할 것이다.

2.1.2 효과 크기

효과 크기란 비교하려는 집단의 차이 혹은 관계를 수치로 나타낸 것이다. 검정력 분석에서 효과 크기는 필수적인 요소이므로 다음의 3가지 방법을 이용해 효과 크기를 반드시 추정해야 한다[6].

ⅰ. 기존의 비슷한 연구에서 추정
ⅱ. 기존의 연구가 없다면 pilot study를 통해 추정
ⅲ. pilot study가 어려운 경우 효과 크기가 클 것인지, 중간일 것인지, 작을 것인지 추정

ⅲ의 방법의 경우 Cohen이 제시한 방법으로 현재에도 다양하게 활용되고 있다[5].

다양한 검정 방법에 따라 사용되는 효과 크기에도 차이가 있으며 Cohen은 자신의 연구에서 이를 다음 Table 2와 같이 제시하였다[2,4-6].

Table 2. Effective size

test	effective size index	effective size		
		large	medium	small
t- test	$d = \frac{\lvert\mu_1 - \mu_0\rvert}{\sigma}$	0.8	0.5	0.2
correlation test	$r = \frac{\sigma_{01}}{\sigma_0\sigma_1}$	0.5	0.3	0.1
one-way ANOVA	$f = \frac{\sigma_1}{\sigma_0}$	0.4	0.25	0.1
regression	$F^2 = \frac{R^2}{1-R^2}$	0.35	0.15	0.02
Wilcoxon signed-rank test	$d = \frac{\lvert\mu_1 - \mu_0\rvert}{\sigma}$	0.5	0.3	0.1

이 논문에서는 각 시험 항목별 시험횟수의 판단에 있어 효과 크기가 어느 정도인지 추정이 어려움으로, 모두 중간의 효과 크기를 가정하고 횟수를 판단하도록 하겠다.

2.2 시험평가 항목별 시험횟수 제안

앞서 설명한 검정력 분석의 여러 요소를 바탕으로 무기체계 및 전력지원 체계의 유형별 시험평가 계획을 수립함에 있어 3가지 형식의 시험 항목에 대해 시험횟수를 G*power, Minitab, Develve 등의 프로그램을 활용하여 계산하고 결과를 제시할 것이다.

2.2.1 집단의 평균과 특정 기준을 비교하는 경우

소총 시험평가 항목 중 최대 사거리가 500m 이상인 성능을 시험하기 위해 적절한 시험횟수를 판단하는 상황을 고려해 보자. 이 항목은 집단의 전체적인 값(평균)과 특정 기준을 비교하는 경우에 사용 가능한 방법으로, 귀무가설과 대립가설을 식으로 표현하면 다음 Eq. (3)과 같다.

$$H_0 : \mu = \mu_0 \qquad (3)$$
$$H_a : \mu > \mu_0 \text{ or } \mu < \mu_0$$

where, μdenote mean, $\mu_0$denote constant

즉 집단의 평균이 μ_0 에 비해 '크다' 혹은 '작다'와 같은 대한 가설을 검정하는 것이다.

위의 가설을 검정하기 위해서 사용하는 방법은 크게 2가지가 존재한다. 첫 번째는 t-검정이고 두 번째는 Wilcoxon 부호순위 검정이다. 두 검정법의 가장 큰 차이는 데이터의 정규성 여부이다. 평균을 특정 기준과 비교하는 경우에, 시험횟수 판단에 있어서 가장 먼저 고려해야 할 사항으로는 해당 데이터의 정규성 판단이다. 이 단계에서 정규성을 검정하지 않고 t-검정으로만 시험횟수를 판단한다면, 통계적으로 담보하고 싶었던 유의수준과 검정력을 유지할 수 없으며 이는 잘못된 판단으로 이어질 가능성이 있으므로 주의해야 한다. 정규성에 대한 검정결과에 따라 정규분포를 따른다면 t-검정을, 그렇지 않다면 Wilcoxon 부호순위 검정을 사용해서 횟수를 판단하게 된다. 각각의 경우에 대한 검정 통계량은 다음의 Eq. (4), Eq. (5)의 형태로 표현된다[7,8].

