

AHP를 이용한 대안 평가의 유의성 분석: 비모수적 통계 검정 적용

The Significance Test on the AHP-based Alternative Evaluation: An Application of Non-Parametric Statistical Method

박준수(Joonsoo Park)*, 김성철(Sung-Chul Kim)**

초 록

공공사업 타당성 평가나 대안 선정에서는 대부분 AHP 기법을 활용하여 대안별 가중합 방식으로 최종 점수를 산정하고 그 값이 가장 큰 대안을 선택하고 있다. 특히 타당성 분석과 같은 경우에는 최종 점수가 0.5보다 큰 대안을 선택하게 되는데, 그 값이 0.5보다 얼마나 커야 의사결정이 의미있고 설득력있는 판단이라고 할 수 있는지에 대한 합리적인 기준 없이 적용되고 있다. 한국개발연구원(KDI)에서 제시한 방법론에는 사업진행 대안의 종합 점수가 0.5 근처에 있는 경우를 회색 영역으로 구분하여 신중한 결정을 하도록 제시하고 있는데, 세부 기준에 관한 이론적 검토는 빈약하다. 반면, 통계적 검정의 개념을 도입하여 시도된 분석 사례에서는 가중합 평가 점수의 확률분포로서 정규분포 또는 베타분포를 가정하였으나, 이에 대한 분포적 타당성은 제시되지 않았다. 본 논문에서는 현재 다양한 분야에서 적용되고 있는 가중합 평가 방식의 사례를 검토하여 그 결과의 통계적 검정의 필요성을 제기하고, 통계적 검정을 위하여 특정 분포를 가정하지 않는 비모수적 검정 방법을 제시한다. 그리고 본 연구에서 제시한 방법을 국방 분야의 사례에 적용하고 그 함의와 함께 향후 연구의 발전방향을 제안한다.

ABSTRACT

The method of weighted sum of evaluation using AHP is widely used in feasibility analysis and alternative selection. Final scores are given in forms of weighted sums and the alternative with largest score is selected. With two alternatives, as in feasibility analysis, the final score greater than 0.5 gives the selection but there remains a question that how large is large enough. KDI suggested a concept of 'grey area' where scores are between 0.45 and 0.55 in which decisions are to be made with caution, but it lacks theoretical background. Statistical testing was introduced to answer the question in some studies. It was assumed some kinds of probability distribution, but did not give the validity on them. We examine the various cases of weighted sum of evaluation score and show why the statistical testing has to be introduced. We suggest a non-parametric testing procedure which does not assume a specific distribution. A case study is conducted to analyze

* Department of Statistics & Actuarial Science, Soongsil University; Center for Defense Acquisition, Korea Institute for Defense Analyses(jspark@kida.re.kr)

** Corresponding Author, Department of Statistics & Actuarial Science, Soongsil University(sckim@ssu.ac.kr)

Received: 2016-12-29, Review completed: 2017-01-10, Accepted: 2017-02-03

the validity of our suggested testing procedure. We conclude our study with remarks on the implication of analysis and the future way of research development.

키워드 : 대안 선택, 상대 평가, 계층적 의사결정 분석, 통계적 유의성, 비모수 통계, 윌콕슨 부호순위 검정
Alternative Selection, Relative Evaluation, AHP, Statistical Significance, Non-Parametric Statistics, Wilcoxon Signed-Rank Test

1. 서 론

정부가 수행하는 공공개발 또는 공공투자 사업은 관리 주체인 정부의 업무 영역을 넘어서 궁극적으로 정부 재정의 근간을 형성하는 일반 납세자들에게 그 혜택이 귀속된다. 그러므로 공공사업의 평가는 경제적 손익뿐만 아니라 사회적 관점에서의 편익을 고려해야 한다. 이러한 사업을 추진할 때는 초기 단계에서부터 여러 가지 정책적 요건을 고려하여 사업의 타당성을 평가하고 효율적인 예산집행이 이루어지도록 하는 것이 중요하다. 아울러서 대안 및 사업자를 선정함에 있어서도 다양한 정량적, 정성적 요소들을 고려하여 공정하고 합리적으로 수행되도록 하여야 한다.

일반적으로 사업의 타당성을 검토할 때 가중합 평가 방식이 통용되고 있는데, 여기에는 여러 가지 유념해야 할 점들이 있다. 평가 항목들 사이에 절대적 기준과 상대적 기준 그리고 정량적 평가와 정성적 평가 내용이 합리적으로 종합되도록 하여야 하며, 평가 결과의 신뢰성과 대표성도 보장되어야 한다. 무엇보다도 궁극적인 관심사 중에 하나는 대안들의 평가 점수 사이에 유의미한 차이를 판별하는 문제이다. 다시 말해서 평가 결과의 변별력을 판단하는 문제가 중요하다. 일반적인 기준에 따르면 사업

진행 자체의 타당성 여부와 같이 양자 대안간 상대적인 지지도를 판별하는 경우에는 평가 항목별 점수 구간을 0~1 사이로 하여 최종 종합된 평가 결과가 0.5 이상인 대안을 선택하게 된다. 그리고 여러 대안들의 지지율을 비교하는 경우에는 최종 점수가 가장 높은 대안을 선택한다. 그런데, 그런 선택이 의미있는 변별력을 보장하는가 하는 점에 관해서는 좀 더 세밀한 분석이 요구된다. 예를 들어서, 양자 대안간 선택 상황에서의 판별 기준이 0.5보다 얼마나 커야 하는지 또는 여러 대안들의 상대적 지지율을 비교할 때 최고 점수의 평가 대안이 다른 대안들의 점수보다 얼마나 차이가 나야 하는지에 대해서는 별도의 판단 근거와 기준이 필요하다.

본 연구는 다수의 평가자가 참여하여 집단 의사결정을 수행하는 상황에서 개별 대안마다 매겨진 최종 가중합 점수들의 표본 분포에 대한 분석을 시도한다. 그리고 대안별 평가 점수들의 분포가 변별력있는 차이를 보이는지에 관해서 통계적인 검정 방법을 적용해보고자 한다. 본 연구에서 상정하는 집단 의사결정의 변별력 문제는 실제로 빈번하게 나타나며 동시에 매우 중요한 문제이다. 전문가 집단을 활용한 의사결정은 거의 대부분 정책 결정에 반영되고 있으며 최종 평가 결과의 변별력이 얼

마나 될 것인가 하는 점은 의사결정 전반의 신뢰성을 좌우하는 사안이다. 특히나 대안간 결합이 치열한 경우에는 더욱 그렇다. 본 연구는 이러한 상황과 문제 인식을 감안하되, 가능한 단순한 문제 형태로부터 출발하여 실증적인 연구를 수행하고자 한다. 그런 관점에서 본 연구는 양자 대안간의 택일 상황에서 상대적 지지도를 비교 평가하는 상황을 가정한다.

본 연구 내용의 구성은 다음과 같다. 먼저 제2장에서는 전문가 집단의 평가 결과에 관한 통계적 유의성 문제를 기존 연구에서 어떻게 접근하고 다루어왔는지를 살펴본다. 그리고 기존의 접근법에서 나타난 문제점을 짚어본다. 다음으로 제3장에서는 본 연구에서 시도하는 통계적 접근 방법을 소개한다. 본 연구는 비모수적 방법론을 적용하고 있으며 사례의 유형에 따라서 검정 방법을 다르게 접근하는 기준도 고려한다. 그리고 제4장에서는 현재 국방 분야에서 구축되어 있는 전문가 평가 사례에 적용하고 본 연구가 시도하는 방법의 유용성을 실증적으로 분석한다. 끝으로 제5장에서 본 연구 결과의 함의와 추후 연구 발전 과제를 제시한다.

2. 기존 연구 사례와 문제점

정부는 대규모 공공투자 사업의 타당성을 검토하기 위해서 여러 가지 단계와 과정을 거쳐서 객관적인 분석 평가를 실시하고 있다. 그 중에서 대표적인 것이 기획재정부가 주관하여 차년도 예산반영 직전에 실시하는 예비타당성조사이다. 예비타당성조사는 국토, 문화관광, 자원/에너지 개발 등과 같은 사회간접자본 구

축을 비롯하여 정부 재정이 투입되는 모든 공공투자 사업의 분석평가에 관한 표준적인 준거가 되고 있다. 현재 우리나라의 공공부문 투자사업의 분석 평가에 관한 표준 지침으로서 시행중인 예비타당성조사의 수행 절차와 분석 방법론에 관해서는 한국개발연구원(KDI)의 ‘예비타당성조사 수행을 위한 일반지침 연구’가 제시되어 있다[10].

