Journal of the Korean Data & Information Science Society
2013, **24**(6), 1253–1262

http://dx.doi.org/10.7465/jkdi.2013.24.6.1253
한국데이터정보과학회지

"통계학은 어렵다"? - 통계학교재의 문제점[†]

이원우[1]

[1]세종대학교 수학통계학부

접수 2013년 8월 15일, 수정 2013년 9월 21일, 게재확정 2013년 10월 1일

요 약

본 연구에서는 "통계학은 어렵다"라는 인식이 어느 정도인지를 알아보고 어렵게 느끼고 있는 이유를 찾는다. 이를 위하여 리서치회사에서 연구원으로 일하고 있는, 통계학을 배웠던 경험이 있는 사람들을 대상으로 설문조사하였다. 그 결과, 응답자 중 80.8%에 이르는 사람들이 "통계학은 어렵다"라고 인식하는 것으로 파악되었다. 그리고 어려운 이유가 대학에서 배운 통계학 교재가 어렵게 쓰여 졌기 때문 (62.5%)이라는 것이다. 따라서 시중에 유통되는 통계학 교재들을 분석한 결과, 대부분의 교재들이 확률변수를 나타내는 대문자 X와 확률변수들이 취하는 값을 나타내는 소문자 x를 구분하지 않고 있다는 점을 밝혀내고, 이런 구분이 왜 통계학교육에 있어 간과되어서는 안 되는가를 지적하였다. 이 연구 결과가 향후 교재를 집필하거나 교육하는데 참고가 되어, 사회적으로 "통계학이 어렵기만 한 분야는 아니다"라는 인식이 확산되기를 바란다. 통계학을 폭넓게 보급하는 것은 통계학자들의 책임이기 때문이다. 이를 위하여 통계학교육에 대한 연구도 활발해지기를 기대한다.

주요용어: 대문자, 소문자, 통계학교육, 확률변수.

1. 서론

통계학은 대학교육 과정 중 많은 전공분야에서 필요로 하는 기초 학문이다. 그러나 통계학을 배웠던 많은 사람들이 "통계학은 어렵다"거나 "무엇을 배웠는지 모르겠다"라는 반응을 보인다. 강조하건데 대학에서 통계학을 전공한 사람들조차 통계학을 제대로 이해하지 못하고 졸업했다고 토로하는 경우도 많다. 본 연구자의 경험에 의하면

통계학을 전공하는 석·박사 과정 학생들조차 통계학 기초이론을 정확하게 이해하지 못한 채 학습의 반복 효과에 따라 고급 통계학 과목들을 배우고 있음이 현실이다. 예컨대 석·박사 과정의 학생들이 확률표본, $X_1, X_2, \cdots, X_n \sim i.i.d.\ f_X(x)$의 의미를 제대로 충분히 이해하고 있지 못하고, 반복적으로 공부해 왔던 것이기 때문에 그저 알고 있는 것이라면 기초이론 교육에 문제가 있었다는 방증이라고 할 것이다.

비통계학 전공분야에서의 통계학 기초이론 교육 문제는 당연히 더욱 심각하다고 할 것이다. 매우 많은 비통계학 전공분야에서 - 교과과정 (curriculum) 편성에 어려움이 있는 것도 불가피하겠지만 - 통계학의 기초이론을 다루지 않고 (통계적)조사방법론을 가르치는 경우, 통계적으로 처리한 결과를 잘못 해석하거나 잘못 사용하는 경우가 많은 것 또한 현실이다. 실제로 비통계학 전공분야의 대학원생들이 학위논문에서 통계적 실증분석을 하는 경우가 적지 않은데, 통계학이론에 대한 이해 부족으로 통계분석 결과에 대해 오류를 범한 논문들이 많다. 게다가 해당 전공분야 선행 연구자들의 오류를 그대로 답습하여 논문을 작성하는, 오류의 악순환이 반복되기도 하는 실정이다.

[†] 이 논문은 세종대학교 2011년 9월~2012년 8월의 연구년 실적 논문임.
[1] (143-747) 서울시 광진구 군자동 98, 세종대학교 수학통계학부, 교수. E-mail: wwlee@sejong.ac.kr

우리나라에서 출판된 통계학 기초이론을 다루고 있는 대학 교재는 약 200여 권이며 그 교재들의 저자는 통계학전공 교수, 경영학/경제학전공 교수가 거의 반반의 비율을 차지한다. 그리고 약간의 기타전공 교수들의 저서가 있다. 그러나 상당수 통계학 기초이론 교재들이 이론을 설명하는 데에 부족하거나, 정확하게 이해해야 하는 부분들에 대해 표현을 잘못함으로써, 배우는 이들이 통계학 이론을 체계적으로 이해하지 못하여 "통계학은 어렵다"고 인식하는 것은 아닌가에 대한 관점을 살펴보는 것이 본 연구의 목적이다.

