응용통계연구 제 12권 2호, 1999년, pp. 585-591

반복된 확률화 응답모형에서 일관성 없는 응답에 대한 검정*

이관제 [1)]

요 약

Warner(1965)의 확률화 응답 모형을 두번 연속사용하여 응답자들이 일관성 있는 응답을 했다는 가설을 검정하는 검정통계량을 제안했다. 이것은 양측과 단측 대립가설 모두 검정하는데 이용할 수 있으며, 제안된 검정통계량의 조건분포는 정규분포에 근사한다. 이 검정통계량의 조건부 검정력 함수와 비조건부 검정력 함수를 구하였다.

1. 서론

Warner(1965)에 의해 개발된 확률화 응답과정은 민감한 질문의 조사에서 생기는 응답오차를 줄이기 위해 이용되어왔다. 확률화 응답기법의 기본 개념은 응답자의 신분을 노출시키지 않고 질문에 대한 정보를 얻기 위하여 개별응답과 민감한 질문사이에 확률장치를 이용하는 것이다.

Warner(1965)가 확률화 응답 기법을 제안한 이후, 많은 논문들이 이 기법을 수정, 확장, 그리고 발전시켜왔다. Abul-Ela et al.(1967)은 확률화 응답기법을 다항 모집단으로 확장한 방법을 제안했고, Greenberg et al.(1969)는 민감한 질문과 무관한 민감하지 않은 질문(unrelated unsensitive question)을 이용하여 Warner 과정을 변형한 방법을 제안했다. 그리고 Mangat와 Singh(1990) 그리고 Mangat(1994)는 단순화한 확률화 응답모형을 제안하였다.

이와 같이 수정, 확장, 그리고 발전된 확률화 응답과정의 대부분은 각 응답자가 민감한 범주에 속하는 확률의 추정에 관심을 두고, 이 확률의 비편향 추정값와 분산을 구한다. 그러나, Warner(1965)는 응답자들이 신뢰할 수 있게 답할 때만 비편향추정치를 얻을 수 있다고 했다. 즉, 응답자들이 신뢰할 수 없게 답했을 때 더 이상 비편향 추정값를 얻을 수 없다. Mangat와 Singh(1990)은 Warner가 제안한 확률장치보다 더 효율적인 두 확률장치를 사용하는 방법을 제안하였으며, 또한 응답자들이 신뢰할 수 없는 응답을 했을 때 추정값의 편향와 최소제곱 오차를 구하였다.

Krishnamoorphy와 Raghavarao(1993)는 Warner 과정을 두번 이용하여 응답자가 신뢰할 수 있는 응답을 했다는 가설을 검정하는 방법을 제안하였다. 이는 점근적 정규성을 이용한 방법이며, Lakshimi와 Raghavarao(1992)는 같은 조건에서 점근적 카이자승 검정을 제안했다. 본 논문에서는 Krishnamoorphy와 Raghavarao(1993)에 의하여 고려된 조건과 유사하나, 두 확률화 응답이 독립이라는 가정없이 조건부 점근검정을 제안한다.

* 본 연구는 동국대학교 전문학술지 논문게재연구비 지원으로 이루어졌음.

1) (100-715) 서울시 중구 필동 3가 26, 동국대학교 통계학과, 부교수

2. 확률화 응답(RANDOMIZED RESPONSE)