$$T = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \qquad (4)$$

where, $\overline{X}$ denote sample mean, s denote sample standard deviation, n denote sample size

$$W^+ = \sum_{i=1}^{n} rank_i I(X_i > \mu_0) \qquad (5)$$

where, $rank_i$ denote rank of $|X_i - \mu_0|$, $I(X_i > \mu_0)$ denote indication function

위의 Eq. (4), Eq. (5)과 효과 크기 및 분포에 대한 몇 가지 가정을 하고 표본 수에 대한 식을 구하면 t-검정의 경우 Eq. (6), Wilcoxon 부호순위 검정의 경우 Eq. (7)의 식을 얻을 수 있다[7,8].

$$n = \left[\frac{(t_{\alpha,n-1} + t_{\beta,n-1})\hat{\sigma}}{\mu - \mu_0} \right]^2 \qquad (6)$$

where, $t_{\alpha,n-1}$ denote α quantile of t-distribution under n-1 degree of freedom, $t_{\beta,n-1}$ denote β quantile of t-distribution under n-1 degree of freedom, $\hat{\sigma}$ denote estimator of standard deviation.

$$n = \frac{(z_\alpha + z_\beta)^2}{3(p' - 1/2)} \qquad (7)$$

where, z_α denote α quantile of normal distribution, z_β denote β quantile of normal distribution, p' denote P(X+X' > median) with X and X' denoting two independent observations.

주어진 식에 따라 t-검정과 Wilcoxon 부호순위 검정의 유의수준 0.1, 0.05, 0.1과 검정력 0.7, 0.8, 0.9에 대한 표본의 수를 계산하면 다음의 Table 3과 Table 4와 같다.

Table 3. Sample size of t test

α \ power	0.7	0.8	0.9
0.1	14	19	28
0.05	21	27	36
0.01	36	43	55

Table 4. Sample size of Wilcoxon signed-rank test

α \ power	0.7	0.8	0.9
0.1	16	20	28
0.05	22	28	37
0.01	38	46	57

예를 들어 시험평가관이 유의수준 0.05, 검정력 0.8로 시험평가를 진행하고자 하면, 최소 27회 혹은 28회의 시험을 통해 평가를 진행하여야 하고, 유의수준 0.01, 검정력 0.9로 시험평가를 수행하고자 하면, 55회 혹은 57회의 시험을 진행해야 한다.

위 Table 3과 Table 4를 보면 Wilcoxon 부호순위 검정의 시험횟수가 t-검정에 대한 시험횟수에 비해 약간 더 큰 값을 갖는 것을 확인할 수 있다. 하지만 그 값의 차이가 크지 않으므로, 일반적인 상황에서 경제적인 차이 또한 크지 않을 것이다. 따라서 정규성이 보장되는 상황이 아니라면, 즉 과거의 비슷한 체계에 대한 경험이나 정규분포를 따를 것이 명확한 경우가 아니면 기본적으로 Wilcoxon 부호순위 검정의 횟수를 선택하여 시험을 진행하는 것이 바람직하다.

2.2.2 어떤 항목의 성공률에 대한 검정을 하는 경우

폭발물 제거 지뢰탐지 로봇의 성능을 시험평가 하는 상황을 고려해 보자. 이런 시험평가에서는 '지뢰탐지 성공률이 90%를 초과해야 한다.' 같은 시험 항목이 있으며, 이처럼 특정 항목의 성공률에 대한 가설검정을 식으로 표현하면 다음 Eq. (8)과 같다.

