IAAOJ, Scientific Science, 2013,1(1),3-8

KADINLARDA FG SKORLARI İÇİN KATEGORİK VARYANS ANALİZİ

Yrd.Doç.Dr. Nuri ÇELİK[1]

[1]Bartın Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Bartın, Türkiye
e-posta: ncelik@bartin.edu.tr

ÖZET

İstatistik uygulamalarında birçok analiz ve yöntem verilerin sürekliliği varsayımına dayanır. Bununla birlikte varyans analizinin de en önemli varsayımlarından biri de verilerin sürekliliğidir. Ancak günlük hayatta verilerin daha çok kategorik, isimsel ya da sıralı olduğu gözlenmiştir. Bu çalışmada kategorik veri setleri için tek yönlü varyans analizi (CATANOVA) metodu tanıtılacak, kareler toplamının parçalanışı ve dağılımı verilecek ve gerçek veri seti ile uygulaması yapılacaktır.

Anahtar Kelimeler: Kategorik Veri, Kategorik Varyans Analizi, Kareler Toplamı, Hipotez Testi

CATEGORİCAL ANALYSIS OF VARIANCE FOR WOMEN FG SCORE

ABSTRACT

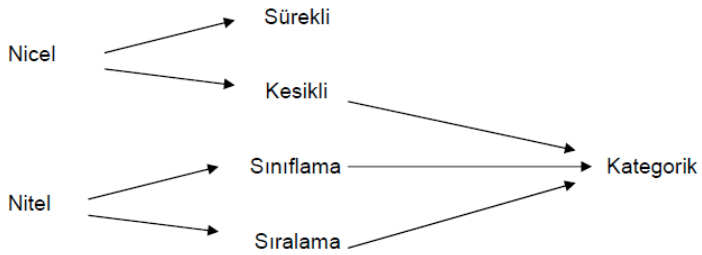
In statistical analysis, lots of application and methods are about on normality assumption. In the same way, in the analysis of variance method too, the most important assumption is normality. However, It is observed that there are more categorical, nominal or sequential data than continious data in everday life. In this work, one way analysis of variance method for categorical data (CATANOVA) will be determined, components of variation and their distributional behaviour will be constructed and numerical example will be applied to real data set.

Keywords: Categorical Data, Categorical Variance Analysis, Sum of Squares, Hypothesis Testing

1. Giriş

Tıp, biyoloji, psikoloji ve birçok sosyal bilim dallarında yapılan araştırmalarda kullanılan değişkenler çoğunlukla sayım ile belirlenmektedir. Değişkenler nitel (kalitatif) ve nicel (kantitatif) olmak üzere ikiye ayrılmaktadır. Daha çok sosyal bilimlerde karşılaşılan nitel değişkenler, sözcüklerle ifade edilirken, nicel değişkenler sayılarla ifade edilirler. Kişilerin adları, cinsiyetleri ve oturdukları şehirler nitel değişkenlere; boy uzunlukları ve kiloları ise nicel değişkenlere örnek teşkil eder. Nicel değişkenler de kendi aralarında sürekli ve kesikli olarak ikiye ayrılırlar. Sürekli değişkenlerde, iki değişken değeri arasına sonsuz çoklukta değer yerleştirilebilirken, kesikli değişkenlerde değişken sadece belirli değerler alabilmektedir. Ailedeki çocuk sayısı kesikli bir değişken iken, ailenin toplam kazancı sürekli bir değişkendir. Nitel değişkenler ise, sıralama ve sınıflama ölçme düzeylerinin genel bir adıdır.

Bu doğrultuda, kategorik değişken, nitel değişkenler ve kesikli nicel değişkenleri kapsamaktadır. Kategorik değişken kapsamı, şekil 1.'deki gibi özetlenebilir.



Şekil 1. Kategorik Veri Tanımı (Powers ve Xie, 2000).

