Fleiss Kappa ve Krippendorff Alpha Uyum Katsayılarının Örneklem Genişliği, Değerlendirici Sayısı ve Kullanılan Ölçeğin Kategori Sayısından Etkilenme Durumları Üzerine Bir Benzetim Çalışması

Effect of Sample Size, The Number of Raters and the Category Levels of Diagnostic Test on Krippendorff Alpha and the Fleiss Kappa Statistics for Calculating Inter Rater Agreement: A Simulation Study

E. Arzu KANIK,[a]
Gülhan OREKİCİ TEMEL,[a]
İrem ERSÖZ KAYA[b]

[a]Biyoistatistik AD,
Mersin Üniversitesi Tıp Fakültesi,
[b]Bilgisayar Sistemleri AD,
Mersin Üniversitesi Teknik Eğitim Fakültesi,
Mersin

Geliş Tarihi/*Received:* 12.02.2010
Kabul Tarihi/*Accepted:* 15.04.2010

Yazışma Adresi/*Correspondence:*
E. Arzu KANIK
Mersin Üniversitesi Tıp Fakültesi,
Biyoistatistik AD, Mersin,
TÜRKİYE/TURKEY
arzukanik@gmail.com

ÖZET Amaç: Bu çalışmanın amacı, aynı materyal üzerinde hastanın durumu hakkında klinik karar veren birden fazla değerlendirici arasındaki uyum hesaplanırken kategorik veriler için kullanılan Krippendorff Alpha ve Fleiss Kappa istatistiklerinin, örnek genişliği, değerlendirici sayısı ve kullanılan ölçeğin kategori sayısından nasıl etkilendiklerini ortaya koymaktır. **Gereç ve Yöntemler:** Fleiss Kappa ve Kripendorff Alpha uyum katsayılarının değerlendiricilerin kararları arasında hiç uyum olmadığı (değerlendiriciler arasındaki karar rasgele) ve yüksek uyumun (değerlendiriciler arasındaki ortak karar 0.90 düzeyinde) olduğu durumlarda sonuçların örnek genişliklerinden, değerlendirici sayısından ve tanı testinin kategori sayısından nasıl etkilendiğini incelemek amacı ile bir Monte Carlo benzetim çalışması yapılmıştır. Benzetim çalışması Matlab 7 paket programı kullanılarak yapılmıştır. **Bulgular:** Örnek büyüklüğünün küçük, orta ve yüksek olması sonuçları değiştirmemekle birlikte değerlendiriciler arasında çok güçlü bir uyum varken değerlendirici sayısının en az 5 ve tanı testinin kategori sayısının 10'a çıkma durumunda Krippendorff Alpha katsayısının beklenen değerin (0.90) daha büyük tahminler yaptığı gözlenmiştir. **Sonuç:** Bu çalışmada örnek büyüklüğünün 30 olarak alınmasının Fleiss Kappa ve Kripendorff Alpha uyum katsayılarının her ikisi için de parametreyi doğru tahmin etmek için yeterli bir büyüklük olduğu saptanmıştır. Değerlendiriciler arası uyum hesaplanırken değerlendirici sayısı ve kategori sayısının 5 ten fazla olduğu durumlarda Krippendorff Alpha kullanırken tahminlerin gerçek değerinden yaklaşık 1,05 kat daha fazla olabildiği dikkate alınmalıdır.

Anahtar Kelimeler: Gözlemci değişkenliği; Sonuçların yeniden üretilebilirliği

ABSTRACT Objective: The aim of the study is to introduce how the Krippendorff Alpha and the Fleiss Kappa statistics that are designed for the categorical data used for calculating the measurement of agreement of more than one raters who make clinical decisions about the state of patients, were affected by the sample size, the number of raters and the category levels of diagnostic test **Material and Methods:** A Monte Carlo simulation study was performed to assess how the Fleiss Kappa and Kripendorff Alpha coefficients were affected by the sample sizes, number of the raters and the category number of the diagnostic test for two situations; no consistency and high consistency. Simulation study was done by using Matlab 7.0. **Results:** It was observed that sample size's being small, medium and large didn't change the results. While there was high agreement between readers, Krippendorff Alpha Coefficient estimate higher than expected value (0.90) in case of the minimum number of reader was 5 and the number of diagnostic test's category was 10. **Conclusion:** In this study, the sample size as 30 would be sufficient to make accurate estimation for the two methods. In calculating the measurement of inter-rater agreement, it must be considered that when using Krippendorff Alpha for the models including more than 5 raters and categories, the estimates are made about 1, 05 times much more than the actual values.

