شواهد آماری: تفسیر نتایج آزمون‌های آماری کلاسیک [۱]

ناصر رضا ارقامی [۲]، احسان زمان‌زاده [۳]، سعید راحتی [۴]

چکیده:

این مقاله اولین از دنباله مقاله‌هایی است که طی آن‌ها روش درست اندازه‌گیری میزان شواهد آماری [۵] موجود در داده‌ها بر علیه یک فرضیه آماری مورد بررسی قرار می‌گیرد. در مقاله حاضر ابتدا به تفسیرهای رایج ولی نادرست نتایج آزمون‌های آماری کلاسیک پرداخته می‌شود و سپس روش درست تفسیر نتایج آزمون‌ها را با توجه به هدف محقق مورد بحث قرار می‌دهیم.

واژه‌های کلیدی: آزمون‌های نیمن – پیرسون، خطای نوع اول، خطای نوع دوم، قانون درستنمایی، شواهد آماری، استنباط شواهدی

۱ مقدمه

جملات زیر نمونه‌هایی از تفسیرهایی است که در گزارش‌های طرح‌های پژوهشی کاربردی به چشم می‌خورد:

۱ – فرضیهٔ $H_{\circ}$ در سطح ۵% رد می‌شود پس ۹۵% اطمینان داریم که $H_{\circ}$ نادرست است.

۲ – فرضیهٔ صفر پذیرفته می‌شود و چون آزمون در اندازه $\alpha = \%۱$ بوده است، به درستی این فرضیه که "میانگین افسردگی در زنان و مردان یکسان است" اطمینان زیادی داریم.

۳ – نتیجه آزمون عبارت است از $p = \%۳$– مقدار، پس ۹۷% اطمینان داریم که "احتمال ابتلا به دیابت به جنس بستگی دارد".

۴ – چون در خروجی $SPSS$، نتیجه آزمون همبستگی به صورت $sig = ۰/۰۰۰۰۰$ آمده است، "رتبه کنکور شدیداً به درآمد پدر بستگی دارد".

هدف از ارائه این مقاله این است که نشان دهیم چرا تفسیرهایی از انواعی که در بالا ذکر شد نادرست است و چرا نتیجه یک آزمون کلاسیک آماری (رد یا قبول $H_{\circ}$) باید با توجه به هدف محقق تفسیر شود.

۲ هدف آزمون‌های نیمن – پیرسون

مبنای آزمون‌های آماری کلاسیک از نوعی که غالب نرم‌افزارهای آماری (منجمله $SPSS$) بر مبنای آن ساخته شده‌اند، نظریه نیمن – پیرسون است. در روش

---

[۱] این مقاله به عنوان بخشی از قرارداد پژوهشی شماره ۳۰۰۸ مورد حمایت دانشگاه امام رضا (ع) بوده است.

[۲] دانشگاه فردوسی مشهد

[۳] دانشگاه فردوسی مشهد

[۴] دانشگاه امام رضا (ع)

[۵] Statistical edvidence

نیمن — پیرسون آزمون طوری انجام می‌شود که احتمال خطای نوع اول (احتمال رد شدن فرض H_0 وقتی فرض H_0 درست است) برابر مقدار از قبل تعیین شده کوچکی مانند α باشد. در این آزمون‌ها احتمال خطای نوع دوم (احتمال پذیرفتن H_0 وقتی H_1 درست است) معمولاً از قبل تعیین نمی‌شود ولی مقدار آن (β) هر چه حجم نمونه بیشتر باشد، کمتر است. البته می‌توان مقدار آن را (لااقل در تئوری) با دانستن مقادیر α و n (حجم نمونه) معین کرد، ولی با حجم نمونه ثابت هر چه α را کوچکتر اختیار کنیم مقدار β بزرگتر خواهد شد. این‌که چه مقداری را محقق برای α و در نتیجه برای β باید انتخاب کند، بستگی به این دارد که چه زیان‌هایی را محقق هنگام ارتکاب خطاهای نوع اول و دوم متحمل خواهد شد. سوالی که در پارادایم نیمن — پیرسون مطرح می‌باشد این است که وقتی α و حجم نمونه معلوم است چگونه آزمون را باید انجام داد طوری که β مینیمم شود. [۶] چون توان آزمون برابر $\beta - ۱$ است، آزمونی که از لم نیمن — پیرسون نتیجه می‌شود، زمانی‌که فرض‌های H_0 و H_1 ساده هستند "تواناترین" (MP) نامیده می‌شود. نکته مهمی که باید در مورد آزمون‌های نیمن — پیرسون به آن توجه داشت این است که گرچه ویژگی‌های یک آزمون (احتمالات خطاهای آزمون) قابل تفسیراند (به این معنی که آزمون، تحت H_0، $۱۰۰\alpha\%$ موارد و تحت H_1، $۱۰۰\beta\%$ موارد، ما را اشتباه راهنمایی می‌کند) ولی نتیجه آزمون به‌همین صورت قابل تفسیر نیست. نتیجه یک آزمون نیمن — پیرسون چیزی غیر از یکی از دو جمله "H_0 رد شد" و "H_0 پذیرفته شد" نیست و هیچگونه تفسیری بر آن مترتب نمی‌باشد. مسلماً همه می‌دانیم که "رد شدن H_0" به معنی "نادرست بودن H_0" و "پذیرفتن H_0" به معنی "نادرست بودن H_1" نیست. پس تفسیر نتیجه یک آزمون نیمن — پیرسون چیست؟

