دانشکده علوم پایه

گروه آمار

عنوان پایان نامه

آزمون فرض استقلال در جداول رسته‌ای دو طرفه با استفاده از تشکیل ناحیه باورپذیر همزمان بیزی برای اثرات متقابل

پایان نامه:

برای دریافت درجه کارشناسی ارشد

در رشته آمار ریاضی

مولف

مرتضی علیاری

استاد راهنما

دکتر پرویز نصیری

ماه و سال انتشار

شهریور ۱۳۸۷

دانشگاه پیام نور

دانشکده علوم پایه

گروه آمار

عنوان پایان نامه

آزمون فرض استقلال در جداول رسته‌ای دو طرفه با استفاده از تشکیل ناحیه باورپذیر همزمان بیزی برای اثرات متقابل

پایان نامه:

برای دریافت درجه کارشناسی ارشد

در رشته آمار ریاضی

مولف

مرتضی علیاری

استاد راهنما

دکتر پرویز نصیری

استاد مشاور

دکتر علی شادرخ

ماه و سال انتشار

شهریور ۱۳۸۷

بسم الله

الرحمن الرحيم

از زحمات بی‌دریغ استاد گرانقدرم، جناب آقای دکتر پرویز نصیری

که در تمام لحظات سخت تحقیق مرا همراهی کردند، صمیمانه

سپاسگزارم.

کمک‌ها و دلداری‌های برادر مهربانم و همچنین دوست خوبم که

لحظه لحظه مرا دلگرم می کردند نیز، قابل قدردانی است.

تقدیم به مادر مهربان

و

پدر دلسوزم

فهرست مطالب

فهرست مطالب

فهرست جداول و شکل‌ها

فهرست جداول

فهرست شکل‌ها

چکیده

آزمون کی‌دو برای بررسی فرض استقلال در جداول متقاطع دوطرفه، وابسته به این فرض است که مشاهدات خانه‌های جدول دارای توزیع چندجمله‌ای باشند. بر این اساس، آزمون‌های دیگری، برای مواقعی که فرض توزیع چندجمله‌ای برقرار نیست، پیشنهاد می‌شود. در این راستا، ابتدا عامل بیز- که جهت آزمون‌های فرض در آمار بیزی مورد استفاده قرار می گیرد- به عنوان جایگزین آزمون کی‌دو در نظر گرفته می‌شود. متاسفانه مشکلی که در استفاده از عامل بیز وجود دارد، حساس بودن مقدار آن به انتخاب توزیع‌های پیشین متفاوت است.

بنابراین، استفاده از یک برآورد بیزی که در حالت کلی مثل عامل بیز به تعیین توزیع پیشین حساس نیست، پیشنهاد می‌شود. رویکرد ما ساختن یک ناحیه باورپذیر همزمان ۹۵ درصد (یک ابَر مستطیل) برای اثرات متقابل است. آزمون صفر بودن همه اثرات متقابل هم‌ارز آزمون استقلال در جداول رسته‌ای دوطرفه است. بنابراین، یک ناحیه باورپذیر همزمان ۹۵ درصد از اثرات متقابل در واقع یک آزمون استقلال از نوع برگردان است.

کلید واژه‌ها: عامل بیز، آماره‌ی کی‌دو، اثرات متقابل، همبستگی درون رده‌ای، توزیع پیشین، استنباط همزمان

فصل اول

مقدمه

فصل اول– مقدمه

بررسی فرض استقلال در جداول توافقی دوطرفه با استفاده از آزمون کی‌دو وابسته به فرض چند جمله‌ای بودن توزیع فراوانی مشاهدات خانه‌های جدول است. اما گاهی اوقات، به ویژه در مواقعی که مشاهدات خانه‌های جدول از یک خوشه جمع‌آوری می‌شوند (مثل نظرسنجی از اعضای یک خانواده)، به دلیل اثرپذیری اعضای یک خانواده از هم، میزانی از همبستگی بین مشاهدات به وجود خواهد آمد که به همبستگی درون‌رده‌ای مشهور است. وجود همبستگی درون‌رده‌ای باعث می‌شود که فرض چندجمله‌ای بودن توزیع فراوانی مشاهدات دیگر برقرار نباشد. در واقع در این حالت، مشاهدات داخل هر خانه از هم مستقل نخواهند بود. از این رو هنگامی که فراوانی مشاهدات خانه‌ها از توزیع چند جمله‌ای پیروی نمی‌کنند یا تعداد مشاهدات بعضی از خانه‌ها کوچک یا صفر هستند، طبیعی است که درباره رویکردهای دیگری اندیشیده شود. این پایان‌نامه ضمن این‌که مروری بر نواقص روش‌های بنیان‌گذاری شده بر اساس آماره‌ی کی‌دو تعدیل شده و عامل بیز دارد روش ساده‌ای را نیز جهت آزمون استقلال در جدول‌های رسته‌ای دوطرفه ارائه می‌کند که بیش از این‌که به آزمون‌های فرض متّکی باشد بر اساس برآورد یک ناحیه‌ی باورپذیر بیزی، استوار است.

