# شواهد آماری[۱]
# قسمت دوم[۲] : قانون درستنمایی و معایب p—مقدار به‌عنوان معیار پشتیبانی داده‌ها از فرضیه صفر

ناصررضا ارقامی[۳] ، علی دست برآورده [۴] ، احسان زمان زاده[۵]

چکیده:

همان‌گونه که در مقاله اول[۱] به آن اشاره شد، در این مقاله ابتدا عدم مشروعیت و مناسبت استفاده شواهدی از p—مقدار مورد بررسی قرار می‌گیرد. سپس بر مبنای قانون درستنمایی به معرفی معیاری مناسب و منطقی به منظور استفاده شواهدی از داده‌ها پرداخته می‌شود.

**واژه‌های کلیدی:** شواهد آماری، معیار پشتیبانی، p—مقدار، قانون درستنمایی.

## ۱ مقدمه

p—مقدار یکی از پرکاربردترین مفاهیم در تحلیل آماری است. تقریباً در همه نرم افزارهای تحلیل آماری برای ارائه نتایج آزمون فرضیه‌ها از p—مقدار استفاده می‌کنند. با وجود قاعده کلی و ساده‌ای که p—مقدار در انجام آزمون فرضیه‌ها ارائه می‌دهد، کاربران غیرمتخصص، اغلب آن را مورد استفاده و تفسیر نادرست قرار می‌دهند. در این مقاله ضمن ارائه تعریفی دقیق و کاملاً کلی از p—مقدار، معایب استفاده از p—مقدار به‌عنوان معیار پشتیبانی داده‌ها از فرضیه صفر مورد بررسی قرار می‌گیرد. سپس بر مبنای قانون درستنمایی به معرفی معیاری مناسب و منطقی به‌منظور استفاده شواهدی از داده‌ها پرداخته می‌شود.

## ۲ تعریف p—مقدار

**تعریف ۱** فرض کنید $X \sim P_\theta$، $\theta \in \Theta$ و $C$ کلاسی از آزمون‌ها(توابع آزمون) برای آزمون فرضیه $H_0 : \theta \in \Theta_0$ در مقابل فرضیه $H_1 : \theta \in \Theta_1$ باشد. در این‌صورت کلاس $C$ را کامل[۶] گوییم، هرگاه در شرط زیر صدق کند:

$$\forall\, \alpha \in [0,1] \quad \exists\, \varphi_\alpha \in C \quad s.t.\ \alpha = \sup_{\theta \in \Theta_0} E_\theta\left(\varphi_\alpha\left(X\right)\right).$$

---

[۱] Statistical Evidence
[۲] مقاله دوم از سری مقاله‌هایی که طی آنها روش درست اندازه‌گیری میزان شواهد موجود در داده‌ها مورد بررسی قرار می‌گیرد.
[۳] استاد گروه آمار دانشگاه فردوسی
[۴] دانشجوی دکتری آمار دانشگاه فردوسی مشهد
[۵] دانشجوی دکتری آمار دانشگاه فردوسی مشهد
[۶] Complete

به‌عبارت دیگر، یک کلاس از آزمون‌ها را کامل می‌گوییم، هرگاه برای هر $\circ \leq \alpha \leq ۱$ حداقل یک آزمون در اندازه $\alpha$ در این کلاس موجود باشد.

**تعریف ۲** فرض کنید $X \sim P_\theta$، $\theta \in \Theta$ و $C$ کلاسی از آزمون‌ها برای فرضیه $H_\circ : \theta \in \Theta_\circ$ در مقابل فرضیه $H_۱ : \theta \in \Theta_۱$ باشد. در این‌صورت، کلاس $C$ را تو در تو[۷] می‌گوییم، هرگاه در شرط زیر صدق کند:

$$\alpha_۱ < \alpha_۲ \ , \ \varphi_{\alpha_۱}, \varphi_{\alpha_۲} \in C \ \Rightarrow \ \forall\, x \in S_X \quad \varphi_{\alpha_۱}(x) \leq \varphi_{\alpha_۲}(x)\,.$$

که در آن $S_X$ تکیه‌گاه متغیر تصادفی $X$ است.
در حالت خاص، اگر $C$ کلاسی از آزمون‌های غیر تصادفی باشد، آنگاه به‌جای شرط فوق به‌طور معادل می‌توان نوشت:

$$\alpha_۱ < \alpha_۲ \ \Rightarrow \ R_{\alpha_۱} \subseteq R_{\alpha_۲}$$

که در آن $R_\alpha$ ناحیه رد آزمون در اندازه $\alpha$ است.

