International Journal of Scientific and Technological Research
ISSN 2422-8702 (Online)
Vol 4, No.7, 2018

Comparing Performances of Homogeneity Tests Used for Intraclass Version of Kappa

Harika Gozde Gozukara Bag (Corresponding author)
Department of Biostatistics and Medical Informatics
School of Medicine, Inonu University, Malatya, Turkey
E-mail: harika.gozukara@inonu.edu.tr

Celal Reha Alpar
Department of Biostatistics
School of Medicine, Hacettepe University, Ankara, Turkey
E-mail: ralpar@hacettepe.edu.tr

Abstract
The reliability of a measure is an important component of the quality of the measurement. Reliability can be defined as repeatability or consistency of duplicates in a measurement process. In many fields, some studies are reliability studies which are based on assessment of agreement between observations or observers. In this study, we considered the most common usage of intraclass kappa statistic which has been the widely accepted measure for assessing the reliability between two ratings on a binary trait. In a meta–analysis of kappa statistics obtained from multiple studies using the same measure, in multicenter studies or in a stratified study, we would like to compare kappa statistics and present a common or summary kappa agreement using all available information. A homogeneity test is required for an overall kappa estimation of two or more independent kappa coefficients. In this study, the aim was to compare the Fleiss, Donner's goodness-of-fit, Likelihood Score, Modified Score and the Pearson's goodness-of-fit test statistics which are used to test the homogeneity of two or more independent intraclass kappa statistics. The test procedures were evaluated separately under the assumption of equal prevelances and unequal prevelances. To be able to compare the tests by taking Type I error rate and power into the consideration, Monte Carlo approach with 10000 simulations was used. Under the assumption of equal prevelances; Pearson's goodness-of-fit test indicated the best performance in terms of Type I error rate, Fleiss test was tend to be liberal because it is based on large sample variance. Under unequal prevelances; Donner's goodness-of-fit and Modified Score tests displayed better performance than under the assumption of equal prevelances, Fleiss test found to be liberal for testing the homogeneity of more than two kappa statistics, Type I error rate of Likelihood Score test was at nominal level and exhibited the best performance.
Keywords: Agreement, intraclass kappa statistic, homogeneity test, chi–square distribution, reliability

Sınıf İçi Kappa İstatistiği İçin Kullanılan Homojenlik Testlerinin Performanslarının Karşılaştırılması

Özet
Bir ölçümün güvenirliği, ölçümün kalitesinin önemli bir bileşenidir. Güvenirlik, bir ölçüm sürecinde, ölçüm işleminin tekrarlanabilirliği ya da tekrarlardaki tutarlılık olarak tanımlanmaktadır. Birçok alanda, gözlemler ya da gözlemciler arasındaki uyumun değerlendirilmesi temeline dayanan güvenirlik çalışmaları yapılmaktadır. Bu çalışmada, güvenirliğin değerlendirilmesinde yaygın olarak kullanımı kabul edilmiş olan sınıf içi kappa istatistiğinin, en sık karşılaşılan şekli olan iki kategorili sonuçlar üzerinde, iki değerlendirme arasındaki uyumun incelenmesinde kullanıldığı durum dikkate alınmıştır.

Aynı temel ölçümü içeren birçok uyum çalışmasından elde edilen kappa istatistiklerinin meta analizinde, çok merkezli çalışmalarda ya da tabakalanmış tek bir çalışmada, elde edilen kappa istatistikleri karşılaştırılmak ve eldeki tüm bilgi kullanılarak ortak ya da özet bir kappa uyumu verilmek istenir. Bağımsız iki ve daha fazla kappa istatistiği için tümel bir kestirim elde edilmek istendiğinde öncellikle bu katsayıların homojen olup olmadığının test edilmesi gerekir. Bu çalışmada, iki ve daha çok bağımsız sınıf içi kappa istatistiğinin homojenliğinin test edilmesinde kullanılan Fleiss, Donner uyum iyiliği, Olabilirlik Skor, Modifiye Edilmiş Skor ve Pearson uyum iyiliği test istatistiklerinin performanslarının karşılaştırılması amaçlanmıştır. Homojenlik testleri, evren prevelanslarının eşit olduğu ve olmadığı varsayımları altında ayrı ayrı incelenmiştir. Tip I hata ve güç açısından testlerin karşılaştırılması için 10000 tekrar ile Monte Carlo benzetim tekniği kullanılmıştır. Evren prevelanslarının eşit olduğu varsayımı altında; Pearson uyum iyiliği testi Tip I hata açısından en iyi performansı göstermiştir, Fleiss testi ise büyük örneklem varyansı temeline dayandığı için küçük örneklemlerde liberal olma eğilimdedir. Evren prevelanslarının eşit olmadığı varsayımı altında ise; Donner uyum iyiliği ve Modifiye Edilmiş Skor testleri daha iyi performans göstermiştir, Fleiss testi ikiden fazla katsayısının test edilmesinde liberal bulunmuştur, Skor homojenlik testi ise Tip I hata açısından nominal seviyede kalarak en iyi performansı göstermiştir.
Anahtar kelimeler: Uyum, sınıf içi kappa istatistiği, homojenlik testi, ki–kare dağılımı, güvenirlik

1. GİRİŞ

Güvenirlik; bir ölçüm sürecinde, ölçüm işleminin tekrarlanabilirliği ya da tekrarlardaki tutarlılık olarak tanımlanabilir (Alpar, 2003). Bir ölçüm belli derecede güvenirliğe sahip olmadığı sürece geçerli bir ölçüm olamayacağı için (Alpar, 2010), güvenirliğin ispatı bir ölçümün kalitesinin onaylanmasında önemli bir ilk adım olarak görülür (Kraemer et al., 2002). Birçok alanda, gözlemler ya da gözlemciler arasındaki uyumun değerlendirilmesi temeline dayanan güvenirlik çalışmaları yapılmaktadır.

Cohen (1960), gözlemcilerin cevaplarının gözlenen marjinal dağılımlarının ve gözlemci değerlendirmelerinin istatistiksel olarak bağımsız olduğu varsayımı ile gözlenen uyum oranını beklenen uyum seviyesine göre azaltarak kappayı şans–düzeltmeli bir uyum ölçüsü olarak önermiştir (Banerjee, 1999). Bu ölçü, Cohen kappa olarak bilinir. Cohen kappa gözlemcilerin pozitif olarak yaptıkları değerlendirmenin marjinal olasılıklarının farklı olmasına izin verir (Bishop et al., 1989; Nam, 2002). Bloch ve Kraemer (1989) tarafından ele alınan alternatif bir yaklaşım, her gözlemcinin pozitif olarak değerlendirme olasılığının aynı olduğunu varsayar. Bu yaklaşım, tek yönlü varyans analizinden hesaplanan sınıf içi korelasyon katsayısı kestirimi gibi elde edilen ve cebirsel olarak Scott uyum indeksine (Scott, 1955) eşit olan, kappanın sınıf içi versiyonunu ortaya çıkarır. Özet olarak, bu iki kappa katsayısı arasındaki fark; Cohen kappa istatistiğinde birinci ve ikinci değerlendirmelerde pozitif sınıflama olasılıkları farklı olabilirken, sınıf içi kappa istatistiğinde ise bu olasılıkların eşit olduğunun varsayılmasıdır (Nam, 2002).

Bu çalışmada, güvenirliğin değerlendirilmesinde yaygın olarak kullanımı kabul edilmiş olan sınıf içi kappa istatistiğinin, en sık karşılaşılan durum olan iki kategorili sonuçlar üzerinde, iki değerlendirme arasındaki uyumun incelenmesinde kullanıldığı durum dikkate alınmıştır.

Bazı durumlarda, güvenirliğin ya da değerlendirmeler arasındaki uyumun incelenmesi için tek bir sınıf içi kappa istatistiğinin doğrudan elde edilmesi uygun olmayabilir. Örneğin; J tane bağımsız kappa istatistiği, çok merkezli bir klinik denemede olduğu gibi bağımsız çalışmalardan ya da bireylerin J tabakaya ayrıldığı tek bir çalışmadan elde edilebilir. Aynı temel ölçüyü içeren birçok uyum çalışmasından elde edilen kappa istatistiklerinin meta analizinde, çok merkezli çalışmalarda ya da tabakalanmış tek bir çalışmada, elde edilen kappa istatistikleri karşılaştırılmak ve eldeki tüm bilgi kullanılarak ortak ya da özet bir kappa uyumu verilmek istenir. Eğer elde edilen bu J tane kappa istatistiğinin homojenlik testi reddedilmezse, bütün çalışmaları özetleyen tek bir kappa istatistiği üzerinden çıkarımlar yapma olanağı elde edilmiş olur. Tabakalar arasında hem kappaların homojenliği hem de eşit prevalans varsayımı sağlandığında birleştirilmiş veri üzerinden istatistiksel analizler yapılabilir (Nam, 2003). Ancak, eşit prevalans varsayımı geçerli değilse, birleştirilmiş veriden elde edilen ortak kappa kestirimi tutarlı değildir. Bu durumda, ortak kappa değerlendirmesi için veri birleştirirken dikkatli olunması gerekir (Nam, 2003). Bağımsız iki ve daha fazla kappa istatistiği için tümel bir kestirim elde edilmek istendiğinde öncellikle bu katsayıların homojen olup olmadığının test edilmesi gerekir.

Bu çalışmada, iki ve daha çok bağımsız sınıf içi kappa istatistiğinin homojenliğinin test edilmesinde kullanılan Fleiss, Donner uyum iyiliği, Olabilirlik Skor, Modifiye Edilmiş Skor ve Pearson uyum iyiliği test istatistiklerinin performanslarının Tip I hata ve güç açısından karşılaştırılması amaçlanmıştır.

2. MATERYAL VE METOD

2.1. Sınıf İçi Kappa İstatistiği

Ölçülen sonuç değişkeni 1 (pozitif) ve 0 (negatif) gibi iki durumlu olduğunda iki değerlendirme arasındaki uyumun incelenmesi için 4 gözlü bir tablo hazırlanır. Sınıf içi kappa istatistiği hesaplanırken, gözlemciler arasında yanlılık olmadığı varsayıldığı için, diğer bir deyişle 1. ve 2. değerlendirmelerde pozitif olarak sınıflama olasılıkları eşit kabul edildiği için uyumsuz gözeler birleştirilir. Bu durumda, uyumun incelendiği 4 göze, 3 gözeye indirgenmiş olur. Bu 3 gözedeki frekanslara bağlı olarak hesaplanan sınıf içi kappa istatistiği için yeniden oluşturulan uyum tablosu J çalışma için Tablo 1'deki gibidir.

