ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

K Bağımsız Grubu Karşılaştırmak İçin Dayanıklı Yöntemler

Robust Methods for Comparing K Independent Groups

İrem YILMAZ,[a]
Fırat ÖZDEMİR[a]

[a]İstatistik AD,
Dokuz Eylül Üniversitesi Fen Fakültesi,
İzmir

Geliş Tarihi/*Received:* 15.02.2016
Kabul Tarihi/*Accepted:* 17.05.2016

Yazışma Adresi/*Correspondence:*
İrem YILMAZ
Dokuz Eylül Üniversitesi Fen Fakültesi
İstatistik AD, İzmir,
TÜRKİYE/TURKEY
iremyilmaz0@gmail.com

ÖZET Amaç: Tek yönlü ANOVA-F testi, ikiden fazla bağımsız grubun kitle ortalamaları arasında istatistiksel olarak anlamlı bir fark olup olmadığını incelemek için kullanılan en yaygın yöntemdir. Ancak bu yöntem, normal dağılım ve homojen varyanslılık varsayımlarının sağlanmadığı durumlarda yanıltıcı sonuçlar verebilmektedir. Güvenilir sonuçlar elde etmek için ANOVA-F testi yerine dayanıklı konum kestiricilerinin kullanıldığı yöntemler tercih edilebilir. Bu çalışmanın amacı, ANOVA-F testi ile birlikte altı farklı yöntemin nominal 1. Tip hatanın korunması ölçütüne göre 28 farklı deneysel durum altında karşılaştırılmasıdır. Bu deneysel durumlar; simetrik ve simetrik olmayan dağılımlar, homojen ve heterojen varyanslılık, eşit ve eşit olmayan örneklem genişliği ve örneklem genişliği ile varyansların pozitif ve negatif eşleşmesi şeklindeki bileşenlerin kombinasyonlarından oluşmaktadır. Ayrıca, altı farklı yöntem, Parkinson hastalarının konuşma egzersizlerinden elde edilmiş iki farklı gerçek veri seti içinde karşılaştırılmıştır. .**Gereç ve Yöntemler:** Karşılaştırılan yöntemler; ANOVA-F testi, Welch testi, budanmış ortalama ile Welch testi, budanmış ortalama ve bootstrap-t ile Welch testi, Box testi ve Kruskal Wallis testidir. **Bulgular:** Farklı deneysel durumlar altında gerçekleştirilen bir benzetim çalışması ile bu yöntemlere ait gerçekleşen 1. Tip hata değerleri hesaplanmış ve bir dayanıklılık ölçütüne göre başarılı bulunan dört yöntem için güç değerleri hesaplanmıştır. **Sonuç:** Bu çalışmada elde edilen tüm sonuçlara göre varsayımların sağlandığı durumlarda Welch testi, varsayımların sağlanmadığı durumlarda ise budanmış ortalama ile Welch testi önerilmektedir.

Anahtar Kelimeler: I. tip hata; güç; Welch testi; Box yöntemi; bootstrap-t; budanmuş ortalama

ABSTRACT Objective: One-way ANOVA-F test is the most common test to examine whether there is a statistically significant difference between population means of more than two independent groups or not. However, ANOVA-F test can give misleading results when normality and homoscedasticity assumptions are violated. In order to get reliable results, robust methods can be preferred instead of ANOVA-F test. The purpose of this study is comparing six different methods including ANOVA-F test in terms of controlling actual Type I error rates under 28 different experimental conditions. These experimental conditions comprise of the combinations of the components which are symmetric and asymmetric distributions, homoscedasticity and heteroscedasticity, equal and unequal sample sizes and positive and negative pairings of variances with sample sizes. Furthermore, these six different methods were compared for two real data sets which were obtained from speech exercises of Parkinson's patients. **Material and Methods:** Compared methods are ANOVA-F test, Welch test, Welch test with trimmed mean, Welch test with trimmed mean and a bootstrap-t, Box method, and Kruskal-Wallis test. **Results:** After computing actual Type I error rates of the methods with a simulation study, power rates are calculated for the four methods which controlled the actual Type I error rate according to a robustness criterion. **Conclusion:** Based on all results obtained from this study, Welch test is recommended when the assumptions are satisfied whereas Welch test with trimmed mean is recommended when the assumptions are violated.

Key Words: Type I error rate; power; Welch test; Box method; bootstrap-t; trimmed mean

doi: 10.5336/biostatic.2016-50837

Turkiye Klinikleri J Biostat 2016;8(2):143-51

Hipotez testi yöntemleri tıp, biyoloji, ziraat ve psikoloji gibi birçok alanda sıklıkla kullanılmaktadır. İkiden fazla kitlenin ortalamaları arasında istatistiksel olarak anlamlı bir fark olup olmadığı araştırılmak istendiğinde, en çok bilinen ve kullanılan hipotez testi yöntemi tek yönlü ANOVA-F testidir. ANOVA-F testi temel varsayımlarının sağlandığı durumlarda güçlü bir yöntemken, varsayımlarının ihlal edilmesinden çok etkilendiği için yanlış ve yanıltıcı sonuçlar verebilmektedir.[1]

ANOVA-F testinin ilk varsayımı, karşılaştırılan grupların elde edildiği kitlelerin normal dağılama sahip olmasıdır. Gerçek verilerde, kitle dağılımının normal olması oldukça nadir görülen bir durumdur.[2] Hatta ilgilenilen kitle dağılımları çok sayıda uç değerlere sahip, çarpık ve ağır kuyruklu da olabilirler. Bu gibi durumlarda örneklem ortalamasının standart hatası fazlasıyla büyüdüğü için ANOVA-F testine ait gerçekleşen 1. Tip hata da nominal değerinden uzaklaşır.