ثابت می‌شود کلاس آزمون‌های MP که در آن توزیع $f_۱(x)/f_\circ(x)$ تحت فرضیه ساده $H_\circ$ پیوسته است همواره تو در تو خواهد بود. اما، اگر توزیع $f_۱(x)/f_\circ(x)$ تحت فرضیه $H_\circ$ پیوسته نباشد لزومی ندارد که حتی کلاس آزمون‌های MP نیز تو در تو باشد. اما، در چارچوب لم نیمن—پیرسون همواره زیرکلاسی از آزمون‌های MP وجود دارد که تو در تو باشد [۵].

**تعریف ۳** فرض کنید $X \sim P_\theta$، $\theta \in \Theta$ و $C$ کلاسی از آزمون‌های تو در تو برای فرضیه $H_\circ : \theta \in \Theta_\circ$ در مقابل $H_۱ : \theta \in \Theta_۱$ باشد، آنگاه p—مقدار به‌ازای مقدار مشاهده شده $x$ بر اساس کلاس $C$ عبارت است از:

$$\hat{p}(x) = \inf\left\{\alpha \mid E_\theta(\varphi(x)) \leq \alpha \ \ \forall\, \theta \in \Theta_\circ \quad , \quad \varphi(x) = ۱ \quad , \quad \varphi \in C\right\}.$$

و در حالتی که برای مشاهده $x$، مجموعه

$$A = \{\alpha \mid E_\theta(\varphi(X)) \leq \alpha \ \ \forall\, \theta \in \Theta_\circ, \quad \varphi(x) = ۱,\ \varphi \in C\}. \quad (۱)$$

تهی باشد تعریف می‌کنیم:

$$\hat{p}(x) = ۱.$$

به‌عبارت دیگر، p—مقدار کوچک‌ترین سطح آزمونی است که در آن مشاهده حاصل از نمونه با احتمال یک منجر به رد فرضیه صفر می‌شود.

**نکته ۱** با توجه به این‌که اندازه‌ی آزمون کوچک‌ترین سطح آزمون است، اگر به‌جای سطح آزمون از اندازه آزمون جهت محاسبه p—مقدار استفاده شود تغییری در p—مقدار حاصل به‌وجود نمی آید. به‌عبارت دیگر، p—مقدار را به‌طور معادل می‌توان براساس اندازه آزمون تعریف کرد؛ یعنی

$$\hat{p}(x) = \inf\{\alpha \mid \alpha = \sup_{\theta \in \Theta_\circ} E_\theta(\varphi_\alpha(X))\ , \quad \varphi_\alpha(x) = ۱,\ \varphi_\alpha \in C\}.$$

---
[۷] Nested

**نکته ۲** گرچه در تعریف p—مقدار، کامل نبودن کلاس آزمون‌ها اشکال منطقی (جهت ارائه نتیجه آزمون) ایجاد نمی‌کند؛ اما، منطقی‌تر این است که برای محاسبه p—مقدار، کلاس آزمون‌ها کامل باشد. علاوه بر این، اگر کلاس آزمون‌ها کامل باشد، آنگاه مجموعه A در رابطه (۱) هیچ‌گاه تهی نخواهد بود[۸].

**نکته ۳** پرسش اساسی که مطرح می‌شود این است که چرا تو در تو بودن کلاس آزمون‌ها برای تعریف p—مقدار الزامی است. در جواب می‌توان این‌گونه بیان کرد: به صرف اینکه p—مقدار کوچک‌ترین سطح آزمونی است که در آن مشاهده حاصل از نمونه با احتمال یک منجر به رد فرضیه صفر می‌شود، نمی‌توان مطمئن بود که برای یک سطح آزمون بزرگ‌تر از p—مقدار نیز الزاماً فرضیه صفر با احتمال یک رد شود. اما تودرتو بودن کلاس آزمون‌ها این مطلب را تضمین می‌کند. برای درک بهتر موضوع به مثال زیر توجه کنید. در این مثال، تناقضی که عدم تو در تو بودن کلاس آزمون‌ها موجب می‌شود به‌خوبی نشان داده شده است.

