Normal ve Karma-Normal Dağılımlı Veri Grupları İçin İlişki Katsayılarının Eşitliğinde Permütasyon Testi

Ufuk EKİZ, Meltem EKİZ*

Gazi Üniversitesi Fen Edebiyat Fakültesi İstatistik Bölümü, Ankara

Özet: İki bağımsız yığına ilişkin korelasyonların eşitliği hipotezinin testi için kullanılan parametrik testler, örnek çaplarının büyük ve tesadüfi değişkenlerin dağılımlarının normal olması varsayımını gerektirmektedir. Bu hipotezin testinde hem küçük örnek çapı durumunda hem de normallik varsayımının sağlanmaması durumunda kullanılabilen permütasyon testinin, birinci tip hata miktarının incelenmesi makalenin amacını oluşturmaktadır. Bu amaçla yapılan simülasyon çalışması örnek çapının küçük olması durumunda, hem normal hem de karma-normal dağılımlı veri üretilmesine dayalı olarak gerçekleştirilmiştir.

Anahtar Kelimeler: Karşılıklı değişebilirlik, Permütasyon testi, Karma-normal protein

Permutation Test For The Equality of Correlation Coefficients of Normal and Mixed-Normal Distributed Data Sets

Abstract: Parametric tests, using for the equality of correlation coefficients among two independent populations, requires the assumption of the normallity of variables and large sample sizes. For testing the hypothesis of the equality of correlation coefficients the permutation test can be used. These test is useful in case of small sample sizes and also non-normallity In this study we calculated the amount of the Type I errors. In order to do this, a simulation study is made in case of small sample sizes and the data sets are generated from normally and mixed-normal distributions.

Key Words: Exchangeability, Permutation test, Mixed-normal

Giriş

Permütasyon testi, gözlemlerin karşılıklı değişebilir olmasına dayalı parametrik olmayan bir yöntemdir. Bu metodun, yokluk hipotezi altında, kesin ve sapmasız çözümler ürettiği ve asimptotik olarak parametrik testler kadar güçlü olduğu bilinmektedir. İki bağımsız yığına ilişkin korelasyon katsayılarının eşitliği hipotezinin permütasyon testi ile test edilmesi durumunu göz önünde bulunduralım. Her bir yığından gelen veriler, iki değişkenli dağılıma sahiptirler. $X^{(i)} = \left(X_1^{(i)}, X_2^{(i)}\right)$, i. yığına ait iki-değişkenli tesadüfi değişkenler vektörü olsun. Bu tesadüfi değişkenin merkezi parametresinin ve kovaryans matrisinin de i=1,2 olmak üzere,

$$\Sigma^{(i)} = \begin{bmatrix} \sigma_{11}^{(i)} & \rho^{\mathrm{i}}\sqrt{\sigma_{11}^{(i)}}\sqrt{\sigma_{22}^{(i)}} \\ \rho^{\mathrm{i}}\sqrt{\sigma_{11}^{(i)}}\sqrt{\sigma_{22}^{(i)}} & \sigma_{22}^{(i)} \end{bmatrix}$$

şeklinde olduğu varsayılsın. Yokluk hipotezi $H_0 : \rho^{(1)} = \rho^{(2)}$ ve karşıt hipotez de $H_0 : \rho^{(1)} \neq \rho^{(2)}$ şeklinde ifade edilsin. Eğer tesadüfi vektör normal dağılıma sahipse,

$$T = \frac{z^{(1)} - z^{(2)}}{\sqrt{1/\left(n_1 - 3\right) + 1/\left(n_2 - 3\right)}}$$

