

R을 활용한 관광데이터 분석 처리 향상 방법: SPSS 프로그램과의 비교를 중심으로

How to Improve Tourism Data Analysis Processing Using R: Focusing on Comparison with SPSS Program

김철민* · 황철현** · 송학준***

Kim, Cheol-Min · Hwang, Chul-Hyun · Song, Hak-Jun

Abstract

Data analysis has recently become a big trend in modern society, which naturally gave rise to the term of data science. In particular, in the era of big data, the importance of scientific decision-making and policy implementation, to make decisions and strategies using data, will continue to increase in the field of tourism. In this situation, interest in R as an alternative to SPSS is increasing significantly. As a free open source program, R is excellent in package extensibility, the introduction of the latest techniques, data visualization, and reproducible research, so it can be usefully employed in data analysis in tourism. In tourism field, R is expected to be actively used while sufficiently replacing SPSS. R would be one of the good choices to deal with statistics more broadly and to bring scalability in many academic fields, including tourism.

주제어: R, SPSS, 관광데이터(Tourism Data), 데이터 분석(Data analysis),
데이터 과학(Data Science), 빅데이터(Big Data), 오픈소스(Open
Source)

* 배재대학교 관광·호텔경영학과 석사과정, 연구 관심 분야: 관광법규, 관광레저,
e-mail: kcm4677@kakao.com

** 경북대학교 소프트웨어학과 부교수, 연구 관심 분야: AI·소프트웨어, 빅데이터 분석,
e-mail: chwang@kbu.ac.kr

*** 배재대학교 호텔항공경영학과 부교수, 교신저자, 연구 관심 분야: 소비자 행동, AI·소프트웨어,
빅데이터 분석, e-mail: bloodia@pcu.ac.kr

I. 서 론

과거에 통계 및 데이터 분석(이하 “데이터 분석”)은 복잡한 계산을 통해 도출된 분석 결과를 해석하고 타당성을 검증하는 과정이 어려워 많은 이들이 쉽게 도전하지 못하던 분야였다(송학준 외, 2020). 하지만 데이터 분석은 최근 현대사회의 큰 추세가 되었고 이는 데이터 과학이라는 용어를 자연스럽게 등장시켰다. 논리적이지 않고 검증이 불가능하며 직관에 주로 의존하던 많은 의사결정이 이제 데이터 과학을 통해 이루어지면서 공학, 의학 등의 자연과학은 물론 교육, 서비스, 금융, 마케팅 등의 사회과학 분야에서도 데이터 과학이 관심받고 있다(Tang & Ji, 2014; 신영송 외, 2018). 특히 여러 분야에서 수집된 대용량 데이터를 다루고 고급 통계기법을 통해 관련 시사점을 도출하여 최적의 의사결정을 도출하는 빅데이터 시대의 도래는 데이터 과학의 발전 가능성을 더욱 높이고 있다. 관광분야에서도 관련 데이터들이 통계청, 한국은행, 관광지식정보시스템 등 여러 곳에서 축적됨에 따라 경제학, 경영학, 심리학, 교육학 등에서 사용되는 고급 통계분석의 활용 가능성은 커져 왔다. 고급 통계분석의 활용은 관광 데이터 분석에 다양하고 새로운 방법을 적용하는 계기를 마련하여 괄목할만한 발전을 가져올 것으로 예상된다(김주향 외, 2017).

그동안 SPSS와 같은 전통적 통계 소프트웨어 패키지는 관광 데이터 분석에서 주류로 자리 잡아 왔다. 하지만 최근 들어 앞서 설명된 변화를 대비하기에 전통적 소프트웨어는 최신 통계분석 방법을 수용하기 어렵고 반복 작업 시행 시 번거로우며 통계분석에 대한 지식을 공유하고 함께 발전시키는 데 있어 한계점을 보이고 있다. 또한, SPSS는 제한된 데이터 저장으로 대규모 데이터를 처리하는데 적합하지 않고 비공개 소스에 기반하고 있으며 구매 비용과 유지비가 많이 들어 어느 정도 규모의 기업과 조직만이 SPSS를 지속해서 사용할 수 있어서 일반 사용자들은 이를 연습해 보고 사용하는 데 있어서 많은 어려움을 겪어왔다(이승덕 외, 1998).

한편 최근 관심 받고 있는 오픈소스 프로그램 중 R은 이러한 SPSS의 부족한 점을 훌륭하게 보완할 수 있는 대안으로 평가받고 있다. 미국 AT&T사 Bell 연구소의 S언어를 기반으로 개발된 객체지향 프로그램 R은 데이터 처리 및 분석작업을 대화형으로 처리하고 고품질의 그래프 기반 자료 분석작업도 용이하다. Ross Ihaka와 Robert Gentleman에 의해 초기 발전을 시작한 이후 R은 전 세계의 개발 팀에 의해 꾸준히 확장 및 발전되고 있다. 또한, 같은 S 언어를 기반하는 S-PLUS와 달리 R은 GNU 규약 하의 소스 개방으로 누구나 자유롭게 쓸 수 있어서 꾸준히 발전하고 있다(Tang & Ji, 2014). 높은 발전 가능성에도 불구하고 국내에서는 아직 사용자들이

많지 않고 특히 이를 SPSS와 같은 기존 패키지를 대체할 정도로는 발전되지 못하고 있는데 이러한 상황은 오픈소스 프로그램으로서의 R이 사용자들의 편의를 최대한 고려하지 못하고 있고 20년 전부터 차근차근 성장해온 SAS나 SPSS와 같은 통계 패키지보다 R의 인지도가 높지 않기 때문에 만들어진 것으로 보인다. R을 통해 호텔관광 분야 데이터를 분석한 연구를 살펴보면 송학준과 박혜미(2020)는 R의 기본 패키지를 통해 소확행 여행에 대한 탐색적 연구를 진행하였고 김주향·송학준(2017)과 신영송·김효은·송학준(2018)은 R의 lavaan 패키지를 이용해 HMR 소비자와 소셜미디어 외식 소비자에 대한 구조방정식 분석을 하였다. 박경열과 이현주(2019)는 R의 rpart 패키지를 통해 관광정책 결정지원을 위한 의사나무결정분석 연구를 하였다. 이처럼 그동안 관광데이터를 실증적으로 분석하는 데에서도 R의 활용은 극히 제한되어 왔다. 하지만 관광분야 여러 연구자가 관심을 보이는 고급 통계기법(조건부 프로세스 분석, 빅데이터 분석, 패널 데이터 분석, 최신 시계열 기법 등)을 SPSS를 통해 시행하기는 쉽지 않고 불가능한 경우도 있다. 예를 들어 Andrew Hayes 교수가 개발한 조건부 프로세스 분석은 관련 모듈을 애드온 형식으로 SPSS에서 추가하여 분석할 수 있지만, 이는 R에서도 유사하게 이루어질 수 있을 뿐만 아니라 관련 분석 결과를 시각적으로 표현하는 추가작업도 보다 용이하다는 장점이 있다. 이에 본 연구는 관광 데이터 분석에서 SPSS 대신 R이 유용하게 사용될 수 있음을 보여주는 것을 주요 연구목적으로 삼고자 한다.

