Türkiye Klinikleri Biyoistatistik Dergisi
*Turkiye Klinikleri Journal of Biostatistics* Turkiye Klinikleri J Biostat. 2020;12(3):280-92

ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

DOI: 10.5336/biostatic.2020-77071

# Uyum Katsayıları ve Nominal Verilerde Kategorik Uyum İçin Yeni Bir Yöntem Önerisi

## Agreement Coefficients and a Novel Method Proposal for Categorical Agreement in Nominal Data

İsmet DOĞAN[a], Nurhan DOĞAN[a]

[a]Afyonkarahisar Sağlık Bilimleri Üniversitesi Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişim ABD, Afyonkarahisar, TÜRKİYE

**ÖZET Amaç:** Makalenin amacı, nominal verilerde kategorik uyum için tarafımızdan önerilen yeni bir yöntem sunmak, uyum kavramına genel bir bakış sağlamak ve bilimsel araştırmalardaki önemini vurgulamaktır. **Gereç ve Yöntemler:** Sosyal, davranışsal, fiziksel, biyolojik ve tıbbi bilimlerde güvenilir ve doğru ölçümler, değerlendirme için temel oluşturmaktadır. Yöntem karşılaştırma ve güvenilirlik çalışmalarında, farklı gözlemciler veya enstrümanlarla yapılan çoklu ölçümler arasındaki uyumu değerlendirmek önemlidir. Literatürde, 2 ölçüm arasındaki ilişkiyi veya uyumu özetlemek için çok sayıda indeks geliştirilmiştir. Bu çalışmada, aynı denek ya da örnekteki farklı gözlemciler, yöntemler, araçlar, laboratuvarlar, tahliller, cihazlar vb. ile yapılan ölçme ya da okumaların karşılaştırılmasında kullanılan yöntemler dikkate alınmış, sürekli ve kategorik verilerin söz konusu olduğu durumlarda uyumu değerlendirmek için kullanılan istatistiksel yaklaşımlar gözden geçirilmiştir. Çalışmada, örnek olarak kullanılan veriler tamamen rastgele türetilmiş verilerdir. **Bulgular:** Gerek Helldén tarafından gerekse tarafımızdan önerilen yöntemler, uyum değerlerinin 0-1 arasında kalmasını garanti etmektedir. Üstelik her 2 yöntemin uyumsuzluktan diğer yöntemler kadar etkilenmediği, gerçeğe daha yakın sonuçlar verdiği söylenebilir. **Sonuç:** Kategori uyumu için "uyum oranı" kriterine göre 2 karar verici için Helldén tarafından önerilen yöntemin, 3 ve daha fazla sayıda karar verici için ise tarafımızdan önerilen yöntemin kullanılmasının uygun olacağı sonucuna ulaşılmıştır. Bazı durumlarda alternatif yöntemlerin, değerlendiriciler arasındaki uyumun belirlenmesinde daha uygun olabileceği unutulmamalıdır.

**Anahtar kelimeler:** Uyum; kategorik uyum; yöntem karşılaştırma; güvenilirlik

**ABSTRACT Objective:** The aim of the article is to present a new method proposed by us for categorical agreement in nominal data, to provide an overview of the concept of agreement and to emphasize its importance in scientific research. **Material and Methods:** In social, behavioral, physical, biological, and medical sciences, reliable and accurate measurements serve as the basis for evaluation. In method comparison and reliability studies, it is often important to assess agreement between multiple measurements made by different observers or devices. The literature contains a vast amount of coefficients for summarizing association or agreement between two measurements. In this study, the comparison of measurement or reading methods performed by different observers, methods, instruments, laboratories, tests, devices, etc. on the same subject or sample is dealt with. In addition, statistical approaches were reviewed to evaluate the agreement with continuous and categorical responses. The data used as an example in the study are completely randomly derived data. **Results:** The methods suggested by both Helldén and us ensure that the agreement values remain between 0 and 1. It can be said that both methods are not affected by disagreements as much as other methods and give results closer to the truth. **Conclusion:** According to the "agreement ratio" criterion for category agreement, it was concluded that the method suggested by Helldén would be appropriate for two decision makers and the method suggested by us would be appropriate for three or more decision makers. However, alternative measures of interrater agreement may be more appropriate in certain instances.

**Keywords:** Agreement; categorical agreement; method comparison; reliability

**Correspondence:** Nurhan DOĞAN
Afyonkarahisar Sağlık Bilimleri Üniversitesi Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişim ABD, Afyonkarahisar, TÜRKİYE/TURKEY
**E-mail:** nurhandogan@hotmail.com

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

***Received:*** 02 Jun 2020 ***Received in revised form:*** 15 Oct 2020 ***Accepted:*** 11 Nov 2020 ***Available online:*** 21 Dec 2020