$$H_0 : p = p_0 \qquad (8)$$
$$H_a : p > p_0 \text{ or } p < p_0$$

where, p denote proportion, p_0 denote constant

이러한 가설을 검정하는 것에 사용되는 방법은 크게 2가지가 있다. 첫 번째는 이항분포의 정규분포로의 근사성질을 이용한 정규분포를 활용하는 방법이고, 두 번째는 이항분포를 이용한 정확 방법(Exact method)이다. 시험횟수가 많다면 계산적인 복잡성이 덜 한 정규분포로의 근사적 방법을 이용하는 것이 좋고, 그렇지 않다면 이항분포를 이용한 정확 방법을 활용하는 것이 좋다. 일반적으로 시험평가의 횟수는 많음을 보장할 수 없으므로 정확 방법만 설명하도록 하겠다. 정확 방법을 기반으로 한 검정 통계량은 이항분포를 따르는 확률변수 그 자체가 된다. 즉, 다음 Eq. (10)의 확률변수 B가 검정 통계량의 역할을 하게 된다.

$$B \sim B(n, p_0) \qquad (9)$$

where, $B(n,p_0)$ denote binomial distribution with number of total event n and proportion p_0

정확 방법에 기반하여 횟수를 판단하는 방식은 다음과 같은 절차에 따라서 구하게 된다. 여기에서 효과 크기는 중간을 가정하고 있다[9].

i. $P(B \geq b) \leq \alpha$를 만족하는 b를 구하고 이 b를 $b_{\alpha,n}$이라고 하자.

ii. $B' \sim B(n,p_1)$인 확률변수에 대해 $P(B' \geq b_{\alpha,n}) \geq 1-\beta$를 만족하는 최소크기 n을 반복적인 계산을 통해 구한다.

표본의 크기와 검정력을 계산해 주는 프로그램인 G*power를 시험횟수를 산출하면 다음과 같다.

Table 5. Sample size of binomial test

α \ power	0.7	0.8	0.9
0.1	15	18	25
0.05	19	25	33
0.01	30	37	49

위 결과를 보면 같은 검정력과 유의수준에서 t-검정 혹은 Wilcoxon 부호순위 검정보다 대부분 시험횟수가 적은 것을 확인할 수 있다. 큰 차이는 아니지만, 경제적인 부분에 대한 고려가 필수적인 시험평가에서 이러한 변화는 상당히 바람직하다고 볼 수 있다.

2.2.3 3개 이상의 집단의 평균을 비교하는 경우

새롭게 개발된 00제독제-Ⅱ의 성능에 대한 시험평가 상황을 고려해 보자. 이러한 시험평가에 대한 시험 항목

으로 아무런 제독제를 사용하지 않은 집단과, 구형(Ⅰ형) 제독제를 사용한 집단, 신형 제독제를 사용한 집단을 비교하여 평가하는 시험이 항목이 있다. 이 항목은 위와 같이 3개 이상의 집단에 평균을 비교하는 경우에 대한 가설검정을 식으로 적으면 다음 Eq. (10)과 같다.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k \qquad (10)$$
$$H_a : At\ least\ one\ \mu\ differs\ from\ another$$

where, μ_i is mean of i-th group

즉 k개의 집단의 평균이 모두 같은지, 아니면 적어도 다른 한 개의 값은 존재하는지를 검정하는 가설검정이다. 이러한 가설을 검정하는 방법은 크게 2가지가 있다. 첫 번째는 ANOVA이고 두 번째는 Kruskal-Wallis 검정 방법이다. 첫 번째 방법인 ANOVA의 경우 정규성, 등분산성, 독립성을 만족해야 한다. 이 3가지 가정을 만족하지 못하는 상황이라면, 두 번째 방법인 Kruskal-Wallis 검정 방법을 사용해서 가설검정을 해야 한다. 이후 귀무가설이 기각되게 되면 다중비교 방법을 통해 분석을 진행해야 한다[10,11].

ANOVA의 검정 통계량 F_0은 다음의 Table 6의 ANOVA table에 따라 쉽게 구할 수 있다. 이때 집단은 총 k개를 가정하고 있으며, i번째 집단의 표본 수 n_i에 대해 $n = \sum_{i=1}^{k} n_i$을 의미한다.

