فصل ۵

آزمون‌های نیکویی برازش و مستقل‌بودن داده‌ها

۱.۵ مقدمه

برای اینکه بدانیم مجموعه‌ای از داده‌ها دارای توزیع خاصی هستند یا نه می‌توانیم از آزمون نیکویی برازش استفاده کنیم. برای توزیع‌های گسسته می‌توان این آزمون را با استفاده از آماره پیرسن $X^۲$، آماره نسبت درستنمایی $G^۲$ یا آماره کولموگروف[۱]-اسمیرنوف[۲] برای توزیع‌های پیوسته انجام داد. در این فصل یک آماره جدید مبتنی بر فاصله آنتروپی معرفی می‌کنیم که برای آزمون نیکویی برازش نمونه‌های با اندازه بزرگ مناسب است. نتایج حاصل از آماره آزمون جدید را با نتایج حاصل از برخی آماره‌های دیگر مقایسه می‌کنیم. همچنین یک آماره جدید برای بررسی مستقل بودن داده‌های آماری معرفی خواهیم کرد. آنتروپی نسبی برای اولین بار در سال ۱۹۵۱ توسط کولبک[۳] و لیبلر[۴] معرفی شد که تحت عناوین مختلفی از جمله فاصله کولبک-لیبلر یا واگرایی داده‌ها شناخته شده است. در ادامه به معرفی آنتروپی نسبی و اطلاعات متقابل می‌پردازیم.

آنتروپی نسبی اندازه‌گیری فاصله بین دو توزیع می‌باشد که با استفاده از امید ریاضی لگاریتم نسبت درستنمایی به‌دست می‌آید. به عبارت دیگر، آنتروپی نسبی $D(q||p)$ معیاری برای ناکارآمدی توزیع p در مقابل توزیع واقعی q می‌باشد.

تعریف ۱.۱.۵. آنتروپی نسبی یا فاصله کولبک-لیبلر بین توابع چگالی احتمال $p(x)$ و $q(x)$ به‌صورت زیر می‌باشد

$$D(q||p) = \sum_{x\in S(x)} q(x)\log_۲\left(\frac{q(x)}{p(x)}\right) = E\Big[\log_۲\left(\frac{q(x)}{p(x)}\right)\Big], \qquad (۱.۵)$$

که در آن $S(x)$ تکیه‌گاه متغیر تصادفی X است.

با استفاده از خاصیت پیوسته‌بودن توابع چگالی احتمال داریم

$$q(x) = ۰ \Longrightarrow ۰\log\left(\frac{۰}{p}\right) = ۰,$$

$$p(x) = ۰ \Longrightarrow q\log\left(\frac{q}{۰}\right) = \infty.$$

[۱] Kolmogorov
[۲] Smirnov
[۳] Kullback
[۴] Leibler

آنتروپی نسبی همواره نامنفی است. همچنین $D(q||p) = ۰$ اگر و تنها اگر $p = q$ باشد. از طرفی آنتروپی نسبی فاصله واقعی بین دو توزیع نیست زیرا آنتروپی نسبی متقارن نیست $\Big(D(q||p) \neq D(p||q)\Big)$ و در نامساوی مثلثی صدق نمی‌کند [۱۲].

اطلاعات متقابل اندازه‌گیری مقدار اطلاعاتی است که یک متغیر تصادفی درباره متغیر تصادفی دیگر دارد. همچنین کاهش عدم قطعیت یک متغیر تصادفی به‌دلیل آگاهی از اطلاعات متغیر تصادفی دیگر است.

تعریف ۲.۱.۵. متغیرهای تصادفی X و Y با تابع چگالی احتمال توأم $p(x,y)$ و توابع چگالی احتمال حاشیه‌ای $p(x)$ و $p(y)$ را در نظر بگیرید. آنتروپی نسبی بین تابع چگالی احتمال توأم و حاصل‌ضرب $p(x)p(y)$ را اطلاعات متقابل می‌نامیم و به‌صورت زیر محاسبه می‌کنیم

$$\begin{aligned} I(X,Y) &= D\Big(p(x,y)||p(x)p(y)\Big) \\ &= \sum_x \sum_y p(x,y) \log_۲ \Big(\frac{p(x,y)}{p(x)p(y)}\Big) \\ &= E\Big[\log_۲ \Big(\frac{p(x,y)}{p(x)p(y)}\Big)\Big]. \end{aligned} \tag{۲.۵}$$