بهتر است نظر خود نیمن را در این مورد بدانیم:

"مسئله آزمون یک فرضیهٔ آماری زمانی پیش می آید که شرایط ما را وادار کرده است که از بین دو عمل A و B یکی را انتخاب نماییم[۹].

"بسیار مهم است که معنی دقیق "پذیرفتن H_1" را (که عبارت است از این که عمل A را و نه عمل B باید انجام دهیم) به خاطر بسپاریم[۹].

یعنی در نظریهٔ نیمن — پیرسون فرض بر این است که محقق مردد است که از دو عمل A و B کدام را انجام دهد ولی می داند که اگر H_0 درست باشد بهتر است عمل A و اگر H_1 درست باشد بهتر است عمل B را انجام دهد. در نتیجه برای خروج از تردید اقدام به جمع آوری داده و انجام دادن آزمون می‌کند.

به مثال‌های زیر توجه کنید:

- بیمار را تحت درمان قرار دهیم یا نه؟
- اجازه پخش فلان دارو را صادر کنیم یا نه؟
- نسبت به حفر چاه آب در محل مورد نظر اقدام کنیم یا نه؟
- محموله رسیده را بپذیریم یا نه؟

[۶] در نوع خاصی از آزمون‌های نیمن — پیرسون (آزمون نسبت درستنمایی) مینیمم کردن β، گر چه جز روش محسوب نمی‌شود ولی اغلب خود به خود تحقق می‌یابد.

در مثال‌های فوق محقق فرضیه H_0 (بیمار بودن شخص مورد نظر، بی ضرر بودن دارو، وجود آب، قابل قبول بودن محموله) را در مقابل نقیض آن (H_1) آزمون می‌کند. در صورت پذیرش H_0 عمل A (تحت درمان قرار دادن، اجازهٔ پخش دارو، حفر چاه، پذیرش محموله) و در صورت رد آن، عمل B (نقیض A) را انجام خواهد داد. در اینجا هدف محقق (که ما آن را هدف ۱ می‌نامیم) پاسخ به سوال زیر است:

سوال ۱ — "با توجه به داده‌ها کدامیک از دو عمل A و B را باید انجام داد؟"

در جهت مینیمم کردن ریسکی که از این طریق متوجه محقق است، محقق می‌خواهد احتمالات خطاهای نوع اول و دوم هر چه کمتر باشد.

این نکته را مجدداً تاکید می‌کنیم که ویژگی‌هایی که بهینه بودن آزمون‌های کلاسیک بر آن‌ها استوار است، یعنی اندازه و توان آزمون، قبل از انجام آزمون و حتی قبل از آنکه به جمع آوری داده‌ها اقدام کنیم، معلوم و مشخص‌اند و نتیجه این آزمون‌ها غیر از این‌که "H_0 رد شد" یا "H_0 پذیرفته شد" تفسیر دیگری ندارد. بنابراین هیچ یک از چهار تفسیری که در مقدمه ذکر شد در پارادیم نیمن — پیرسون وجود ندارد.