Laboratuvar performansı, enstrüman veya test validasyonu, yöntem karşılaştırmaları, istatistiksel proses kontrolü, uyum iyiliği ve bireysel biyo eş değerlik alanlarında yeni veya jenerik bir prosesin, metodolojinin ve formülasyonun kabul edilebilirliğini değerlendirmek için uyum ölçümlerine ihtiyaç vardır. İki (veya daha fazla) gözlemcinin, 2 (veya daha fazla) cihazın veya aynı gözlemcinin, 2 ve daha fazla sayıdaki ölçümleri arasındaki uyum, istatistikçiler, klinisyenler, epidemiyologlar, psikologlar ve diğer birçok bilim insanı için önemli bir konudur. Ölçümler arasındaki uyum, geçerlilik çalışmaları, randomize kontrollü çalışmalar ve sağlık hizmetlerinin verimli bir şekilde sunulmasını sağlamak için temeldir. Ölçüm hatalarının tamamını kontrol etmek ve ortadan kaldırmak imkânsız olduğu için ölçümlerde uyum derecesi dikkate alınmalı ve değerlendirilmelidir. Bununla birlikte literatürde, az da olsa uyum çalışmaları ile ilgili yanlış analiz veya yanıltıcı sonuçlar ile karşılaşılmaktadır.[1] Girişimsel ve gözlemsel çalışmalarda, anlamlı ve güvenilir sonuçlar elde etmek için ölçümlerin geçerlilik ve güvenilirliğinin yüksek olması çok önemlidir. Geçerlilik, çalışmanın ilgilenilen konu ile ilgili ölçmeyi ne kadar iyi yakaladığı şeklinde tanımlanırken, yüksek güvenilirlik, bir ölçümün zaman içinde, farklı ortamlarda ve farklı değerlendiriciler tarafından tekrarlanabilir olduğu anlamına gelir. Bu, hem farklı değerlendiriciler arasındaki uyumu (değerlendiriciler arası güvenilirlik) hem de aynı değerlendirici tarafından gerçekleştirilen tekrarlanan ölçümlerin (değerlendirici içi güvenilirlik) uyumunu içermektedir.[2] Geleneksel olarak davranışsal değerlendirmelerde, doğrudan davranış gözleme temel bir bileşendir. Araştırmacılar tarafından doğrudan gözlemden elde edilen veriler için psikometrik özellikleri ölçmede kullanılabilecek prosedürler belirlenmiştir. Bu prosedürlerin en yaygın olanları gözlemciler arası uyumdır. Uyum (veya fikir birliği) terimi, derecelendirmelerin ne derece özdeş olduğunu ifade etmektedir ve genellikle özdeş ile ıraksak derecelendirme çiftlerinin oranı kullanılarak hesaplanmaktadır.[3] İki (veya daha fazla) gözlemci veya 2 (veya daha fazla) cihazdan elde edilen yanıtlar arasındaki uyumun homojenliğini ölçmek için çok sayıda davranışsal araştırma uygulaması söz konusudur. Sayısal ölçek kullanılarak elde edilen veriler için klasik sınıf içi korelasyon katsayısı [intraclass correlation coefficient (ICC)] ve son zamanlarda sıklıkla kullanılan uyum indeksleri alternatif yöntemlerdir. Korelasyon katsayılarının kullanılması, değerlendiriciler arasındaki uyumun aşırı yüksek olarak elde edilmesine veya değerlendiriciler arasındaki gerçek uyum seviyesinin göz ardı edilmesine yol açabileceği için değerlendiriciler arasındaki uyum miktarının ancak zayıf bir yansıması olabilir.[4] Uyum, ölçümler arasındaki yakınlığı ölçtüğünden hem doğruluk hem de kesinlik içeren daha geniş bir terimdir. Ölçümlerden biri referans olarak kabul edildiğinde uyum, geçerlilik ile ilgilidir. Bundan dolayı araştırmacılardan bir kısmı uyumu, bir kısmı korelasyonu, bir kısmı ise hem uyumu hem de kovaryasyonu 2 farklı güvenilirlik göstergesi olarak kabul etmektedirler.[5] Uyum, birlikteliğin özel bir durumu olarak görülebilir. Neredeyse tüm yaşam veya sosyal bilim araştırmalarında konular; değerlendiriciler, görüşmeciler veya gözlemciler tarafından kategorilere ayrılmaktadır. Bu araştırmalardan elde edilen yanıtlardan, hem ilişki hem de uyum değerleri elde edilebilir. İlişki, bir yanıtın kategorisinin diğer yanıtın kategorisinden tahmin edilmesini, uyum ise 2 yanıtın da aynı kategoride yer almasını gerektirmektedir. Bu nedenle, 2 yanıt arasında bir uyum varsa doğal olarak ilişki de söz konusudur, ancak 2 yanıt arasında güçlü bir uyum olmaksızın güçlü bir ilişki de olabilir.[6] Farklı bir birliktelik göstergesi olarak uyum ölçüsü, istatistiklerde özel bir rol oynar. İyi tanımlanmış bir uyum ölçüsü, gözlemciler arası veya test-tekrar test güvenilirliğini açıklamaktadır. Sonuç olarak iyi tanımlanmış bir uyum ölçüsü, tek bir derecelendirmenin sınırlarının ötesinde, bir gözlemin genelleştirilebilirliğinin bir göstergesidir. Dolayısıyla bir bilimsel çalışmada yer alan derecelendirmeler için böyle bir uyum ölçüsünün bildirilmesi, sonuçların başka zamanlarda veya başka yerlerde ne kadar tekrarlanabilir olabileceğini gösterir.[7] Literatürde, farklı durumlarda kullanılmak üzere çok sayıda uyum indeksleri önerilmektedir. Benzer konularda yapılan derecelendirmeler, bir değerlendiriciden diğerine büyük farklılıklar gösterebilir. Çeşitli araştırma alanlarındaki birçok araştırmacı, bu sorunu uzun zamandır kabul etmektedir. Bu makalenin temel amacı, uyum kavramına geniş bir bakış açısı sunmak ve bilimsel araştırmalardaki önemini vurgulamaktır. Makalede, Helsinki Deklarasyonu Prensipleri dikkate alınmıştır.

## GEREÇ VE YÖNTEMLER

Bireyleri/birimleri, birbiri ile örtüşmeyen kategorilerden birine sınıflandıran çeşitli değerlendiriciler arasındaki uyum miktarının ölçülmesi biyoistatistik ve epidemiyoloji için giderek daha önemli hâle gelen bir sorundur. Uygulamalı istatistik ile ilgili literatürde, 2 veya daha fazla sınıflandırma mekanizması arasındaki uyum miktarını belirlemek için farklı uyum indeksleri yer almaktadır.[8] Evrensel bir uyum indeksinin;

- İstatistiksel bir temeli olmalıdır,
- Çok değişkenli verileri analiz eden uyum ölçüleri, tek değişkenli verileri analiz eden uyum ölçülerine göre tartışmasız bir avantaja sahiptir,
- Herhangi bir ölçüm düzeyine (nominal, sıra, aralık veya oran) sahip verileri analiz edebilmelidir,
- İkiden fazla değerlendiriciden gelen bilgileri değerlendirebilmelidir,

biçiminde sıralanan niteliklere sahip olması gerekmektedir.[9] İstatistik, biyoistatistik, psikoloji, psikiyatri, eğitim ve sosyoloji alanlarındaki araştırmacılar tarafından önerilen indeksler Tablo 1 ve Tablo 2'de görülmektedir.

**TABLO 1:** İki değerlendirici için uyum indeksleri ve kronoloji.

| Uyum indeksi | Kronoloji |
|---|---|
| $a_0$ | Unknown, pre 1901[10] |
| $\beta$ | Benini, 1901[11] |
| $Yule_1$ | Yule, 1912[12] |
| $G_1, G_2, G_3$ | Gini, 1914-1915[13] |
| $\rho$ | Guttman, 1946[14] |
| $\lambda$ | Goodman & Kruskal, 1954[15] |
| $ICC$ | Fisher, 1954[16] |
| $S$ | Bennett, et al., 1954[17] |
| $\pi$ | Scott, 1955[18] |
| $CR$ | Osgood, 1959[19] |
| $\kappa$ | Cohen, 1960[20] |
| $G$ | Holley & Guilford, 1964[21] |
| $A_1, A_2$ | Rogot & Goldberg, 1966[22] |
| $CR$ | Holsti, 1969[23] |
| $\alpha$ | Krippendorff, 1970[24] |
| $\kappa_c$ | Cicchetti, 1976[25] |
| $RE$ | Maxwell, 1977[26] |
| $C$ | Janson & Vegelius, 1979[27] |
| $\kappa_n$ | Brennan & Prediger, 1981[28] |
| Bland-Altman | Altman & Bland, 1983[29] |
| $D_2$ | Popping,1983[30] |
| $B$ | Bangdiwala, 1985[31] |
| $I_r$ | Perreault & Leigh, 1989[32] |
| $\rho_C$ | Lin, 1989[33] |
| $C_{AB}$ | Kupper & Hafner, 1989[34] |
| $\alpha$ | Aickin, 1990[8] |
| $PABAK$ | Byrt, et al., 1993[35] |
| $\kappa_T, \kappa_S$ | Donner & Eliasziw, 1997[36] |
| $PA$ | Svensson, 1997[37] |
| $\delta$ | Andrés & Marzo, 2004[38] |
| $\kappa_S$ | von Eye, 2006[39] |
| $\psi^N$ | Barnhart, et al., 2007[40] |
| $AC_1$ | Gwet, 2008[41] |
| $AC_2$ | Gwet, 2012[42] |
| $\kappa_W$ | Warrens, 2013[43] |