در حالت گسسته، اطلاعات متقابل مانند آنتروپی نسبی همواره نامنفی است. این معیار در واقع میزان وابستگی دو متغیر تصادفی را نشان می‌دهد. از طرفی $I(X,Y) = ۰$ اگر و تنها اگر متغیرهای تصادفی X و Y مستقل باشند. همچنین اطلاعات متقابل متقارن است. یعنی

$$I(X,Y) = I(Y,X).$$

۲.۵ آزمون نیکویی برازش

فرض کنید نتیجه یک آزمایش تصادفی به یکی از دسته‌های دو به دو مجزای $A_۱$، $A_۲$، ... ،A_k تعلق داشته باشد و احتمال قرارگرفتن نتیجه آزمایش تصادفی در دسته A_i برابر q_i باشد که

$$۰ \leq q_i \leq ۱, \qquad \sum_{i=۱}^{k} q_i = ۱.$$

این توزیع احتمال مجهول را با $q(x)$ نمایش می‌دهیم. می‌خواهیم بررسی کنیم که آیا این آزمایش تصادفی دارای توزیع $p(x)$ تحت شرایط

$$p(A_i) = p_i, \qquad ۰ \leq p_i \leq ۱, \qquad \sum_{i=۱}^{k} p_i = ۱,$$

می‌باشد یا خیر. به‌عبارت دیگر می‌خواهیم برای هر x فرض

$$H_۰ : \ q(x) = p(x),$$

را در مقابل فرض

$$H_۱ : \ q(x) \neq p(x),$$

برای حداقل یک مقدار x آزمون کنیم. لذا آزمایش تصادفی را n بار تکرار می‌کنیم و فرض می‌کنیم O_i فراوانی حالت‌هایی باشد که نتیجه آزمایش به A_i تعلق داشته باشد. اگر فرض $H_۰$ برقرار باشد آن‌گاه

$$e_i = E(O_i) = np_i.$$

برای بررسی فرضیه فوق می‌توانیم از آماره‌های زیر استفاده کنیم.

(۱) آماره پیرسن:

آماره پیرسن به‌صورت زیر می‌باشد

$$X^۲ = \sum_{i=۱}^{k} \frac{(O_i - e_i)^۲}{e_i}.$$

برای نمونه‌های با اندازه بزرگ $X^۲$ دارای توزیع کای‌دو با درجه آزادی $k - ۱$ می‌باشد. بنابراین اگر

$$X^۲ > \chi^۲(k - ۱, \alpha),$$

باشد آن‌گاه فرض $H_۰$ رد می‌شود که در آن $\chi^۲(k - ۱, \alpha)$ چندک $\alpha - ۱$ از توزیع کای‌دو است [۲۱].

(۲) آماره نسبت درستنمایی:

آماره نسبت درستنمایی به‌صورت زیر می‌باشد

$$G^۲ = ۲ \sum_{i=۱}^{k} O_i \log_۲ \left(\frac{O_i}{e_i}\right).$$

برای نمونه‌های با اندازه بزرگ $G^۲$ دارای توزیع کای‌دو با درجه آزادی $k - ۱$ می‌باشد. بنابراین اگر

$$G^۲ > \chi^۲(k - ۱, \alpha),$$

باشد آن‌گاه فرض $H_۰$ رد می‌شود. آماره $G^۲$ به‌صورت مجانبی با آماره پیرسن $X^۲$ معادل است [۱۲].

(۳) آماره کولموگروف–اسمیرنوف:

آماره کولموگروف–اسمیرنوف به‌صورت زیر می‌باشد

$$T_n = \sup_{x \in S(x)} \sqrt{n} |F_n(x) - F(x)|,$$

که در آن $F_n(x)$ توزیع تجربی از نمونه و $F(x)$ تابع توزیع متغیر تصادفی پیوسته X است. اگر T_n از مقدار بحرانی جداول مربوطه بیشتر باشد آنگاه فرض $H_۰$ رد می‌شود [۲۱].