در اینجا باید به نکتهٔ مهم دیگری اشاره کنیم که نقش p–مقدار (که در خروجی $SPSS$ با علامت اختصاری sig نشان داده می‌شود) در جواب به سوال (۱) باید گفت که اولاً برای رسیدن به جواب نهایی آزمون (یعنی رد یا قبول H_0) از مراجعه به جداول آماری بی نیاز باشیم و ثانیاً کاربر بتواند با مقایسهٔ αی مورد نظر خود، آزمون را انجام دهد، به این ترتیب که اگر $p < \alpha$–مقدار باشد H_0 را رد و در غیر این صورت آن را بپذیرد. با توجه به دو نقل قول از نیمن، در جهت نیل به هدف (۱)، p–مقدار هیچ گونه تفسیری ندارد.

۳ احتمال درستی H_0 به عنوان هدف

در عمل موارد دیگری نیز ممکن است پیش آید. مثلاً ممکن است هدف از انجام آزمون آماری (که آن را هدف ۲ می‌نامیم) پاسخ به سوال زیر باشد:

سوال ۲ — با توجه به داده‌ها، احتمال درستی H_0 و (H_1) چقدر است؟

استفاده از آزمون‌های نیمن — پیرسون برای هدف (۲) نادرست یا دست کم ناکارآمد است، زیرا برای نیل به هدف (۲) باید احتمال درستی H_0 (یا H_1) را (از نظر محقق)، قبل از انجام آزمون و حتی قبل از جمع آوری داده‌ها، بدانیم. چون بنا بر قانون بیز داریم:

$$\begin{aligned} & P(H_0|AH_0) \\ &= \frac{P(AH_0|H_0)P(H_0)}{P(AH_0|H_0)P(H_0)+P(AH_0|H_1)P(H_1)} \\ &= \frac{(1-\alpha)P(H_0)}{(1-\alpha)P(H_0)+\beta P(H_1)} \end{aligned}$$

که در آن AH_0 به معنی پذیرش H_0 است.

ملاحظه می‌شود که برای یافتن احتمال درستی H_0 پس از انجام دادن آزمون، باید نسبت $O = \frac{P(H_0)}{P(H_1)}$ (بخت پیشین H_0 نسبت به H_1) را (در نزد محقق) بدانیم. در صورتی‌که در نظریهٔ نیمن — پیرسون جایی برای احتمالات پیشین وجود ندارد. از این گذشته حتی اگر بخت پیشین را برابر۱ (یعنی اگر $P(H_0) = \frac{1}{2}$) بگیریم هنوز استفاده از آزمون‌های نیمن —پیرسون برای هدف

(۲) ناکارآمد است، چون اگرآزمون را به‌صورت زیر تعریف کنیم:

$$\varphi(X)=\begin{cases}۰ & \text{پذیرش } H_۰\\ ۱ & \text{رد } H_۰\end{cases}$$

آن‌گاه داریم:

$$P(H_۰|RH_۰)=P(H_۰|\varphi(X)=۱)$$
$$P(H_۰|AH_۰)=P(H_۰|\varphi(X)=۰)$$

که در آن φ تابع بحرانی آزمون و $RH_۰$ به‌معنی رد فرض $AH_۰$ است. چون معمولاً $\varphi(X)$ (که فقط مقادیر ۰ و ۱ را می‌پذیرد) یک آماره بسنده نیست. کاهش داده‌ها از X به $\varphi(X)$ قابل توجیه نبوده و باعث عدم استفاده از کلیه اطلاعات موجود در داده‌ها می‌شود. به عبارت دیگر لازم است از $P(H_۰|X=x)$ استفاده شود نه از $P(H_۰|\varphi(X)=\varphi(x))$.

در هر صورت، نتیجه یک آزمون کلاسیک، چه به صورت رد یا قبول $H_۰$ باشد و چه به‌صورت ارائه p–مقدار، مقدار احتمال مطلوب در سوال ۲ را در بر ندارد.

در این جا لازم است به استدلال نادرست دیگری که متداول‌ترین نوع استفاده نادرست از آزمون‌های نیمن –پیرسون است، اشاره کنیم. چنین استدلال می‌شود که "چون اگر $H_۰$ درست باشد احتمال رد $H_۰$ (α) کوچک می‌باشد (مثلا ۰/۰۵)، وقتی $H_۰$ رد می‌شود، به احتمال زیاد $H_۰$ نادرست است."

