ضرایب توافق بین درجه بندی‌کنندگان

Interrater Agreement Coefficients

Frits E. Zegers*

University of Groningen

ترجمه و تلخیص دکتر حیدرعلی هومن

دانشگاه آزاد اسلامی

واحد تهران مرکز

Translated by Haidar Ali Hooman, PhD

Islamic Azad University

Center Tehran Branch

چکیده

کمتر اتفاق می‌افتد که داوران مختلف در داوریهای خود حتی نسبت به یک امر واحد مجموعه یکسانی به دست دهند. اگر مجموعه نمره‌های دو داور دقیقاً برابر با یکدیگر باشد، بیشینه مقدار ضرایب همخوانی بین نمره‌ها برابر با ۱+ خواهد بود. اما به عکس، ضریب همخوانی ۱+ دلالت بر این ندارد که دو مجموعه داده‌های مورد مطالعه دقیقاً برابر با یکدیگر است. داوران مختلف در واقع تنها درباره وضع نسبی افراد، چیزها یا پدیده‌ها داوری می‌کنند، نه وضع مطلق آنها. مقصود از این مقاله معرفی ضرایبی است که میزان توافق مطلق و همانندی واقعی بین نمره‌های داوران مختلف را به نمایش می‌گذارد. برخی از ویژگیهای همبستگی گشتاوری ممکن است برای یک موقعیت معین نامناسب باشد. برای این موقعیتها نوعی ضرایب همخوانی ارائه شده است که بسیاری از آنها متعلق به طبقه ضرایب اقلیدسی است. بحث درباره ویژگیهای مطلوب این گونه ضرایب نشان می‌دهد که چگونه ضرایب همانندی و تعمیم‌های مربوط به آن می‌توانند برای ارزیابی توافق درجه‌بندی کنندگان به کار روند و افزون بر این مشخص می‌کند که کدام یک از اطلاعات مربوط به داده‌ها باید از طریق ضرایب مختلف تبیین و از کدام یک از آنها باید صرف‌نظر شود.

واژه‌های کلیدی : ضرایب همخوانی، همبستگی، ضرایب اقلیدسی، ضریب همانندی تعمیم‌یافته، توافق بین درجه‌بندی‌کنندگان.

Abstract

The degree of agreement between two or more judges who rate a number of objects on certain characteristics can be expressed by means of an association coefficient (e. g., product moment correlation, PMC). The agreement among judges can rarely be perfect. If two judges produce two identical sets of scores the PMS will attain its maximum value, which equals +1. However, a PMC that equals +1 does not mean that the two sets of data are exactly equal. Different judges actually make judgments only based on the relative states of individuals, objects, or phenomena, and not based on their absolute states. This article aims to introduce association coefficients that are indicative of the degree to which there is an absolute agreement and real similarity among the ratings by different judges. Some properties of the PMC in specific situations may be undesirable. Many association coefficients have been purposed for those situations, many of which belong to the class of Euclidian coefficients. A discussion of the desirable properties of these coefficients demonstrates that how the identity coefficient and its generalizations can be used to assess the interrater agreements.

Key words : association coefficients, correlation, Euclidian coefficients, generalized identity coefficient, interrater agreement.

Contact information e-mail : haidarali70@yahoo.com

** Applied psychological Measurement 15 (4), December, 1991, pp. 321-328*

نکات کلی

اگر اعداد یا برچسبهایی که به مقوله‌های مورد مطالعه منتسب می‌شوند بدون ترتیب خاصی باشند، داده‌ها اسمی و در غیر این صورت داده‌ها غیراسمی خواهند بود. انتخاب ضریب مناسب برای ارزشیابی میزان توافق بستگی به این دارد که داده‌ها اسمی یا غیراسمی باشند.