Varyans analizi (ANOVA), gözlenen varyansı çeşitli kısımlara ayırma yöntemiyle bazı değişkenlerin başka bir değişken üzerindeki etkisini incelemeye yarayan bir grup modelleme türü ve bu modellerle ilişkili işlemlere verilen genel isimdir (Lindman, 1974). Varyans analizindeki en önemli varsayım verilerin normal dağıldığı varsayımıdır. Ancak bu varsayım ortadan kalktığında bir takım problemler söz konusu olup çeşitli alternatif yöntemler geliştirilmiştir. Veri seti kategorik olduğunda varyans analizi yöntemi için değişik yöntemler geliştirilse de, ilk olarak Light ve Margoin (1971) tarafından yayımlanan makale ile literatüre dahil olmuştur. Ayrıca Singh (1992) CATANOVA ile standart ki-kare analizinin güçleri karşılaştırmış ve CATANOVA'nın daha güçlü olduğunu göstermiştir. Onukogu (1985) CATANOVA analizini iki boyutlu çoklu veri setine genişletmiştir. Singh (1996) , iki boyutlu çapraz sınıflandırılmış kantitif veri setleri için CATANOVA analizini tanımlamış ve test istatistiğini elde etmiştir.

Bu çalışmada öncelikle CATANOVA analizinin teorik altyapısı özetlenmiş daha sonra Türkiye için kadınların FG skorları bölgelere göre ayrışmış çapraz tabloda uygulaması yapılmıştır.

2. Materyal ve Metot

2.1. Veri Yapısı

Bir yönlü varyans analizi için kullanılacak olan verinin yapısı, ki-kare analizinde kullanılan $n \times k$ boyutlu tablolara benzemektedir. Buna göre veri yapısı aşağıdaki tablodaki gibidir.

		Kategoriler				
		1	2	…	i	**Toplam**
Denemeler	1	n_{11}	n_{21}	$\cdots$	n_{i1}	n_{+1}
	2	n_{12}	n_{22}	$\cdots$	n_{i2}	n_{+2}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	j	n_{1j}	n_{2j}	$\cdots$	n_{ij}	n_{+j}
	Toplam	n_{1+}	n_{2+}	$\cdots$	n_{i+}	n

Burada, i, kategori sayısını ve j, deneme sayısını göstermektedir. Örnek olarak denemeler, bölgeler olarak alınırsa kategorilerde cinsiyet alınılabilir. Bununla beraber, n_{ij}, *j.* denemede *i.* kategoride yer alan örneklem sayısıdır. Burada n_{i+}, *i.* satır toplamı ve n_{+j} *j.* sütun toplamıdır. Böylece, bütün gözelerdeki toplam sayı, örneklem sayısını vermektedir.

2.2. Kategorik Veri İçin Kareler Toplamı

Varyans analizi metodunda, değişim ölçüsü olarak, her bir veri setinin ortalamadan sapmalarının karesi kullanılmaktadır ve bu kareler toplamının parçalanması ile test istatistiği geliştirilmiştir. Ancak kategorik veri seti için ortalamadan sapma anlamsız olacağı için başka ölçüler geliştirilmiştir. Gini (1936), değişim ölçüsünü her bir veri çifti için tanımlanan $\binom{n}{2}$ adet farkların toplamı şeklinde ifade etmiştir. X_i ve X_j her bir gözlemi göstermek üzere, kareler toplamı,

$$SS = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(X_i - X_j\right)^2$$
$$= \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}^{\ 2}$$

şeklindedir. Burada

$$d_{ij} = X_i - X_j$$

şeklinde tanımlanmaktadır. Daha sonra, aynı mantıkla, kategorik veri seti için değişim ölçüsü,

$$d_{ij} = \begin{cases} 1 & , X_i \ ve \ X_j \ farklı \ kategoride \ ise \\ 0 & , X_i \ ve \ X_j \ aynı \ kategoride \ ise \end{cases}$$

biçiminde ifade edilmiştir. Buna göre, *i* kategori için, veri seti $\Phi = (n_1, n_2, \ldots, n_i)$ şeklinde özetlenebilir. Burada n_i, i. kategorideki yanıt değişkeninin sayısıdır. Burada

$$\sum_{i=1}^{k} n_i = n$$

şeklinde olup, yanıt değişkenlerinin değişim ölçüsü,

$$\frac{1}{2n}\left[\sum_{i \neq j} n_i n_j\right] = \frac{1}{2n}\left[n^2 - \sum_{i=1}^{k} n_i^2\right]$$
$$= \frac{n}{2} - \frac{1}{2n}\sum_{i=1}^{k} n_i^2$$

şeklinde hesaplanmaktadır.