۳.۵ آزمون نیکویی برازش مبتنی بر آنتروپی نسبی

در این بخش با استفاده از آنتروپی نسبی یک آماره جدید برای آزمون نیکویی برازش معرفی می‌کنیم. فرض $H_۰ : q = p$ را در مقابل $H_۱ : q \neq p$ در نظر بگیرید که با فرضیه زیر معادل است

$$H_۰ : \ D(q||p) = ۰$$

$$H_۱ : \ D(q||p) > ۰$$

که در آن

$$D(q||p) = \sum_{i=۱}^{k} q_i \log_۲ \Big(\frac{q_i}{p_i}\Big) = E\Big[\log_۲ \Big(\frac{q}{p}\Big)\Big].$$

فرض کنید O_i فراوانی حالت‌هایی باشد که نتیجه آزمایش به A_i تعلق دارد و $e_i = np_i$ باشد. برآوردگر درستنمایی ماکسیمم q_i به‌صورت $\hat{q}_i = \dfrac{O_i}{n}$ است. چون $p_i = \dfrac{e_i}{n}$ می‌باشد لذا برآوردگر درستنمایی ماکسیمم برای $D(q||p)$ برابر است با:

$$\hat{D}(q||p) = \frac{۱}{n} \sum_{i=۱}^{k} O_i \log_۲ \Big(\frac{O_i}{e_i}\Big). \qquad (۳.۵)$$

می‌دانیم که $O = (O_۱, O_۲, ..., O_k)$ دارای توزیع چندجمله‌ای است. به‌عبارت دیگر

$$O = (O_۱, O_۲, ..., O_k) \sim M_k(n, (q_۱, q_۲, ..., q_k)).$$

برای یک نمونه با اندازه بزرگ O دارای توزیع نرمال چندمتغیره متقارن $N_k(nq, n(D_q - qq'))$ است که در آن D_q یک ماتریس قطری با عناصر قطری q_i می‌باشد و $q = (q_۱, q_۲, ..., q_n)$ است. در نتیجه

$$\sqrt{n}\Big(\frac{۱}{n}O - q\Big) \longrightarrow N_k(۰, D_q - qq').$$

با استفاده از (۳.۵) و قضایای حد داریم

$$Z = \sqrt{n}\Big(\frac{\hat{D}(q||p) - D(q||p)}{\hat{\sigma}}\Big) \longrightarrow N(۰, ۱), \qquad (۴.۵)$$

که در آن

$$\hat{\sigma}^۲ = \frac{۱}{n}\left[\sum_i O_i \Big(\log_۲ \frac{O_i}{e_i}\Big)^۲ - \Big(\sum_i O_i \log_۲ \frac{O_i}{e_i}\Big)^۲\right]. \qquad (۵.۵)$$

اکنون آماره زیر را در نظر می‌گیریم

$$Z_۰ = \frac{\sqrt{n}\hat{D}(q||p)}{\hat{\sigma}}. \qquad (۶.۵)$$

اگر $Z_۰ > Z_\alpha$ باشد فرض $H_۰$ را رد می‌کنیم که در آن Z_α چندک $\alpha - ۱$ از توزیع نرمال استاندارد می‌باشد.

تذکر ۱.۳.۵. آماره $Z_۰$ به آماره $G^۲$ بسیار نزدیک است اما حساسیت آن از $G^۲$ بیشتر می‌باشد. از طرفی نمی‌توانیم آماره جدید را با آزمون نیکویی برازش مبتنی بر آماره‌های $X^۲$، $G^۲$ و T_n مقایسه کنیم زیرا آماره جدید برای نمونه‌های با اندازه بسیار بزرگ مناسب است و دیگر نمی‌توان آن را به‌صورت یک مدل پارامتری توصیف کرد [۲۱]. لذا در ادامه از مطالعات شبیه‌سازی برای مقایسه این آماره‌ها استفاده می‌کنیم و حساسیت آن‌ها را مقایسه می‌کنیم.

۴.۵ مثال‌ها

در این بخش با استفاده از چند مثال به بررسی روش‌های معرفی شده در بخش قبل می‌پردازیم.

مثال ۱.۴.۵. از بین یک جمعیت بزرگ نمونه‌ای ۲۰۰ نفری را انتخاب کرده و تعداد دفعات مراجعه آن‌ها به یک شرکت بیمه در دوره ۴ ساله را در نظر می‌گیریم. نتایج حاصل در جدول ۱.۵ ارائه شده است. فرض کنید X تعداد دفعاتی باشد