استدلال فوق (که "قانون احتمال کم" نام دارد) تعمیم غیر مجاز "اصل عکس نقیض" در منطق است که می‌گوید: "اگر وقوع D در صورت درست بودن H امکان پذیر نباشد، آن‌گاه مشاهده D فرضیه H را باطل می‌کند". به‌عبارت دیگر

اگر $H\Rightarrow\sim D$ آنگاه $D\Rightarrow\sim H$

قانون احتمال کم (تعمیم نادرست اصل فوق) می‌گوید:

"اگر $P(D|H)=\varepsilon$ و ε عدد کوچکی باشد، آن‌گاه مشاهدهٔ D باعث می‌شود اطمینان زیادی به نادرست بودن H داشته باشیم."

"اطمینان زیاد" چیزی غیر از "احتمال زیاد" نیست (یا اگر هست، کسی تاکنون آن را تعریف نکرده است)، و در نتیجه بنا به آنچه در بخش ۲ ذکر شد، این تفسیر از درجه اعتبار ساقط است.

فقط زمانی که نتیجه آزمون مورد استفاده شخصی و فوری محقق است، اهداف (۱) و (۲) موجودیت پیدا می‌کنند و هدف ۲ بدون مولفه شخصی "احتمالات پیشین" و هدف ۱ بدون مولفه شخصی "زیان های نسبی دو عمل ذیربط" امکان وجود ندارند.

۴ شواهد آماری به عنوان هدف

در اکثر پژوهش‌هایی که نتایج آنها در مجلات علمی به چاپ می‌رسد، هدف محقق (که آن را هدف ۳ نامیم) پاسخ به سوال زیر است:

سوال ۳ – "داده‌ها از کدام فرضیه ($H_۰$ یا $H_۱$) بیشتر پشتیبانی می‌کنند و به چه میزان؟"

برای اثبات نادرستی قانون احتمال کم، حتی وقتی که هدف ۳ مورد نظر باشد، یعنی برای اینکه ببینیم صرفاً کوچک بودن $P(D|H_۰)$ شواهد قوی بر علیه $H_۰$ (وقتی D مشاهده می‌شود) نیست، به مثال‌های زیر توجه کنید.

مثال ۱ – اگر $X\sim b(۱۲۰,p)$, $H_۰:p=\frac{۱}{۲}$ و $D:X=۵۰$ باشد، آن‌گاه $P(X=۵۰|H_۰)=۰/۰۱۳$

ولی علیرغم کوچکی این احتمال، نه تنها مشاهده $X = ۵۰$ شاهدی بر نادرست بودن $H_۰$ نیست بلکه از آن حمایت نیز می‌کند.

مثال ۲ — اگر $X \sim N(\theta, ۱)$ ، $H_۰ : \theta = ۰$ و $D : X = ۰/۳۸۶$ باشد (که D یک مشاهده از توزیع $N(\theta, ۱)$ است که تا ۳ رقم اعشار گرد شده است)، آنگاه $P(D|H_۰) = ۰/۰۰۰۳۷۰$. چون این احتمال کوچک است، آیا می‌توان گفت مشاهده فوق شاهدی قوی برعلیه $H_۰$ است؟ مسلماً خیر.

مثال ۳ — فرض کنید یکی از دو جعبه A و B و به طریقی که بر ما مجهول است انتخاب می شود و سپس یک مهره از داخل آن به تصادف برداشته می شود. اگرمهره انتخاب شده قرمز باشد و بدانیم جعبه A، ۲ مهره قرمز و ۹۸ مهره سیاه دارد. آیا می توان گفت که قرمز بودن مهره انتخاب شده قویاً بر رد فرضیه "$H_۰$: مهره از جعبه A انتخاب شده است" دلالت می‌کند؟ مسلماً نمی‌توان به این سوال پاسخ داد مگر این‌که بدانیم در جعبه B ترکیب مهره‌ها چیست.

اگر B دارای ۵۰ مهره سفید و ۵۰ مهره قرمز باشد، البته قرمز بودن مهره انتخاب شده، شاهدی قوی بر علیه $H_۰$ است. ولی اگر جعبه B دارای ۱ مهره قرمز و ۹۹ مهره سفید باشد، مسلماً قرمز بودن مهره انتخاب شده نه تنها شاهدی بر علیه $H_۰$ (در مقابل $H_۱$) نیست بلکه از آن حمایت نیز می‌کند. اگر جعبه B دارای هیچ مهره قرمز نباشد، شک نیست که قرمز بودن مهره انتخاب شده، فرضیهٔ $H_۰$ را نه تنها رد نمی‌کند بلکه اثبات می‌کند!