이 지침에서는 세부 분석에 대한 종합적 검토 방법론으로 계층적 의사결정 분석(AHP: Analytic Hierarchy Process)을 적용하도록 규정하고 있다. AHP는 여러 가지 정량적, 정성적 속성을 갖고 있는 의사결정 문제를 계층구조로 표현하고 속성간 또는 대안간 쌍대비교를 통해 계량화된 방법으로 최적 대안을 결정하는 의사결정 방법이다[19]. 최종 의사결정은 가중합 형태로 계산된 각 대안의 최종 점수에 의해 결정되는데, 이 때 최종 점수는 모든 대안의 점수의 합이 1이 되도록 상대평가 방식으로 산출된다. 경우에 따라서는 대안별 최대 점수가 각각 1이 되도록 절대평가 방식이 적용되기도 한다.

본 연구에서 다루고자 하는 전문가 평가 결과의 통계적 유의성 문제와 관련하여, KDI에서 수행한 최근 연구에서도 ‘예비타당성조사 일반지침’의 후속 보완 방안이 제시되었다. 해당 연구에서 세부적으로 제시된 절차와 방법론에 따르면 사업진행 대안의 종합 점수가 0.5에서 크게 벗어나지 않는 경우를 ‘회색 영역’으로 구분하여 신중한 결정을 하도록 권고하고 있다[1]. 그 내용은 다음 <Table 1>과 같이 회색 영역을 0.45~0.55 구간으로 설정하고 있다.

이렇게 설정된 회색 영역의 구분과 판단 기

〈Table 1〉 Criteria for the Judgement of ‘Grey Area’ Suggested by KDI’s Research on the General Guidelines of the Preliminary Feasibility Study

Total weighted sum of evaluation scores Proportion of supportive evaluation	AHP < 0.45	$0.45 \leq \text{AHP} < 0.5$	$0.5 \leq \text{AHP} < 0.55$	$0.55 \leq \text{AHP}$
4 : 0			feasibility validated	feasibility validated
3 : 1	“feedback” required	strongly requested to be cautious in making a decision	slightly requested to be cautious in making a decision	feasibility validated
2 : 2	if AHP > 0.42, slightly requested to be cautious in making a decision; if AHP < 0.42, feasibility invalidated	moderately requested to be cautious in making a decision	moderately requested to be cautious in making a decision	if AHP > 0.58, feasibility validated; if AHP < 0.58, slightly requested to be cautious in making a decision
1 : 3	feasibility invalidated	slightly requested to be cautious in making a decision	strongly requested to be cautious in making a decision	“feedback” required
0 : 4	feasibility invalidated	feasibility invalidated		

Note) The above criteria is supposed to be applied to the situation having four evaluators.

준이 주는 의미를 개괄해서 살펴보면 다음과 같다. 먼저 $0.45 \leq \text{AHP} < 0.5$ 인 경우에는 종합 평점이 0.5보다 작기는 하지만 평가자 중에서 0.5점 이상을 부여한 비율이 다수(예를 들어, 3:1)이면 ‘아주 신중’, 소수(예를 들어, 1:3)이면 ‘약간 신중’하게 판단하도록 권고한다. 동일한 논리로서 $0.5 \leq \text{AHP} < 0.55$ 인 경우에는 반대의 기준이 적용된다. 한편, 평가자 그룹의 선호도가 동률일 때는 종합 평점의 구간에 따라 차등적인 고려가 필요함을 명시하고 있다. 0.45~0.55점 이내일 때는 ‘신중’, 0.42~0.45, 0.55~

0.58 구간일 때는 ‘약간 신중’한 판단을 권고한다. 그리고 AHP 종합 평점이 0.45 미만인 상황에서 0.5 이상의 평가자 응답 비율이 다수이거나 반대로 종합 평점이 0.55 이상인데도 0.5 이상의 응답 비율이 소수로 나타난다면 평가 내용이 왜곡되었을 가능성이 있기 때문에 이런 경우는 재평가를 실시하도록 권고한다.

이러한 판정 기준은 AHP 종합평가 결과에 대한 회색 영역의 의미를 설득력있게 제시하고 있으며 직관적으로 받아들여지기에 무리가 없는 것으로 보인다. 하지만 회색 영역 설정

의 구간값과 관련 영역에서의 판정 기준값 지정에 대한 이론적인 분석이 뒷받침되지 않아서, 그것이 과연 객관적으로 타당한지를 확인하기가 어렵다. 또한 회색 영역의 판단과 적용 기준에 있어서도 다분히 주관적인 설명과 해석을 부여하고 있는 실정이다.

이 외에도 AHP 기법이 적용된 평가 결과의 통계적 특성을 분석한 여러 가지 연구 사례가 있다. 최근 국내에서 수행된 주요 연구 사례를 살펴보면 다음과 같다. 먼저 Kim et al.[11]의 연구는 전문가 집단의 유형을 구분하고 집단별 평가 결과의 통계적 차이를 비교 분석하였다. Lee et al.[16]의 연구에서는 쌍대비교 응답 결과에 대한 일관성 판단을 위한 방법으로서 모의실험에 의한 일관성지수의 통계적 분포 검정이 적용되었다. 그리고 Jeong et al.[7] 연구는 평가 항목의 가중치를 추정하는 여러 가지 방법을 소개하고 평균제곱오차를 비롯한 통계적인 기준을 적용하여 추정 방법들을 비교하였다[7]. 한편, Eo and Kim[5]의 연구는 공공사업의 경쟁 입찰에서 사업자 결정방법을 개선하기 위한 방법의 하나로써 가중합 평가 결과에 대한 통계적 유의성 검정의 필요성을 제기하고 비모수검정과 베타분포를 가정한 근사검정 방법을 제안하였다. 이렇듯 AHP 기법을 적용한 가중합 평가 방법의 통계적 유의성을 검증하기 위해서 여러 가지 방법들이 탐색되고 있지만, 아직까지 실증적인 논거가 뒷받침된 연구 결과는 미미하다. 아울러서 통계적 검정의 대상이나 조건을 설정하는데 있어서도 이론적인 연구와 함께 보다 현실적인 관점에서의 활용성을 높이는 방안도 고려할 필요가 있다.

기존 연구 문헌에 나타난 AHP 평가의 대안

선정 및 타당성분석에 관한 사례를 정리하면 다음 <Table 2>에서 보는 바와 같다. 총 11개의 사례 중에서 타당성 분석 평가에 관한 연구 사례는 5가지이며, 대안선정 문제를 다룬 것은 6가지이다.

타당성 분석 사례에서는 대부분 최종 평가 결과에 유의한 차이가 있는 것으로 나타났다. 국내 지역공항 개발타당성 분석[13]과 에너지 어링 프로젝트 타당성 분석[8]은 시나리오에 대한 대안선정 사례이다. 한편, 충주댐 유역관리 타당성 분석[6]에서는 경제성을 고려하지 않은 중간점수가 제시되어 있고 최종 결정은 추가적인 분석이 필요함을 언급하고 있다. 항만 리모델링사업 타당성 분석[22]과 우편물 통합처리시스템 예비타당성 조사[3] 결과에서는 사업진행의 타당성을 보여줄 만한 점수 차이를 보이고 있다.

대안선정 문제의 연구 사례에서는 몇 가지 특징적으로 살펴볼 점들이 발견된다. 에너지 시스템 대안 선정[15]과 소프트웨어 개발방법론 평가[12]는 주요 대안들의 최종 점수에 충분한 차이를 보이지 않았다. 친환경 군차량 선정[14]에서는 대안간 최종 점수가 어느 정도 차이를 보이지만 기술적, 경제적 사항을 고려하여 전략적으로 선택해야 한다고 검토되었다. 한편, 모바일기기 선정 연구[2]는 대안들의 최종 평가 점수에 대한 분산분석을 실시하였는데, 결과는 대안간 유의한 차이가 없는 것으로 나타났다. 또한 정보기술 프로젝트 선정 문제를 다룬 연구[17]에서는 프로젝트 대안들의 평가 점수만으로는 유의미한 선별이 어려운 상황에서 예산 한도를 감안하여 투자 가치와 예산 집행율을 극대화하는 최적 조합을 선정한 사례가 제시되었다.