بسیار کم اتفاق می‌افتد که دو داور در داوری خود نسبت به یک امر واحد ارزشیابی یکسانی به عمل آورند. به گونه کلی اگر مجموعه نمره‌های دو داور دقیقاً برابر با یکدیگر باشد، بیشینه مقدار ضرایب همخوانی برابر با ۱+ خواهد بود. اما به عکس، ضریب همخوانی ۱+ دلالت بر این ندارد که دو مجموعه داده‌های مورد مطالعه دقیقاً برابر با یکدیگر است. برای مثال، چنانچه در استفاده از یک مقیاس ۱۰ درجه‌ای برای ارزشیابی سه مقاله، یک معلم نمره‌های ۹ و ۸ و ۷ و معلم دیگری نمره‌های ۴ و ۳ و ۲ بدهد، ضریب همبستگی بین این نمره‌ها (با یک انتقال خطی مثبت) دقیقاً برابر با ۱+ خواهد بود، در حالی که بین این دو معلم توافق کامل وجود ندارد. در واقع این دو معلم تنها درباره وضع نسبی، و نه وضع مطلق سه مقاله توافق کامل داشته‌اند. ضریب همانندی[1] که توسط زیگرز و تن برگ (۱۹۸۵) معرفی شده یک ضریب توافقی برای مقیاسهای مطلق است که مقدار آن وقتی برابر با ۱+ است که اگر و تنها اگر دو مجموعه نمره‌ها دقیقاً یکسان باشد.

این ضریب که با e_{xy} نمایش داده می‌شود چنین است :

$$e_{xy} = ۱ - \frac{\sum_{i=۱}^{n} (X_i - Y_i)^۲}{\sum_{i=۱}^{n} X_i^۲ + \sum_{i=۱}^{n} Y_i^۲}$$

که در آن n تعداد چیزها یا افرادی است که داوری شده، و X_i و Y_i معرف نمره‌های دو داور به فرد i است. فرمول معادل آن چنین است :

$$e_{xy} = \frac{۲\sum_{i=۱}^{n} X_i Y_i}{\sum_{i=۱}^{n} X_i^۲ + \sum_{i=۱}^{n} Y_i^۲}$$

اینک به نظر می‌رسد مسئله‌ای که در بالا بدان اشاره شد (یعنی داوری دو معلم) حل شده باشد. زیرا به رغم آنکه همبستگی گشتاوری آن برابر با ۱+ است، ضریب همانندی آن برابر با ۱+ نیست، و بواقع کمتر از $e_{xy} = ۰/۶۶$ است که نشان می‌دهد دو معلم توافق کامل ندارند. اما اکنون فرض کنید درجه‌بندیهای مذکور بدین دلیل انجام گرفته است که جوائز یکم، دوم و سوم به سه شاگرد داده شود. در این حالت معنای معمولی نمره‌ها نامناسب و نامربوط خواهد بود، و فقط ترتیب نمره‌های داوران دارای اهمیت است. ضریب همانندی که در اینجا به دست آمده است ($e_{xy} = ۰/۶۶$)، توافق بین دو معلم را از لحاظ توزیع جوائز به شاگردان به غلط عرضه می‌دارد. در این مورد مسلماً ضریب همبستگی اسپیرمن روش بهتری برای ارزشیابی است.

نکته مهم آن است که در برخی از موارد روابط بین نمره‌های یک داور به لحاظ تجربی مهم، مفید و با معنا، و در برخی از موارد این روابط نامناسب و نامربوط‌اند. ضریب توافق باید میزان توافق بین داوران را برحسب روابط بین نمره‌های (از لحاظ تجربی) بامعنا و مفید بیان کند و میزان توافق روابط نامربوط و نامناسب را نادیده انگارد. به روابطی که از لحاظ تجربی بامعنا و مفید باشند، اطلاعات بامعنا[2] و به روابطی که نامربوط باشند اطلاعات نامناسب[3] گفته می‌شود.