Key Words: Observer variation; reproducibility of results

Turkiye Klinikleri J Biostat 2010;2(2):74-81

Tıp biliminde bir hastaya ait materyal üzerinde hastalığın varlığı yokluğu ya da derecesi konusunda karar veren birden çok değerlendiricinin uyumu özellikle altın standart bir testin bulunmadığı durumlarda oldukça önemlidir. Latent Class yöntemiyle tanı testi etkinliğinin saptanmasında değerlendiricilerin uyumunun yüksek olması başarıyı arttırmaktadır.[1,2] Değerlendirici uyumunun ölçülmesinde kullanılan yöntemler, kullanılan tanı testinin sürekli ya da kategorik olmasına ve değerlendirici sayısına bağlı olarak değişir. Tanı testi sonuçları kategorik yapıda ise kullanılan uyum katsayılarından birisi de Cohen Kappa istatistiğidir.[3,4] Bu katsayısı, eşit sayıda kategorisi olan tanı testinin kullanıldığı durumlarda iki değerlendirici arasındaki uyumu ölçmek için kullanılır. Eğer değerlendirici sayısı ikiden fazla ise uygulamada genellikle değerlendiricileri ikili olarak çoklu Kappa testleriyle karşılaştırmak, çalışmaya ait I. tip hatanın artmasına neden olmaktadır. Birden çok değerlendiricinin kategorik test sonuçları arasındaki uyum ölçülmesinde yaygın kullanılan iki yöntem Fleiss Kappa[5-7] ve Krippendorff's Alpha[8,9] katsayılarıdır. Bu çalışmada Krippendorff Alpha ve Fleiss Kappa istatistiklerinin örnek genişliğinden, değerlendirici sayısından ve kullanılan ölçeğin kategori sayısından nasıl etkilendikleri araştırılmıştır.

GEREÇ VE YÖNTEMLER

FLEİSS KAPPA UYUMLULUK KATSAYISI

1971'de Fleiss, ikiden fazla değerlendirici arasındaki uyumu genellenmiş bir Kappa istatistiği ile ortaya koymuştur. Fleiss Kappa istatistiği ikiden fazla değerlendiricinin uyumunu kategorik ya da sıralı yapıda olan tanı testi sonuçlarını ölçmek amacıyla kullanılır.[10,11]

1'den k'ya kadar sonucu olan bir tanı testini, n değerlendirici N adet vaka için yorumlar ve tanı testinin her bir kategorisi, her bir vaka için ortak karar sonucu olarak başlangıç tablosu olarak düzenlenir (Tablo 1).

TABLO 1: Fleiss Kappa uyumluluk katsayısı için ortak karar sonuçlarının gösterilmesi.

	Tanı Testleri				
Vaka Sayısı	**1**	**2**	**....**	**k**	**P_i**
1	n_{11}	n_{12}		n_{1k}	P_1
2	n_{21}	n_{22}		n_{2k}	P_2
N	n_{N1}	n_{N2}		n_{Nk}	P_N
p_j	p_1	p_2		p_k	

Tablo (1)'de

N: toplam hasta sayısını

n: Değerlendirici sayısını

k: tanı testinin kategori sayısını

n_{ij}: i. hastanın tanı testi sonucuna j kararını veren değerlendirici sayısını

P_i: i. birey satır toplamı için toplam orantıyı

P_j: *j.* Kategori sütun toplamı için toplam orantıyı göstermektedir.

Fleiss Kappa istatistiği 1 nolu eşitlikte olduğu gibi hesaplanmaktadır.

$$K = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e} \quad (1)$$

i=1,2,….,N ve j=1,2…k olmak üzere K değeri Fleiss Kappa katsayısını göstermektedir.

Fleiss Kappa $0 \le k \le 1$ değer alır.

(1) Nolu Eşitlikte;

$\overline{P}$, P_i'lerin ortalamasını göstermek üzere $\overline{P}$ ve $\overline{P}_e$ ve aşağıdaki eşitliklerde verilmiştir:

$$\overline{P} = \frac{1}{N}\sum_{i=1}^{N} P_i = \frac{1}{Nn(n-1)}\left(\sum_{i=1}^{N}\sum_{j=1}^{k} n_{ij}^2 - Nn\right) \quad (2)$$

$$\overline{P}_e = \sum_{j=1}^{k} p_j^2 \quad (3)$$

$$1 = \frac{1}{n}\sum_{j=1}^{k} n_{ij} \quad (4)$$

$$P_i = \frac{1}{n(n-1)}\sum_{j=1}^{k} n_{ij}(n_{ij}-1) = \frac{1}{n(n-1)}\sum_{j=1}^{k}(n_{ij}^2 - n_{ij}) = \frac{1}{n(n-1)}\left[\left(\sum_{j=1}^{k} n_{ij}^2\right) - (n)\right] \quad (5)$$

$$P_j = \frac{1}{Nn}\sum_{i=1}^{N} n_{ij} \quad (6)$$

KRİPPENDORFF'S ALPHA GÜVENİRLİK KATSAYISI

Krippendorff Alpha katsayısı tüm ölçekler için kullanılabilen bir uyum katsayısıdır. Bu katsayının en önemli avantajı tamamlanmamış veri ya da eksik veri bulundurabilir.[12,13]

Bu katsayının genel formülü eşitlik (7)'deki gibidir.

$$\alpha = 1 - \frac{D_o}{D_e} \tag{7}$$

Eşitlik (7) de, D_o gözlenen uyumsuzluk, D_e ise beklenen uyumsuzluktur:

$$D_o = \frac{1}{n}\sum_c \sum_k O_{ck\ metric}\delta^2_{ck} \tag{8}$$

$$D_e = \frac{1}{n(n-1)}\sum_c \sum_k n_c n_{k\ metric}\delta^2_{ck} \tag{9}$$

Gözlenen uyumsuzluk $D_o = 0$ olduğunda, değerlendiricilerin mükemmel uyumlu olduğu sonucuna varılır. Bu durumda güvenirlik katsayısı $\alpha = 1$ çıkar. $D_o = D_e$ olur ise güvenirlik katsayısı $\alpha = 0$ olacaktır. Güvenirlik katsayısı $0 \le \alpha \le 1$ değer alır.