istatistiği asimptotik olarak normal dağılır. Burada

* ozmeltem@gazi.edu.tr

$$z^{(i)} = \frac{1}{2}\log\frac{1+r^{(i)}}{1-r^{(i)}}$$

Fisher'in z'si olarak bilinir ve $r^{(i)}$ de $\rho^{(i)}$ korelasyon katsayısının En Küçük Kareler tahminidir [1]. Ancak tesadüfi değişken vektörü normal dağılmıyor ve (veya) örnek çapı küçükken parametrik olmayan permütasyon testi kullanılır. Bu çalışmada yukarıda verilen hipotezin testinde permütasyon testinin kullanımı ve simülasyon yoluyla birinci tip hatasının belirlenmesi üzerinde durulmuştur. Bu amaçla ikinci bölümde yokluk hipotezinin testinde kullanılacak permütasyon testi, üçüncü bölümde de permütasyon testinin sahip olduğu birinci tip hata miktarlarına ilişkin simülasyon çalışmaları yer almaktadır.

Permütasyon testi

Permütasyon testinin kesin ve sapmasız bir test olabilmesi için gerekli tek koşul, tesadüfi değişkenin karşılıklı değiştirilebilir olmasıdır. Eğer,

$$P\left(\bigcap_{i=1}^{n}\left(X_i \le x_i\right)\right) = P\left(\bigcap_{i=1}^{n}\left(X_i \le x_{c_i}\right)\right)$$

ise $X_1, X_2, ..., X_n$ tesadüfi değişkenleri karşılıklı değiştirilebilirdir. Buradaki $c_1, c_2, ..., c_n$ indisleri $(1,2,...,n)$'nin permütasyonlarından herhangi biridir. Ayrıca bağımsız ve aynı dağılımlı tesadüfi değişkenler karşılıklı değiştirilebilirdir [2].

$x_k^{(i)}$, $k = 1,...,n_i$, i. yığına ilişkin tesadüfi bir örnek, $\bar{x}^{(i)}$, $s_{jj}^{(i)}$ ve $r^{(i)}$ sırasıyla $\mu^{(i)}$, $\sigma_{jj}^{(i)}$ ve $\rho^{(i)}$ parametrelerinin tahminleri, $\Sigma^{(i)}$ kovaryans matrisi ve $D^{(i)} = diag\left(\sigma_{11}^{(i)}, \sigma_{22}^{(i)}\right)$ olsun. $\mu^{(i)}$ ve $\sigma_{jj}^{(i)}$ biliniyor olmak üzere,

$$y_k^{(i)} = D^{(i)-\frac{1}{2}}\left(x_k^{(i)} - \mu^{(i)}\right)$$

bağımsız ve aynı dağılımlıdırlar ve $y_k^{(1)}$ ile $y_k^{(2)}$ yokluk hipotezi altında aynı dağılıma sahiptirler. $\mu^{(i)}$ ve $\sigma_{jj}^{(i)}$'nin bilinmiyor olması durumunda,

$$y_k^{(i)} = \hat{D}^{(i)-\frac{1}{2}}\left(x_k^{(i)} - \bar{x}^{(i)}\right)$$

dönüşümü uygulanabilir. $\hat{D}^{(i)} = diag\left(s_{11}^{(i)}, s_{22}^{(i)}\right)$'dır.

Teorem (*Slutsky*): $\theta_n \in R^d$, θ'ya olasılıkta yakınsayan bir dizi olmak üzere, R^d'de tanımlanan her sürekli $h(\theta_n)$ fonksiyonu olasılıkta $h(\theta)$'ya yakınsar.

$\theta_n^{(i)} = \left(\bar{x}^{(i)'} s_{11}^{(i)} s_{22}^{(i)}\right)$'nın olasılıkta $\theta^{(i)} = \left(\mu^{(i)'} \sigma_{11}^{(i)} \sigma_{22}^{(i)}\right)$'ye yakınsaması nedeniyle $y_1^{(1)},...,y_{n_1}^{(1)}$, $y_2^{(2)},...,y_{n_2}^{(2)}$'den asimptotik bağımsızdır.