**TABLO 2:** İki veya daha fazla değerlendirici için uyum indeksleri ve kronoloji.

| Uyum indeksi | Kronoloji |
|---|---|
| $w$ | Kendall, 1962[44] |
| $r_1$ | Maxwell & Pilliner, 1968[45] |
| $\kappa_L$ | Light, 1971[46] |
| $\kappa_F$ | Fleiss, 1971[47] |
| $T$ | Tinsley & Weiss, 1975[48] |
| $\kappa_H$ | Hubert, 1977[49] |
| $\rho$ | Landis & Koch, 1977[50] |
| $\kappa_{FC}$ | Fleiss & Cuzick, 1979[51] |
| $\kappa_C$ | Conger, 1980[52] |
| $\kappa_0$ | Kraemer, 1980[53] |
| $k_W$ | Schouten, 1980[54] |
| $P_{e(m/h)}$ | Craig, 1981[55] |
| $\kappa_{DF}$ | Davies & Fleiss, 1982[56] |
| $S_{av}$ | O'Connell & Dobson, 1984[57] |
| $Rev-K$ | Siegel & Castellan, 1988[58] |
| $Rdf-P_i$ | Potter & Levine-Donnerstein, 1999[59] |
| $\kappa_{free}$ | Randolph, 2005[60] |
| $\kappa_M$ | Mielke, et al., 2009[61] |
| Sklar's ω | Hughes, 2018[62] |
| $I_r^2$ | van Oest, 2018[63] |

Tablo 1 ve Tablo 2'de kronolojisi verilen indekslere ait formüller, aşırı derecede yer kapladığından ve farklı uyarlamalarından dolayı karışıklığa sebep olmaması için bu çalışmada özellikle verilmemiştir. Ayrıntılı bilgi için ilgili literatürün okunması tavsiye edilmektedir. Tablo 1'de verilen uyum indeksleri, 2 değerlendirici arasındaki uyum hesaplamaları için; Tablo 2'de verilenler ise 2 ve daha fazla değerlendirici arasındaki uyum hesaplamaları için kullanılabilir. Bu indekslerin çoğu ($\psi^N$, $\rho_C$, $r_1$ *ve* $T$ *hariç*), kesikli ve sıralı yanıtlar üreten bir ölçüm aracı üzerinde 2 ve daha fazla değerlendiricinin uyumuna ilişkin hesaplamalar için kullanılabilir. Sürekli ölçekte yanıtlar üreten bir ölçüm aracı üzerinde, 2 ve daha fazla değerlendirici arasındaki uyum ise genellikle ICC'den biri ile değerlendirilir. Gözlemci uyumunun, aynı konudaki farklı gözlemcilerin derecelendirmelerinin farkı (mutlak uyum) olarak tanımlanması durumunda, kullanılacak uyum indeksi aynı zamanda bir ICC olan $\rho_C$'dir. $\rho_C$, aynı konudaki gözlemcilerin derecelendirmeleri arasındaki ortalama kare farkını, gözlemciler arasında bağımsızlık olarak tanımlanan "şans uyumu" varsayımı altında bu miktarın beklenen değeri ile karşılaştırmaktadır.[64] $\rho_C$ ve ICC, benzer indeksler olmalarına rağmen aralarında bazı farklılıklar bulunmaktadır. ICC, hem sabit hem de rastgele gözlemciler için; $\rho_C$ ise genellikle sabit gözlemciler için önerilmektedir. Ayrıca ICC'nin, varyans analizi model varsayımlarını sağlaması gerekirken; $\rho_C$ için buna gerek duyulmamaktadır. Bütün bunlara rağmen belirli durumlarda ICC ve $\rho_C$ benzer değerlere sahiptir.[65] Ancak ICC ve $\rho_C$, boylamsal (longitudinal) tip ardışık ölçümler içeren çalışmalarda uyum hesaplamaları için kullanılamaz. İki veya daha fazla değerlendiricinin, ardışık ölçümlerinin uyum hesaplamasında kullanılabilecek tek indeks $\psi^N$ idi.[66] Derecelendirmelerin nasıl analiz edilmesi gerektiği, büyük ölçüde verilerin türüne ve analizin nihai hedeflerine bağlıdır. Derecelendirmeler nominal, sıra, aralık veya oran türlerinde olabilir. Nominal ölçekler için önerilen uyum indeksleri sıralı, aralıklı veya oran ölçekleri için veya bunun tersine oran verilerinin analizi için uygun uyum indeksleri, nominal verilerin analizinde yetersiz olacaktır.[67] Kategorik sınıflandırmalarda uyum indekslerinin değeri ve yorumu, kategori sayısı ile ilişkili olduğundan kategori sayısı belirtilmelidir. Sürekli ölçümler söz konusu olduğunda, değerlendiriciler arasındaki uyum indekslerinin değeri, veri aralığına (range) bağlıdır. Sürekli ölçümlerin, kategorik verilere dönüştürülmesi durumunda, her bir kategori için belirlenen sınır değerlerinin açıkça belirtilmesi gereklidir.[68]

## KATEGORİK UYUM VE BİR UYGULAMA

Tablo 1 ve Tablo 2'de verilen uyum indeksleri, değerlendiricilerden elde edilen tüm bilgileri dikkate alarak hesaplanmaktadır. Oysa bu bilgilerden yararlanarak belirli bir i'nci kategori için değerlendiriciler arasındaki uyumun belirlenmesi istenebilir. Bu tür durumlarda kullanılan indeksler ve kronolojisi Tablo 3'te görülmektedir.

**TABLO 3:** Nominal verilerde $i$'nci kategoriye ait iki değerlendirici için uyum indeksleri ve kronoloji.

| Uyum indeksi | Kronoloji |
|---|---|
| $\hat{\kappa}_i$ | Bishop, et al., 1975[69] |
| $\vartheta_i$ | Türk, 1979[70] |
| $MA_H$ | Helldén, 1980[71] |
| $MA_S$ | Short, 1982[72] |

Tablo 3'te görüldüğü üzere i'nci kategori için 2 değerlendirici arasındaki uyumun hesaplanmasında literatürde farklı indeksler bulunmaktadır. $X_{ii}$: i'nci satır ve i'nci sütundaki gözlem sayısı, $X_{i+}$: i'nci satır toplamı, $X_{+i}$: i'nci sütun toplamı ve $N$: toplam gözlem sayısı olmak üzere, Bishop ve ark. tarafından;

$$\hat{\kappa}_i = \frac{NX_{ii} - X_{i+}X_{+i}}{NX_{i+} - X_{i+}X_{+i}} \quad (1)$$