جدول ۱.۵: دفعات مراجعه نمونه ۲۰۰ نفری به شرکت بیمه

شماره	۰	۱	۲	۳	۴	۵	۶	۷
فراوانی	۲۲	۵۳	۵۸	۳۹	۲۰	۵	۲	۱

که یک شخص به شرکت بیمه مراجعه می‌کند. می‌خواهیم فرضیه زیر را بررسی کنیم

$$H_۰ : \ X \sim P(\lambda)$$

$$H_۱ : \ X \not\sim P(\lambda)$$

آماره آزمون $Z_۰ = ۰/۸۰۱۲$ و $Z_{۰/۰۵} = ۱/۶۴۵$ است. چون $Z_{۰/۰۵} > Z_۰$ است لذا فرض $H_۰$ رد نمی‌شود. همچنین آماره پیرسون $X^۲ = ۲۳$ و $\chi^۲(۴, ۰/۰۵) = ۹/۴۸۷$ می‌باشد و در این آزمون نیز $H_۰$ رد نمی‌شود.

مثال ۲.۴.۵. نتایج یک نظرسنجی درباره موضوعی خاص در یک کشور به‌صورت جدول ۲.۵ است. یک نمونه صدتایی از این جامعه انتخاب شده است که نتایج حاصل در جدول ۳.۵ ارائه شده است. آیا نتایج حاصل در جدول ۳.۵ با نتایج نظرسنجی در جامعه مطابقت دارد؟

جدول ۲.۵: نتایج نظرسنجی در جامعه

نظر	خیلی موافق	موافق	ممتنع	مخالف	خیلی مخالف
درصد	۲۰	۳۰	۲۰	۲۰	۱۰

جدول ۳.۵: نتایج نظرسنجی در نمونه ۱۰۰ تایی

نظر	خیلی موافق	موافق	ممتنع	مخالف	خیلی مخالف
فراوانی	۱۴	۱۸	۱۸	۲۶	۲۴

با استفاده از (۶.۵) آماره $Z_۰ = ۲/۳۱۵۶$ حاصل می‌شود. همچنین از جدول توزیع نرمال مقدار $Z_{۰/۰۵} = ۱/۶۴۵$ به‌دست می‌آید. با مقایسه این مقادیر $Z_۰ > Z_{۰/۰۵}$ است. در نتیجه فرض $H_۰$ رد می‌شود و نتایج حاصل در نمونه ۱۰۰ تایی با نتایج کلی جامعه مطابقت ندارد. همچنین آماره پیرسن به‌صورت $X^۲ = ۴۶/۵$ و $\chi^۲(۴, ۰/۰۵) = ۹/۴۸۷$ در سطح معنی‌داری $\alpha = ۰/۰۵$ است. در این حالت نیز فرض $H_۰$ رد می‌شود.

مثال ۳.۴.۵. یک تاس ۱۲۰ بار پرتاب و نتایج حاصل در جدول ۴.۵ ارائه شده است.

جدول ۴.۵: نتایج حاصل از ۱۲۰ بار پرتاب یک تاس

تاس	۱	۲	۳	۴	۵	۶
فراوانی	۱۸	۲۳	۱۶	۲۱	۱۸	۲۴

می‌دانیم که

$$p_i = \frac{۱}{۶} \qquad e_i = ۲۰ \times \frac{۱}{۶} = ۲۰ \qquad i = ۱, ۲, ..., ۶.$$

در نتیجه آماره $Z_۰ = ۰/۷۹۱۴$ حاصل می‌شود. از جدول توزیع نرمال نیز $Z_{۰/۰۵} = ۱/۶۴۵$ است. چون $Z_{۰/۰۵} > Z_۰$ می‌باشد لذا فرض $H_۰$ رد نمی‌شود. همچنین آماره پیرسن به‌صورت $X^۲ = ۲/۵$ و $\chi^۲(۵, ۰/۰۵) = ۱۱/۱$ می‌باشند که در این حالت نیز فرض $H_۰$ رد نمی‌شود.

۵.۵ شبیه‌سازی

کلاس فرضیه‌های $H_۱$ خیلی بزرگ است و همه توزیع‌ها به جز توزیع‌های $H_۰$ را شامل می‌شود. لذا نمی‌توانیم تابع توان آماره جدید را با روش‌های معمول محاسبه و مقایسه کنیم. در این بخش با استفاده از توزیع‌های پواسن و نمایی اعداد تصادفی را تولید کرده و به کمک آن‌ها آزمون نیکویی برازش را انجام می‌دهیم. برای مقایسه توابع آزمون از مقادیر p

استفاده می‌کنیم.