구체적으로 본 연구에서는 (1) R의 장단점, (2) R의 간단한 사용 소개, (3) R 및 SPSS의 기본 통계분석 비교, (4) R이 더 잘 수행 할 수 있고 SPSS가 수행하지 못하는 분석 등에 대해 살펴볼 것이다. 한편, 본 연구는 R의 매뉴얼을 위해 이루어지는 것이 아니므로 보다 복잡한 통계분석 방법이나 R의 본격적 활용을 원하는 경우에는 관련 논문이나 서적을 참고하는 것이 좋을 것이다. 본 연구는 향후 관광분야 데이터 분석에서 R 사용에 대한 인식을 확대하고 당위성을 기여하며 향후 R을 통해 관광분야 데이터 분석이 이루어질 수 있는 토대를 제공해 줄 것으로 기대된다.

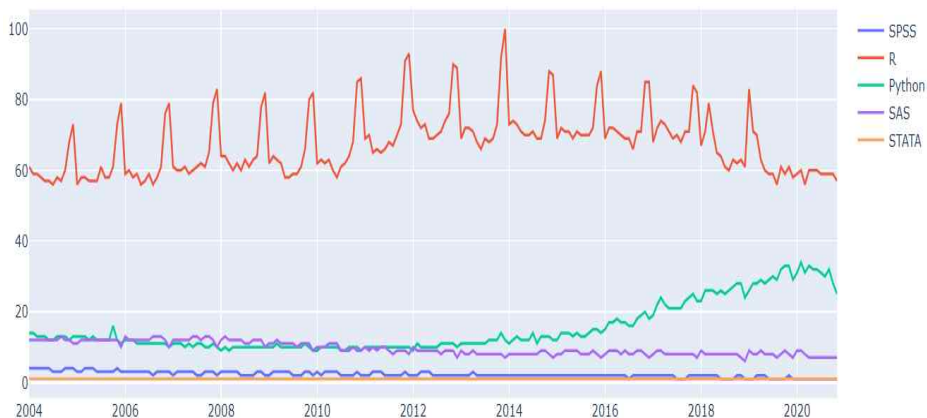
II. 이론적 고찰

1. R을 배우고 사용하는 이유

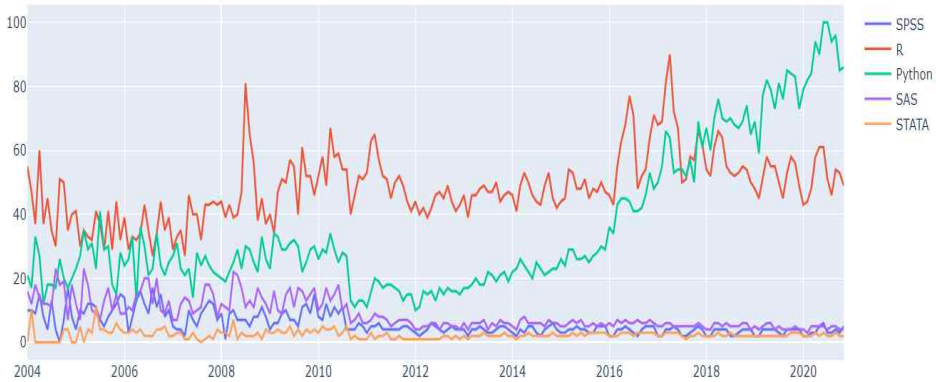
그동안 널리 사용되어온 통계 패키지로는 SPSS, SAS, Stata, Matlab, Minitab 등이 있다(조미순·김순귀, 2003). 관광 데이터 분석에서 SPSS는 정량 데이터 분석

을 위한 가장 인기 있는 통계 소프트웨어로 자리 잡아 왔다. SPSS는 주로 사회과학을 위한 통계 패키지로 1968년 처음 개발되어 SPSS Inc.가 개발해오다(이승덕 외, 1998) 2009년부터 IBM SPSS Statistics로 변경되면서 IBM이 이를 개발 및 판매하고 있다. SPSS는 특별한 교육 없이 버튼만 눌러서 통계분석 결과를 도출하고 만족할만한 시각적 효과도 얻을 수 있는 편리성으로 인해 실무자들에게 인기를 얻어왔다(Muenchen, 2011). 하지만 SPSS는 분석 가능한 통계분석 기법이 다양하지 못하고 매크로를 프로그래밍할 수 있는 기능이 유연하지 못하다는 단점이 있다. 많은 관광 데이터 분석 연구자들이 SPSS를 선호하는 중요한 이유 중 하나는 직관적인 인터페이스와 엑셀과 같은 top-down 메뉴 기반의 분석 방법이다. 관광 데이터 분석에 대한 기본 통계분석은 SPSS에서 메뉴와 대화상자에서 마우스를 여러 번 클릭하여 비교적 쉽게 수행될 수 있다.

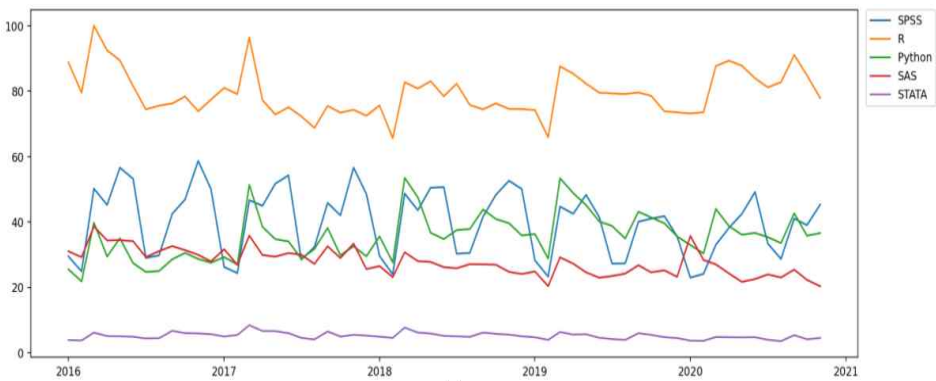
한편 SPSS에는 적절히 처리하지 못하는 몇 가지 문제가 있다. 예를 들어 회귀분석에서 여러 독립변수의 유의도를 개별적으로 또는 동시에 살펴보기 위해 여러 번의 회귀분석을 실시하여 이를 1개의 표에 정리하는 경우 R의 추가 패키지를 사용하면 이를 보다 간편하고 효율적으로 수행할 수 있어서 R이 더 나은 선택이 될 수 있다. SPSS에서도 회귀분석에서 옵션을 선택하면 중요한 독립변수를 기준으로 결과를 제시하지만, 이는 보통 사용되는 표 형식이 아니어서 결과를 논문이나 보고서를 활용할 때 추가로 작업을 해주어야 한다는 번거로움이 있다. 이를 처치하더라도 최근 SPSS에 대한 관심은 R이나 다른 오픈소스 프로그램인 Python보다 떨어지고 있다. 이에 대한 근거는 구글 트렌드나 네이버 트렌드를 통해 살펴볼 수 있다(<그림 1, 2, 3 참조>).