연구 모집단에 있는 각 응답자는 민감한 범주(C)와 민감하지 않은 범주($\bar{C}$)중 하나에 속해 있으며 모집단에서 민감한 범주에 속해 있는 사람들의 비율을 π라 하자. 이 모집단으로부터 복원추출 방법으로 표본의 크기가 n인 확률표본을 얻는다. 각 응답자에 두 확률장치 R_1과 R_2가 주어지고, 각 확률장치에는 (1) '나는 범주 C에 속한다'와 (2) '나는 범주 $\bar{C}$에 속한다'의 두 질문이 있다. 두 확률장치에서 민감한 질문 (1)이 선택될 확률은 모두 p이다. 여기서 $p \neq 1/2$로 가정한다. 선택된 질문이 어느 것인지 밝히지 않고 응답자들은 R_1과 R_2에 대해 '예' 또는 '아니오'의 대답을 할 것이다. 이 때 응답자들은 민감한 질문에 항상 진실한 응답을 하지는 않을 것이다. 본 논문에서는 응답자가 민감하지 않은 범주에 속할 때는 거짓 응답을 할 확률은 0이라고 가정한다. 두번의 확률화 응답에서 확률변수 X와 Y를 다음과 같이 정의한다. i번째 응답자가 R_1에 대해 '예' 또는 '아니오'라고 응답할 때 X_i는 각각 1 또는 0으로 정의하고, i번째 응답자가 R_2에 대해 '예' 또는 '아니오'라고 응답할 때 Y_i는 각각 1 또는 0으로 정의한다. 한 응답자가 두번 응답함으로 X와 Y는 독립이라는 가정을 하지 않는다. 예를 들면, $p = 0.1$이고 $n = 100$이라 하면 확률장치 R_1에서 약 90명은 질문(2)에 응답하게 되고, 또한 확률장치 R_2에서도 90여명이 질문(2)에 답하게 됨으로 약 80여명이 같은 문항에 답하게 되어 신뢰할 수 있는 응답을 얻는 경우에 변수 X와 Y사이에 상관(correlation)이 존재하게 된다. 상관이 존재하면 두 변수사이에 종속관계가 성립한다. 두 변수가 독립이라는 가정이 없으면 Krishnamoorphy와 Raghavarao(1993)와 Lakshimi 와 Raghavarao(1992)의 (2.3)처럼 확률 $Pr(X_i = 1, Y_i = 0)$ 과 $Pr(X_i = 0, Y_i = 1)$을 계산할 수 없으며, 두편의 논문에서 제안한 검정통계량을 사용하기 어렵게 된다. T_1을 확률장치 R_1에서 신뢰할 수 있는 응답을 얻을 확률이라 하고 T_2를 확률장치 R_2에서 신뢰할 수 있는 응답을 얻을 확률이라 한다. 그러면 확률장치 R_1에서 '예'라는 답을 얻을 확률은

$$Pr(X = 1) = T_1 p\pi + (1-p)(1-\pi) + (1-T_1)(1-p)\pi \tag{2.1}$$

이며 또한, 확률장치 R_2에서 '예'라는 답을 얻을 확률은

$$Pr(Y = 1) = T_2 p\pi + (1-p)(1-\pi) + (1-T_2)(1-p)\pi \tag{2.2}$$

이다.

3. 일관성 없는 답에 대한 검정

표본의 크기가 n인 확률표본의 각 응답자에 대하여 두번의 확률장치가 주어진다. 이때 가능한 결과들은 $(X = 1, Y = 1)$, $(X = 1, Y = 0)$, $(X = 0, Y = 1)$, 그리고 $(X = 0, Y = 0)$이며 이들의 확률응답빈도변수는 N_{11}, N_{10}, N_{01}, 그리고 N_{00}이라 한다. 이들의 각 실제관측치(realization)를 n_{11}, n_{10},n_{01}, 그리고 n_{00}이라고 하자. 그러면 자료는 다음과 같은 분할표로 표현된다.

$X\backslash Y$	1	0	Total
1	N_{11}	N_{10}	N_{1+}
0	N_{01}	N_{00}	N_{0+}
Total	N_{+1}	N_{+0}	n

(X,Y)의 결합분포의 모수와 X와 Y의 주변분포의 모수는 다음과 같이 나타낼 수 있다.

$$\lambda_{fg} \stackrel{\text{def}}{=} Pr(X=f, Y=g) \tag{3.1}$$

$$\lambda_{1+} \stackrel{\text{def}}{=} Pr(X=1) = \lambda_{11} + \lambda_{10}, \tag{3.2}$$

그리고

$$\lambda_{+1} \stackrel{\text{def}}{=} Pr(Y=1) = \lambda_{11} + \lambda_{01}. \tag{3.3}$$

여기서 확률장치 R_1에서 '예'라는 답을 얻을 확률과 확률장치 R_2에서 '예'라는 답을 얻을 확률이 같다는 가설을 고려하여 보자. 즉, 확률장치 R_1과 확률장치 R_2에서의 일관성있는 답을 한다는 가설이다. 이 가설은 다음과 같이 쓸 수 있다.