본 연구에서는 대학에서 통계학을 배운 경험이 있는 직장인들을 대상으로 통계학은 얼마나 어렵다고 느끼고 있는지, 그리고 그 이유는 무엇인가를 밝혀 본다. 그 이유가 대학에서 가르치는 교수 (강사)들의 설명 부족인지, 대학 통계학교재의 어려움 때문인지를 밝혀보는 것은 흥미로운 일일 것이다. 그리고 교재가 어렵다는 것은 다른 이유도 있겠지만, 교재가 통계학 이론을 체계적으로 명확하게 설명하지 못하고 있는 것은 아닌지, 그래서 혼란 (어려움)을 가중시키는 것은 아닌지 살펴보고자 한다.

본 연구가 향후 통계학이 쉽게 폭넓게 보급될 수 있는 방향 제시의 기초가 되기를 바란다.

2. 통계학 기초이론 전개의 핵심 포인트

통계학은 기본적으로 모집단으로부터 얻어진 자료 (표본)를 분석함으로써 모집단에 대한 특성을 파악해보자는 것이다. 그러나 모집단 전체의 자료들을 알 수 없기 때문에 모집단에 대해서는 이론적으로 접근하고 설명할 수밖에 없다. 여기서, 모집단을 설명하는 필요한 수단이 확률변수이며 일반적으로 확률변수를 X(대문자)로 표현한다. 그리고 X는 분포를 하는데 이 분포를 확률분포라고 하는 것이다. 여기서, 확률변수를 대문자 X로 표현하는 것은 매우 중요하다. 왜냐하면 소문자 x는 대문자 X와 별도로 사용되어야만 하기 때문이다. 즉, 확률변수 X의 실제로 얻어진 - 확률변수 X가 갖게 되는 - 구체적인 값을 소문자 x로 나타내기 때문에 통계학 이론에서 소문자와 대문자의 구분은 매우 중요한 것이다. 다시 말하면, 통계학 이론을 처음 배우는 사람들에게 대문자와 소문자를 명확하게 구분할 수 있도록 가르치는 것은 일관성 있는 기초이론의 정리와 이해에 절대적으로 필요한 요소이다. 그러나 이 부분에 대한 중요성에 비중을 두지 않는 통계학자들도 있음이 현실이고 통계학 교재들 내용에 구분하여 설명되어 있지 않은 경우가 상당하다. 이는 저자의 취향에 따라 아무렇게나 할 수 있는 것이 아니다. 통계학 교재에서 이 점을 분명하게 하지 않으면 저자 자신조차 일관된 이론 전개를 해 나가기 어렵다. 예컨대 통계학 기초이론 교재에서 기댓값을 $E(X)$로 표현하지 않고 $E\,(x)$로 나타내는 것은 매우 잘못된 것이다. 이렇듯 X와 x의 구분을 제대로 하지 않은 경우, 체계적인 이론 전개, 설명이 되기 어렵고 통계학이론이 뒤죽박죽된다.

더욱이, 모집단으로부터 얻어진 표본이 두 가지로 표현될 수 있다는 것이 설명되어야 한다. 다시 말하면, 모집단으로부터 얻어지는 표본은

$$\{x_1, x_2, \cdots, x_n\} \text{과 } \{X_1, X_2, \cdots, X_n\}$$

으로 표현될 수 있다는 것을 반드시 이해시키는 것이 매우 중요하다. 그 이유는 전자는 통계학을 배우기 전에라도 누구나 알고 있는 표본이라는 (구체적으로 얻어진 값들) 것에 대한 표현이기 때문이다. 그러므로 통계학 교재에서는 후자, 즉, 확률변수들로 표현된 표본으로서 $X_i, i = 1, 2, \cdots, n$들이 각각 모집단과 같은 분포를 가지며 서로 독립적으로 분포한다는 것을 나타내는 (이론적)표본이라는 것을 설명하고 있어야 한다. 통계학 교재에 이러한 구분이 명확하게 설명되어 있지 않거나, 기초이론 교육 내용에 포함되지 않을 경우, 표본 $\{X_1, X_2, \cdots, X_n\}$으로부터 얻어지는 표본평균, $\overline{X}$의 분포를 설명하는 중심극한정리를 제대로 이해할 수 없다 ($\overline{x}$는 실제로 얻어지게 되는 어떤 값이다). 더욱이 중심극한정리를 충분히 이해하지 못하는 경우, 추정량의 분포, 검증통계량의 분포 등에 대한 이해가 어려워지며 결국은

통계학 이론을 제대로 소화하지 못하게 되는 것이다. 예를 들면, 몇몇 저자들의 교재에서 중심극한정리를 설명하는 내용에 기댓값을 $E(\overline{X})$가 아니고 $E(\overline{x})$라고 표현하여 설명하고 있는 것은 잘못된 (틀린) 것이다. 즉, 중심극한정리를

$$\overline{x} \sim N(\mu, \sigma^2/n)$$

으로 소개 한다면, 이후의 이론 전개는 매우 혼란스러워질 수밖에 없다. 이러한 기본적이고 단순한 문제의 이해 부족으로 인하여 통계학 교재의 이론 전개가 체계적이지 못하다면, 배우는 이들이 더욱 혼란을 일으키게 되고 "통계학은 어렵다"고 하게 되는 것 아니겠는가?