Helldén tarafından;

$$MA_H = \frac{2 * X_{ii}}{(X_{i+}+X_{+i})} \quad (2)$$

Short tarafından;

$$MA_S = \frac{X_{ii}}{(X_{i+}+X_{+i} - X_{ii})} \quad (3)$$

eşitliklerinin kullanılması önerilmiştir. $\hat{\kappa}_i$'nin, Cohen kappa indeksiyle; $MA_S$'nin, Jaccard indeksiyle benzer özelliklere sahip olduğu, $MA_H$ indeksinin ise herhangi bir olasılık veya matematiksel temelinin olmadığı Helldén tarafından tamamen mantıksal (sezgisel) olarak ortaya konulan bir indeks olduğu ifade edilmektedir.[73] Eşitlik 1'de verilen indeks, uyumun 0 (tam uyumsuzluk) ile 1 (tam uyum) arasında kalmasını garanti etmemekte, negatif değerler alabilmektedir. Eşitlik 3'te verilen indeks ise değerlendiriciler arasındaki yüksek uyumu sayısal olarak Eşitlik 2'de verilen indeks kadar yansıtamamaktadır. Dolayısı ile Helldén tarafından önerilen indeks, diğer indekslere tercih edilmelidir. Kategorik uyum, örnek bir uygulama ile aşağıda gösterilmiştir. Örnek için kullanılan veriler tamamen rastgele türetilmiş verilerdir. Bir hastanenin psikiyatri polikliniğine müracaat eden hastaların, 2 farklı psikiyatrist tarafından konulan tanılarına göre dağılımları Tablo 4'te verildiği gibi olsun.

**TABLO 4:** Kategorik uyum için örnek veri.

| | | Psikiyatrist 1 | | | |
|---|---|---|---|---|---|
| | | Depresyon | Kişilik bozukluğu | Şizofreni | Nevroz |
| Psikiyatrist 2 | Depresyon | 9 | 1 | 4 | 6 |
| | Kişilik bozukluğu | 1 | 10 | 2 | 7 |
| | Şizofreni | 4 | 2 | 17 | 0 |
| | Nevroz | 23 | 7 | 1 | 6 |

Tablo 4'te verilen bilgiler kullanılarak, 2 psikiyatrist arasındaki genel uyum için kappa katsayısı hesaplanmış ve uyum değeri 0,237 bulunmuştur. Kategori uyum değerleri ise;

| | Bishop, et al. | Helldén | Short |
|---|---|---|---|
| Depresyon | 0,1270 | 0,3158 | 0,1875 |
| Kişilik bozukluğu | 0,3750 | 0,5000 | 0,3333 |
| Şizofreni | 0,6568 | 0,7234 | 0,5667 |
| Nevroz | -0,0344 | 0,2143 | 0,1200 |

olarak elde edilmiştir. Kategori uyumu için Eşitlik 1, Eşitlik 2 ve Eşitlik 3 ile verilen yöntemlerden hangisinin en güvenilir yöntem olduğunu belirlemek amacıyla performans kriteri olarak "uyum oranı" dikkate alınmış ve uyum için elde edilen değerler uyum oranı ile karşılaştırılmıştır. Uyum oranı, bir çapraz tablonun ana köşegeninde yer alan değerlerin toplamının, toplam gözlem sayısına bölümünden elde edilen değerdir. Farklı kategorilere ait uyum oranı hesaplamalarında tablolar, her bir kategori için yeniden düzenlenmelidir. Örneğin Tablo 4'te verilen değerlerden yararlanarak, depresyon kategorisi için uyum oranı hesaplamasında kullanılacak çapraz tablo,

| Depresyon | | Psikiyatrist 1 | | |
|---|---|---|---|---|
| | | Evet | Hayır | Toplam |
| Psikiyatrist 2 | Evet | 9 | 11 | 20 |
| | Hayır | 28 | 52 | 80 |
| | Toplam | 37 | 63 | 100 |

şeklinde olacaktır. Bu tablodan yararlanarak depresyon kategorisi için hesaplanan uyum oranı 0,61'dir. Ancak uyum oranının yüksek olması, uyum değerinin de yüksek olacağı anlamına gelmemelidir. Hazırlanan tablodan da anlaşılacağı üzere uyum oranının elde edilmesinde kullanılan köşegen öğelerinden ilk köşegen değerinin büyük olması, söz konusu kategori için elde edilecek uyum değerinin yüksek olacağını, küçük olması ise bu kategori için elde edilecek olan uyum değerinin düşük olacağını göstermektedir.

## KATEGORİK UYUM İÇİN YENİ BİR ÖNERİ

Üç ve daha fazla sayıda değerlendiricinin söz konusu olduğu durumlarda, kategorik uyum indeksi ile ilgili literatürde herhangi bir öneri bulunmamaktadır. Dolayısıyla bu tür durumlar için Eşitlik 4 ile verilen indeksin hesaplanması tarafımızdan önerilmektedir.

$$IND_{Kategori-i} = \frac{\sum_{k=1}^{DS-1}\sum_{l=k+1}^{DS} w_{kl} MA_{H_{kl}}}{\sum_{k=1}^{DS-1}\sum_{l=k+1}^{DS} w_{kl}} \qquad (4)$$

$DS$ : Değerlendirici sayısı;

$w_{kl}$ : $k$ ve $l$ değerlendiricilerinin i'nci kategoriye ait ortak gözlem sayısı;

$MA_{H_{kl}}$ : $k$ ve $l$ değerlendiricilerinin i'nci kategoriye ait Eşitlik 2'den elde edilen uyum miktarı.

Tarafımızdan önerilen indeksin hesaplanmasındaki adımlar, örnek bir uygulama ile aşağıda anlatılmıştır. Örnek için kullanılan veriler tamamen rastgele türetilmiş verilerdir. P1, P2 ve P3 ile gösterilen 3 farklı değerlendiricinin 4 farklı kategori için vermiş oldukları kararlar Tablo 5'te, ikili olarak verdikleri kararlar ise Tablo 6a, Tablo 6b ve Tablo 6c'de verildiği gibi olsun.

**TABLO 5:** Karar vericilerin kararlarının kategorilere göre dağılımı.

| Kategori | P1 | P2 | P3 |
|---|---|---|---|
| Depresyon | 55 | 48 | 44 |
| Kişilik bozukluğu | 33 | 38 | 38 |
| Şizofreni | 48 | 45 | 50 |
| Nevroz | 27 | 32 | 31 |
| Toplam | 163 | 163 | 163 |

**TABLO 6a:** Örnek veri.

| Kategori | | P1 | | | | |
|---|---|---|---|---|---|---|
| | | Depresyon | Kişilik bozukluğu | Şizofreni | Nevroz | Toplam |
| | Depresyon | 48 | 0 | 0 | 0 | 48 |
| | Kişilik bozukluğu | 0 | 33 | 5 | 0 | 38 |
| P2 | Şizofreni | 5 | 0 | 40 | 0 | 45 |
| | Nevroz | 2 | 0 | 3 | 27 | 32 |
| | Toplam | 55 | 33 | 48 | 27 | 163 |