Teorem: $y_1^{(1)},...,y_{n_1}^{(1)}$ ile $y_2^{(2)},...,y_{n_2}^{(2)}$ asimptotik olarak karşılıklı değiştirilebilirdir [3].

Teoremden hareketle $y_1^{(1)},...,y_{n_1}^{(1)}$ ve $y_2^{(2)},...,y_{n_2}^{(2)}$ için permütasyon testi uygulanabilir. Yokluk hipotezinin testi için,

$$T_D = \left|r^{(1)} - r^{(2)}\right| \tag{1}$$

$$T_F = \left|z^{(1)} - z^{(2)}\right| \tag{2}$$

test istatistikleri kullanılmaktadır. $y_1^{(1)},...,y_{n_1}^{(1)}$ ve $y_2^{(2)},...,y_{n_2}^{(2)}$ değerleri için $y_1^{(1)*},...,y_{n_1}^{(1)*}$ ve $y_2^{(2)*},...,y_{n_2}^{(2)*}$ permütasyon verisi üzerinden eşitlik 1 ve 2'de verilen T_D ve T_F istatistikleri hesaplanarak yokluk hipotezi test edilir.

Uygulama

Permütasyon testi uygulamalarında birinci tip hata olasılığı kesin olarak bilinmemektedir. İki yığına ilişkin korelasyon katsayılarının eşitliği hipotezinin testinde kullanılan bu test yönteminin anlamlılık düzeyinin ne olduğuna ilişkin yaklaşık değerler simülasyon çalışması ile elde edilmiştir. Permütasyon testinin sağladığı iki önemli avantaj bulunmaktadır. Bunlar örnek çapının küçük olması durumunda da kullanılabilirliği ve parametrik testlerin gerektirdiği normallik varsayımını gerektirmemesidir. Bu amaçla, örnek çapı $n_1 = n_2 = 6$ gibi küçük bir değer alınmış ve örneklerin üretileceği dağılımların normal ve karmaşık-normal olduğu iki farklı durum altında simülasyon çalışmaları gerçekleştirilmiştir.

Birinci durumda, $n_1 = n_2 = 6$ çaplı örnekler iki-değişkenli normal dağılımdan, $N\left(0, \Sigma^{(i)}\right)$, rasgele üretilmişlerdir. Burada,

$$\Sigma^{(i)} = \begin{pmatrix} 1 & \rho^{(\mathrm{i})} \\ \rho^{(\mathrm{i})} & 1 \end{pmatrix} \qquad \mathrm{i}=1,2$$

i.yığına ait varyans-kovaryans matrisidir. H_0 hipotezinin doğruluğu koşulu altında, yani iki yığının korelasyon katsayılarının eşit olması durumunda, üretilen bir örnek için T_D ve T_F değerleri hesaplanmış daha sonra tüm mümkün permütasyonlar oluşturulup bu permütasyonlara ait T_D ve T_F değerleri hesaplanmıştır. Permütasyonlara ilişkin T_D ve T_F değerlerinden örneğin T_D ve T_F değerlerinden büyük olanların oranı tespit edilmiştir. Eğer bu oran α 'dan küçük ise H_0 hipotezi red kararı verilmiştir. Bu işlem 10000 defa tekrar edilmiştir. Elde edilen bu 10000 denemede H_0 hipotezinin hangi oranda red edildiği tespit edilmiştir. Korelasyon katsayılarının çeşitli değerleri için elde edilen sonuçlar Tablo 1'de yer almaktadır.