또한 회귀분석 모형을

$$\begin{aligned} & y_i = \beta_0 + \beta_1 X_i + \epsilon_i,\ i = 1, 2, \cdots, n \\ & (\text{가정})\ \epsilon_i \sim N(0, \sigma^2) \end{aligned} \tag{2.1}$$

으로 표현하고 있는 것은 부적합하며

$$\begin{aligned} & Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1, 2, \cdots, n \\ & (\text{가정})\ \epsilon_i \sim N(0, \sigma^2) \end{aligned} \tag{2.2}$$

로 나타내야 한다. 식 (2.1)과 식 (2.2)에서 ϵ_i가 정규분포한다는 가정을 하기 때문에 식 (2.1)은 엄밀하게 말하면 틀린 표현이다. 즉, 식 (2.1)의 모형에서 y_i로 표현하고 있는 것은 Y_i로 표현해야 하며, 설명변수에 대한 부분도 X_i가 아니라 x_i로 표현하는 것이 적절하다. 물론 X_i가 주어진 값이라고 별도로 설명을 하더라도 혼란이 따르기 마련이다. 엄밀하게 말하면, 회귀모형을 식 (2.2)로 제시하고 설명해야 회귀분석에서 반드시 다루어야만 하는 $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$에 대한 가설검증이 설명될 수 있다. 즉, 이 가설에 대한 검증통계량으로 $\widehat{B}_1$(확률변수)를

$$\widehat{B}_1 = \frac{\Sigma(x_i - \overline{x})(Y_i - \overline{Y})}{\Sigma(x_i - \overline{x})^2} \sim N\left(\beta_1, \frac{\sigma^2}{\Sigma(x_i - \overline{x})^2}\right) \tag{2.3}$$

으로 설명할 수 있고, 가설검증의 절차를 이해시킬 수 있는 것이다. 그러나 이 부분에 대해서 거의 모든 통계학 교재가 올바로 설명하지 못하고 있는(미흡한) 것이 현실이다. 즉, 자료들로써 계산되어 얻어진

$$\widehat{\beta}_1 = \frac{\Sigma(x_i - \overline{x})(y_i - \overline{y})}{\Sigma(x_i - \overline{x})^2} \tag{2.4}$$

만을 제시하고 설명하면 검증통계량의 분포를 토대로 설명되는 가설검증 절차에 대해 이해하기 어려운 것이다. 왜냐하면 식 (2.4)의 $\widehat{\beta}_1$표현은 얻어진 표본값들로 계산된 기울기의 값을 나타내는 것이기에 이를 확률변수라고 설명할 수 없다.

추가로, 회귀분석에서도 다루게 되는 분산분석표를 이해하기 위해 분산분석 모형도 통계학 교재에서 설명하고 있어야 한다. 통계학 기법 중 가장 기본적이면서도 중요한 분산분석은 통계학 기초이론에서 대부분의 사람들이 이해하기 어려울 정도의 난이도를 가지고 있다. 그렇기에 분산분석 모형을 제대로 설명하지 않은 교재들이 상당수 있다고 믿어진다. 분산분석을 제대로 이해시키려면 비교하고자하는 모집단들 (k개)에 대한 기본모형 (1인자 분산분석)으로

$$\begin{aligned} Y_{ij} =& \mu_{ij} + \epsilon_{ij} \\ =& \mu + \alpha_i + \epsilon_{ij},\ i = 1, 2, \cdots, k,\ j = 1, 2, \cdots, n_i \\ & (\text{가정})\ \epsilon_{ij} \sim N(0, \sigma^2) \end{aligned} \tag{2.5}$$

을 표현하고 설명하여야만 분산분석에서의 가설은 F검증통계량으로 검증된다는 것을 설명할 수 있다. 아무리 난이도가 있다고 하더라도 F-분포에 대한 언급 없이 분산분석을 설명할 수 없다. 즉, 제곱합의 관계에 대해서

$$\Sigma\Sigma(Y_{ij} - \overline{\overline{Y}})^2 = \Sigma\Sigma(\overline{Y_i} - \overline{\overline{Y}})^2 + \Sigma\Sigma(Y_{ij} - \overline{Y}_i)^2 \text{와} \quad (2.6)$$
$$\Sigma\Sigma(y_{ij} - \overline{\overline{y}})^2 = \Sigma\Sigma(\overline{y_i} - \overline{\overline{y}})^2 + \Sigma\Sigma(y_{ij} - \overline{y}_i)^2$$

의 차이를 설명해야 한다. 확률변수들로 표현된 식과 (전자) 자료로부터 계산되어 얻어지는 값에 (후자) 대해 구분할 수 있어야 가설검증의 이론적배경이 되는 F- 분포를 형식적으로라도 소개할 수 있을 것이다. 다시 말하면, F-분포에 대해서는 식 (2.6)의 전자를 가지고 이론적으로 설명하고, 후자는 실제 자료를 얻었을 때 전자에 대해 계산된 값들로서 분산분석표가 작성된다는 점을 설명해야 할 것이다.