**TABLO 6b:** Örnek veri.

| Kategori | | P1 | | | | |
|---|---|---|---|---|---|---|
| | | Depresyon | Kişilik bozukluğu | Şizofreni | Nevroz | Toplam |
| | Depresyon | 44 | 0 | 0 | 0 | 44 |
| | Kişilik bozukluğu | 0 | 33 | 5 | 0 | 38 |
| P3 | Şizofreni | 6 | 0 | 40 | 4 | 50 |
| | Nevroz | 5 | 0 | 3 | 23 | 31 |
| | Toplam | 55 | 33 | 48 | 27 | 163 |

**TABLO 6c:** Örnek veri.

| Kategori | | P2 | | | | |
|---|---|---|---|---|---|---|
| | | Depresyon | Kişilik bozukluğu | Şizofreni | Nevroz | Toplam |
| | Depresyon | 44 | 0 | 0 | 0 | 44 |
| | Kişilik bozukluğu | 0 | 38 | 0 | 0 | 38 |
| P3 | Şizofreni | 3 | 0 | 45 | 2 | 50 |
| | Nevroz | 1 | 0 | 0 | 30 | 31 |
| | Toplam | 48 | 38 | 45 | 32 | 163 |

Üç değerlendiricinin, her bir kategori için vermiş oldukları kararlar bakımından uyumları aşağıdaki gibi hesaplanabilir;

**Adım 1.** i'nci kategori için tüm değerlendiricilerin (örneğimiz için P1, P2 ve P3) ikili kombinasyonları dikkate alınarak, Eşitlik 2 ile verilen formül yardımıyla tüm ikili kombinasyonlar için uyum hesaplanır.

| Karar vericiler | Kategori | Uyum | Kappa |
|---|---|---|---|
| P1-P2 | Depresyon | 0,9320 | 0,8773 |
| | Kişilik bozukluğu | 0,9296 | |
| | Şizofreni | 0,8602 | |
| | Nevroz | 0,9153 | |
| P1-P3 | Depresyon | 0,8889 | 0,8118 |
| | Kişilik bozukluğu | 0,9296 | |
| | Şizofreni | 0,8163 | |
| | Nevroz | 0,7931 | |
| P2-P3 | Depresyon | 0,9565 | 0,9509 |
| | Kişilik bozukluğu | 1,0000 | |
| | Şizofreni | 0,9474 | |
| | Nevroz | 0,9524 | |

**Adım 2.** i'nci kategori için ikili kombinasyonlarda yer alan her 2 karar vericinin i'nci kategoriye ait ortak gözlem sayıları ağırlık katsayısı olarak belirlenir.

| Karar vericiler | Kategori | Ağırlık katsayısı |
|---|---|---|
| P1-P2 | Depresyon | 48 |
| | Kişilik bozukluğu | 33 |
| | Şizofreni | 40 |
| | Nevroz | 27 |
| P1-P3 | Depresyon | 44 |
| | Kişilik bozukluğu | 33 |
| | Şizofreni | 40 |
| | Nevroz | 23 |
| P2-P3 | Depresyon | 44 |
| | Kişilik bozukluğu | 38 |
| | Şizofreni | 45 |
| | Nevroz | 30 |

**Adım 3.** i'nci kategori için her bir ikili kombinasyondan elde edilen uyum katsayıları ilgili ikili için belirlenen ağırlık katsayısı ile çarpılarak ağırlıklandırılır. Ağırlıklandırılmış uyum değerleri toplanır. Her bir kategori için elde edilen toplam uyum değeri, her bir kategori için ağırlıklandırma amacıyla kullanılan gözlem sayılarının toplamına bölünür.

| Kategori | Karar vericiler | Uyum (1) | Ağırlık katsayısı (2) | (1)*(2) | Kappa | $IND_{Kategori-i}$ |
|---|---|---|---|---|---|---|
| Depresyon | P1-P2 | 0,9320 | 48 | 44,7360 | 0,8932 | 0,9259 |
| | P1-P3 | 0,8889 | 44 | 39,1116 | | |
| | P2-P3 | 0,9565 | 44 | 42,0860 | | |
| Kişilik bozukluğu | P1-P2 | 0,9296 | 33 | 30,6768 | 0,9410 | 0,9553 |
| | P1-P3 | 0,9296 | 33 | 30,6768 | | |
| | P2-P3 | 1,0000 | 38 | 38,0000 | | |
| Şizofreni | P1-P2 | 0,8602 | 40 | 34,4080 | 0,8221 | 0,8775 |
| | P1-P3 | 0,8163 | 40 | 32,6520 | | |
| | P2-P3 | 0,9474 | 45 | 42,6330 | | |
| Nevroz | P1-P2 | 0,9153 | 27 | 24,7131 | 0,8638 | 0,8940 |
| | P1-P3 | 0,7931 | 23 | 18,2413 | | |
| | P2-P3 | 0,9524 | 30 | 28,5720 | | |

$$IND_D = \frac{44{,}736 + 39{,}1116 + 42{,}0860}{48 + 44 + 44} = \frac{125{,}9336}{136} = 0{,}9259$$

$$IND_{KB} = \frac{30{,}6768 + 30{,}6768 + 38}{33 + 33 + 38} = \frac{99{,}3536}{104} = 0{,}9553$$

$$IND_{\c{S}} = \frac{34{,}408 + 32{,}652 + 42{,}633}{40 + 40 + 45} = \frac{109{,}693}{125} = 0{,}8775$$

$$IND_N = \frac{24{,}7131 + 18{,}2413 + 28{,}572}{27 + 23 + 30} = \frac{71{,}5264}{80} = 0{,}8940$$

## BULGULAR

Her bir kategori için ayrı ayrı uyumun hesaplanması, karar vericilerin en az veya en fazla uyumlu oldukları kategorilerin hangileri oldukları konusunda detaylı bilgi sunmasından dolayı önem arz etmektedir. Kategorik uyumun hesaplanması, özellikle genel uyumun düşük olduğu durumlarda karar vericilerin tüm kararlarından ziyade en az uyumlu oldukları kategori ile ilgili kararlarını yeniden gözden geçirmeleri konusunda yol gösterici olacaktır. Tarafımızdan önerilen yöntem, Helldén tarafından önerilen yönteme dayalı bir yöntemdir. Dolayısıyla Helldén tarafından önerilen yöntemin, gerek avantajları gerekse dezavantajları tarafımızdan önerilen yöntem için de geçerlidir. Tablo 6a, Tablo 6b ve Tablo 6c ile verilen değerler dikkate alınarak elde edilen sonuçlar, Tablo 7'de gösterilmiştir. Tablo 7'de verilen sonuçlar dikkate alındığında, Bishop ve ark. tarafından önerilen eşitlikten elde edilen sonuçların güvenirliğinin, diğer yöntemlerden elde edilen sonuçlara göre daha düşük olduğu söylenebilir. Çünkü Bishop ve ark. tarafından önerilen eşitlik, satır toplamından etkilenmektedir. Dolayısıyla karar vericilerin çapraz tabloda yerlerinin değişmesi, elde edilecek sonucu doğrudan etkilemektedir. Ayrıca 2 karar verici arasındaki uyum katsayısının 1 değerini alması, ancak kararların tıpa tıp aynı olması durumunda söz konusudur. Oysa Tablo 5'ten de görüldüğü üzere karar vericilerin kararları tıpa tıp benzerlik göstermemektedir. Short tarafından önerilen eşitlikten elde edilen değerler ise uyum oranı değerlerine göre oldukça düşüktür. Tablo 7'den de görüldüğü üzere uyum oranına en yakın uyum değerleri Helldén tarafından önerilen eşitlikten elde edilmektedir.

