

## PERMÜTASYON TESTLERİNİN DOĞRUSAL REGRESYONDA KULLANILABİLİRLİĞİNİN İRDELENMESİ

Hasan ÖNDER

Ondokuz Mayıs Üniversitesi, Ziraat Fakültesi, Zootekni Bölümü, Samsun

Sorumlu yazar: hasanonder@gmail.com

Geliş Tarihi: 06.02.2006

Kabul Tarihi: 30.03.2007

**ÖZET:** Genellikle  $F$  ve  $t$  testleri deneysel veri analizinde doğrusal modellerin ve/veya parametrelerin önemini test etmek için kullanılır. Bu testler çoğu durumda oldukça etkili olsa da modelin ihtiyaç duyduğu bir ya da daha fazla varsayım sağlanmadığında etkilerini kaybetmektedir. Bu durumda, varsayımlardan etkilenmeyen permütasyon testleri parametrik olmayan bir yöntem olarak uygulanabilmektedir. Bu çalışmada, doğrusal regresyon analizi için permütasyon testleri incelendi. Testin regresyon tekniği ile birlikte kullanımı biyolojik çalışmalardan elde edilen ve yapay olarak üretilen veri kümeleri üzerinde gerçekleştirildi. Ayrıca, permütasyon testlerinin iki türü (ham verinin tam permütasyonu ve kalıntılarının permütasyonu) Normal, Ki-kare ve Poisson dağılımları gibi farklı dağılımla sahip veri setleri için karşılaştırılmalı olarak incelendi. Sonuç olarak, bu çalışmada ilgilenilen tüm dağılımlarda permütasyon testlerinin I. Tip hatayı engellemek için kullanılabilmesi anlaşıldı.

**Anahtar Kelimeler:** Permütasyon testleri, doğrusal regresyon, I. Tip hata oranı

### TO EXAMINE THE USABILITY OF PERMUTATION TESTS ON LINEAR REGRESSION

**ABSTRACT:** Generally,  $F$  and  $t$  tests are used to test significance of linear models and/or their parameters in experimental data analysis. Although these tests are considerably effective in most cases they may be ineffective for some data sets when one or more of the assumptions belongs to the model are not satisfied. In these cases, permutation tests that are not affected by the assumptions can be applied as non-parametric test methods. In this paper, the permutation tests for linear regressions were introduced, and their uses were demonstrated on real biometrical and hypothetically produced data sets. Additionally, two types of permutation (permutation of raw data and permutation of residuals) were also compared for data sets which have Normal, Chi-square, Poisson distributions. As a result, it was obtained that permutation tests can be used to avoid Type I error for linear regression models in all forms of distributions concerned in this study.

**Key Words:** Permutation tests, linear regression, Type I Error rate

### 1. GİRİŞ

Doğrusal regresyon modelleri için açıklayıcı (bağımsız) değişkenlerin birbirlerinden bağımsız olduğu ve benzer dağılıma sahip oldukları varsayılmaktadır. Parametrik regresyon analizinin bir diğer varsayımı ise; regresyon katsayısının test edilmek istenmesi durumunda hata terimlerinin normal dağılıma sahip olması gerekliliğidir. Bu varsayımlar altında regresyon katsayısının En Küçük Kareler (EKK) tahmin edicisi  $n-k$  serbestlik dereceli  $t$  dağılımına sahiptir (burada,  $n$  ve  $k$  sırasıyla örnek büyüklüğü ve açıklayıcı değişken sayısını göstermektedir).

Örnek büyüklüğü arttıkça  $t$  dağılımı standart normal dağılıma  $N(0,1)$  yaklaşmaktadır. Merkezi limit teoreminden de bilindiği üzere regresyon katsayısı uygun örnek büyüklüğünde normal dağılıma sahip olmaktadır. Bununla birlikte, hata terimlerinin normal dağılım göstermediği ve örnek büyüklüğünün yeterli düzeyde olmadığı durumlarda, EKK tahmin edicisi  $t$  dağılımına sahip değildir. Bu durumda  $t$ -testi önem testi için uygun değildir. Bu sorunu çözmek için Fisher (1935) tarafından önerilen permütasyon testleri kullanılabilir. Permütasyon testlerinde, korelasyon katsayıları olası tüm kombinasyonlar için hesaplanmaktadır ve ardından elde edilen bu değerler önem düzeyini belirlemek için orijinal veriden elde edilen korelasyon katsayıları ile karşılaştırılmaktadır.