۱.۵.۵ آزمون پواسن

در آزمون پواسن ۱۰۰۰ عدد تصادفی از توزیع $P(\lambda_1)$ را به‌ازای $\lambda_1 = 1,2$ شبیه‌سازی کرده و فرضیه‌های زیر را در نظر می‌گیریم

$$H_0 : \ X \sim P(\lambda_2)$$

$$H_1 : \ X \not\sim P(\lambda_2)$$

که مقادیر λ_2 در جدول‌های ۵.۵ و ۶.۵ آمده است. همچنین

$$\hat{d} = \hat{D}(q||p), \qquad \hat{\sigma}^2 = S_d^2.$$

داده‌های ارائه شده در جدول ۵.۵ از توزیع پواسن به‌ازای $\lambda = 1$ تولید شده‌اند. در سطرهای I مقادیر λ_2 نزدیک یک هستند و فرض H_0 پذیرفته می‌شود. اما در سطرهای III مقادیر λ_2 از مقدار یک فاصله دارند و فرض H_0 رد می‌شود. اگر مقادیر λ_2 بین ۱/۰۱۹۶ و ۱/۰۲ باشند آن‌گاه آماره جدید مبتنی بر توزیع Z فرض H_0 را به درستی رد می‌کند ولی آماره‌های X^2 و G^2 فرض H_0 را به اشتباه قبول می‌کنند. در نتیجه آماره جدید در مقایسه با آماره‌های موجود در رد فرض H_0 حساس‌تر است. به‌طور مشابه داده‌های ارائه شده در جدول ۶.۵ از توزیع پواسن به‌ازای $\lambda = 2$ تولید شده‌اند. در این حالت به‌ازای $\lambda_2 \in (2/005, 2/03)$ آماره جدید فرض H_0 را رد و آماره‌های دیگر فرض H_0 را قبول می‌کنند.

جدول ۵.۵: نتایج آزمون پواسن برای ۱۰۰۰ عدد تصادفی تولید شده با $P(1)$

$p(G^2)$	$p(X^2)$	$p(Z)$	G^2	X^2	Z	S_d^2	$\hat{d}$	λ_2	
۰/۸۸	۰/۸۹	۰/۱۲۵۱	۲/۳۶	۲/۳۲	۱/۱۵	۰/۰۰۴۸	۰/۰۰۰۸	۱	I
۰/۷۷۴	۰/۷۷۸۸	۰/۱۰۷۵	۳/۲۹	۳/۲۱	۱/۲۴	۰/۰۰۴۹	۰/۰۰۰۸	۱/۰۰۵	
۰/۵۸۴	۰/۶	۰/۰۸۵۵	۴/۷۱	۴/۵۸	۱/۳۵	۰/۰۰۵	۰/۰۰۰۹	۱/۰۱	
۰/۳۷	۰/۴	۰/۰۶۸	۶/۶۱	۶/۴۳	۱/۴۹	۰/۰۰۵	۰/۰۰۱۱	۱/۰۱۵	
۰/۲۱۵	۰/۲۲	۰/۰۵۲۶	۸/۲۸	۸/۲۵	۱/۶۳	۰/۰۰۶	۰/۰۰۱۲	۱/۰۱۹	
۰/۲۰۱	۰/۲۱	۰/۰۵۰۵	۸/۷۴	۸/۴۹	۱/۶۴۶	۰/۰۰۶	۰/۰۰۱۳	۱/۰۱۹۵	
۰/۱۹۸	۰/۲۱	۰/۰۴۹۵	۸/۷۹	۸/۵۴	۱/۶۵	۰/۰۰۶	۰/۰۰۱۳	۱/۰۱۹۶	II
۰/۱۹۶	۰/۲۰۹	۰/۰۴۹۵	۸/۸۳	۸/۵۹	۱/۶۵۳	۰/۰۰۶	۰/۰۰۱۳	۱/۰۱۹۷	
۰/۱۸۸	۰/۲	۰/۰۴۸۵	۸/۹۹	۸/۷۵	۱/۶۶	۰/۰۰۶	۰/۰۰۱۳	۱/۰۲	
۰/۰۴	۰/۰۲۳	۰/۰۲۰۲	۱۵/۲	۱۲/۷۷	۲/۰۵	۰/۰۰۷	۰/۰۰۱۷	۱/۰۳	III
۰	۰	۰/۰۰۶۲	۲۳/۲۷	۲۲/۶	۲/۵	۰/۰۰۸۵	۰/۰۰۲۳	۱/۰۴	
۰	۰	۰/۰۰۱۵	۳۳/۱۸	۳۲/۱۸	۲/۹۶	۰/۰۰۱	۰/۰۰۳	۱/۰۵	