سطح اندازه‌گیری داده‌ها مشخص می‌کند که کدام یک از روابط بین داده‌ها (از لحاظ تجربی) بامعنا و کدام یک نامناسب است (البته از لحاظ ریاضی نمی‌توان این مطلب را ثابت کرد، اما از لحاظ عملی می‌توان گفت کدام یک مناسب و کدام نامناسب است). به عنوان مثال برای توزیع جوائز در مدرسه، آن گونه که در بالا اشاره شد، اطلاعات ترتیبی مفید و بامعناست. اما اگر نمره‌های مدرسه به منظور قبولی و مردودی (P[4] و F[5]) به کار رود،

1. identity coefficient
2. meaningful
3. irrelevant
4. pass
5. fail

اطلاعات مفید آن است که تعیین کنیم یک نمره به خصوص بالاتر یا پایین‌تر از ملاک قبولی / مردودی (حد نصاب) قرار دارد.

به این پرسش که روش داوری تا چه حد از لحاظ به دست دادن اطلاعات بامعنا اعتبار دارد می‌توان ازطریق محاسبه ضریب توافق مناسب (ضریبی که اطلاعات معنادار را به حساب آورد و اطلاعات نامناسب را نادیده انگارد) پاسخ داد. مفهوم اطلاعات معنادار را می‌توان در موقعیتهایی به کار برد که در آنها لازم است یک ضریب توافق با بیشینه مقدار برابر با ۱/۰ به دست آید، و ضریب توافق فقط هنگامی باید برابر با ۱/۰ باشد که مجموعه نمره‌های دو داور (از لحاظ اطلاعات بامعنا) دقیقاً یکسان باشد. یک راه برای پیداکردن ضریبی که این شرط را تأمین کند آن است که نمره‌های هر داور به مقدار معنادار تبدیل و سپس ضریب همانندی بین این مقادیر حساب شود. با این روش، مسئله انتخاب ضریب نامناسب توافق جای خود را به مسئله تعیین مقادیر معنادار نمره‌های داوران می‌دهد.

روش تبدیل نمره‌ها به مقادیر معنادار ضمن آنکه اطلاعات نامربوط و نامناسب را از میان می‌برد، به منظور فراهم ساختن نوعی از مقادیر استاندارد برای نمره‌ها باید اطلاعات بامعنا و مفید را حفظ کند. یک مثال ساده برای این گونه تبدیلها همان استانداردکردن متغیرهاست که در آن مقیاس، نمره‌ها با میانگین صفر و واریانس ۱/۰ بیان می‌شوند. استانداردکردن، اطلاعات فاصله‌ای را حفظ می‌کند، اما اطلاعات موجود درباره میانگین و مقیاس‌ـ پردازی را از میان می‌برد.

به گونه کلی برای تعیین مقدار معنادار باید به سه پرسش جواب داده شود.

۱- آیا نمره‌ها باید تبدیل به رتبه شود؟ اگر اطلاعات معنادار فقط شامل اطلاعات ترتیبی باشد، در این صورت نمره‌ها باید تبدیل به رتبه شود.

۲- آیا نمره مرجع باید از نمره‌ها کم شود؟ نمره‌ها از طریق کم‌کردن مقدار یک نقطه مرجع به صورت نمره‌های تفاوت نسبت به این نقطه مرجع بیان می‌شود. نقطه مرجع ممکن است مطلق یا نسبی باشد. نقطه مرجع مطلق مستقل از نمونه[1] است که نشان می‌دهد بستگی به نمره‌های مشاهده شده ندارد. یک نقطه مرجع مطلق را می‌توان به عنوان نقطه صفر طبیعی یا نقطه خنثای یک مقیاس داوری (مثلاً نقطه قبول / رد در مقیاس نمره‌های کلاسی) در نظر گرفت. اما تعیین نقطه مرجع نسبی به نمره‌های مشاهده شده (برای مثال بـرحسب نمونه، میانگین یا اندازه‌های دیگری از مقادیر مرکزی تعیین‌کنندهٔ نقطه مرجع خواهند بود) بستگی دارد. در اینجا میانگین نمونه را مورد بررسی قرار می‌دهیم.