〈Table 2〉 Case Review of Project Evaluation Studies Using the Weighted Sum of Evaluation Scores

Problem type	Project title	Weighted sum of evaluation score for each alternative	Evaluation method
Feasibility Analysis	Non-Market Benefit of Port Remodelling[22]	{0.822, 0.178}	Relative evaluation
	Construction of Integrated Mail Processing System[3]	{0.788, 0.212}	
	Engineering Project Selection[8]	{0.497, 0.268, 0.235}	
	Validity Analysis on Regional Airport Development[13]	{0.698, 0.238, 0.234}	Absolute evaluation
	Feasibility of Watershed Management for Dam Chungju[6]	{0.6114, 0.1747}	
Alternative Selection	Alternative Selection of Energy System[15]	{0.918, 0.903, 0.883, 0.633, 0.618}	Absolute evaluation
	Software Development Methodology Applicability[12]	{0.317, 0.313, 0.311}	
	Green Car Technologies for Military Transport Use[14]	{0.469, 0.333, 0.198}	Relative evaluation
	Deploying Automated Gate System in Container Terminal[4]	{0.477, 0.206, 0.162, 0.083, 0.073}	
	Evaluation Model of Internet Portal Sites[9]	{0.442, 0.273, 0.194, 0.091}	
	Selecting a Mobile Device[2]	{0.254, 0.252, 0.250, 0.244}	
	IT Project Selection with Budget Constraints[17]	{0.115, 0.108, 0.102, 0.098, 0.095, 0.089, 0.085, 0.083, 0.081, 0.076, 0.069}	

Note) The number in the bracket of project title is indicated by the list of references.

이와 같이 기존에 AHP 평가 방법을 위주로 하여 여러 분야에서 전문가 집단의 의사결정 문제를 분석하고 있지만, 최종적인 평가 결과의 유의성 여부를 판단하는 기준과 근거에 관해서는 체계적으로 연구되어 있지 않다. 앞에서 살펴보았던 한국개발연구원의 예비타당성조사 수행지침에서 제시한 회색 영역 개념이나 <Table 2>의 모바일기기 선정에서 수행한 분산분석[2] 등이 통계적 검증의 개념을 도입

하여 분석한 결과이다. 이러한 사례들은 평가자별로 산출되는 평가 결과가 확률적 속성을 가질 것으로 보고, 그 분포를 정규분포로 가정하여 분석하였다. 그런데 이에 관한 실증적인 근거나 이론적 검증이 뒷받침되지 않았기 때문에 일반화해서 적용하기 곤란하다. 또한 확률 분포적 특성이 식별되지 않거나 적절한 분포의 가정이 성립하지 않는 경우에는 마땅한 접근 방법이 제시되지 않고 있다.

3. 통계적 접근 방법: 비모수적 방법론

공공사업의 평가는 대체로 기술적 요소, 경제적 요소, 정책요소 등에 대한 전문가들의 판단을 종합하여 이루어진다. 경우에 따라서 정책요소를 별도로 분리하여 고려하거나 기술적 요소와 경제적 요소를 분리하여 기술적 요소만을 별도로 평가하기도 한다. 그럴 경우에는 정책적 또는 기술적 평가 이후에 추가적으로 비용을 고려한 비용대효과 분석을 수행한다. 그런가 하면 경제적 요소를 기술적 요소와 함께 고려하여 하나의 평가를 수행하는 경우도 있다.

이 중에서 어떤 접근 방법을 사용하더라도 실제 평가는 한 명의 전문가보다는 다수 전문가의 의견을 종합하여 수행하는 경우가 대부분이다. 이 때, 다수 전문가의 의견을 종합하는 방법에는 산술평균이나 기하평균을 적용하는가 하는 문제, 전문가들이 공통의 가중치 또는 개별적인 가중치를 사용하는가 하는 문제와 같이 여러 가지 고려 사항이 있다.

여기서는 중간 과정의 다양한 접근 방법에 상관없이 n 명의 전문가가 개별적인 최종 점수를 산출하고 이 최종 점수들이 크기 n 의 확률표본을 구성하는 것으로 가정한다. X_{ij} 를 전문가 i 가 대안 j 를 평가한 최종 점수라고 하자. $i=1, 2, \dots, n$ 이고 $j=1, 2, \dots, k$ 이다. X_{ij} 는 전문가 i 가 대안들에 대하여 평가한 가중합 점수이며 0과 1 사이의 값을 가진다. 이 때 X_{ij} 의 확률적 속성에 관해서 정규분포는 물론 다른 특정 분포도 가정하기 어려운 상황이라고 하자.

그렇다면 평가 결과의 유의성 검정은 k 개

대안의 최종 점수가 의미있는 차이를 보이는지에 대한 검정으로서 통계적으로는 k 개 모집단의 평균이 모두 같은가를 검정하는 것이다. 그런데, 동일 전문가가 k 개 대안을 모두 평가하므로 k 개 확률표본이 서로 독립이라고 가정할 수 없다. $\mu_j = E(X_{ij})$, $j=1, \dots, k$ 라 하면 $H_0: (\mu_1 = \dots = \mu_k)$ 과 $H_1: (\text{적어도 하나의 평균은 다르다})$ 에 대한 비모수적 일원배치검정은 확률표본의 독립성을 보장할 수 없기 때문에 3개 이상의 집단을 비교할 때 일반적으로 적용되는 Kruskal-Wallis 검정이나 Jonckheere 검정을 적용하기에 적합하지 않다. 또한 $k \geq 3$ 인 경우에는 전문가 평가의 주된 관심은 모든 대안의 점수가 같은지 여부가 아니라 최고점수의 대안을 찾는 것이 일반적이다.

본 연구에서는 이와 같은 이유로 $k=2$ 에 해당하는 타당성분석 또는 대안선정의 최종 점수에 대한 비모수적 유의성 검정을 고려한다. 최종 점수를 산출하는 방식과 특성에 따라서 다음과 같이 유의성 검정의 귀무가설이 다르고 검정방법이 달라진다. 첫 번째는 기술적 요소와 경제적 요소의 평가를 분리한 방식에 관한 것이고 두 번째는 기술적 요소와 경제적 요소의 평가를 합한 방식에 관한 것이다.

3.1 기술적 요소와 경제적 요소의 평가를 분리한 경우: 평균의 비율 검정

경제적 요소(비용)를 포함하지 않고 기술적 요소(효과)만의 평가결과를 총비용으로 나누어서 분석하는 비용대효과 분석의 경우에는 분모에 적용되는 총비용의 단위에 따라 총비용의 차이가 많이 희석될 수 있다. 따라서 이 경우에는 두 모평균의 차이보다 모평균의 비

율에 대한 검정이 더 효과적이다. Park and Kim[11] 연구는 비열등성시험의 비모수적 검정을 위한 방법으로, 신뢰구간을 이용하여 활성대조군과 시험군의 효과 평균들의 비율에 관한 검정 방법을 제시하였다[18]. 여기서는 이 방법을 다음과 같이 비모수적 대응비교에 적용하였다.

E_{iA} 를 전문가 i 가 대안 A 에 매긴 효과점수라 하고 C_A 를 대안 A 의 총비용이라 하면 비용대효과 비율은 $X_{iA} = E_{iA}/C_A$ 로 계산된다. 같은 방법으로 대안 B 에 대하여 E_{iB} , C_B , X_{iB} 를 구한 후 X_{iA} 와 X_{iB} 중에서 평균이 큰 대안을 선택하는 것으로 비용대효과 분석을 정리할 수 있다. 이 선택에 대한 유의성 검정을 위하여 확률변수를 다음과 같이 정의한다.

$$X_{iA}, i = 1, 2, \dots, n; \mu_A = E(X_{iA})$$

$$X_{iB}, i = 1, 2, \dots, n; \mu_B = E(X_{iB})$$

여기에서 X_{iA} 와 X_{iB} 는 독립이 아니며 $\mu_A - \mu_B$ 보다는 $\frac{\mu_A}{\mu_B}$ 가 의미있는 모수이므로 다음과 같이 가설을 설정하고 검정한다. 비용대효과 분석에서 B 가 선택되었다면 $\mu_A \leq \mu_B$ 라고 가정할 수 있고, 검정의 가설은 다음과 같이 설정된다.

$$H_0: \frac{\mu_A}{\mu_B} \geq M;$$

$$H_1: \frac{\mu_A}{\mu_B} < M, 0 \leq M \leq 1$$

위의 식에서 M 은 허용범위라고 볼 수 있고 (예를 들어 $M=0.9$), 가설의 양변에서 각각 1을 빼서 정리하면 검정 가설을 아래와 같이 다

시 쓸 수 있다.

$$H_0: \frac{\mu_A - \mu_B}{\mu_B} \geq M - 1;$$

$$H_1: \frac{\mu_A - \mu_B}{\mu_B} < M - 1, 0 \leq M \leq 1$$

그런 다음, 대응비교에 관한 Wilcoxon 부호 순위검정에 기초하여 $\mu_A - \mu_B$ 의 신뢰구간을 구한다. $Y_i = X_{iA} - X_{iB}$, $i = 1, 2, \dots, n$ 에 대하여 대응비교를 위한 Wilcoxon 부호순위 통계량 W^+ 의 분포를 이용하면 다음과 같은 표본 통계량이 정의된다.