3. 직장인들의 통계학에 대한 이해

대학에서 통계학을 배운 직장인들이 통계학을 어떻게 인식하고 있는가를 설문조사를 통하여 알아본다. 특히, 통계학을 어렵다고 인식하는 중요한 이유가 대학교재 내용의 문제인지, 교·강사의 설명이 문제인지를 확인해 보고자한다. 이 연구에서는 직장인을 통계학의 활용도가 높은 리서치회사 연구원으로 한정하고 우리나라에서 일하고 있는 전체 1,000여명 (2012년도 한국조사협회 등록자 기준)의 연구원들 중에서 통계학을 배운 경험이 있는 125명의 설문응답 결과를 분석하였다 (통계학전공 27명, 사회계열 전공 87명, 기타전공 11명). 주요 조사내용은 (1)통계학 과목에서 배운 내용은? (이론, 방법론만, 이론과 방법론) (2)통계학 및 관련과목을 몇 학기나 배웠는가? (3)강의는 누가 (교수 또는 강사) 했는가? (4)통계학을 배울 때 어느 정도 어려웠다고 생각하는가? 그리고 그 이유는 무엇이었는지? (5)통계학을 어느 정도 이해했다고 생각하는가? (6)통계학의 재수강 기회에 대한 반응 등이다. 여기서는 문항 (4)에 대한 결과를 중점적으로 다룬다.

Table 3.1에서 통계학이 어느 정도 어려웠느냐는 질문에 대해서는 '어려웠다' 70.4%, '매우 어려웠다' 10.4%로 전체의 80.8%가 어려웠다고 답을 하고 있다. '전혀 어렵지 않았다'는 응답은 전무하다. 통계학 전공자들 만에 대해서는 '어려웠다' 74.1%, '매우 어려웠다' 11.1%로 85.2%가 어려웠다고 응답하고 있다. 물론 통계학 전공자들은 수리통계학이나 통계학 각론에서의 어려웠던 기억이 응답에 영향을 준 것일 수도 있지만, 이들 중 92.6%가 통계학 재교육의 기회가 있다면 '다시 한 번 배워보고 싶다'에 응답하고 있음을 볼 때 (Table 3.3), 전공에 따른 차이는 없다고 보는 것이 타당하다 (카이제곱 0.329 [유의확률 0.848] - Cramer 규칙 적용).

Table 3.1 Responses to "How much difficult was Statistics?"

Degree of difficulty	Not at all	Not difficult	So-so	Difficult	Very difficult
Stat major (27)	0.0%	11.1%	3.7%	74.1%	11.1%
Social sci major (87)	0.0%	1.1%	18.4%	69.0%	11.5%
Total (125)	0.0%	4.0%	15.2%	70.4%	10.4%

Table 3.2는 통계학이 어려웠던 이유에 대한 결과이다. '교재가 어려웠다'가 62.5%로 압도적으로 많고, '교수의 설명이 어려웠다'가 24.0%, '본인이 공부를 하지 않았다'가 8.7%를 차지한다. 통계학 전공자들도 '교재의 어려움'이 63.2%, '교수설명이 어려웠음'이 26.3%, '공부를 안 했음'이 10.5%를 나타내고 있어 사회계열 전공자들의 분포와 크게 다르지 않다 (카이제곱 0.099 [유의확률 0.953] - Cramer 규칙 적용). 여기서 교재가 어려우면 교수의 설명도 어려울 수밖에 없다는 점을 감안하면 통계학이 어려웠던 주 이유는 '교재가 어려웠다'이라고 할 수 있다.

Table 3.2 Responses to "Reason of difficulty"

Difficulty reason	Lecturer's explanation	Lecturer's ability	Non-stat major lec.	Textbook	Did not study
Stat major (27)	26.3%	0.0%	0.0%	63.2%	10.5%
Social sci major (87)	23.7%	1.3%	5.3%	61.8%	7.9%
Total (125)	24.0%	1.0%	3.8%	62.5%	8.7%

이 조사 결과로 대학 과정에서 통계학을 배운 대부분의 사람들이 통계학은 어렵다고 인식하고 있다는 것이 밝혀졌고, 그 주요 이유는 통계학 교재가 어려웠기 때문이라는 것으로 정리된다. 요약하면, 이해하기 어렵게 쓰인, 문제가 있는 통계학 교재를 가지고 강의하는 교·강사들의 설명이 "통계학은 어렵다"는 인식을 갖게 한다는 것이다. 그러므로 통계학 교재에 어떤 문제가 있는가하는 문제를 살펴볼 필요성이 대두된다.

Table 3.3 Response to necessity of re-education of Statistics

Re-education	Not necessary	Necessary
Stat major (27)	7.4%	92.6%
Social sci major (87)	5.7%	94.3%
Total(125)	8.0%	92.0%

참고로 Table 3.3은 "통계학 과목을 다시 수강할 수 있는 기회가 있다면 어떻게 하시겠습니까?"에 대한 응답이다. 전체 응답자들의 92.0%가 '다시 배우고 싶다'에 응답하고 있다. 물론 리서치회사에 근무하는 응답자들로서는 통계학의 필요성을 더 절실하게 느낄 수 있겠지만 통계학 전공자들의 92.6%가 재교육의 기회를 원한다는 것은 매우 놀라운 결과가 아닐 수 없다.