İkinci varsayım homojen varyanslılıktır ve kitle varyanslarının istatistiksel olarak eşit olması anlamına gelir. Ancak Normal dağılım varsayımı gibi bu varsayımda nadiren sağlanır. Kitle varyanslarının eşit olmaması durumuna heterojen varyanslılık denir. Heterojen varyanslılığın etkisi özellikle dengesiz tasarımlarda daha çok ortaya çıkar.[3] Örneğin, varyanslar ve örneklem genişlikleri arasında pozitif bir eşleşme (varyansı büyük olan kitleden çok, varyansı küçük olan kitleden az sayıda örneklemin seçilmesi) var ise ANOVA-F testi tutucu (gerçekleşen anlam düzeyinin nominal anlam düzeyinden küçük çıkması) sonuçlar verirken negatif eşleşme (varyansı büyük olan kitleden az, varyansı küçük olan kitleden çok sayıda örneklemin seçilmesi) durumunda bu test, liberal (gerçekleşen anlam düzeyinin nominal anlam düzeyinden büyük çıkması) sonuçlar vermektedir.

Normalliğin ve homojen varyanslılığın sağlanmadığı durumlarda ANOVA-F testine (F) alternatif ve varsayım bozulmalarına karşı dayanıklı birçok yöntem önerilmiştir. Bu çalışmada önerilen yöntemler arasından Welch testi (W), budanmış ortalama ile Welch testi (TW), budanmış ortalama ve bootstrap-t ile Welch testi (BTW), Box testi (BX) ve Kruskal Wallis (K) testi incelenmiştir.[3-5] Çalışmanın amacı, belirtilen 6 farklı yöntemin 1. Tip hatanın korunması ölçütüne göre 28 farklı deneysel durum altında karşılaştırılmasıdır. Yapılan benzetim çalışması ile elde edilen sonuçlar açısından en başarılı dört yöntem için güç değerleri de hesaplanmıştır.

GEREÇ VE YÖNTEMLER

WELCH TESTİ

Homojen varyanslılık varsayımını sağlamayan normal dağılıma sahip iki kitle konum ölçüsü arasında fark olup olmadığı incelendiğinde ortaya çıkan probleme Behrens-Fisher problemi denir.[6,7] Welch bu probleme çözüm olarak önerdiği iki bağımsız grubun karşılaştırması için kullanılan yöntemi, daha sonra ikiden fazla bağımsız grup karşılaştırması için geliştirmiştir.[4,8]

$$H_0 : \mu_1 = ... = \mu_k = \mu \quad (1)$$

$$F_w = \frac{\sum_{i=1}^{k} w_i(\bar{x}_i - \bar{x})^2}{k-1}\left[1 + \frac{2(k-2)}{k^2-1}\sum_{i=1}^{k}\left(\frac{1}{n_i-1}\right)\left(1-\frac{w_i}{\sum_{i=1}^{k} w_i}\right)^2\right]^{-1} \quad (2)$$

$$w_i = \frac{n_i}{s_i^2} \quad (3)$$

$$s_i^2 = \frac{\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2}{n_i - 1} \quad (4)$$

Sıfır hipotezi doğru olduğunda F_w, k-1 ve ν_w serbestlik dereceleri ile yaklaşık olarak F dağılır. $F_w \geq F_{k-1,\nu_w,\alpha}$ ise H_0 reddedilir.

$$\nu_w = \frac{k^2-1}{3\sum_{i=1}^{k}\left(\frac{1}{n_i-1}\right)\left(1-\left(\frac{1}{\sum_{i=1}^{k} w_i}\right)\right)^2} \quad (5)$$

BUDANMIŞ ORTALAMA İLE WELCH TESTİ

Welch testi heterojen varyanslılığa karşı dayanıklı bir yöntem iken, kitle dağılımlarının normallikten uzaklaşması durumuna karşı oldukça hassastır. Ancak Welch testi, örneklem ortalaması ve örneklem varyansı yerine budanmış ortalama ve winsorized varyans gibi dayanıklı kestiriciler ile kullanılırsa, iki varsayım ihlaline karşı da dayanıklı hale gelir.[5]

$$H_0 : \mu_{t1} = ... = \mu_{tk} = \mu_t \quad (6)$$

Budanmış ortalamayı hesaplamak için ilk olarak budama yüzdesi (γ) [0, 0,5] aralığından seçilir. Her bir kuyruktan budanacak gözlem sayısı λ ile gösterilir ve $\lambda = [\gamma n]$ şeklinde hesaplanır. n örneklem genişliğidir. Tüm gözlemler küçükten büyüğe doğru sıralandıktan sonra budama yapılır. Budama sonrası kalan örneklem sayısına etkin örneklem sayısı (h) denir ve $h = n - 2\lambda$ şeklinde hesaplanır. Budama işleminden sonra eşitlik (7)'de belirtildiği şekilde budanmış ortalama hesaplanır.