$$C_\alpha = \frac{n(n+1)}{2} + 1 - w^+(n, \alpha)$$

여기서 만약 n 이 크면, C_α 는 다음과 같이 근사적으로 정의될 수 있다.

$$C_\alpha \approx \frac{n(n+1)}{4} - z_\alpha \sqrt{\frac{n(n+1)(2n+1)}{24}},$$

(C_α 가 정수가 아니면 가까운 정수값 사용)

$N = n(n+1)/2$ 개의 모든 (i, j) 쌍에 대해서 Walsh 평균 $W_{ij} = (Y_i + Y_j)/2$ 를 구하고 W_{ij} 의 순서통계량 $W_{(1)}, W_{(2)}, \dots, W_{(N)}$ 이라 하자. 귀무가설의 조건에서 Wilcoxon 부호순위검정에 기초한 $100(1-\alpha)\%$ 신뢰구간의 상한은 다음과 같다.

$$UL_W = W_{(N+1-C_\alpha)}$$

다음으로 $\frac{\mu_A - \mu_B}{\mu_B}$ 의 신뢰구간은 위 신뢰구간을 μ_B 의 추정량으로 나누어서 구한다. μ_B 의

추정량은 비모수적 방법인 Hodges-Lehmann 추정량 $\hat{\mu}_{BHL}$ 을 사용한다. $\hat{\mu}_{BHL}$ 은 X_{iB} 자료의 Walsh 평균 W_{ij}' 의 중앙값이다[21].

그러므로 $\frac{\mu_A - \mu_B}{\mu_B}$ 의 $100(1-\alpha)\%$ 단측 신뢰구간의 상한은 $UL_R = \frac{W_{(N+1-C_\alpha)}}{\hat{\mu}_{BHL}}$ 이고, 신뢰구간 $(-\infty, UL_R)$ 이 $M-1$ 을 포함하지 않으면 H_0 를 기각한다. 즉, $UL_R < M-1$ 이면 유의수준 α 로 H_0 를 기각하고 μ_A 와 μ_B 의 의미있는 차이를 확인하고 비용대효과 분석의 선택을 뒷받침할 수 있다.

3.2 기술적 요소와 경제적 요소의 평가를 합한 경우: 평균의 차이 검정

전문가의 평가 결과로 산출되는 대안별 최종 점수가 기술적 요소와 경제적 요소를 모두 포함하여 평가한 결과인 경우에는 두 모집단의 평균의 차가 있는지를 검정하는 대응비교의 Wilcoxon 부호순위검정을 수행한다[21]. 두 개의 모집단을 각각 A와 B로 칭하면 대응표본은 X_{iA} 와 X_{iB} , $i = 1, 2, \dots, n$ 이고 그들의 모평균은 μ_A 와 μ_B 이다. $H_0: \mu_A = \mu_B$ 와 $H_1: \mu_A \neq \mu_B$ 의 검정을 위해서 $Y_i = X_{iA} - X_{iB}$, $i = 1, \dots, n$ 이라고 하면, $|Y_i|$ 값들 사이의 순위 R_i 에 대한 양의 순위들의 합으로서 정의되는 Wilcoxon 부호순위 검정통계량은 $W^+ = \sum_{i=1}^n \psi_i R_i$ 이다. 이 때, ψ_i 는 $Y_i > 0$ 일 때 1의 값을 갖는 지표함수이다. 결론적으로 귀무가설 하에서 W^+ 분포의 상위 100α 백분위수를 $w^+(n, \alpha)$ 라고 할 때, 유의수

준 α 에서의 검정 방법은 $W^+ \geq w^+(n, \frac{\alpha}{2})$ 또는

$W^+ \leq w^+(n, 1 - \frac{\alpha}{2})$ 이면 H_0 를 기각한다.

다음 장에서 살펴볼 실증 분석 사례는 평가자들이 매기는 대안별 평가 점수에 기술적 요소와 경제적 요소가 종합되어있는 경우로서 위와 같이 대안별 평가 점수에 대한 평균의 차이를 검정하는 방법을 적용한다. 그 중에서도 본 연구가 다루는 사례는 $X_{iA} = 1 - X_{iB}$ 인 특수한 상황이다. 여기에서 검정할 가설은 $H_0: \mu_A = 0.5$ 와 $H_1: \mu_A \neq 0.5$ 이다. 이 때, $Z_i = X_{iA} - 0.5$, $i = 1, \dots, n$ 이라고 하고 $|Z_i|$ 값들의 순위를 R_i 라고 하면, Wilcoxon 부호순위 검정통계량 W^+ 는 앞에서 정의한 것과 동일하게 쓸 수 있다. 아울러서 귀무가설 H_0 하에서 W^+ 의 평균과 분산을 각각 $E_0(W^+)$ 와 $Var_0(W^+)$ 라고 할 때, 본 연구에서는 대표본 근사를 고려하여 다음과 같이 표준화된 W^+ 통계량 $Z_{W^+} = \frac{W^+ - E_0(W^+)}{[Var_0(W^+)]^{1/2}}$ 을 이용하여 Z-검정을 실시한다. 그 결과, 유의수준 α 에서의 검정 기준은 $|Z_{W^+}| \geq z_{\alpha/2}$ 이면 H_0 를 기각하는 것으로 설정된다.

4. 실증 분석: 기존 사례에의 적용

이 장에서는 앞서 제시한 비모수적 방법론을 실제 문제에 적용하고 그 결과의 함의를 살펴해보도록 한다. 본 연구의 적용 대상은 현재 국방획득사업의 분석 평가를 위해 내부 정책적

판단을 지원하는 목적으로 구축된 전문가 평가 자료이다. 이 사례는 기술적 요소와 경제적 요소를 모두 포함하여 대안을 평가한 경우이며, 본 연구에서 고려하는 통계적 분석을 위해서는 제3장에서 논의한 두 가지 방법 중에서 평균의 차에 대한 검정을 적용한다.

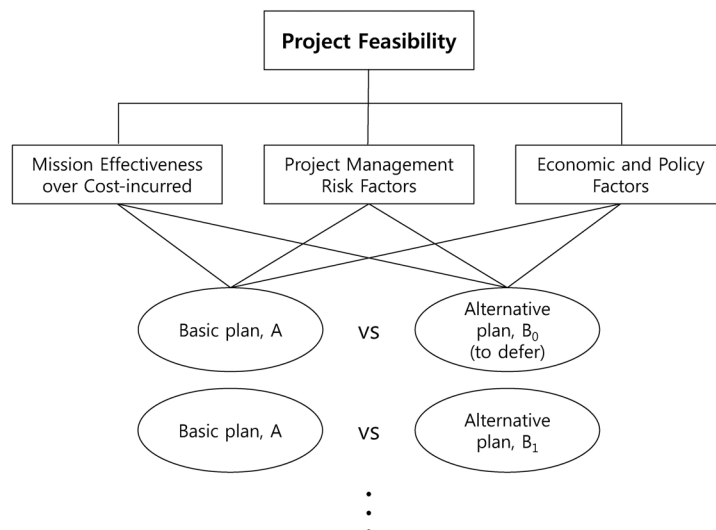
4.1 분석 사례의 구성

본 연구의 적용 대상 사례는 다음 <Figure 1>에서 보는 것과 같은 평가 구조를 가지고 있다. 각각의 사안마다 설정된 기본안의 타당성을 검증하기 위해서 여러 가지 대안을 고려하여 비교 평가한다. 대안은 사전에 별도의 절차를 거쳐서 식별되며, 전문가 그룹은 기본안과 각각의 대안을 비교하여 상대적인 지지도를 평가한다.

평가 항목은 세 가지이며 비용대비 작전 효

과도, 사업 위험 요소, 경제 및 정책 요소로 구성된다. 평가 항목마다 하부 요소가 정의되어 있으며, 필요할 경우에 하부 요소를 기준으로 평가할 수 있다. 예를 들어, 비용대비 작전 효과도의 하부 요소는 능력, 임무, 비용으로 구성되고 사업 위험 요소의 하부 평가 항목은 획득 용이성, 군수지원성, 상호운용성으로 구성된다. 그리고 경제 및 정책 요소는 경제적 파급효과, 기술적 파급효과, 기타 정책요소 등과 같은 하부 요소를 포함한다.