4. 통계학 교재 분석

통계학 기초이론을 이해하기 위한 대학 교재는 주로 대학 1학년 또는 2학년생들을 대상으로 한다. 이 과정의 대학 교재 저서명은 주로 '통계학', '일반통계학', '기초통계학' 등이다. 물론 '경제경영 통계학', '전산통계학', '통계패키지 (EXCEL, SAS, SPSS 등)를 이용하는 통계학' 등의 타이틀을 가지고 있는 통계학 교재들도 있다. 현재 우리나라에서 대학 교재로 활용되는 통계학 이론을 포함하고 있는 교재의 수가 약 200여 권인 데, 이 중26권에 대해 그 내용을 조사하였다 (참고문헌 참조). 저자들의 전공에 따른 차이가 있는가를 살펴보기 위해 통계학을 전공한 저자들의 교재를 13권, 경영/경제를 전공한 교수들이 저자인 교재 13권으로 나누어 조사하였다. 여기서 '통계적 조사방법론'에 대한 교재들은 주로 이론 설명 없이 방법론만을 다루고 있기 때문에 조사대상에서 제외한다.

4.1. 조사 내용

통계학 교재들에 대해 조사한 항목은 (1) 확률변수와 확률분포 (2) 기댓값 (3) 표본 (4) 중심극한정리 (5) 신뢰구간 (6) 검증통계량 (7) 분산분석 (8) 회귀분석 등이다. 이 연구에서는 앞에 열거한 8가지 항목들에 대해 누락되어 있지는 않은지, 올바로 이해할 수 있도록 구성되어 있는지에 대해 다음과 같은 내용들을 조사하였다.

(1) [확률변수와 확률분포] 확률변수를 올바로 표현하고 설명하고 있는지를 조사한다. 모집단의 분포가 곧 확률변수의 분포라는 설명이 있어야 할 것이다. 그리고 확률변수가 확률분포를 한다는 것은 모집단의 자료들이 어떤 분포를 한다는 의미이기 때문에 반드시 알아야 하는 개념이다.

(2) [기댓값] 기댓값이라는 개념은 모집단의 평균을 나타내는 것으로서 통계학 이론의 핵심인 중심극한정리, 추정, 그리고 가설검증에서도 반드시 알아야 하는 개념이다. 기댓값의 설명과 표현이 올바로 되어 있어야 한다.

(3) [표본] 표본을 두 가지로 설명하는 것은 매우 중요하다. 누구나 상식적으로 알고 있는 표본은 $\{x_1, x_2, \cdots, x_n\}$이며 통계학 이론을 이해하기 위해서는 표본을 $\{X_1, X_2, \cdots, X_n\}$로 표현하여 설명하는 것이 적절하다.

(4) [중심극한정리 (표본평균의 표본분포)] 표본평균들이 분포한다는 것을 올바로 설명하고 있는지를 조사한다. 통계학 이론을 이해하기 위해서는 중심극한정리의 이해가 반드시 필요한 개념이기 때문에 이 정리에 대한 설명이 누락되거나 올바로 설명되어 있지 않으면 안 된다.

(5) [신뢰구간] 신뢰구간에 대한 설명과 그 표현이 올바른지를 조사한다.

(6) [검증통계량] 가설검증에 사용되는 검증통계량에 대한 분포를 설명하고 있는지와 그 표현의 적정성 여부를 알아본다.

(7) [분산분석] 분산분석 모형에 대한 올바른 설명과 가설검증의 절차가 이해될 수 있도록 검증통계량과 F-분포를 설명하고 있는지를 조사한다.

(8) [회귀분석] 회귀분석의 이론적 배경을 이해해야만 여러 가지 회귀분석을 올바로 수행할 수 있다. 특히, 추정량으로서 $\widehat{B}_1$의 표현과 그의 분포를 설명하고 있는지를 조사한다.

4.2. 대학통계학 교재 조사결과

다음 Table 4.1과 Table 4.2는 각각 통계학전공 저자인 교재들과 (1-13, 무순) 비통계학전공 (주로 경영학, 경제학) 저자의 교재들에 (14-26, 무순) 대한 평가를 적절 (***), 미흡 (**), 부적절 (*), 누락 (x)으로 표기한 것이다.