**مثال ۱** فرض کنید $X$ متغیر تصادفی باشد که مقادیر ۱، ۲، و ۳ را با احتمال‌های ۰/۸۵، ۰/۱۰، و ۰/۰۵ تحت فرضیه $H_0$ و احتمال‌های ۰/۷، ۰/۲، و ۰/۱ را تحت فرضیه $H_1$ اختیار می‌کند. براساس کلیه آزمون‌های MP برای مشاهده $x = ۳$، p—مقدار آزمون عبارت است از

$$\hat{p}(۳) = \inf\{\alpha | \varphi_\alpha(۳) = ۱\} = ۰/۰۵$$

در نتیجه، بر مبنای p—مقدار از آن‌جا که $\hat{p}(۳) < ۰/۱$، با مشاهده $x = ۳$ فرضیه صفر در سطح معنی‌داری ۰/۱ رد می‌شود، در حالی که بر مبنای آزمون MP ی

$$\varphi_{۰/۱}(x) = \begin{cases} ۱ & x = ۲ \\ ۰ & x = ۱, ۳ \end{cases}$$

مشاهده $x = ۳$ در سطح معنی‌داری ۰/۱ منجر به رد فرضیه صفر نمی‌شود.

## ۳ p—مقدار به‌عنوان معیار پشتیبانی داده‌ها از فرضیه صفر

با توجه به این‌که رد فرضیه صفر در سطح معنی‌داری کوچک‌تر حاکی از شواهد متقاعدکننده‌تری بر نادرستی فرضیه صفر می‌باشد، بنابراین p—مقدار کوچک‌تر حاکی از شواهد قوی‌تری بر نادرستی فرضیه صفر خواهد بود. لذا می‌توان p—مقدار را به‌عنوان معیاری برای پشتیبانی داده‌ها از فرضیه صفر درنظر گرفت (برای شرح بیشتر به مرجع [۳] می‌توان مراجعه کرد)؛ هر چند در ادامه نشان خواهیم داد که p—مقدار برای این کار ابزار مناسبی نیست.

## ۴ معایب p—مقدار به‌عنوان معیار پشتیبانی داده‌ها

اشکالاتی که بر p—مقدار (به‌عنوان معیار پشتیبانی از فرضیه‌ی صفر در مقابل فرضیه جانشین) وارد است، عبارتند از:

[۸] در ادامه هرگاه سخن از p—مقدار شود منظور p—مقداری است که براساس کلاس کاملی از آزمون‌های تو در تو به دست آمده است

۱) اتکای p—مقدار به قانون نادرست احتمال کم

در مقاله قبل به تفصیل به بیان قانون احتمال کم و چرایی نادرستی آن پرداختیم، اولین ایرادی که به p—مقدار وارد است اتکای این معیار به قانون احتمال کم می‌باشد. برای درک بهتر این مطلب، مثال زیر را در نظر گیرید.

**مثال ۲** فرض کنید $x = ۱۶$ یک تک مشاهده از متغیر تصادفی $X$ با توزیع دوجمله‌ای $bin(۴۰, p)$ باشد. می‌خواهیم میزان پشتیبانی داده‌ها از فرضیه $H_۰ : p = \frac{۱}{۴}$ را در مقابل فرضیه $H_۱ : p = \frac{۳}{۴}$ تعیین کنیم. p—مقدار حاصل عبارت است از:

$$\hat{p}(۱۶) = P\left(X \geq ۱۶ | p = \frac{۱}{۴}\right) \approx ۰/۰۲۶$$

حال اگر بخواهیم از p—مقدار فوق استفاده شواهدی بکنیم به این نتیجه می‌رسیم که مشاهده $x = ۱۶$ قویاً از فرضیه $H_۱$ پشتیبانی می‌کند. اما به‌طور منطقی و شهودی انتظار می‌رود که مشاهدات کمتر از ۲۰ از فرضیه $H_۰$ بیشتر از فرضیه $H_۱$ پشتیبانی کنند، که p—مقدار خلاف این مطلب را نشان می‌دهد. همچنین اگر در این مثال نسبت احتمال مشاهده $x = ۱۶$ تحت فرضیه $H_۰$ به احتمال این مشاهده تحت فرضیه $H_۱$ (نسبت درستنمایی) را محاسبه کنیم، خواهیم داشت:

$$\frac{P(X = ۱۶ | H_۰)}{P(X = ۱۶ | H_۱)} = ۳^۸$$

همان‌طور که ملاحظه می‌شود با وجود اینکه احتمال رخ دادن مشاهده $x = ۱۶$ تحت فرضیه $H_۰$ بسیار کم است اما در مقایسه با فرضیه $H_۱$ بسیار محتمل‌تر است. لذا این مشاهده از فرضیه $H_۰$ پشتیبانی بسیار بیشتری می‌کند (نادرستی قانون احتمال کم)(برای شرح بیشتر به مرجع [۸] مراجعه شود).

۲) وابستگی به فضای نمونه (عدم وابستگی صِرف به داده‌های مشاهده شده)

یکی از ویژگی‌های مطلوب یک معیار پشتیبانی عدم وابستگی آن به فضای نمونه است. به‌عبارت دیگر، معیار پشتیبانی باید تنها به داده‌های مشاهده شده بستگی داشته باشد و نه به داده‌هایی که مشاهده نشده‌اند. p—مقدار به‌عنوان یک معیار پشتیبانی فاقد این ویژگی مهم است.

**مثال ۳** فرض کنید دستگاهی داریم که فشار هوا در آن دارای توزیع نرمال با میانگین $\theta$ و واریانس ۱۰۰ است و می‌خواهیم فرضیه $H_۰ : \theta = ۸۰$ را در مقابل $H_۱ : \theta = ۸۵$ آزمون کنیم. برای سنجش فشار هوای دستگاه از فشارسنجی استفاده می‌کنیم که در فشار بالاتر از ۱۰۰ از کار می‌افتد. در این‌صورت، فشار هوایی که فشارسنج نشان می‌دهد دارای توزیع نرمال $N(\theta, ۱۰۰)$، بریده شده در نقطه ۱۰۰ خواهد بود. حال فرض کنید $x = ۹۰$ را مشاهده کرده‌ایم. بنابراین p—مقدار عبارت است از

$$\hat{p}(۹۰) = P_{H_۰}(X \geq ۹۰) = \frac{\Phi(۲) - \Phi(۱)}{\Phi(۲)}.$$

در حالی‌که اگر ما از خراب بودن فشارسنج مطلع نباشیم، p−مقدار به‌صورت

$$\hat{p}(90) = P_{H_0}(X \geq 90) = 1 - \Phi(1).$$

محاسبه می‌شود. در این مثال بدیهی است که تفسیر مشاهده $x = 90$، نباید تحت تأثیر خرابی دستگاه باشد؛ چون $100 > 90$ است. ولی همان‌طور که مشاهده شد p−مقدار به خرابی دستگاه بستگی دارد و این نشان می‌دهد که p−مقدار به فضای نمونه بستگی دارد.

۳) بستگی داشتن به قاعده توقف نمونه‌گیری

به‌طور کلی استنباط ما در مورد جامعه نباید به قاعده توقف نمونه‌گیری[9] بستگی داشته باشد، برای درک بهتر حالتی را در نظر گیرید که متغیرهای تصادفی مستقل و هم‌توزیع پی‌درپی مشاهده می‌شوند، آیا اگر این متغیرها بدانند که چه برنامه‌ای برای متوقف کردن طرح نمونه‌گیری داریم، مقادیرشان را تغییر می‌دهند؟ مسلماً خیر. پس نوع و مقدار اطلاعی که در مقادیر مشاهده شده وجود دارد به قاعده توقف بستگی ندارد و بنابراین استنباط ما نیز نباید به قاعده توقف بستگی داشته باشد (برای شرح بیشتر به مرجع [۸] مراجعه شود).