Krippendorff Alpha katsayısının hesaplanmasında 1. adım olarak *r* tane vakaya ait *m* tane değerlendiricinin sonuçlarından oluşan veri matrisi oluşturulur (Tablo 2).

Tablo (2) de;

r: Toplam vaka sayısını,

m: Toplam değerlendirici sayısını,

C_{iu}: *i.* değerlendiricinin *u*. vaka için değerlendirme sonucunu,

TABLO 2: Krippendorff Alpha katsayısı için veri matrisi.

	Vaka Sayısı					
Değerlendirici Sayısı	**1**	**2**	**.**	**u**	**.**	**r**
1	C_{11}	C_{12}		C_{1u}		C_{1r}
i	C_{i1}	C_{i2}		C_{iu}		C_{ir}
j	C_{j1}	C_{j2}		C_{ju}		C_{jr}
.						
m	C_{m1}	C_{m2}		C_{mu}		C_{mr}
Toplam	m_1	m_2		m_u		m_r

TABLO 3: Krippendorff Alpha katsayısı için uyum matrisi.

Sonuçlar	**1**	**.**	**k**	**.**	**.**	
1	O_{11}		O_{1k}			n_1
.	.	.	.			.
.	.	.	.			.
c	O_{c1}		O_{ck}	.	.	$n_c = \sum_k O_{ck}$
	n_1		n_k			$n \sum_c \sum_k O_{ck}$

m_u: u.vakayı değerlendiren değerlendiricilerin toplamını gösterir.

Eksik veri olmadığı durumda m_u değeri değerlendirici sayısına eşit olacaktır.

Krippendorff Alpha katsayısını hesaplamada 2. adım olarak 1. adımda oluşturulan veri matrisi, tüm eşleşen değerlendirme çiftlerinin frekanslarını içeren uyum matrisine dönüştürülür (Tablo 3).

Tablo (3)'de eşleşen değerlendirme çiftlerinin frekansları eşitlik (10)'da ki gibi hesaplanmaktadır:

$$O_{ck} = \sum_u \frac{u.\text{ vakadaki } c\text{-}k \text{ çiftlerinin sayısı}}{m_u - 1} \tag{10}$$

Eşitlik (10)'da O_{ck}, *c-k* değerlendirme çiftinin u. vakadaki gözlenme frekansını göstermektedir. Ve sonuçta eşitlik 11 yardımı ile Krippendorff Alpha katsayısı hesaplanır.

$$no\min al\alpha = 1 - \frac{D_o}{D_e} = \frac{(n-1)\sum_c O_{cc} - \sum_c n_c(n_c-1)}{n(n-1) - \sum_c n_c(n_c-1)} \tag{11}$$

BENZETİM DENEMELERİ

Bu çalışmada Fleiss Kappa ve Kripendorff Alpha katsayılarıyla hesaplanan uyumların değerlendiricilerin kararları arasında hiç uyum yokken (rasgele) ve uyumun 0.90 olduğu iki farklı durum için, örnek genişliklerinden, değerlendirici sayısından ve tanı testinin sonuçlarından nasıl etkilendiğini incelemek amacı ile bir Monte Carlo benzetim çalışması yapılmıştır. Benzetim denemeleri; veri üretimi ve katsayıların hesaplanması MatLab 7 paket programında yapılmış ve benzetim çalışmasının kodları Ek 1'de verilmiştir. Ayrıca yöntemlerin tanımlayıcı