$$\bar{X}_t = \frac{1}{h}\sum_{\lambda+1}^{n-\lambda} X_i \quad (7)$$

Eşitlik (10)'da verilen Winsorized varyans, eşitlik (8)'de verilen winsorized gözlemler ve eşitlik (9)'da verilen winsorized ortalama kullanılarak hesaplanır.

$$Y_i = \begin{cases} X_{(\lambda+1)}, & X_i < X_{(\lambda+1)} \\ X_i, & X_{(\lambda+1)} < X_i < X_{(n-\lambda)} \\ X_{(n-\lambda)}, & X_i \geq X_{(n-\lambda)} \end{cases} \quad (8)$$

$$\bar{X}_w = \frac{1}{n}\sum_{j=1}^{n} Y_j \quad (9)$$

$$s_w^2 = \frac{1}{n-1}\sum_{j=1}^{n} (Y_j - \bar{X}_w)^2 \quad (10)$$

Yönteme ait test istatistiği, (11) ve (17) nolu eşitlikler arasında verildiği biçimdedir.

$$F_t = \frac{A}{B+1} \quad (11)$$

$$A = \frac{1}{k-1}\sum_{i=1}^{k} w_i (\bar{X}_{ti} - \tilde{X})^2 \quad (12)$$

$$B = \frac{2(k-1)}{k^2-1}\sum_{i=1}^{k} \frac{\left(1 - \frac{w_i}{u}\right)^2}{h_i - 1} \quad (13)$$

$$\tilde{X} = \frac{\sum_{i=1}^{k} w_i \bar{X}_{ti}}{U} \quad (14)$$

$$U = \sum_{i=1}^{k} w_i \quad (15)$$

$$w_i = \frac{1}{d_i} \quad (16)$$

$$d_i = \frac{(n_i - 1)s_{wi}^2}{h_i(h_i - 1)} \quad (17)$$

Sıfır hipotezi doğru olduğunda F_w , k-1 ve ν_{wt} serbestlik dereceleri ile yaklaşık olarak F dağılır. $F_t \geq F_{k-1,\nu_{wt},\alpha}$ ise H_0 reddedilir.

$$\nu_{wt} = \left(\frac{3}{k^2-1}\sum_{i=1}^{k} \frac{\left(1 - \frac{w_i}{U}\right)^2}{h_i} \right)^{-1} \quad (18)$$

BUDANMIŞ ORTALAMA VE BOOTSTRAP-T İLE WELCH TESTİ

%20 budanmış ortalama ve bootstrap-t yönteminin Welch testi ile beraber kullanımı 1. Tip hatanın kontrolü açısından başarılı sonuçlar vermektedir.[5,9] İlgilenilen hipotez (6)'daki gibidir.

Bootstrap-t yönteminin adımları aşağıda sıralanmıştır.

1. Tekrarlı seçim yapılarak k sayıda grubun herbiri için bir bootstrap örneklem türetilir. $X_{i1}^*, X_{i2}^*, ..., X_{in}^*$, $i = 1,2,...,k$

2. Her bir gruptaki gözlemlerden kendi grubuna ait budanmış ortalama çıkartılarak C_{ij}^{*} değerleri hesaplanır. $C_{ij}^{*} = X_{ij}^{*} - \bar{X}_{ti}$, $j = 1,2,...,n_i$

3. F_t^* ile gösterilen ve C_{ij}^{*} değerleri ile hesaplanan test istatistiği bulunur.

4. İlk üç adım B kez tekrarlanır.

5. Elde edilen B tane test istatistiği küçükten büyüğe sıralanır. $F_{t(1)}^{*} \leq ... \leq F_{t(B)}^{*}$

6. $U = (1-\alpha)B$ değeri hesaplanır ve en yakın tam sayıya yuvarlanır.

7. Kritik değer olarak U. test istatistiği $F_{t(U)}^{*}$ bulunur.

8. Orijinal örneklemden elde edilen ve budanmış ortalama ile Welch testindeki gibi hesaplanan F_t, bootstrap-t ile elde edilen kritik değerden büyük ya da eşit ise sıfır hipotezi reddedilir.