**مثال ۴** فرض کنید که $\theta$ نسبت یک ویژگی در جامعه باشد. همچنین فرض کنید می‌خواهیم میزان پشتیبانی داده‌ها را از فرضیه $H_0 : p = \frac{1}{2}$ در مقابل فرضیه $H_1 : p = \frac{3}{4}$ تعیین کنیم. فرض کنید در یک تحقیق از ۱۴ مورد مشاهده شده، ٦ مورد ویژگی مورد نظر را دارند. ابتدا فرض کنید که این مشاهده‌ها چنین به‌دست آمده که از یک نمونه تصادفی انتخاب شده به حجم ۱۴ از جامعه، ٦ مورد دارای ویژگی مورد نظر هستند، در این‌صورت اگر متغیر تصادفی $X$ تعداد واحدهایی از نمونه باشند که دارای ویژگی مورد نظراند، آنگاه $X$ دارای توزیع دوجمله‌ای $bin(14, \theta)$ است. بنابراین میزان پشتیبانی داده‌ها از فرضیه $H_0$ در مقابل فرضیه $H_1$ بر اساس معیار p−مقدار عبارت است از:

$$\hat{p}(6) = P_{H_0}(X \geq 6) \approx 0/788$$

بار دیگر فرض کنید که نمونه‌گیری از جامعه را تا آن‌جا ادامه داده‌ایم که ٦ مورد از واحدهای نمونه دارای ویژگی موردنظر باشند و چهاردهمین نمونه به دست آمده، ششمین نمونه دارای ویژگی موردنظر است؛ در این‌صورت، متغیر تصادفی $Y$، تعداد نمونه‌های لازم تا مشاهده ٦ مورد دارای ویژگی موردنظر، دارای توزیع دوجمله‌ای منفی $Nbin(6, \theta)$ است. بنابراین میزان پشتیبانی داده‌ها از فرضیه $H_0$ در مقابل فرضیه $H_1$ بر اساس معیار p−مقدار عبارت است از:

$$\hat{p}(14) = P_{H_0}(Y \geq 14) \approx 0/29$$

همان‌طور که ملاحظه می‌شود میزان پشتیبانی داده‌ها از فرضیه $H_0$ در مقابل فرضیه $H_1$ بر اساس

[9] Stopping rule

معیار p—مقدار در دو آزمایش فوق با یکدیگر متفاوت است؛ در صورتی که داده‌های مشاهده شده در هر دو مورد یکی است. و این نشان می‌دهد که p—مقدار به قاعده توقف نمونه‌گیری بستگی دارد.

۴) صدق نکردن در اصل درستنمایی
اصل درستنمایی[10] که مورد قبول اغلب قریب به اتفاق آماردانان، به‌عنوان یک اصل در استنباط آماری است، بیان می‌کند:

فرض کنید که $L_1(\theta)$ تابع درستنمایی حاصل از مشاهده $x$ از آزمایش $E_1$ و $L_2(\theta)$ تابع درستنمایی حاصل از مشاهده $y$ از آزمایش $E_2$ باشد که $\theta$ پارامتر مشترک در هر دو آزمایش است. اگر $L_1(\theta)/L_2(\theta)$ به $\theta$ بستگی نداشته باشد، آنگاه هرگونه استنباط در مورد $\theta$ از $x$ و $y$ باید یکسان باشد [۳].

به‌عبارت دیگر، اگر $L_1(\theta)/L_2(\theta)$ به $\theta$ بستگی نداشته باشد، در حقیقت میزان اطلاع مشاهده‌های $x$ و $y$ در مورد $\theta$ یکسان است. بنابراین منطقی به‌نظر می‌رسد که استنباط بر اساس $x$ و $y$ برای $\theta$ یکسان باشد.

اما همان‌طور که در مثال قبل دیدیم با اینکه $x = ۶$ و $y = ۱۴$ در دو طرح نمونه‌گیری از نظر درستنمایی با یکدیگر معادل هستند اما، p—مقدار حاصل از آنها با یکدیگر تفاوت دارد. بنابراین p—مقدار در اصل درستنمایی صدق نمی‌کند.