Bu tablodaki kategoriler (2, 1 ve 0), bir değerlendirme çiftindeki pozitif değerlendirme sayısını temsil etmektedir ve kullanılan ilk alt indis kategoriyi, ikinci alt indis ise çalışmayı temsil etmektedir. n_j, birey sayısını göstermek üzere, bu çalışmada prevalans olarak adlandırılan, j. çalışmadaki k. gözlemcinin i. birey için yaptığı değerlendirmenin pozitif(1) olma olasılığı aşağıdaki eşitlik ile elde edilir.

$$\hat{p}_j = \tilde{p}_j = \Pr\left(X_{ikj}=1\right) = \frac{\left(2x_{2j}+x_{1j}\right)}{\left(2n_j\right)} \qquad (1)$$

Tablo 1. J çalışma için uyum tablosu

Kategori	Değerlendirme	Çalışma				Toplam	Olasılık
		1	2	…	J		
2	(1,1)	x_{21}	x_{22}	…	x_{2J}	$x_{2.}$	$P_{2j}(\kappa)$
1	(1,0) veya (0,1)	x_{11}	x_{12}	…	x_{1J}	$x_{1.}$	$P_{1j}(\kappa)$
0	(0,0)	x_{01}	x_{02}	…	x_{0J}	$x_{0.}$	$P_{0j}(\kappa)$
Toplam		n_1	n_2	…	n_J	$n_.$	1

Üç kategoride toplanan göze olasılıkları, n'e koşullu multinomial dağılım gösteren ve Eşitlik 2'de verilen ortak korelasyon modeli yardımıyla kestirilir (Mak, 1988; Bloch & Kraemer, 1989; Nam, 2003). Bu model, herhangi bir çift (X_{i1}, X_{i2}) arasındaki korelasyonun (κ) aynı değere sahip olduğunu varsaydığı için, ortak korelasyon modeli olarak adlandırılır (Bloch & Kraemer, 1989; Donner et al., 1996; Ridout et al, 1999). Her bir çalışmaya ($j=1,\ldots,J$) özel göze olasılıklarının gösterimi Eşitlik 2'deki gibidir (Donner et al., 1996).

$$\begin{aligned} P_{2j}\left(\kappa_j, p_j\right) &= \Pr\left(X_{i1j}=1, X_{i2j}=1\right) = p_j^2 + p_j\left(1-p_j\right)\kappa_j \\ P_{1j}\left(\kappa_j, p_j\right) &= \Pr\left(X_{i1j}=0, X_{i2j}=1 \; yada \; X_{i1j}=1, X_{i2j}=0\right) = 2p_j\left(1-p_j\right)\left(1-\kappa_j\right) \\ P_{0j}\left(\kappa_j, p_j\right) &= \Pr\left(X_{i1j}=0, X_{i2j}=0\right) = \left(1-p_j\right)^2 + p_j\left(1-p_j\right)\kappa_j \end{aligned} \qquad (2)$$

Ortak korelasyon modeli altında sınıf içi kappa istatistiğinin ve prevalansın en çok olabilirlik kestirimleri aşağıdaki gibidir (Bloch & Kraemer, 1989).

$$\tilde{\kappa} = \frac{\left(4x_2x_0 - x_1^2\right)}{\left(2x_2+x_1\right)\left(2x_0+x_1\right)} \qquad \tilde{p} = \frac{2x_2+x_1}{2n} \qquad (3)$$

2.2. Homojenlik Testleri

Farklı çalışmalardan ya da tek bir çalışmanın farklı tabakalarından elde edilen sınıf içi kappa istatistiklerinin homojenliğinin test edilmesinde kullanılması önerilen testler aşağıdaki başlıklar altında toplanabilir.

- Fleiss homojenlik testi (Fleiss, 1981; Donner et al., 1996; Nam, 2003; Nam, 2006)
- Donner uyum iyiliği testi (Donner GOF-goodness-of-fit) (Donner et al., 1996)
- Olabilirlik Skor Homojenlik testi (Nam, 2003; Nam, 2006)
- Modifiye Edilmiş Skor Homojenlik testi (MES) (Nam, 2003)
- Pearson uyum iyiliği testi (Pearson GOF-goodness-of-fit) (Nam, 2006)

Sınıf içi kappa istatistiği ile ölçülen *J* tane uyumun çalışmalar ya da tabakalar arasında homojen kabul edilip edilemeyeceğini test etmek için kurulacak hipotezler ise aşağıdaki gibidir.

$$H_0 : \kappa_1 = \kappa_2 = \cdots = \kappa_J$$
$$H_1 : \kappa_j \neq \kappa$$

Bağımsız çalışmalardan ya da tabakalara ayrılmış tek bir çalışmadan birçok sınıf içi kappa istatistiği elde edildiğinde, bu kappaların homojen olup olmadığı araştırılmak istenir. Bir homojenlik testinin yapılması ise ortak kappa kestirimini gerektirir. Önerilen farklı ortak kestirim yaklaşımları vardır. Burada, her test prosedürünün kullandığı ortak kappa kestirimi, test alt başlıkları altında verilmiştir.

2.2.1. Fleiss Homojenlik Testi

Fleiss (1981), *J* tane bağımsız kappa'nın homojenliğinin test edilmesi için bir ki–kare testi önermiştir. Donner ve ark. (1996), sınıf içi kappa istatistiklerinin homojenliğinin test edilmesi için, Fleiss (1981) metodolojisinin kullanılabileceğini göstermiştir. Bu metodoloji kappanın büyük örneklem varyansı ile yakından bağlantılıdır.

Evren prevalanslarının eşit olduğu varsayımına gerek duymaksızın, her bir *j* için $\hat{p}_j = \tilde{p}_j$ ve $\hat{\kappa}_j = \tilde{\kappa}_j$ olmak üzere, diğer bir deyişle tabakaya ya da örnekleme özel en çok olabilirlik kestirimlerinin kullanılmasıyla *j*. çalışma için kappanın varyans kestirimi aşağıdaki eşitlikle elde edilebilir.

$$\hat{V}\left(\hat{\kappa}_j\right) = \frac{1-\hat{\kappa}_j}{n_j}\left[\left(1-\hat{\kappa}_j\right)\left(1-2\hat{\kappa}_j\right) + \frac{\hat{\kappa}_j\left(2-\hat{\kappa}_j\right)}{2\hat{p}_j\left(1-\hat{p}_j\right)}\right] \tag{4}$$

Fleiss homojenlik testi için ortak kappa kestirimi ise her bir tabakadan elde edilen katsayıların ağırlıklı ortalaması ile tanımlanmıştır ve her bir tabakanın ağırlığı ise varyanslarının tersi alınarak elde edilir.

$$\hat{\kappa}_\omega = \frac{\sum_j \hat{\omega}_j \hat{\kappa}_j}{\sum_j \hat{\omega}_j} \qquad \hat{\omega}_j = \frac{1}{\hat{V}\left(\hat{\kappa}_j\right)} \tag{5}$$

Evren prevalanslarının eşit olma varsayımına gerek duymayan Fleiss (Dif_Fleiss, χ^2_{difF}) test istatistiği, *(J–1)* serbestlik dereceli ki-kare dağılımı gösterir.

$$\chi^2_{difF} = \sum_{j=1}^{J} \hat{\omega}_j \left(\hat{\kappa}_j - \hat{\kappa}_\omega\right)^2 \tag{6}$$

Evren prevalanslarının eşit olduğu varsayımı geçerli iken Nam (2006), her bir tabakanın kendine özel en çok olabilirlik kestirimleri ($\hat{\kappa}_j$ ve $\hat{p}_j$) ile tabakaların varyanslarını ayrı ayrı hesaplamak yerine, birleştirilmiş veriden elde edilen en çok olabilirlik kestirimlerini ($\tilde{\kappa}_\bullet$ ve $\tilde{p}_\bullet$) kullanarak tabakaları sadece örneklem genişliklerine göre ağırlıklandırmıştır.

$$\tilde{\kappa}_\bullet = \frac{4x_{2\bullet}x_{0\bullet} - x_{1\bullet}^2}{\left(2x_{2\bullet} + x_{1\bullet}\right)\left(2x_{0\bullet} + x_{1\bullet}\right)} \tag{7}$$

$$\tilde{p}_{\bullet} = \frac{2x_{2\bullet} + x_{1\bullet}}{2n_{\bullet}} \tag{8}$$

$$\hat{V} = (1-\tilde{\kappa}_{\bullet})^2\left[(1-2\tilde{\kappa}_{\bullet}) + \frac{\tilde{\kappa}_{\bullet}(2-\tilde{\kappa}_{\bullet})}{2\tilde{p}_{\bullet}(1-\tilde{p}_{\bullet})(1-\tilde{\kappa}_{\bullet})}\right] \qquad \hat{\omega}_j(\tilde{\kappa}_{\bullet}, \tilde{p}_{\bullet}) = \frac{n_j}{\hat{V}} \tag{9}$$

Eşitlik 7-9 gösterimleriyle, *(J–1)* serbestlik dereceli ki-kare dağılımı gösteren, eşit prevalans varsayımı altında çalışan Fleiss (Eq_Fleiss) homojenlik test istatistiği aşağıdaki gibidir.

$$\chi^2_{eqF} = \sum_{j=1}^{J} \hat{\omega}_j(\tilde{\kappa}_{\bullet}, \tilde{p}_{\bullet})(\hat{\kappa}_j - \tilde{\kappa}_{\bullet})^2 \tag{10}$$

2.2.2. Donner Uyum İyiliği Testi

Donner ve ark. (1996), *J* sınıf içi kappa istatistiğinin homojenliğinin test edilmesi için uyum iyiliği yaklaşımını kullanarak evren prevalanslarının eşit olduğu varsayımına gerek duymayan bir homojenlik testi geliştirmişlerdir.

Tablo 1 gösterimlerinin kullanılmasıyla; her *j* tabaka için n_j'ye koşullu multinomial dağılım gösteren üç kategorinin olasılıkları $\hat{P}_{ij}(\kappa)$, H_0 hipotezinin doğru olduğu varsayımı altında ortak korelasyon modeli yardımı ile, $P_{ij}(\kappa)$'da p_j yerine $\hat{p}_j$ ve κ_j yerine $\hat{\kappa}$ (ortak kappa kestiriminin) kullanılmasıyla kestirilir. Donner ve ark. (1996), üç kategorinin olasılıkların kestirilmesinde kendi önerdikleri ortak kappa kestirimini $(\hat{\kappa}_D)$ kullanmışlardır. Çalışmaya özel en çok olabilirlik kestirimlerinin ($\hat{p}_j = \tilde{p}_j$ ve $\hat{\kappa}_j = \tilde{\kappa}_j$) kullanılmasıyla ağırlıklandırılmış ortalama alınarak elde edilen ortak kappa kestirimi Eşitlik 11 ile elde edilir.

$$\hat{\kappa}_D = \frac{\sum_j n_j \hat{p}_j \hat{q}_j \hat{\kappa}_j}{\sum_j n_j \hat{p}_j \hat{q}_j} = 1 - \left[\frac{x_{1\bullet}}{2\sum_{j=1}^{J} n_j \hat{p}_j (1-\hat{p}_j)}\right] \tag{11}$$

Ortak korelasyon modeliyle kestirilen olasılıklar yardımıyla elde edilen, Donner ve ark. (1996) tarafından önerilen test istatistiği, χ^2_D (Donner GOF), H_0 altında *(J–1)* serbestlik dereceli ki-kare dağılımı gösterir.

$$\chi^2_D = \sum_{i=0}^{2}\sum_{j=1}^{J} \frac{\left[x_{ij} - n_j \hat{P}_{ij}(\hat{\kappa}_D, \hat{p}_j)\right]^2}{\left[n_j \hat{P}_{ij}(\hat{\kappa}_D, \hat{p}_j)\right]} \tag{12}$$

2.2.3. Olabilirlik Skor Homojenlik Testi

Nam (2003), evren prevalanslarının eşit olma varsayımına gerek duymayan Donner uyum iyiliği testinin, parametrelerin en çok olabilirlik kestirimlerine dayandırılmaması nedeniyle tümüyle optimal olmayabileceğini düşünmüştür. Bu nedenle, evren prevalanslarının eşit olmadığı varsayımı altında parametrelerin en çok olabilirlik kestirimlerinin Olabilirlik Skor yöntemiyle kullanılmasıyla, Olabilirlik Skor Homojenlik test istatistiğini önermiştir.