در واقع رویکرد ما در این پایان نامه جهت انجام دادن آزمون فرض استقلال، ساختن یک ناحیه‌ی باورپذیر همزمان بیزی (یک ابر مستطیل[1]) برای اثرات متقابل است. آزمون صفر بودن همه اثرات متقابل، هم‌ارز آزمون استقلال در جداول رسته‌ای دوطرفه است. بنابراین، یک **ناحیه‌ی باورپذیر همزمان ۹۵ درصد**[2] از اثرات متقابل در واقع یک آزمون استقلال برگردان[3] است.

---

[1] Hyper rectangle

[2] 95% Simultaneous Credible Region

[3] A test of independence by inversion

مطالب پایان‌نامه در قالب فصل‌های زیر ارائه شده‌اند:

۱. مقدمه

۲. آزمون کی‌دو برای استقلال (آزمون کی‌دو پیرسون، آزمون کی‌دو تعدیل شده)

۳. عامل بیز

۴. تشکیل ناحیه‌ی باورپذیر همزمان بیزی برای اثرات متقابل

۵. مثال‌ها و شبیه‌سازی

۶. بحث و نتیجه‌گیری

۱-۱ هدف از انجام پایان نامه

هدف، هشدار به افرادی است که برای انجام دادن آزمون استقلال سطر از ستون در جداول رسته‌ای، طبق روال معمول از آماره‌ی کی‌دو پیرسون یا عامل بیز استفاده می‌کنند. علاوه براین، ارائه‌ی راهکارهایی برای رویارویی با مشکلات آماره‌ی کی‌دو پیرسون یا عامل بیز، هدف دیگری است که این پایان نامه دنبال می‌کند.

از این رو دلایل تدوین پایان نامه عبارتند از:

الف) مرور بعضی از نواقص آزمون کی‌دوی استاندارد

ب) نشان دادن حساسیت عامل بیز، به عنوان روش دیگری در مقابل آزمون کی‌دو استاندارد، در برابر توزیع‌های متفاوتی که به عنوان توزیع پیشین انتخاب می شوند.

پ) بنا نهادن یک ناحیه‌ی باورپذیر همزمان برای اثرات متقابل در یک جدول $r \times c$

۱-۲ پیشینه‌ی تحقیق

محققان بسیاری، عدم صحت آزمون کی‌دو برای داده‌های همبسته را تصدیق کرده‌اند. تلاش‌های انجام‌یافته برای تصحیح تورم ساختگی در چنین آماره‌ی آزمونی به طور اساسی بر دو رویکرد استوار است. رویکرد اول، رویکرد طرح پایه[4] نامیده می‌شود. این رویکرد، با تکرار طرح نمونه، استنباطی را با توجه به چند جمله‌ای بودن توزیع مجانبی برآوردها، فراهم می‌آورد (فلگی[۱]، هلت و همکاران[۲]، رائو

[4] Design Based

و اسکات [۳] و [٤]، بدریک و فای [٥]). این محققان از اثر طرح، که نسبت واریانس برآوردگرهای حاصل از دو طرح است، استفاده کردند. یکی از این برآوردگرها از طرحی پیچیده‌تر از طرح نمونه‌گیری تصادفی ساده و دیگری از نمونه‌گیری تصادفی ساده به دست می‌آیند. به عنوان مثال رائو و اسکات [۳] و [٤]، جهت آزمون نیکویی برازش و استقلال در جداول رسته‌ای چندطرفه، اثر طبقه‌بندی و خوشه‌بندی را بر توزیع مجانبی آماره کی‌دو پیرسون، ارزیابی کردند. آنها اثر طرح‌های تعمیم‌یافته را که برای تعدیل آماره کی‌دو استفاده می‌شد، پیشنهاد کردند.