Permütasyon testleri regresyon problemlerine doğrudan uygulanabilmektedir.

Doğrusal modellerde önem testleri model parametrelerinin bir alt kümesi üzerine  $F$  veya  $t$  testleri uygulanarak gerçekleştirilmektedir. Hata terimleri normal dağılıma sahip olduğunda parametrik yöntemler olan  $F$  ve  $t$  testleri çok güçlü sonuçlar üretebilmektedir. Ancak, hata terimleri normal dağılıma sahip değilse bu testler güçlerini kaybetmektedir. Bu durumda, hata terimlerinin dağılımı genellikle bilinmediğinden, testin gücünün hesaplanması önemli bir sorun teşkil etmektedir. Normallik varsayımının bozulduğu durumlar için parametrik olmayan testler sıklıkla parametrik testlerin yerine başarıyla kullanılabilir (O'Gorman, 2001; Önder ve Cebeci, 2005). Bu çalışmada permütasyon testleri dağılımın normal ve normal olmayan durumları için ayrı ayrı irdelendi. Permütasyon testleri 20. yüzyılın ortalarında tanımlanmış olmasına rağmen, hesaplanmasının uzun zaman alması nedeniyle ucuz ve göreceli olarak daha güçlü bilgisayarların yaygınlaşması ile beraber ancak günümüzde popüler hale gelebilmiştir (Nichols and Holmes, 2001). Bu çalışma Normal, Poisson ve Ki-kare dağılımına sahip hipotetik veriler ve hayvan besleme denemesinden elde edilen gerçek biyolojik bir veri seti kullanılarak ve permütasyon testlerinin I.

Tip hata üzerinde yaptığı etkileri incelemek üzere yürütüldü.

## 2. MATERYAL VE YÖNTEM

### 2.1. Materyal

Permütasyon testlerinde kullanılacak veri kümeleri bir açıklayıcı ve bir sonuç değişkeni olmak üzere 5, 10, 15, 20 ve 25 örnek büyüklüklerinde Normal, Poisson ve Ki-kare dağılımına sahip olmak üzere hipotetik olarak Minitab istatistik paket programı yardımı ile üretilmiştir. Permütasyon testlerinin gerçek biyolojik bir veri kümesi üzerinde de uygulanabilirliğinin gösterilebilmesi amacıyla etlik piliçler üzerinde gerçekleştirilen bir besi denemesinin sonuçları kullanıldı. Genellikle biyolojik çalışmalardan elde edilen veriler parametrik test yöntemlerinin ihtiyaç duyduğu varsayımları sağlayamamaktadır. Bu amaçla EKK, ham verinin tam permütasyonu (HVTP) ve kalıntıların tam permütasyonu (KTP) yöntemleri dağılımı bilinmeyen biyolojik veri kümesi üzerinde uygulandı. Kullanılan veri kümesinde karın yağı ağırlığı (g) sonuç değişkeni olarak kullanılırken, karkas ağırlığı (g) açıklayıcı değişken olarak kullanıldı. Permütasyon testleri Marti Jane Anderson tarafından yazılan DISTLM yazılımı kullanılarak uygulandı (Anderson, 2003).

### 2.2. Yöntem

Basit doğrusal regresyon modeli,

$$Y_i = X_i\beta + u_i, \quad i = 1,2,\dots,n$$

şeklinde yazılabilmektedir. Bu modelde  $F$  değeri,

$$F = \frac{MS_{reg}}{MS_{error}}$$

denklemleri kullanılarak hesaplanmaktadır (Kleinbaum ve ark, 1998).

$X_1$  sabit olduğunda ve  $n$  kombinasyonun varlığında (eksik gözlemler ve aynı değer yinelenmesi önemsizdir), olası kombinasyonlar  $(X_1, Y_j)$ ,  $j = 1,2,\dots,n$  şeklinde gösterilebilir. Benzer şekilde  $X_2$  sabit olduğunda,  $(X_2, Y_j)$ ,  $j = 2,3,\dots,n$  şeklinde gösterilebilir. Böylece,  $n - 1$  kombinasyon var olacaktır. Bu işlemin devamında,  $X_3$  ve  $X_4$  için sırasıyla  $n - 2$  ve  $n - 3$  kombinasyon elde edilecektir. Böylece  $X$  ve  $Y$  arasında olası tüm kombinasyonların toplamı  $n!$  olacaktır.