$\tilde{p}_j \neq p$ olduğu zaman ortak kappanın en çok olabilirlik kestirimi iteratif bir işlem ile elde edilir ve kapalı bir formda gösterilemez.

$\ln L_j(\kappa, p_j) = x_{2j} \ln\{p_j(p_j + q_j\kappa)\} + x_{1j} \ln\{2p_jq_j(1-\kappa)\} + x_{0j} \ln\{q_j(q_j + p_j\kappa)\}$ olmak üzere, *J* tablo için ortak kappa ile log–olabilirlik aşağıdaki gibi yazılabilir:

$$\ln L(\kappa, p) = \sum_{j=1}^{J} \ln L_j(\kappa, p_j) \tag{13}$$

$S_\kappa(\kappa, p_j) \equiv \frac{\partial \ln L_j}{\partial \kappa}$ ve $S_j(\kappa, p_j) \equiv \frac{\partial \ln L_j}{\partial p_j}$ kısmı türevler olmak üzere ve $S_\kappa(\kappa, p) = \sum_{j=1}^{J} S_\kappa(\kappa, p_j)$ olarak gösterildiğinde, κ ve p'nin en çok olabilirlik kestirimleri $(J+1)$ kısmi denklemin çözümüdür $\{j=1,2,...,J$ için $S_\kappa(\kappa, p) = 0$ ve $S_j(\kappa, p_j) = 0\}$ (Nam, 2003).

İteratif bir süreç ile elde edilen $\tilde{\kappa}$ ve $\tilde{p}_j$, sırasıyla κ ve p_j'nin en çok olabilirlik kestirimlerini göstermek üzere, $j=1,2,...,J$ çalışma için skor istatistiği ve varyansı sırasıyla aşağıdaki eşitliklerde verilmiştir.

$$S_\kappa(\tilde{\kappa}, \tilde{p}_j) = \frac{\frac{x_{2j}}{(\tilde{p}_j + \tilde{q}_j\tilde{\kappa})} + \frac{x_{0j}}{(\tilde{q}_j + \tilde{p}_j\tilde{\kappa})} - n_j}{(1-\tilde{\kappa})} \tag{14}$$

$$\upsilon(\tilde{\kappa}, \tilde{p}_j) = \frac{n_j}{\left[(1-\tilde{\kappa})\left\{(1-\tilde{\kappa})(1-2\tilde{\kappa}) + \frac{\tilde{\kappa}(2-\tilde{\kappa})}{(2\tilde{p}_j\tilde{q}_j)}\right\}\right]} \tag{15}$$

$z_j(\tilde{\kappa}, \tilde{p}_j) = \frac{S_\kappa(\tilde{\kappa}, \tilde{p}_j)}{\sqrt{\upsilon(\tilde{\kappa}, \tilde{p}_j)}}$ olmak üzere, evren prevalanslarının eşit olma varsayımına gerek duymayan Olabilirlik Skor istatistiği (Dif_Skor), *(J–1)* serbestlik dereceli ki-kare dağılımı gösterir.

$$\chi^2_{difS} = \sum_{j=1}^{J} z_j^2(\tilde{\kappa}, \tilde{p}_j) \tag{16}$$

Ayrıca, evren prevalanslarının eşit olduğu varsayımı geçerli olduğunda, Nam (2006) eşit prevalans varsayımı altında birleştirilmiş veriden elde edilen ortak prevalans $(\tilde{p}_\bullet)$ ve ortak kappanın $(\tilde{\kappa}_\bullet)$ en çok olabilirlik kestirimlerinin skor yöntemiyle kullanılmasıyla elde ettiği χ^2_{eqS} (Eq_Skor) test istatistiğini önermiştir.

Eşitlik 14 ve 15'de, Eşitlik 7 ve 8'de verilen $(\tilde{\kappa}_\bullet)$ ve $(\tilde{p}_\bullet)$'nin kullanılmasıyla eşit prevalans varsayımı altında *(J–1)* serbestlik dereceli ki-kare dağılımı gösteren χ^2_{eqS} skor istatistiği aşağıdaki gibi gösterilebilir.

$$\chi^2_{eqS} = \sum_{j=1}^{J} \frac{\{S_j(\tilde{\kappa}_\bullet, \tilde{p}_\bullet)\}^2}{\upsilon(\tilde{\kappa}_\bullet, \tilde{p}_\bullet)} \tag{17}$$

2.2.4. Modifiye Edilmiş Skor Homojenlik Testi

Nam (2003), skor homojenlik testinin teorisinin, Donner ve ark. (1996) tarafından önerilen tutarlı ve iteratif olmayan ağırlıklandırılmış ortak kappa kestirimi ile kullanılmasıyla Modifiye Edilmiş Skor homojenlik testini geliştirmiştir.

Eşitlik 14 ve 15'de, Eşitlik 11 ile verilen $(\hat{\kappa}_D)$ ve tabakaya özel prevalans kestirimlerinin kullanılması ile asimptotik olarak *(J–1)* serbestlik dereceli ki–kare dağılımı gösteren Modifiye Edilmiş Skor (MES) istatistiği aşağıdaki gibidir.

$$\chi^2_M = \sum_{j=1}^{J} \frac{\{S_\kappa(\hat{\kappa}_D, \hat{p}_j)\}^2}{\upsilon(\hat{\kappa}_D, \hat{p}_j)} - \frac{\sum_{j=1}^{J}\{S_\kappa(\hat{\kappa}_D, \hat{p}_j)\}^2}{\sum_{j=1}^{J}\upsilon(\hat{\kappa}_D, \hat{p}_j)} \tag{18}$$

2.2.5. Pearson Uyum İyiliği Testi

Nam (2006), eşit prevalans $p_j = p$ varsayımı altında Pearson uyum iyiliği testinin kappa istatistiklerinin homojenliğinin test edilmesinde kullanılabileceğini göstermiştir.

Prevalansların eşit olduğu varsayımı altında, birleştirilmiş veriden elde edilen (Eşitlik 7 ve 8) ortak kappanın $(\tilde{\kappa}_\bullet)$ ve ortak prevalansın $(\tilde{p}_\bullet)$ en çok olabilirlik kestirimlerinin kullanılmasıyla ortak korelasyon modeli altında Pearson uyum iyiliği istatistiği aşağıdaki gibi yazılabilir. H_0 altında Pearson uyum iyiliği test istatistiği *2(J−1)* serbestlik dereceli ki-kare dağılımı gösterir.

$$\chi^2_P = \frac{\sum_{j=1}^{J} x_{2j}^2/n_j}{\tilde{p}_\bullet(\tilde{p}_\bullet + \tilde{q}_\bullet\tilde{\kappa}_\bullet)} + \frac{\sum_{j=1}^{J} x_{1j}^2/n_j}{2\tilde{p}_\bullet\tilde{q}_\bullet(1-\tilde{\kappa}_\bullet)} + \frac{\sum_{j=1}^{J} x_{0j}^2/n_j}{\tilde{q}_\bullet(\tilde{q}_\bullet + \tilde{p}_\bullet\tilde{\kappa}_\bullet)} - n_\bullet \qquad (19)$$

2.3. Benzetim Çalışması

Sınıf içi kappa katsayılarının homojenliğinin test edilmesinde kullanılan Donner uyum iyiliği, Fleiss, Olabilirlik Skor, Modifiye Edilmiş Skor ve Pearson uyum iyiliği test istatistiklerinin değişik koşullar altında performanslarının incelenmesi amacıyla yapılan bu çalışmada, R (2.11.1) programı kullanılmıştır. Test istatistiklerinin performanslarının incelenmesi için yapılan Monte Carlo benzetim çalışmasında gözlemler ortak korelasyon modeli altında multinomial dağılımdan türetilmiştir. Çalışmada dikkate alınan her bir senaryo için 10000 tekrar kullanılmıştır.

Evren prevalanslarının eşit olduğu varsayımı altında, Donner uyum iyiliği, Eq_Fleiss, Pearson uyum iyiliği, Eq_Skor ve Modifiye Edilmiş Skor homojenlik testleri dikkate alınmıştır. Evren prevalansları eşit olmadığı durumda ise Donner uyum iyiliği, Dif_Fleiss, Dif_Skor ve Modifiye Edilmiş Skor testlerinin performansı incelenmiştir.

Bu çalışmada prevalans olarak adlandırılan pozitif sınıflama olasılığının 0 ya da 1 olması durumda sınıf içi kappa istatistiği tanımsızdır. Bu nedenle, bu durumda hiçbir test hesaplanamamaktadır. En uçtaki prevalans değerlerinde bu tanımsızlıkla karşılaşma olasılığı yüksek olduğu için yapılan benzetim çalışmasında, prevalans değerleri 0,2 ile 0,8 arasında alınmıştır.

Homojenliği test edilecek olan herhangi bir örneklemdeki κ değeri 1 olduğunda evren prevalansları eşit olmadığı durumda dikkate alınan Dif_Fleiss homojenlik testi hesaplanamamaktadır. Kappanın 1'e yakın değerleri için, özellikle büyük olmayan örneklem genişliklerinde bu olasılığın yüksek olması nedeniyle dikkate alınan tüm test istatistiklerinin aynı koşullar altında karşılaştırılması amacıyla çalışmada en yüksek kappa değeri 0,80 olarak alınmıştır. Ayrıca, tüm örneklemlerden elde edilen kappa değerlerinin 1 olması durumunda ise hiçbir test istatistiği hesaplanamamaktadır. Ancak, Donner ve ark. (1996) bu koşul altında H_0 kabul şeklinde karara vardıklarını belirtmişlerdir.

