

Rastgeleleştirilmiş Cevap Modelinde Oran Tahmini

Sevil Bacanlı

Hacettepe Üniversitesi, İstatistik Bölümü,
06800, Beytepe, Ankara, Türkiye
sevil@hacettepe.edu.tr

Tuğçe Tuncel

Sosyal Güvenlik Kurumu, Teknoloji İş
Geliştirme Dairesi Başkanlığı Balgat,
Ankara, Türkiye
tugcetuncel@gmail.com

Özet

Bu çalışmada, rastgele cevap modelinde oran tahmini basit ve tabakalı rastgele örnekleme için incelenmiştir. Modelin uygulanmasını göstermek amacıyla sayısal örnek verilmiştir.

Anahtar sözcükler: Rasgele cevap modeli, Warner-rasgele cevap modeli

Abstract

Randomized Response Model for Proportion

In this study, randomized response model for proportion is examined in simple and stratified random sampling. Numerical example is given to illustrate the application of the model.

Key words: *Randomized response model, Warner's randomized response model*

1.Giriş

Rastgeleleştirilmiş cevap (RC), modeli cevap verenin kimliğini açığa çıkarmadan hassas konular hakkında bilgi toplamak için kullanılmaktadır. Hassas konuların katılımcılara doğrudan sorulması, cevap vermeyi reddetme ya da uydurma cevap verme şeklinde sonuçlanmaktadır.

Sosyal araştırmalarda özellikle insanların hassas oldukları kişisel soruların yer aldığı araştırmalarda kişilerin cevap vermeyi reddetmeleri, uydurma ya da yanıltıcı cevap vermeleri gibi problemler ortaya çıkmaktadır. Geleneksel yöntemlerle yapılan araştırmalarda ortaya çıkan bu problemleri önlemek araştırmacıların kişilerin güvenini kazanmasıyla sağlanmaya çalışılmıştır. Ancak bu konuda başarılı sonuçlar elde edilememektedir. Warner[1] çalışmasında, sosyal olarak hoş karşılanmayan, suçlayıcı davranışsal sorularda cevap hatasını azaltmak, cevap verenin gizliliğini korumak ve cevap oranını artırmak için alternatif bir araştırma yöntemi olan RC modelini basit rastgele örnekleme (BRÖ) için geliştirmiştir. Bu yöntem ile hassas özelliğe sahip kitle oranı tahmin edilebilmektedir.

Rastgeleleştirme önceden hazırlanmış bir rastgeleleştirme cihazı ile sağlanmaktadır. Böylelikle cevap veren kişinin istediği gizlilik ve yüz yüze durumda özgürce cevap verebilmesi sağlanmaktadır.

Hong K. et al.[2], Warner-RC modelini tabakalı rastgele örneklemede (TRÖ) her tabakaya aynı, Kim-Warde[3] her tabakaya farklı rastgeleştirme cihazlarının uygulandığı tabakalı RC modelini önermiştir. Kim-Elam[4] ise her tabakaya farklı rastgeleştirme cihazlarının uygulandığı iki aşamalı tabakalı RC modelini önermiştir. Bacanlı ve Tuncel[5] ise sonradan tabakalamalı örnekleme yönteminde RC modelini incelemiştir.

Bu çalışmada, RC modellerinde oran tahmini BRÖ ve TRÖ için incelenmiştir. Ayrıca, Türkiye Nüfus ve Sağlık Araştırması; doğurganlık düzeyi ve değişimi, bebek ve çocuk ölümlülüğü, aile planlaması anne ve çocuk sağlığı konularında bilgi sağlamak üzere tasarlanmış ulusal düzeyde örneklem araştırması[6,7] bilgilerinden yararlanılarak ve Hong K. et al.[2], RC modeli ile kadınlarda düşük yapma oranı tahmin edilmiştir[8].

2. Basit Rastgele Örneklemede Warner-RC modeli

Warner[1], sosyal olarak hoş karşılanmayan, suçlayıcı davranışsal sorularda cevap hatasını azaltmak, cevap verenin gizliliğini korumak ve cevap oranını artırmak için alternatif bir araştırma yöntemi olarak RC modeli önermiştir.