Table 6. ANOVA table

source	sum of squares	degree of freedom	mean square	F_0
factor	SS_f	k-1	$MS_f = \frac{ss_f}{k-1}$	$F_0 = \frac{MS_f}{MSE}$
residual	SSE	n-k	$MSE = \frac{SSE}{k(n-1}$	
total	SST	n-1		

Kruskal-Wallis의 검정 통계량은 다음 Eq. (11)와 같다[12].

$$S_{KW} = \frac{12}{n(n+1)} \sum_{i=1}^{K} n_i \left(\frac{R_i}{n_i} - \frac{n+1}{2}\right)^2 \qquad (11)$$

where, n_i denote sample size of i-th group, n denote sum of n_i for all i, R_i denote sum of ranks of the i-th sample

위의 검정 통계량과 효과 크기에 대한 가정을 이용하여 표본의 수를 G*power와 Develve를 이용해서 구하면 다음의 Table 7, Table 8과 같다.

Table 7. Sample size of ANOVA

α \ power	0.7	0.8	0.9
0.1	99	126	171
0.05	129	159	207
0.01	192	228	285

Table 8. Sample size of Kruskall-Wallis

α \ power	0.7	0.8	0.9
0.1	149	189	257
0.05	194	239	311
0.01	288	342	428

위의 시험횟수는 모든 집단의 표본 수를 더한 것을 의미한다. 그러나 한 집단에 표본 대부분이 몰려있다면 정확한 검정이 이루어지기 어려움으로, 시험횟수를 판단할 때 집단 간 적절한 표본의 분배가 필요하다. ANOVA와 Kruskal-Wallis 방법을 비교해 보면 비 모수적 방법인 Kruskal-Wallis의 시험횟수가 더 많은 것을 확인할 수 있다. 하지만 일반적으로, ANOVA의 가정들을 모두 만족하는 경우가 아니면 검정의 결과는 Kruskal-Wallis가 더 적절하고 알려져 있다[7,12].

3. 결론

지금까지의 시험평가 횟수판단과는 달리, 몇 가지 상황에 대한 통계적 검정 방법과 시험횟수 판단 방법에 대해 알아보았다. 단순히 27회로 모든 시험 항목에 대한 시험평가를 수행하는 것보다 앞서 통계적으로 구한 시험횟수를 만족하는 정도로 시험평가를 수행하면 무기체계 및 전력지원 체계의 성능을 보장할 수 있도록 유의한 시험평가가 가능할 것이고, 이는 국방전력 발전에 있어서 긍정적인 영향을 줄 것이다. 다양한 검정력과 유의수준에 대해 의사결정 자료를 제공한 이유는, 시험평가에 있어서 시험횟수를 최우선으로 고려하지 못할 상황을 염두에 두었기 때문이다. 이를 통해 시간적, 금전적 경제성과 통계적인 신뢰성 사이에서 적절한 타협점을 찾아 시험횟수를 판단할 수 있을 것이다.

이 논문에서 유의해야 할 점은 효과 크기를 모두 중간에 맞춘 상태로 횟수를 판단했다는 것이다. 효과 크기가 변하면 시험횟수도 변하게 된다. 이는 시험횟수가 감소할 수 있음을 의미한다. 예를 들어, t-검정의 경우 유의수준 0.05와 검정력 0.8에서, 효과 크기가 중간(0.5)이면 시험횟수가 27회이지만 효과 크기가 큰(0.8) 경우에는 단 12회만의 시험으로도 통계적으로 비슷한 유의성을 확보할 수 있다. 그러므로 앞으로의 시험평가 결과를 통계적으로 적절하게 기록하는 것이 필요하고, 이에 관한 연구가 필수적이다.