به‌طور کلی روش‌هایی که به فضای نمونه و یا قاعده توقف بستگی دارند در اصل درستنمایی صدق نمی‌کنند.

۵) فرضیه جانشین در p—مقدار نقش ثانویه دارد.
فرضیه جانشین در p—مقدار تنها در جهت تعیین شواهد آماری ایفای نقش می‌کند و تأثیری در تعیین شدت شواهد آماری ندارد.

**مثال ۵** فرض کنید که $x = ۰/۷۵$ یک تک مشاهده از توزیع نرمال $N(\theta, ۱)$ باشد، در این‌صورت منطقی به‌نظر می‌رسد که p—مقدار فرضیه $H_0 : \theta = ۰$ در مقابل $H_1 : \theta = ۱۰۰$ خیلی بیشتر از p—مقدار فرضیه $H_0 : \theta = ۰$ در مقابل $H_1 : \theta = ۱$ باشد، اما p—مقدار این دو آزمون مساوی و برابر است با

$$\hat{p}(۰/۷۵) = P\left(X > ۰/۷۵ | \theta = ۰\right) \approx ۰/۲۲۷.$$

٦) p—مقدار نسبت به $H_0$ و $H_1$ متقارن نیست. یکی دیگر از ویژگی‌های مطلوب یک معیار پشتیبانی متقارن بودن آن نسبت به فرضیه‌های $H_0$ و $H_1$ است؛ به این معنی که با عوض شدن نقش فرضیه‌ها به‌عنوان فرضیه صفر یا فرضیه جانشین، میزان پشتیبانی داده‌ها از هر فرضیه تغییر نکند. بنابراین انتظار می‌رود که اگر $\hat{p}_{01}$، p—مقدار حاصل از آزمون فرضیه $H_0$ در مقابل $H_1$ و $\hat{p}_{10}$، p— مقدار حاصل از آزمون فرضیه $H_1$ در مقابل

[10] Law of likelihood

$H_0$ باشد آنگاه، $\hat{p}_{10} = 1 - \hat{p}_{01}$[11]. متأسفانه p—مقدار در این شرط صدق نمی‌کند.

**مثال ۶** فرض کنید که $X \sim N(\theta, 1)$، $H_0 : \theta = 0$ و $H_1 : \theta = 1$ باشد. در این‌صورت، برای مشاهده داده شده $x$،

$$\hat{p}_{01} = P(X \geq x | \theta = 0) = 1 - \Phi(x)$$

در صورتی که

$$\hat{p}_{10} = P(X \leq x | \theta = 1) = \Phi(1 - x)$$

و البته بدیهی است که $\Phi(1 - x) \neq 1 - \Phi(x)$.

۷) p—مقدار منسجم نیست.

یکی از ویژگی‌های مطلوب یک معیار پشتیبانی انسجام[12] آن است؛ به این معنی که اگر فرضیه $H'_0$ فرضیه $H_0$ را در بر گیرد ($H_0 \subset H'_0$)، آنگاه میزان پشتیبانی از فرضیه $H'_0$ حداقل به میزان پشتیبانی از فرضیه $H_0$ باشد. اما p—مقدار به‌عنوان یک معیار پشتیبانی فاقد این ویژگی است [۱۰].

**مثال ۷** فرض کنید $X \sim N(\theta, 1)$. همچنین فرض کنید $H_0 : \theta = \theta_0$ و $H'_0 : \theta \leq \theta_0$. فرضیه جانشین $H_0$ و $H'_0$ را به ترتیب به‌صورت $H_1 : \theta \neq \theta_0$ و $H'_1 : \theta > \theta_0$ در نظر گیرید. p—مقدار در حمایت از فرضیه $H$ را برای مقدار مشاهده شده $x$ با $\hat{p}_H(x)$ نمایش می‌دهیم. فرض کنید $x > \theta_0$ مشاهده شده باشد. در این‌صورت

$$\hat{p}_{H_0}(x) = 2\Phi(\theta_0 - x)$$

$$\hat{p}_{H'_0}(x) = \Phi(\theta_0 - x).$$

که در آن $\Phi$ تابع توزیع نرمال استاندارد است. بنابراین با وجود این که $H_0 \subset H'_0$ است اما $\hat{p}_{H_0}(x) \geq \hat{p}_{H'_0}(x)$.