رویکرد دوم، رویکردی است که بر مدل پایه[5] استوار است. این رویکرد، برای مدل دادن به داده‌های نمونه یک توزیع احتمالی فرضی را در نظر می‌گیرد (آلتام [٦]، کوهن [٧]، بری‌یر [٨]، فین برگ [٩] و چوی و ام‌سی‌هوگ [١٠]). در این راستا کوهن [٧] چگونگی تعدیل آماره‌ی کی‌دو پیرسون به هنگام وجود همبستگی درون رده‌ای در خانوارهای دوعضوی را نشان داد و آلتام [٦] نتایج کوهن [٧] را به خانوارهای K عضوی تعمیم داد. پس از آلتام [٦] و کوهن [٧] ، بری‌یر [٨] نتایج این دو را به خانوارهایی با تعداد اعضای نامساوی تعمیم داد.

همچنین در سال‌های بعد، چوی و ام‌سی‌هوگ [١٠]، با اعمال یک توسعه‌ی احتمالاتی به نتایج به دست آمده توسط آلتام [٦] و کوهن [٧]، چگونگی تعدیل آماره‌ی آزمون کی‌دو را به هنگام وجود همبستگی درون‌رده‌ای نشان دادند.

سرانجامناندرام و چوی [١١] پس از چند سال با ارائه‌ی یک روش قابل اعتماد برای انجام آزمون استقلال در جداول رسته ای دو طرفه، با استفاده از روش برآوردی بیزی، توانستند مشکلات مربوط به روشهای پیشین را تا حدود زیادی مرتفع نمایند.

---

[5] Model Based

فصل دوم

آزمون کی‌دو برای فرض استقلال

فصل دوم– آزمون کی‌دو برای فرض استقلال

در استنباط آماری، دو رویکرد برای انجام دادن آزمون ارتباط[6] (استقلال) در یک جدول رسته‌ای $r \times c$ وجود دارد. رویکرد اول استفاده از روش شناخته شده‌ی آماره‌ی کی‌دو پیرسون است. البته باید توجه داشت که این آزمون بر اساس نمونه‌گیری تصادفی ساده است (این همان نمونه‌گیری چند جمله‌ای است). این رویکرد در این فصل بررسی خواهد شد.

رویکرد دوم استفاده از عامل بیز است که به عنوان یک روش در مقابل آزمون کی‌دو مطرح می‌شود و به آن در فصل سوم پرداخته خواهد شد.

در این فصل، ابتدا مفاهیمی که در کل پایان‌نامه به صورت دائم تکرار خواهند شد، آورده می‌شوند، سپس به بحث در مورد آماره‌ی کی‌دو کاهش یافته پرداخته خواهد شد.

۲-۱ تعاریف

۲-۱-۱ توزیع چند جمله‌ای

این توزیع، تعمیم توزیع دوجمله‌ای است. فرض کنید که در اجرای یک آزمایش برآمدهای ناسازگار $A_1,\ldots,A_k$ رخ دهند به گونه‌ای که

$$P\left(A_i\right) = p_i, i = 1,\ldots,k$$

و

[6] Association

$$\sum_{i=1}^{k} p_i = 1 \, , \, p_i \geq 0$$

حال اگر این آزمایش را به تعداد n بار تکرار کنیم و از این n بار x_1 بار برآمد A_1، x_2 بار برآمد A_2 و در نهایت x_k بار برآمد A_k رخ دهد، آنگاه تابع احتمال توأم $X_1,\dots,X_k$ به صورت زیر خواهد بود

$$P\left(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k\right) = pr\left(x_1, \dots, x_k \mid n, p_1, \dots, p_k\right)$$
$$= \frac{n!}{x_1!x_2!\dots x_k} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k} \, ,$$

که در آن عبارت ترکیبی $\frac{n!}{x_1!x_2!\dots x_k}$، یک ثابت نرمال‌سازی و $\sum_{i=1}^{k} x_i = n$ است.

توزیع چند جمله‌ای اغلب در علوم اجتماعی برای مدل دادن به متغیرهایی به کار می‌رود که دارای برآمدهای مختلفی از نظر کیفی هستند. اعتقادات مذهبی، گرایشات سیاسی، نژاد و غیره نمونه‌ای از این متغیرها هستند.