2.3. Kareler Toplamının Parçalanışı

Kategorik veri seti için düzenlenen varyans analizinde toplam kareler toplamı,

$$TSS = \frac{n}{2} - \frac{1}{2n}\sum_{i=1}^{k} n_{i+}^2$$

şeklindedir. Grup içi kareler toplamı,

$$WSS = \sum_{j=1}^{g}\left(\frac{n_{+j}}{2} - \frac{1}{2n_{+j}}\sum_{i=1}^{k} n_{ij}^2\right)$$
$$= \frac{n}{2} - \frac{1}{2}\sum_{j=1}^{g}\frac{1}{n_{+j}}\sum_{i=1}^{k} n_{ij}^2$$

gruplar arası kareler toplamı ise,

$$BSS = \frac{1}{2}\left(\sum_{j=1}^{g}\frac{1}{n_{+j}}\sum_{i=1}^{k} n_{ij}^2\right) - \frac{1}{2n}\sum_{i=1}^{k} n_{i+}^2$$

şeklinde hesaplanmaktadır. Bu bilgiler ışığında açıklanan değişimin oranı,

$$R^2 = \frac{BSS}{TSS}$$
$$= \frac{\left(\sum_{j=1}^{g} \frac{1}{n_{+j}} \sum_{i=1}^{k} n_{ij}^2\right) - \frac{1}{n}\sum_{i=1}^{k} n_{i+}^2}{n - \frac{1}{n}\sum_{i=1}^{k} n_{i+}^2}$$

biçimindedir.

2.4. Test İstatistiği

Bir önceki bölümde kareler toplamının parçalanışı ile elde ettiğimiz metot, kategorik veri seti için varyans analizini oluşturmaktadır ve CATANOVA adını almaktadır. Buna göre, bilinen varyans analizi yönteminde kullanılan ortalamaların karşılaştırılması yerini, oranların karşılaştırılmasına bırakacaktır. Dolayısıyla test edilecek hipotez,

$$H_0 : p_1 = p_2 = \ldots = p_g$$

şeklinde elde edilecektir. Sıfır hipotezinin doğruluğu altında elde edilecek test istatistiği ise,

$$C = \frac{(n-1)(k-1)BSS}{TSS}$$

şeklindedir (Light ve Margolin, 1971). Kategorik varyans analizi için kullanılan test istatistiği ise asimptotik olarak $(g-1)(k-1)$ serbestlik dereceli ki-kare dağılıma sahiptir. Dolayısıyla,

$$C = (n-1)(k-1)\left[\frac{\left(\sum_{j=1}^{g} \frac{1}{n_{+j}} \sum_{i=1}^{k} n_{ij}^2\right) - \frac{1}{n}\sum_{i=1}^{k} n_{i+}^2}{n - \frac{1}{n}\sum_{i=1}^{k} n_{i+}^2}\right] \sim \chi^2_{(g-1)(k-1)}$$

olacaktır. (Çıkarımlar için bkz, Light ve Margolin, 1971)

2.5. CATANOVA Tablosu

Önceki bölümlerdeki bilgiler ışığında CATANOVA tablosu aşağıdaki tabloda verilmektedir.

Değişim Kaynağı	Kareler Toplamı	R^2	C
Gruplar Arası	BSS	$\frac{BSS}{TSS}$	$\frac{(n-1)(k-1)BSS}{TSS}$
Gruplar İçi	WSS		
Toplam	TSS		

2.6. Hipotez Testi ve Karar

CATANOVA tablosuna göre, hipotez testleri

$$H_0 : p_1 = p_2 = \ldots = p_k$$
$$H_1 : En\ az\ biri\ \ farklidir$$

şeklinde kurulmaktadır. Buna göre, C istatistiğinin değeri ki-kare tablo değerinden büyükse sıfır hipotezi reddedilir. Yani karar kuramı,

$$C > \chi^2{}_{\alpha,(g-1)(k-1)} \Rightarrow H_0 \; red.$$

şeklindedir.

3. Uygulama

Uygulamada, 4 farklı bölgede kadınlar üzerinde yapılan bir ankette FG skorları ile ilgili veriler toplanmıştır. FG skoru, kadın vücudundaki erkeklik hormonuna bağlı olarak 11 farklı bölgedeki tüylenme oranını göstermektedir. (Ferriman and Gallwey, 1961).FG skorları, uluslar arası standartlara sahip skorlama sistemiyle elde edilmiş olup Türkiye uygulaması henüz yapılmamıştır. Buna göre FG skorları uluslar arası skorlama sistemi göz önüne alınarak düşük, orta, yüksek şeklinde üç farklı kategoriye ayrılmıştır. Bölgeler ise, Marmara, Ege ve Akdeniz Bölgeleri Batı, İç Anadolu Bölgesi, Karedeniz Bölgesi ve Güneydoğu Anadolu ve Doğu Anadolu Bölgesi olarak dört farklı bölge belirlenmiştir. 758 kadın üzerine yapılan anketten elde edilen veriler aşağıdaki tabloda özetlenmiştir.