۸) وقتی $H_1$ به $H_0$ میل می‌کند، p—مقدار تغییر نمی‌کند.

وقتی $H_1$ به $H_0$ میل می‌کند، باید معیار پشتیبانی به سمت مقداری میل کند که نشان‌دهنده پشتیبانی یکسان از هر دو فرضیه است. p—مقدار به‌عنوان یک معیار پشتیبانی فاقد این ویژگی است.

**مثال ۸** فرض کنید $X \sim N(\theta, 1)$. همچنین فرض کنید $H_0 : \theta = 0$ و $H_1 : \theta = a \ (a > 0)$. در این‌صورت p—مقدار آزمون برای مقدار مشاهده شده $x$ عبارت است از

$$\hat{p}(x) = P(X \geq x | \theta = 0).$$

از آن‌جا که p—مقدار فوق به $a$ بستگی ندارد بنابراین وقتی $a \to 0$ ($H_1 \to H_0$)، p—مقدار تغییر نمی‌کند.
علت این که p—مقدار فاقد این ویژگی است، ناشی از این موضوع است که فرضیه جانشین در تعیین شدت شواهدی که براساس p—مقدار سنجیده می‌شود، تأثیر ندارد.

[11] با توجه به اینکه p—مقدار معیاری در بازه‌ی $[0, 1]$ است انتظار می‌رود که $\hat{p}_{01}$ و $\hat{p}_{10}$ نسبت به ۱/۲ (پشتیبانی یکسان از هر دو فرضیه) متقارن باشد؛ یعنی، $\hat{p}_{10} - 1/2 = -(1/2 - \hat{p}_{01})$.
[12] Coherence

۹) p—مقدار با مشکل پارادوکس لیندلی[13] روبرو است.

به موجب این پارادوکس، برای یک توزیع پشین و p—مقدار ثابت، با افزایش حجم نمونه احتمال پسین درستی فرضیه صفر به عدد یک میل می‌کند [۲ و ٦].

**مثال ۹** فرض کنید $X_1,\ldots,X_n \overset{iid}{\sim} N\left(\theta,\sigma^2\right)$ با $\sigma^2$ معلوم باشند. فرضیه $H_0:\theta=\theta_0$ در مقابل $H_1:\theta\neq\theta_0$ را در اختیار داریم. آماره آزمون (متداول) و p—مقدار آزمون به‌ترتیب عبارتند از

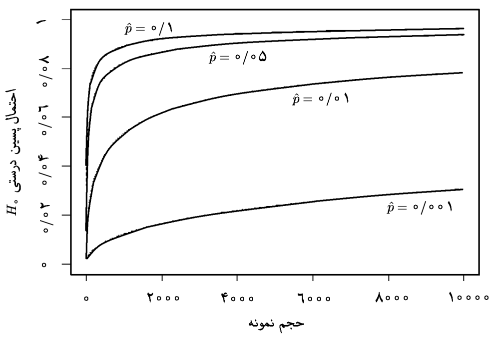
$$T = T(X)=\sqrt{n}\left|\bar{X}-\theta_0\right|/\sigma$$
$$\hat{p} = \hat{p}(x)=2\left(1-\Phi(t)\right).$$

همچنین فرض کنید $P(H_0)=\pi_0$، $P(H_1)=1-\pi_0$ و $\theta$ تحت فرضیه $H_1$ دارای توزیع $N\left(\theta_0,\tau^2\right)$ با $\tau^2$ معلوم باشد. در این‌صورت احتمال پسین درستی $H_0$ عبارت است از

$$P(H_0|x)=\frac{1}{1+\frac{1-\pi_0}{\pi_0}(1+c)^{-1/2}exp\{\frac{t^2}{2(1+1/c)}\}}$$
$$=\frac{1}{1+\frac{1-\pi_0}{\pi_0}(1+c)^{-1/2}exp\{\frac{(\Phi^{-1}(1-\hat{p}/2))^2}{2(1+1/c)}\}}$$

که در آن $c=n\left(\tau^2\big/\sigma^2\right)$.

