통계연구(2017), 제22권 제1호, 27-43

표본조사에서 베이지안적 추정방법[1)]

강승택[2)] · 이상은[3)]

요약

직종별 훈련 수요는 직종별 훈련 프로그램을 효율적으로 운영하기 위해 매우 정확히 추정되어야 한다. 따라서 이를 위한 표본조사가 실시되고 있다. 이때 수요조사는 H-T(Horvitz-Thompson)추정 방법을 사용함으로써 매우 과대 추정되는 것으로 알려져 있다. 과대 추정을 해결하기 위한 방법으로 베이즈 정리를 이용한 추정량이 제안되었으며 이를 위해 한국고용정보원에서 제공하는 자료인 HRD-net을 이용하였다. 그러나 현재 한국고용정보원에서 제공하는 정보가 부족하여 이 방법을 적용할 경우 과소 추정되는 것으로 나타났다. 이에 본 논문에서는 베이즈 추론을 이용하여 현실적으로 타당한 수요 예측이 가능한 방법을 제안하였다. 또한 실제 자료 분석을 위해 그 타당성을 검증하였다.

주요용어 : 사건 중심 베이즈 정리, 베이즈 추론, 훈련수요예측

1. 서론

표본조사에서 베이즈 추정방법을 적용하는 사례가 늘어나고 있다. 이는 조사환경이 어려워짐에 따라 무응답이 늘어나고 있으며 조사대상의 수, 즉 표본규모를 충분히 설정하기에는 많은 부담이 되기 때문이다. 또한 조사항목 역시 민감하거나 난해한 항목은 다수의 무응답을 초래하므로 행정자료 등을 이용한 추정값을 활용하고자 하는 경향이 있다. 이러한 모든 경우에 베이즈 추정방법은 표본조사 추정량의 효율성을 향상시키는 데 도움이 된다. 그 외에도 표본조사에 베이즈 추정방법의 적용을 제안하고 있다. Hamner(2001, 2003)은 표본조사에서 베이즈 방법에 의한 추정방법을 다양하게 소개하고 있으며 Rao(2011)는 표본조사에서 기존의 전통적 방법과 베이지안 접근방법을 비교하였다. 이때 기존의 설계/사후 가중치를 적용하는 방법과 모형기반(model-base) 추정법을 비교하였다. 그리고 최근 Little(2015)은 30주년 JOS conference에서 표본조사의 무응답 대체(imputation)에 사용된 베이지안 방법을 표본으로 추출 않은 자료의 추정에 적용하는 방법을 제안하였다. 이는 최근 정보가 주어진 표본추출법(informative sampling)에서의 추정방법에서 베이즈 추정방법의 활용 가능성을 주는 연구라 판단된다.

최근 청년실업을 위한 취업 혹은 인재개발 교육사업이 각 분야에서 다양하게 이루

1) 본 연구는 2015학년도 경기대학교 학술연구비(일반연구과제)지원에 의하여 수행되었음.
2) 경기도 수원시 영통구 광교산로 154-42 , 경기대학교 응용통계학과 석사과정. E-mail: pooooo100@naver.com
3) 교신저자, 경기도 수원시 영통구 광교산로 154-42, 경기대학교 응용통계학과, 교수 E-mail: sanglee62@kgu.ac.kr

어지고 있다. 일반적으로 직종훈련 관련 교육프로그램의 지원에 앞서 직종별 교육훈련 수요조사가 이루어지며 수요조사에서 얻은 추정 훈련수요자수에 따라 훈련프로그램이 운영된다. 그러나 훈련과정 후 훈련받은 혹은 그 외의 직군으로 취업하는 비율은 기대보다 높지 않다. 이에 현재까지 사용되고 있는 훈련수요조사에서의 H-T추정량에 직종 훈련 후 직종별 취업 자료를 사전 정보로 이용한 베이즈 정리(Bayes' Theory)를 적용한 추정방법이 제안되었다.(최영섭과 양정승(2015)). 이 방법에서 훈련수요는 조사 항목 '채용계획인원'에서 추정된 전체 채용자 중 '훈련경험 비중'을 고려하여 추정된다. 즉 '채용계획인원'을 관측값/우도함수로 '훈련경험 비중'을 사전정보/사전함수로 사용하여 신규채용자 중 훈련을 받았을 사후확률을 계산한 후 이를 이용하여 H-T방법으로 얻어진 추정 훈련수요자수를 조정한다. 이때 적용한 베이즈 정리(Bayes' Theory)에서 사전정보로 사용된 자료는 한국고용정보원에서 제공한 HRD-net 자료로 각 직종별 훈련자수와 취업자 수의 요약자료이다. 따라서 사건중심의 베이즈 정리가 적용될 수 있다. 그러나 모든 직종별 사전확률이 존재하지 않으며 이러한 경우에는 평균 혹은 최근방의 자료를 이용하여 대체한 후 사전확률을 계산한다. 그러나 사전확률의 대체 방법/조건에 따라 사후확률의 변동이 매우 커지기 때문에 베이즈 정리를 이용한 추정방법은 사정정보에 따라 강건(robust)하지 않은 추정 결과를 나타내게 된다. 따라서 본 연구에서는 HRD-net의 요약자료에서 사전확률이 아닌 사전분포를 유도하여 베이지안적 추론(Bayesian Inference)에 의한 추정량을 제안하였다. 먼저 2장에서 본 연구를 하게 된 근본적인 문제를 설명하였으며 3장에서는 베이즈 정리와 베이지안적 추론에 의한 추정방법을 설명하였다. 이어서 4장에서는 실제 자료를 적용하여 각 방법의 성능을 비교하였다. 5장에 본 연구의 요약과 결론이 있다.