Homojenliği test edilecek olan sınıf içi kappa sayısı 2 olduğunda, örneklem büyüklüğünün testlerin performanslarına etkisini incelemek amacıyla, n_j=50 ve n_j=100 eşit örneklem büyüklükleri, farklı örneklem genişlikleri için ise n_1=50 ve n_2=100 düzenlerinde çalışılmıştır. 3 bağımsız tabakadan ya da çalışmadan elde edilen kappa istatistiklerinin homojenliğinin test edilmesinde ise dikkate alınan örneklem genişlikleri, n_j=50, n_j=100 ve n_1=50, n_2=100, n_3=150 şeklindedir. Özellikle meta analizi çalışmalarında homojenliği test edilecek olan kappa istatistiklerinin sayısının fazla olabileceği görülmüştür. Bu nedenle, testlerin performansları ayrıca 6 bağımsız sınıf içi kappa istatistiğinin homojenliğinin test edilmesinde de incelenmiştir. Çalışmaların örneklem genişliklerinin eşit olması durumda n_j=50 ve n_j=100, eşit olmaması durumunda ise n_1=50, n_2=60, n_3=70, n_4=80, n_5=90 ve n_6=100 düzenleri kullanılmıştır.

Öngörülen senaryolarda, üretilen verinin prevalansının 0 ya da 1 olması durumunda sınıf içi kappa tanımsız olduğu için bu veriler silinerek iterasyon sayısı tamamlanana kadar yeni veriler türetilmiştir.

Prevalansların eşit olmadığı senaryolarda ise Dif_Fleiss homojenlik testinin hesaplanamaması nedeniyle herhangi bir örneklemdeki sınıf içi kappa kestirimi 1 olduğu durumda bu veriler silinerek iterasyon sayısı tamamlanana kadar yeni veriler türetilmiştir.

Prevalansların eşit olmadığı koşulda kullanılan Olabilirlik Skor Homojenlik testinde (Dif_Skor) parametrelerin en çok olabilirlik kestirimleri kullanılmaktadır. Bazı veriler için olabilirlik fonksiyonunu maksimize eden parametrelerin elde edilmesi aşamasında yapılan iteratif sürecin sonuçsuz kaldığı gözlemlenmiştir. Seyrek karşılaşılan bir durum olmasına rağmen iterasyon sayısını tamamlamak amacıyla fonksiyonun çözümsüz kaldığı veriler silinerek yerine yenisi türetilmiştir.

3. BULGULAR

Testlerin performanslarının incelenmek istendiği koşullar altında olası senaryo sayısının çok fazla olması nedeniyle, benzetim çalışmasının sonuçlarını içeren grafiklerde, sunum kolaylığı olması açısından sınırlı sayıda kombinasyon seçilmiştir.

3.1. İki bağımsız sınıfiçi kappa istatistiği için elde edilen bulgular

İki bağımsız sınıf içi kappanın homojenliğinin test edilmesinde tüm örneklem genişliklerinde Pearson uyum iyiliği testi dışında tüm testlerin uç prevalans değerlerinde (p_j=0,2 ve p_j=0,8) Tip I hatalarının nominal seviyeden (0,05) yüksek olma eğilimleri dikkat çekmektedir. Eşit prevalans varsayımı altında hesaplanan Fleiss (Eq_Fleiss) ve Modifiye Edilmiş Skor (MES) testlerinin Tip I hatası benzer bir eğilim göstermekle beraber nominal seviyeden biraz yüksek bulunmuştur. Eşit prevalans varsayımı altında hesaplanan Olabilirlik Skor (Eq_Skor) testi ise Tip I hata açısından genel olarak nominal seviyededir. Eşit örneklem genişliklerinde, örneklem genişliği artınca testlerin Tip hataları nominal seviyeye yaklaşmıştır. Eşit olmayan örneklem genişliği n_1=50, n_2=100 kombinasyonunda ise testler n_j=50'ye göre daha iyi, n_j=100'e göre daha kötü performans sergilemişlerdir. Pearson uyum iyiliği testi ise tüm prevalans ve örneklem genişliği düzenlerinde nominal seviyede kalarak eşit prevalans varsayımı altında 2 sınıf içi kappa istatistiğinin homojenliğinin test edilmesinde Tip I hata açısından en iyi performansı göstermiştir.

Kappa katsayıları arasındaki fark sabitken kappa değerleri büyüdükçe ve prevelans 0,5'e yaklaştıkça tüm testlerin güçlerinin arttığı gözlenmiştir (Şekil 1). Örneklem genişliğinin eşit olmadığı n_1=50 ve n_2=100 düzenlerindeki güç değerleri ise n_j=50'ye göre yüksek n_j=100'e göre daha düşük bulunmuştur. Pearson uyum iyiliği testi en düşük güce sahipken, Donner GOF ile Eq_Fleiss testleri ve Eq_Skor ile MES testleri güç açısından benzer bulunmuştur. Kappalar arasındaki fark az ve örneklem genişliği küçükken tüm testlerin güç değeri 0,80'den düşük bulunmuştur.

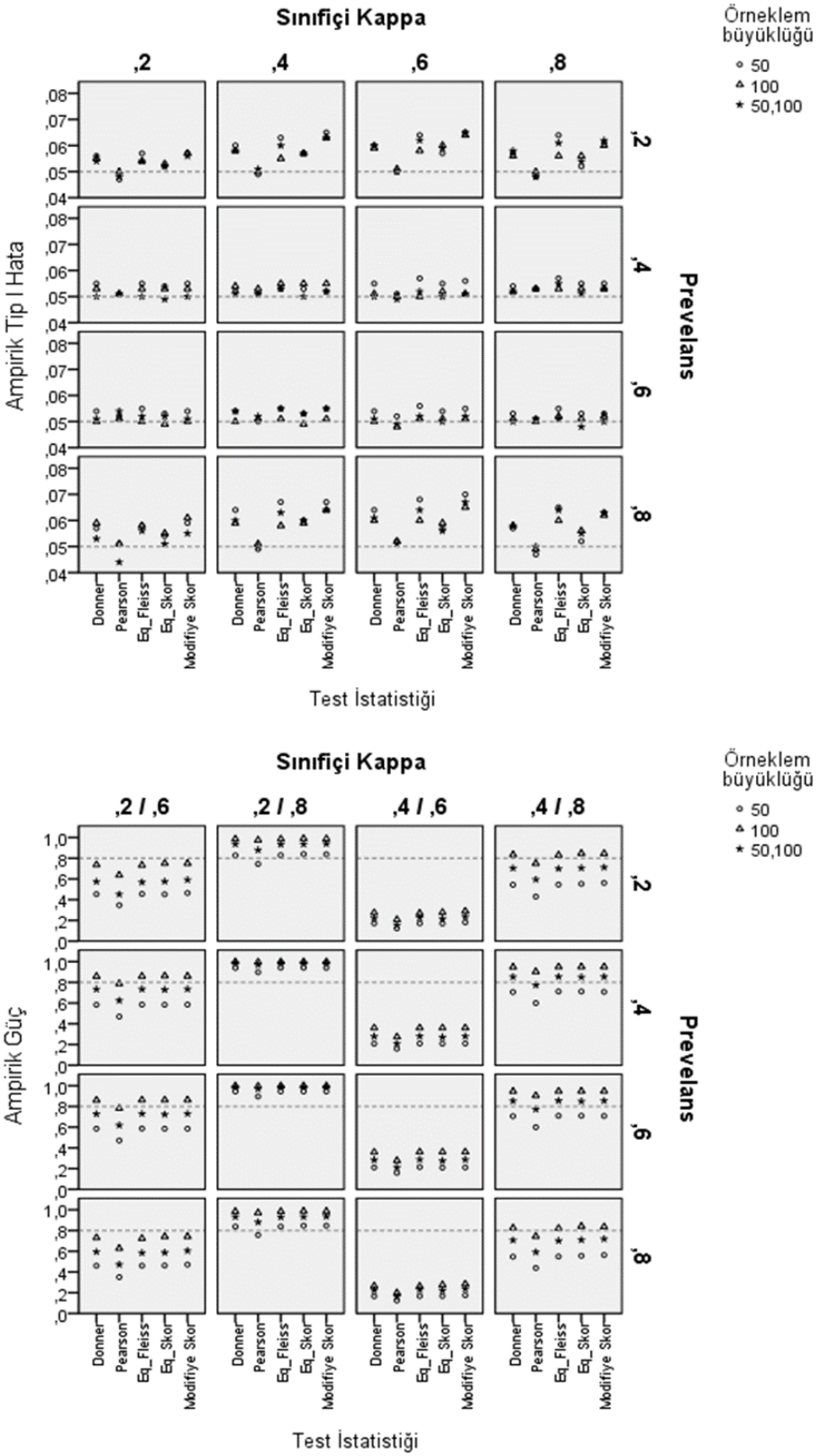
Prevalanslar eşit olmadığında iki sınıf içi kappa istatistiğinin homojenliğinin test edilmesinde prevalanslar arasındaki fark 0,3'ü geçmediği sürece Donner GOF testinin nominal seviyede, prevalanslar arasındaki fark 0,6 olduğu durumda bile nominal seviyeye yakın ve eşit prevalans düzeninden daha iyi performansa sahip olduğu görülmüştür. Evren prevalanslarının eşit olma varsayımına gerek duymayan Dif_Fleiss testi Tip I hata açısından n_j=100 dışında nominal seviyeden yüksek, prevalanslar arasındaki fark 0,2 ve daha yüksek olduğunda ise liberal bulunmuştur. Prevalanslar arasındaki fark sabit olsa bile 0,5'den uzak prevalans değerlerinde Dif_Fleiss testinin Tip I hatası daha da yükselmiştir. Dif_Skor testi prevalanslar arasındaki fark 0,6 olsa bile tüm örneklem genişliği kombinasyonlarında nominal seviyede ve Tip I hata açısından en iyi testtir (Şekil 2). Modifiye Edilmiş Skor testi ise prevalanslar arasındaki fark 0,4'den fazla olduğunda ve eşit olmayan örneklem genişliğinde nominal seviyeden biraz yüksek olmakla beraber diğer koşullarda genel olarak nominal seviyeye yakındır ve eşit prevalans altında çalıştığı kombinasyonlara göre Tip I hata düzeyi daha iyidir.

Prevalanslar arasındaki fark sabit olduğunda düşük prevalans ve küçük kappa değerlerinde Dif_Fleiss testi diğer testlerden daha güçlü olmakla beraber genel olarak küçük prevalanslarda en güçlü test Modifiye Edilmiş Skor testidir. Düşük prevalanslarda tüm örneklem genişliklerinde Donner GOF testi, Dif_Skor testinden daha güçlü bulunmuştur. Ancak, yüksek prevalans değerlerinde Dif_Skor testinin gücü, Donner GOF testine yaklaşmıştır. Prevalans değerleri büyüdüğünde n_j=100'de kappalar arasındaki fark en yüksek seviyeye ulaştığında tüm testler eşit güce sahiptir.