۲-۱-۲ انواع مدل‌های آماری برای داده‌های یک جدول توافقی

مدل آماری برای داده‌های جداول توافقی بستگی به روش جمع‌آوری داده‌ها دارد. لکن در بیشتر موارد می‌توان داده‌ها را در قالب سه مدل آماری مختلف که در ارتباط نزدیک با یکدیگر قرار دارند بررسی کرد. این سه مدل عبارتند از:

۲-۱-۲-۱ نمونه‌گیری پواسن

در این نوع نمونه‌گیری نمی‌توان کنترلی در حجم نمونه یا تعداد مشاهدات داشت. همچنین نمی‌توان هیچ‌گونه اطلاعاتی در خصوص توزیع‌های حاشیه‌ای به‌دست آورد. در این حالت تابع احتمال توأم مشاهدات عبارتست از:

$$f\left(x_{11}, x_{12}, \dots, x_{IJ}; \lambda_{11}, \lambda_{12}, \dots, \lambda_{IJ}\right) = \prod_{i=1}^{I}\prod_{j=1}^{J} e^{-\lambda_{ij}} \frac{\lambda_{ij}^{x_{ij}}}{x_{ij}!}$$

۲-۱-۲-۲ نمونه‌گیری چند جمله‌ای

در این نمونه‌گیری، نمونه‌ها به تعداد از پیش تعیین شده‌ای جمع‌آوری می‌شوند و سپس در دسته‌های مقتضیِ مربوط به دو متغیر رسته‌ای کلاس‌بندی می‌شوند. چون در اینجا حجم نمونه (n)، از قبل تعیین شده است لذا منطقی است مقادیر X_{ij}، به عنوان یک نمونه‌ی تصادفی از توزیع چند جمله‌ای با

پارامترهای p_{ij} در نظر گرفته شوند. در چنین حالتی مدل کلی به صورت چند جمله‌ای با تابع احتمال زیر است:

$$f\left(x_{11},x_{12},\ldots,x_{IJ};p_{11},p_{12},\ldots,p_{IJ}\right)=\binom{n}{x_{11},x_{12},\ldots,x_{IJ}}\prod_{i=1}^{I}\prod_{j=1}^{J}p_{ij}^{x_{ij}}$$

۲-۱-۲-۳ نمونه‌گیری چند جمله‌ای ترکیبی

در این نمونه‌گیری، داده‌ها به تعداد از پیش تعیین شده‌ای از افرادی که متعلق به رسته‌ای از یک متغیر هستند، جمع‌آوری می‌شوند و سپس براساس متغیرهای دیگر کلاس‌بندی می‌شوند. این رویکرد زمانی مفید است که برخی از رسته‌ها کمیاب هستند و در نتیجه با نمونه‌گیری چند جمله‌ای یا پواسن نمی‌توان به تعداد کافی از این رسته‌ها نمونه استخراج کرد. به عنوان مثال فرض کنید، می‌خواهید ارتباط بین مصرف سیگار و یک بیماری کمیاب را بررسی کنید. برای این منظور بهتر است، تعداد ثابتی از افراد مبتلا و غیر مبتلا در نظر گرفته شود و سپس آزمایش سیگاری بودن در مورد هر یک از آنها انجام شود. اگر شما از تعداد زیادی از افراد نمونه‌گیری و سپس آنها را با توجه به بیماری و سیگاری بودن دسته‌بندی کنید، این امکان وجود دارد که تعداد کمی از افراد مبتلا به بیماری در نمونه‌ی گرفته شده وجود داشته باشد و با توجه به آن نتوان نتایج معناداری گرفت.

نتیجه‌ی نمونه‌گیری از هر یک از طرح‌های نمونه‌گیری ذکر شده، در یک جدول شمارشی قابل بیان است. صرفاً از طریق نگاه کردن به داده‌ها نمی‌توان تشخیص داد، کدام طرح نمونه‌گیری به کار گرفته شده است. با این حال، طرح نمونه‌گیری از این جهت با اهمیت است که برآورد بعضی از پارامترهایی که به آسانی با یک طرح نمونه‌گیری قابل انجام است، با طرح‌های نمونه‌گیری دیگر غیرممکن است.

هر چند، طرح نمونه‌گیری، برآورد انجام شده را تحت تاثیر قرار می‌دهد، با این وجود هر سه طرح نمونه‌گیری، از یک آماره‌ی آزمون و یک توزیع مرجع جهت تصمیم‌گیری راجع به ارتباط یا عدم ارتباط بین متغیرهای سطر و ستون استفاده می‌کنند. البته، نامی که برای مسئله انتخاب می‌شود، بر اساس طرح نمونه‌گیری انتخاب شده تغییر خواهد کرد.