2. 문제제기

현재 고용노동부에서는 교육훈련기관중심 교육훈련 패러다임을 기업중심 교육훈련으로 전환하기 위해 전국 16개 지역에 인적자원개발위원회를 출범시켰으며 "지역·산업 맞춤형 인력양성체계" 구축을 추진하고 있다. 각 지역의 인적자원개발위원회에서는 기업에서 필요로 하는 인력을 맞춤형으로 양성하여 공급하기 위해 산업계의 수요파악에 많은 노력을 기울이고 있다. 이를 위해 기업에서 원하는 신규인력을 양성(양성훈련)하여 공급(공급훈련)하고, 재직인력의 역량을 향상시키기 위한 수요조사를 실시하고 있다. 이 조사의 주된 목적은 각 지역에서 직종 전체의 훈련 수요자수를 파악하는 것이다. 이에 본 논문에서는 양성훈련 수요자 추정에 중점을 두고 연구하였다.

최영섭과 양정승(2015)에서 지역별 양성훈련 규모는 사업체 조사결과인 직종별 '신규채용인원과 HRD-net 계좌제 훈련의 직종별 취업 비율을 이용하여 추정하고 있다. 이는 HRD-net 정보를 사전정보로 사용하는 베이즈 정리를 이용한 추정방법이다. 추정법에서 사용한 베이즈 정리는 다음과 같다.

$$P(T_j|E) = \frac{P(E|T_j)P(T_j)}{\sum_{i=1}^{K} P(E|T_i)P(T_i)}$$

여기서 E는 취업자수이며 T_j는 j직종의 훈련자수 이다. K는 전체 훈련 직종 수가 된다. 따라서 취업자(E)에서 j직종의 훈련(T_j)을 받았을 확률을 계산하기위해 분모의 전확률 공식에서 서로 배반 사건인 T_i가 $P(T_i)>0$을 만족해야 한다. 그러나 현실적으로 모든 직종 i 에서 T_i가 존재하기 어려울 수 있다. 여기서 HRD-net의 자료가 이에 해당된다. 즉 양성훈련수요의 경우 i 직종별 총 훈련자수인 T_i가 발생할 확률인 $P(T_i)=0$인 경우가 많이 있다. 따라서 본 연구에서는 타당하고 합리적인 사전확률 분포를 적용한 양성훈련 수요자 예측 방법을 제안하였다.

3. 추정방법

일반적으로 표본조사에서 가장 많이 적용되는 총계 추정법은 조사된 값에 설계가중치를 곱한 후 합하는 방법을 사용한다. 즉 i 직종의 총 훈련 수요자수를 Y_i라 할 때 H-T 추정법에 의한 추정량 $\hat{Y}_{iHT}$는 다음과 같다.

$$\hat{Y}_{iHT} = \sum_{j=1}^{n_i} \frac{y_{ij}}{\pi_i} = \sum_{j}^{n_i} w_i y_{ij} \tag{3.1}$$

여기서 y_{ij}는 i 직종 j번째 관측값 이며 N_i는 i 직종의 산업분류의 모집단 사업체수 n_i는 i 직종의 산업분류의 표본 사업체수 $w_i = N_i/n_i = 1/\pi_i$로 설계 가중치이며 $\hat{Y}_{iHT}$의 분산 추정량은 다음과 같다.

$$Var(\hat{Y}_{i_{HT}}) = \sum_{j}^{n_i} w_i^2 \frac{(y_{ij} - \hat{\bar{Y}}_{iHT})^2}{n_i - 1}(1 - f_i) \tag{3.2}$$

여기서 $f_i = \frac{n_i}{N_i}$이고, $\hat{\bar{Y}}_{iHT} = \frac{\hat{Y}_{iHT}}{N_i}$이다.

그러나 사업체 수요조사에서는 각 사업체 부서별로 필요 또는 희망하는 모든 직종별 훈련수요가 조사가 된다. 이 자료에 H-T 추정량을 적용하면 총 훈련수요 추정은 매우 과대추정 된다. 이는 수요조사의 조사표에 따르면 원하는 모든 훈련을 기입하도록 되었기 때문이다. 즉, 다중응답의 형태이다. 따라서 기존의 직종별 훈련자수와 그에 따른 취업자 수를 이용한 베이즈 추정 방법을 적용하여 합리적인 추정량을 도출할 필요가 있다. 이에 이 절에서는 한국직업능력개발원에서 제안한 사건 중심의 베이즈

정리(Bayes Theory)를 이용하는 경우와 본 연구에서 제안한 변수 중심의 베이즈 정리, 즉 베이지안 추론(Bayesian Inference)을 이용한 방법을 설명하였다.

3.1 사건 중심의 베이즈 정리(Bayes Theory)를 적용하는 경우

베이즈 정리는 사전 정보와 자료에서 얻어진 새로운 정보를 결합한 사후 정보를 얻는 방법으로 사건 중심의 베이즈 정리는 다음과 같다. 즉 직종별 총 훈련자 수를 $T_1, T_2, T_3, \cdots, T_K$라 하고

i) $T_i \cap T_j = \varnothing \ , i \neq j$

ii) $T_1 \cup \cdots \cup T_k = S$

를 만족한다면 조건부 확률 $P(T_i|E_i)$는 다음과 같다.

$$P(T_i|E_i) = \frac{P(E_i|T_i)P(T_i)}{\sum_{k=1}^{K} P(E_i|T_k)P(T_k)}$$

여기서 E_i는 직종 i의 취업자 수이고 $P(T_i)$는 i 직종 훈련을 받은 확률(사전확률)이며 $P(T_i|E_i)$는 조건부확률(사후확률)이다. 따라서 사건 중심의 베이즈 정리를 이용한 i 직종의 총 훈련 수요자수 Y_i의 추정량은 다음과 같이 얻어진다.

$$\hat{Y}_{i_{BT}} = \hat{Y}_{i_{HT}} \times P(T_i|E_i) \tag{3.3}$$

여기서 $\hat{Y}_{i_{HT}}$는 H-T 추정량이다. 이 방법은 기존의 i 직종 훈련을 받은 후 취업한 경우들의 사전정보, 즉 직종별 훈련자 수를 기준으로 한 취업자 수를 이용하여 H-T 추정량을 보정하기 때문에 과대 추정을 방지할 수 있다.