Table 4.1 Evaluations of textbooks authored by statisticians

	Random variable	E(X)	Sample	CLT	C.I.	Test stat	ANOVA	Reg.
1	***	***	***	***	***	***	***	**
2	***	***	**	***	**	**	*	**
3	***	***	***	***	***	***	*	**
4	***	***	**	***	***	***	***	**
5	***	**	*	**	**	**	*	*
6	***	***	**	***	***	***	***	**
7	***	***	**	***	**	**	**	**
8	***	***	**	***	***	***	**	**
9	***	***	**	***	**	**	***	**
10	***	***	*	***	**	**	**	**
11	**	*	**	**	**	**	*	*
12	**	**	**	***	**	**	**	**
13	***	***	***	***	***	***	***	***

Table 4.2 Evaluations of textbooks authored by non-statisticians

	Random variable	E(X)	Sample	CLT	C.I.	Test stat	ANOVA	Reg.
14	***	***	*	***	*	*	*	**
15	***	*	**	***	**	**	*	**
16	***	**	**	*	*	*	*	*
17	***	***	*	*	*	*	*	*
18	***	***	**	***	***	***	**	**
19	***	***	**	***	***	***	*	**
20	***	***	*	**	**	**	**	**
21	***	***	***	***	***	***	**	**
22	***	**	*	**	**	**	**	**
23	***	***	*	**	**	**	*	*
24	***	**	*	**	**	**	**	*
25	***	**	*	***	**	**	**	**
26	***	***	*	***	**	**	**	**

조사 대상인 교재들 중에서 확률변수를 소문자 x로 표현한 교재가 한 권도 없었음에도 불구하고, 기댓값을 $E(X) = \Sigma X_i P(X_i)$로 표현하거나 $E(x)$로 표현하는 것은 참으로 이해할 수 없다 (Cho와 Kim, 2005; Kang, 2006; Ryu 등, 2011; Park 등, 2013). 더욱이 중심극한정리에 대해서도 설명을 하고는 있지만 대문자, $\overline{X}$와 소문자, $\overline{x}$를 구분하지 않은 경우 (미흡 또는 부적절)도 8권 (30.7%)이 되었다. 게다가 $E(\overline{x}) = \mu$ 혹은 $E(\overline{X}) = E[(x_1 + x_2 + \cdots + x_n)/n] = \mu$로 표현하는 교재가 있다는 것은 참으로 놀라운 일이 아닐 수 없다 (Lee, 2004; Lee와 Lee, 2006; Moon, 2004; Shin과 Choi, 2006; Yum 등, 2003).

주목할 점은 거의 모든 교재에 있어 대체적으로 표본이론에서부터 문제가 되기 시작한다는 것이다. 그 문제란 표본을 두 가지로 표현할 수 있다는 점, 즉, 표본을 $\{X_1, X_2, \cdots, X_n\}$과 $\{x_1, x_2, \cdots, x_n\}$으로 구분하여 이해해야 한다는 점을 설명하고 있지 않다. Table 4.1과 Table 4.2에서 보는 바와 같이 표본에 대한 설명이 미흡 (대문자와 소문자로 구분하지 않은 경우)하거나 부적절한 경우 (표본이라는 개념을 이해하기 어렵게 설명한 경우)가 많다. 표본을 대문자, 소문자로 나누어 설명하고 있는 교재는 단 4권 (15.4%)에 불과하다 (Kim 등, 2012; Lee, 2009; Lee와 Kim, 2012; Rho 등, 2006).

또한 이러한 문제는 통계학 교재를 쓴 저자 자신이 일관적이고 체계적으로 이론 전개를 해나가지 못하게 한다. 즉, 신뢰구간을 구하는데 있어서 자유도 $(n-1)$의 t-분포 정의를

$$t = \frac{(\overline{X} - \mu)}{s/\sqrt{n}} \sim t_{(n-1)}$$

로 한다든지 (Hong, 2008; Kang 등, 2012; Kim 등, 2003; Kim 등, 2012; Kim 등, 2006; Moon, 2004; Shin과 Choi, 2006; Yoon과 Lee, 1996),

$$t = \frac{(\overline{x} - \mu)}{s/\sqrt{n}} \sim t_{(n-1)}$$

로 하는 등 올바르지 못한 교재들이 상당수 있다 (Cho 등, 2005; Kim과 Park, 2007; Kim 등, 2009; Lee, 2004; Lee와 Lee, 2006; Park 등, 2013; Yum 등, 2003). 무려 15권 (57.7%)의 통계학 교재들이 위와 같은 식들로써 이 부분을 올바로 설명하지 못하고 있다. 물론 이에 대한 올바른 표현은

$$T = \frac{(\overline{X} - \mu)}{S/\sqrt{n}} \sim t_{(n-1)}$$

이다. 특히, 신뢰구간을

$$\overline{x} \pm t_{\alpha/2} \frac{S}{\sqrt{n}}$$

으로 표현하고 있는 교재는 대문자와 소문자에 대해 전혀 구분하고 있지 않은데 (Lee, 2004), 일관적이지 못한 표현으로 통계학이론을 가르칠 경우 배우는 사람들이 얼마나 혼란스러울 것인가 걱정되지 않을 수 없다.

분산분석 모형이나 회귀분석 모형을 설명하는데 있어서도 일관성을 유지하지 못하고 있는 교재들이 많다. 비통계학전공 저자들의 경우 식 (2.5)와 같은 분산분석 모형을 소개하고 있지 않거나 모형에 대한 설명 없이 구체적인 예를 통하여 설명하는 경우가 대부분인데 이는 F-분포의 소개가 부담되기 때문으로 이해된다. 그러나 분산분석표를 이해시키려면 F-분포를 정확하게 설명하고 있어야만 한다. 왜냐하면 통계분석에 있어 분산분석표는 매우 기본적이며 다른 분석 기법에도 널리 응용되는 것이기 때문이다.