۲-۱-۳ آزمون استقلال در جداول توافقی دوطرفه[۱۲]

برای نمونه‌گیری چندجمله‌ای با احتمال‌های $\{\pi_{ij}\}$ در یک جدول پیشایندی $I\times J$، فرض استقلال آماری به ازای تمام i و j ها به صورت زیر است:

$$H_0:\pi_{ij}=\pi_{i.}\pi_{.j} \qquad (۱-۲)$$

برای نمونه‌های مستقل چندجمله‌ای در I سطر، استقلال، متناظر با یکسان بودن احتمال هر برآمد، بین سطرها است. بحث اشاره شده در اینجا، به یک تک‌نمونه‌ی چندجمله‌ای اشاره می‌کند. حال آن‌که همان آزمون‌ها برای نمونه‌های چندجمله‌ای مستقل نیز به کار گرفته می‌شوند.

جدول زیر، یک جدول رسته‌ای $I \times J$ است که در آن فرض استقلال ذکر شده در رابطه‌ی ۲-۱، با خانه‌های هاشور زده شده نشان داده شده است.

جدول ۲-۱- جدول رسته‌ای $I \times J$ [۲]

		متغیر کیفی A					
		1	…	j	…	J	مجموع سطری
متغیر کیفی B	1	π_{11}	…	π_{1j}	…	π_{1J}	$\pi_{1.}$
	⋮	⋮		⋮		⋮	⋮
	i	π_{i1}	…	π_{ij}	…	π_{iJ}	$\pi_{i.}$
	⋮	⋮		⋮		⋮	⋮
	I	π_{I1}	…	π_{Ij}	…	π_{IJ}	$\pi_{I.}$
	مجموع ستونی	$\pi_{.1}$	…	$\pi_{.j}$	…	$\pi_{.J}$	۱

۲-۱-۴ آماره‌ی پیرسون و توزیع آن

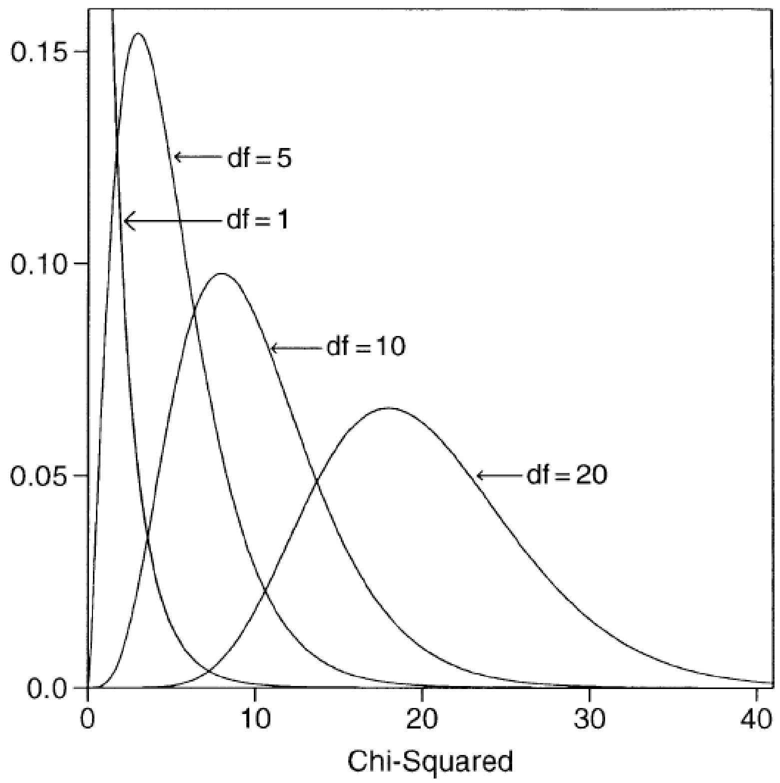
آماره‌ی کی‌دو پیرسون برای آزمون H_0 به صورت زیر است

$$X^2 = \sum \frac{\left(n_{ij} - \mu_{ij}\right)^2}{\mu_{ij}}$$

این آماره کمترین مقدار خود (صفر) را هنگامی اختیار می‌کند که به ازای تمام i و j ها، تساوی‌ها $n_{ij} = \mu_{ij}$ برقرار باشد. به ازای اندازه‌ی نمونه‌ی ثابت، هر چه حاصل تفاضل $\left\{n_{ij} - \mu_{ij}\right\}$ بیشتر باشد، مقادیر بزرگتری از X^2 به دست می‌آید و به طور طبیعی در پی آن شاهد قوی‌تری علیه H_0، فراهم می‌شود.