3.2 베이지안적 추론/모형(Bayesian Inference)을 적용하는 경우

이 절에서는 변수 중심의 베이즈 정리인 베이즈 추론은 일반적으로 관측값에서 얻은 자료의 분포를 우도함수 $Y|\theta \sim f(Y|\theta)$로 하고 우도함수에 포함된 모수 θ의 분포가 사전함수 $\theta \sim g(\theta)$을 따른다고 가정하면 θ의 사후함수는 $p(\theta|Y) \propto f(Y|\theta)g(\theta)$이다. 이때 베이즈 추론에 의한 θ의 베이지안 추정량은 $\hat{\theta}_B = E(\theta|Y)$ 이 된다.

Pfeffermann(1998)는 표본조사에서 관심변수 Y_i의 모집단의 분포를 $Y_i \sim f_p(y_i|\theta)$라 하며 이때 Y_i 의 표본분포, $f(y_i|i \in s)$를 다음과 같이 제안했다.

$$f_s(y_i|\theta^*) = f(y_i|i \in s) = \Pr(i \in s|y_i)\, f_p(y_i|\theta) \,/\, \Pr(i \in s) \tag{3.4}$$

여기서 S는 표본, θ^*은 임의의 모수 θ의 함수로 $\Pr(i \in s|y_i)$을 나타낸다.

이 방법들을 본 연구에 적용하면 총 훈련수요자수의 베이즈 추정량은 $\hat{Y}_{iHT}$에 i 직종에 취업한 취업자가 i 직종의 훈련을 받았을 사후확률인 $f(T_i|E_i,P_i)$를 곱한 식으로 얻어진다. 여기서 T_i가 i 직종의 훈련자수, E_i가 i 직종의 취업자 수, P_i는 취업확률로 매우 자연스럽게 $E_i|T_i,P_i \sim Bin(T_i,P_i)$를 가정할 수 있으며 또한 $P_i = \frac{E_i}{T_i}$가 된다. 따라서 T_i에 대한 사후확률함수는 다음과 같이 정의된다.

$$f(T_i|E_i,P_i) = \frac{f(E|T_i,P_i)f(T_i)f(P_i)}{\int_{t_i} f(E|T_i,P_i)f(T_i)f(P_i)dt_i} \tag{3.5}$$

또한 훈련자수의 분포는 포아송 분포를 따른다고 가정하고, i 직종 훈련자가 i 직종에 취업할 확률인 P_i는 $Beta$ 분포를 따른다고 가정할 수 있다. 즉 $T_i \sim Poi(T_{io})$이고 $P_i \sim Beta(\alpha,\beta)$를 사용할 수 있다. 이 내용을 식 (3.5)에 대입하면 사후확률을 구할 수 있다. 결과적으로 i 직종의 총 훈련수요자수 Y_i의 베이즈 추정량은 다음과 같이 계산된다.

$$\hat{Y}_{iBI} = \hat{Y}_{iHT} \times f(T_i|E_i,P_i) \tag{3.6}$$

4. 실제 자료를 이용한 응용예시

4.1 자료소개

한국직업능력개발원에 의해 기획된 훈련수요조사에서 직종별 양성훈련 수요는 신규 채용자 중 훈련 경험자 비중 결과와 수요조사인 사업체조사의 채용계획인원으로 추정된다. 이때 직종은 직종분류(KECO) 2자리 또는 3자리 단위를 기준으로 정해진다.

여기서 추정방법에 이용되는 사전정보 자료로 한국고용정보원에서 제공하는 HRD-net/고용보험연계 자료의 형태는 다음과 같다.

〈표 4.1〉 HRD-net/고용보험연계 자료의 형태

지역	훈련직종 중분류 (KECO 2자리)	훈련직종 중분류 (KECO 3자리)	취업 업종								총합
			A	B	C	D	E	…		미취업	
11	02	022								24	30
		023			7					18	36
		024			1					77	106
	⋮	⋮			…			…		⋮	⋮

〈표 4.1〉은 각 지역별, 직종별 총 훈련인원수와 훈련 후 직종별 취업현황표이다. 예를 들어 〈표 4.1〉의 서울(11) 지역, KECO 02/KECO 022를 설명하면 훈련교육을 받은 사람 30명 중에 6명은 취업, 24명은 미취업 상태를 말해준다. 또한 전술한 것처럼 본 연구에서 사용된 추정량 계산에 필요한 다른 모든 자료는 표본조사인 수요조사에서 얻어진다.

4.2 추정방법의 비교

훈련수요 추정방법으로는 한국직업능력개발원에서 제안한 1. 베이즈 정리(Bayes' Theory : BT)를 이용하는 방법과 본 연구에서 새롭게 제안한 2. 베이지안적 추론(Bayesian inference : BI)를 이용하는 방법이 있다. 본 연구에서는 각 지역별*직종별(KECO) 총 훈련자수(T)를 현재 HRD-net이 가지고 있는 총 훈련자수를 평균으로 하는 포아송 분포로, 취업자 수(E)는 총 훈련자수(T)를 총 개수로 하고, 성공확률 P ($P=\frac{E}{T}$)로 하는 이항분포를 이용한다. 각 방법의 성능을 비교하기 위해 $R=1000$번의 자료를 생성하여 모의실험을 수행하였다.

