응용통계연구 제 **20**권 *1*호, *2007*년, *pp. 159–165*

고등학교 수학 I「통계」에 대한 고찰

허명회[1)]

요 약

제7차 교육과정 고등학교 수학 I의 통계 단원을 논리의 정합성 측면에서 살펴보았다. 검토 결과 큰 수의 법칙에서 신뢰구간에 이르기까지 수리적 연결이 곳곳에서 끊어져 있어, 고교 수학 수업에 무리가 있는 것으로 보인다. 개선 방안으로, 1) 몇 가지 요목에 대한 고교 수준에서 교수 가능한 논거를 제시하고 2) 교수 요목의 변경 또는 축소를 제안한다.

주요용어: 고교 수학, 정규분포, 모평균에 대한 신뢰구간.

1. 들어가며

제7차 고등학교 교육과정에서 통계단원이 수학에 포함되어 있는 만큼 수학적 관점에서 현재 고등학교 수학 I 교과서들이 통계를 어떻게 다루고 있는가를 살펴보았다. 단순한 지식 전달보다는 논리적 정합성이 수학의 가치일 것이다. 그런데 큰 수의 법칙, 정규분포, 신뢰구간 등 주요 교수항목을 검토해본 결과 수리적 논리 측면에서 다수의 문제점이 발견되었다. 제시된 논거가 불충분하고 항목간 연결성이 떨어진다. 이 글은 수학적 관점에서 고등학교 수학 I의 통계단원을 항목별로 검토하고 몇 가지 개선의견을 제시한 것이다. 제7차 수학 I의 확률 및 통계 단원의 교수항목 및 일부 문제에 대하여는 장대흥과 이효정(2004)에 의해 다루어진바 있다. 검토한 교과서는 수학 I의 경우 우정호 외 5인(2003), 이강섭 외 6인(2003), 선택교과 <확률과 통계>의 경우 한국교원대학교 국정도서편찬위원회(2003) 등이나 사실상 모든 교과서들이 거의 같으므로 이하 지칭 사항들이 어떤 특정 교과서에 국한된 것은 아니다. 특별히 예외가 되는 경우엔 별도로 부기하기로 한다.

2. 현재 교과서, 제시된 논거, 검토 의견 및 제안

2.1. 큰 수의 법칙

현재 교과서: 어떤 한 시행에서 사건 A가 일어날 수학적 확률이 p일 때, n번의 (독립) 시행에서 사건 A가 일어나는 횟수를 X라고 하면, 아무리 작은 양수 ϵ을 택하더라도 다음이 성립한다.

1) (136-701) 서울시 성북구 안암동 5가 1, 고려대학교 정경대학 통계학과, 교수
E-mail: stat420@korea.ac.kr

$$\lim_{n\to\infty} P\left(|\frac{X}{n} - p| < \epsilon\right) = 1.$$

제시된 논거: 한 개의 주사위를 n번 던져서 1의 눈이 나오는 횟수를 X라고 할 때, $n = 10$, 30, 50의 각각에 대하여

$$p\left(|\frac{X}{n} - \frac{1}{6}| < 0.1\right)$$

을 구하여 보면 점점 큰 값이 나온다 (0.614, 0.784, 0.946).

제안 논거: 고등학교 수학에서 모든 진술이 엄밀하게 증명되어야 하는 것은 아니다. 그러나 단편적일 수밖에 없는 수치적 증거의 제시보다는 논리적 증명(logical demonstration)의 제시가 바람직하다고 본다. 이항분포에 국한된 체비셰프(Chebyshev) 부등식은 고교수학 I 범위 내에서 증명이 가능하다.

2.2. 정규분포

현재 교과서: 연속확률변수 X 중에서 확률밀도함수가 다음 식과 같을 때, X의 확률분포를 정규분포라고 하고, 함수 $f(x)$의 그래프를 정규분포곡선이라고 한다.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}, \quad -\infty < x < \infty.$$

여기서 m과 $\sigma(>0)$는 각각 평균과 표준편차를 나타내는 상수이며 e는 그 값이 2.7182...인 무리수이다.

제시된 논거: 자연현상이나 사회현상을 측정할 때 그 확률밀도함수가 종 모양의 곡선에 가까운 경우가 많이 있다.