۵ نسبت درستنمایی

از مثال‌های فوق می‌توان چنین نتیجه گرفت که برای اینکه D شاهدی قوی بر علیه $H_۰$ باشد صرفاً کوچک بودن $P(D|H_۰)$ کافی نیست بلکه باید دید که $P(D|H_۱)$ چقدر است. $H_۱$ فرضیه‌ای است که قوت D را به عنوان شاهدی بر علیه $H_۰$ در مقابل آن باید سنجید. $H_۱$ فرضیه جانشین گفته می‌شود.

می‌توان گفت که میزان پشتیبانی مشاهده D از فرضیه $H_۰$ فقط در مقابل یک فرضیه دیگری مانند $H_۱$ معنی دارد و شدت آن را می‌توان با مقدار

$$r = \frac{P(D|H_۰)}{P(D|H_۱)}$$

که برابر با نسبت درستنمایی است، اندازه‌گیری کرد.

در مثال ۱ هیچ فرضیه جانشینی وجود ندارد که به ازای آن $D : X = ۵۰$ شاهدی قوی بر علیه $H_۰$ باشد.

در مثال ۲ اگر $H_۱ : \theta = a$ فرضیه جانشین باشد آن‌گاه $P(D|H_۱)$ به ازای $a = ۰/۱ \ , \ ۱ \ , \ ۲$ به‌ترتیب عبارت است از

$$۰/۰۰۰۳۸۹ \quad , \quad ۰/۰۰۰۳۳ \quad , \quad ۰/۰۰۰۰۱۰۸$$

در نتیجه نسبت درستنمایی به‌ازای سه مقدار فوق برای a به‌ترتیب برابر است با

$$r = \frac{P(D|H_۰)}{P(D|H_۱)} = ۰/۹۵ \ , \ ۱/۱۱, \ ۳/۴۲$$

یعنی میزان پشتیبانی مشاهده $x = ۰/۳۸۶$ از $H_۰$، شدیداً به مقدار a (یعنی به فرضیه جانشین) بستگی دارد.

حال با توجه به مطالب فوق اگر بخواهیم از نتیجه آزمون‌های نیمن — پیرسون استفادهٔ شواهدی بکنیم، یعنی

بخواهیم بر اساس رد یا قبول H_0، میزان پشتیبانی داده‌ها از H_0 در مقابل H_1 را معین کنیم، نه تنها مقدار α بلکه مقدار β را نیز باید بدانیم یعنی صرفاً به خاطر کوچکی α، رد H_0 را نمی‌توان شاهدی قوی بر علیه H_0 یا پذیرش H_0 را شاهدی قوی بر علیه H_1 تلقی کرد. یعنی تفسیر درست نتیجه آزمون، در جهت نیل به هدف ۳، به مقدار β نیز بستگی دارد. اگر از نسبت درستنمایی به عنوان میزان پشتیبانی از یک فرضیه، در مقابل فرضیهٔ دیگر استفاده کنیم، پشتیبانی داده‌ها از فرضیهٔ H_1 در مقابل H_0 زمانی‌که H_0 رد می‌شود برابر است با

$$r_1 = \frac{P(RH_0|H_1)}{P(RH_0|H_0)} = \frac{1-\beta}{\alpha}$$

و میزان پشتیبانی داده‌ها از H_0 در مقابل H_1 زمانی‌که H_0 پذیرفته می‌شود برابر است با

$$r_0 = \frac{P(AH_0|H_0)}{P(AH_0|H_1)} = \frac{1-\alpha}{\beta}$$

جدول (۱) مقادیر مختلف r_1 و r_0 را به‌ازای $\alpha = 5\%$ و مقادیر مختلف β نشان می‌دهد. همان‌طورکه از جدول (۱) ملاحظه می‌شود زمانی‌که H_0 پذیرفته می‌شود، تفسیر نتیجهٔ حاصل (نسبت به زمانی‌که H_0 رد می‌شود)، خیلی بیشتر متاثر از مقدار β است. به‌عبارت دیگر زمانی‌که H_0 پذیرفته می‌شود در گفتن "داده‌ها از H_1 قویا پشتیبانی می‌کنند" باید احتیاط بیشتری کرد، زیرا وقتی $\alpha = 0/05$ است، شواهد بر له H_0 می تواند از ۱/۹ (به ازای $\beta = 0/5$) تا ∞ (پشتیبانی بسیار بسیار قوی) تغییر کند.