Sonuç değişkeni olan  $Y$  nin sırasının değiştirilmesiyle  $n!$  adet  $F$  değeri elde edilebilmektedir.  $Y$  nin sırasının değiştirilmesiyle diğer bir deyişle permütasyonu ile elde edilen  $F$  değerleri  $F^*$  ile gösterilsin.  $F^*$ ,  $*$  = 1,2,...,n! ve  $F_j^*$ ,  $F^*$  in  $j$ 'inci elemanı olsun.  $F_j^*$  nin sıfır hipotezi altında  $F$  değerler kümesinin  $j$ 'inci elemanı olduğu

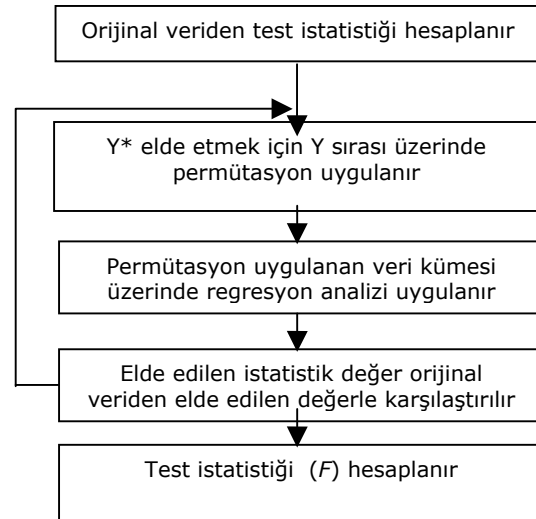
varsayıldığında, EKK ile tahmin edilen  $F$  değerler kümesinin  $j$ 'inci elemanı olan  $F_j$  nin deneysel dağılımı aşağıdaki şekilde verilebilir;

$$P(F_j < F_j^1) = \frac{F_j^1 \text{ den küçük } F^* \text{ sayısı}}{n!}$$

$$P(F_j = F_j^1) = \frac{F_j^1 \text{ ye eşit } F^* \text{ sayısı}}{n!}$$

$$P(F_j > F_j^1) = \frac{F_j^1 \text{ den büyük } F^* \text{ sayısı}}{n!}$$

Tüm  $j = 1,2,\dots,k$  için yukarıda gösterilen hesaplama yöntemi her olasılık için uygulanabilir ve hesaplanabilir. Regresyon denkleminin önem testi  $F_j^1$  nin konumunun belirlenmesiyle ortaya konulabilir.  $P(F_j < F_j^1)$  veya  $P(F_j > F_j^1)$  yeterince küçük ise sıfır hipotezi olan  $H_0 : \beta_j = 0$  iki yönlü test ile reddedilir.



Şekil 1. Permütasyon testi için uygulama çizelgesi (Makarenkov ve Legendre, 2002)

Doğrusal regresyonlarda permütasyon testlerinin uygulanışı Şekil 1'de görüldüğü gibi beş adımda özetlenebilmektedir;

1. Orjinal veri kümesinde regresyon modelinin  $F$  değerini elde etmek üzere EKK yöntemi kullanılarak sonuç değişkeni açıklayıcı değişken üzerinde regresyona tabii tutulur. Elde edilen  $F$  değeri referans değerini oluşturur.
2.  $Y^*$  değerlerini elde etmek için sonuç değişkeni olan  $Y$  nin değerleri şansa bağlı olarak permütasyona tabii tutulur.
3. Permütasyon uygulanan veri kümesinde katsayıların tahmini ve regresyon

denkleminin önem testi için  $Y^*$  değerleri açıklayıcı değişken üzerinde regresyona tabii tutulur.

4. 2. ve 3. adımlar olası tüm permütasyonlar için uygulanır.
5. Olasılıklar yukarıda açıklanan kurallar dahilinde hesaplanır (Anderson ve Legendre, 1999).

Elde edilen  $F^*$  değerlerinden I. Tip hata oranı

$$P = \frac{(F^* \geq F \text{ sayısı})}{(\text{toplam } F \text{ sayısı})}$$

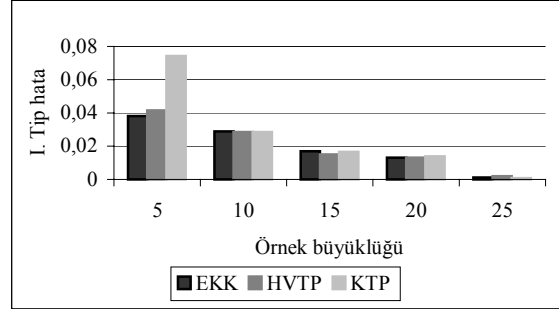
kullanılarak hesaplanabilmektedir. Elde edilen değer seçilen eşik değerinden daha küçük ise sıfır hipotezi reddedilmektedir.