۳- آیا تبدیل مقیاس ضرورت دارد؟ اگر مقیاس‌پردازی نمره‌ها اختیاری باشد، نمره‌ها به مقیاسی دیگر به گونه‌ای تبدیل می‌شود که نمره میانگین مربعات آن برابر با ۱/۰ شود. این عمل از طریق تقسیم نمره‌ها بر ریشه دوم نمره میانگین مربعات انجام می‌شود. معادله‌ای که در بالا داده شد نشان می‌دهد که ضریب همانندی تحت تأثیر ضرب یا تقسیم X و Y بر یک عدد ثابت نیست. یعنی اگر هر دو مجموعه بر یک عدد ثابت ضرب یا تقسیم شود مقدار ضریب تغییر نمی‌کند.

بنابر آنچه که گفته شد، فرآیند تبدیل به مقیاس هنگامی در ضریب همانندی حاصل اثر می‌گذارد که اگر و تنها اگر نمره‌های X و Y بعد از انجام مرحله ۲ بالا دارای میانگین مربعات متفاوتی باشند. وجود چنین تفاوتی ممکن است نتیجه اعمال روشهای مختلف توسط دو داور و یا نتیجه تمایل یک داور به کاربرد نقاط افراطی‌تر در مقیاس جدید نسبت به داور دیگر باشد. اگر نمره‌های هر دو داور (از لحاظ سازمان) دارای اهمیت یکسان باشـد، باید تبدیل به یک مقیاس شود تا تفاوتهای نامربوط از میان برود (مگر آنکه بخواهیم برای یکی اهمیت بیشتری قائل شویم).

نکته مهم آن است که به منظور تعبیر و تفسیر ضرایب توافق بتوانیم مقدار ضریب همانندی را در شرایط شانس و تصادف (یعنی فرضیه صفر) تعیین کنیم. یک فرمول ساده

1. sample - independent

که توسط زیگرز (ب ۱۹۸۶) پیشنهاد شده چنین است :

$$g'_{xy}=\frac{۲n^{-۱}\sum_{i=۱}^{n}X_i\sum_{i=۱}^{n}Y_i}{\sum_{i=۱}^{n}X_i^۲+\sum_{i=۱}^{n}Y_i^۲}$$

مقدار ضریب همانندی را می‌توان با مقدار معادله بالا مقایسه و بررسی کرد که آیا مقدار مورد انتظار ضرایب مبتنی بر نمره‌هائی که نقطه مرجع نسبی آنها برابر با میانگین است، تحت شرایط شانس، صفر خواهد بود یا نه؛ تحت این شرایط، صورت کسر بالا که مجموع نمره‌های تفاوتها را در بر دارد برابر با صفر می‌شود. مقایسه مقدار ضریب همخوانی g_{xy} با مقدار مورد انتظار در شرایط شانس g'_{xy} مبتنی بر تفاوت $g_{xy}-g'_{xy}$ است. این تفاوت را می‌توان برای تعیین مقدار تصحیح شده یک ضریب معین به کار برد.

روش متعارف برای تصحیح یک ضریب همخوانی (g_{xy}) آن است که تفاوت آن با ضریب حاصل در شرایط شانس (یعنی g'_{xy}) را با بیشینه مقدار نظری این تفاوت (یعنی با مقدار $۱-g'_{xy}$) مربوط سازیم. این مطلب ضریب تصحیح شده به خاطر شانس را بر پایه فرمول زیر به دست می‌دهد (زیگرز، ۱۹۸۶ الف).

$$g''_{xy}=\frac{g_{xy}-g'_{xy}}{۱-g'_{xy}}$$

اینک فرض کنید دو معلم با به کار بردن یک مقیاس ۱۰ امتیازی درباره کیفیت ۴ مقاله داوری کرده‌اند و نتایج جدول ۱ (با نقطه مرجع ۵/۵، یعنی مرکز مقیاس) به دست آمده است (این نمره‌ها نشان‌دهنده موقعیتی است که در آن تفاوت بین ضرایب تصحیح شده و تصحیح نشده زیاد است).