از آنجایی که مقادیر بزرگتر $X^۲$ در تناقض بیشتری با H_0 هستند، p مقدار، احتمال آن است که $X^۲$ دست‌کم به بزرگی مقدار مشاهده شده باشد. آماره‌ی $X^۲$ برای n بزرگ، به طور تقریبی دارای توزیع کی‌دو است. بر این اساس، p مقدار، احتمال آن است که دُم سمت راست توزیع، بالاتر از مقدار مشاهده شده $X^۲$ قرار گیرد. با افزایش $\{\mu_{ij}\}$، دقت تقریب کی‌دو افزایش پیدا می‌کند و معمولاً اگر $\{\mu_{ij} \geq ۵\}$ باشد یک تقریب خوب به دست خواهد آمد.

توزیع کی‌دو، بر روی مقادیر نامنفی متمرکز است. میانگین این توزیع با درجه‌ی آزادی آن برابر است و انحراف استاندارد آن برابر با مقدار $\sqrt{۲\text{df}}$ می‌باشد. با بزرگتر شدن درجه‌ی آزادی، توزیع در اطراف مقادیر بزرگتر، متمرکزتر و پهن‌تر می‌شود. این توزیع، چوله به راست است، اما شکل آن با بزرگتر شدن درجه‌ی آزادی، به شکل زنگوله‌ای (نرمال) نزدیک‌تر می‌شود. شکل زیر چگالی‌های کی‌دو با درجات آزادی ۱، ۵، ۱۰ و ۲۰ را نشان می‌دهد. مشاهده می‌شود که به ازای درجه‌ی آزادی ۲۰، شکل توزیع، به شکل زنگوله‌ای در آمده است.



شکل ۲-۱- نمودار توزیع کی‌دو به ازای درجات آزادی مختلف

۲-۱-۵ آماره‌ی نسبت درستنمایی

ابتدا یادآوری می‌شود که آزمون نسبت درستنمایی، مقادیری از پارامترها را تعیین می‌کند که تابع درستنمایی را تحت فرض‌های زیر بیشینه می‌کنند:

الف) تحت فرض درست بودن H_0

ب) تحت شرط کلی‌تری که طبق آن H_0 ممکن است درست یا نادرست باشد

آماره‌ی آزمون، از نسبت درستنمایی‌های ماکزیمم شده به صورت زیر بهره می‌گیرد

$$\Lambda = -2\log\frac{L(\hat{\Omega}_0)}{L(\hat{\Omega})}$$

که در آن، $L(\hat{\Omega}_0)$، مقدار تابع درستنمایی به ازای مقادیر پارامترها تحت فرض صفر است و $L(\hat{\Omega})$، مقدار تابع درستنمایی به ازای مقادیر پارامترها تحت فرض کلی‌تری است که در آن محدودیتی برای پارامترها موجود نیست.

توجه شود که مقدار آماره‌ی آزمون نامنفی است. زمانی که فرض صفر نادرست است، نسبت درستنمایی‌های ماکزیمم شده کمتر از ۱ و در پی آن، مقدار لگاریتم آن، منفی خواهد شد. از این‌رو ۲- بار لگاریتم نسبت درستنمایی، یک مقدار بزرگ مثبت خواهد شد و با بزرگتر شدن اندازه‌ی نمونه، این مقدار بیشتر نیز خواهد شد.

۲-۱-۶ آزمون کی‌دوی پیرسون

به هنگام استفاده از آماره‌ی X^2 در آزمون استقلال H_0، از n_{ij} به جای n_i و از μ_{ij} به جای μ_i استفاده می‌شود. در اینجا تحت فرض H_0

$$\mu_{ij} = E(n_{ij})$$

تحت فرض استقلال، پارامترهای $\{\pi_{i.}\}$ و $\{\pi_{.j}\}$ معمولاً ناشناخته هستند. برآوردهای ماکزیمم درستنمایی این پارامترها، که نسبت‌های حاشیه‌ای در نمونه هستند، برابر هستند با

$$\hat{\pi}_{i.} = \frac{n_{i.}}{n},$$

و

$$\hat{\pi}_{.j} = \frac{n_{.j}}{n}$$

بر این اساس، برآورد مقادیر مورد انتظار به صورت زیر خواهد بود

$$\hat{\mu}_{ij} = n\hat{\pi}_{i.}\hat{\pi}_{.j} = \frac{n_{i.}n_{.j}}{n}$$