〈그림 1〉 구글 트렌드를 이용한 세계 프로그램 검색량 지수



〈그림 2〉 구글 트렌드를 이용한 한국 프로그램에 대한 검색량 지수



〈그림 3〉 네이버 트렌드를 이용한 한국 프로그램에 대한 검색량 지수

2. R 프로그램의 장·단점

1) 장점

(1) 무료이고 오픈소스

Tang과 Ji(2014)는 R의 장점 중 가장 강력한 것으로 R이 무료이고 오픈소스라는 것을 강조했다. R은 오픈소스 언어로써 기본 프로그램은 물론 특정 목적을 위해 개발된 패키지도 누구나 사용할 수 있다. R은 또한 Windows, Macintosh 및 Unix/Linux를 포함한 대부분 운영체제에서 잘 실행된다. 이는 R을 사용하는 연구

자가 더는 사용하는 통계 패키지의 가용성과 비용을 걱정할 필요가 없다는 것을 얘기해 준다. 유료 통계 패키지의 가격은 연구자가 부담하기에는 어려운 가격인 경우가 있다. 예를 들어 IBM SPSS v26에는 Base, Standard, Professional, Premium 4가지 Pack이 있는데 1년 이용가격을 살펴보면 Base는 \$1,290 USD(약 151만 원)이고 Standard는 \$2,850 USD(약 335만 원), Professional은 \$5,730 USD(약 647만 원), Premium은 \$8,540 USD(약 1,006만 원)로 이용가격이 매우 높다(<https://www.g2.com/products/ibm-spss-statistics/pricing>). 매달 이용할 때도 사용자는 \$99 USD를 지급해야 한다. 회사, 기관, 개인의 상황에 따라 이용 가격의 부담 정도가 다를 수 있지만 이러한 비용은 통계분석의 가치와 인식확산 그리고 통계 패키지 이용에 있어 충분한 장애요인이 되어 왔다.

SPSS에는 무료평가판이 존재하고 있지만, 이는 한 달 동안만 사용할 수 있다. R은 통계분석에 관심이 있지만, 상업적 통계 소프트웨어에 쉽게 접근할 수 없는 사람들에게 유용하게 사용될 수 있다(Fox & Andersen, 2005). 이들을 대상으로 하는 통계 패키지를 통한 적극적인 통계교육은 관광은 물론 여러 분야에서 데이터 분석의 미래인재를 길러내는데 장기적 관점에서 기여할 수 있다. R은 또한 오픈소스 프로그래밍 언어이지만 SPSS 또는 SAS 소프트웨어만큼 정확하다(Kleinman & Horton, 2009). 성공적인 오픈소스 프로젝트(Linux, MySQL 등)와 마찬가지로 통계 전문가들의 지속적 모니터링에 기반하여 활발한 코드 개선을 통해 R은 높은 통계 품질표준과 수치 정확도가 유지되도록 하고 있다. 뿐만 아니라 R은 개방형 인터페이스를 통해 애플리케이션 및 시스템과 쉽게 통합될 수 있다는 장점이 있다.

(2) 패키지

CRAN(<https://www.r-project.org/>)에서 다운로드 받을 수 있는 것과 깃허브에서 받을 수 있는 것을 고려하면 R 패키지는 2만 개 이상이 될 것으로 판단된다. 뿐만 아니라 R에서는 숙련된 개인이 필요한 패키지를 스스로 만들어 공유할 수도 있다. 이러한 R의 확장성은 상용 패키지인 SPSS, SAS, STATA 등과 비교할 때 강력한 장점이 된다. 특히 빅데이터 분석 및 데이터 시각화를 위한 R의 확장성은 R의 인기가 지속해서 높아지는 데 기여하고 있다.

(3) 업데이트

R은 오픈소스 프로그래밍 언어로서 대규모 커뮤니티를 가지고 있는데 이는 R이 소프트웨어 업데이트를 빠르게 제공하고 사용자가 더 나은 기능을 이용할 수 있도록 관련 패키지를 추가, 변경, 사용법 논의 등이 지속해서 이루어지는 데 기여하고 있다.

〈표 1〉 SPSS와 R 프로그램 통계분석 방법 비교

통계분석 방법	R	SPSS	통계분석 방법	R	SPSS
Bayesian Statistics	Yes	No	Nonparametric Tests	Yes	Yes
Experimental Design	Yes	No	T-test	Yes	Yes
Univariate Time Series	Yes	Yes	ANOVA & MANOVA	Yes	Yes
Multivariate Time Series	Yes	No	ANCOVA & MANCOVA	Yes	Yes
Hidden Markov Models	Yes	No	Linear Regression	Yes	Yes
Random Forests	Yes	No	Generalized Linear Models	Yes	Yes
Support Vector Machines	Yes	No	Logistic Regression	Yes	Yes
Wavelet Analysis	Yes	No	Mixed Effects Models	Yes	Yes
Bagging	Yes	No	Factor & Principal Components Analysis	Yes	Yes
Meta-analysis	Yes	No	Canonical Correlation Analysis	Yes	Yes

자료: 연구자가 요약정리

반면 SPSS는 오픈소스 프로그래밍 언어가 아닌 IBM의 상용제품으로서 R과 같은 커뮤니티가 많지 않아서 빠른 업데이트도 제공되지 않을 뿐만 아니라 많은 경우 유료로 진행되고 있다(Muenchen, 2011). 업데이트와 관련하여 R에는 수많은 메일링 리스트, 포럼 및 블로그가 있다. 예를 들어 자습서, 토론 및 문제해결을 포함한 리소스, R-help 메일링 리스트 (<https://stat.ethz.ch/mailman/listinfo/r-help>), 로컬 R 사용자 그룹 디렉터리(<https://blog.revolutionanalytics.com/local-rgroups.html>)가 있다. 이 외에도 R을 대상으로 하는 Stack Overflow (<https://stackoverflow.com>), Cross Validated (<https://stats.stackexchange.com>) 및 Talk Stats (<http://www.talkstats.com/>) 등의 다양한 전문 블로그 및 사이트가 있다. 이처럼 R은 사용자, 패키지 개발자 및 서적 저자로 구성된 대규모 지원 커뮤니티를 보유하고 있고 서로 돕고 통계 소프트웨어의 발전을 촉진할 수 있는 플랫폼을 성공적으로 구축시켜 왔다. SPSS와 비교해볼 때 R에서 가능한 통계분석 방법을 간단히 정리해보면 <표 1>과 같다.