전문가 그룹에 선정된 평가자들은 각자 개별적으로 평가 항목의 중요도를 매기고 평가 항목마다 기본안과 대안의 상대적 지지도를 쌍대 비교한다. 기본안과 대안의 양자간 지지도는 9점 척도로 평가되며 1(대등), 3(대체로 지지), 5(강력하게 지지), 7(매우 강력하게 지지), 9(전폭적으로 지지) 등으로 구분된다. 지지도 평가 결과는 대등인 경우를 0.5점 기준으



<Figure 1> Structure of Problem Reviewed in this Research: Overall Evaluation Scheme

로 하여 0~1점 구간으로 변환된다. 기본안이 지지되는 경우는 각각 3/4, 5/6, 7/8, 9/10점으로 환산되고 대안이 지지되는 경우에는 1/4, 1/6, 1/8, 1/10점으로 환산된다. 이렇게 변환되는 이유는 지지도 평가가 AHP의 쌍대비교 척도를 사용하고 있기 때문이다. 두 개 대안을 비교할 경우에 쌍대비교행렬의 고유값을 산출하면 위와 같은 방식으로 대안간의 상대적인 지지도 점수가 매겨진다.

개별 평가자는 기본안과 대안의 비교 조합마다 별도의 가중치와 지지도 점수를 부여하게 되는데, 그 결과를 전문가 그룹의 종합 평가로 통합하는 과정이 필요하다. 이 때, 전문가 집단의 평가치를 수치적으로 통합하기 위해서는 평가자마다 평가 항목에 부여하는 가중치와 대안간 비교 자료가 모두 상이하다는 점을 고려해야 한다. 그리고 일반적으로 AHP의 쌍대비교 평가를 적용하는 경우에는 평가 결과의 역수성 조건이 충족되어야 한다. 한편, 평가자별 점수 중에서 이상치(outlier)가 주는 영향을 배제할 필요도 있다. 그런 점을 감안하여 다음과 같이 전문가 그룹의 종합 평가 점수를 산출한다[20]. 먼저 개별 평가자마다 기본안 지지도에 관한 가중합 점수를 산출한다. 그런 다음에 평가자의 가중합 점수 중에서 최대값과 최소값을 제외하고, 그 결과를 기하평균하여 최종적으로 전문가 집단의 종합 평점을 산출한다.

이와 함께 평가자 그룹 내 편차를 고려하기 위해서 평가자별 가중합 점수가 0.5점 이상인 경우를 집계하여 전문가 집단 내에서의 기본안 지지 비율을 산출한다. 궁극적으로 기본안 지지 여부를 판정하는 척도는 두 가지이다. 하나는 개별 평가자의 가중합 점수를 기하평균한 점수(이하, 기본안 종합 평점이라 함), 다른

하나는 전문가 그룹 내에서 기본안 지지도에 관한 가중합 점수가 0.5를 넘는 평가자의 비율(이하, 기본안 지지 비율이라 함)이다. 이 두가지 척도의 산출 개념을 보다 명확하게 정의하기 위해서 수식으로 표현하면 다음과 같다.

평가자가 모두 n 명이고 평가 항목이 m 개인 상황에서 X_{ijA} 를 기본안 A 에 대해서 평가자 i 가 평가 항목 j 에 부여한 지지도 점수라고 하자. 그리고 w_{ij} 를 평가자 i 가 평가 항목 j 에 부여한 가중치라고 하자. 여기에서 $i \in N = \{1, \dots, n\}$, $j \in M = \{1, \dots, m\}$ 이다. 그러면 기본안 A 에 대한 평가자 i 의 가중합 점수 X_{iA} 와 최종적인 종합 평점 X_A 는 다음과 같은 수식으로 정의된다.

$$X_{iA} = \sum_{j \in M} w_{ij} X_{ijA}; \quad \forall i \in N,$$

$$X_A = \left(\prod_{i \in N / \{n_1, n_2\}} X_{iA} \right)^{\frac{1}{n-2}},$$

$$n_1 = \operatorname{argmax}_{i \in N} \{X_{iA}\}, \quad n_2 = \operatorname{argmin}_{i \in N} \{X_{iA}\}$$

이러한 방식으로 총 k 개의 대안 $B_1, \dots, B_k$ 에 대해서 각각 기본안 A 의 지지도 평가를 실시하게 되는데, 대부분의 경우에 사업마다 기본안과 필수적으로 비교하는 대안은 ‘유보안(B_o)’이다. 즉, 사업의 소요기획을 기본안대로 계획에 반영할 것이냐 아니면 유보할 것이냐를 평가하는 것이 필수 고려 사항이다. 그런 다음에 소요를 다른 방법으로 대체하거나 또는 수량, 성능, 시기 등 소요의 결정 변수를 변경하는 대안들을 고려한다.

4.2 문제 정의

현재 구축된 전문가 평가 사례에서는 개별

대안과 비교했을 때 기본안의 우위에 관한 판정 기준을 다음 <Table 3>과 같이 제시하고 있다. 이러한 판정 기준에 내재되어 있는 개념은 기본안에 대한 전문가 집단의 전체적인 종합 평점도 높고 선호 비율도 높아야 기본안의 우위를 보장할 수 있다는 점이다. 반대로 대안의 경우에도 마찬가지이다. 그렇지 않고 선호 비율이 어느 한 쪽으로 뚜렷하게 두드러지지 않거나 종합 평점이 0.5점 근처에 위치할 경우에는 양자 대안간의 지지도 차이에 우열을 단정하기 어렵다는 점을 고려하고 있다.

이 기준은 현재 평가에 참여하는 전문가들과 최종 결정을 내리는 위원회의 구성원들에게 전반적으로 타당성과 신뢰성이 있는 방법론으로 받아들여지고 있다. 특히, 지지도와 지지율을 모두 고려함으로써 편향되거나 왜곡된 평가를 배제하도록 했다는 점에서 특징이 있다. 또한 확실한 변별력을 보장하기 어려운 상황은 재평가 영역으로 묶으로써 추가적인 검증과 보완이 되도록 한 점도 기준에 평가 점수

에만 과도하게 의존하던 상황을 개선한 것이다. 이러한 기준은 앞서 제2장에서 살펴보았던 KDI의 연구[1]에서 ‘회색 영역’ 개념을 설정한 것과 같은 맥락으로 볼 수 있다. 다만, KDI 연구는 통계적 유의성의 수준을 차등적으로 고려하여 신중한 결정을 권고하고 있는데, 본 연구의 대상 사례에서는 확실한 변별력이 보장되지 않는 경우를 모두 재평가 대상으로 설정한 점이 다르다.

아래 <Table 3>과 같이 정립된 기준은 직관적인 의미를 파악하기 쉽고 경험적으로 타당성이 인정되고 있다는 점에서 현실에서의 적용성을 갖춘 것으로 평가된다. 다만, 평가 결과의 유의성을 판단하는데 필요한 지지도와 지지율의 구간 설정에 관해서 이론적인 근거가 있었는가 하는 점에서는 약점이 있다. 개념상으로 이해하는데 무리가 없는 기준이 적용되기는 했지만, 그것이 일반화할 수 있는 기준인지 혹은 과소, 과대 설정된 기준인지에 관해서는 추가적으로 검증해 보아야 할 여지가 있다. 그러므로 현재

<Table 3> Structure of Problem Reviewed in this Research: Empirical Criteria for the Judgement of Significance on the Relative Evaluation

Total weighted sum of evaluation scores Proportion of supportive evaluation	[0, 0.45)	[0.45, 0.5)	[0.5, 0.55)	[0.55, 1]
(75%, 100%]			Basic plan supported	Basic plan supported
(60%, 75%]	reevaluation needed	reevaluation needed	reevaluation needed	Basic plan supported
[40%, 60%]	reevaluation needed	reevaluation needed	reevaluation needed	reevaluation needed
[25%, 40%)	Alternative plan supported	reevaluation needed	reevaluation needed	reevaluation needed
[0%, 25%)	Alternative plan supported	Alternative plan supported		

〈Table 4〉 The Cases Tested in this Research: Judgement Results on the Significance of Evaluation by Applying Empirical Criteria