〈표 4.2〉 생성된 모의실험 자료의 형태

HRD-net			$R=1$			$R=2$			..	$R=1000$		
x	t	p	x	t	p	x	t	p		x	t	p
24	37	0.730	22	27	0.815	22	25	0.880	..	27	37	0.730
18	34	0.677	18	40	0.450	21	35	0.600	..	23	34	0.677
77	118	0.712	80	109	0.734	77	100	0.770	..	84	118	0.712
.	.	.	.	.	.	.	.	.	..	.	.	.
95	122	0.853	91	124	0.734	89	132	0.674	..	104	122	0.853
108	139	0.813	107	132	0.811	101	131	0.771	..	113	139	0.813

4.2.1 베이즈 정리(Bayes' Theory : BT)를 이용하는 경우

사건 중심의 베이즈 정리를 이용한 방법은 최영섭과 양정승(2015)의 '훈련수요조사 결과 가이드'에 소개되었다. 이를 간단히 설명하면 다음과 같다. 먼저 수요조사를 통해 각 사업체의 근로자를 대상으로 각 직종별 (양성)훈련자수 추정을 위한 조사를 실시한다. 즉 수요조사에서는 각 사업체에서 조사된 '채용계획인원' 전체에 대해 '채용자 중 훈련 경험자 비중'을 이용하여 (양성)훈련자수를 추정한다. 따라서 수요조사를 통해 각 사업체의 직종별 채용 예정인원을 추정할 수 있으며 한국고용정보원에서 제공하는 HRD-net 자료에서는 훈련자의 직종별 취업 비중을 계산 할 수 있다. 따라서 직종분류(KECO) 기준으로 i 직종의 예상 (양성)훈련수요(Expected Initial Training : $\widehat{IT}$)수는 다음과 같이 계산된다.

$$\widehat{IT_i^{BT}} = \hat{T}_i \times P(T_i|E_i)$$

이때 $\widehat{T}_i$(Expected Training)는 수요조사 자료에 H-T 방법을 적용하여 추정한 i 직종의 총 훈련 수요자수이며, $P(T_i|E_i)$는 T_i의 사후확률이다. 즉 $P(T_i|E_i)$는 i 직종의 신규 취업자 중 훈련 경험자 비중을 나타내는 확률이다.

여기서 $\widehat{IT_i^{BT}}$를 구하는 단계는 다음과 같다.

단계 1: $\widehat{T}_i$은 사업체 수요조사 결과인 조사표 문항 '채용계획 인원/훈련수요인원'에서 조사된 값 d_i에 설계(사후)가중치 sw_{ij}를 곱하여 얻는다. 즉, $\widehat{T}_i = d_i \times sw_{ij}$로 구한다.

단계 2: $P(T_i|E_i)$은 사업체 수요조사에서 i 직종의 신규 취업자 중 훈련 여부를 알 수 없기 때문에 HRD-net에서 얻은 연도별 직종분류별 훈련 후 취업 분포를 Bayes 정리에 적용하여 구한다. 즉 다음의 식 (4.1)을 이용한다.

$$P(T_i|E_i) = \frac{P(E_i|T_i)P(T_i)}{\sum_k P(E_i|T_k)P(T_k)} = \frac{P(E_i|T_i)P(T_i)}{P(E_i)} \tag{4.1}$$

이때 $P(E_i|T_i)$은 i 직종 훈련을 받은 훈련 경험자가 i 직종에 취업할 확률로 수요조사 자료에서는 결과를 얻을 수 없기 때문에 HRD-net의 자료로 부터 얻어진 $P(\widetilde{E}_i|T_i) = \widetilde{E}_{ii}|T_i$ ($\widetilde{E}_i$는 i직종 훈련 후 취업된 사람 수)를 이용한다. 즉, $\widetilde{E}_{ii}|T_i$는 i 직종 훈련 후에 취업한 비중이며, 이는 HRD-net 자료에서 i 직종 훈련 후 취업한 비중, $\widetilde{E}_{ii}|T_i = \widetilde{E}_{ii}|\sum_j \widetilde{E}_{ij}$로 계산된다. 따라서

$$P(E_i|T_i) = P(\widetilde{E}_i|T_i) = \widetilde{E}_{ii}|T_i \tag{4.2}$$

로 구하게 된다.

단계 3: $P(T_i)$는 사전확률로 i 직종 훈련을 받았을 확률이며 이는 HRD-net에서 직종 i 총 훈련자수 T_i와 사업체 수요조사에서 얻어진 전체 신입사원의 수요 대비 비중을 이용하여 얻는다. 즉 HRD-net에는 훈련을 받은 사람의 수가 T_i로 전체 취업자 중 훈련을 받았을 확률을 구할 수 없다. 이에 사업체 수요조사 자료에서 가중치를 적용한 신규 채용자수를 총 훈련자수로 가정한다. 따라서 $P(T_i) = \frac{T_i}{\sum_i E_i}$로 계산된다. 여기서 $\sum_i E_i = \sum_i \sum_j sw_{ij} C_{2_{ij}}$이고 C_2는 조사항목 '신규 졸업자 및 경력 1년 미만자' 수로 조사된 자료에서 얻는다. 이때 T_i와 E_i는 서로 다른 자료에서 얻어진 결과이므로

$P(T_i)$ 확률의 합이 "1"이 되도록 조정한다. 즉 $\sum_i P(T_i)=1$이 되도록 조정한다.

단계 4: 식 (4.1)의 분모에서 $\sum_k P(E_i|T_k)P(T_k)$과 $P(E_i)$를 비교해보면 직접적으로 $P(E_i)$를 구하는 것이 매우 편리하다. 여기서 $P(E_i)$는 i 직종의 취업자의 비중을 의미하는 확률이므로 조사항목 '신규 졸업자 및 경력 1년 미만자', C_{2ij}에서 얻은 결과에 사후 가중치, sw_{ij}를 적용 적용하여 i 직종의 취업할 확률 $P(E_i)$을 구한다. 즉 다음의 수식으로 구한다.

$$P(E_i)=\frac{\sum_j sw_{ij}C_{2ij}}{\sum_i\sum_j sw_{ij}C_{2ij}}$$

단계 5: 최종적인 $P(T_i|E_i)$는 다음과 같이 계산된다.

$$P(T_i|E_i)=\min(1,\ \frac{P(\widetilde{E}_i|T_i)P(T_i)}{P(E_i)})$$

이는 확률이 "1" 이상인 것을 제거하기 위함이다.