Warner-RC modelinde kişiler, kitleden yerine koyarak BRÖ ile çekilmiştir ve hassas özellik “A karakteristiği” olarak tanımlanmıştır. A karakteristiğine sahip olan kişilerin oranı (π) tahmini için bir rastgeleştirme cihazı (Şekil-1) kullanılmıştır.



Şekil 1. Warner modeli rastgeleştirme cihazı

Rastgeleştirme cihazı A karakteristiğine sahip olma durumunu araştıran iki ifadeden oluşmaktadır: 1-) A karakteristiğine sahibim, 2-) A karakteristiğine sahip değilim.

Örneklemedeki her kişiyle görüşmeden önce, her katılımcıya P olasılıkla “A karakteristiğine sahibim.” ve $1-P$ olasılıkla “A karakteristiğine sahip değilim.” ifadelerini gösteren özdeş ibrelere sahip özdeş rastgeleştirme cihazları sağlanmaktadır. Katılımcı, araştırmacı tarafından gözlenilmeden ibreyi çevirmektedir. İbrenin gösterdiği ifadeyi söylemeden ve bu ifadeye bağlı olarak sadece “evet” ya da “hayır” cevabını vermesi istenmektedir. 1. ya da 2. ifadeye “evet” cevabını veren katılımcıların oranı

$$P(X_i = 1) = Z = P\pi + (1 - P)(1 - \pi) \quad (1)$$

şeklinde tanımlanabilmektedir.

Basit rastgele örneklemeyle yerine koyarak çekilen n kişiden n_1 tanesi “evet” cevabını vermiş olsun. Dolayısıyla $n-n_1$ tanesi de “hayır” cevabını vermiş olacaktır. X_i rastgele değişkeni

$$X_i = \begin{cases} 1; & \text{eğer } i. \text{ katılımcı evet cevabını verirse} \\ 0; & \text{d. d.} \end{cases}$$

şeklinde tanımlanır. Bu durumda X_i r.d binom dağılımına sahiptir ve X_i r.d'nin olasılık fonksiyonu

$$L = \frac{n!}{n_1!(n-n_1)!} P^{n_1} (1-P)^{n-n_1} \quad (2)$$

dir. A karakteristiğine sahip katılımcıların oranı (π)' nin olasılık fonksiyonu ise

$$L = \frac{n!}{n_1!(n-n_1)!} [P\pi + (1-P)(1-\pi)]^{n_1} [P(1-\pi) + (1-P)\pi]^{n-n_1} \quad (3)$$

biçiminde tanımlanır. (3) den yararlanılarak (π)' nin tahmini,

$$\hat{\pi} = \frac{\hat{Z} - (1-P)}{2P-1} \quad P \neq 0.5 \quad (4)$$

olarak bulunmaktadır. $\hat{\pi}$ ' nin varyansı ise

$$\begin{aligned} V(\hat{\pi}) &= V\left(\frac{\hat{Z} - (1-P)}{2P-1}\right) \\ &= \frac{P(1-P)}{n(2P-1)^2} + \frac{\pi(1-\pi)}{n} \end{aligned} \quad (5)$$

biçiminde tanımlanmaktadır. Geliştirilen RC modellerinin amacı birinci terimi minimum yapmaktır. Böylece modelin etkinliğinin artması sağlanmaktadır[8,9].

3. Tabakalı Rastgele Örneklemede Warner-RC modeli

3.1. Hong K. v.d. RC Modeli

Hong K. et al.[2], TRÖ'de her bir tabakaya Warner-RC modelinin ve aynı rastgeleştirme cihazının uygulandığı RC modelini önermiştir.

Modelde kitle, N_h genişliğinde l tabakaya ayrılmaktadır. Her tabakadan yerine koyarak basit rastgele örnekleme ile n_h genişliğinde örneklem seçilmektedir. h . tabakadaki her katılımcıya sırasıyla P ve $1-P$ olasılıklarına sahip hassas soru kartı (A) ve bunun negatifi (\bar{A}) ifadelerinden oluşan rastgeleştirme cihazı temin edilmektedir. Katılımcıdan hangi soru kartına sahip olduğunu belirtmeden rastgeleştirme cihazının çıktısına ve kendi gerçek durumuna göre sadece “evet” ya da “hayır” cevabını vermesi istenmektedir.