EK 1

Fleiss Kappa code:

```
%----Fleiss Uygulamasi Basliyor----%
 function fleissK(data_in,conf)

    [m,N] = size(data_in);
    % m = kodlayici sayisi
    % N = ornek sayisi

    caseTypes = unique(data_in);
    fleiss_mat = [ ];
    sum_cases = 0;

    for i=1:N
       for j=1:m
          sum_cases = sum_cases + ismember(caseTypes, data_in(j,i))';
       end
       fleiss_mat(i,:) = sum_cases;
       sum_cases = 0;
    end

    n=size(fleiss_mat,1);
    m=sum(fleiss_mat(1,:)); %raters
    a=n*m;
    pj=(sum(fleiss_mat)./(a)); %overall proportion of ratings in category j
    b=pj.*(1-pj);
    c=a*(m-1);
    d=sum(b);
    kj=1-(sum((fleiss_mat.*(m-fleiss_mat)))./(c.*b)); %the value of Kappa for the j-th
category
    FleissAlpha=sum(b.*kj)/d %Fleiss'es (overall) Kappa
    sek=realsqrt(2*(d^2-sum(b.*(1-2.*pj)))/sum(b.*realsqrt(c));
%Kappa standard error
    ci=FleissAlpha+([-1 1].*(abs(0.5*erfc(-conf/2/realsqrt(2)))*sek)); %FleissAlpha
confidence interval
    z=FleissAlpha/sek; %normalized Kappa
    p=(1-0.5*erfc(-abs(z)/realsqrt(2)))*2;

    fid = fopen('sonuc.txt','a');
    fprintf(fid,' \t  %0.4f \t (%d%%) = %0.4f   %0.4f \t z = %0.4f   p = %0.4f',FleissAl-
pha,(1-conf)*100,ci,z,p);
    fclose(fid);

%----Krippendorf Uygulamasi Basliyor----%
 function krippendorff(data_in)

 [m,N] = size(data_in);
 % m = kodlayici sayisi
 % N = ornek sayisi

 for i=1:N
    u = data_in(:,i);
    % u.ornek aliniyor.
    miss = find(isnan(u)==1);
    % eksik olan hucrelerin indisleri bulunuyor.
    u(miss) = [ ];
    % eksik hucreler siliniyor.
    mu = length(u);
    % u.ornekte kodlayici sayısı bulunuyor.
    if mu==1
       allComb = [ ];   %Tek kodlayıcı oldugunda kombinasyon alinmiyor.
    elseif mu==0
       allComb = [ ];
    else
       allComb = combntns(u,2);
    end
    % u.ornekte tum kombinasyonlar bulunuyor.
    hucre{i,1} = allComb;
    hucre{i,2} = mu;
    % u.ornek icin bulunan kombinasyon ve
    % kodlayici sayisi hucrede saklaniyor.
 end

case_types = unique(data_in);
 x  = find(isnan(case_types)==1);
 case_types(x) = [ ];
 [v,bir] = size(case_types);
 % v = durum sayisi

 diag_comb = [case_types case_types];
 triu_comb = combntns(case_types,2);
 join_comb = combine(diag_comb, triu_comb);
 all_comb = sortrows(join_comb);
 % arastirilacak tum kombinasyonlar biraraya getirilip siralaniyor.

 kripp_mat = [ ];
 count = 0;

 for c=1:v
    for k=c:v
       count = count + 1;
       lookC = int2str(all_comb(count,:));
       % tum kombinasyonlar sirayla alinip string yapiliyor.
       Ock = 0;
       for t=1:N
          forC = int2str(hucre{t});
          % t.hucre alinip string yapiliyor.
          findC = strmatch(lookC, forC);
          % mevcut kombinasyon t.hucrede araniyor.
          rev_forC = seqreverse(forC);
          % t.hucrenin tersi aliniyor.
          find_revC = strmatch(lookC, rev_forC);
          % mevcut kombinasyon t.hucrenin tersinde araniyor.
          sizeFinds = length(combine(findC, find_revC));
          % tum bulunan ihtimaller sayiliyor.
          if sizeFinds == 0
             Ock_hucre = 0;
          else
             Ock_hucre = sizeFinds /(hucre{t,2}-1);
          end
          Ock = Ock + Ock_hucre;
          % Ock degerleri hesaplaniyor.
       end

       kripp_mat(c,k) = Ock;
       if c~=k
          kripp_mat(k,c) = Ock;
       end
       % hesaplanan Ock degerleri matrise yaziliyor.
    end
 end

 nk = sum(kripp_mat);
 nc = sum(kripp_mat);
 n = sum(nk);

 Do = 0;
 De = 0;
 for c=1:v
    for k=c:v
       Do = Do + (kripp_mat(c,k)*metric(c,k));
       De = De + (nc(c)*nk(k)*metric(c,k));
    end
 end

 KrippAlpha = 1 - (Do/(De/(n-1)))

 fid = fopen('sonuc.txt','a');
 fprintf(fid,' \t  %0.4f',KrippAlpha);
 fclose(fid);

 function xxx = metric(c,k)
 if c==k
    xxx=0;
 else
    xxx=1;
 end
 %----Krippendorf Uygulamasi Bitti---
```

istatistikleri ve yöntemler arasındaki uyum için çizilen Youden Plot grafikleri ise Medcalc®v11.2 paket programı ile elde edilmiştir.

Benzetim denemelerinde, değerlendirici sayısının 2, 5 ve 7 olduğu 3 durum ile tanı testinin 2, 5, 7 ve 10 kategorili olduğu 4 durum, örnek genişliğinin 30, 100 ve 1000 olduğu 3 durumu ile toplam 36 farklı kombinasyon kullanılmıştır. Bu kombinasyonlar değerlendiriciler arasında hiç uyumun olmadığı durum ile uyumun 0.90 olduğu durumlar için Krippendorff Alpha ve Fleiss Kappa'nın değerleri 1000 benzetim denemesi için kaydedilmiş ve her kombinasyon için ortalama ve standart sapmaları hesaplanmıştır. Her iki katsayı için hesaplanan ortalamalar benzetim denemelerinde tekrar sayısının 1000 olması nedeniyle popülasyon değeri kabul edilmiş ve hipotez testi ile karşılaştırma yapılmamıştır.