회귀분석 모형을 식 (2.5)와 같이

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \ i = 1, 2, \cdots, n$$
$$(\text{가정}) \ \epsilon_{ij} \sim N(0, \sigma^2)$$

로 표현하고 있는 교재는 전체 26권 중 7권 (26.9%)에 지나지 않는다 (Ko 등, 2010; Kim 등, 2003; Kim 등, 2012; Kim 등, 2007; Lee, 2009; Rho 등, 2006; Park 등, 2013). 더욱이 기울기, β_1에 대한 추정량을

$$\widehat{B}_1 = \frac{\Sigma(x_i - \overline{x})(Y_i - \overline{Y})}{\Sigma(x_i - \overline{x})^2}$$

으로 표현하여 가설검증의 절차를 설명하고 있는 교재는 단 한권에 불과하다 (Lee, 2009). 통계학 교재들이 추정치인 식 (2.4)만을 제공한 후, 가설검증의 절차를 설명하고 있는데 (미흡), 회귀분석 이론을 이해하기에 부족할 수밖에 없다.

Table 4.1과 Table 4.2에서 보는 바와 같이 통계학전공 저자들의 이론 전개가 비통계학전공 저자들의 그것에 비해서는 충실하다고 할 것이다. 그러나 Table 4.1의 경우에도 (통계학전공 저자), 확률분포부터 가설검증까지는 대부분 적절 (***)이지만 분산분석과 회귀분석에서는 체계적인 설명이 미흡 (**)한데 그 원인은 바로 저자들이 대문자와 소문자의 구분을 명확하게 인식하고 있지 않기 때문이다.

Table 4.2의 비통계학전공 저자들의 교재들은 대부분 소문자와 대문자를 구분하지 않고 이론을 전개하고 있기 때문에 신뢰구간과 가설검증 부분에서도 적절한 평가를 받지 못하고 있다. 그리고 분산분석에 대해서도 거의 이론에 대한 설명 없이 분산분석의 과정을 설명하고 있는 것은 통계기법의 오용, 남용으로 연결될 수도 있어 문제의 소지가 크다고 할 것이다.

더욱이 회귀분석을 분산분석보다 먼저 설명하는 교재들은 그 이유를 납득하기 어렵다. 분산분석표에 대한 이해 없이 어떻게 회귀분석을 설명하는가?

5. 정리 및 제언

통계학의 기초이론은 매우 단순하다. 모집단과 표본에 대한 차이를 알고, 모집단의 평균, μ와 이론적 표본평균, $\overline{X}$의 차이를 이해하고 중심극한정리를 확실하게 이해하면, 추정, 가설검증의 절차까지 매우 체계적으로 간단히 정리할 수 있다. 물론 이론을 정리해가는 과정에서 대문자와 소문자를 확실하게 구분할 수 있어야 한다. 하지만, 많은 통계학 교재들이 대문자와 소문자를 구분하지 않고 있어 통계학이론을 체계적으로 학습할 수 없고, 결국 "통계학은 어렵다"고 인식하고 있다고 생각된다. 대문자와 소문자를 명확하게 구분하고 있지 않은 것은 외국 통계학 교재에서도 거의 마찬가지라고 생각된다 (Kim 등, 2009).

대학의 많은 전공 분야에서 통계학이 필요한 만큼, 비통계학전공 교·강사들이 통계학 기초이론을 확실하게 이해하고 강의할 수 있는 인프라 구축을 제안한다. 이를 위해서는 통계학 교재에 대해 객관적으로 평가할 수 있는 시스템이 구축되고 통계학 기초이론 부분의 가이드라인이 만들어져야 한다고 생각한다. 이에 관련하여 우리나라에서도 통계학교육에 대한 연구가 활발해지기를 제안한다. 예컨대 미국의 Journal of Statistics Education에서는 통계학과의 교과목에 대한 연구를 포함하여 교육 방법론 등에 대해 활발한 연구가 진행되고 있다. 아마도 비통계학전공 분야의 석·박사학위 논문 실증분석에서 얼마나 많은 통계기법들의 오용, 남용이 있는지를 연구해 보는 것도 통계학교육에 대한 좋은 테마일 것이다.

결국 통계학을 폭넓게 보급하는 것은 통계학을 전공한 사람들의 책임이며, 통계학 교육에 대한 활발한 연구가 밑바탕이 되어 통계학이 쉽고 튼튼하게 발전되기를 감히 기대한다.

References

Cho, G. H. and Kim, T. G. (2005). *Statistics*, Global Books Press, Seoul.
Hong, C. S. (2008). *Statistics*, Jiin Books, Seoul.
Kang, K. S. (2006). *Modern Statistics*, 2nd Ed., Pakyoungsa, Seoul.