단계 6: 최종 i 직종의 양성훈련수요자 수의 추정은 다음과 같다.

$$\widehat{IT_i^{BT}}=\widehat{T}_i\times P(T_i|E_i) \tag{4.3}$$

4.2.2 베이지안적 추론(Bayesian Inference : BI)을 이용하는 경우

사전정보가 요약자료(summary data)로 존재하는 경우에는 사건 중심의 베이즈 정리를 이용하여 사후확률을 구하는 것은 일반적이다. 이때 주어진 사전 정보가 모든 경우/셀에서 값이 존재하는 경우에는 모든 셀의 사후확률 계산이 가능하다. 즉 베이즈 정리에서는 i) $T_i\cap T_j=\varnothing$,$i\neq j$과 ii) $T_1\cup\cdots\cup T_K=S$를 가정하고 있다. 따라서 요약자료에서 특정 셀에 자료가 없는 경우 혹은 확률이 '0'인 경우에는 그 셀의 사후확률을 계산하는 것은 불가능하다. 결국 사후확률 결과는 사전정보에 매우 민감하게 작용하게 된다. 같은 맥락에서 주어진 자료의 특정 직종에 취업자가 없거나 매우 작은 경우에는 수요조사에서 얻은 훈련수요의 추정값이 큰 값으로 추정되었음에도 불구하고 '0'에 가까운 값으로 추정된다. 따라서 본 연구에서는 요약자료로 얻어진 사전정보를 사건의 확률로만 이용하는 베이즈 정리(BT) 방법을 확장하여 변수 중심의 확률분포를 이용한 베이지안 추론적(BI) 방법을 제안하였다. 즉 베이지안 추론적(BI) 방법에서는 직종분류(KECO)를 기준으로 한 i 직종의 예측 (양성)훈련수요자 (Expected Initial Training : $\widehat{IT_i^{BI}}$)는 다음과 같이 계산된다.

$$\widehat{IT_i^{BI}} = \hat{T}_i \times f(T_i|E_i,P_i) \tag{4.4}$$

여기서 전술한 것처럼 P_i는 직종 i의 취업 확률이고 E_i는 직종 i의 취업자 수 그리고 T_i는 직종 i의 총 훈련자 수를 나타낸다. 이때 $\hat{T}_i$는 BT방법과 동일하게 조사자료에서 H-T 방법에 의해 얻어지며 기존의 취업자들(E_i)로부터 양성훈련을 받은 사람의 수(T_i)의 사후확률분포인 $f(T_i|E_i,P_i)$는 HRD-net 자료로 부터 i 직종의 직종훈련 후 취업자 수 E_i가 i 직종의 훈련을 받았을 확률을 구하게 된다. 이때 시전확률 분포로 지난해 총 훈련자수에 맞는 포아송 분포를 사용하였으며 이때 평균은 HRD-netm에서 얻은 직종별 평균 훈련자수를 활용하였다. 따라서 $T_i|E_i,P_i$의 사후확률함수는 다음과 같다.

$$f(T_i|E_i,P_i) \propto f(E_i|T_i,P_i) f(T_i) f(P_i)$$

여기서 $E_i|T_i,P_i \sim Bin(T_i,P_i)$으로 i 직종 훈련 후의 i 직종에 취업한 비중의 분포이다. 또한 사전함수, $T_i \sim Poi(T_{i0})$에서 T_{i0}은 HRD-net 자료에서 작년 한해 직종 i의 총 훈련자수 평균이며, $P_i \sim Beta(\alpha,\beta)$를 사용한다. 이때 초모수 α, β는 타당한 값을 이용한다.

4.2.3 각 방법을 이용한 모의실험 결과

모의실험은 특정 A 지역의 조사 자료와 사전정보를 이용하였으며 이때 다음은 각 KECO i 별로 $\hat{Y}_{HT} = \hat{Y}_{iHT} = \sum_{j=1}^{n_i} \frac{y_{ij}}{\pi_i} = \sum_{j}^{n_i} w_i y_{ij}$, $\hat{Y}_{BT} = \hat{Y}_{iBT} = \frac{1}{R}\sum_{r=1}^{R} \widehat{IT_i^{BT}}_r$, $\hat{Y}_{BI} = \hat{Y}_{iBI} = \frac{1}{R}\sum_{r=1}^{R} \widehat{IT_i^{BI}}_r$ 으로 총 훈련수요자수의 추정 결과 이다. 본 연구에서는 SAS를 활용하였습니다. 이때 SAS에서 가능도함수인 $E_i|T_i,P_i \sim Bin(T_i,P_i)$은 HRD-net에서 얻은 취업지수 E값과 관측값에서 얻은 훈련 총 수요자수 T 값으로부터 얻은 P를 이용하여 얻어지며 사전함수, $T_i \sim Poi(T_{i0})$는 HRD-net 자료에서 작년 한해 직종 i의 총 훈련자수 평균을 사용한다. 이때 초모수 α, β는 임의 값으로 모의실험이 이루어 졌다.

〈표 4.3〉 직종 i 별 $\hat{Y}_{HT}$, $\hat{Y}_{BT}$, $\hat{Y}_{BI}$ 의 추정결과

KECO(i)	$\hat{Y}_{HT}$	$\hat{Y}_{BT}$	$\hat{Y}_{BI}$		
			$(\alpha,\beta)=(1,0.2)$	$(\alpha,\beta)=(1,1)$	$(\alpha,\beta)=(1,5)$
01	313.830	0	26.558	54.507	12.309
02	1478.422	46.964	86.070	186.532	13.937
04	200.215	1	13.918	24.791	0.617
06	2586.379	13.982	130.603	326.622	51.214