Normallik varsayımının sağlanmadığı durumlarda da permütasyon testinin anlamlılık düzeyinin ne olacağına ilişkin simülasyon çalışmasında ise $n_1 = n_2 = 6$ çaplı örnekler karışım oranı $\varepsilon = 0.30$ olmak üzere,

$$(1-\varepsilon) N\left(0, \textstyle\sum_1^{(i)}\right) + \varepsilon N\left(0, \textstyle\sum_2^{(i)}\right)$$

karmaşık-normal dağılımdan üretilmiştir. Burada

$$\Sigma_1^{(i)} = \begin{pmatrix} 1 & \rho^{(\mathrm{i})} \\ \rho^{(\mathrm{i})} & 1 \end{pmatrix}, \quad \Sigma_2^{(i)} = \begin{pmatrix} 9 & 9\rho^{(\mathrm{i})} \\ 9\rho^{(\mathrm{i})} & 9 \end{pmatrix}$$

şeklinde tanımlanmıştır. H_0 hipotezinin doğruluğu altında yukarıdaki karmaşık-normal dağılımdan üretilen $n_1 = n_2 = 6$ çaplı örneklerden T_D ve T_F değerleri elde edilmiştir. Elde edilen bu örneğin tüm mümkün permütasyonlarına ait T_D ve T_F değerleri elde edilmiştir. Permütasyonlara ilişkin T_D ve T_F değerlerinden, örneğin T_D ve T_F değerlerinden büyük olanların oranı tespit edilmiştir. Bu işlem 10000 defa tekrar edilmiştir. Elde edilen bu 10000 denemede H_0 hipotezinin hangi oranda red edildiği tespit edilmiştir. Korelasyon katsayılarının çeşitli değerleri için elde edilen sonuçlar Tablo 2'de yer almaktadır.

Tablo 1. H_0 hipotezinin doğruluğu altında T_D ve T_F test istatistiklerine ilişkin anlamlılık düzeyleri (Örnekler normal dağılımdan üretilmiştir).

$\rho^{(1)} = \rho^{(2)}$		$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
-0.9	T_D	0.0067	0.0422	0.0928
	T_F	0.0093	0.0473	0.0978
0	T_D	0.0084	0.0468	0.0997
	T_F	0.0082	0.0488	0.0996
0.9	T_D	0.0074	0.0449	0.0918
	T_F	0.0074	0.0481	0.0919

Tablo 2. H_0 hipotezinin doğruluğu altında T_D ve T_F test istatistiklerine ilişkin anlamlılık düzeyleri (Örnekler karmaşık-normal dağılımdan üretilmiştir).

$\rho^{(1)} = \rho^{(2)}$		$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
-0.9	T_D	0.0002	0.0077	0.0310
	T_F	0.0097	0.0443	0.0846
0	T_D	0.0060	0.0346	0.0754
	T_F	0.0069	0.0394	0.0803
0.9	T_D	0.0008	0.0080	0.0299
	T_F	0.0109	0.0383	0.0797

Sonuç

İki bağımsız yığına ilişkin korelasyon katsayılarının eşitliği hipotezinin testinde kullanılabilen parametrik testlerde, örnek çapının yeterince büyük olması ve(veya) normallik varsayımının sağlanması gerekmektedir. Aynı hipotezin, gözlemlerin normal dağılımdan geldiği varsayımının sağlandığı ancak örnek çapının küçük olması durumunda parametrik olmayan permütasyon testi ile test edilmesi durumunda en az parametrik testlerin sahip olduğu anlamlılık seviyesine sahip olduğu Tablo 1'den gözlemlenmiştir. Gözlemlerin karmaşık-normal dağılımdan gelmesi durumunda ise T_D istatistiğinin anlamlılık düzeyi T_F istatistiğinin anlamlılık düzeyinden çok daha küçük olarak belirlenmiştir.

Kaynaklar

[1] Anderson, T.W., **An Introduction to Multivariate Statistical Analysis**. 2nd Ed., Wiley, New York, (1984)

[2] Fay, M.P. ve Follmann, D.A., **Designing Monte Carlo implementations of permutation or bootstrap hypothesis tests**. The American Statistician, 56, 1, (2002).

[3] Sakaori, F., **Permutation test for equality of correlation coefficients in two populations**. Commun. Statist.- Simula., 31(4), 641-651, (2002).