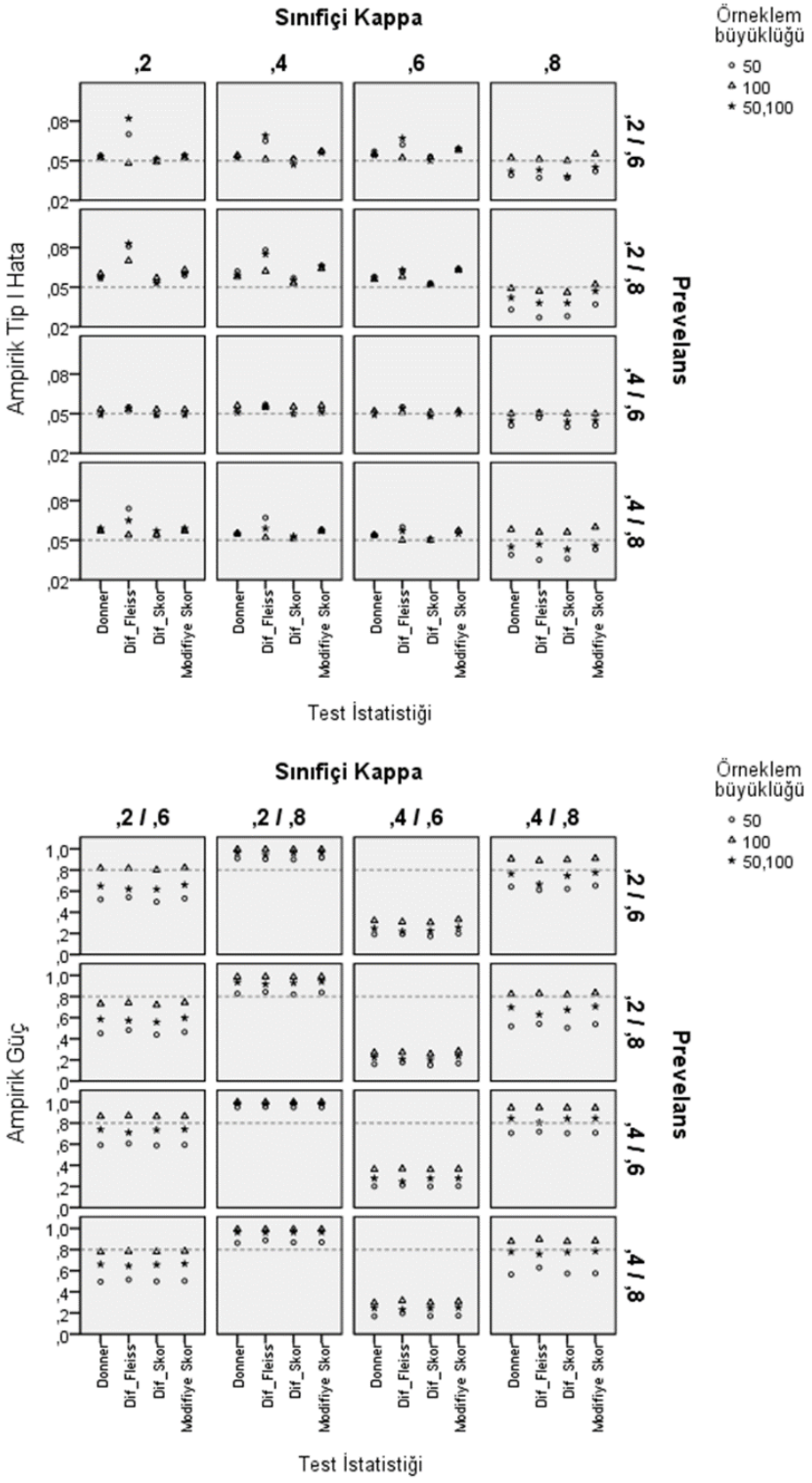
3.2. Üç bağımsız sınıfiçi kappa istatistiği için elde edilen bulgular

Eşit prevalans varsayımı altında 3 bağımsız sınıf içi kappa istatistiğinin homojenliğinin test edilmesinde Donner GOF ve Eq_Skor testlerinin, prevalansın uç değerleri dışında genel olarak nominal seviyeye yakın Tip I hata değerlerine sahip olduğu görülmüştür. Eq_Fleiss ve MES testi ise özellikle uç prevalans değerlerinde n_j=100'de bile liberal bulunmuştur. Pearson GOF testinin ise tüm prevalans ve tüm örneklem genişliği kombinasyonlarında Tip I hatası nominal seviyede olup diğer tüm testlerden üstün olduğu görülmüştür (Şekil 3). Genel olarak MES testi en yüksek Tip I hata değerleri sergilemiştir.

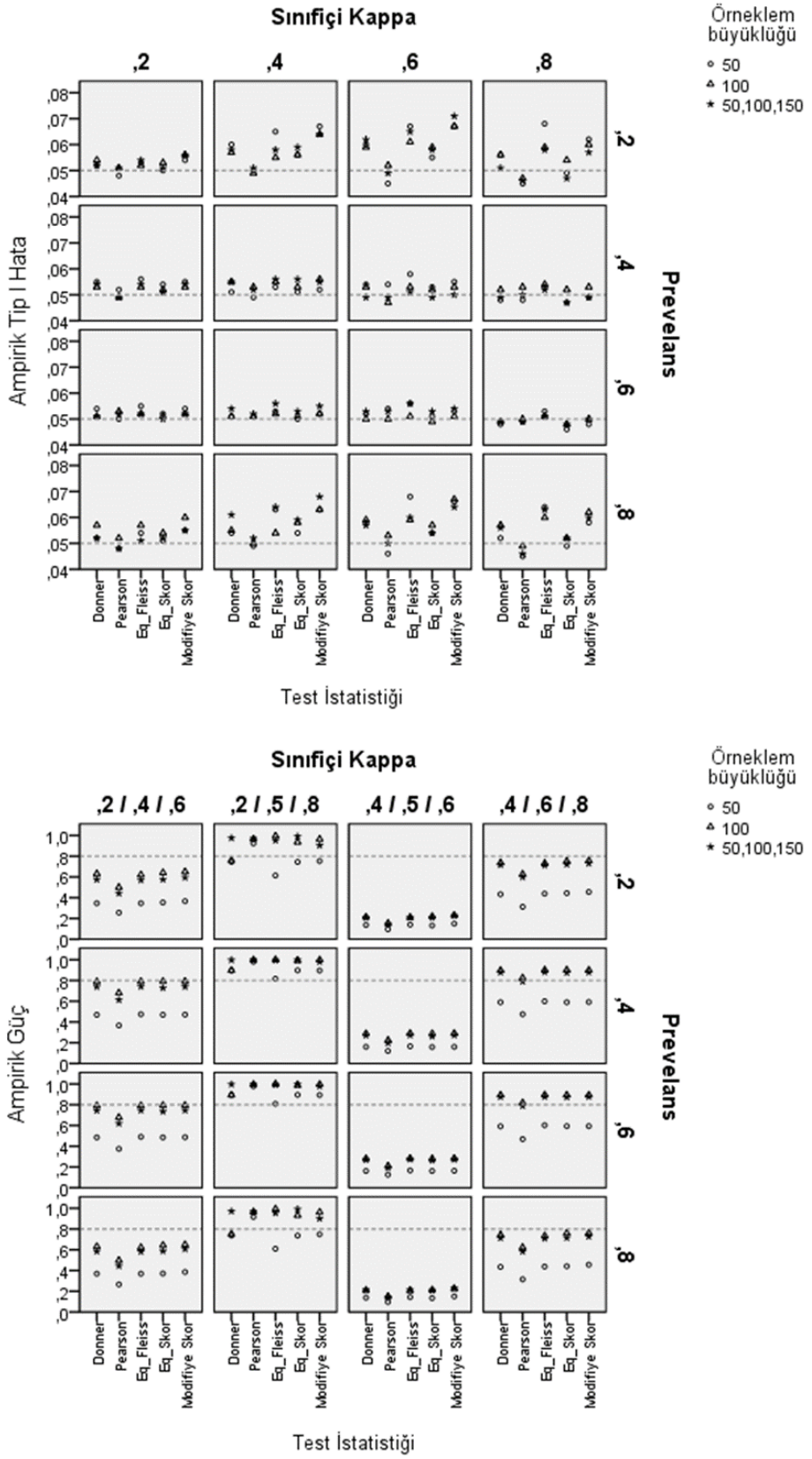
Tüm prevalans değerlerinde kappalar arasındaki fark 0,3 olduğunda n_j=50'de Pearson GOF testinin, n_j=100'de Eq_Fleiss testinin, n_1=50, n_2=100 ve n_3=150 koşulunda ise Eq_Skor testinin en güçlü test olduğu görülmüştür. p_j=0,2 ve p_j=0,8'de kappalar arasındaki fark 0,1 ve 0,2 olduğunda tüm örneklem genişliklerinde en güçlü test Modifiye Edilmiş Skor testidir. p_j=0,4 ve 0,6 arasında olduğunda ise tüm örneklem genişliklerinde en güçlü test Eq_Fleiss testidir, eşit örneklem genişliklerinde Donner GOF ve Eq_Skor testleri hemen hemen aynı güç değerlerine sahipken, eşit olmayan örneklem genişliğinde Donner GOF testi daha güçlü bulunmuştur. Pearson GOF testinin ise genel olarak gücü en düşük olan test olduğu görülmüştür.



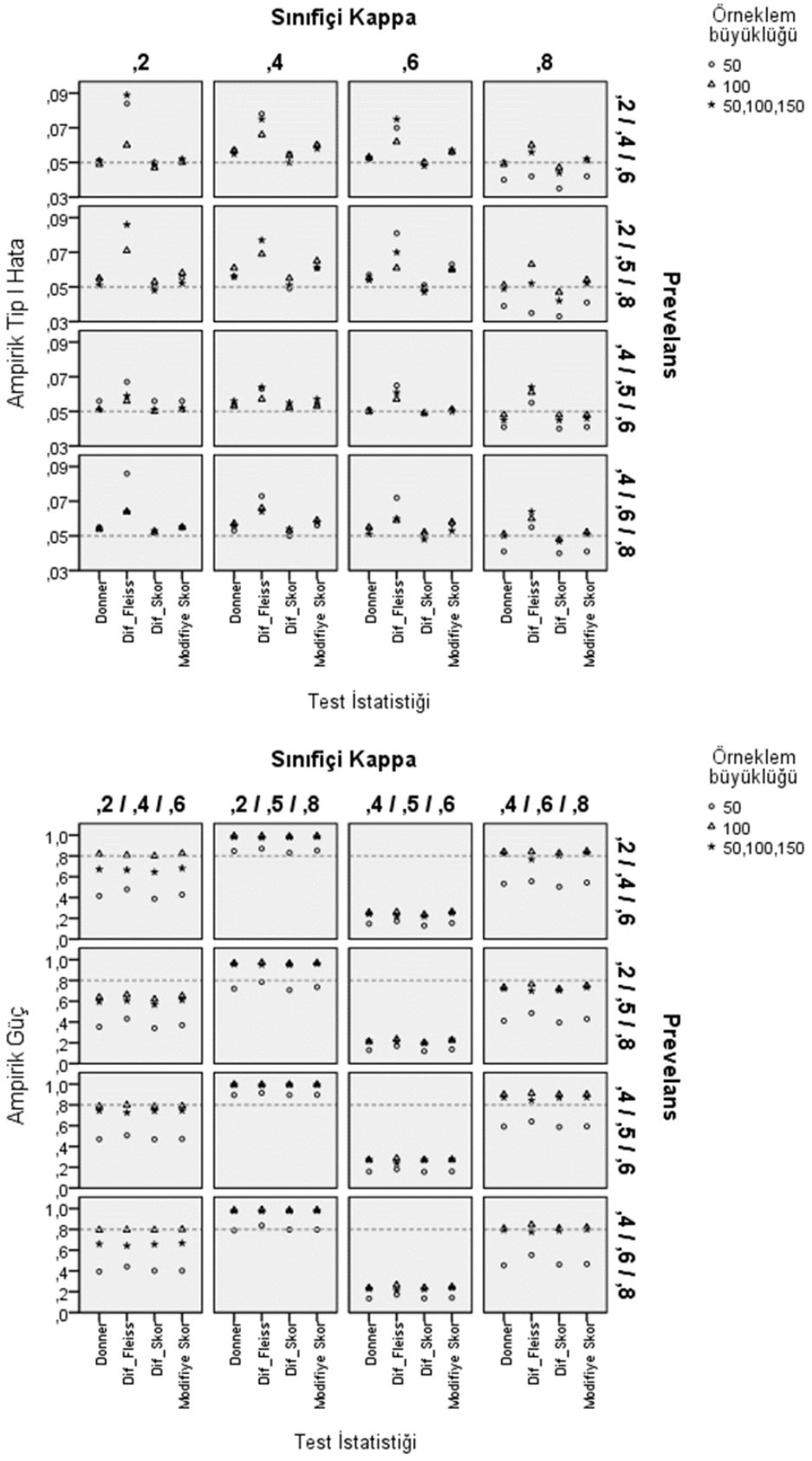
Şekil 1. Evren prevelanslarının eşit olduğu varsayımı altında 2 sınıf içi kappa istatistiği için homojenlik testlerinin ampirik Tip I hata ve güç düzeyleri