### 3. ARAŞTIRMA SONUÇLARI VE TARTIŞMA

EKK, HVTP ve KTP yöntemlerinin 5, 10, 15, 20 ve 25 örnek büyüklüklerinde Normal, Ki-kare, Poisson dağılımlarında ve biyolojik veriye uygulanması ile elde edilen I. Tip hata oranları Çizelge 2'de gösterilmiştir. Çizelge 2 incelendiğinde görülmektedir ki, örnek büyüklüğü arttıkça beklendiği şekilde I. Tip hata oranları azalmaktadır. Örnek büyüklüğü 15'den küçük iken Normal dağılım dışındaki verilerde en küçük I. Tip hata oranı HVTP yöntemi ile elde edilmiştir. Konunun daha iyi değerlendirilebilmesi için veri kümesinde bulunan anormal gözlemler (outliers) Çizelge 1'de verilmiştir. Yorum kolaylığı açısından Normal, Ki-kare, Poisson ve gerçek biyolojik verilere ait I. Tip hata oranları sırasıyla Şekil 2, Şekil 3, Şekil 4 ve Şekil 5'te grafik olarak gösterilmiştir.

Çizelge 1. Veri kümelerinde bulunan anormal gözlem sayıları

Dağılımlar	Örnek büyüklüğü				
	5	10	15	20	25
Normal	0	1	1	0	1
Ki-kare	0	1	1	1	2
Poisson	0	1	1	1	2
Biyolojik	0	0	0	0	1

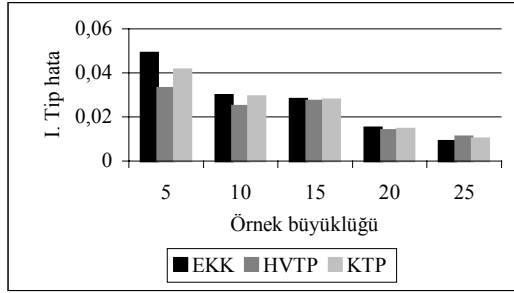


Şekil 2. Normal dağılım için farklı örnek büyüklüklerinde I. Tip hata oranları.

Şekil 2 incelendiğinde, örnek büyüklüğü arttıkça I. Tip hata oranlarının azaldığı görülmektedir. 5 örnek büyüklüğünde en yüksek I. Tip hatayı KTP yönteminin, en düşük I. Tip hatayı ise EKK yönteminin ürettiği görülmektedir. Ancak burada anormal gözlem bulunmadığından HVTP yöntemi KTP yöntemine göre daha güvenilirdir. Bu durumda KTP yöntemi 0.07438 değeri üretmekle yanlış olan hipotezi kabul ederek II. Tip hata üretmiştir. Bunun nedeni KTP yönteminin küçük örnek büyüklüklerinde yanlış sonuç üretme eğilimi göstermesi olabilir. Diğer örnek büyüklüklerinde benzer sonuçlar elde edilmiş ve hipotez hakkında verilen kararı etkileyen farklılık gözlemlenmedi. Kullanılan veri kümesi Normal dağılıma sahip olduğundan EKK yöntemi güvenilir sonuçlar üretebilmektedir, dolayısıyla zaman kaybını engellemek açısından permütasyon testlerinin Normal dağılıma sahip olan örneklerde kullanılmaması önerilebilir.

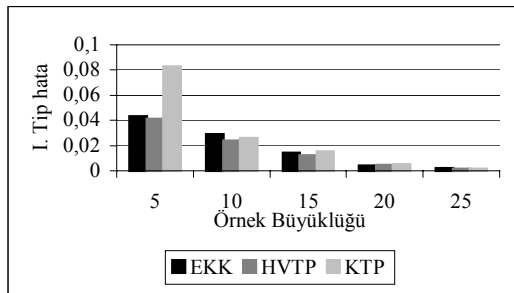
Çizelge 2. Hipotetik ve gerçek biyolojik verilere ait istatistik önem düzeyleri