$$Pr(X=1) = Pr(Y=1)$$

또한 (2.1)식과 (2.2)식으로 부터, 이 가설은 $p \neq 1/2$라 가정하였으므로 $T_1 = T_2$를 의미한다. 즉, 확률장치 R_1과 확률장치 R_2에서의 신뢰할 수 있는 응답을 할 확률이 같다는 의미이다. 그러므로, 이 가설의 채택은 아래 세 경우중 한 경우를 받아들이는 것이다: (i) $T_1 = T_2 = 0$, (ii) $T_1 = T_2 = 1$, 또는 (iii) $T_1 = T_2 = 0.7$(예를 들면). 여기서, 경우 (i)는 현실적이지 못함으로 제외하여도 해석상의 문제를 야기치 않는다. 만일 이 가설이 기각될 경우에는 경우 (ii) 또는 (iii)이 기각되며, 어느 것이 기각되거나 공통적인 것은 두 번의 확률장치중에서 적어도 한번의 확률장치에서는 신뢰할 수 있는 응답에 대한 확률이 1이 아니라는 것이다. 다시 말하면, 상기 가설의 기각은 두 번의 확률장치에서 적어도 한번은 신뢰할 수 있는 응답을 갖지 못하였다는 증거이다. 그러므로 상기 가설는 신뢰할 수 있는 응답에 대한 가설검정을 가능하게 한다. 이와 같은 상기 가설에 대한 해석에 기반을 두고, (3.2)과 (3.3)식을 이용하여 귀무가설과 양측 대립가설을 다음과 같이 설정한다.

$$H_o : \lambda_{1+} = \lambda_{+1} \ , \ H_{11} : \lambda_{1+} \neq \lambda_{+1}.$$

그리고 단측 귀무가설을 다음과 같이 나타낼 수 있다.

$$H_{12} : \lambda_{1+} > \lambda_{+1} \ , \ H_{13} : \lambda_{1+} < \lambda_{+1}.$$

이 귀무가설과 양측 대립가설을 (3.1)을 이용하여 다음과 같이 표현할 수 있다.

$$H_o : \lambda_{10} = \lambda_{01} \ , \ H_{11} : \lambda_{10} \neq \lambda_{01},$$

그리고 단측 귀무가설은 다음과 같이 나타낼 수 있다.

$$H_{12} : \lambda_{10} > \lambda_{01} \ , \ H_{13} : \lambda_{10} < \lambda_{01}.$$

귀무가설이 사실일 때 확률화 장치하에서 관측수 n_{10}과 n_{01}에 대해 대략 같은 빈도를 기대한다. $m = n_{10} + n_{01}$을 주대각 칸(cell)을 제외한 전체관측수라 하자. 즉 확률빈도 $M = N_{10} + N_{01}$에 대한 실제관측수를 m이라고 가정한다. 이들 두 칸에 대한 그들의 할당은 m번 시행한 이항변량의 결과이다. 귀무가설하에서 이들 m관측치의 각각은 n_{10}에 들어갈 기회가 1/2이고 n_{01}에 들어갈 기회가 1/2이다. m이 클 때, 이항분포의 정규근사법(normal approximation to binomial)에 의해 이항분포는 평균 $m/2$이고 분산이 $m/4$인 정규분포와 유사한 형태를 가진다. 즉, N_{10}는 이항분포 $b(m, 1/2)$를 따르므로, m이 충분히 클 때,

$$\frac{N_{10} - \frac{1}{2}m}{\sqrt{\frac{1}{4}m}}$$

의 분포는 정규분포로 근사해 간다.