Şekil 2. Evren prevelanslarının eşit olmadığı varsayımı altında 2 sınıf içi kappa istatistiği için homojenlik testlerinin ampirik Tip I hata ve güç düzeyleri



Şekil 3. Evren prevelanslarının eşit olduğu varsayımı altında 3 sınıf içi kappa istatistiği için homojenlik testlerinin ampirik Tip I hata ve güç düzeyleri

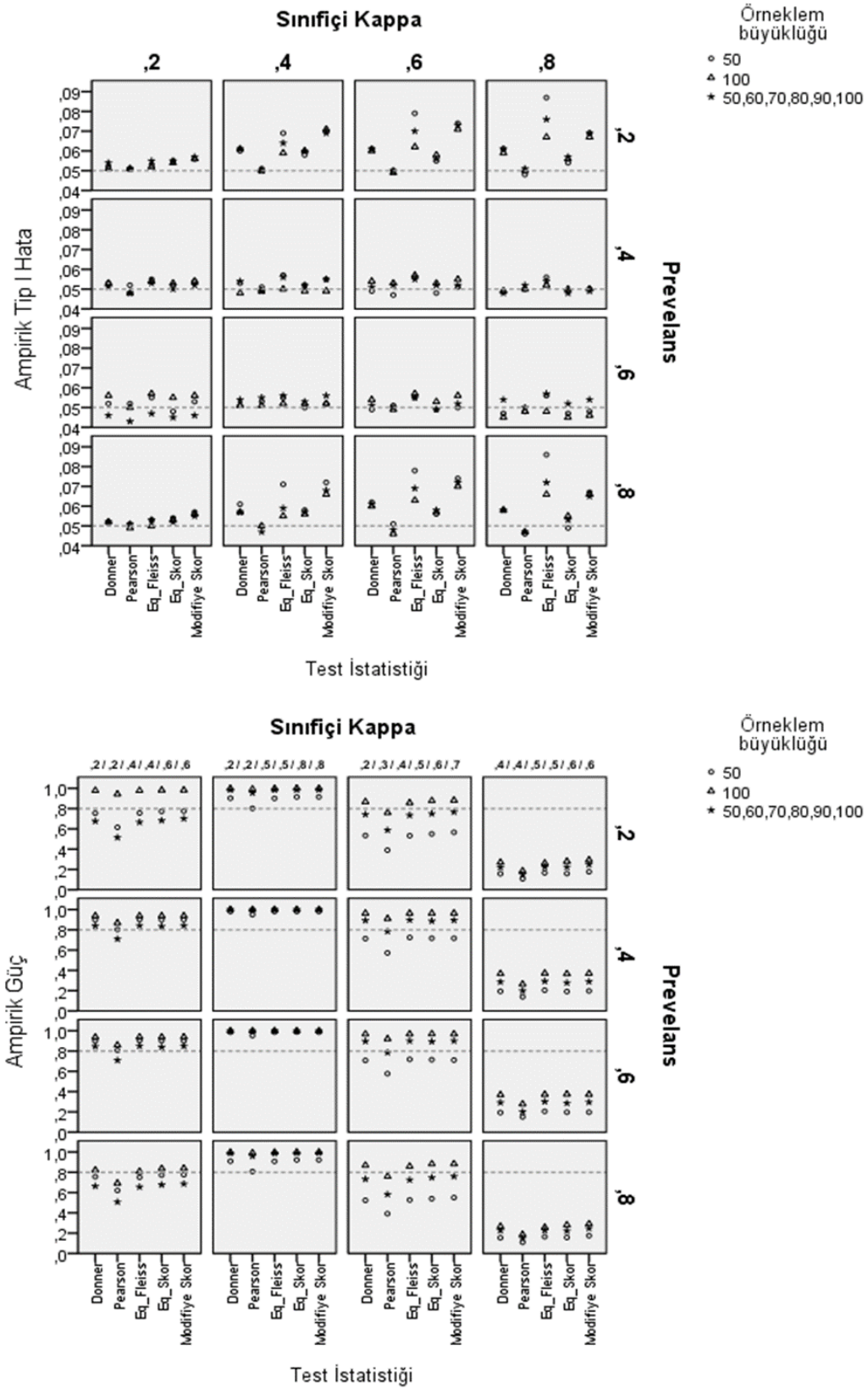


Şekil 4. Evren prevelanslarının eşit olmadığı varsayımı altında 3 sınıf içi kappa istatistiği için homojenlik testlerinin ampirik Tip I hata ve güç düzeyleri

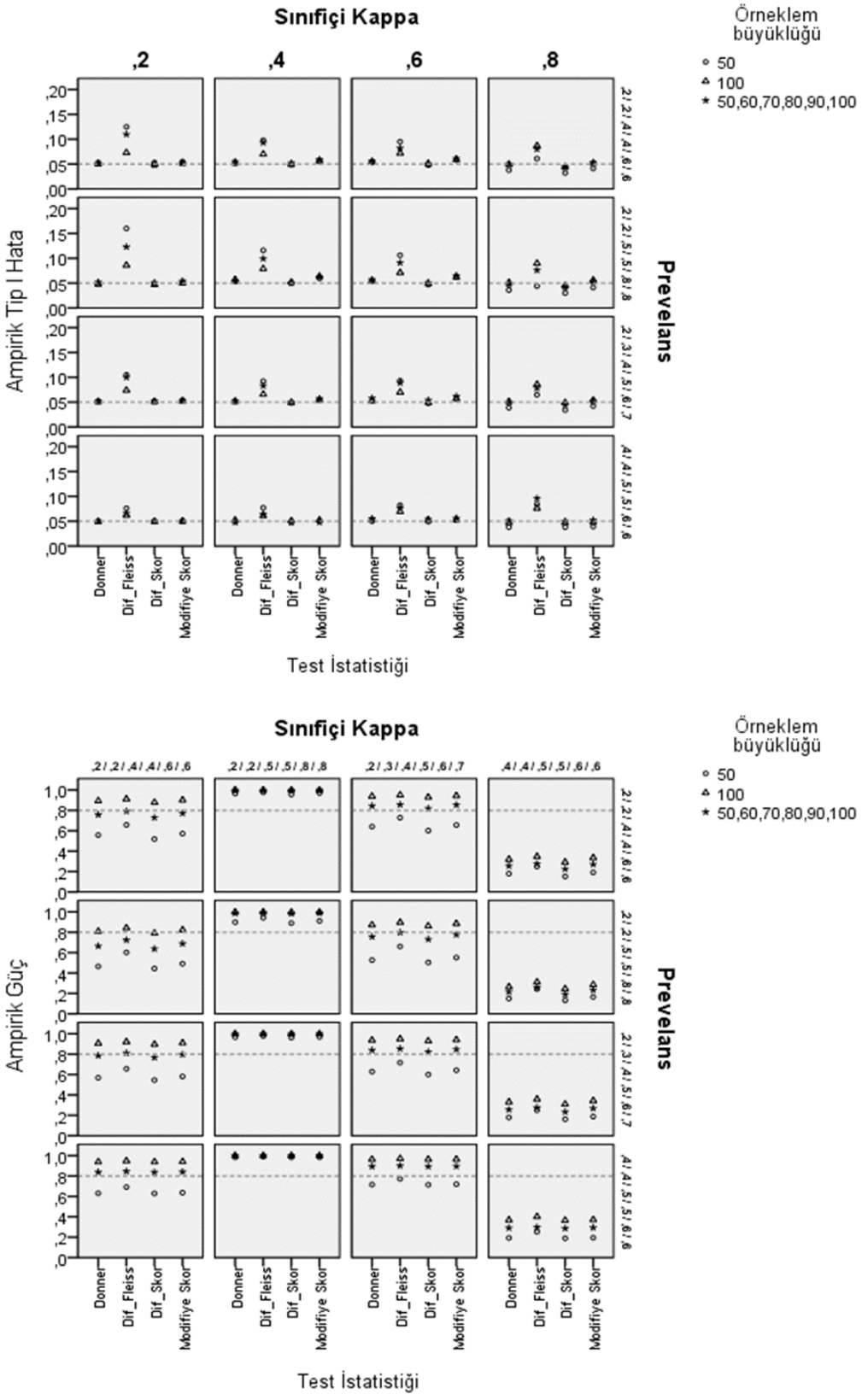
Prevelanslar eşit olmadığında; prevalansların eşit olmaması durumunda daha iyi performans gösteren Donner GOF testinin nominal seviyeye yakın Tip I hata değerlerine sahip olduğu görülmüştür. Dif_Fleiss testinin ise Tip I hatasının n_j=100'de bile nominal seviyeden yüksek olduğu ve en kötü performansa sahip test olduğu belirlenmiştir. Dif_Skor testinin Tip I hatasının nominal seviyede olduğu ve 3 sınıf içi kappa istatistiğinin homojenliğinin test edilmesinde tüm testler içinde en iyi performansı sergilediği görülmüştür. MES testi ise genel olarak Donner GOF testi ile benzer olmakla beraber prevalanslar arasındaki fark 0,3 olduğunda nominal seviyeden biraz yüksek Tip I hata değerlerine sahiptir.
Şekil 4 ile verilen güç değerleri incelendiğinde tüm prevalans ve kappa değerlerinde n_j=50 olduğunda Tip I hata açısından liberal olması nedeniyle güç değerleri şişkinleşen Dif_Fleiss testinin en yüksek güç değerlerine sahip olduğu görülmektedir. Diğer testler arasında ise güç açısından önce MES testi ve daha sonra Donner GOF testi gelmektedir. Tüm çalışmaların örneklem genişliği n_j=100 olduğunda ise düşük prevalans değerlerinde MES testi, yüksek prevalans değerlerinde ise Dif_Fleiss testi en güçlü testtir. Eşit olmayan örneklem genişliği düzeninde ise en güçlü test MES testi olmakla beraber prevalansların 0,5'e yakın olduğu düzenlerde Donner GOF testi de MES testi kadar güçlü bulunmuştur. Prevalansların eşit olmadığı varsayımı altında 3 sınıf içi kappa istatistiğinin homojenliğinin test edilmesinde Donner GOF testi Dif_Skor testinden daha güçlü olmakla beraber, dikkate alınan tüm düzenlerde Dif_Skor testi en güçsüz test olarak karşımıza çıkmaktadır.