List of cases for evaluation	Frequency distribution of weighted sum of evaluation scores									No. of eval.'s in total	Judgement by empirical criteria
	0.1~	0.2~	0.3~	0.4~	0.5~	0.6~	0.7~	0.8~	0.9~		
A1											
B1.0	4	2	2	1		7	3			(19)	REV
B1.1	2	4	5	2		3	3			(19)	B1.1
B1.2					1	5	11	1	1	(19)	A1
A2											
B2.0			1		2	6	9	2		(20)	A2
B2.1	1		1	1	1	7	9			(20)	A2
A3											
B3.0	1	2	1	1	1	2	10	2		(20)	A3
A4											
B4.0				1	1	2	14	2		(20)	A4
B4.1	1		2		1	9	6	1		(20)	A4
A5											
B5.0					1	6	12	2		(21)	A5
B5.1	2	1	3	1	3	6	3	2		(21)	REV
A6											
B6.0				1	1	12	3	4		(21)	A6
B6.1	4	4	6	1	1	1	3	1		(21)	B6.1
B6.2	2	6	7	2	2	1	1			(21)	B6.2
B6.3		4	6	4	3	3	1			(21)	B6.3
A7											
B7.0			1			3	10	4		(18)	A7
B7.1	3	1		1	1	6	6			(18)	REV
A8											
B8.0			1	1		1	10	7		(20)	A8
B8.1			1		1	5	9	4		(20)	A8
B8.2			2	1	5	5	4	3		(20)	A8
A9											
B9.0	3	3	1	1	3	5	1	6		(23)	REV
B9.1	3	4	2	3	4	3	2	2		(23)	REV
B9.2	8	3	3	4	2	1	1	1		(23)	B9.2
A10											
B10.0	2	1			1	3	4	8		(19)	A10
B10.1	3	1		3	3	2	5	1	1	(19)	REV
A11											
B11.0	2	1	3	2	2	6	2	2		(20)	REV
B11.1		5	1	4	2	5		3		(20)	REV
B11.2	7	4	3	2	3	1				(20)	B11.2
B11.3	9	4	4	1	1	1				(20)	B11.3
A12											
B12.0	2	1		2	4	7	3			(19)	A12
B12.1	6	4		2	3	4				(19)	B12.1

Note) BAS. is for 'Basic plan supported', ALT. is for 'Alternative plan supported', REV. is for 'Revaluation needed', and 'eval.' is for 'evaluators'.

본 연구의 적용 사례에서 설정된 경험적 판정 기준이 통계 이론으로도 검증 가능한 것인지를 살펴볼 필요가 있다. 그리고 통계적 방법을 활용하여 현재의 평가 기준을 보완하고 발전시키는 방안을 모색하는 것도 필요하다.

그런 관점에서 본 연구는 다음과 같은 문제를 설정한다. 앞서 제3장에서 살펴보았던 비모수적 방법론을 적용한다면 기존에 <Table 3>의 기준을 활용해서 평가된 사례와 그 결과가 일치되는가? 그리고 그 기준은 통계적으로 볼 때 어느 정도의 엄밀성을 발휘하는 것인가? 만약 통계적 방법을 적용한 결과에 차이가 있다면 그 차이에 관해서 고려해야 할 점은 무엇인가? 앞으로 통계적 방법론과 기존 방법을 상호 보완적으로 활용할 수 있는 부분은 무엇인가?

이러한 연구 문제를 실증적으로 분석하기 위해서 적용할 사례는 앞쪽 <Table 4>에 요약하여 제시된 바와 같다. 이외에도 본 연구의 적용 사례에서 검토된 분석 대상이 더 있었으나, 기본안과 대안과의 비교 자체가 필요하지 않거나 혹은 대안과의 비교가 분석적으로 의미가 없었던 경우를 제외하였다. 다음 소절에서 분석 결과를 살펴보겠지만, <Table 4>에 예시된 사례들 중에서 기존의 경험적 판정 기준에 의해 재평가 대상으로 식별된 경우가 통계적 검정을 통해서는 어떤 결과를 보이는지를 살펴보는 것이 본 연구가 제기하는 핵심 질문이다. 즉 기본안 A1, A9, A11에 대한 유보안 B1.0, B9.0, B11.0의 비교 결과와 기본안 A5, A7, A9, A10, A11에 대한 비교대안 B5.1, B7.1, B9.1, B10.1, B11.1의 평가 결과가 본 연구에서 주목해서 보고자 하는 부분이다. 아울러서 이와 같이 재평가로 판정된 경우가 아니더라도 기본안이나 대안으로 양

자 간에 판정된 결과 내에서도 추가적인 검증이 필요한 부분이 있는가 하는 점도 함께 살펴볼 문제이다.

4.3 분석 결과

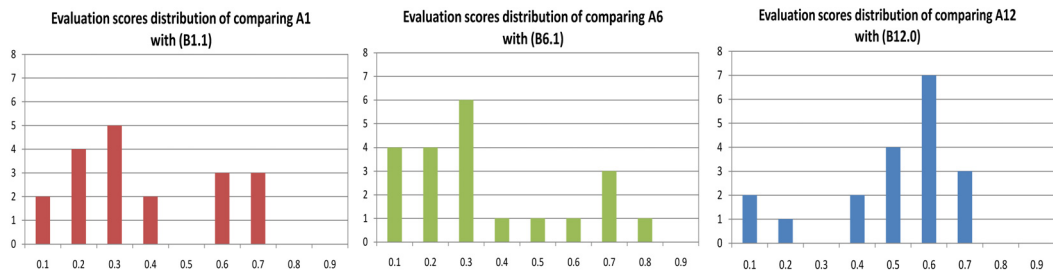
본 연구의 분석 결과를 요약하면 다음 <Table 5>와 같다. 원본 자료에서 최대, 최소 값을 제외하고 분석하였으며, 표의 좌측에는 경험적 기준에 따른 판정 결과를 제시하였고 우측에는 Wilcoxon 부호순위검정에서 귀무가설이 기각된 경우를 표시하였다.

각각의 기본안과 대안의 조합에 대한 비교 검증에 관해서 분석 대상이 되는 모수 μ_A 는 기본안 종합 평점의 모평균이다. 이에 대한 귀무가설 $H_0: \mu_A = 0.5$ 와 대립가설 $H_1: \mu_A \neq 0.5$ 을 검정하기 위해 Wilcoxon 부호순위 검정을 실시하였다.

전반적인 분석 결과를 살펴보면 기존에 경험적으로 설정된 기준에 따른 것과 본 연구에서 비모수적 통계 검정을 적용한 것이 대부분 동일한 결과를 보여주고 있다. 본 연구의 분석 사례 중에서 재평가 대상으로 판정되었던 경우는 모두 8가지이다. 그런 경우들은 모두 Wilcoxon 부호순위 검정에서 귀무가설이 기각되지 않는 것으로 나타났다. 그러므로 경험적인 기준에서 볼 때 양자 대안간 지지도의 판정이 모호한 경우는 통계적 검정에 있어서도 마찬가지로 어느 한 쪽 대안이 선호된다고 간주할 수 없는 것으로 분석되고 있다. 그리고 기본안이나 대안의 어느 한 쪽으로 판정되었던 경우는 Wilcoxon 부호순위 검정에서도 귀무가설이 대부분 기각되고 있어서 통계적으로도 양자 대안의 지지도에 유의한 차이가 있다고 할 수 있다.

〈Table 5〉 The Cases Tested in this Research: Results of the Significance Tests by Applying Wilcoxon Signed-Rank Test Compared with using Empirical Criteria

List of cases for evaluation	No. of eval. for the test	Proportion of supportive evaluation	Total weighted sum of evaluation	Judgement by empirical criteria	Wilcoxon signed-rank test			
					Test Statistics		Test result	
					W+	Z(W+)	$\alpha = 0.05$	$\alpha = 0.1$
A1								
B1.0	17	52.9%	0.396	REV	66	-0.497	x	x
B1.1	17	29.4%	0.397	B1.1	52	-1.159	x	x
B1.2	17	100%	0.723	A1	152	3.574		
A2								
B2.0	18	100%	0.700	A2	170	3.680		
B2.1	18	88.9%	0.651	A2	165	3.462		
A3								
B3.0	18	77.8%	0.593	A3	137	2.242		
A4								
B4.0	18	100%	0.732	A4	169	3.636		
B4.1	18	88.9%	0.639	A4	160	3.244		
A5								
B5.0	19	100%	0.730	A5	189	3.782		
B5.1	19	68.4%	0.519	REV	132	1.489	x	x
A6								
B6.0	19	100%	0.700	A6	189	3.782		
B6.1	19	26.3%	0.349	B6.1	47	-1.931	x	
B6.2	19	15.8%	0.334	B6.2	12	-3.340		
B6.3	19	21.1%	0.412	B6.3	37	-2.334		
A7								
B7.0	16	100%	0.741	A7	134	3.412		
B7.1	16	68.8%	0.531	REV	92	1.241	x	x
A8								
B8.0	18	94.4%	0.746	A8	169	3.636		
B8.1	18	100%	0.724	A8	171	3.723		
B8.2	18	88.9%	0.631	A8	162	3.331		
A9								
B9.0	21	61.9%	0.497	REV	149	1.164	x	x
B9.1	21	47.6%	0.407	REV	91	-0.851	x	x
B9.2	21	19.0%	0.282	B9.2)	23	-3.215		
A10								
B10.0	17	88.2%	0.654	A10	135	2.769		
B10.1	17	64.7%	0.492	REV	95	0.875	x	x
A11								
B11.0	18	61.1%	0.500	REV	106	0.892	x	x
B11.1	18	50.0%	0.459	REV	83	-0.108	x	x
B11.2	18	16.7%	0.277	B11.2	8	-3.375		
B11.3	18	5.6%	0.219	B11.3	1	-3.680		
A12								
B12.0	17	76.5%	0.526	A12	113	1.727	x	
B12.1	17	35.3%	0.321	B12.1	30	-2.201		