BOX TESTİ

Homojen varyanslılık ve normallik varsayımlarının ihlal edildiği durumlara karşı önerilen bir diğer dayanıklı yöntemde Box yöntemidir.[3,5] İlgilenilen hipotez (6)'daki gibidir.

$$F_b = \frac{\sum_{i=1}^{k} h_i (\bar{X}_{ti} - \bar{X}_t)^2}{\sum_{i=1}^{k} 1 - \left(\frac{h_i}{H}\right) S_i^2} \quad (19)$$

$$H = \sum_{i=1}^{k} h_i \quad (20)$$

$$\bar{X}_t = \sum_{i=1}^{k} \frac{h_i \bar{X}_{ti}}{H} \quad (21)$$

$$S_i^2 = \frac{(n_i - 1) s_{wi}^2}{h_i - 1} \quad (22)$$

$$\nu_1 = \frac{\left[\sum_{i=1}^{k} (1-f_i) S_i^2\right]^2}{\left(\sum_{i=1}^{k} f_i S_i^2\right)^2 + \sum_{i=1}^{k} S_i^4 (1-2f_i)} \quad (23)$$

$$\nu_2 = \frac{\left[\sum_{i=1}^{k} (1-f_i) S_i^2\right]^2}{\sum_{i=1}^{k} \frac{S_i^4 (1-f_i)}{(h_i - 1)}} \quad (24)$$

$$f_i = \frac{h_i}{H} \quad (25)$$

Sıfır hipotezi doğru olduğunda F_b, ν_1 ve ν_2 serbestlik dereceleri ile yaklaşık olarak F dağılır. $F_b \geq F_{\nu_1, \nu_2, \alpha}$ ise H_0 reddedilir.

BENZETİM ÇALIŞMASI

Altı farklı yöntemden elde edilen gerçekleşen 1. Tip hata değerleri üç bağımsız grubu karşılaştırmak için 28 farklı deneysel durum altında incelenmiştir. Bu deneysel durumlar; normallik ve homojen varyanslılık varsayımlarının sağlandığı ve sağlanmadığı, örneklem genişliklerinin eşit ya da eşit olmadığı, örneklem genişliği ve varyansların pozitif ya da negatif eşlenmesi gibi durumların kombinasyonlarından elde edilmiştir. Teorik dağılımlardan normal dağılım ve g-h (g=0,5, h=0,2) dağılımı kullanılmıştır. Bunların dışında Parkinson hastalarıyla ilgili iki gerçek veri seti için de tüm karşılaştırmalar yapılmıştır. Nominal anlam düzeyi 0,05 olarak seçilip, her bir durum için 10.000 tekrar yapılmıştır. R (3.1.2) programlama dili ile yapılan benzetim çalışmasında kullanılan tüm örneklem genişlikleri ve varyanslar Tablo 1'de gösterilmiştir.

g-h DAĞILIMI

g-h dağılımı, çarpıklığın kontrol edilebileceği g ($-1<g<1$) ve basıklığın kontrol edilebileceği h ($0<h<1$) parametreleri ile normal dağılımdan uzaklaşmayı derecelendirerek gözleme imkanı veren ve g=h=0 için standart normal dağılıma eşdeğer bir dağılımdır.

Z standart normal dağılımdan gelen bir rassal değişken olmak üzere;

TABLO 1: Örneklem genişlikleri ve varyanslar.

Örneklem genişliği	Homojen varyanslılık	Heterojen varyanslılık
20, 20, 20	1, 1, 1	4, 25, 81
30, 30, 30	1, 1, 1	4, 25, 81
20, 30, 40	1, 1, 1	4, 25, 81; 81, 25, 4

$g \neq 0$ için

$$X = \frac{(\exp(gZ)-1)\exp\left(\frac{hZ^2}{2}\right)}{g} \quad (26)$$

$g = 0$ için

$$X = Z\exp\left(\frac{hZ^2}{2}\right) \quad (27)$$

dönüşümleri kullanılarak g-h dağılımından veri üretilir.[10]

BULGULAR

Bradley, kullanılan yöntemin nominal anlam düzeyini korumada yeterli olup olmadığını belirlemek için liberal ($0,5\alpha \leq \hat{\alpha} \leq 1,5\alpha$) ve tutucu ($0,9\alpha \leq \hat{\alpha} \leq 1,1\alpha$) olarak adlandırdığı iki farklı dayanıklılık ölçütü öne sürmüştür.[11] Bu çalışmada, Bradley'nin liberal ve tutucu dayanıklılık ölçütlerinin arasında olan ve benzer çalışmalarda da görülen ($\mathbf{0,8\alpha \leq \hat{\alpha} \leq 1.2\alpha}$) ölçütü kullanılmıştır.[12] Bu ölçüte göre, $\alpha = 0,05$ iken teste ait gerçekleşen 1. Tip hata değeri [0,040, 0,060] aralığında ise test dayanıklıdır şeklinde yorum yapılabilmektedir. Bulunan gerçekleşen 1. Tip hata değerleri bu ölçüte göre karşılaştırılmış ve en başarılı dört yöntem için güç değerleri hesaplanmıştır.