جدول ۶.۵: نتایج آزمون پواسن برای ۱۰۰۰ عدد تصادفی تولید شده با $P(۲)$

	$\lambda_۲$	$\hat{d}$	$S_d^۲$	Z	$X^۲$	$G^۲$	$p(Z)$	$p(X^۲)$	$p(G^۲)$
I	۲	۰/۰۰۱	۰/۰۰۵	۱/۵۸	۳/۷۲	۳/۸۱	۰/۰۵۷۱	۰/۸۸	۰/۸۶
	۲/۰۰۱	۰/۰۰۱	۰/۰۰۵	۱/۶	۳/۸۷	۳/۹۶	۰/۰۵۴۸	۰/۸۶	۰/۸۵۵
	۲/۰۰۵	۰/۰۰۱۷	۰/۰۰۵	۱/۶۵	۴/۵۳	۴/۶۴	۰/۰۴۹۵	۰/۸	۰/۷۹
II	۲/۰۱	۰/۰۰۱۲	۰/۰۰۵	۱/۷۳	۵/۵۸	۵/۷۲	۰/۰۴۱۸	۰/۶۹	۰/۶۸
	۲/۰۳	۰/۰۰۲	۰/۰۰۶	۲/۱۷	۱۲/۱۷	۱۲/۴۷	۰/۰۱۵۰	۰/۱۶	۰/۱۴
	۲/۰۴	۰/۰۰۲	۰/۰۰۸	۲/۴۵	۱۶/۸۹	۱۷/۳	۰/۰۰۷۱	۰/۰۳۳	۰/۰۲۸
III	۲/۰۵	۰/۰۰۳	۰/۰۰۸	۲/۷۵	۲۲/۵۲	۲۳/۰۸	۰/۰۰۳	۰	۰
	۲/۰۶	۰/۰۰۳	۰/۰۰۹۵	۳/۰۶	۲۹/۰۷	۲۹/۸۲	۰/۰۰۱۱	۰	۰
	۲/۰۷	۰/۰۰۳۵	۰/۰۱۱	۳/۳۸	۳۶/۵۲	۳۷/۴۸	۰	۰	۰

۲.۵.۵ آزمون نمایی

مشابه حالت قبل تعداد ۱۰۰۰ عدد تصادفی از توزیع نمایی $\exp(\lambda_۱)$ را به‌ازای $\lambda_۱ = ۱$ شبیه‌سازی کرده و فرضیه‌های زیر را در نظر می‌گیریم

$$H_۰ : \ X \sim \exp(\lambda_۲)$$

$$H_۱ : \ X \not\sim \exp(\lambda_۲)$$

که مقادیر $\lambda_۲$ در جدول ۷.۵ آمده است. همچنین مقادیر $\hat{d}$، $\hat{\sigma}^۲$ و مقادیر p وابسته به آماره‌های Z، $X^۲$، $G^۲$ و T_n محاسبه شده است. از نتایج ارائه شده در جدول ۷.۵ واضح است که به‌ازای مقادیر $\lambda_۲$ نزدیک یک در سطر IV فرض $H_۰$ پذیرفته می‌شود و به‌ازای $\lambda_۲ = ۱/۰۶$ در سطرهای I که دور از یک می‌باشد فرض $H_۰$ رد می‌شود. همچنین به‌ازای $\lambda \in (۱/۰۳۹, ۱/۰۴۱)$ آماره جدید Z فرض $H_۰$ را رد می‌کند در حالی‌که آماره‌های $X^۲$، $G^۲$ و T_n فرض $H_۰$ را می‌پذیرند. در نتیجه آماره جدید در مقایسه با آماره‌های دیگر حساس‌تر است.

۶.۵ آزمون مستقل بودن مبتنی بر اطلاعات متقابل

یک نمونه تصادفی به اندازه n از یک جامعه را در نظر بگیرید. مشاهدات در نمونه تصادفی براساس دو معیار طبقه‌بندی می‌شود. با استفاده از معیار اول هر مشاهده با یکی از ردیف‌ها (R) و با استفاده از معیار دوم هر مشاهده با یکی از ستون‌ها (C) مرتبط می‌شود. فرض کنید O_{ij} تعداد مشاهدات مرتبط با سطر iام و ستون jام باشد. این مقادیر را در جدول احتمالی از مرتبه $R \times C$ به‌صورت زیر قرار می‌دهیم.