BULGULAR

Değerlendirici sayısı 2, 5 ve 7 (D=2-5-7) olmak üzere, tanı testi sonuçlarının 2'li, 5'li, 7'li ve 10'lu (K=2-5-7-10) ölçülme düzeyleri için beklenen uyumun sıfır ve beklenen uyumun 0.90 olması gereken durumlarda Krippendorff Alpha ve Fleiss Kappa'nın tanımlayıcı istatistikleri örnek genişlikleri 30, 100 ve 1000 için verilmiştir (Tablo 4, 5, 6).

Örnek büyüklüğü 30 olduğu durumda değerlendirici sayısı, karar sayısı ve değerlendiriciler arasındaki uyumdan Krippendorff Alpha katsayısı ile Fleiss Kappa katsayısı benzer sonuçlar vermiştir. Fakat değerlendiriciler arasındaki uyum 0.90

TABLO 4: Örnek büyüklüğü 30 ve beklenen uyumun 0.00 ve 0.90 olduğu durumlarda iki yöntemin tanımlayıcı istatistik tablosu.

	Beklenen Uyum 0,90		Beklenen Uyum 0,00	
	Krippendorrf Alpha Ortalama	Fleiss Kappa Ortalama	Krippendorrf Alpha Ortalama	Fleiss Kappa Ortalama
N=30	S.Sapma	S.Sapma	S.Sapma	S.Sapma
D=2-K=2	0,8974	0,8957	0,003525	-0,01336
	± 0,05925	± 0,06024	± 0,1784	± 0,1815
D=2-K=5	0,9004	0,8987	0,001790	-0,01513
	± 0,03053	± 0,03105	± 0,09113	± 0,09268
D=2-K=7	0,8995	0,8978	0,002820	-0,01408
	± 0,02477	± 0,02519	± 0,07327	± 0,07451
D=2-K=10	0,9027	0,8968	0,009757	-0,01543
	± 0,02742	± 0,02085	± 0,06407	± 0,06074
D=5-K=2	0,8965	0,8958	-0,001492	-0,008210
	± 0,01967	± 0,01980	± 0,05644	± 0,05682
D=5-K=5	0,8973	0,8966	-0,00001530	-0,006727
	± 0,01013	± 0,01019	± 0,02905	± 0,02925
D=5-K=7	0,8973	0,8966	-0,001741	-0,008464
	± 0,007715	± 0,007766	± 0,02387	± 0,02403
D=5-K=10	0,9317	0,8971	0,01178	-0,006612
	± 0,02788	± 0,006888	± 0,02536	± 0,01919
D=7-K=2	0,8974	0,8969	-0,001808	-0,006605
	± 0,01314	± 0,01320	± 0,03795	± 0,03814
D=7-K=5	0,8971	0,8966	-0,0003823	-0,005167
	± 0,006680	± 0,006713	± 0,01982	± 0,01992
D=7-K=7	0,8975	0,8970	0,0002912	-0,004489
	± 0,005370	± 0,005396	± 0,01550	± 0,01557
D=7-K=10	0,9451	0,8967	0,01367	-0,005009
	± 0,03050	± 0,004511	± 0,02093	± 0,01309

TABLO 5: Örnek büyüklüğü 100 ve beklenen uyumun 0.00 ve 0.90 olduğu durumlarda iki yöntemin tanımlayıcı istatistik tablosu.

	Beklenen Uyum 0,90		Beklenen Uyum 0,00	
	Krippendorrf Alpha Ortalama	Fleiss Kappa Ortalama	Krippendorrf Alpha Ortalama	Fleiss Kappa Ortalama
N=100	S.Sapma	S.Sapma	S.Sapma	S.Sapma
D=2-K=2	0,8979	0,8974	-0,002811	-0,007852
	± 0,03153	± 0,03169	± 0,1032	± 0,1037
D=2-K=5	0,8990	0,8985	0,002711	-0,002303
	± 0,01540	± 0,01547	± 0,05087	± 0,05112
D=2-K=7	0,8991	0,8986	0,001118	-0,003902
	± 0,01224	± 0,01230	± 0,04200	± 0,04221
D=2-K=10	0,9042	0,8989	0,008143	-0,0008549
	± 0,01444	± 0,01059	± 0,01122	± 0,01052
D=5-K=2	0,8990	0,8988	-0,0008063	-0,002814
	± 0,01020	± 0,01021	± 0,03099	± 0,03105
D=5-K=5	0,8994	0,8992	0,0008289	-0,001173
	± 0,005124	± 0,005133	± 0,01607	± 0,01610
D=5-K=7	0,8975	0,8968	-0,0001023	-0,002108
	± 0,007390	± 0,007440	± 0,01310	± 0,01313
D=5-K=10	0,9328	0,8992	0,01178	-0,006612
	± 0,01585	± 0,003442	± 0,02536	± 0,01919
D=7-K=2	0,8991	0,8989	-0,0002343	-0,001663
	± 0,006876	± 0,006886	± 0,02192	± 0,02195
D=7-K=5	0,8992	0,8991	0,0002131	-0,001217
	± 0,003605	± 0,003610	± 0,01124	± 0,01126
D=7-K=7	0,8994	0,8992	-0,00005720	-0,001488
	± 0,002989	± 0,002996	± 0,009090	± 0,009105
D=7-K=10	0,9455	0,8999	0,01411	-0,001156
	± 0,005126	± 0,0007450	± 0,01183	± 0,007254

TABLO 6: Örnek büyüklüğü 1000 ve beklenen uyumun 0.00 ve 0.90 olduğu durumlarda iki yöntemin tanımlayıcı istatistik tablosu.