Kang, S. H., Kim, C. E., Kim, H. J., Park, S. H., Park, T. Y., Lee, H. B. and Jeon, Y. H. (2012). *Introduction to Statistics*, Freedom Academy, Kyunggi-do.

Kim, B. W., Choi, K. J., Bae k, H. Y., Kim, H. J., Dong, K. H., Park, T. R. and Jang, I. H. (2002). *Understanding of Statistics*, Freedom Academy, Kyunggi-do.

Kim, D. H., Kim, C. R., Son, G. T., Jung, K. M., Chung, Y. S., Choi, Y. S. and Hong, C. G. (2003). *Statistics*, 2nd Ed., Freedom Academy, Kyunggi-do

Kim, D. W., Namgoong, P., Park, J. S., Huh, M. E. and Hong, J. S. (2003). *Statistics*, Pakyoungsa, Seoul.

Kim, H. S. and Park, H. C. (2007). *Statistics*, Hyungseulsa, Seoul.

Kim, J. K., Park, J. H., Park, H. J., Lee, J. J., Jeon, H. S. and Hwang, J. S. (2012). *Statistics*, 2nd Ed., Freedom Academy, Kyunggi-do

Kim, S. I., Kim, H. M., Suh, H. S., Ahn, B. J., Yeo, S. C., Ryu, K. S. and Lee, S. G. (2012). *Statistics understanding and applications*, 3rd Ed., Minyoungsa, Seoul.

Kim W. C., Kim, J. J., Park, B. W., Park, S. H., Song, M. S., Lee, S. E., Lee, Y. J., Jeon, J. W. and Cho, S. S. (2007). *General Statistics*, 2nd Ed., Youngjisa, Seoul.

Kim, Y. J., Kim, J. I., Ryu, Y. H., Yoon, Y. H., Lee, Y. K., Lee, S. S. and Joo, S. E. (2009). *Introductory Statistics(translation of Introductory Statistics* authored by *Prem Mann)*, 6th Ed., Freedom Academy, Kyunggi-do.

Kim, Y. R., Min, C. K. and Yoon, S. H. (2006). *Business Statistics*, 2nd Ed., Myungkyungsa, Seoul.

Ko, S. N., Kim, N. Y. and Jang, Y. M. (2010). *Applied Statistics*, Chungmok Press, Seoul.

Lee, H. Y. (2004). *Statistics*, 2nd Ed., Cheongram Books, Seoul.

Lee, H. Y. and Lee, P. Y. (2006). *Statistics*, Freedom Academy, Kyunggi-do.

Lee, W. W. (2009). *Statistics*, 2nd Ed., Pakyoungsa, Seoul.

Lee, Y. H. (2003). *Statistics theory and applications*, Hakhyunsa, Seoul.

Lee, Y. G. and Kim, S. Y. (2012). *Understanding of Statistics*, 6th Ed., Yulgok Press, Seoul.

Moon, D. J. (2004). *Statistics*, Moonyoungsa, Seoul.

Rho, B. H., Min, J. H. and Lee, G. H. (2006). *Understanding of Statistics*, 2nd Ed., Beopmunsa, Seoul.

Ryu, S. J., Hong, G. S. and Kim, N. Y. (2011). *Statistics*, Pakyoungsa, Seoul.

Park, J. S., Yoon, Y. S. and Park, R. S. (2013). *Modern Statistics*, 5th Ed., Dasan, Seoul.

Shin, T. G. and Choi, S. C. (2006). *Statistics*, 5th Ed., Beopmunsa, Seoul.

Yoon, S. W. and Lee, T. S. (1996). *Practical Statistics*, Freedom Academy, Kyunggi-do.

Yum J. K., Kim, H. J., Lee, G. J., Kim, J. H., Sim, G. P., Jo, T. K. and Lee, Y. S. (2003). *General Statistics*, 2nd Ed., Gyowoosa, Seoul.

Journal of the Korean Data & Information Science Society
2013, **24**(6), 1253–1262

http://dx.doi.org/10.7465/jkdi.2013.24.6.1253
한국데이터정보과학회지

"Statistics is difficult"? - Textbooks problems[†]

Wonwoo Lee[1]

[1]Department of Statistics, Sejong University

Received 15 August 2013, revised 21 September 2013, accepted 1 October 2013

Abstract

This study observes not only how much those who studied Statistics during the college years feel that Statistics is difficult but also why they felt it was difficult. Most of the targeted researchers, 80.8 percent, say "Statistics was difficult". They selected the item "textbooks were hard to understand" as the main reason (62.5%). Based on the explanatory survey of text books, many textbooks do not distinguish the small letter, x from the capital letter, X. Hence, in this study, one of the main reasons why most of the researchers felt Statistics was difficult must be the ambiguousness of the notations. If authors keep in mind the importance of the difference between capital letters and small letters in Statistics, the Statistics learners' recognition of difficulty of Statistics will decline.

Keywords: Capital letter, random variable, small letter, statistics education.

[†] This study was supported by Sejong University Grant.
[1] Professor, Department of Statistics, Sejong University, Seoul 143-747, Korea.
E-mail: wwlee@sejong.ac.kr