검토 의견: 고교수학 I은 미분과 적분을 포함하지 않는다. 이에 따라 고교수학 I의 통계 단원에서 이산형 변수에 대하여는 기대값과 평균·분산을 정의하지만 연속형 분포에 대하여는 확률밀도함수만 소개할 뿐 기대값을 정의하지 않는다. 이 점은 최수일(2006)에 의하여 지적된 바도 있다. 그런데도 불구하고 정규분포의 평균과 분산을 언급하는 것은 문제이다. m이 분포의 중앙을 나타내고 σ가 분포의 퍼짐을 나타내는 파라미터라고 밖에 말하기 어렵다. 정규분포가 자연현상과 사회현상에서 많이 나타난다는 주장도 받아들이기 어렵다 (몸무게나 키 외에 실제 관측자료가 정규분포로 나타나는 경우가 주변에서 얼마나 자주 있는가?). 이것보다는 이항분포의 극한으로 정규분포를 도입하는 것이 바람직하다. 이런 방식은 한 심사자가 지적한대로 통계학의 역사와도 맥락을 같이 한다. Stigler(1986, 조재근 옮김)에 의하면 정규분포는 확률 게임(chance games)을 탐구하는 과정에서 1733년 De Moivre에 의하여 이항분포의 극한으로 유도되었으며 1809년 Gauss에 의하여 천문학에서 측정오차 자료에 대한 확률모형으로 적용되었다. 따라서 정규분포를 이항분포의 극한으로 도입하는 것이 더 자연스럽다.

2.3. 이항분포와 정규분포와의 관계

현재 교과서: 확률변수 X가 이항분포 $B(n,p)$ 를 따를 때, n 이 충분히 크면 X는 근사적으로 정규분포 $N(np,npq)$를 따른다 $(q=1-p)$.

제시된 논거: 이항분포 $B(n,p)$에서 p를 고정하고 n을 충분히 크게 하면 $B(n,p)$의 그래프가 $N(np,npq)$의 그래프에 가까워진다. 예로서 $p=1/6$로 고정하고 $n=10,20,30,50$인 경우에 대하여 $B(n,p)$의 그래프와 $N(np,npq)$의 그래프를 살펴본다.

제안 논거: $B(n,p)$의 그래프와 $N(np,npq)$의 그래프는 n이 변함에 따라서 중심과 산포가 움직이므로 분포의 수렴성을 말하기 어렵다. 실제로 보여야 하는 것은 이항분포에 적용된 중심극한정리인

> "확률변수 X가 이항분포 $B(n,p)$를 따를 때, n이 충분히 크면 $Z=\frac{X-np}{\sqrt{npq}}$는 근사적으로 정규분포 $N(0,1)$을 따른다."

이다. 그런데 고교 수학 I 범위에서 이것을 증명하기는 어렵다. 엑셀을 활용하여 다음을 실증해보이는 것이 최선일 듯 하다.

[1] $p=0.5$인 경우:

$A_n = P(\frac{X-n/2}{\sqrt{n}/2} \le z)$ 라고 하자. A_n은 이항분포의 누적확률인데, 이것은 엑셀을 활용하여 어렵지 않게 구할 수 있다. 예컨대 $z=1$인 경우 $n=10$, 20, 40, 80에 대하여 $A_n=0.8281$, 0.8684, 0.8659, 0.8428이다. 표준정규분포에서 $P(Z \le 1)=0.8413$이다. n을 크게 하면 $A_n \to 0.8413$에 수렴해간다.

[2] $p \neq 0.5$인 경우:

$B_n = P(\frac{X-np}{\sqrt{npq}} \le z)$ 라고 하자. B_n도 이항분포의 누적확률인데, 이것은 엑셀을 활용하여 어렵지 않게 구할 수 있다. 예컨대 $p=0.6$, $z=1$인 경우 $n=10$, 20, 40, 80에 대하여 $B_n=0.8327$, 0.8744, 0.8715, 0.8479이다. $p=0.1$, $z=1$인 경우 $n=10$, 20, 40, 80에 대하여 $B_n=0.7361$, 0.8670, 0.7937, 0.8266이다. 어느 경우에서도 n을 크게 하면 $B_n \to 0.8413$에 수렴해가지만 p가 0.5에서 떨어진 값일수록 수렴 속도가 늦는 것으로 알려져 있다.

검토 의견: 현재 교과서에는 정규분포를 소개하고 성질을 다룬 다음 이항분포의 정규근사를 기술하고 있으나, 이것보다는 이항분포의 극한을 기술하는 과정에서 정규분포를 유도하고 이후 정규분포의 제 성질을 다루는 것이 자연스러울 것으로 생각한다. 한 심사자가 지적한대로 이산형 분포의 극한이 연속형 분포로 되는 현상을 설명하기가 쉽지 않겠으나, 이는 점들이 누적되어 선이 되는 것과 같은 이치로 이해시킬 수 있을 것이다.