(4) 강력하고 유연

R은 관광 데이터 분석에 필요한 대부분의 기본 통계기능을 가지고 있고 여기에 다양한 분야에서 사용되는 광범위한 분석 방법이 가능한 수많은 추가 패키지가 제공되고 있다. R의 기능은 통계뿐만 아니라 컴퓨팅, 모델링, 기계학습 및 데이터 마이닝에서도 적용할 수 있다. R의 유연성은 사용자가 R과 상호작용이 잘 이루어지

도록 도와준다. 추가 명령을 통해 사용자는 단일 최종 보고서 대신 중요한 중간 출력을 얻는 것이 가능하여 사용자는 이를 통해 통계분석을 수행하는 원칙과 절차를 전반적으로 이해할 수 있고 통계학을 교육하는 데 있어 큰 장점으로 작용한다. 특히 R에서는 모든 명령줄이 저장되어 있으므로 사용자가 이전 작업을 기록하고 유사 분석을 위해 프로그램을 재사용하는 데 있어 매우 편리하게 작용할 수 있다 (Zuur, Ieno & Meesters, 2009).

(5) 데이터 시각화

R은 만들어질 당시부터 데이터 분석, 통계학, 시각화를 더 좋고 용이하게 할 수 있도록 고안되었다. 시각화된 데이터는 원자료가 가진 의미를 보다 효과적으로 전달하는데 기여할 수 있다. 구체적으로 R은 고품질 그래픽 생성이 가능하고 선 스타일, 글꼴, 색상, 축 및 제목을 포함한 기능 옵션을 사용자가 편리하게 지정할 수도 있다. R로 개발된 그래프는 pdf, png, jpeg, postscript 등 다양한 형식으로 저장된다. 이는 R이 연구대상에 대해 특징하고 발전된 분석은 물론 의미전달이나 보고가 주요 목적인 탐색적 연구에서도 용이하게 이용될 수 있는 계기를 마련해 주었다 (이윤환, 2016). 데이터 시각화와 관련된 패키지에는 ggplot2, ggvis, googleVis, rCharts 등이 있다. 이런 패키지들을 통해 R에서는 간단한 그래프뿐만 아니라 전문적인 그래프, 상호작용이 가능한 interactive plot 등도 쉽게 작성이 된다.

(6) 코드 활용 및 재현 가능한 연구

SPSS에서는 자료파일을 로드, 변형, 분석 및 저장하는 과정이 주로 메뉴와 마우스를 통해 진행되지만, R에서는 모든 작업이 코드를 기반으로 행해진다(정철용 외, 2021). 이것이 이용 초기에는 R을 사용하는 데 있어 생소하고 어렵게 느껴지게 할 수도 있지만 모든 분석과정을 코드로 남겨놓도록 하는 R의 이러한 특징은 반복 작업을 용이하게 해주고 유사 연구를 할 때 분석을 더 빠르게 해주고 나아가 해당 부분을 사용자 함수 등을 통해 상당 부분 자동화시키는데 기여할 수 있다는 점에서 향후 장점으로 작용할 수도 있다. 또한 이러한 R의 특성은 R이 재현 가능한 연구(reproducible research)의 대표사례가 되는 데 크게 기여하였다. 재현 가능한 연구는 원시데이터와 데이터 전처리(preprocessing), 분석 코드(analytic code), 통계처리, 테이블, 그림 등을 하나의 문서 안에 재현시킴으로써 통계교육이나 특정 프로젝트를 여러 사람과 공유할 때 큰 생산성을 만들어 낼 수 있다. 또한 Shiny 등의 패키지를 이용하게 되면 R에 기반하여 분석되는 과정을 SPSS와 같은 하나의 웹 프레임워크로 표현할 수 있게 되므로 이는 웹상에서 구동되는 웹 애플리케이션 구축에 기여할 수 있다.

2) 단점

SPSS의 top-down 클릭 인터페이스에 익숙한 연구자에게 R의 명령어 기반의 운용은 명백한 장애요인 및 단점이 된다(Verma, 2012). 대부분의 경우 하나 또는 몇 개의 명령줄로 R은 충분한 통계분석이 이루어질 수 있음에도 불구하고 많은 사람은 여전히 마우스로 이동하고 클릭하는 방식을 선호하고 있다. R에서 코드를 직접 입력하는 것은 오류를 유발할 수 있고 코드를 수정해 나가는 디버깅과정은 특히 초보자에게 매우 어려운 프로세스가 될 수 있다.

(1) 느린 속도

언급하였듯 R은 통계학자들이 좀 더 쉽게 연구하려고 만든 언어이기에 컴퓨터를 효율적으로 활용하는 방법은 그리 깊게 고려하지 않은 듯하다. 다시 말해 R은 읽기 어려운 코드 때문에 좀 느릴 수 있다. 하지만 그런 단점을 상쇄해줄 수 있는 `pqR`, `renjin`, `FastR`, `Riposte` 등의 패키지가 있다.

(2) 어렵다

R을 배우는 것은 그리 쉬운 일이 아니다. 풍부한 패키지가 있지만 그걸 사용하는 데 익숙하지 않다면 활용하는 데 시간이 꽤 오래 걸린다. 입문자가 처음부터 R을 활용해 원하는 분석을 할 수는 없고 꾸준한 학습이 필요하다. 이 외에도 R의 단점으로 지적되는 것은 R의 메모리 관리이다. R은 모든 데이터를 메모리에 저장하여 사용하므로 시스템의 메모리보다 큰 데이터를 다루는 데 한계가 있다. 이런 경우에는 `hadoop`, `spark` 등을 이용한 분산처리로 해결할 수 있다.