GERÇEKLEŞEN1.TİP HATA DEĞERLERİ

Tablo 3, 4, 5, 6 sırasıyla normal, g-h dağılımı ve Parkinson hastalarına ait iki gerçek veri seti kullanılarak hesaplanan 1. tip hata değerlerini göstermektedir. Bu tabloların hepsinde, açık gri ile işaretlenmiş hücreler liberal sonuçları gösterirken, koyu gri ile işaretlenmiş hücreler tutucu sonuçları göstermektedir.

Normal dağılım için elde edilen sonuçlara bakıldığında, F ve K testleri hariç tüm yöntemlerin 1. tip hata kontrolünü her koşul altında sağladığı görülmektedir. Normal dağılım varsayımı altında F ve K testleri pozitif eşleşme hariç, homojen varyanslılık varsayımının sağlanmadığı tüm durumlar için liberal sonuçlar vermişlerdir. Ayrıca, F testi pozitif eşleşme durumunda tutucu sonuçlar vermiştir. Bu sonuçlar F testinin negatif eşleşmede liberal, pozitif eşleşmede tutucu sonuç verme eğilimini desteklemektedir (Tablo 3).

Tablo 4'de g-h dağılım için elde edilen sonuçlar bulunmaktadır. g parametresi 0,5 ve h parametresi 0,2 olarak belirlenmiştir. Bu parametrelere göre ilgilenilen g-h dağılımının çarpıklık ve basıklık kestirimleri sırasıyla $\hat{\kappa}_1 =$ 8,06 ve $\hat{\kappa}_2 =$212,28 şeklindedir. Bu nedenle, bu parametreler ile kullanılan g-h dağılımı oldukça çarpık ve ağır kuyruklu bir dağılımdır. Sonuçlardan da görüldüğü üzere, normallik varsayımının sağlanmadığı ve çarpık bir dağılımın kullanıldığı bu koşullar altında Tablo 3'e göre belirlenen 1. tip hatanın korunma sıklığı azalmıştır. Bu örnekte, budanmış ortalama ile kullanılan yöntemler (TW, BTW ve BX) 1. tip hata kontrolünde başarılı olurken dayanıklı kestiricilere sahip olmayan yöntemlerin (F, K, W) özellikle varyans homojenliğinin sağlanmadığı koşullar altında liberal sonuçlar verdikleri görülmektedir (Tablo 4).

Sıradaki iki uygulama gerçek veriler kulanılarak yapılmıştır. http:/www.archive.ics.uci.edu/ml/ adresinden alınan bu veri setleri Parkinson hastalarının konuşma egzersizlerinden elde edilmiş ve minimum ses perdesi ve titreşim sayısı isimlerini taşımaktadır. Tablo 2'de bu iki veri setine ait tanımlayıcı istatistikler bulunmaktadır.[13]

TABLO 2: Gerçek veri setlerine ait tanımlayıcı istatistikler.

Veri Seti	N	Ortalama	s	Çarpıklık	Basıklık
Minimum SesPerdesi	1040	27.058	36.673	2,469	9,798
TitreşimSayısı	1040	234.876	121.541	1,154	3,056

TABLO 3: Normal dağılım için gerçekleşen 1. tip hata oranları.

Gruplar	n	σ	F	K	W	TW	BTW	BX
k=3	20 20 20	1 1 1	0,0496	0,0463	0,0492	0,0505	0,0437	0,0458
	20 20 20	2 5 9	0,0686	0,0664	0,0499	0,0547	0,0471	0,0537
	30 30 30	1 1 1	0,0506	0,0504	0,0519	0,0538	0,0508	0,0508
	30 30 30	2 5 9	0,0699	0,0697	0,0534	0,0572	0,0508	0,0531
	20 30 40	1 1 1	0,0511	0,0513	0,0516	0,0503	0,0460	0,0482
	20 30 40	2 5 9	0,0318	0,0428	0,0545	0,0532	0,0517	0,0542
	20 30 40	9 5 2	0,1398	0,0965	0,0477	0,0518	0,0437	0,0511

TABLO 4: g-h(g=0,5, h=0,2) dağılım için gerçekleşen 1. tip hata değerleri.

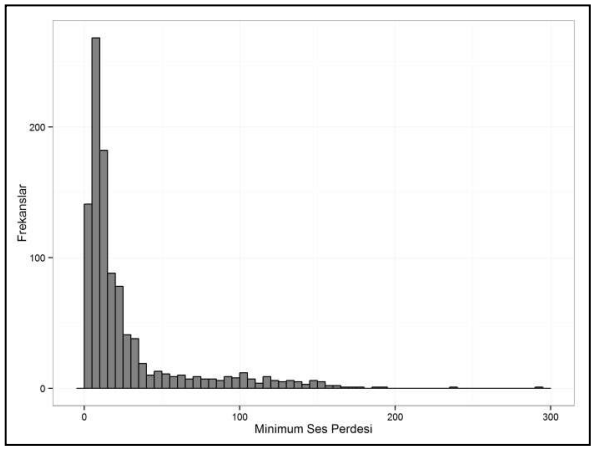
Gruplar	n	σ	F	K	W	TW	BTW	BX
k=3	20 20 20	1 1 1	0,0376	0,0433	0,0404	0,0374	0,0372	0,0334
	20 20 20	2 5 9	0,0770	0,0625	0,0850	0,0503	0,0463	0,0498
	30 30 30	1 1 1	0,0442	0,0514	0,0487	0,0463	0,0453	0,0435
	30 30 30	2 5 9	0,0757	0,0675	0,0854	0,0519	0,0489	0,0492
	20 30 40	1 1 1	0,0430	0,0507	0,0503	0,0476	0,0480	0,0407
	20 30 40	2 5 9	0,0393	0,0406	0,0710	0,0477	0,0482	0,0476
	20 30 40	9 5 2	0,1476	0,0992	0,0947	0,0562	0,0482	0,0491