جدول ۱ : نتایج درجه‌بندی ۴ مقاله توسط دو معلم قبل و بعد از کاهش نقطه مرجع

مقاله	قبل از کاهش		بعد از کاهش	
	X	Y	X	Y
۱	۸	۸	۲/۵	۲/۵
۲	۸	۹	۲/۵	۳/۵
۳	۹	۸	۳/۵	۲/۵
۴	۹	۹	۳/۵	۳/۵

به این پرسش که بین نمره‌های این دو معلم تا چه حد توافق وجود دارد به دو روش می‌توان پاسخ داد. نخست آنکه بین آنها توافق زیادی وجود دارد، زیرا هر دو درباره کیفیت مقاله‌ها با امتیازهای ۹ و ۸ یعنی «خیلی خوب» و «خوب» داوری کرده‌اند. میزان این توافق برحسب ضریب همانندی بیان می‌شود که مقدار آن برابر است با :

$$exy=۱-\frac{\sum(X-Y)^۲}{\sum X^۲+\sum Y^۲}=۱-\frac{۲}{۲۹۰+۲۹۰}=$$

$$\frac{۵۷۸}{۵۸۰}=۰/۹۹۷$$

و با استفاده از فرمول دوم نیز داریم :

$$exy=\frac{۲\sum XY}{\sum X^۲+\sum Y^۲}=\frac{۲\times۲۸۹}{۲۹۰+۲۹۰}=۰/۹۹۷$$

$$\frac{۵۷۸}{۵۸۰}=۰/۹۹۷$$

بعد از کم کردن نقطه مرجع مطلق (یعنی عدد ۵/۵) از هر نمره، مقدار این ضریب هنوز بسیار قوی و برابر با $e_{xy}=۰/۹۷۳$ خواهد بود. اما به عکس این دو معلم عدم توافق نیز نشان داده‌اند : زیرا نمره نسبتاً پایین معلم X (نمره ۸) هم با نمره نسبتاً پایین معلم Y (نمره ۸) و هم با نمره نسبتاً بالای او (نمره ۹)، و نمره نسبتاً بالای معلم X نیز هم با نمره نسبتاً پایین و هم با نمره نسبتاً بالای معلم(نمره ۷)، همراه بوده است. این عدم توافق را می‌توان با ضریب همانندی تصحیح شده برای شانس که مقدار آن $e'_{xy}=۰$ است، به خوبی نشان داد :

$$e'_{xy}=\frac{۲n^{-۱}\sum X\sum Y}{\sum X^۲\sum Y^۲}=\frac{۲\times۰/۵\times۳۵\times۳۴}{۲۹۰+۲۹۰}=۰/۹۹۷$$

که اگر آن را از e_{xy} کم کنیم حاصل صفر می‌شود (یعنی مقدار آن با e_{xy} دقیقاً یکی است). به گونه کلی، ارزشیابی میزان توافق دو داور را در نمره‌های غیراسمی می‌توان به شرح زیر خلاصه کرد.

۱- نمره‌های داوران به ارقام بامعنا تبدیل می‌شود. این امر ممکن است مستلزم رتبه‌بندی، کم کردن یک نمره مرجع مطلق یا نسبی و تبدیل مقیاس باشد.

۲- ضریب همانندی بین مقادیر با معنا محاسبه می‌شود.

مقادیر به دست آمده را می‌توان از طریق یک مدل صفر نسبی (برحسب نمونه) یا مدل مطلق (مستقل از نمونه) با مقدار مورد انتظار مقایسه کرد.

ضریب گاور[1]

فرض کنید دو داور X و Y چهار چیز را در یک مقیاس ۵ امتیازی (با دامنه ۱ تا ۵) با نمره مرجع ۳ (مرکز مقیاس) درجه‌بندی کرده‌اند. دو مجموعه نمره‌های بامعنا در جدول ۲ نشان داده شده که در آنها مقادیر ضریب همانندی برابر با ۰/۶۷ و ۰/۵ است.