(3) 상호작용

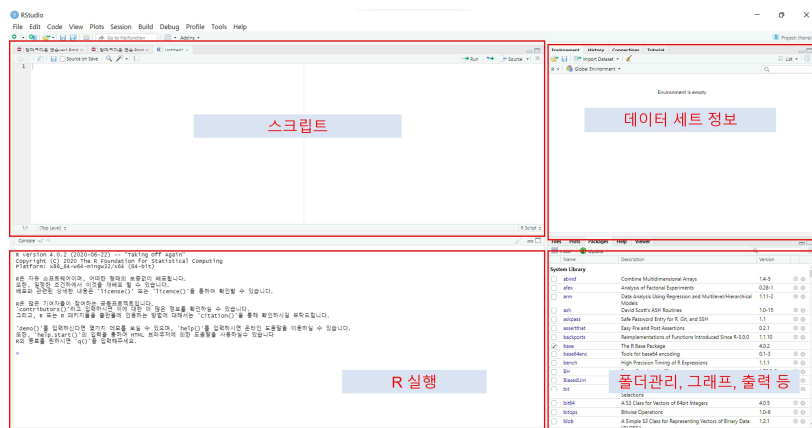
SPSS가 R보다 대화형 분석 도구로는 더 선호된다. 하지만 R에는 GUI 지원이 가능한 다양한 편집기가 있고 분석을 배우고 연습하고자 한다면 R이 분석 단계와 명령을 배우는 데 있어서 유용하게 사용될 수도 있다. 이를 보완하기 위해 기본 R 프로그램을 사용하기보다는 작업 디렉터리 변경, 데이터 가져오기 및 보기와 같은 작업을 위한 메뉴와 대화상자가 있는 R Studio (<http://www.rstudio.com>)와 같은 사용자 친화적인 R 인터페이스의 향상된 추가 프로그램을 통해 작업을 더 쉽게 완료시킬 수 있다. 또한 사용자에게 도움말 파일 형식에 익숙해지도록 하고 예제를 모방하는 방식을 중심으로 프로그램을 교육하면 이러한 문제를 예상보다 쉽게 해결할 수 있다.

Ⅲ. 기초통계 결과 비교분석

1. R 사용 소개

1) 설치 및 인터페이스

R은 Windows, Mac 및 Linux까지 사용자의 컴퓨터 운영체제(OS)에 맞는 설치과 일을 <http://cran.r-project.org>의 CRAN(Comprehensive R Archive Network)에서 다운로드 할 수 있고 설치도 간단하다(정철용 외, 2021). Windows에서 R을 열면 "RGui"라는 프레임이 화면에 나타난다. 그러나 이 인터페이스에는 데이터 셋 가져 오기와 같은 일부 기본작업을 위한 메뉴가 없고 새로운 사용자는 처음부터 어려움을 느낄 수 있다. 따라서 <http://www.rstudio.com>에서 추가 프로그램으로서 R Studio를 이용하는 것이 좋다. R Studio를 시작할 때마다 R은 백그라운드에서 자동으로 시작되는데 네 개의 창으로 구성된 R Studio 인터페이스가 기본적으로 실행된다. 좌측 하단은 명령 프롬프트(>) 다음에 명령을 입력하고 Enter 키를 누르면 실행되는 대화 형 콘솔이다. 좌측 상단은 여러 줄의 명령을 작성하고 저장할 수 있지만, 실행 버튼을 클릭할 때까지 실행되지 않는 스크립트 창이다. 우측 상단은 컨텍스트의 데이터 이름과 값을 표시할 수 있는 작업 영역 창이다. 바닥 오른쪽 창에는 파일을 열고, 플롯을 보고, 패키지를 설치하고, 기능에 대한 도움말 정보를 찾을 수 있는 메뉴가 있다. 모든 창은 재배열 및 크기조정이 가능하다.



〈그림 4〉 R Studio의 화면구성

2) 패키지

R 패키지는 특정 작업을 위한 함수들의 모임을 의미한다. R이 설치되면 기본 패키지 모음이 자동으로 포함되고 설치된 패키지 목록은 R Studio의 오른쪽 아래 창에서 찾을 수 있다. 가장 일반적인 통계기법은 처음부터 시행할 수 있지만 보다 최신의 통계기법이나 시각화 데이터의 추가변환을 할 때는 도구 → R Studio에서 패키지 설치를 통해 추가 패키지를 설치해야 한다. 한편 SPSS와 유사한 인터페이스가 있는 Rcmdr 과 같은 특수 추가 패키지도 있다. SPSS 사용자는 이러한 패키지를 더 편하게 생각할 수도 있지만 여기서 제공되는 기능은 매우 제한적이다. 또한 이를 확장하기 위해서는 다시 R 프로그래밍을 공부해야 하는 경우가 많아서 크게 추천하지는 않는다.

3) R파일

R을 종료하면 모든 개체, 함수, 데이터 및 로드된 패키지가 포함된 현재 작업 공간 이미지를 저장할 것인지 묻는 프롬프트 창이 나타나는데 확장자가 * .RData인 작업 공간 이미지 파일은 간단히 클릭하여 다음에 사용할 수 있다. 또한 콘솔 창이나 스크립트 창에 입력된 모든 명령은 * .Rhistory 확장자를 사용하여 R 히스토리 파일에 저장된다. 이 파일은 일반 텍스트 편집기로 열 수 있으므로 R을 열지 않고도 작업내용을 확인할 수 있다. 또한 이 작업내용을 R에 붙여넣어 반복 입력을 최소화하는 것이 가능하다.

2. R과 SPSS 기초통계 비교

1) 데이터 설명

본 연구의 기초통계 비교에서 사용할 데이터는 2017년 인천 영종도의 파라다이스 시티 복합리조트 개장을 앞두고 지역주민들에게 실시된 설문조사가 코딩된 자료이다. 이에 대한 실제 자료와 설문지는 “<https://github.com/HakJun-Song/class>”에서 다운받을 수 있다. 분석에 사용될 주요 변수에 대한 설명은 다음과 같다.

〈표 2〉 주요 변수

변수명	변수내용	변수형태	코딩형태
Q1	성별	질적변수(명목변수), 카테고리 2개	1:남자, 2:여자
Q10	긍정적 경제인식1 지역주민 고용기회가 확대될 것이다		
Q11	긍정적 경제인식2 지역주민 소득이 증가할 것이다		
Q12	긍정적 경제인식3 지역개발 재원이 마련될 것이다		
Q23	긍정적 사회인식1 지역주민의 위락시설 이용기회가 확대될 것이다	양적변수(등간변수) 리커트 5점 척도	1(매우그렇지않다) ~5(매우그렇다)
Q48	인지1 영종도 복합리조트 개발은 지역사회를 살기 좋은 곳으로 만들 것이다		

2) 데이터 로드

R에서는 다양한 형식의 데이터 로드는 물론 인터넷에서 직접 데이터를 가져올 수도 있다. 본 연구에서는 향후 사용자의 편의를 위해 인터넷에서 존재하는 예제 데이터를 사용하고자 한다. <표 3>의 코드를 통해 R을 통해 인터넷에 존재하는 데이터를 로드할 수 있다. 위 코드 이외에 다른 방법으로서 R에서 데이터를 편하게 사용하기 위해서는 Microsoft Excel에서 데이터를 사전 처리하고 텍스트 파일 (.txt 또는 .csv)로 저장한 다음 R Studio에서 Tools → Import Dataset을 클릭하여 데이터를 가져올 수도 있다.