Dağılımlar	Metotlar	Örnek büyüklüğü				
		5	10	15	20	25
Normal	EKK	0.03800	0.02900	0.01700	0.01300	0.00100
	HVTP	0.04159	0.02832	0.01500	0.01300	0.00200
	KTP	0.07438	0.02832	0.01673	0.01387	0.00093
Ki-kare	EKK	0.04900	0.03000	0.02800	0.01500	0.00900
	HVTP	0.03306	0.02500	0.02720	0.01400	0.01112
	KTP	0.04132	0.02924	0.02797	0.01437	0.01011
Poisson	EKK	0.04300	0.02900	0.01400	0.00400	0.00200
	HVTP	0.04132	0.02400	0.01236	0.00460	0.00167
	KTP	0.08264	0.02594	0.01499	0.00483	0.00148
Biyolojik	EKK	0.11800	0.02200	0.00800	0.00100	0.00000
	HVTP	0.10744	0.02305	0.00886	0.00160	0.00021
	KTP	0.20661	0.02391	0.00943	0.00185	0.00016



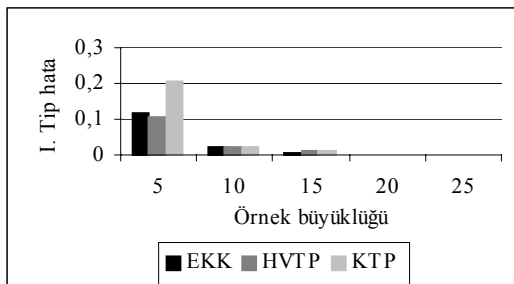
Şekil 3. Ki-kare dağılımı için farklı örnek büyüklüklerinde I. Tip hata oranları.

Ki-kare dağılımına sahip veri kümesi için örnek büyüklüğü 25’den küçük olduğu durumlarda en küçük I. Tip hata oranını HVTP yönteminin ürettiği Şekil 3’ten görülebilmektedir. Tüm örnek büyüklüklerinde sıfır hipotezi hakkında verilen kararda farklılık oluşturan değişimler söz konusu değildir. Ancak 10 ve daha büyük örnek büyüklüklerinde anormal gözlemler bulunması nedeniyle KTP yöntemi önerilebilir.



Şekil 4. Poisson dağılımı için farklı örnek büyüklüklerinde I. Tip hata oranları.

Poisson dağılımı için 20 örnek büyüklüğüne kadar en küçük I. Tip hata oranının HVTP yöntemi ile elde edildiği Şekil 4’de görülmektedir. Örnek büyüklüğü 5 için KTP yöntemi örnek büyüklüğünün yetersiz olması nedeniyle II. Tip hata üretmiştir. Diğer örnek büyüklüğünde sıfır hipotezi hakkında verilen kararı değiştirecek bir farklılık gözlemlenmemiştir. Örnek büyüklüğünün 10 ve daha fazla olduğu durumlarda veri kümesinin anormal gözlemler içermesinden dolayı Poisson dağılımı için KTP yöntemi önerilebilir.



Şekil 5. Biyolojik veri için farklı örnek büyüklüklerinde I. Tip hata oranları.

Gerçek biyolojik veriden elde edilen sonuçlar değerlendirildiğinde, tüm örnek büyüklüklerinde sıfır hipotezi için verilen kararı değiştiren farklılığın bulunmadığı görülmektedir. Örnek büyüklüğünün 5 olduğu durumda yetersiz örnek büyüklüğü nedeniyle KTP yönteminin kullanılması sakıncalıdır. 10 ve daha büyük örnek büyüklüklerinde ise EKK yöntemi en küçük I. Tip hata oranı üretmiştir. Veri kümesinde anormal gözlem değeri bulunmaması nedeniyle mevcut çalışma için HVTP yöntemi önerilebilmektedir.

Permütasyon testleri parametrik metodların ihtiyaç duyduğu varsayımlardan ve veri kümesinin sahip olduğu dağılımın şekline etkilenmemesi nedeniyle varsayımların sağlanmadığı ya da bu konuda yeterli bilginin elde edilemediği durumlarda başarıyla kullanılabilmektedir. Tusell (2001)’inde bildirdiği gibi özellikle normal olmayan dağılımlarda ve örnek büyüklüğü 15’den küçük olduğunda permütasyon testleri daha güvenilir sonuçlar üretebilmektedir.