اما زمانی که H_0 رد می‌شود میزان پشتیبانی بر علیه H_0 از ۱۰ (به ازای $\beta = 0/5$) تا ۲۰ (به ازای $\beta = 0$) تغییر می‌کند. که در همه موارد نسبت درستنمایی بر علیه H_0 نسبتاً قابل توجه است.

پس مطالب فوق را می توان به‌صورت زیر خلاصه کرد:

الف) وقتی که فرضیه صفر رد می‌شود (و α کوچک است) شواهد قوی بر علیه H_0 وجود دارد (و صحت این تفسیر خیلی به مقدار β بستگی ندارد)

ب) زمانی‌که فرضیه صفر پذیرفته می‌شود، شدت شواهد بر علیه H_0 (حتی وقتی α کوچک است) شدیداً به مقدار β بستگی دارد. پس اگر از مقدار β اطلاع نداشته باشیم بهتر است به گفتن این‌که "شواهد موجود در داده‌ها بر علیه H_0 قاطع نیست" اکتفا کنیم. در اینجا لازم است متذکر شویم که هدف این بخش از مقاله این بوده است که با دانستن نتیجه یک آزمون کلاسیک (رد یا قبول H_0) چگونه میزان پشتیبانی داده‌ها را فرضیه‌ایی که پذیرفته شده است، معین کنیم. اما از زمانی که محاسبات سریع و ارزان امکان پذیر شده است، چنین معمول و متداول شده است، که استفاده از آزمون‌های کلاسیک در جهت نیل به هدف ۳ از طریق p–مقدار صورت می‌گیرد.

در مقاله بعدی از این دنباله، به مشروعیت و مناسبت استفاده از استنباط شواهدی از p–مقدار خواهیم پرداخت و اکنون فقط به ذکر این نکته اکتفا می کنیم که p–مقدار نیز معمولاً یک آماره بسنده نیست و اگر قرار است از داده‌ها استنباط شواهدی به عمل آید، بهتر است برمبنای قانون درستنمایی که ذیلاً بیان می‌شود، عمل کنیم.

استفاده از نسبت درستنمایی به عنوان میزان پشتیبانی داده‌ها از یک فرضیه در مقابل فرضیه دیگر "قانون درستنمایی" نام دارد که اولین بار توسط هاکینگ [۴] ارائه

شد و به‌صورت زیر بیان می‌شود.

قانون درستنمایی:[۴] فرض کنید متغیر تصادفی X تحت فرضیهٔ H_0 دارای توزیع با چگالی $f_{H_0}(x)$ و تحت فرضیهٔ H_1 دارای توزیعی با تابع چگالی $f_{H_1}(x)$ است. مشاهدهٔ $X = x$ از H_0 در مقابل H_1 پشتیبانی می کند اگر $f_{H_0}(x) > f_{H_1}(x)$ باشد و شدت این پشتیبانی برابر نسبت درستنمایی یعنی $r = \frac{f_{H_0}(x)}{f_{H_1}(x)}$ است.

با توجه به قانون درستنمایی، استفاده از نتیجه آزمون‌های نیمن —پیرسون برای محاسبه میزان پشتیبانی داده‌ها از H_0 در مقابل H_1 بهترین روش و حتی روش صحیح نیست و ابتدایی‌ترین دلیل آن این است که نه آماره $\varphi(X)$ و نه آماره $-p(X)$ مقدار هیچ‌کدام بسنده نیستند و استفاده از آن‌ها به عدم استفاده کامل از اطلاعات می‌انجامد.

٦ نتیجه گیری

۱ — بدون دانستن احتمالات پیشین برای H_0 و H_1 استفاده از عبارت‌های "به احتمال ۰/۹۵"، "به احتمال زیاد"، "کاملاً مطمئن هستیم" و امثال آن، در تفسیر نتایج آزمون‌های کلاسیک مجاز نیست.