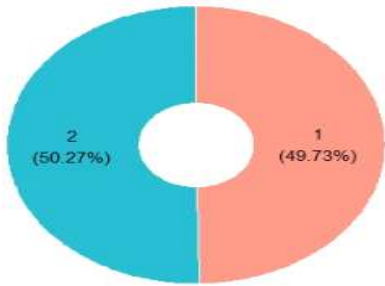
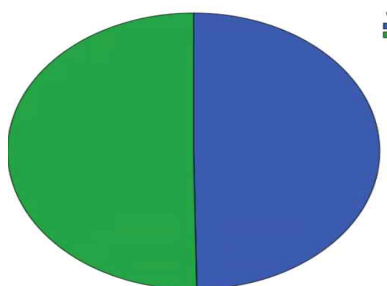
〈표 3〉 데이터 로드 (R과 SPSS 비교)

구분	R	SPSS
코드 및 조작	<pre>### 데이터 로드 if (!require(readxl)) install.packages("readxl") library(readxl) url <- "https://github.com/HakJun-Song/class/blob/master/incheon0427.xlsx?raw=true" destfile <- "incheon0427.xlsx" curl::curl_download(url, destfile) incheon0427 <- read_excel(destfile)</pre>	<pre>1. https://github.com/HakJun-Song/class/blob/master/incheon0427.xlsx 에서 incheon0427.xlsx 파일을 다운 받는다 2. SPSS 실행 3. [파일]-[열기]-[데이터] 4. 데이터 열기 창이 열리면 창 하단의 파일 유형에서 Excel을 선택 5. Excel 데이터 소스 열기에서 '데이터 첫 행에서 변수 이름 읽어오기' 선택</pre>

R은 이러한 각 측정에 대해 간단한 기능을 제공한다. SPSS에서 이러한 통계는 분석 → 기술통계 → 기술 또는 빈도를 선택하여 생성할 수 있다(<표 4>~<표 6> 참조). 기술통계 분석 결과 R과 SPSS에서 유사한 결과값이 도출되었다.

수치 설명 통계 외에도 R 및 SPSS는 데이터의 표 및 그래픽 표현을 만드는 간단한 방법을 제공한다. 이러한 표와 그래픽을 통해 연구자들은 데이터 패턴(예 : 분포)을 시각화할 수 있다. 빈도분석에서 R과 SPSS에서 유사한 값이 도출되었다.

〈표 6〉 빈도분석 그래프 (R과 SPSS 비교)

구분	R	SPSS
코드 및 조작	<pre>### 빈도분석 if (!require(webr)) { install.packages("webr") library(webr)} PieDonut(incheon0427,aes(pies=q1),showPieName=FALSE)</pre>	<ol style="list-style-type: none"> 1. [그래프]-[차트작성기]-[갤러리]-[원형차트] 2. 갤러리 차트를 끌어서 시작점으로 사용 3. 조각기준 q1 선택 후 확인버튼 클릭
분석 결과		

파이 차트, 막대 차트, 꺾은 선형 차트, 산점도, 상자 그림 및 히스토그램이 가장 일반적으로 사용되는 그래프이다. SPSS의 그래프 메뉴에서 모두 만들 수 있으며 R은 각각에 대해 단일 명령을 사용하여 이러한 그래프를 생성할 수 있다.

4) 기본 추리통계

표본을 통해 모집단을 추리하는 추리통계 기법은 크게 연관성 분석(카이제곱 검정, 상관분석 등), 차이분석(t-test, 분산분석(ANOVA) 등), 인과성 분석(회귀분석 등)으로 나누어 볼 수 있다. 본 연구에서는 연관성 분석으로서 상관분석을 차이분석으로서 독립표본 t-test를 인과성 분석으로써 회귀분석에 대해 각각 R과 SPSS로서 분석을 하여 이를 비교 분석해 보고자 한다.

(1) 상관분석

상관분석은 양적변수 간 연관성을 알아보는 추리통계 기법이다. R에서는 Hmisc 패키지의 rcorr 함수를 통해 변수 간 상관관계와 이에 대한 유의성을 확인해 볼 수 있다. SPSS에서 이러한 통계는 분석 → 상관관계 → 이변량 상관계수를 선택하여 시행될 수 있다. 상관분석 결과 R과 SPSS에서 유사한 결과값이 도출되었다.

〈표 7〉 상관분석 (R과 SPSS 비교)

구분

R

SPSS

상관분석

```
if (!require(Hmisc))
  install.packages("Hmisc")
library(Hmisc)
res2 <- rcorr(as.matrix(target_vars))
print(res2$r, digits=3)
print(res2$p, digits=3)
```

코드
및
조작

1. [분석]-[상관분석]-[이변량 상관계수]
2. 긍정적 경제인식과 관련된 q10, q11, q12를
오른쪽 변수창에 넣고 확인버튼 클릭

```
> print(res2$r, digits=3)
      q10  q11  q12
q10 1.000 0.676 0.555
q11 0.676 1.000 0.600
q12 0.555 0.600 1.000
> print(res2$p, digits=3)
      q10 q11 q12
q10 NA   0   0
q11 0   NA   0
q12 0   0  NA
```

분석
결과
(R)

Correlations

		q10	q11	q12
q10	Pearson Correlation	1	.676**	.555**
	Sig. (2-tailed)		.000	.000
	N	563	563	563
q11	Pearson Correlation	.676**	1	.600**
	Sig. (2-tailed)	.000		.000
	N	563	563	563
q12	Pearson Correlation	.555**	.600**	1
	Sig. (2-tailed)	.000	.000	
	N	563	563	563

**. Correlation is significant at the 0.01 level (2-tailed).

(2) 독립표본 t-test

독립표본 t-test는 2개의 그룹(질적변수)에 따라 동일한 양적변수의 평균이 차이가 있는지를 알아보는 추리통계 기법이다. R에서는 lawstat 패키지의 levene.test 함수를 통해 분산의 동질성을 살펴보고 t.test 함수에 따라 t-test의 분석 결과를 알 수 있는데 이때 var.equal의 값을 “T”로 지정하면 분산의 동질성을 만족하는 t.test 분석값을 그렇지 않으면 분산의 이질성에 대한 t.test 분석값을 얻을 수 있다. SPSS에서 이러한 통계는 분석 → 평균 비교 → 독립표본 T 검정을 선택하여 시행될 수 있다. 분석 결과 등분산검정과 평균 비교 모두에서 R과 SPSS에서 유사한 결과값이 도출되었다.