마지막으로 이 논문은 더 많은 시험 항목들에 대한 고려가 부족하고, 그 이전에 다양한 시험 항목들에 대해 어떤 방식으로 시험 항목을 설정하여 시험평가를 진행하는 것이 통계적으로 더 적절한지 등에 관한 연구가 부족하다. 그러므로 차후에 더 많은 시험 항목들에 관한 연구가 이루어져야 하고, 시험 항목의 통계적 선정 방법에 관한 연구가 필요할 것이다.

References

[1] Defense Acquisition Program Act.

[2] H.E Jang, I.K Jeon, “A Study on the Sample Size Determination In the Defense System Using Various Statistical Tests”, *Society of Korea Industrial and Systems Engineering spring conference,* Society of Korea Industrial and Systems Engineering, Jeju, Korea, pp. 648-653, Jun. 2022.

[3] Y.B Lee, I.K. Jeon, S.W Kang, H.E Jang, “Guide Book for scientific test and evaluation”, ROKA Test and evaluation group, pp. 29-64, 2022.

[4] S.W Kang, I.K Jeon, Y.B Lee, "Proposal for method to decide an appropriate number of prototype and tests using effect size", *Society of Korea Industrial and Systems Engineering spring conference,* Society of Korea Industrial and Systems Engineering, Seoul, Korea, pp. 203-208, May 2021.

[5] Cohen. J, “Statistical power analysis for the behavioural sciences”, New York: Academic Press, pp. 12-209, 1988. DOI: https://doi.org/10.4324/9780203771587

[6] F.S Nahm, "Understanding Effect Sizes", *Hanyang Medical Reviews*, 35(1), pp. 40-43, 2015. DOI: https://dx.doi.org/10.7599/hmr.2015.35.1.40

[7] Gottfried E. Noether, "Sample Size Determination for Some Common Nonparametric Test.", *Journal of the American Statistical Association,* vol. 82(398), pp. 645-647, 1987. DOI: https://doi.org/10.1080/01621459.1987.10478478

[8] Thomas P. Ryan, “Sample Size Determination and Power”, John Wiley & Sons, pp. 66-326, 2013, DOI: https://doi.org/10.1002/9781118439241

[9] I.K.Yeo, "Sample Size Determination for One-Sample Location Test", *The Korean Journal of Applied Statistics*, vol. 28, no. 3, pp. 573-581, Jun. 2015. DOI: https://dx.doi.org/10.5351/KJAS.2015.28.3.573

[10] Tukey J.W, "Comparing Individual Means in the Analysis of Variance", *Biometrics*, vol. 5, no. 2, pp. 99-114, 1949. DOI: https://doi.org/10.2307/3001913

[11] Dunnett C.W, "A multiple comparisons procedure for comparing several treatments with a control", *Journal for the American Statistical Association*, 50, pp. 1096-1121, Dec. 1955. DOI: https://doi.org/10.2307/2281208

[12] Fan C, Zhang D, Zhang C. H, "On sample size of the Kruskal-Wallis test with application to a mouse peritoneal cavity study", *Biometrics*, vol. 67, no. 1, pp. 213-224, 2011. DOI: https://doi.org/10.1111/j.1541-0420.2010.01407.x

정 승 원(Sueng-Won Chung) [준회원]

- 2021년 9월 : 고려대학교 통계학과 (석사 재학)
- 2022년 5월 ~ 현재 : 육군 시험평가단 S/W 연구병

〈관심분야〉
시험평가, 머신러닝, 딥러닝, 강화학습, FDA

전 일 국(Il-Kuk Jeon) [정회원]

- 2007년 1월 : 국방대학교 국방관리학과(석사)
- 2016년 2월 : 충남대학교 군사학과 (박사 수료)
- 2020년 12월 ~ 2022년 10월 : 육군 시험평가단 적합성평가장교
- 2022년 11월 ~ 현재 : 방위사업청 230mm 다련장 사업관리장교

〈관심분야〉
무기체계, 신속획득, 시험평가, 방위산업, 사업관리