جدول ۷.۵: نتایج آزمون نمایی برای ۱۰۰۰ عدد تصادفی تولید شده با exp(۱)

$p(G^۲)$	$p(X^۲)$	$p(Z)$	T_n	$G^۲$	$X^۲$	Z	$S_d^۲$	$\hat{d}$	$\lambda_۲$	
۰/۷۶۶	۰/۷۶۷	۰/۱۲۵۱	۰/۷۶	۴/۱	۴/۰۹	۱/۱۵	۰/۰۰۱۲	۰/۰۰۰۴	۱	I
۰/۷	۰/۶۷۳	۰/۱۲۱	۰/۴۰	۴/۶۶	۴/۹	۱/۱۷	۰/۰۰۱۴	۰/۰۰۰۴	۱/۰۲	
۰/۳۸	۰/۳۳۷	۰/۰۷۷۸	۰/۶۴	۷/۶۲	۸/۱	۱/۴۲	۰/۰۰۲۱	۰/۰۰۰۷	۱/۰۳	
۰/۱۱	۰/۰۸۶۲	۰/۰۴۲۷	۰/۸۷	۱۱/۷۵	۱۲/۵۸	۱/۷۲	۰/۰۰۳	۰/۰۰۰۹	۱/۰۳۹	II
۰/۰۹	۰/۰۷۲	۰/۰۴۰۱	۰/۹۰	۱۲/۳	۱۳/۱۸	۱/۷۵	۰/۰۰۳۲	۰/۰۰۱	۱/۰۴	
۰/۰۸۳	۰/۰۵۷۳	۰/۰۳۶۷	۰/۹۲	۱۲/۸۶	۱۳/۷۹	۱/۷۹	۰/۰۰۳۳	۰/۰۰۱	۱/۰۴۱	
۰/۰۰۹	۰/۰۰۵۵	۰/۰۱۶۶	۱/۱۰	۱۸/۶۶	۲۰/۱۳	۲/۱۳	۰/۰۰۴۶	۰/۰۰۱۴	۱/۰۵	III
۰	۰	۰/۰۰۵۹	۱/۳۸	۲۶/۶۷	۲۸/۹۷	۲/۵۲	۰/۰۰۶۵	۰/۰۰۲	۱/۰۶	IV

Y / X	۱	۲	$\cdots$	C	Sum
۱	$O_{۱۱}$	$O_{۱۲}$	$\cdots$	$O_{۱C}$	$O_{۱}$
۲	$O_{۲۱}$	$O_{۲۲}$	$\cdots$	$O_{۲C}$	$O_{۲}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
R	$O_{R۱}$	$O_{R۲}$	$\cdots$	O_{RC}	O_R
Sum	$O_{\cdot ۱}$	$O_{\cdot ۲}$	$\cdots$	$O_{\cdot C}$	n

مجموع مقادیر O_{ij} در سطر iام برابر $O_{i\cdot}$، مجموع مقادیر $O_{\cdot j}$ در سطر jام برابر $O_{\cdot j}$ و مجموع کل آن‌ها برابر n است. فرض $H_۰$ را مستقل بودن یک مشاهده در سطر i ام از برخی مشاهدات در ستون jام در نظر می‌گیریم. بنابراین

فرض $H_۰$:

$$p(i \text{ سطر}, j \text{ ستون}) = p(i \text{ سطر}) \times p(j \text{ ستون}).$$

فرض $H_۱$:

$$p(i \text{ سطر}, j \text{ ستون}) \neq p(i \text{ سطر}) \times p(j \text{ ستون}).$$

آماره آزمون به‌صورت

$$X^۲ = \sum_j \sum_i \frac{(O_{ij} - e_{ij})^۲}{e_{ij}},$$

می‌باشد که در آن

$$e_{ij} = \frac{O_{i\cdot} \times O_{\cdot j}}{n}.$$

این آماره دارای توزیع کای‌دو با درجه آزادی $(R-۱)(C-۱)$ است. لذا اگر $X^۲ > \chi^۲((R-۱)(C-۱), \alpha)$ باشد آنگاه فرض $H_۰$ رد می‌شود. با استفاده از تعریف اطلاعات متقابل فرض‌های $H_۰$ و $H_۱$ را می‌توانیم به‌صورت زیر بیان کنیم

$$H_۰ : \ I(X,Y) = ۰$$

$$H_۱ : \ I(X,Y) \neq ۰$$

یا به‌طور معادل

$$H_۰ : \ D(p(x,y)||p(x)p(y)) = ۰$$

$$H_۱ : \ D(p(x,y)||p(x)p(y)) > ۰$$

همچنین برآوردگر ماکسیمم درستنمایی برای $I(X,Y)$ برابر است با:

$$\hat{I}(X,Y) = \frac{۱}{n} \sum_i \sum_j O_{ij} \log_۲ \Big(\frac{O_{ij}}{e_{ij}}\Big),$$