〈표 8〉 독립표본 t-test (R과 SPSS 비교)

구분	R	SPSS
코드 및 조작	<pre>### t-test if (!require(lawstat)) {install.packages("lawstat")} library(lawstat) levene.test(incheon0427\$q10, incheon0427\$q1, location = "mean", correction.method = "zero.correction") t.test(incheon0427\$q10 ~ incheon0427\$q1, var.equal=T) t.test(incheon0427\$q10 ~ incheon0427\$q1, var.equal=F)</pre>	<ol style="list-style-type: none"> 1. [분석]-[평균비교]-[독립표본 T검정] 2. 집단변수에는 q1(성별), 검정 변수에는, q10(긍정적 경제인식 1)을 오른쪽 변수창에 넣는다 3. 집단변수의 물음표를 해결하기 위해 집단정의를 눌러서 숫자 1과 2를 각각 집단1과 집단2에 넣어주고 확인버튼 클릭

<p>분석 결과 (R)</p>	<pre>> levene.test(incheon0427\$q10, incheon0427\$q1, + location = "mean", correction.method = "zero.correction")</pre> <p>Classical Levene's test based on the absolute deviations from the mean (zero.correction not applied because the location is not set to median)</p> <p>data: incheon0427\$q10 Test Statistic = 5.4004, p-value = 0.0204</p>	<pre>> t.test(incheon0427\$q10 ~ incheon0427\$q1, var.equal=F)</pre> <p>Welch Two Sample t-test</p> <p>data: incheon0427\$q10 by incheon0427\$q1 t = -0.58499, df = 550.96, p-value = 0.5588 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -0.17345399 0.09384773 sample estimates: mean in group 1 mean in group 2 3.578571 3.618375</p>
--------------------------	--	---

Independent Samples Test										
		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
q10	Equal variances assumed	5.408	.020	-.585	561	.559	-.040	.068	-.173	.094
	Equal variances not assumed			-.585	550.961	.559	-.040	.068	-.173	.094

(3) 회귀분석

SPSS에서 Analyze → Regression → Linear를 클릭하면 Linear Regression 창이 나타난다. R에서는 lm 함수를 사용하여 선형회귀 분석을 수행할 수 있다. 아래 표에서 종속변수는 q48이고 독립변수는 q10과 q23이다. 출력에는 계수, R-제곱, 잔차 및 잔차 표준오차, t 및 F 통계 등이 포함된다. 분석 결과 R과 SPSS에서 유사한 결과값이 도출되었다.

〈표 9〉 회귀분석 (R과 SPSS 비교)

구분

R

SPSS

```

if (require(stargazer))
{install.packages("stargazer")
  library(stargazer)}
코드 및 요약 mod<-lm(q48~q10+q23, data=incheon0427)
summary(mod)
stargazer(mod, type = "text", report =
"vct*", star.cutoffs = c(0.05, 0.01, 0.001),
          column.labels = c("coef"),
          single.row = TRUE)

```

1. [분석]-[회귀분석]-[선행]
2. 독립변수에는 q10(긍정적 경제인식1), q23(긍정적 사회인식1)을 종속변수에는 q48(인지1)을 넣고 확인버튼 클릭

분석
결과

```

=====
Dependent variable:
-----
               q48
               coef
-----
q10              0.343
               t = 8.849***
q23              0.228
               t = 5.782***
Constant         1.095
               t = 7.323***
-----
Observations      563
R2                0.244
Adjusted R2       0.242
Residual Std. Error 0.680 (df = 560)
F Statistic      90.487*** (df = 2; 560)
=====
Note:             *p<0.05; **p<0.01; ***p<0.001

```

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.494 ^a	.244	.242	.680

a. Predictors: (Constant), q23, q10

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	83.669	2	41.834	90.487	.000 ^a
	Residual	258.900	560	.462		
	Total	342.568	562			

a. Predictors: (Constant), q23, q10

b. Dependent Variable: q48

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.095	.150		7.323	.000
	q10	.343	.039	.354	8.849	.000
	q23	.228	.039	.231	5.782	.000

a. Dependent Variable: q48

한편 R에서는 stargazer 함수를 사용하여 회귀분석 결과를 정리된 표의 형태로 표현하는 것이 가능하다. 이는 회귀분석 결과를 비교적 쉬운 방법으로 시각화하는데 기여함으로써 통계분석 초보자가 회귀분석의 원리를 이해하고 분석 결과를 그림으로 표현하는 데 있어 좋은 도구가 될 수 있다.

〈표 10〉 회귀분석 그래프 그리기 (R과 SPSS 비교)

구분	R	회귀분석 결과 그래프 그리기 결과
코드 및 조작	<pre> ### 회귀분석 결과 그래프 그리기 if (!require(DiagrammeR)) {install.packages("DiagrammeR")} library(DiagrammeR) grViz(" digraph research_model { graph [layout = neato, overlap = true, outputorder = edgesfirst] A [pos='-2, 0', label='긍정적\\n경제효과', shape=square, fixedsize = true, width = 1.5, height=0.2] C [pos='-2, 4', label='긍정적\\n사회효과', shape=square, fixedsize = true, width = 1.5, height=0.2] D [pos='2, 2', label='R개발 인식\\nR-square = 0.244', shape=square, fixedsize = true, width = 2.0, height=0.2] A->D [headlabel = '0.343(8.849)@{***}', labeldistance=12, labelangle=22, headport = 'w', tailport = 'e'] C->D [headlabel = '0.228(5.782)@{***}', labeldistance=14, labelangle=-13, headport = 'w', tailport = 'e'] } ") </pre>	

IV. 결 론

빅데이터 시대에서 데이터를 통해 의사결정을 하고 전략을 만드는 과학적 의사결정과 정책 시행의 중요성은 지속해서 높아질 것인데 여기에는 관광분야도 포함된다. 이러한 상황에서 SPSS는 그동안 유용한 통계분석 도구로서 인식됐는데 최근 들어 SPSS의 대체재로서 R에 관한 관심은 크게 높아지고 있다. R은 분석처리 속도가 느리고 처음 배울 때 어려우며 UI 측면에서 친근감이 떨어지는 단점이 있다. 이러한 단점에도 불구하고 R은 무료 오픈소스 프로그램으로서 패키지의 확장성, 최신기법 도입, 데이터 시각화, 재현 가능성 연구 측면에서 우수하여 관광분야 데이터 분석에서 유용하게 사용될 수 있다. 수학적 개념이나 코드 활용에 비교적 익숙하지 않은 관광분야의 학생과 연구자들은 대부분 C나 java 등 다른 언어를 배운 적이 많지 않고 SPSS에 익숙해져 있어 R로 데이터 분석을 대체하는 데 있어 많은 어려움을 겪고 포기할 가능성도 크다고 생각된다. 하지만 이러한 어려움이 있다고 해서