TABLO 5: Minimum ses perdesi veri seti için gerçekleşen 1. tip hata değerleri.

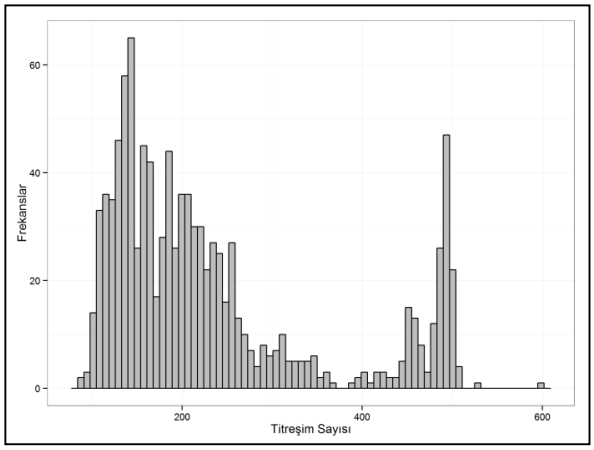
Gruplar	n	σ	F	K	W	TW	BTW	BX
k=3	20 20 20	1 1 1	0,0410	0,0429	0,0589	0,0270	0,0218	0,0161
	20 20 20	2 5 9	0,0976	0,0878	0,1364	0,0757	0,0607	0,0447
	30 30 30	1 1 1	0,0434	0,0486	0,0599	0,0260	0,0201	0,0152
	30 30 30	2 5 9	0,0878	0,1032	0,1069	0,0648	0,0554	0,0403
	20 30 40	1 1 1	0,0401	0,0441	0,0677	0,0285	0,0241	0,0150
	20 30 40	2 5 9	0,0425	0,0612	0,0820	0,0489	0,0446	0,0310
	20 30 40	9 5 2	0,1583	0,1487	0,1284	0,0873	0,0671	0,0525

TABLO 6: Titreşim sayısı very seti için gerçekleşen 1. tip hata değerleri.

Gruplar	n	σ	F	K	W	TW	BTW	BX
k=3	20 20 20	1 1 1	0,0485	0,0472	0,0531	0,0346	0,0257	0,0222
	20 20 20	2 5 9	0,0734	0,0791	0,0725	0,0494	0,0382	0,0403
	30 30 30	1 1 1	0,0514	0,0487	0,0555	0,0321	0,0244	0,0241
	30 30 30	2 5 9	0,0685	0,0865	0,0584	0,0470	0,0388	0,0390
	20 30 40	1 1 1	0,0500	0,0480	0,0594	0,0346	0,0290	0,0264
	20 30 40	2 5 9	0,0283	0,0518	0,0534	0,0389	0,0337	0,0302
	20 30 40	9 5 2	0,1399	0,1233	0,0698	0,0537	0,0394	0,0461



ŞEKİL 1: Minimum ses perdesine ait histogram.



ŞEKİL 2: Titreşim sayısına ait histogram.

Tablo 5'de minimum ses perdesi veri seti ile elde edilmiş 1. tip hata değerleri bulunmaktadır. Şekil 1'den de görüldüğü üzere bu veri seti oldukça çarpık ve ağır kuyrukludur. Bu nedenle tüm yöntemler 1. tip hata kontrolünü sağlamakta zorlanmışlardır. Dayanıklı kestiricilere sahip olmayan yöntemler genellikle liberal sonuçlar verme eğilimdeyken, dayanıklı kestiriciler ile kullanılan yöntemler çoğunlukla tutucu sonuçlar vermişlerdir (Tablo 5).

Tablo 6'da ise titreşim sayısı veri setine ait 1. tip hata değerleri bulunmaktadır. Şekil 2' den de görüldüğü üzere titreşim sayısı, minimum ses perdesine göre çarpıklığı daha az bir veri setidir. Bu yüzden de Tablo 5'deki sonuçlara göre Tablo 6'da 1. tip hata kontrolü daha iyi sağlanmıştır. Bu koşullar altında 1. tip hata kontrolünü en iyi sağlayan yöntem W testidir (Tablo 6).