3.3. Altı bağımsız sınıfiçi kappa istatistiği için elde edilen bulgular

Evren prevalanslarının eşit olduğu varsayımı altında 6 sınıf içi kappa istatistiği için homojenlik testlerinin Tip I hata düzeyleri incelendiğinde, Donner GOF testi, prevalansın uç değerlerinde kappa değerinin 0,3'den büyük olması durumunda tüm örneklem genişliği düzenlerinde nominal seviyeden biraz yüksek, diğer prevalans ve kappa değerlerinde ise Tip I hata açısından nominal seviyede bulunmuştur (Şekil 5). Prevelans 0,2 ve 0,8 olduğunda, Eq_Fleiss ve MES testleri, n_j=50 ve eşit olmayan örneklem genişliğinde, n_j=100 olduğunda ise kappanın 0,2'den büyük değerleri için liberal olma eğilimi göstermişlerdir. Eq_Skor testinin ise tüm örneklem genişliklerinde prevalansın uç değerlerinde nominal seviyeden biraz yüksek olmakla beraber, diğer tüm düzenlerde nominal seviyede Tip I hata değerlerine sahip olduğu gözlenmiştir. Pearson GOF testi ise homojenliği test edilen kappa istatistiği sayısı 2 veya 3 olduğu durumdakine benzer şekilde eşit prevalans varsayımı altında 6 sınıf içi kappanın homojenliğinin test edilmesinde de Tip I hata açısından nominal seviyede kalmış ve tüm testler arasında en iyi performansı göstermiştir (Şekil 5).
Prevalanslar eşit olduğunda testlerin güç değerleri incelendiğinde prevalansların 0,4'den küçük ve 0,6'dan büyük değerleri için n_j=50 olduğunda en güçlü testin Modifiye Edilmiş Skor Homojenlik testi olduğu görülmektedir. Prevalanslar 0,4–0,6 arasında olduğunda ise n_j=50'de liberal olması nedeniyle Eq_Fleiss en güçlü testtir. Yine n_j=50'de Donner GOF ve Eq_Skor testleri güç açısından benzerken, MES testinden düşük, en güçsüz test olan Pearson GOF testinden ise yüksek güç değerlerine sahip oldukları görülmektedir. Örneklem genişliği büyüdüğünde (n_j=100), 0,4–0,6 arasındaki prevalans değerleri için Pearson GOF testi dışında tüm testlerin güç değerleri benzerdir. Ancak, n_j=100 olduğunda tüm prevalans değerlerinde kappalar arasındaki fark 0,3 olduğunda tüm testler eşit güce sahiptir. Eşit olmayan örneklem genişliği ($n_j \neq n$) düzeninde ise testler arasında n_j=50'ye benzer ancak daha yüksek bir güç eğilimi gözlenmekle beraber n_j=100 koşulana göre daha düşük güç değerleri elde edilmiştir.
Prevalansların eşit olmaması durumunda 6 sınıf içi kappa istatistiğinin homojenliğinin test edilmesinde, Donner GOF ve Dif_Skor testleri Tip I hata açısından nominal seviyede bulunmuştur. Modifiye Edilmiş Skor Homojenlik testi prevalansın 0,5'den uzak olduğu koşullarda, kappanın 0,5'e yakın değerleri için prevalanslar arasındaki fark 0,3'e çıktığı zaman nominal seviyeden biraz yüksek Tip I hata değerlerine sahipken, diğer koşullarda ise nominal seviyededir. Dif_Fleiss testi ise n_j=100 olduğunda bile tüm prevalans düzenleri ve kappa değerleri için oldukça liberal bulunmuştur (Şekil 6).
Dif_Fleiss testinin prevalanslar eşit olmadığında homojenliği test edilen kappa istatistiğinin sayısının artmasıyla Tip I hata açısından oldukça liberalleşmesine bağlı olarak güç değerleri şişkinleşmiştir. Diğer testlerin performanslarına bakılacak olursa, tüm prevalans düzenlerinde ve tüm örneklem genişliklerinde Donner GOF testi, Dif_Skor testinden daha güçlüdür. Modifiye Edilmiş Skor testi ise tüm düzenlerde bu iki testten daha güçlü bulunmuştur.
Örneklem genişliği açısından güç değerleri değerlendirilecek olursa tüm testler için n_j=100>$n_j \neq n$>n_j=50 şeklinde bir sıralama yapılabilir. Testler farklı prevalans değerlerine göre incelendiğinde ise prevalanslar arasındaki fark sabit kalmak koşuluyla küçük prevalans değerlerinde büyük prevalans değerlerine göre daha güçlü oldukları, en güçlü oldukları prevalans koşulunun ise 0,5'e yakın değerler olduğu gözlemlenmiştir. Sınıf içi kappa değerlerine bağlı olarak güç değerlerinin değişimine bakıldığında ise kappalar arasındaki fark sabit olduğunda büyük kappa değerleri için gücün daha yüksek olduğu görülmektedir.



Şekil 5. Evren prevelanslarının eşit olduğu varsayımı altında 6 sınıf içi kappa istatistiği için homojenlik testlerinin ampirik Tip I hata ve güç düzeyleri



Şekil 6. Evren prevelanslarının eşit olmadığı varsayımı altında 6 sınıf içi kappa istatistiği için homojenlik testlerinin ampirik Tip I hata ve güç düzeyleri

4. TARTIŞMA

Prevalanslar eşit olmadığında 2 ve 3 bağımsız sınıf içi kappanın homojenliğinin test edilmesinde Nam (2003) küçük örneklem genişliklerinde Tip I hata açısından Dif_Fleiss testini oldukça liberal bulmuştur. Bu çalışmada ise $J=2$ olduğunda Dif_Fleiss testinin $n_j<100$ olduğunda liberal, prevalanslar arasındaki fark çok büyük olmadıkça $n_j=100$'de ise Tip I hata açısından nominal seviyede olduğu görülmüştür. Ancak, yine yaptığımız bu çalışmada homojenliği test edilen kappa katsayısının 2'den büyük olması durumunda özellikle $J=6$ olduğunda $n_j=100$'de bile çok liberal olduğu gözlenmiştir. Dif_Fleiss testi prevalanslar arasındaki fark sabitken homojenliği test edilen kappa katsayısının sayısının artmasıyla nominal seviyeden daha da uzaklaşmaktadır. Donner ve ark. (1996) tarafından yapılan çalışmada ise Dif_Fleiss testi homojenliği test edilen kappa istatistiği sayısı 2 ve 3 olduğunda büyük örneklemlerde incelenmiş, $p_j=0,1$ koşulu için $n_j=200$'de bile liberal bulunmuş ancak, $J=3$ için farklı prevalans düzenlerinde performansı incelenmemiştir. Güç açısından incelenecek olursa Donner ve ark.'nın (1996) çalışmasında $J=2$ olduğunda yine büyük örneklemlerde incelenmiş ve Donner GOF testinin az da olsa daha güçlü olduğu söylenmiştir. Çalışmamızda ise homojenliği test edilen sınıf içi kappa katsayısı 2 olduğunda Tip I hatasının nominal seviyede kaldığı $n_j=100$ koşulu dışında, $J=3$ ve $J=6$ olduğunda liberal bulunması nedeniyle Dif_Fleiss testinin güç değerlerinin şişkinleştiği görülmüştür.

Nam (2006) eşit prevalans varsayımı altında önerdiği Eq_Fleiss testini sadece $J=2$ olduğunda incelemiş ve küçük örneklemlerde liberal bulmuştur. Bu çalışmada ise 2, 3 ve 6 kappanın homojenliğinin test edilmesinde prevalansın uç değerlerinde (0,2 ve 0,8) $n_j=100$'de bile liberal bulunurken, diğer koşullarda Tip I hata açısından nominal seviyeye yakın bulunmuştur. Eq_Fleiss testi, Dif_Fleiss testi gibi homojenliği test edilen kappa istatistiğinin sayısının artmasıyla daha kötü performans sergilememiştir. Eq_Fleiss testi güç açısından değerlendirilecek olursa sadece $J=2$'de küçük örneklemlerde incelendiği Nam'ın (2006) çalışmasında yüksek Tip I hata değerlerine bağlı olarak şişkinleşmiş güç değerleri sergilemiştir. Çalışmamızda ise $J=2$'de ve ayrıca ilk defa performansının değerlendirildiği $J=3$ ve 6 koşullarında diğer testlerle yarışabilecek düzeyde ve prevalansın 0,5'e yakın değerleri için az da olsa daha avantajlı güç değerlerine sahiptir.

Nam (2006) tarafından eşit prevalans varsayımı altında kullanılması önerilen Pearson GOF testi küçük ve orta örneklem genişliklerinde sadece $J=2$ koşulunda incelenmiştir. Tip I hatası nominal seviyede bulunmuş ve güç açısından karşılaştırıldığı Eq_Fleiss, Donner GOF ve Eq_Skor testlerinden güçsüz bulunmuştur (Nam, 2006). Bu çalışmada da yine tüm örneklem genişliklerinde, tüm prevalans değerlerinde ve homojenliği test edilen kappa istatistiği sayısı 2, 3 ve 6 olduğunda Tip I hata açısından nominal seviyede olduğu görülmüştür. Eşit prevalans varsayımı altında her koşulda Tip I hata açısından tüm testler arasında en iyi performansı göstermesine rağmen, çalışmamızda da yukarıdaki testlere ek olarak Modifiye Edilmiş Skor Homojenlik testi de dahil olmak üzere tüm testlerden daha güçsüz bulunmuştur.