계속해서 SPSS만을 고집하는 것은 다른 사회과학 분야와 비교해 볼 때 관광분야 데이터 분석의 발전 입장에서 바람직하다고만은 얘기할 수 없을 것이다. SPSS는 연구자 의도에 따라 자유롭게 데이터를 정리, 통합, 확인하기가 쉽지 않고 시계열 분석, 베이지안 분석, 텍스트 마이닝 등의 최신 분석기법 등을 충분히 활용하는 데 있어 한계점이 있다. SPSS는 유료 제품이므로 이를 지속해서 이용하는 데는 금전적 상황도 고려해야 하는 단점도 있다. 이러한 상황에서 본 연구는 R의 장단점, 간단한 사용소개, R 및 SPSS의 기본 통계분석 비교 등을 수행하면서 R을 통해 관광분야 데이터 분석이 이루어질 수 있는 토대를 마련하고자 하였다. 구체적으로 기초 통계(빈도분석, 기술통계, t-test분석, 상관분석, 회귀분석) 분석을 R과 SPSS로 병행하여 실행하면서 도출된 결과값이 동일함을 확인하였다. 관광분야에서 R이라는 통계분석 오픈소스 프로그램은 외부에서 전하여 들어온 ‘전래’는 이루어졌지만 이를 전하여 널리 퍼뜨리는 ‘전파’의 단계에는 아직 이루지 못한 상태라고 평가될 수 있다. 하지만 다른 사회과학 분야에서는 관광분야보다 R의 사용이 더 일반화되고 있으므로 관광분야에서도 R은 SPSS를 충분히 대체하면서 활발하게 사용되게 될 것으로 예상된다. 그동안 R은 상업적 통계 프로그램 대체를 위한 저렴한 대안 정도로만 여겨져 왔다. 하지만 그동안 R은 충분히 발전하여 이와 같은 인식을 능가했고 이제는 기능, 유연성 및 다른 응용 프로그램과의 통합성 측면에서 충분히 SPSS와 같은 상업적 통계 프로그램을 능가할 수 있는 잠재성을 갖추고 있다. R이 상업적 경쟁자보다 사용하기 어렵다는 한계점은 R Studio, R markdown 등의 사용으로 충분히 보완될 수 있다. 이제 R은 관광분야를 포함하는 많은 학문 분야에서 통계를 더 광범위하게 다루고 확장성을 가져가기 위해서 좋은 선택 중 하나가 될 수 있다. 한편 R의 장점과 발전 가능성을 고려하더라도 데이터 분석에서 어떤 통계 프로그램을 사용해야 하는지를 결정하는 것은 서로의 강점, 약점 및 처리 측면에서 서로 아주 다르고 변경에 따른 시간과 비용도 적잖게 소모되므로 어떤 프로그램이 적합한지에 대한 결정은 신중하게 이루어져야 할 것이다. 본 연구가 가지고 있는 한계점 및 향후 연구 방향은 다음과 같다. 첫째, 본 연구는 코로나19 발병 이전에 수집한 데이터를 분석에 적용하였는데 추후 코로나19 발병 시부터 그 이후의 관광산업 데이터 수집을 통한 연구가 필요할 것으로 판단된다. 둘째, R은 많은 장점을 갖고 있지만, 그에 반하여 언어기반으로 코딩을 해야 하고, R에 익숙하지 못한 사람이 코딩하다 보면 에러도 많이 발생한다는 단점을 갖은 프로그램이기도 하다. 또한, 대다수 연구자가 아직도 SPSS를 이용하고 있으므로 기존의 연구자들이 SPSS와 R이 각각 갖는 장단점을 보완하면서 관광분야 연구에 어떻게 활용할 것인지에 관한 아이디어를 도출해보는 연구도 이루어질 필요가 있을 것이다.

참고문헌

- 김주향 · 송학준 (2017). HMR(Home Meal Replacement) 판매업체의 서비스품질에 따른 지각된 가치가 고객만족도에 미치는 영향에 관한 연구. 『관광레저연구』, 29(1), 317-333.
- 박경열 · 이현주 (2019). 관광정책 결정지원을 위한 의사결정나무분석의 적용-관광개발정책 우선순위에의 실증 분석. 『관광학연구』, 43(5), 11-28.
- 송학준 · 박혜미 (2020). 소화행 여행에 대한 탐색적 연구: 육하원칙(5W1H)에 따른 개념정리를 중심으로. 『관광레저연구』, 32(6), 449-468.
- 신영송 · 김효은 · 송학준 (2018). 소셜미디어 외식정보 탐색 행동의도에 관한 연구: 확장된 목표지향적 행동모형을 중심으로. 『관광레저연구』, 30(12), 285-306.
- 이승덕 · 최윤정 · 김갑성 (1998). 통계분석용 package software 비교 고찰-SAS와 SPSS를 중심으로. 『대한한의정보학회지』, 4(1), 17-34.
- 이윤환 (2016). 『제대로 알고 쓰는 R 통계분석』. 서울: 한빛아카데미(주).
- 조미순 · 김순귀 (2003). 생존분석을 위한 통계패키지의 비교 연구-SAS, SPSS, STATA. 『2003 한국통계학회 학술대회』, 335-340.
- 정철용 · 유진호 · 김민정 · 김진 (2021). 『SPSS와 Python, R 그리고 Excel을 활용한 데이터 분석 따라하기』. 서울: 북넷.
- Fox, J. & R. Andersen (2005). *Using the R statistical computing environment to teach social statistics courses*. Unpublished paper. Retrieved from <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.308.13&rep=rep1&type=pdf>.
- Kleinman, K. & N. J. Horton (2009). *SAS and R: Data management, statistical analysis, and graphics*. London: Chapman and Hall/CRC.
- Muenchen, R. A. (2011). *R for SAS and SPSS users*. Berlin: Springer Science & Business Media.
- Tang, H. & P. Ji (2014). *Using the statistical program R instead of SPSS to analyze data*. American Chemical Society. Retrieved from <https://pubs.acs.org/doi/abs/10.1021/bk-2014-1166.ch008>.
- Verma, J. P. (2012). *Data analysis in management with SPSS software*. Berlin: Springer Science & Business Media.
- Zuur, A., E. N. Ieno & E. Meesters (2009). *A Beginner's Guide to R*. Berlin: Springer Science & Business Media.

2021년 10월 06일 원고 접수

2021년 11월 02일 수정본 접수

2021년 11월 10일 최종 수정본 접수 및 게재 확정

3인 익명심사 필