**TABLO 7:** Uyum oranı ve katsayıları.

| Kategori | Karar vericiler | Uyum oranı | Bishop | Helldén | Short | Kappa | $IND_{Kategori-i}$ |
|---|---|---|---|---|---|---|---|
| Depresyon | P1-P2 | 0,9571 | 1,0000 | 0,9320 | 0,8727 | 0,9009 | 0,9259 |
| | P1-P3 | 0,9325 | 1,0000 | 0,8889 | 0,8000 | 0,8413 | |
| | P2-P3 | 0,9755 | 1,0000 | 0,9565 | 0,9167 | 0,9395 | |
| Kişilik bozukluğu | P1-P2 | 0,9693 | 0,8350 | 0,9296 | 0,8684 | 0,9101 | 0,9553 |
| | P1-P3 | 0,9693 | 0,8350 | 0,9296 | 0,8684 | 0,9101 | |
| | P2-P3 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | |
| Şizofreni | P1-P2 | 0,9202 | 0,8425 | 0,8602 | 0,7547 | 0,8045 | 0,8775 |
| | P1-P3 | 0,8896 | 0,7165 | 0,8163 | 0,6897 | 0,7374 | |
| | P2-P3 | 0,9693 | 0,8619 | 0,9474 | 0,9000 | 0,9258 | |
| Nevroz | P1-P2 | 0,9693 | 0,8127 | 0,9153 | 0,8438 | 0,8967 | 0,8940 |
| | P1-P3 | 0,9264 | 0,6907 | 0,7931 | 0,6571 | 0,7486 | |
| | P2-P3 | 0,9816 | 0,9599 | 0,9524 | 0,9091 | 0,9410 | |

Tarafımızdan önerilen yöntemden elde edilen uyum değerlerinin, güvenilir olup olmadığının belirlenmesi amacıyla her bir kategoriye ait ortalama uyum oranları, önerilen yöntemden elde edilen uyum değerleri ile karşılaştırılmıştır. Tablo 7'de verilen uyum oran değerleri dikkate alındığında depresyon için ortalama uyum oranı 0,9550; önerilen yöntemden elde edilen uyum değeri 0,9259; kişilik bozukluğu için ortalama uyum oranı 0,9795; önerilen yöntemden elde edilen uyum değeri 0,9553; şizofreni için ortalama uyum oranı 0,9263; önerilen yöntemden elde edilen uyum değeri 0,8775; nevroz için ortalama uyum oranı 0,9591; önerilen yöntemden elde edilen uyum değeri 0,8940 olarak hesaplanmıştır. Helldén tarafından önerilen yöntem, herhangi bir karar verici ikilisinin uyumsuzluğundan diğer yöntemler kadar etkilenmemekte, gerçeğe daha uygun sonuçlar vermektedir. Helldén tarafından önerilen yöntemin bu üstünlüğü, tarafımızdan önerilen yönteme de doğal olarak yansımaktadır.

## TARTIŞMA

Değerlendiriciler arası uyum, farklı değerlendiricilerin aynı şeyi değerlendirirken birbirinin yerine geçebildiği zaman gerçekleşir. Bu nedenle, 2 veya daha fazla bağımsız değerlendiricinin aynı şeyi değerlendirdiği herhangi bir durumda değerlendiriciler arası varyasyon ölçülebilir. Değerlendiriciler arasındaki uyumun değerlendirilmesi, gerek yöntem karşılaştırma çalışmalarında gerekse güvenilirlik çalışmalarında önemlidir. Bilimsel araştırmalardaki bazı istatistiksel problemler, 2 veya daha fazla değerlendirici arasındaki ilişkilendirme veya korelasyondan ziyade uyumun ölçülmesini gerektirir. Gözlemsel çalışmalarda toplanan verilerin kalitesinin en yaygın göstergesi olarak değerlendiriciler arası uyum yüzdesi kullanılmaktadır. Değerlendiriciler arası uyum analizleri ile ilgili literatür, son yıllarda büyük ölçüde artmıştır. Tıp ve diğer ilgili bilim dallarında, değerlendiriciler arasındaki uyumu değerlendirmek için birçok istatistiksel yaklaşım önerilmiştir. Uyum, derecelendirmeler arasındaki yakınlığı ölçtüğünden, hem doğruluk hem de kesinlik içeren daha geniş bir terimdir. Derecelendirmelerden birinin referans olarak kabul edilmesi durumunda uyum, geçerlilik ile ilgilidir. Tüm derecelendirmelerin aynı temel dağılımdan geldiği varsayıldığında ise uyum, derecelendirmelerin ortalama etrafında kesinliğini değerlendirmektedir. Tıbbi araştırmalarda uyum, genellikle yöntem karşılaştırmaları, test doğrulaması ve bireysel biyo eş değerlik için kullanılmaktadır.[65] Uyum indeksleri, çok sayıda araştırmada kullanılmış olmasına rağmen çeşitli yazarlar tarafından da eleştirilmektedir. Uyum çalışmalarının sonuçları herhangi bir tanı, puan veya ölçümde bulunan hata miktarı hakkında bilgi sağlamayı amaçlamaktadır. Ölçek, enstrüman veya sınıflandırma kullanıcıları arasındaki uyum düzeyi yaygın olarak bilinmemektedir. Bu nedenle, titizlikle yürütülen değerlendiriciler arası ve değerlendiriciler içi güvenilirlik ve uyum çalışmalarına ihtiyaç vardır. Literatürde, birçok farklı uyum indeksi bulunmasına rağmen farklı indeksler farklı sonuçlara yol açabilir ve veri kalitesinin genel bir göstergesi olarak hangi indeksin tercih edilen seçenek olması gerektiği konusu açık değildir.[74] Ne yazık ki Tablo 1 ve Tablo 2'de verilen yöntemlerin hiçbiri evrensel bir uyum ölçüsünün taşıması gereken niteliklerin tamamına sahip değildir. Bununla birlikte, kappa ve ağırlıklı kappa, sırasıyla kesikli ve sıralı derecelendirmeler için ICC, $\psi^N$ ve $\rho_C$ indeksleri ise sürekli ölçeğe sahip derecelendirmeler için en popüler uyum indeksleridir.[75] Literatürde, 2 karar verici olması durumunda kategori uyumu ile ilgili yöntem karşılaştırmaları içeren bir adet çalışma bulunmaktadır. Bu çalışmada, Helldén tarafından önerilen yöntemin, diğer yöntemlere göre gerçeğe daha yakın sonuçlar verdiği ifade edilmektedir. Benzer sonuç, çalışmamız için de geçerlidir. Tarafımızdan önerilen yöntem, karar verici ikililerinden elde edilen sonuçların uygun bir ağırlıklandırma işlemi ile tek bir sonuca indirgenmesinden ibarettir. Dolayısıyla 3 ve daha fazla sayıda karar verici için tarafımızdan önerilen yöntem Helldén tarafından önerilen yöntemin bütün avantajlarını taşımaktadır.