GÜÇ DEĞERLERİ

1. tip hata kontrolünü en iyi sağlayan W, TW, BTW ve BX testleri için güç değerleri hesaplanmıştır. Teorik dağılımlar için etki büyüklüğü (Δ) 1 seçilmiştir. Minimum ses perdesi ve titreşim sayısı için ise sırasıyla 25 ve 100 seçilmiştir. Tablo 7 ve 8 bu dört yöntem için güç değerlerini göstermektedir.

Tablo 7'de normal dağılım altında yedi farklı koşul için dört yönteme ait güç değerleri bulunmaktadır. Heterojen varyanslılık, yöntemlere ait güç değerlerini düşürmüştür. Ancak, varyansların homojen olduğu koşullarda dört yöntemde yüksek ve birbirine yakın güç değerlerine sahip olmuşlardır. Karşılaştırılan dört yöntem içerisinden, W testinin diğer üç yönteme göre daha yüksek güç değerlerine sahip olduğu gözlenmektedir (Tablo 7).

Tablo 8'de ise g-h dağılımı kullanılarak elde edilen güç değerleri bulunmaktadır. TW testi, bu dağılım altındaki tüm koşullarda diğer üç yönteme göre daha yüksek güç değerlerine sahip olmuştur. Dayanıklı kestiricilere sahip olmayan W testinin gücü, normallikten uzaklaşma olduğu durumda diğer yöntemlere göre daha düşük çıkmıştır (Tablo 8).

Gerçek veri setlerine ait gerçekleşen 1. tip hata değerleri çoğunlukla belirlenen (0.04, 0.06) limitlerinin dışında kaldığı için bu veri setlerine ait güç tablolarına yer verilmemiştir.

SONUÇ

ANOVA-F testi, ikiden fazla grubun karşılaştırılmasında en çok kullanılan hipotez testi yöntemi olmasına rağmen, varsayımlarının ihlal edildiği durumlarda güvenilir bir test değildir. Bu çalışmada, ANOVA-F testi ve alternatif olarak geliştilen

TABLO 7: Normal dağılım için güç değerleri.

Gruplar	n	σ	F	K	W	TW	BTW	BX
k=3	20 20 20	1 1 1	1	0,8971	0,8221	0,7931	0,8291	20 20 20
	20 20 20	2 5 9	1	0,0972	0,0949	0,0808	0,0613	20 20 20
	30 30 30	1 1 1	1	0,9795	0,9555	0,9475	0,9574	30 30 30
	30 30 30	2 5 9	1	0,1310	0,1206	0,1095	0,0735	30 30 30
	20 30 40	1 1 1	1	0,9793	0,9536	0,9440	0,9566	20 30 40
	20 30 40	2 5 9	1	0,1246	0,1153	0,1106	0,0764	20 30 40
	20 30 40	9 5 2	1	0,1286	0,1190	0,0991	0,0749	20 30 40

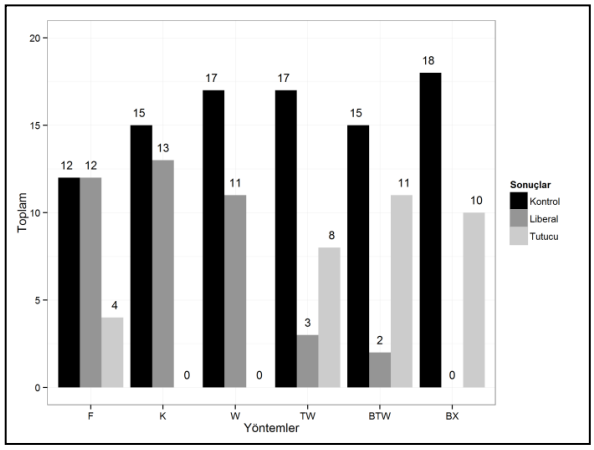
TABLO 8: g-h (g=0,5, h=0,2) dağılımı için güç değerleri.

Gruplar	n	σ	F	K	W	TW	BTW	BX
k=3	20 20 20	1 1 1	1	0,8971	0,8221	0,7931	0,8291	20 20 20
	20 20 20	2 5 9	1	0,0972	0,0949	0,0808	0,0613	20 20 20
	30 30 30	1 1 1	1	0,9795	0,9555	0,9475	0,9574	30 30 30
	30 30 30	2 5 9	1	0,1310	0,1206	0,1095	0,0735	30 30 30
	20 30 40	1 1 1	1	0,9793	0,9536	0,9440	0,9566	20 30 40
	20 30 40	2 5 9	1	0,1246	0,1153	0,1106	0,0764	20 30 40
	20 30 40	9 5 2	1	0,1286	0,1190	0,0991	0,0749	20 30 40

bir parametrik olmayan ve dört dayanıklı yöntem, gerçekleşen 1. Tip hata değerlerine göre 28 farklı deneysel durum altında bir benzetim çalışması ile karşılaştırılmışlardır.