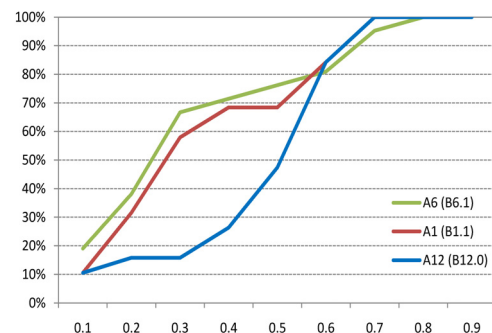


〈Figure 2〉 Histograms of Evaluation Scores Having Different Results for a Significance Test between using Empirical Criteria and Wilcoxon Signed-Rank Test

여기에서 주목해서 볼 점은 경험적으로 판정된 결과와 통계적 검정 결과에 차이가 있는 몇 가지 경우이다. 기본안 A1과 비교대안 B1.1을 대조한 경우에는 기존 판정 결과는 대안을 채택하는 것이었는데, 통계적 검정 결과는 유의하지 않은 것으로 나타났다. 그리고 기본안 A6와 비교대안 B6.1을 비교한 경우는 대안을 채택할 수 있는지에 관해서 통계적 유의성이 다소 모호한 것으로 보여진다. 유의수준 10%에서는 유의한 차이가 있다고 할 수 있지만, 유의수준 5%에서는 유의하지 않은 결과이기 때문이다. 마찬가지로 기본안 A12와 유보안 B12.0을 비교한 경우도 기본안 채택 여부에 관해서 유의수준 10%와 5%의 조건에 따라 검정 결과가 다르다.

이 세 가지 경우의 상황에 대해서 평가자의 점수 분포를 기준으로 주요 특징을 살펴보면 위에 도식된 〈Figure 2〉와 같다. 기본안 A1과 A6에 대한 평가 점수는 0.5 이하에 각각 68%, 76% 가량 분포하고 있다. 다만 기본안 지지도의 0.5~0.8점 구간에도 20~30%의 평가자 의견이 분포하고 있다는 점에서 기본안 대비 대안의 우위를 단정하기는 어렵다고도 할 수 있다. 그리고 기본안 A12에 대한 유보안의 비교 상황은 기본안 지지도가 0.6~0.8점 구간에서

높은 비율로 분포하고 있지만, 0.4~0.6 구간에 많은 의견이 있고 0.1~0.3 구간에도 다소 의견이 있기 때문에 기본안을 강력하게 지지하기는 어려운 점이 있다.



〈Figure 3〉 Cumulative Plots of Evaluation Scores Having Different Results for a Significance Test between using Empirical Criteria and Wilcoxon Signed-Rank Test

앞서 살펴본 세 가지 경우에 대해서 평가자 그룹의 점수 분포를 누적곡선으로 도식하면 〈Figure 3〉과 같다. 기본안 A1과 A6의 대안 비교에서는 각각 기본안의 지지도가 0.5점 이하에서 누적 증가율이 가파르고 0.5점 이상 구간에서

서는 증가율이 완만하다. 그만큼 0.5점 이하에 많은 비율의 의견이 분포하고 있다는 점을 의미한다. 다만, 두 경우의 차이점은 기본안 A6의 대안 비교 상황은 0.4~0.6점 구간에서 일종의 변곡점과 같은 형태를 보이며 0.6점 이상에서 증가율이 높아진 모습을 보인다는 점이다. 그런 점에서 기본안 A6에 관한 대안의 상대적 우위가 유의하지 않은 것으로도 볼 수 있다. 그리고 기본안 A12과 유보안의 비교에서는 기본안 지지도가 성장 곡선(growth curve)과 같은 모습을 보인다. 대체로 0.3점 이하와 0.7점 이상에서는 분포 빈도가 미미하지만 그 사이 0.4~0.6점 구간에서 점수 분포의 누적 증가율이 높게 나타난다. 이렇게 양자간 지지도의 중간 지대에 평가

의견이 많은 상황에서는 어느 한 쪽의 선호를 단정하기는 어려울 것이다.

한편, 이상과 같이 살펴본 세 가지 경우가 기존의 판정 기준에서는 어떤 조건에 해당되는 것이었는지를 살펴볼 필요도 있다. 다음 <Table 6>은 본 연구의 적용 사례에서 설정되었던 경험적 판정 기준표를 놓고 비모수 검정 결과와 차이가 있는 세 가지 경우의 상황을 표시한 것이다. 전반적으로 이 경우들은 구간별 분류 기준의 경계 근처에 위치한 상황임을 볼 수 있다. 이런 상황에서는 경계선과의 거리가 얼마나 되는지를 어떻게 판단하느냐에 따라 최종 판정 결과의 변별력이나 유의성에 대한 평가도 달라질 수 있다.

<Table 6> The Cases Having Different Results for a Significance Test between using Empirical Criteria and Wilcoxon Signed-Rank Test: Conditions of Empirical Criteria Fitted to Those Cases

Total weighted sum of evaluation scores Proportion of supportive evaluation	[0, 0.45)	[0.45, 0.5)	[0.5, 0.55)	[0.55, 1]
(75%, 100%]			Basic plan supported (0.526, 76.5%)*	Basic plan supported
(60%, 75%]	reevaluation needed	reevaluation needed	reevaluation needed	Basic plan supported
[40%, 60%]	reevaluation needed	reevaluation needed	reevaluation needed	reevaluation needed
[25%, 40%)	Alternative plan supported (0.397, 29.4%)** (0.349, 26.3%)*	reevaluation needed	reevaluation needed	reevaluation needed
[0%, 25%)	Alternative plan supported	Alternative plan supported		

Note) Three cases marked with '*' and '**' indicate that null hypothesis $H_0: \mu_A = 0.5$ is not rejected at the level of significance α with 0.05 and 0.1 respectively.

5. 결 론

전문가 집단의 평가 결과에 관한 통계적 유의성 문제를 분석하기 위해서 본 연구에서 비모수적 검정 방법을 적용한 결과는 기존에 실시된 경험적 판정 결과와 거의 대부분 일치하는 것으로 나타났다. 국방 분야의 특정 사례에 한정해서 분석한 것이기 때문에 이 결과가 일반적으로도 적용 가능한지를 단정할 수는 없다. 하지만, 의사결정 대안의 지지도를 평가하고 그 유의성을 판단하는 문제에 관해서 지금까지 경험적으로 또는 개념적으로 사용되었던 기준이 이론적으로도 설명이 가능한 지를 확인했다는 점에서 본 연구의 기본적인 의미가 있다.

앞서 제2장에서 기존의 연구 사례를 살펴보았듯이, 이전에도 여러 가지 상황에서 전문가 집단의 평가 방법을 통계적인 관점에서 보완하려는 시도와 제안이 있었다. 그러나 개념 수준에서의 틀을 제시하거나 또는 경험적인 기준을 적용하는 상황이 대부분이었다. 본 연구에서 시도한 이론적 접근은 기존에 적용되고 있는 방법들의 객관성과 엄밀성을 뒷받침하는 근거가 된다. 또한 여러 가지 평가 방법들이 통계적으로 어느 정도의 검정력이나 변별력을 가지는지를 비교하는 척도로서도 활용될 수 있다.