Warner-RC modelinde (1)' de tanımlanan “evet” cevabını veren katılımcıların oranı h . tabaka için

$$\begin{aligned} Z_h &= P(X = 1)_h \\ &= P\pi_h + (1 - P)(1 - \pi_h) \quad h = 1, \dots, l \end{aligned} \quad (6)$$

biçiminde tanımlanır. Burada

Z_h : h . tabakada “evet” cevabı veren katılımcıların oranı

π_h : h . tabakadaki hassas özelliğe sahip katılımcıların oranı

P : Katılımcının hassas soru kartı (A)'na sahip olma olasılığıdır.

Warner-RC modelinde (4)' de verilen hassas özelliğe sahip katılımcıların oranı h . tabaka için,

$$\hat{\pi}_h = \frac{Z_h - (1 - P)}{2P - 1} \quad P \neq 0.5 \quad (7)$$

olarak hesaplanmaktadır. $\hat{\pi}_h$ 'nin varyansı ise

$$V(\hat{\pi}_h) = \frac{\pi_h(1 - \pi_h)}{n_h} + \frac{P(1 - P)}{n_h(2P - 1)^2} \quad (8)$$

biçiminde tanımlanır. Dolayısıyla, Hong K. et al.[2] TRÖ için önerdiği RC modeli ile kitledeki hassas özelliğe sahip katılımcıların oran tahmini

$$\begin{aligned} \hat{\pi}_{tBH} &= \sum_{h=1}^l W_h \hat{\pi}_h \\ &= \sum_{h=1}^l W_h \left[\frac{Z_h - (1 - P)}{2P - 1} \right] \end{aligned} \quad (9)$$

olarak tanımlanır. $\hat{\pi}_{tBH}$ 'nin varyansı ise

$$\begin{aligned} V(\hat{\pi}_{tBH}) &= \sum_{h=1}^l W_h^2 V(\hat{\pi}_h) \\ &= \sum_{h=1}^l \frac{W_h^2}{n_h} \left(\pi_h(1 - \pi_h) + \frac{P(1 - P)}{(2P - 1)^2} \right) \end{aligned} \quad (10)$$

eşitliği ile verilir.

Hong K. et al.[2], önerdiği RC. modelinde örneklem genişliği (n_h), belirleyebilmek için, Varyans ($V(\hat{\pi}_{tBH})$) maliyet fonksiyonu $c = c_0 + \sum_{h=1}^l c_h n_h$ için en küçük yapmak gerekir. Lagrange çarpanlar yönteminden yararlanılarak, n_h , aşağıdaki gibi elde edilir:

$$\begin{aligned} V(\hat{\pi}_{tBH}) + \lambda \left(\sum_{h=1}^l c_h n_h - c + c_0 \right) &= \sum_{h=1}^l \frac{W_h^2}{n_h} \left[\pi_h(1 - \pi_h) + \frac{P(1 - P)}{(2P - 1)^2} \right] \\ &+ \lambda \left(\sum_{h=1}^l c_h n_h - c + c_0 \right) \end{aligned}$$

ifadesini en küçük yapan λ çarpanının bulunması istenir. Bunun için, n_h ' a göre türev alınıp sıfıra eşitlendiğinde

$$-\frac{W_h^2}{n_h^2} \left[\pi_h(1-\pi_h) + \frac{P(1-P)}{(2P-1)^2} \right] + \lambda c_h = 0 \quad h = 1, \dots, l$$

$$n_h \sqrt{\lambda} = \frac{W_h \left[\pi_h(1-\pi_h) + \frac{P(1-P)}{(2P-1)^2} \right]^{1/2}}{\sqrt{c_h}} \quad (11)$$

elde edilir. l tabaka üzerinden toplam alınırsa

$$n \sqrt{\lambda} = \sum_{h=1}^l \frac{W_h \left[\pi_h(1-\pi_h) + \frac{P(1-P)}{(2P-1)^2} \right]^{1/2}}{\sqrt{c_h}} \quad (12)$$