Elde edilen sonuçlara göre, varsayımların sağlandığı ve örneklem genişliklerinin eşit olduğu durumlarda, F testi ve diğer tüm alternatif yöntemler 1. Tip hata kontrolünü sağlamışlardır. Ancak varsayım ihlallerinin olduğu, örneklem genişliklerinin eşit olmadığı, özellikle negatif ve pozitif eşleşmenin bulunduğu durumlarda F testi, 1. Tip hata kontrolünü sağlayamamıştır. Dayanıklı yöntemlerden TW, BTW ve BM özellikle teorik dağılımlar altında incelenen 14 durum içinden sadece bir tanesinde tutucu sonuç vermiş, geri kalan 13 durumda belirlenen nominal 1. Tip hata değerini korumuşlardır. Kullanılan gerçek veri setlerinden birincisi için yöntemler çoğunlukla 1. Tip hata kontrolünde başarılı olamamıştır. Ayrıca, dayanıklı kestiricilere sahip yöntemler tutucu sonuçlar vermişken, dayanıklı kestiricilere sahip olmayan yöntemler de liberal sonuçlar gözlenmiştir. Buradan da anlaşılacağı üzere, bu çalışmada yapılan benzetim çalışmasında, dayanıklı yöntemlerin, 1. Tip hata kontrolünü sağlayamadıkları durumlarda tutucu sonuç verme eğiliminde oldukları görülmektedir. İkinci gerçek veri setine göre ise yöntemler, 1. gerçek veri setine göre daha iyi performans göstermişlerdir.

28 durum altında, yöntemlerin gösterdiği genel performanslar Şekil 3'de gösterilmiştir. Yapılan benzetim çalışması sonuçlarına göre W (17), TW (17) ve BX (18) 1. Tip hata kontrolünü en iyi sağlayan üç yöntemdir (Şekil 3).



ŞEKİL 3: Yöntemlerin nominal anlam düzeyini korudukları deneysel durumların sayıları.

1. Tip hata kontrolü ölçütüne göre en iyi sonuçları veren dört yöntem (W, TW, BTW, BX) için güç hesaplanmıştır. W yöntemi, normallik ve homojen varyanslılık varsayımlarının sağlandığı tüm durumlar altında, diğer üç yönteme göre daha yüksek güç değerleri vermiştir. Ancak, normallik varsayımının sağlanmadığı g-h dağılımı altında W yönteminin budanmış ortalama ve winsorized varyans ile kullanılan şekli olan TW en yüksek güce sahiptir. Normalliğin bozulduğu durumlarda, dayanıklı kestiricilerin kullanıldığı yöntemlere ait güç değerleri artmıştır.

Genel olarak elde edilen tüm sonuçlara bakıldığında, belirlenen 1. Tip hatayı korumada en başarılı olan W (17), TW (17) ve BX (18) yöntemleri içerisinden, varsayımların sağlandığı durumlar altında W testi, güç açısından en iyi sonuçlara sahipken, varsayımların ihlal edildiği durumlar altında TW testi en iyi güce sahiptir. Bu sonuçlara bağlı olarak varsayımların sağlandığı durumlar için Welch (W) ve varsayımların sağlanmadığı durumlar için ise Budanmış ortalama ile Welch (TW) testlerinin kullanılması önerilmektedir.

KAYNALAR

1. Cribbie AR, Fiksenbaum L, Keselman JH, Wilcox RR. Effect of non-normality on test statistics for one-way independent groups designs. Br J Math Stat Psychol 2012;65(1):56-73.
2. Micceri T. The unicorn, the normal curve, and other improbable creatures. Psychological Bulletin 1989;105(1):156-66.
3. Box GEP. Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. Ann Math Statist 1954;25(2):290-302.
4. Welch BL. On the comparison of several mean values: an alternative approach. Biometrika 1951;38(3-4):330-6.
5. Wilcox RR. One-way and higher desings for independent groups. Introduction to Robust Estimation and Hypothesis Testing. 3rded. San Diego, California: Academics Press; 2012. p.291-377.
6. Behrens WV. Ein beitrag zur fehlerberechnung bei wenigen beobachtungen. Landwirtsch Jahrbücher 1929;68:807-37.
7. Fisher RA. The fiducial argument in statistical inference. Annals of Eugenics 1935;6(4):391-8.
8. Welch BL. The generalization of student's problem when several different population variances are involved. Biometrika 1947;34(1-2):28-35.
9. Westfall PH, Young SS. Resampling-based adjustments: basic concepts. Resampling Based Multiple Testing: Examples and Methods for P-Value Adjustment. 1sted. New York: Wiley; 1993. p.32-75.
10. Hoaglin DC, Mosteller F, Tukey JW. Summarizing shape numerically:the g-and-h distributions. Exploring data tables, trends and shapes. New York: Wiley; 1985. p.461-513.
11. Bradley JV. Robustness? Br J Math Stat Psychol 1978;31(2):144-52.
12. Koh A, Cribbie R. Robust tests of equivalence for k independent groups. Br J Math Stat Psychol 2013;66(3):426-34.
13. Sakar BE, Isenkul ME, Sakar CO, Sertbas A, Gurgen F, Delil S, et al. Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. IEEE J Biomed Health Inform 2013; 17(4):828-34.