۲ — اگر مقادیر احتمالات پیشین $P(H_0)$ و $P(H_1)$ معلوم باشند و علیرغم ذهنی بودن آن‌ها بخواهیم از آن‌ها استفاده کنیم، محاسبه "احتمال درستی H_0" پس از اطلاع یافتن از "رد یا پذیرش H_0" در یک آزمون نیمن — پیرسون امکان پذیر است ولی استفاده کافی از اطلاعات موجود در داده‌ها به عمل نمی آید.

۳ — اگر بخواهیم از دانستن رد یا قبول H_0، میزان پشتیبانی داده‌ها را از فرضیهٔ ساده H_0 در مقابل H_1 معین کنیم، لازم است علاوه بر مقدار α مقدار β را نیز بدانیم. در مقاله دوم به معایب استفاده از p–مقدار به عنوان "میزان شواهد آماری" موجود در داده‌ها(آزمون‌های معنی‌داری) خواهیم پرداخت. آزمون‌های معنی‌داری (که ترجمه درست آن اندازه‌گیری معنی‌داری است.) در واقع نشان می‌دهد که نتیجه آزمون تا چه حد معنی‌دار است. این عمل نباید با آزمون فرضیه‌های آماری، که نتیجه آن یکی از دو عمل "رد H_0" یا "پذیرش H_0" است، اشتباه شود. نتیجه یک "آزمون معنی‌داری" یک عدد بین صفر و یک است که همانا p–مقدار حاصل از محاسبات می‌باشد. چون هر چه p–مقدار کوچکتر باشد معنی‌داری آزمون بیشتر است، محاسبه و ارائه p–مقدار به نام آزمون معنی‌داری معروف شده است.

جدول ۱. مقادیر مختلف r_1 و r_0 به‌ازای $\alpha = ۰/۰۵$ و مقادیر مختلف β

β	۰	۰/۰۱	۰/۰۵	۰/۱	۰/۲	۰/۵	۰/۹۵
r_1	۱۲۰	۱۹۸	۱۹	۱۸	۱۶	۱۰	۱
r_0	∞	۹۵	۱۹	۹/۵	۴/۷۵	۱/۹	۱

در مقاله دوم همچنین در مورد قانون درستنمایی و آنچه که "استنباط شواهدی" نام دارد، توضیح بیشتری خواهیم داد. تعمیم نتایج مقاله دوم به حالت فرضیه‌های مرکب در مقاله سوم به انجام خواهد رسید. خواننده برای اطلاعات بیشتر می‌تواند به فرهادی[۱]، بلوم[۲]، عمادی[۳] و رویال[۴, ۵, ۶, ۷, ۸] مراجعه کند.

مراجع

[۱] فرهادی، ا. (۱۳۸۵)، رهیافت استنباط شواهدی جهت تعیین حجم نمونه در آزمون فرضیه‌های بیزی، پایان نامه کارشناسی ارشد، دانشگاه فردوسی مشهد.

[2] Blume, J. D. (2002), Tutorial In Biostatistics: likelihood methods for measuring statistical evidence, *Journal of Statistics in Medicine*, 21, 2563-2599.

[3] Emadi, M., Ahmadi, J. and Arghami, N. R. (2007), Comparison of record data and random observations based on statististical evidence, *Statistical Papers*, 48, 1-25.

[4] Hacking, I. (1965), *Logic of Statistical Inference*, Cambridge University Press, New York.

[5] Royall, R. (1992), The elusive concept of statistical evidence, *Bayesian Statistics*, 4, 405-418.

[6] Royall, R. (1997), *Statistical Evidence: A Likelihood Paradigm*, Chapman & Hall, New York.

[7] Royall. R. (2000), On the probability of observing misleading statistical evidence, *Journal of American Statisitcal Association*, 45(451), 760-769.

[8] Royall, R. (2003), Interpreting statististical evidence by using imperfect models: robust adjusted likelihood functions, *Journal of Royal Statistical Society*, B.65(2), 391-404.

[9] Neyman, J. (1950), *First course in probability and statistics*, Henry and Holt Company, New York.

تشکر و قدردانی: مولفین مراتب تشکر خود را از داوران محترم مقاله به خاطر پیشنهادات ارزنده‌شان که در بهبود مقاله بسیار موثر بوده است، ابراز می دارند.