07	684.314	4.026	50.097	86.083	1.889
08	66.125	19	3.897	7.895	0.030
09	759.460	3	37.12	95.600	16.608
10	441.644	1	46.344	53.486	0.109
11	1902.279	0	180.588	214.104	1.275
12	53.555	24	3.835	6.299	0.015
13	452.488	39.119	34.412	56.762	0.588
14	6181.368	11	479.466	781.293	14.373
15	2879.445	5.403	170.535	363.605	25.683
16	1942.850	4	115.766	245.091	16.677
17	1088.955	0	55.734	134.284	30.297
18	843.364	9	91.573	106.192	0.059
19	1590.830	3	99.869	200.742	10.691
20	251.259	6	16.155	31.276	0.964
21	400.844	8.7	31.995	50.211	0.330
22	2192.696	6	182.961	276.833	3.457
23	64.250	2	3.935	7.641	0.031

〈표 4.3〉의 수요조사 결과에서 얻은 추정량 $\hat{Y}_{HT}$은 과대추정되고 있음은 이미 알려져 있다. 이를 보정하기 위해 사용된 두 추정량에서 $\hat{Y}_{BT}$를 보면 KECO 01($i=01$) 인 경우 $\hat{Y}_{HT}$은 약 134명이 추정되었음에도 '0'명으로 추정되었다. 이는 주어진 사전정보에서는 01 직종의 훈련생에서 취업자 수가 매우 적어서 취업확률이 '0'에 가깝게 나타났기 때문에 얻어진 결과이다. 그러나 $\hat{Y}_{BI}$의 경우에는 주어진 α,β 값에 따라 직종 i의 취업 확률의 달라지며 이에 따라 추정결과가 달라진다. 즉 사정정보에 따라 추정의 결과에 차이가 있음을 알 수 있다. 따라서 최적의 초모수 값을 이용하는 것은 중요하다. 이는 $\hat{Y}_{HT}$이 과대추정 되었음은 분명하지만 조사결과에서 요구 수요가 있는 한 이를 최대한 고려하는 것이 바람직하기 때문이다.

이를 확인하기 위해 〈표 4.4〉에서 $\hat{Y}_{BI}$의 경우 사전분포에 주어진 초모수의 값에 따라 추정값이 얼마나 변화하는지 살펴보았다.

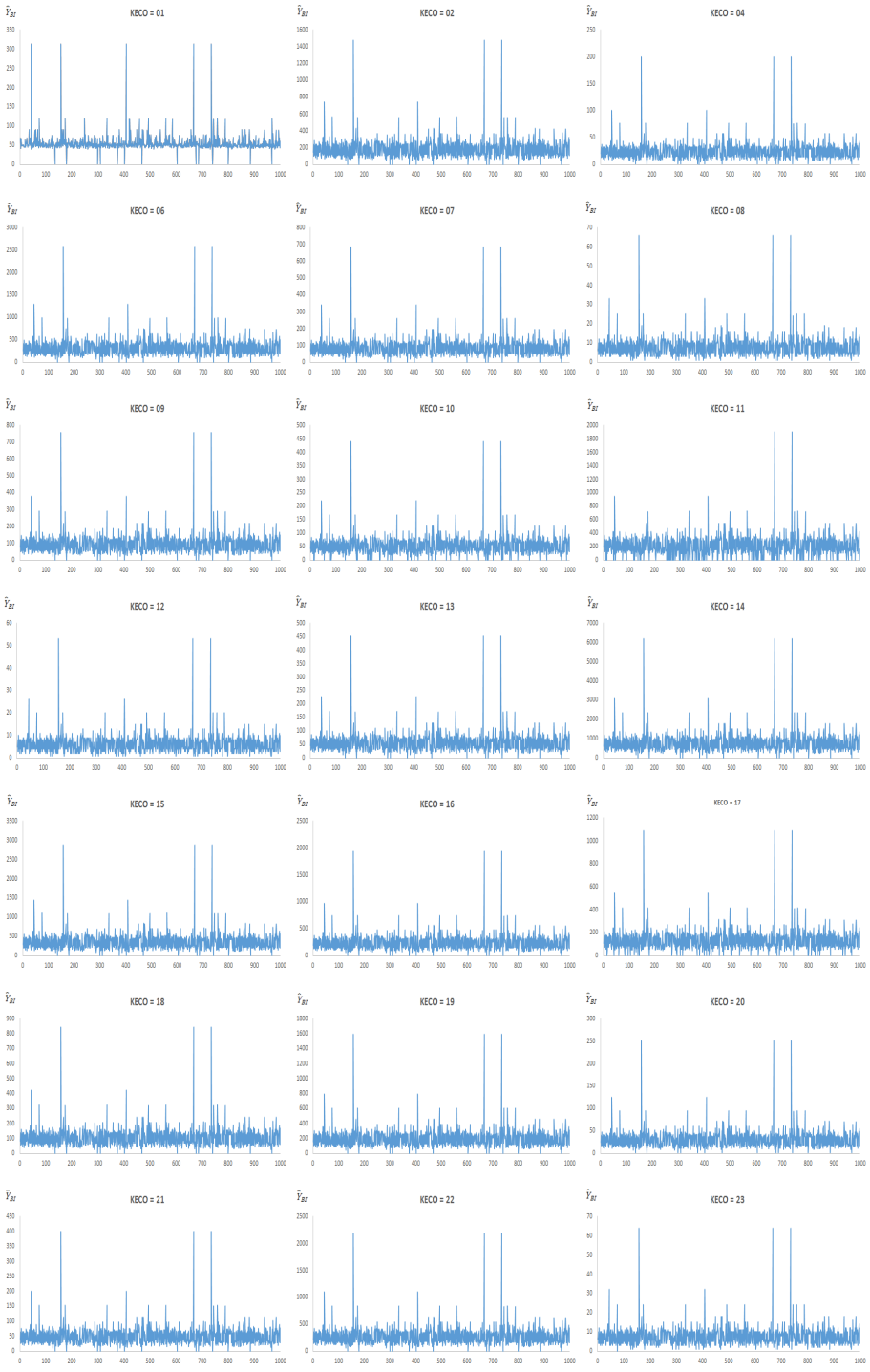
〈표 4.4〉 직종 i 별 α 와 β에 따른 $\widehat{Y_{BI}}$ 추정 결과