olarak bulunur. (11) ve (12)' nin taraf tarafa oranlanmasıyla,

$$\frac{n_h}{n} = \frac{N_h \left[\pi_h(1-\pi_h) + \frac{P(1-P)}{(2P-1)^2} \right]^{1/2} / \sqrt{c_h}}{\sum_{h=1}^l N_h \left[\pi_h(1-\pi_h) + \frac{P(1-P)}{(2P-1)^2} \right]^{1/2} / \sqrt{c_h}} \quad (13)$$

biçiminde elde edilir. Tabakalar arası birim başına yapılacak maliyet ve tabaka varyansları farklı değilse $c_1 = c_2 = \dots = c_l = c$ ve $V(\hat{\pi}_1) = V(\hat{\pi}_2) = \dots = V(\hat{\pi}_l) = V(\hat{\pi})$ yazılabileceğinden, tabaka örneklem genişliği, $n_h = n \frac{N_h}{N}$ orantılı dağıtıma dönüşür[2].

3.2. Kim&Warde RC Modeli

Kim&Warde[3], her tabakaya farklı rastgeleleştirme cihazlarının uygulandığı tabakalı R.C. modelinde her tabakadan basit rastgele örnekleme ile n_h hacimli örnekler seçilmektedir. Modelde, h . tabakada “evet” cevabını veren katılımcıların oranı

$$Z_{hKW} = P_h \pi_h + (1 - P_h)(1 - \pi_h) \quad h = 1, \dots, l \quad (14)$$

biçiminde tanımlanır. π_h ' in tahmin edicisi ve varyansı

$$\hat{\pi}_{hKW} = \frac{z_h - (1 - P_h)}{2P_h - 1} \quad P_h \neq 0.5 \quad (15)$$

$$V(\hat{\pi}_{hKW}) = \frac{\pi_h(1-\pi_h)}{n_h} + \frac{P_h(1-P_h)}{n_h(2P_h-1)^2} \quad (16)$$

biçiminde yazılır. Kitledeki hassas özelliğe sahip katılımcıların oranı olarak tanımlanan π_{td} ' nin tahmin edicisi

$$\hat{\pi}_{tbKW} = \sum_{h=1}^l W_h \left[\frac{z_h - (1-P_h)}{2P_h - 1} \right] \quad (17)$$

biçiminde elde edilir. Varyansı ise

$$\begin{aligned} V(\hat{\pi}_{tbKW}) &= \sum_{h=1}^l W_h^2 V(\hat{\pi}_h) \\ &= \sum_{h=1}^l \frac{W_h^2}{n_h} \left[\pi_h(1-\pi_h) + \frac{P_h(1-P_h)}{(2P_h-1)^2} \right] \end{aligned} \quad (18)$$

eşitliği ile hesaplanır[3]. Kim&Warde- R.C. modelinde ise n_h örneklem genişliği, Hong v.d –RC modelinde olduğu gibi elde edilir. Varyans eşitliğinin tanımı değiştiğinden ,

$$\frac{n_h}{n} = \frac{N_h \left[\pi_h(1-\pi_h) + \frac{P_h(1-P_h)}{(2P_h-1)^2} \right]^{1/2} / \sqrt{c_h}}{\sum_{h=1}^l N_h \left[\pi_h(1-\pi_h) + \frac{P_h(1-P_h)}{(2P_h-1)^2} \right]^{1/2} / \sqrt{c_h}} \quad (19)$$

olarak elde edilir.

3.3. Kim&Elam RC Modeli

Kim-Elam[4], her tabakaya farklı rastgeleleştirme cihazlarının uygulandığı iki aşamalı tabakalı RC modelini önermiştir.