	Beklenen Uyum 0,90		Beklenen Uyum 0,00	
	Krippendorrf Alpha Ortalama S.Sapma	Fleiss Kappa Ortalama S.Sapma	Krippendorrf Alpha Ortalama S.Sapma	Fleiss Kappa Ortalama S.Sapma
N=1000				
D=2-K=2	0,8994 ± 0,01025	0,8993 ± 0,01026	-0,001485 ± 0,03102	-0,001984 ± 0,03104
D=2-K=5	0,8997 ± 0,005013	0,8997 ± 0,005016	0,0008283 ± 0,01585	0,0003283 ± 0,01586
D=2-K=7	0,8991 ± 0,01224	0,8986 ± 0,01230	0,0002834 ± 0,01270	-0,0002169 ± 0,01271
D=2-K=10	0,9042 ± 0,01444	0,8989 ± 0,01059	0,01000 ± 0,03629	-0,003341 ± 0,03396
D=5-K=2	0,9000 ± 0,001241	0,8999 ± 0,001241	-0,0003415 ± 0,009957	-0,0005414 ± 0,009958
D=5-K=5	0,8973 ± 0,01013	0,8966 ± 0,01019	0,0001190 ± 0,005035	-0,00008140 ± 0,005036
D=5-K=7	0,8999 ± 0,0009172	0,8999 ± 0,0009154	-0,00006020 ± 0,004103	-0,0002602 ± 0,004105
D=5-K=10	0,9329 ± 0,004888	0,9000 ± 0,001059	0,01148 ± 0,004472	-0,00006780 ± 0,003238
D=7-K=2	0,9000 ± 0,002214	0,9000 ± 0,002212	-0,0005548 ± 0,006781	-0,0006989 ± 0,006782
D=7-K=5	0,9000 ± 0,001115	0,8999 ± 0,001116	-0,00001440 ± 0,003424	-0,0001578 ± 0,003425
D=7-K=7	0,8999 ± 0,0009172	0,8999 ± 0,0009154	-0,00009020 ± 0,002807	-0,0002323 ± 0,002811
D=7-K=10	0,9455 ± 0,005126	0,8999 ± 0,0007450	0,01344 ± 0,003627	-0,00002250 ± 0,002273

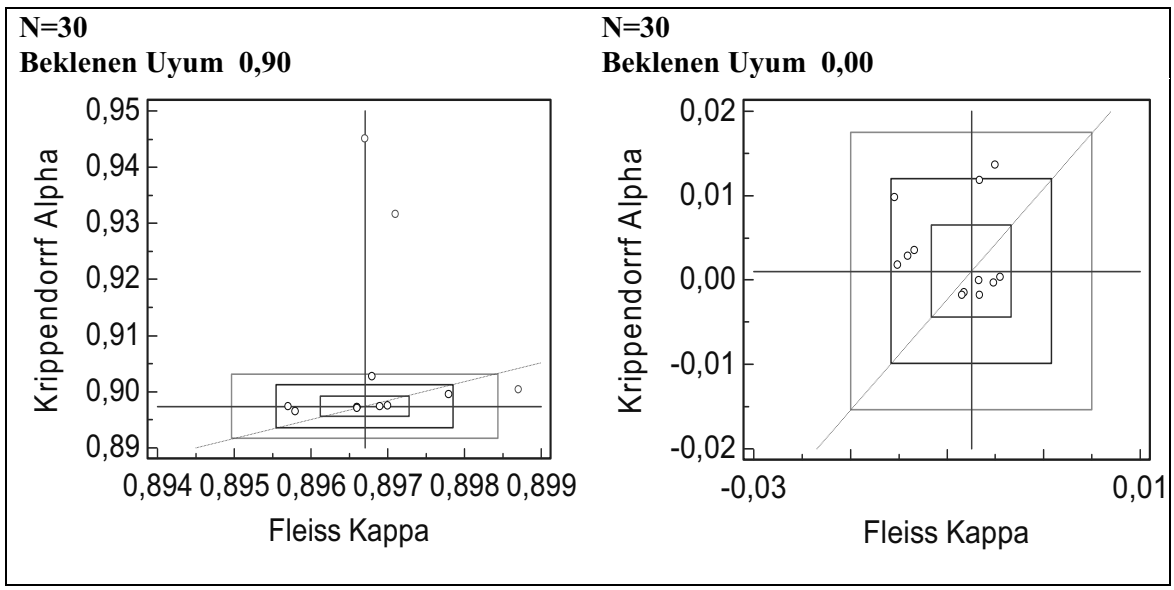
çıkması beklenen durumda, değerlendirici artışından ve tanı testinin kategorisi sayısının 10'a çıkma durumunda Krippendorff Alpha katsayısının etkilendiği ve Fleiss Kappa'ya göre arttığı gözlenmektedir.