KECO(i)	$\alpha=1$					
	$\beta=0.2$	$\beta=0.5$	$\beta=0.7$	$\beta=1$	$\beta=3$	$\beta=5$
01	26.558	44.067	49.932	54.507	27.514	12.309
02	86.070	157.453	178.750	186.532	68.842	13.937
04	13.918	22.925	24.481	24.791	5.938	0.617
06	130.603	252.011	296.614	326.622	173.275	51.214
07	50.097	82.393	85.972	86.083	18.263	1.889
08	3.897	7.172	7.841	7.895	1.801	0.030
09	37.120	72.601	86.020	95.600	53.648	16.608
10	46.344	53.444	53.486	53.486	3.730	0.109

11	180.588	213.11	214.064	214.104	18.479	1.275
12	3.835	6.283	6.299	6.299	0.611	0.015
13	34.412	56.32	56.762	56.762	10.172	0.588
14	479.466	774.689	781.293	781.293	143.097	14.373
15	170.535	309.795	350.359	363.605	129.774	25.683
16	115.766	209.811	236.625	245.091	85.981	16.677
17	55.734	103.558	120.936	134.284	78.892	30.297
18	91.573	106.192	106.192	106.192	7.967	0.059
19	99.869	177.437	196.799	200.742	62.270	10.691
20	16.155	28.391	30.908	31.276	8.428	0.964
21	31.995	50.046	50.211	50.211	7.942	0.330
22	182.961	275.986	276.833	276.833	42.891	3.457
23	3.935	7.037	7.565	7.641	1.622	0.031

〈표 4-4〉의 결과를 살펴보면 주어진 초모수를 이용한 베타분포의 평균과 분산이 작은 경우에는 수요 추정값이 작아지게 된다. 결국 수요조사 결과에서 수요 요구가 있는 한 이를 고려하는 것이 바람직하기 때문에 본 연구에서는 사전확률함수로 $\alpha=1, \beta=1$인 $P_i \sim Beta(1,1)$, 즉 무정보적 사전분포(non-informative prior)를 사용할 것을 제안한다.

다음의 〈그림 4.1〉과 〈그림 4.2〉는 $R=1000$의 모의실험에서 $\hat{Y}_{BT}$와 $\alpha=1$, $\beta=1$인 경우의 $\hat{Y}_{BI}$ 흔적(history) 결과이다.

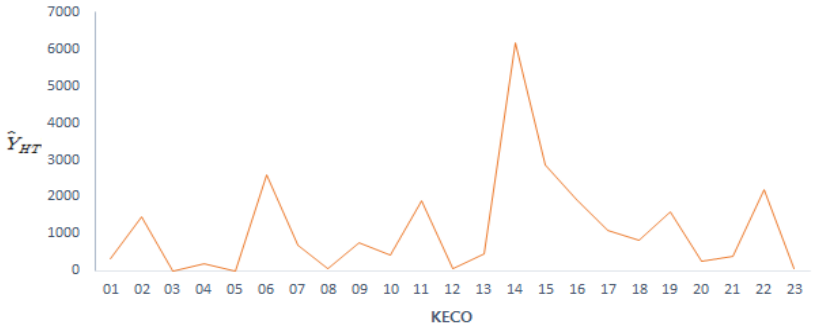


〈그림 4.1〉 KECO별 $\hat{Y}_{BI}$의 1000번의 모의실험 흔적

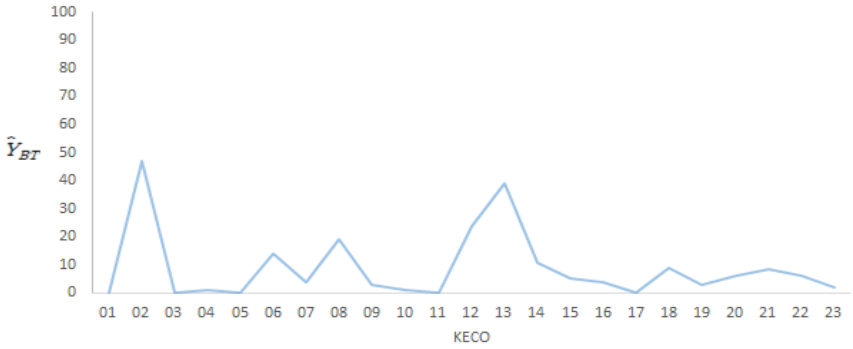
〈그림 4.2〉 KECO별 $\hat{Y}_{BT}$의 1000번의 모의실험 흔적

〈그림 4.1〉과 〈그림 4.2〉를 살펴보면 1,000번의 모의실험 흔적에서 $\hat{Y}_{BI}$의 경우 추정량의 변동을 볼 수 있는 반면 $\hat{Y}_{BT}$는 추정량의 변동을 거의 볼 수 없다. 이는 훈련자수의 요약자료로 사후확률은 정하는 $\hat{Y}_{BT}$는 1,000번의 모의실험에서의 사후확률의 변동이 거의 없게 된다. 반면 훈련자의 개별 자료 역할을 할 수 있는 변수에 분포도를 적용한 $\hat{Y}_{BI}$의 경우 1,000번의 모의실험에서 사후확률의 변동을 볼 수 있게 된다.

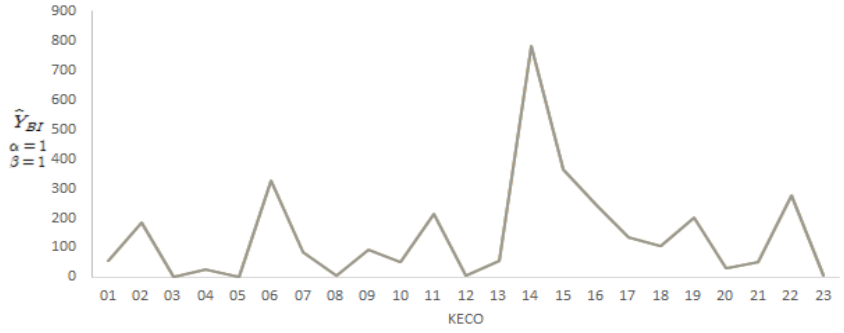
다음은 〈그림 4.3〉에서 〈그림 4.5〉는 $\hat{Y}_{HT}$, $\hat{Y}_{BT}$ 그리고 $\hat{Y}_{BI}$의 추정 결과이다.