Verilen “evet” , “hayır” cevaplarının doğru olduğu varsayımı altında ve M_h ve P_h olasılıkları araştırmacı tarafından belirlenmek üzere h . tabakada “evet” cevabını veren katılımcıların oranı

$$z_h = M_h \pi_h + (1 - M_h) [P_h \pi_h + (1 - P_h)(1 - \pi_h)] \quad h = 1, \dots, l \quad (20)$$

eşitliği ile ifade edilmektedir. Burada,

M_h : Birinci aşamada h . tabakadan seçilen örneklemdeki bir katılımcının hassas soru kartına sahip olma olasılığı

P_h : İkinci aşamada h . tabakadan seçilen örneklemdeki bir katılımcının hassas soru kartına sahip olma olasılığıdır.

π_h 'in tahmin edicisi ve varyansı

$$\hat{\pi}_{hKE} = \frac{z_h - (1-M_h)(1-P_h)}{2P_h - 1 + 2M_h(1-P_h)} \quad h = 1, 2, \dots, l \quad (21)$$

$$V(\hat{\pi}_{hKE}) = \frac{\pi_h(1-\pi_h)}{n_h} + \frac{(1-M_h)(1-P_h)[1-(1-M_h)(1-P_h)]}{n_h [2P_h - 1 + 2M_h(1-P_h)]^2} \quad (22)$$

biçiminde tanımlanmaktadır. Tüm kitledeki hassas özelliğe sahip katılımcıların oranı olarak tanımlanan π_{tb} 'nin tahmin edicisi ve varyans

$$\hat{\pi}_{tbKE} = \sum_{h=1}^l W_h \hat{\pi}_h$$

$$= \sum_{h=1}^L W_h \left[\frac{Z_h - (1 - M_h)(1 - P_h)}{2P_h - 1 + 2M_h(1 - P_h)} \right] \quad (23)$$

$$V(\hat{\pi}_{\text{TRÖ}}) = \sum_{h=1}^L \frac{W_h^2}{n_h} \left[\pi_h(1 - \pi_h) + \frac{(1 - M_h)(1 - P_h)[1 - (1 - M_h)(1 - P_h)]}{[2P_h - 1 + 2M_h(1 - P_h)]^2} \right] \quad (24)$$

olarak bulunmaktadır. Kim&Elam-R.C. modeli için n_h örneklem genişliği Hong v.d.- RC modelinde olduğu gibi elde edilir. Varyans eşitliğinin tanımı değiştiğinden,

$$\frac{n_h}{n} = \frac{N_h \left[\pi_h(1 - \pi_h) + \frac{(1 - M_h)(1 - P_h)[1 - (1 - M_h)(1 - P_h)]}{[2P_h - 1 + 2M_h(1 - P_h)]^2} \right]^{1/2}}{\sum_{h=1}^L N_h \left[\pi_h(1 - \pi_h) + \frac{(1 - M_h)(1 - P_h)[1 - (1 - M_h)(1 - P_h)]}{[2P_h - 1 + 2M_h(1 - P_h)]^2} \right]^{1/2}} \quad (25)$$

biçiminde elde edilir.

4. Sayısal Örnek

Türkiye Nüfus ve Sağlık Araştırması; doğurganlık düzeyi ve değişimi, bebek ve çocuk ölümlülüğü, aile planlaması ve anne ve çocuk sağlığı konularında bilgi sağlamak üzere tasarlanmış ulusal düzeyde bir örneklem araştırmasından yararlanılarak hassas durumlarla ilgili araştırmalarda kullanılan RC modelinin TRÖ uygulaması için bir örnek verilmiştir. Kadınların isteyerek düşük yapma durumları hassas özellik olarak tanımlanmıştır. Gebeliği önleyici yöntemlere ilişkin hizmetlere ulaşmada sorunlar yaşayan, istenmeyen gebelikten korunmak için yöntem kullanması konusunda psikososyal engelleri olan ve/veya gebeliği önleyici yöntem başarısızlığı nedeniyle gebe kalan kadınlar isteyerek düşüğe başvurumaktadırlar. Tezcan ve Omran[6]'ın çalışmasında rastgeleleştirme cihazından "İsteyerek düşük yaşadım" ifadesinin gelme olasılığı ise $P=0,7$ olarak belirlenmiştir[7,8]. Dolayısıyla, çalışmada konuya ilişkin olarak her bir tabakaya ilişkin P_h ve M_h değerleri bilinmediğinden sadece Hong v.d.- RC modelinin TRÖ için uygulaması verilmiştir.