이와 같은 사실에 근거해서 상기 귀무가설에 대한 대표본 검정의 통계량을 다음과 같이 설정할 수 있다.

$$Z = \frac{N_{10} - N_{01}}{\sqrt{N_{10} + N_{01}}}.$$

4. 점근적 조건부 검정력 함수

$M = m$이 주어진 조건하에서의 Z의 조건부분포 $f(Z|m)$는 평균 $\sqrt{m}(\lambda_{10} - \lambda_{01})/(\lambda_{10} + \lambda_{01})$과 분산 $4\lambda_{10}\lambda_{01}/(\lambda_{10} + \lambda_{01})^2$을 갖는 점근적 정규분포를 한다(부록 참조). 다음과 같은 식을 고려해보자.

$$\delta = \lambda_{10} - \lambda_{01},$$

$$\psi = \lambda_{10} + \lambda_{01}.$$

그러면 다음과 같이 나타낼 수 있다.

$$\lambda_{10}\lambda_{01} = \frac{1}{4}(\psi^2 - \delta^2).$$

이러한 표현을 사용하면 $f(Z|m)$은 평균 $\sqrt{m}\delta/\psi$과 분산 $(\psi^2 - \delta^2)/\psi^2$을 갖는 점근적 정규분포를 한다. 위의 결과로부터 양측검정의 점근적 조건부 검정력은 다음과 같다.(부록 참조)

$$\beta(\delta|m) = \Phi[\frac{-z_{\alpha/2}\psi + \sqrt{m}|\delta|}{\sqrt{(\psi^2 - \delta^2)}}|m] + \Phi[\frac{-z_{\alpha/2}\psi - \sqrt{m}|\delta|}{\sqrt{(\psi^2 - \delta^2)}}|m].$$

위의 결과는 관측치 m이 알려졌을 때 검정력 평가에 이용된다. 비조건부 검정력은 이중기대값이론(double expectation theorem)에 의해 얻어진다. 즉, 조건부 검정력의 기대값 $\beta(\delta) = E_M[\beta(\delta|M)]$이다. 그러면 양측검정에 대한 비조건부 검정력 함수는 다음과 같다.

$$\beta(\delta) = \Phi[\frac{-z_{\alpha/2}\psi + \sqrt{n\psi}|\delta|}{\sqrt{(\psi^2 - \delta^2)}}] + \Phi[\frac{-z_{\alpha/2}\psi - \sqrt{n\psi}|\delta|}{\sqrt{(\psi^2 - \delta^2)}}].$$

단측검정에대한 조건부 검정력 함수와 비조건부 검정력 함수는 유사한 방법으로 쉽게 얻어 질 수 있다. 이 검정력 함수를 이용하여 검정력을 구할 때는 대립가설에서 δ의 값이 주어지므로 ψ의 최대우도추정량을 사용하여 검정력을 구할 수 있다.

5. 결론

각 응답자에게 두번의 Warner의 과정을 적용할 때 Krishnamoorphy와 Raghavarao(1993)와 Lakshimi와 Raghavarao(1992)는 두 확률장치로부터 반복측정의 결과를 독립이고 각 응답자의 신뢰할 수 없는 응답의 확률이 모두 같다는 조건하에 응답자의 신뢰할 수 없는 응답을 관측하는 과정을 개발했다. 제안된 검정은 두 번의 응답이 독립적이라는 조건없이, 두 번의 연속된 질문에 대한 응답이 일관되지 않았다는 대립가설이 채택되는 경우에 간접적으로 신뢰할 수 없는 응답에 대한 검정을 할 수 있다.

참고문헌

[1] Abul-Ela, A.A., Greenberg,B.G., and Horvits, D.G.(1967) A Multi-Proportions Randomized Response Model, *Journal of the American Statistical Association*, 62, 990-1000.

[2] Greenberg, B.G., Abul-Ela, A.A., Simmons,W.R., and Horvitz,D.G.(1969) The Unrelated Question Randomized Reponse Model: theoretical Framwork, *Journal of the American Statistical Association*, 64, 520-539.

[3] Mangat, N.S. and Singh, R. (1990) An alternative randomized response procedure. *Biometrika*, 77,439-442.

[4] Mangat, N.S.(1994) An Improved Randomized Response Strategy, *Journal of the Royal Statitical Society. SeriesB*, 56, 93-95.

[5] Krishnamoorphy, K. and Raghavarao, D.(1993) Untruthful answering in repeated randomized response procedures, *The Cannadian Journal of Statistics*, 22, No.2, 233-236.

[6] Lakshmi, D. and Raghavarao, D.(1992) A test for detecting untruthful answering in randomized response procedures, *Journal of Statistical Planning and Inference*, 31, 387-390.

[7] Warner, S.L. (1965) Randomized response: a survey technique for eliminating evasive answer bias, *Journal of the American Statistical Association*, 60, 63-69.