		FG Skorları			
Bölgeler		Düşük	Orta	Yüksek	**Toplam**
	Doğu	175	68	88	331
	Batı	29	52	20	101
	İç Anadolu	57	31	39	127
	Karadeniz	93	92	14	199
	Toplam	354	243	161	758

Burada, bölgeler denemeler olarak düşünülecektir. Hipotez ise bölgeler arasında FG skorları açısından anlamlı bir fark olup olmadığıdır. Buna göre, CATANOVA tablosunu oluşturmak için,

$$\begin{aligned} TSS &= \frac{n}{2} - \frac{1}{2n}\sum_{i=1}^{k} n_{i+}^2 \\ &= \frac{758}{2} - \frac{1}{2(758)}\left[354^2 + 243^2 + 161^2\right] \\ &= 240.289 \end{aligned}$$

$$\begin{aligned} WSS &= \sum_{j=1}^{g}\left(\frac{n_{+j}}{2} - \frac{1}{2n_{+j}}\sum_{i=1}^{k} n_{ij}^2\right) \\ &= \frac{n}{2} - \frac{1}{2}\sum_{j=1}^{g}\frac{1}{n_{+j}}\sum_{i=1}^{k} n_{ij}^2 \\ &= \frac{758}{2} - \frac{1}{2}\left[\left(\frac{1}{331}\left(175^2 + 68^2 + 88^2\right)\right) + \ldots + \left(\frac{1}{199}\left(93^2 + 92^2 + 14^2\right)\right)\right] \\ &= 228.473 \end{aligned}$$

ve

$$\begin{aligned} BSS &= TSS - WSS \\ &= 240.289 - 228.473 \\ &= 11.815 \end{aligned}$$

şeklinde hesaplanacaktır. Buna göre, CATANOVA tablosu,

Değişim Kaynağı	Kareler Toplamı	R^2	C
Gruplar Arası	11.815	0.049	74.44
Gruplar İçi	228.473		
Toplam	240.289		

şeklinde düzenlenmiştir. Buna göre $\chi_h^2 > \chi_6^2 = 12.59$ olduğu için hipotez reddedilir. Yani bölgelere göre FG oranları arasında istatistiksel olarak anlamlı bir fark vardır denilmektedir.

4. Sonuç ve Tartışma

Varyans analizi üç ya da daha fazla grup ortalaması arasında istatistiksel olarak farklılık olup olmadığını test etmek için kullanılan bir yöntemdir. Varyans analizinin uygulanabilmesi için verilerin sürekli olması gerekmektedir. Veri seti kategorik olduğunda ise denemeler arasında anlamlı bir fark olup olmadığının araştırılması için Kategorik varyans analizi teknikleri kullanılmaktadır. Bu çalışmada kategorik veri setleri için CATANOVA tekniği tanıtılmış ve avantajları vurgulanmıştır.

Kaynaklar

Gini, C. (1936) On the measure of Concentration with special reference to incomme and statistics, Colorado College Publication, Generel Series no: 208, 73-79.

Light, J.R. and Margolin, B.H.,(1971) An Analysis of Variance for Categorical Data,Journal of American Stastical Assosication, 66, pp:329-335.

Lindman H.R, (1974), Analysis of Variance in complex experimental designs, San Fransisco: W. H. Friman and J.O.

Ferriman, D. and Gallwey, J.D. (1961), Clinical Assesment of Body Hair Growth in Women, J. Clin. Endocrinology, 21, pp: 1440-1447

Onukogu, I.B. (1985), An analysis of variance of nominal data, Biom. J., 27(4), pp: 375-384

Powers, D.A., Xie, Y. (2000), Statistical methods for Categorical data Analysis, John Emerald Group Publishin Ltd, England

Singh, B. (1992), A comparison of catanova and chi-square tests for nominal data, Ind, J. Appl. Statistics, 1,pp: 1-11.

Singh, B. (1995), On catanova method for analysis of two-way classified nominal data, Indian Journal of Statistics, Series B (1960-2002), 58(3), pp: 379-388.