Nam (2003) prevalanslar eşit olmadığında $J=2$ ve 3 olması durumunda sadece Tip I hata açısından değerlendirdiği Donner GOF testinin küçük örneklemlerde nominal seviyeden yüksek olduğunu belirtmiştir. Yaptığımız bu çalışma ile Donner GOF testinin uç prevalans değerlerinde $n_j=100$'de bile nominal seviyeden yüksek olduğu diğer koşullarda ise nominal seviyeye daha yakın olduğu görülmüştür. Nam'ın (2006) eşit prevalans varsayımı altında yaptığı çalışma ile benzer olarak elde edilen sonuçlara göre Donner GOF testinin prevalanslar eşit olmadığında, eşit olduğu duruma göre daha iyi Tip I hata performansı gösterdiği belirlenmiştir. Donner ve ark. (1996) $J=2$ ve $p_j=p$ olduğunda $J=3$ için büyük örneklemlerde yaptıkları çalışmalarında, Donner GOF ile Dif_Fleiss testinin genel olarak benzer olduğunu ancak, az da olsa güç açısından Donner GOF testinin daha avantajlı olduğunu belirtmişlerdir. Bu çalışmada ise, $J=2$'de orta öneklem genişliklerinde güç değerleri şişkinleşen Dif_Fleiss testi Donner GOF testinden daha güçlü bulunmuştur. İlk defa güç değerlerinin incelendiği $J=3$ ve 6 olması durumunda ise Donner GOF testi $p_j=p$ olduğunda Eq_Skor testi ile benzer ya da daha güçlü, Pearson GOF testinden güçlü, Eq_Fleiss ve MES testlerinden daha güçsüz bulunmuştur. Prevalanslar eşit olmadığında ise Donner GOF testinin genel olarak Dif_Skor testinden güçlü, Dif_Fleiss ve MES testinden yine güçsüz olduğu görülmüştür.

Eşit prevalans varsayımı altında çalışan Olabilirlik Skor testi (Eq_Skor) Nam'ın (2006) yaptığı çalışmada sadece $J=2$ olması durumda incelenmiş ve Tip I hata açısından nominal seviyeye yakın bulunmuştur. Çalışmamızda ise Eq_Skor testi sadece $J=2$ durumu için değil aynı zamanda homojenliği test edilen bağımsız sınıf içi kappa istatistiği sayısı 3 ve 6 olduğunda da incelenmiştir. Bu çalışmada, Eq_Skor testinin prevalansın uç değerlerinde nominal seviyeden biraz yüksek olmakla beraber genel olarak nominal seviyede Tip I hata değerlerine sahip olduğu görülmüştür. Nam (2006) aynı zamanda Eq_Skor testini $J=2$ olması durumunda güç açısından incelediğinde küçük ve orta örneklem genişlikleri için karşılaştırdığı Pearson GOF testinden daha güçlü ancak güç değerleri şişkinlenmiş olan Eq_Fleiss testinden daha güçsüz bulduğunu belirtmiştir. Örneklem genişliği arttığında ise Eq_Skor testinin, Donner

GOF ve Eq_Fleiss testi ile benzer olduğunu ve Pearson GOF testinden güçlü olduklarını ifade etmiştir. Çalışmamızda güç açısından performansı ise J=2 için benzer bulunmuş, karşılaştırıldığı bir diğer test olan Modifiye Edilmiş Skor testinden genel olarak güçsüz bulunmuştur. Ayrıca, ilk defa performansının değerlendirildiği J=3 ve J=6 olması durumunda ise güç açısından genel olarak Donner GOF testi ile benzer, MES ve Eq_Fleiss testlerinden güçsüz ve Pearson GOF testinden güçlü olduğu görülmüştür.
Nam (2003) farklı prevalans değerlerinde J=2 ve J=3 olması koşulunda sadece Tip I hata açısından değerlendirdiği Dif_Skor testinin küçük ve orta örneklem genişliklerinde genel olarak nominal seviyede olduğunu belirtmiştir. Yapılan bu çalışmada da Dif_Skor testinin homojenliği test edilen sınıf içi kappa istatistiği sayısının 2, 3 ve 6 olması durumunda dikkate alınan tüm örneklem genişliği düzenlerinde Tip I hatası nominal seviyede bulunmuştur. Prevalanslar eşit olmadığında bu çalışma kapsamında karşılaştırıldığı Donner GOF, Dif_Fleiss ve MES testleri içinde Tip I hata açısından en iyi performansı gösteren test istatistiği olduğu belirlenmiştir. Güç açısından performansının ilk defa incelendiği bu çalışma ile homojenliği test edilen sınıf içi kappa sayısı 2, 3 ve 6 olduğunda genel olarak Donner GOF testi ile benzer ya da karşılaştırıldığı tüm testler içinde en güçsüz test olduğu görülmüştür.
Modifiye Edilmiş Skor testi, Nam (2003) tarafından J=2 ve 3 olması durumunda farklı prevalanslardaki performansının değerlendirildiği çalışmada Tip I hata açısından nominal seviyeden yüksek bulunmuştur. Bu çalışmada ise J=2, 3 ve 6 olması durumunda p_j=p olduğunda nominal seviyeden biraz yüksek, $p_j \neq p$ olduğunda ise prevalanslar arasındaki fark çok fazla olmadığı sürece ve eşit örneklem genişliklerinde nominal seviyeye yakın bulunmuştur. Yine ilk defa bu çalışmada değerlendirildiği güç açısından performansından söz edilecek olursa MES testinin, eşit prevalansta Tip I hatası biraz yüksek olduğu için genel olarak en güçlü test olduğu görülmüştür. Prevalansların eşit olmadığı durumda ise güç değerleri şişkinleşmiş olan Dif_Fleiss testinden daha düşük ancak, genel olarak karşılaştırıldığı diğer testlerden daha yüksek güç değerlerine sahip olduğu gözlenmiştir.

5. SONUÇ VE ÖNERİLER

Kappa istatistiklerinin homojenliğinin test edilmesi için geliştirilen Fleiss testi kappanın büyük örneklem varyansını kullanmaktadır. Bununla birlikte, kappa istatistiğinin örneklem genişliğine bağlı olarak normale yakınsaması oldukça yavaş (Nam, 2006) olduğu için küçük örneklemlerde ve n büyük olsa bile küçük prevalans değerlerinde kullanılmaması önerilmektedir. Ayrıca, Dif_Fleiss testinin 2'den fazla sınıf içi kappa istatistiğinin homojenliğinin test edilmesinde yüksek Tip I hata değerlerine sahip olması nedeniyle kullanılmamasının daha uygun olacağı düşünülmektedir. Eq_Fleiss testinin performansı ise homojenliği test edilen sınıf içi kappa istatistiği sayısından etkilenmemiştir.
Aynı ortak kappa kestiricisini kullanan Donner GOF ve Modifiye Edilmiş Skor Homojenlik testleri, Tip I hata açısından eşit prevalans varsayımı olmadığında, diğer bir deyişle farklı prevalans düzenlerinde daha iyi performans göstermişlerdir. Bu nedenle, her iki testin de prevalanslar farklı olduğunda kullanılmalarının daha uygun olacağı düşünülmektedir. MES testinin genel olarak nominal seviyenin biraz üzerinde olan Tip I hatası, güç açısından diğer testlere göre üstün konuma gelmesini sağlamıştır.
Eşit prevalans varsayımı altında çalışan Pearson uyum iyiliği testi, bu varsayım geçerli olduğunda homojenliği test edilen kappa istatistiğinin sayısından bağımsız olarak tüm prevalans değerlerinde ve örneklem büyüklüklerinde nominal seviyede kalarak Tip I hata açısından en iyi performansı gösteren test olmuştur. Ancak, güç açısından diğer testlere göre dezavantajlı olduğu görülmüştür.
En çok olabilirlik kestirimlerine dayanması nedeniyle Nam (2003) tarafından önerilen Dif_Skor testinin, incelendiği tüm düzenlerde Tip I hata açısından nominal seviyede kaldığı ve eşit prevalans varsayımına gerek duymayan testler içinde performansı en iyi olan test olduğu görülmüştür. Bununla birlikte, güç açısından incelendiğinde genel olarak Donner GOF testi ile benzerlik göstermekle beraber diğer testlere göre biraz daha güçsüz kalmıştır.
Evren prevalanslarının eşit olduğu varsayımı altında çalışan Eq_Skor testi, genel olarak Tip I hata açısından nominal seviyeye yakın bulunmuştur. Tip I hatasının genel olarak nominal seviyeden yüksek olması nedeniyle en güçlü test olan MES testinden güçsüz olmakla beraber, en güçsüz test olan Pearson GOF testinden güçlü olduğu görülmüştür. Donner GOF ve Eq_Fleiss testleri ile karşılaştırıldığında ise güç açısından benzer performans göstermiştir.
Sınıf içi kappa istatistiklerinin homojenliğinin test edilmesinde kullanılan testlerin performanslarının incelenmesi amacıyla yapılan bu benzetim çalışmasının sonuçlarına göre, örneklem genişliği, prevalans ve homojenliği test edilen sınıf içi kappa istatistiğinin sayısı gibi faktörlerin testlerin performansları üzerinde etkili oldukları anlaşılmıştır.

Kaynaklar

Alpar, R. (2010). Spor, Sağlık ve Eğitim Bilimlerinden Örneklerle Uygulamalı İstatistik ve Geçerlik-Güvenirlik (1. bs.). Ankara: Detay Yayıncılık.

Alpar, R. (2003). Uygulamalı Çok Değişkenli İstatistiksel Yöntemlere Giriş 1 (2. bs.). Ankara: Nobel Yayın Dağıtım.

Banerjee, M. (1999) Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics-Revue Canadienne De Statistique*, 27 (1), 3-23.

Bishop, Y.M.M.F., S.E.; Hollanda, P.W. (1989). Discrete Multivariate Analysis: Theory and Practice (Tenth Printing bs.). Cambridge, Massachusetts, and London, England: The MIT Pres.

Bloch, D.A.,Kraemer, H.C. (1989) 2x2 Kappa-Coefficients - Measures of Agreement or Association. *Biometrics*, 45 (1), 269-287.

Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.

Donner, A., Eliasziw, M.,Klar, N. (1996) Testing the homogeneity of kappa statistics. *Biometrics*, 52 (1), 176-183.

Fleiss, J.L. (1981). Statistical methods for rates and proportions. New York, John Wiley & Sons, Inc.

Kraemer, H.C., Periyakoil, V.S.,Noda, A. (2002) Kappa coefficients in medical research. *Statistics in Medicine*, 21 (14), 2109-2129.Nam, J.M. (2002) Testing the intraclass version of kappa coeffcient of agreement with binary scale and sample size determination. *Biometrical Journal*, 44 (5), 558-570.

Mak, T.K. (1988) Analyzing Intraclass Correlation for Dichotomous-Variables. *Applied Statistics-Journal of the Royal Statistical Society Series C*, 37 (3), 344-352.

Nam, J.M. (2002) Testing the intraclass version of kappa coeffcient of agreement with binary scale and sample size determination. *Biometrical Journal*, 44 (5), 558-570.

Nam, J.M. (2003) Homogeneity score test for the intraclass version of the kappa statistics and sample-size determination in multiple or stratified studies. *Biometrics*, 59 (4), 1027-1035.

Nam, J. (2006) Assessment on homogeneity tests for kappa statistics under equal prevalence across studies in reliability. *Statistics in Medicine*, 25 (9), 1521-1531.

Ridout, M.S., Demetrio, C.G.B.,Firth, D. (1999) Estimating intraclass correlation for binary data. *Biometrics*, 55 (1), 137-148.

Scott, W.A. (1955) Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19, 321-325.