〈그림 4.3〉 KECO별 $\hat{Y}_{HT}$



〈그림 4.4〉 KECO별 $\hat{Y}_{BT}$



〈그림 4.5〉 KECO별 $\hat{Y}_{BI}$

〈그림 4.3〉은 $\hat{Y}_{HT}$는 비록 과대 추정된 값이기는 하나 조사시점에서 각 직종별 훈련수요의 분포를 보여 준다. 따라서 〈그림 4.3〉과 〈그림 4.5〉의 분포가 〈그림 4.4〉의 분포보다 흡사하게 나타나고 있다. 따라서 본 연구에서 제안한 $\hat{Y}_{BI}$는 실제 자료를 잘 설명하면서도 과소추정을 방지할 수 있는 합리적인 방법이다.

5. 요약 및 결론

본 연구에서는 수요 추정을 위해 베이즈 추론을 제안하였으며 일반 추정법인 $\hat{Y}_{HT}$와 베이즈 정리를 이용한 결과인 $\hat{Y}_{BT}$ 그리고 베이즈 추론 $\hat{Y}_{BI}$을 이용한 수요 추정 결과를 비교하였다. 결과를 살펴보면 $\hat{Y}_{HT}$의 경우 KECO, 01의 경우 313명이 얻어졌으며 $\hat{Y}_{BT}$의 경우 '0'명이 얻어졌다. 이 두 추정결과는 모두 현실과 맞지 않는다. 그러나 본 연구에서 제안한 방법인 $\hat{Y}_{BI}$의 경우 54명이 예측되었다. 따라서 다른 두 방법에 비해 매우 현실적인 결과가 얻어졌다고 판단된다. A 지역의 합계를 살펴보더라도 $\hat{Y}_{HT}$의 경우 26,374명으로 추정되었으나 $\hat{Y}_{BT}$의 경우 207명으로 두 추정값 모두 현실적으로 타당하다고 할 수 없다. 반면 $\hat{Y}_{BI}$의 경우 3,309명으로 추정되어 현실적으로 타당한 결과를 준다. 이는 변수 중심의 베이즈 추론을 이용하는 경우는 요약자료에 분포도를 가정함으로써 추정량의 분표를 볼 수 반면 $\hat{Y}_{BT}$의 경우 사후확률은 자료값만을 중심으로 계산되어 추정량의 변동의 거의 일어나지 않는다. 따라서 제공된 자료가 요약자료로 사전정보의 분포도가 존재하지 않더라도 타당한 분포 가정을 통해 보다 합리적으로 추정하는 것이 바람직하다. 따라서 본 사업은 계속되기 때문에 향후 분석에서는 $\hat{Y}_{BI}$를 사용하는 것이 타당하다고 판단된다. 다만 어떤 사전분포를 사용하는 것이 현실적인 결과를 줄 수 있는지와 관련된 추가 연구가 필요하다. 하지만 현재까지 얻어진 결과를 종합해 보면 무정보적 사전분포를 사용하는 것이 가장 타당하다고 판단된다.

(2016년 9월 7일 접수, 2016년 10월 14일 수정, 2016년 11월 25일 채택)

참고문헌

김달호 (2013). R과 Winbugs을 이용한 베이지안 통계학, 자유아카데미.

최영섭, 양정승 (2015). 훈련수요조사 결과 가이드, 한국직업능력개발원.

Hamner, M. S., Seaman, J. W. and Young, D.M. (2001). Bayesian Methods in Finite Population Sampling. Proceedings of the Annual Meeting of the American Statistical Association, August 5-9.

Hamner, M.S., Seaman, J. W. and Young, D.M. (2003). A Bayes and Empirica Bayes Prediction for a Finite Population total usign Auxiliary Information. Joint Statistical Meeting-section on Survey Research Methods, 1731-1738.

Little, Roderick(2015). Bayesian Inference for sample survey, Journal of Official Statistics - Anniversary Conference 2015.

Pfeffermann, Danny., Krieger., Abba, M. and Rinott, Yosef. (1998). Parametric Distributions of Complex Survey data Under Informative Probability Sampling. Statistics Sinica 8, 1087-1114

Rao, J, K, N. (2011). Impact of Frequentist and Bayesian Methods on Survey sampling Practice: A Selective Appraisal. Statistical Science, Vol. 26, No. 2, 240-256.

Bayesian Approach in Sample Survey[1)]

Sung Taek Kang[2)] · Sang Eun Lee[3)]

Abstract

In sample survey, mostly H-T method is used for estimation. Recently Bayesian inferences are imposed to various kind of sample surveys. For example, labour demands are estimated from labour demands survey. In general demand survey results are overestimated. For adjusting the overestimated results, Bayes theory is applied. For this study we use the training demand survey data and use the data from korea Employment Information Service as a prior information. In this study Bayes theory and Bayesian inference are used for estimating the training demands and compared the two methods.

Key words : Bayes Theory, Bayesian Inference , Expected Initial Training

1) This work was supported by Kyonggi University Research Grand 2015.
2) Graduate Student , Dept. of Applied Statistics, Kyonggi University, 154-42 Gwanggyosan-Ro Yeongton-Gu, Suwon-Si, Gyeonggi-Do, Korea. E-mail : pooooo100@naver.com
3) (corresponding author) Professor, Dept. of Applied Statistics, Kyonggi University, 154-42 Gwanggyosan-Ro Yeongton-Gu, Suwon-Si, Gyeonggi-Do, Korea. E-mail: sanglee62@kgu.ac.kr.