TNSA-2008[7] raporuna göre Çizelge 1.'de belirtilen her bir yaş grubuna yönelik isteyerek düşük yapma oranları (π_h) kullanılarak TRÖ'de örneklem büyüklüğü 564 olarak elde edilmiş ve tabakalara dağıtılmıştır. Yaş gruplarına göre kitle ve örneklem tabaka büyüklükleri, tabaka ağırlıkları Çizelge 1.'de verilmiştir. Ayrıca tabakalara göre "evet" cevabını veren katılımcıların oranları ve varyansları hesaplanarak Çizelge 1.'de verilmiştir.

Çizelge-1 den, (9) ve (10) eşitliklerinden isteyerek düşük yaşayan kadınların oranı $\hat{\pi}_{\text{TRÖ}} = 0,238$ varyans ise $V(\hat{\pi}_{\text{TRÖ}}) = 0,002560$ olarak hesaplanmıştır.

5. Sonuç

RC modelleri, aile içi şiddet, uyuşturucu kullanımı gibi hassas karaktere sahip konular hakkındaki araştırmalarda katılımcının gizliliğini koruyarak cevap hatalarını azaltmayı amaçlayan bir yöntemdir. Bu çalışmanın, hassas özelliklerin oranının hesaplanmasına ilişkin yapılan araştırmalar için bir kılavuz niteliği taşıması hedeflenmektedir.

Çizelge 1. Yaş gruplarına göre kitle ve örneklem tabaka büyüklükleri, tabaka ağırlıkları ve TNSA–2008 raporuna göre her bir tabakaya yönelik isteyerek düşük yapan kadınların oranları, tabaka varyansları

Yaş grupları	N_h	π_h	n_h	W_h	“Evet” cevabını veren kişilerin sayısı	“Evet” cevabını veren katılımcıların oranı (\hat{Z}_h)	$V(\hat{\pi}_h)$
9	3077387	0,029	87	0,154250	45	0,517241	0,015410
20-24	3050973	0,036	86	0,152926	23	0,267442	0,015665
25-29	3095890	0,117	88	0,155177	34	0,386364	0,016089
30-34	3210247	0,188	91	0,160909	59	0,648352	0,016101
35-39	2795560	0,259	79	0,140124	21	0,265823	0,019043
40-44	2339933	0,327	66	0,117286	18	0,272727	0,023221
45-49	2380649	0,392	67	0,119327	23	0,343284	0,023147
			$n=564$				

Kaynaklar

- [1] S.L. Warner, 1965, Randomized Response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63-69.
- [2] K. Hong, J. Yum, H. Lee, 1994, A stratified randomized response technique. *The Korean Journal of Applied Statistics*, 7, 141-147.
- [3] J. Kim, W.D. Warde, 2004, A stratified Warner’s randomized response model. *Journal of Statistical Planning and Inference*, 120, 155-165.
- [4] J. Kim, M. Elam, 2005, A two-stage stratified Warner’s randomized response model using optimal allocation. *Metrika*, 61, 1-7.
- [5] S. Bacanlı, T. Tuncel, 2014, A Post- Stratified randomized response model for proportion, *American Journal of Mathematics and Statistics*, 4(3), 156-161.
- [6] S. Tezcan, A.R. Omran, 1981, Prevalence and Reporting of Induced abortion in Turkey: *Two Survey Techniques. Studies in Family Planning*, Vol.12, No.6/7, 262-271.
- [7] TNSA Türkiye Nüfus ve Sağlık Araştırması 2008, 2009, Hacettepe Üniversitesi Nüfus Etüpleri Enstitüsü.
- [8] T. Tuncel, 2012, *Sonradan Tabakalı Örneklem Yönteminde Rasgeleleştirilmiş Cevap Modelleri*. Yüksek Lisans Tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, Ankara.
- [9] Y. Bek, 1977, Şansa bağlı cevap verme tekniği, *Atatürk Üniversitesi Ziraat Fakültesi Dergisi*, 8(1), 35-45