Örnek büyüklüğünün 100 olduğu durumda Krippendorff Alpha katsayısı ile Fleiss Kappa katsayısı benzer sonuçlar vermiştir. Fakat değerlendiriciler arasındaki uyumun 0.90 olması beklenen durumda, değerlendirici sayısının ve tanı testi kategorisinin birlikte artmasının Krippendorff Alpha katsayısını etkilediği ve Fleiss Kappa'yı ise etkilemediği gözlenmektedir.

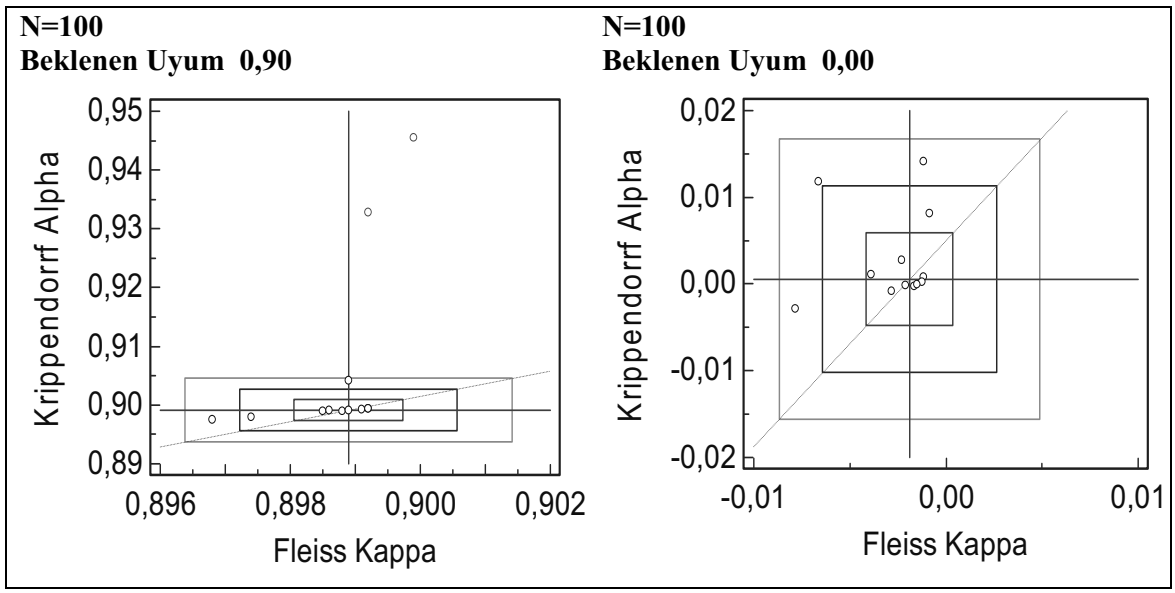
Örnek büyüklüğü 1000 olduğu durumda da sonuçların 30 ve 100 olduğu duruma oldukça yakın bulunduğu gözlenmiştir. Ayrıca örnek büyüklüklerine göre her iki yöntemin uyumluluklarının görsel gösterimi beklenen uyumların 0.00 ve 0.90 olduğu durumlarda 1SD, 2SD ve 3SD alanları için Youden Plot grafiği ile sunulmuştur (Şekil 1, 2, 3). Grafikler incelendiğinde Krippendorff Alpha değerinde D=5, K=10 ve D=7, K=10 durumlarının 3SD sınırları dışında kaldığı görülmektedir.

TARTIŞMA

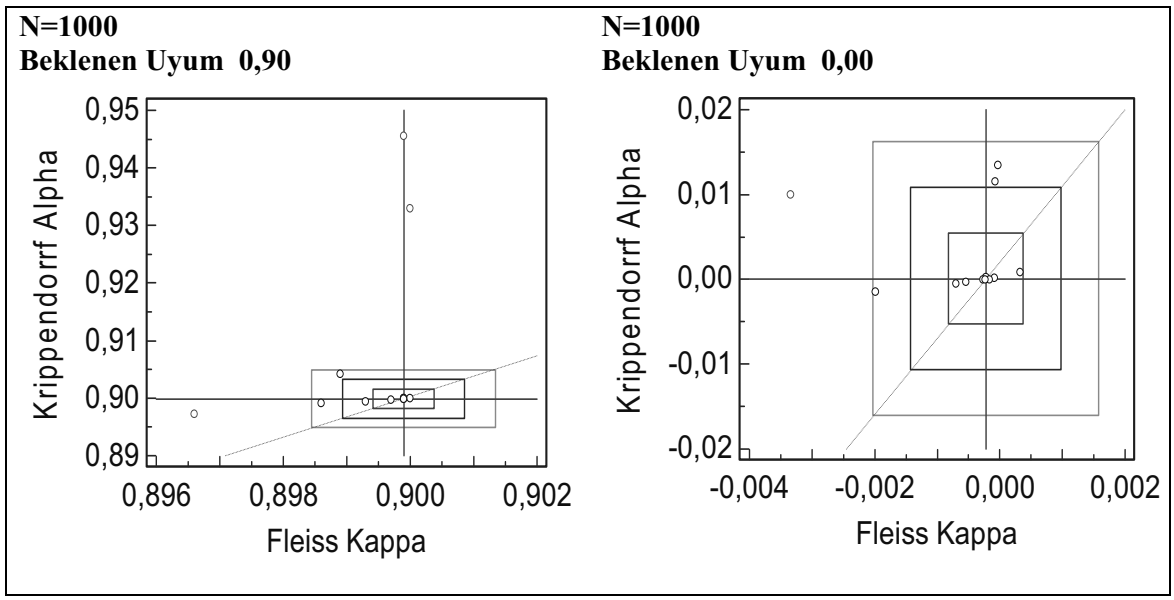
Değerlendiriciler arasındaki uyum hesaplanırken değerlendirici sayısı, tanı testinin alt kategorileri