[1999년 3월 접수, 1999년 8월 최종수정]

부록

$Z|m$은 평균 $\sqrt{m}\delta/\psi$과 분산 $(\psi^2-\delta^2)/\psi^2$을 갖는 점근적 정규분포를 하는 것을 보인다. 점근적 조건부 평균과 분산은 다음과 같이 얻어진다.

$$\begin{aligned} E(Z|m) &= E(\frac{N_{10}-N_{01}}{\sqrt{m}}|m) \\ &= \frac{1}{\sqrt{m}}E(N_{10}-N_{01}|m) \\ &= \frac{1}{\sqrt{m}}(\frac{m\lambda_{10}}{\lambda_{10}+\lambda_{01}}-\frac{m\lambda_{01}}{\lambda_{10}+\lambda_{01}}) \\ &= \frac{\sqrt{m}(\lambda_{10}-\lambda_{01})}{\lambda_{10}+\lambda_{01}} \end{aligned}$$

$$\begin{aligned} Var(Z|m) &= Var(\frac{N_{10}-N_{01}}{\sqrt{m}}|m) \\ &= \frac{1}{m}\{Var(N_{10}|m)+Var(N_{01}|m)-2Cov(N_{10},N_{01})\} \\ &= \frac{1}{m}\{\frac{m\lambda_{10}\lambda_{01}}{(\lambda_{10}+\lambda_{01})^2}\}+\frac{1}{m}\{\frac{m\lambda_{01}\lambda_{10}}{(\lambda_{10}+\lambda_{01})^2}\}+2\frac{1}{m}\{\frac{m\lambda_{10}\lambda_{01}}{(\lambda_{10}+\lambda_{01})^2}\} \\ &= 4\frac{\lambda_{10}\lambda_{01}}{(\lambda_{10}+\lambda_{01})^2} \end{aligned}$$

$\delta \neq 0$에 대한 조건부 검정력 함수 $\beta(\delta|m)$는 다음과 같다.

$$\begin{aligned} \beta(\delta|m) &= Pr_\delta(|Z| >= Z_{\alpha/2}|m) \\ &= Pr_\delta(Z <= -Z_{\alpha/2}|m)+Pr_\delta(Z >= Z_{\alpha/2}|m) \\ &= Pr_\delta(\mathcal{N} <= \frac{-Z_{\alpha/2}-\sqrt{m}|\delta|/\psi}{\sqrt{\psi^2-\delta^2}/\psi}|m)+Pr_\delta(\mathcal{N} >= \frac{Z_{\alpha/2}-\sqrt{m}|\delta|/\psi}{\sqrt{\psi^2-\delta^2}/\psi}|m) \\ &= Pr_\delta(\mathcal{N} <= \frac{-Z_{\alpha/2}-\sqrt{m}|\delta|/\psi}{\sqrt{\psi^2-\delta^2}/\psi}|m)+Pr_\delta(\mathcal{N} <= \frac{-Z_{\alpha/2}+\sqrt{m}|\delta|/\psi}{\sqrt{\psi^2-\delta^2}/\psi}|m) \\ &= \Phi(\frac{-Z_{\alpha/2}\psi-\sqrt{m}|\delta|}{\sqrt{\psi^2-\delta^2}}|m)+\Phi(\frac{-Z_{\alpha/2}\psi+\sqrt{m}|\delta|}{\sqrt{\psi^2-\delta^2}}|m), \end{aligned}$$

여기서 $\mathcal{N}$ 표준정규분포를 따르는 변량이다.

A test for detecting consistent answering in repeated randomized response model*

Kwan Jeh Lee[1)]

ABSTRACT

In the repeated response setting of Warner(1965), under some proper conditions a procedure is given to test the hypothesis that the respondents are giving truthful answers. Both two-sided and one-sided alternative hypotheses are avaiable in the testing procedure. The asymptotic conditional distribution of the suggested test statistic is normal. The asymptotic conditional and unconditional power functions are founded.

* This research is supported by the Dongguk Univerisity Research Fund.

1) Associate Professor,Department of statistics, Dongguk University, 26 3-Ga Pil-Dong Choong-Gu Seoul 100-715 Korea.