2.4. 정규분포의 표준화

현재 교과서: 확률변수 X가 정규분포 $N(m,\sigma^2)$을 따를 때, 확률변수 Z를 $Z=\frac{X-m}{\sigma}$ 이라고 하면 Z는 표준정규분포 $N(0,1)$을 따른다.

제시된 논거: $E(Z) = 0$, $V(Z)$=1 이기 때문이다.

제안 논거: 순차적으로 다음을 그래프로 보일 필요가 있다 (우정호 외 5인(2003)은 이런 방식으로 기술하고 있다).

[1] $X - m$은 $N(0, \sigma^2)$을 따른다: $P(a \le X \le b)$=$P(a - m \le Y \le b - m)$, $Y \sim N(0, \sigma^2)$.

[2] $\frac{Y}{\sigma}$는 $N(0,1)$을 따른다: $P(a' \le Y \le b')$=$P(a'/\sigma \le Z \le b'/\sigma)$, $Z \sim N(0,1)$.

2.5. 표본평균 $\bar{X}$의 분포

현재 교과서: 평균이 m이고 표준편차가 σ인 모집단에서 크기 n인 임의표본을 복원추출할 때, 표본평균 $\bar{X}$에 대하여 다음이 성립한다.

[1] $E(\bar{X}) = m$, $V(\bar{X}) \ = \ \sigma^2/n$.

[2] 모집단의 분포가 정규분포이면 $\bar{X}$는 $N\left(m, \frac{\sigma^2}{n}\right)$을 따른다.

[3] 모집단의 분포가 정규분포가 아닐 때도 표본의 크기 n이 충분히 크면 $\bar{X}$는 $N\left(m, \frac{\sigma^2}{n}\right)$에 가까워진다.

제시된 논거: 모집단이 $\{1,2,3,4\}$, $n = 2, 3$인 경우에서 [1]을 확인한다.

제안 논거: 앞서 논의한대로 연속형 분포에 대하여 평균과 표준편차를 말하기 어렵다. 표본평균 $\bar{X}$의 분포를 표본비율 $\hat{p}$의 분포로 대치하면 문제가 해결된다 (여기서 $\hat{p} = X/n$, X는 크기 n의 임의표본에서 성공 수). 즉 비율이 p인 모집단에서 크기 n인 임의표본을 복원추출하여 얻는 표본비율 $\hat{p}$에 대하여 다음이 성립한다.

[1] $E(\hat{p}) = p$, $V(\hat{p}) = p\ q/n$.

[2] 표본의 크기 n이 충분히 크면 $\hat{p}$의 분포는 $N\left(p, \frac{pq}{n}\right)$에 가까워진다.

2.6. 모평균 m에 대한 신뢰구간

현재 교과서: 모평균 m에 대하여

[1] 신뢰도 95 %의 신뢰구간: $\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \le m \le \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$.

[2] 신뢰도 99 %의 신뢰구간: $\bar{x} - 2.58\frac{\sigma}{\sqrt{n}} \le m \le \bar{x} + 2.58\frac{\sigma}{\sqrt{n}}$.

제시된 논거: 모집단의 분포가 정규분포이면 $\bar{X}$는 $N\left(m, \frac{\sigma^2}{n}\right)$을 따른다.

제안: 앞의 신뢰구간을 실제로 쓰려면 모표준편차 σ를 표본표준편차 s로 대치하여야 하나 마땅한 논리제시가 어렵다. 따라서 모평균 m에 대한 신뢰구간은 다루지 말고 이를 모비율 p에 대한 신뢰구간으로 대체하는 것이 좋겠다 (아래에서 $\hat{q} = 1 - \hat{p}$).

[1] 신뢰도 95%의 신뢰구간: $\hat{p} - 1.96\sqrt{\hat{p}\ \hat{q}/n} \le p \le \hat{p} + 1.96\sqrt{\hat{p}\ \hat{q}/n}$.

[2] 신뢰도 99%의 신뢰구간: $\hat{p} - 2.58\sqrt{\hat{p}\ \hat{q}/n} \le p \le \hat{p} + 2.58\sqrt{\hat{p}\ \hat{q}/n}$.

여기서도 $\sqrt{\hat{p}\hat{q}}$로 $\sqrt{pq}$를 대치하는 문제가 있으나 이것은 큰 수의 법칙으로 논거가 된다고 본다. 현실적으로 모비율 p에 대한 신뢰구간은 신문과 방송 등에서 선거예측조사를 포함 각종 여론조사 보도시 다루어지므로 고등학교 교과과정에 포함되어야 한다.