그리고 경험적 평가와 이론적 분석이 일치하지 않는 상황에 대해서 정책적 함의를 살펴보는 것도 중요하다. 통계적 검정을 통해 경험적 평가에 대한 이론적 설명을 제공하는 것뿐만 아니라 그 사이에서 차이가 발생하는 조건과 특성을 분석하는 것이 필요하다. 예를 들어서 평가 점수의 분포적 특성에 따라서 평가 결

과에 주는 영향 관계를 고려해볼 수 있다. 그 중에서 어떤 정형화된 유형이 존재한다면 평가 점수의 분포 특성에 영향을 줄 수 있는 외부적 요인을 식별하고 통제하는 방안을 구상하는 것도 가능할 것이다. 그 결과는 궁극적으로 전문가 평가 결과만으로는 파악되지 않는 추가 검증이나 정책적 판단이 필요한 사안을 식별하는데 기여할 수 있을 것이다.

앞으로 본 연구 결과를 더 많은 사례와 넓은 연구 범위에 확장해서 적용해볼 수 있다. 그러기 위해서는 다음과 같은 연구 문제에 관한 분석이 보완되어야 할 것이다. 가장 우선적으로는 본 연구에서 시도한 비모수적 검정의 신뢰성을 평가해보아야 한다. 실제로 다양한 문제와 상황을 반영했을 때 전문가 평가 결과의 유의성을 확인하는 방법으로서 비모수적 검정과 경험적 판정의 차별성을 고려하여 상호 보완적으로 활용할 수 있는 방법을 모색해보아야 할 것이다. 분석 사례가 축적되면 모수적 검정 방법을 적용하는 것도 시도해볼 수 있다.

다음으로 기존 사례에서 적용된 판정 기준의 구간 설정이 최적이냐 하는 점을 고려해야 한다. 예를 들면, 앞에서 <Table 3>에 제시되었던 판정 기준에서 구간 경계값의 변화에 따라서 통계적 검정 결과가 어떻게 달라지는지를 실험해볼 수 있을 것이다. 분석 결과에 따라서는 통계적 검정과 완전히 일치시킬 수 있는 경험적 판정 기준의 구간 경계값을 탐색해볼 수도 있다. 또는 구간 경계값이 반드시 대칭적으로 설정될 필요가 있는지 하는 점도 분석 대상이 될 것이다.

아울러 기존 사례에서 재평가 대상으로 판정되었거나 통계적으로 유의하지 않은 상황에 대해서는 그 상황의 특수한 유형과 조건을 일

반화할 수 있는 분석 척도를 개발하는 방안도 구상해볼 수 있다. 예를 든다면, 경험적 판정 기준의 구간 경계값과의 거리(distance)를 통계적으로 정의하고 분석하는 방법도 가능할 것이다. 그렇게 된다면 평가 결과의 유의성 여부에 관해서도 그 정도와 수준을 고려하여 차등적인 판정이 가능하다. 그것은 궁극적으로 평가 결과에 따른 민감도를 고려할 수 있는 정보를 제공함으로써 의사결정의 품질과 신뢰성을 높이는데 기여하게 될 것이다.

References

- [1] Ahn, S. H. et al., "A Study on General Guidelines for Pre-feasibility Study: The 5th revision," Korea Development Institute, 2008.
- [2] Byun, D.-H., "Selecting a Mobile Device Using the Analytic Hierarchy Process," *Journal of Digital Convergence*, Vol. 9, No. 4, pp. 1-8, 2011.
- [3] Chang, T., Jeong, H., Seo, Y., and Lee, S., "A Feasibility Study on Construction of Integrated Mail Processing System," *Entrue Journal of Information Technology*, Vol. 7, No. 2, pp. 151-163, 2008.
- [4] Choi, H., Lee, C., Shon, J., Lee, H., Lee, H., and Shin, J., "Selection of Optimum Alternative for Deploying Automated Gate System in Container Terminal," *Entrue Journal of Information Technology*, Vol. 8, No. 2, pp. 171-181, 2009.
- [5] Eo, H. and Kim, S.-C., "Contract Awarding Process and its Reasonable Improvement for Defense Acquisition," *Journal of the Korea society of IT services*, Vol. 14, No. 2, pp. 69-86, 2015.
- [6] Jeong, D. H., Kim, H. Y., Gi, B. J., Park, K. H., and Jin, Y. S., "Application of analytic hierarchy process on the feasibility of watershed management program by construction and operation of an integrated sewerage system in the upper watershed of Dam Chung-Ju," *Proceedings of the KSEE Fall Conference*, Korean Society of Environmental Engineers, pp. 98-103, 2004.
- [7] Jeong, H. C., Lee, J. C., and Jhun, M., "A Study for Obtaining Weights in Pairwise Comparison Matrix in AHP," *The Korean Journal of Applied Statistics*, Vol. 25, No. 3, pp. 531-541, 2012.
- [8] Jun, H., Cho, S., Kim, S., and Kim, D.-G., "A Study on Engineering Project Selection," *Entrue Journal of Information Technology*, Vol. 12, No. 2, pp. 91-105, 2013.
- [9] Kim, B., Park, J.-R., and Kim, H., "Development and Application of the Evaluation Model of Internet Portal Sites as an Advertising Medium: Employing Analytic Hierarchy Process(AHP)," *The Korean Journal of Advertising*, Vol. 23, No. 6, pp. 29-51, 2012.
- [10] Kim, J. H. et al., A Study on General Guidelines for Pre-feasibility Study: the

- 1st version, Korea Development Institute, 1999.
- [11] Kim, K.-H., Park, G.-W., and Hwang, S.-J., "An Evaluation Model of IT PMO Performance Using Pentagon Model," The Journal of Society for e-Business Studies, Vol. 21, No. 4, pp. 119-136, 2016.
- [12] Kim, K. O. and Yoo, H. Y., "An Evaluation of Software Development Methodology Applicability at Medium and Small Business through AHP," KIPS Transactions on Software and Data Engineering, Vol. 2, No. 10, pp. 691-696, 2013.
- [13] Kim, W. Y., "A Study on Validity Analysis about the Regional Airport Development," Journal of the Korean Society for Aeronautical and Flight Operation, Vol. 16, No. 3, pp. 21-30, 2008.
- [14] Kwon, K.-M., Seo, S.-D., and Jung, B.-H., "Evaluation of Green Car Technologies for Military Transport Use with AHP Analysis," Journal of Transport Research, Vol. 17, No. 2, pp. 97-111, 2010.
- [15] Lee, D. K., Park, S. U., Yang, J. T., and Gim, B. J., "An Assessment of the Alternative Selection of Energy System using AHP," Environmental and Resource Economics Review, Vol. 12, No. 4, pp. 611-635, 2003.
- [16] Lee, J. C., Jhun, M., and Jeong, H. C., "A Statistical Testing of the Consistency Index in Analytic Hierarchy Process," The Korean Journal of Applied Statistics, Vol. 27, No. 1, pp. 103-114, 2014.
- [17] Park, J., Cho, N.-W., and Kim, W., "A Study on the IT Project Selection Considering Budget Constraints," The Journal of Society for e-Business Studies, Vol. 18, No. 4, pp. 327-338, 2013.
- [18] Park, S. and Kim, D., "Nonparametric Method for a Non-inferiority Test using Confidence Interval," The Korean Journal of Applied Statistics, Vol. 27, No. 5, pp. 833-842, 2014.
- [19] Saaty, T., The Analytic Hierarchy Process, McGraw-Hill, New York, 1980.
- [20] Saaty, T., Multicriteria Decision Making: The Analytic Hierarchy Process, McGraw-Hill, New York, 2003.
- [21] Song, M. S. and Park, C. S., Introduction to Non-parametric Statistics, Free Academy Publishing Co., 1989.
- [22] Yang, C. H., Yeo, G. T., Jang, H. J., Kim, H. K., and Son, Y. L., "Analysis on Non-Market Benefit of Port Re-Modeling Factor Using AHP," Journal of Korean Society of Coastal and Ocean Engineers, Vol. 23, No. 1, pp. 57-62, 2011.

저 자 소 개



박준수

1997년

1999년

2001년~현재

2008년~현재

2015년

관심분야

(E-mail : jspark@kida.re.kr)

한국과학기술원 산업경영학부 (학사)

고려대학교 경영학과 (석사)

한국국방연구원 국방획득연구센터 연구위원

숭실대학교 정보통계보험수리학과 박사과정

미국 코넬대학교 PACS 연구소 방문 연구원

의사결정분석, 기술경영전략, 국방경제학



김성철

1982년

1984년

1988년

1989년~1991년

1991년~1992년

1993년~현재

관심분야

(E-mail : sskim@ssu.ac.kr)

서울대학교 금속공학과 (학사)

Case Western Reserve University, OR학과 (석사)

U. C. Berkeley 산업공학과 (박사)

한국국방연구원 선임연구원

수원대학교 응용통계학과 조교수

숭실대학교 정보통계보험수리학과 교수

베이지안 분석, 국방통계모형, 지식재산권 가치평가 등