## SONUÇ

Bu çalışma ile "uyum oranı" kriterine göre kategori uyumu ile ilgili 2 karar verici için Helldén tarafından önerilen yöntemin, 3 ve daha fazla sayıda karar verici için ise tarafımızdan önerilen yöntemin kullanılmasının uygun olacağı düşünülmektedir. Gerek Helldén tarafından gerekse tarafımızdan önerilen yöntemler,

uyum değerlerinin 0-1 arasında kalmasını garanti etmektedir. Üstelik her iki yöntem, herhangi bir karar verici ikilisi arasındaki aşırı uyumsuzluktan diğer yöntemler kadar etkilenmemekte, gerçeğe daha yakın sonuçlar vermektedir. Ancak, Helldén tarafından önerilen yöntem tamamen mantıksal (sezgisel) olarak ortaya konulan bir indeks olduğu için güven aralığı ve yansızlık gibi performans kriterlerinin elde edilmesi konusunda yetersiz kalmaktadır. Bu durum, tarafımızdan önerilen yöntem için de geçerlidir.

### *Finansal Kaynak*

*Bu çalışma sırasında, yapılan araştırma konusu ile ilgili doğrudan bağlantısı bulunan herhangi bir ilaç firmasından, tıbbi alet, gereç ve malzeme sağlayan ve/veya üreten bir firma veya herhangi bir ticari firmadan, çalışmanın değerlendirme sürecinde, çalışma ile ilgili verilecek kararı olumsuz etkileyebilecek maddi ve/veya manevi herhangi bir destek alınmamıştır.*

### *Çıkar Çatışması*

*Bu çalışma ile ilgili olarak yazarların ve/veya aile bireylerinin çıkar çatışması potansiyeli olabilecek bilimsel ve tıbbi komite üyeliği veya üyeleri ile ilişkisi, danışmanlık, bilirkişilik, herhangi bir firmada çalışma durumu, hissedarlık ve benzer durumları yoktur.*

### *Yazar Katkıları*

*Bu çalışma hazırlanırken tüm yazarlar eşit katkı sağlamıştır.*

## REFERENCES

1. Lin L, Hedayat AS, Sinha B, Yang M. Statistical methods in assessing agreement: models, issues and tools. J Am Stat Assoc. 2002;97(457):257-70.[Crossref]
2. Zapf A, Castell S, Morawietz L, Karch A. Measuring inter-rater reliability for nominal data - which coefficients and confidence intervals are appropriate? BMC Med Res Methodol. 2016;16:93.[Crossref] [PubMed] [PMC]
3. Bryington AA, Palmer DJ, Watkins MW. The estimation of interobserver agreement in behavioral assessment. The Behavior Analyst Today. 2002;3(3):323-8.[Crossref]
4. Ato M, López JJ, Benavente A. A simulation study of rater agreement measures with 2x2 contingency tables. Psicológica. 2011;32(2):385-402.
5. Xinshu Z, Liu JS, Deng K. Assumptions behind intercoder reliability indices. In: Solmon CT, ed. Communication Yearbook 36. New York: Routledge; 2013. p.419-80.[Crossref]
6. Adejumo AO, Heumann C, Toutenburg H. A review of agreement measure as a subset of association measure between raters. SFB386-Discussion Paper 385, Ludwig-Maximilians-Universitat, München, 2004.[Link]
7. Bloch DA, Kraemer HC. 2 x 2 kappa coefficients: measures of agreement or association. Biometrics. 1989;45(1):269-87.[Crossref] [PubMed]
8. Aickin M. Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's kappa. Biometrics. 1990;46(2):293-302.[Crossref] [PubMed]
9. Berry KJ, Mielke PW. A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters. Educ Psychol Meas. 1988;48(4):921-33.[Crossref]
10. Krippendorff K. Reliability in content analysis: some common misconceptions and recommendations. Hum Commun Res. 2004;30(3):411-33.[Crossref]
11. Benini R. No. 29 of Manuali Barbèra di Scienza Giuridiche Sociale e Politiche. In: Firenze GB, ed. Principii di Demografia. 1901.
12. Yule GU. On the methods of measuring the association between two attributes. J R Stat Soc. 1912;75(6):579-652.[Crossref]
13. Gini C. Atti del Reale Istituto veneto di scienze, lettere ed arti. Nuovi contribute alla teoria delle relazioni statistiche. 1914-1915;74:1903-42.
14. Guttman L. The test-retest reliability of qualitative data. Psychometrika. 1946;11:81-95.[Crossref] [PubMed]
15. Goodman LA, Kruskal WH. Measures of association for cross classifications. J Am Stat Assoc. 1954;49(268):732-64.[Crossref]
16. Fisher RA. Statistical Methods for Research Workers.12th ed. Edinburgh: Oliver and Boyd; 1954.
17. Bennett EM, Alpert R, Goldstein AC. Communications through limited-response questioning. Public Opinion Quarterly. 1954;18:303-8.[Crossref]
18. Scott WA. Reliability of content analysis: the case of nominal scale coding. Public Opin Q. 1955;19(3):321-5.[Crossref]
19. Osgood CE. Representational model and relevant research methods. In: Pool I, ed. Trends in Content Analysis. Urbana: Illinois Press; 1959. p.33-88.
20. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37-46.[Crossref]
21. Holley JW, Guilford JP. A note on the G-index of agreement. Educ Psychol Meas. 1964;24(4):749-53.[Crossref]
22. Rogot E, Goldberg ID. A proposed index for measuring agreement in test-retest studies. J Chronic Dis. 1966;19(9):991-1006.[Crossref] [PubMed]
23. Holsti OR. Content Analysis for the Social Sciences and Humanities. Reading, MA: Addison-Wesley; 1969.
24. Krippendorff K. Bivariate agreement coefficients for reliability data. In: Borgatta ER, Bohrnstedt GW, eds. Sociological methodology. San Fransisco, CA: Jossey Bass; 1970. p.139-50.[Crossref]
25. Cicchetti DV. Assessing inter-rater reliability for rating scales: resolving some basic issues. Br J Psychiatry. 1976;129:452-6.[Crossref] [PubMed]