Veri kümesinin normal dağılıma sahip olduğu varsayılsa bile varyansların homojen olmaması ya da değişkenler arasında istatistiksel olarak önemli korelasyonların bulunması durumunda *F* ve *t* testleri güçlerini kaybetmektedir. Permütasyon testlerinin bu sorunun çözümünde kullanılabilecek alternatif bir yöntem olduğu Lin ve Lee (2003), Bracken (2001), Legendre (2000), Tanizaki (2001) çalışmalarında da belirtilmiştir. Maggini ve ark (2002) normal olmayan dağılımlarda ve küçük örnek durumunda permütasyon testlerinin daha doğru I. Tip hata üretme eğiliminde olduğunu yinelemiş ve ayrıca veri kümesinde anormal gözlemlerin bulunması durumunda KTP, aksi durumda ise HVTP yönteminin kullanılması gerektiğini bildirmiştir.

Verinin normal dağılım gösterdiği ve değişkenler arasında yüksek korelasyon bulunmadığı durumda doğrusal regresyon analizi için bilgisayar işlem zamanı açısından EKK yöntemi permütasyon yöntemine tercih edilebilir.

#### 4. KAYNAKLAR

- Abecasis, G.R., Cardon, L.R. ve Cookson, W.O.C., 2000. A General Test of Association for Quantitative Traits in Nuclear Families. *Am. J. Hum. Genet.* 66: 279-292
- Anderson, M.J. ve Legendre, P., 1999. An Empirical Comparison of Permutation Methods for tests of Partial Regression Coefficients in a Linear Model. *J. Statist. Comput. Simul* 62: 271-303
- Anderson, M.J., 2001. Permutation Tests for Univariate or Multivariate Analysis of Variance and Regression. *Can. J. Fish. Aquat. Sci.* 58:626-639
- Anderson, M.J., 2003. DISTLM v.2.: A FORTRAN Computer Program to Calculate a Distance-Based Multivariate Analysis for Linear Model. Department of Statistics, University of Auckland, New Zealand.
- Anderson, M.J. ve Robinson, J., 2001. Permutation Tests for Linear Models. *Aust. N. Z. J. Stat.* 43(1): 75-88
- Bracken, M.B., 2001. On Stratification, Minimization and Protection Against Types 1 and 2 Error. *Journal of Clinical Epidemiology*, 54: 104 – 107.

- Fisher, R.A., 1935. Design of Experiments. Oliver and Body, Edinburgh.
- Gonzalez, L. ve Manly, B.F.J., 1998. Analysis of Variance by Randomization with Small Data Sets. *Environmetrics* 9: 53-65.
- Kleinbaum, D.G., Kupper, L., Muller, K.E. ve Nizam, A., 1998. Applied Regression Analysis and Other Multivariable Methods, Duxbury Press, Pacific Grove, 798 page.
- Legendre, P., 2000. Comparison of Permutation Methods for Partial Correlation and Partial Mantel Tests. *J. Statist. Comput. Simul.* 67: 37 – 73.
- Lin, S. ve Lee, J.C., 2003. Exact Test in Simple Growth Curve Models and One-Way ANOVA with Equicorrelation Error Structure. *Journal of Multivariate Analysis* 84: 351 – 368.
- Maggini, R., Guisan, A. ve Cherix, D., 2002. A Stratified Approach to Modeling the Distribution of a Threatened Ant Species in the Swiss National Park. *Biodiversity and Conservation* 11: 2117 – 2141.
- Makarenkov, V. ve Legendre, P., 2002. Nonlinear Redundancy Analysis and Canonical Correspondence Analysis Based on Polynomial Regression. *Ecology* 83: 1146-1161.
- Manly, B.F.J., 1997. Randomization, Bootstrap and Monte Carlo methods in biology, 2nd edition. Chapman and Hall, London.
- Nichols, T.E. ve Holmes, A.P., 2001. Nonparametric permutation tests For Functional Neuroimaging: A Primer with Examples. *Human Brain Mapping* 15:1-25
- O'Gorman, T.W., 2001. An Adaptive Permutation Test procedure for Several Common Tests of Significance. *Computational Statistics & Data Analysis* 35: 335 - 350
- Önder, H. ve Cebeci, Z., 2005. Use of Permutation Test on Nested Models. *International Congress on Information Technology in Agriculture, Food and Environment* 312-315. October 12-15 2005, Adana, Turkey.
- Tanizaki, H., 2001. On Small Sample Properties of Permutation Tests: An Independence Test between Two Samples and Significance Test for Regression Models. Accessed at [<http://ht.econ.kobe-u.ac.jp/~tanizaki/cv/working/permute.pdf>] Son erişim tarihi: 19.06.2003
- Tusell, F., 2001. A Permutation Test for Randomness with Power Against Smooth Variation. *Statistics and Computing* 11: 147 – 154.