2.7. 선택교과「확률과 통계」

선택교과「확률과 통계」에도 수학 I의 통계 단원과 거의 같은 문제가 있다. 수학 II에서 정적분의 기초 단원을 이식하여 연속형 확률변수에 대하여 평균과 분산을 정의하여야 할 것이다. 정규분포 $N(m, \sigma^2)$의 평균과 분산이 각각 m과 σ임을 보이기 위하여

$$\int_{-\infty}^{+\infty} z\frac{1}{\sqrt{2\pi}}e^{\frac{-z^2}{2}}dz = 0, \quad \int_{-\infty}^{+\infty} z^2\frac{1}{\sqrt{2\pi}}e^{\frac{-z^2}{2}}dz = 1$$

임을 제시해주어야 할 것이다. 표본평균에 대한 큰 수의 법칙이 필요하다. 이러한 몇 가지가 사전에 조정된다면 모평균 m에 대한 신뢰구간이 큰 무리 없이 도입 가능하다.

3. 맺음 말

김희곤과 손중권(2004)은 고등학교 통계학 교육의 목표를 "지식 정보화 사회에 사는 시민으로서 통계를 근거로 생산된 각종 정보의 해독 이해 능력을 갖게 하고 다양한 정보로 지식을 생산할 수 있는 능력을 함양하는 것"으로 보았다. 이상복 등(2005)도 이와 유사한 입장에 있다. 그러나 이상복(2004)은 우리나라 고등학교 수학 I의 교육 목표가 "수학의 기본 개념, 원리, 법칙을 이해하고 수학적 사고력, 논리적 추론 능력을 키우는" 데 있다고 하였다. 전자는 실용성, 후자는 논리성에 비중을 둔 것이다. 이 글은 고등학교 교육과정에서 통계단원이 수학에 포함되어 있는 만큼 통계학의 지식 전달보다는 논리적 정합성이 중요하다고 보는 관점에서 본 것이다.

줄여 말하자면, 수리적 논리 측면에서 현재 고등학교 수학 I의 통계 단원은 많은 문제를 갖고 있어 개선이 필요하다. 특히 정규분포에 대한 설명과 모평균에 대한 신뢰구간에서 그렇다. 총량이 제한되어 있으므로 범위를 축소하되 축소된 범위 내에서 보다 엄밀한 수리 전개를 하는 것이 바람직하다고 본다. 현재 교과서는 논리적이지도 않고 실용적이지도 않다.

참고문헌

김희곤, 손중권 (2004). 고등학교의 통계교육의 문제점 및 개선방향, <한국통계학회 추계학술논문발표회>.

우정호 외 5인 (2003). <수학 I>, 서울, 대한교과서.

이강섭 외 6인 (2003). <수학 I>, 서울, 지학사.

이상복 (2004). 한국과 일본의 고등학교 수학 교육과정과 확률통계 교육, <한국통계학회 추계학술논문발표회>.

이상복, 손중권, 정성석 (2005). 수학 I 검정교과서 확률통계 영역에 대한 연구, <응용통계연구>, **18**, 197–210.

장대홍, 이효정 (2004). 제 7차 수학과 교육과정에 따른 실용수학 및 수학 I 확률 및 통계 단원 분석, <한국통계학회 추계학술논문발표회>.
최수일 (2006). 제 7차 수학과 교육과정에 따른 고등학교 확률 및 통계단원 교육과 개선방향, <한국통계학회 춘계학술논문발표회>.
한국교원대학교 국정도서편찬위원회 (2003). <확률과 통계>, 교육인적자원부.
Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*, Harvard University Press. (조재근 옮김, 「통계학의 역사」 한길사, 2005)

[2006년 7월 접수, 2006년 9월 채택]

Critical Review of Statistics Chapter in High School Mathematics I

Myung-Hoe Huh[1)]

ABSTRACT

The statistics chapter in High School Mathematics I as implemented in The 7th Curriculum is reviewed critically. In views from mathematical integrity or logic, the current contents are not satisfactory in several key issues. Specific instances are the law of large numbers, normal distribution and confidence intervals for population mean. We suggest alternative teaching points to handle such difficulties and propose re-structuring the course syllabus with reduced items.

Keywords: High school mathematics, normal distribution, confidence interval.

1) Professor, Department of Statistics, Korea University, Anam-Dong 5, Sungbuk-Gu, Seoul 136-701, Korea
E-mail: stat420@korea.ac.kr