که در آن $O = (O_{۱۱}, O_{۱۲}, ..., O_{RC})$ مقادیر مشاهده شده است که دارای توزیع چندجمله‌ای می‌باشد و اگر اندازه نمونه به اندازه کافی بزرگ باشد دارای توزیع نورمال خواهد بود. به‌عبارت دیگر

$$Z = \frac{\sqrt{n}\Big(\hat{I}(X,Y) - I(X,Y)\Big)}{\hat{\sigma}} \longrightarrow N(۰,۱),$$

که در آن

$$\hat{\sigma}^۲ = \frac{۱}{n}\left[\sum_i \sum_j O_{ij}\Big(\log_۲ \frac{O_{ij}}{e_{ij}}\Big)^۲ - \Big(\sum_i \sum_j O_{ij} \log_۲ \frac{O_{ij}}{e_{ij}}\Big)^۲\right].$$

سرانجام اگر

$$Z_۰ = \frac{\sqrt{n}(\hat{I}(X,Y))}{\hat{\sigma}} > Z_\alpha,$$

آنگاه فرض $H_۰$ رد می‌شود.

مثال ۱.۶.۵. تعداد ۱۲۰۰ نفر در چهار گروه و دو دیدگاه متفاوت طبقه‌بندی شده‌اند که نتایج حاصل در جدول ۸.۵ ارائه شده است. آیا این دو نوع دیدگاه مستقل از هم هستند؟

با استفاده از داده‌های جدول ۸.۵ آماره $X^۲ = ۲۰/۵۹$ و از جدول توزیع خی‌دو $\chi^۲(۳, ۰/۰۵) = ۷/۸۱۵$ حاصل می‌شود. چون $X^۲ > \chi^۲(۳, ۰/۰۵)$ است لذا فرض مستقل بودن رد می‌شود. همچنین $Z_۰ = ۲/۵۱۴$ و $Z_{۰/۰۵} = ۱/۶۴۵$ است و در نتیجه فرض مستقل بودن نیز با این آماره رد می‌شود. برای این مثال هر دو آماره نتیجه یکسانی ارائه می‌دهند.

جدول ۸.۵: ۱۲۰۰ نفر در چهارگروه با دو دیدگاه

دیدگاه / گروه	I	II	مجموع
A	۳۲	۲۶۹	۳۰۰
B	۵۱	۱۹۹	۲۵۰
C	۶۷	۲۳۳	۳۰۰
D	۸۳	۲۶۷	۳۵۰
مجموع	۲۳۳	۹۶۷	۱۲۰۰

مثال ۲.۶.۵. برای بررسی تأثیر یک داروی شیمیایی روی بذر، ۲۵۰ بذر را انتخاب می‌کنیم که برای ۱۰۰ مورد از آن‌ها این دارو استفاده شده است و برای ۱۵۰ بذر دیگر از این دارو استفاده نشده است. نتایج حاصل در جدول ۹.۵ ارائه شده است.

جدول ۹.۵: تأثیر داروی شیمیایی روی بذر

	مثبت	منفی
با داروی شیمیایی	۸۴	۱۶
بدون داروی شیمیایی	۱۳۲	۱۸

با استفاده از داده‌های جدول ۹.۵ آماره $X^۲ = ۰/۸۱۷$ و از جدول توزیع خی‌دو $\chi^۲(۱, ۰/۰۵) = ۳/۸۴۱$ حاصل شده است. چون $\chi^۲(۳, ۰/۰۵) > X^۲$ است لذا فرض مستقل بودن قبول می‌شود. همچنین آماره جدید $Z_۰ = ۰/۴۴۵۶$ و $Z_{۰/۰۵} = ۱/۶۴۵$ است و در نتیجه فرض مستقل بودن نیز با این آماره رد نمی‌شود.