جدول ۲ : نمره‌های بامعنای مربوط به دو داور

چیز	مجموعه ۱		مجموعه ۲	
	X	Y	X	Y
۱	۲	۱	۲	۱
۲	۱	۲	۰	۱
۳	۰	۱	-۱	۰
۴	۰	۱	-۱	۰

این ضرایب، پس از اعمال تصحیح به خاطر شانس (برپایه فرمول‌های بالا) برابر با ۰/۶۳ و ۰/۵ خواهد بود. بدون تردید، میزان توافق برای این دو مجموعه داده‌ها با هم متفاوت است. اما می‌توان گفت که توافق بین این دو مجموعه نیز یکسان است. زیرا تفاوتهای بین داوران برای هر چیز منفرد دقیقاً یکی است. صورت کسر ضریب همانندی (در معادله یکم) شامل مجموع مجذور تفاوتهاست و این مجموع برای هر دو مجموع داده‌ها یکسان است. اما مخرج آنها با هم برابر نیست، زیرا در مخرج کسر مجموع کلی مجذورات قرار دارد، و بنابراین مقادیر ضریب همانندی برای هر دو یکی نخواهد بود. ضریبی که برای داده‌های جدول ۲ مقادیر یکسان به دست می‌دهد توسط گاور (۱۹۹۶، ۱۹۷۱) برپایه فرمول زیر پیشنهاد شده است.

$$G_{xy} = ۱ - \frac{\sum_{i=۱}^{n} |X_i - Y_i|}{nR}$$

در این فرمول X_i و Y_i معرف نمره‌های دو داور، n تعداد چیزها، و R دامنه مقیاس درجه‌بندی یعنی بیشینه قدر مطلق تفاوت $X_i - Y_i$ (برای جدول بالا R = ۴) است. برای هر دو مجموعه داده‌های جدول ۲ داریم : $G_{xy} = ۰/۷۵$. روشن است که در محاسبه G_{xy}، مجموع قدر مطلق تفاوتها، و نه مجموع مجذور تفاوتها، دخالت دارد. برای به دست آوردن یک ضریب نرم‌گیری شده، این مجموع بر بیشینه مقدار نظری آن، یعنی nR تقسیم می‌شود. این بیشینه مقدار هنگامی به دست می‌آید که داوران برای داوری درباره هر فرد، این بیشینه اختلاف را داشته باشند. در این شرایط داوران در مقیاس درجه‌بندی، نمره‌های افراطی مخالف را درباره هر فرد به کار می‌برند و در نتیجه مقدار G_{xy} برابر با صفر خواهد شد. آشکار است که دامنه تغییرات G_{xy} می‌تواند مقادیری را در فاصله (۰ و ۱) اختیار کند. تفاوت اساسی بین ضریب گاور و ضریب همانندی در روش نرم‌گیری آن است. ضریب همانندی با استفاده از یک روش نسبی یا برحسب نمونه، با به کاربردن مجذور ارقام با معنا متوسط‌گیری می‌شود، در حالی که ضریب گاور با به کاربردن دامنه مقیاس درجه‌بندی به صورت مطلق یا متوسط از نمونه متوسط‌گیری می‌شود.

منابع

Gawer, J. C. (1966). Some distance properties of latent root and vector methods used multivariate analysis. *Biometrika, 53*, 315-328.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, *27*, 857-871.

Zegers, F. E. (1986 a). A family of chance-corrected association coefficients for metric scales. *Psychometrika, 51*, 559-569.

Zegers, F. E. (1986 b). A general family of association coefficients. Groningen Netherlands : Boomker.

Zegers, F. E., & Ten Berge, J. M. F. (1985). A family of association for metric scales. *Psychometrika, 50*, 17-24.

1. Gower's coefficient