ŞEKİL 1: Örnek büyüklüğü 30 olduğu durumda oluşturulan Youden Plot grafiksel gösterimi.



ŞEKİL 2: Örnek büyüklüğü 100 olduğu durumda oluşturulan Youden Plot grafiksel gösterimi.



ŞEKİL 3: Örnek büyüklüğü 1000 olduğu durumda oluşturulan Youden Plot grafiksel gösterimi.

ve örnek büyüklüğünün etkisi sık tartışılan bir konudur. Bu çalışmada her iki yönteminin de örnek büyüklüğünden etkilenmediği saptanmıştır.

Örnek büyüklüğünün 30'dan 100'e ve hatta 1000 çıkması durumunda bile değerlendiriciler arasındaki uyumun tahmininde değişiklik olmamıştır. Fakat değerlendiriciler arasındaki uyumun 0.90 çıkması beklenen durumda, değerlendirici sayısının en az 5 ve tanı testinin kategori sayısının 10'a çıkma durumunda Krippendorff Alpha katsayısının beklenen değerin (0.90) daha büyük tahminler yaptığı gözlenmiştir. Bu dezavantajının yanında literatürde Krippendorff Alpha katsayısı eksik verilerde de kullanılabildiği için değerlendiriciler arasındaki uyum hesaplamasında önemli bir yere sahip olduğu bildirilmektedir.[12,13]

SONUÇ

Değerlendiriciler arasındaki uyum hesaplanırken kullanılan Krippendorff Alpha ve Fleiss kappa istatistikleri üzerinde durulan özelliğin örnek büyüklüğünden etkilenmemekle birlikte Krippendorff Alpha katsayısı değerlendirici sayısı ve tanı testinin kategori sayısından etkilenmekte ve beklenenden daha yüksek tahminlerde bulmaktadır. Bunun yanında Krippendorff Alpha katsayısı eksik verilerde de kullanılabiliyor olması özelliği bu çalışmada dikkate alınmamıştır. Bu konu bir sonraki çalışmada ele alınacaktır.

KAYNAKLAR

1. Conaway RC. Latent class analysis. In: Armitage P, Colton T. Encyclopedia of Biostatistics. 2nd ed. NewYork; John Wiley & Sons: 2005. p.2730-3.
2. Uebersax JS, Grove WM. Latent class analysis of diagnostic agreement. Stat Med 1990;9(5):559-72.
3. Cohen J. A coefficient of agreement for nominal scales. Educational and psychological measurement. 1990;20(1):37-46.
4. Shoukri M. Measures of 2×2 association and agreement of cross-classified data. Samar Haddad. Measures of Interobserver Agreement. 1st ed. New York: Chapman & Hall/Crc; 2004. p.25-7.
5. Peyré SE, Peyré CG, Hagen JA, Sullivan ME. Reliability of a procedural checklist as a high-stakes measurement of advanced technical skill. Am J Surg 2010;199(1):110-4.
6. Ersland K, Kvaløy JT, Styr BM, Helland EB, Espeland A. Do radiologists agree on the quality of computed tomography enterography? J Med Imaging Radiat Oncol 2009;53(4):353-60.
7. Wadsten MA, Sayed-Noor AS, Sjödén GO, Svensson O, Buttazzoni GG. The Buttazzoni classification of distal radial fractures in adults: interobserver and intraobserver reliability. Hand (NY) 2009;4(3):283-8.
8. Haderlein T, Riedhammer K, Nöth E, Toy H, Schuster M, Eysholdt U, et al. Application of automatic speech recognition to quantitative assessment of tracheoesophageal speech with different signal quality. Folia Phoniatr Logop 2009;61(1):12-7.
9. Dedouit F, Bindel S, Gainza D, Blanc A, Joffre F, Rougé D, et al. Application of the Iscan method to two- and three-dimensional imaging of the sternal end of the right fourth rib. J Forensic Sci 2008;53(2):288-95.
10. Fleiss JL. Measuring nominal scale agreement among many raters. Psychological Bulletin.1971;76(5):378-82.
11. Krippendorff K. Reliability. In: Seawell MH, Hoffman CA. Selhort J, eds. Content Analysis, an Introduction to its Methodology, 2nd ed. California. Sage Publications; 2004. p.211-56.
12. Krippendorff K. Reliability in content analysis some common misconceptions and recommendations. Human Communication Research 2004;30(3):411-33.
13. Hayes AF, Krippendorff K. Answering the call for a standard reliability measure for coding data. Communication Methods and Measures 2007;1(1):77-89.