26. Maxwell AE. Coefficients of agreement between observers and their interpretation. Br J Psychiatry. 1977;130:79-83.[Crossref] [PubMed]
27. Janson S, Vegelius J. On generalizations of the G index and the phi coefficient to nominal scales. Multivariate Behav Res. 1979;14(2):255-69.[Crossref] [PubMed]
28. Brennan RL, Prediger DJ. Coefficient kappa: some uses, misuses, and alternatives. Educ Psychol Meas. 1981;41(3):687-99.[Crossref]
29. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. The Statistician. 1983;32(3):307-17.[Crossref]
30. Popping R. Traces of agreement: on the dot-product as a coefficient of agreement. Qual Quant. 1983;17:1-18.[Crossref]
31. Bangdiwala S. The agreement chart. University of North Carolina Institute of Statistics, Mimeo Series No. 1859, 1988.
32. Perreault WD, Leigh LE. Reliability of nominal data based on qualitative judgments. J Mark Res. 1989;26(2):135-48.[Crossref]
33. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989;45(1):255-68.[Crossref] [PubMed]
34. Kupper LL, Hafner KB. On assessing interrater agreement for multiple attribute responses. Biometrics. 1989;45(3):957-67.[Crossref] [PubMed]
35. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol. 1993;46(5):423-9.[Crossref] [PubMed]
36. Donner A, Eliasziw M. A hierarchical approach to inferences concerning interobserver agreement for multinomial data. Stat Med. 1997;16(10):1097-106.[Crossref] [PubMed]
37. Svensson E. A coefficient of agreement adjusted for bias in paired ordered categorical data. Biometrical Journal. 1997;39:643-57.[Crossref]
38. Andrés AM, Marzo PF. Delta: a new measure of agreement between two raters. Br J Math Stat Psychol. 2004;57(Pt 1):1-19.[Crossref] [PubMed]
39. von Eye A. An alternative to Cohen's κ. Eur Psychol. 2006;11(1):12-24.[Crossref]
40. Barnhart HX, Kosinski AS, Haber MJ. Assessing individual agreement. J Biopharm Stat. 2007;17(4):697-719.[Crossref] [PubMed]
41. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol. 2008;61(Pt 1):29-48.[Crossref] [PubMed]
42. Gwet KL. Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters. 3rd ed. Maryland: Advanced Analytics, LLC; 2012.
43. Warrens MJ. Weighted kappas for 3×3 tables. Journal of Probability and Statistics. 2013;1:1-9.[Crossref]
44. Kendall MG. Rank Correlation Methods. 3rd ed. London: Griffin; 1962.
45. Maxwell AE, Pilliner AEG. Deriving coefficients of reliability and agreement for ratings. Br J Math Stat Psychol. 1968;21(1):105-16.[Crossref] [PubMed]
46. Light RJ. Measures of response agreement for qualitative data: some generalizations and alternatives. Psychol Bull. 1971;76(5):365-77.[Crossref]
47. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971;76(5):378-82.[Crossref]
48. Tinsley HE, Weiss DJ. Interrater reliability and agreement of subjective judgments. Journal of Counseling Psychology. 1975;22(4):358-76.[Crossref]
49. Hubert L. Kappa revisited. Psychol Bull. 1977;84(2):289-97.[Crossref]
50. Landis JR, Koch GG. A one-way components of variance model for categorical data. Biometrics. 1977;33(4):671-9.[Crossref]
51. Fleiss JL, Cuzick J. The reliability of dichotomous judgments: unequal numbers of judges per subject. Appl Psychol Meas. 1979;3(4):537-42.[Crossref]
52. Conger AJ. Integration and generalization of kappas for multiple raters. Psychol Bull. 1980;88(2):322-8.[Crossref]
53. Kraemer HC. Extension of the kappa coefficient. Biometrics. 1980;36(2):207-16.[Crossref] [PubMed]
54. Schouten HJA. Measuring pairwise agreement among many observers. Biom J. 1980;22(6):497-504.[Crossref]
55. Craig RT. Generalization of Scott's index of intercoder agreement. Public Opin Q. 1981;45(2):260-4.[Crossref]
56. Davies M, Fleiss JL. Measuring agreement for multinomial data. Biometrics. 1982;38(4):1047-51.[Crossref]
57. O'Connell DL, Dobson AJ. General observer-agreement measures on individual subjects and groups of subjects. Biometrics. 1984;40(4):973-83.[Crossref]
58. Siegel S, Castellan NJ. Nonparametric Statistics for the Behavioural Sciences. 2nd ed. New York: McGraw-Hill; 1988. p.284-90.
59. Potter WJ, Levine-Donnerstein D. Rethinking validity and reliability in content analysis. J Appl Commun Res. 1999;27(3):258-84.[Crossref]
60. Randolph JJ. Free-marginal multirater Kappa (multirater K[free]): an alternative to Fleiss' fixed-marginal multirater kappa. Joensuu Learning and Instruction Symposium. Joensuu, Finland, Oct 14-15, 2005.
61. Mielke PW, Berry KJ, Johnston JE. Unweighted and weighted kappa as measures of agreement for multiple judges. Int J Manag. 2009;26(2):213-23.[Link]
62. Hughes J. Sklarsomega: an R package for measuring agreement using Sklar's Omega coefficient. 2018; arXiv:1809.10728v1.[Link]
63. van Oest R. A new coefficient of interrater agreement: The challenge of highly unequal category proportions. Psychol Methods. 2019;24(4):439-51.[Crossref] [PubMed]
64. Haber M, Barnhart HX. Coefficients of agreement for fixed observers. Stat Methods Med Res. 2006;15(3):255-71.[Crossref] [PubMed]
65. Barnhart HX, Haber MJ, Lin LI. An overview on assessing agreement with continuous measurements. J Biopharm Stat. 2007;17(4):529-69.[Crossref] [PubMed]
66. Örekici Temel G, Erdoğan S, Selvi H, Kaya Ersöz İ. Investigation of coefficient of individual agreement in terms of sample size, random and monotone missing ratio, and number of repeated measures. KUYEB Dergisi. 2016;16(4):1381-95.[Crossref]
67. Gwet KL. Handbook of Inter-rater Reliability. 4th ed. Gaithersburg: Advanced Analytics, LLC; 2014. p.3-25.
68. Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol. 2011;64(1):96-106.[Crossref] [PubMed]
69. Bishop YMM, Feinberg SE, Holland PW. Discrete Multivariate Analysis-Theory and Practice. Cambridge, Mass: The MIT Press; 1975.
70. Türk G. Gt index: a measure of the success of prediction. Remote Sens Environ. 1979;8:65-75.[Crossref]

71. Helldén U. A test of landsat-2 imagery and digital data for thematic mapping illustrated by an environmental study in northern Kenya, Lund University. 1980.[Link]

72. Short NM. The Landsat Tutorial Workbook-Basics of Satellite Remote Sensing. Vol. 1078. NASA reference publication. 1982. p.553.

73. Rosenfield GH, Fitzpatrick-Lins K. A coefficient of agreement as a measure of thematic classification accuracy. Photogramm Eng Rem S. 1986;52(2):223-7.[Link]

74. Barnhart HX, Yow E, Crowley AL, Daubert MA, Rabineau D, Bigelow R, et al. Choice of agreement indices for assessing and improving measurement reproducibility in a core laboratory setting. Stat Methods Med Res. 2016;25(6):2939-58.[Crossref] [PubMed]

75. Lin HM, Kim HY, Williamson JM, Lesser VM. Estimating agreement coefficients from sample survey data. Surv Methodol. 2012;38(1):63-72.[Link]