بدیهی است که $n\to\infty \Rightarrow c\to\infty$. حال اگر n افزایش یابد ولی $t=\sqrt{n}|\bar{x}-\theta_0|/\sigma$ ثابت بماند، احتمال پسین درستی $H_0$ به سمت یک میل می‌کند، در صورتی‌که p—مقدار تغییر نمی‌کند. در این‌صورت اگر p—مقدار کوچک باشد نشان‌دهنده شواهد قوی برعلیه $H_0$ است، در صورتی‌که به احتمال قریب به یقین $H_0$ درست می‌باشد. توجه کنید که این مطلب به‌ازای هر $1>\pi_0>0$ و $\tau^2>0$ درست است.




**شکل ۱.** نمودار پارادوکس لیندلی

جدول (۱) و نمودار (۱) رفتار احتمال پسین درستی $H_0$ هنگام افزایش حجم نمونه را برای چند p—مقدار در مثال فوق نمایش می‌دهند (هنگامی که $\tau^2=\sigma^2, \pi_0=1/2$).

۱۰) p—مقدار کالیبره[14] نیست.

یکی دیگر از ویژگی‌های مطلوب یک معیار پشتیبانی کالیبره بودن آن است؛ به این معنی که مقادیر یکسان آن در موقعیت‌های مختلف معنی یکسان داشته باشند. به‌عنوان مثال دماسنج را به‌عنوان یک ابزار سنجش دما کالیبره گوییم زیرا اگر در یک مکان دماسنج اختلاف دمای بین

[13] Lindley
[14] Calibration

دوجسم را ۱۰ درجه سانتی‌گراد نشان دهد، این میزان اختلاف دما را در هر زمان و مکان دیگری با همان ۱۰ درجه سانتی‌گراد نشان خواهد داد. متأسفانه، این مطلب برای p—مقدار صادق نمی‌باشد؛ زیرا ممکن است در دو مساله متفاوت که شواهد واقعی یکسان هستند، p—مقدار دو مقدار متفاوت را ارائه دهد و بالعکس. ما در اینجا دو فرضیه را ساده فرض می‌کنیم و شواهد واقعی را عامل بیز[15] در نظر می‌گیریم (برای شرح بیشتر به مرجع [۸] مراجعه شود). عامل بیز که به صورت نسبت بخت پسین به بخت پیشین تعریف می‌شود ( $b = \frac{P(H_0|X=x)/P(H_1|X=x)}{P(H_0)/P(H_1)}$ ) نشان می‌دهد که مشاهده $X = x$ چند برابر نسبت بخت‌ها را افزایش (کاهش) می‌دهد. در حالتی‌که فرضیه‌های $H_0$ و $H_1$ ساده هستند عامل بیز همان نسبت درستنمایی دو فرضیه خواهد بود. مثال زیر به‌خوبی نشان می‌دهد که p—مقدار کالیبره نیست (البته، در مورد کالیبره کردن p—مقدار کوشش‌هایی به‌عمل آمده از آن جمله می‌توان به مرجع [۱۱] اشاره کرد).

**مثال ۱۰** فرض کنید که $X_1, \cdots, X_n$ یک نمونه تصادفی مستقل و هم‌توزیع از توزیع پیوسته $f$ باشد. برای آزمون فرضیه $H_0 : f = f_0$ در مقابل $H_1 : f = f_1$ کلاس آزمون‌های MP و تو در توی

$$\varphi(x) = \begin{cases} 1 & b \leq k \\ 0 & b > k \end{cases}$$

که در آن $b = f_0(\mathbf{x})/f_1(\mathbf{x})$. در نتیجه p—مقدار آزمون عبارت است از:

$$\begin{aligned} \hat{p}(x) &= P_{H_0}\left(\frac{f_1(X)}{f_0(X)} \geq \frac{f_1(x)}{f_0(x)}\right) \\ &= P_{H_0}(B \leq b) = F_{B|H_0}(b). \end{aligned}$$

که در آن $B = f_0(\mathbf{X})/f_1(\mathbf{X})$. در این‌صورت، اگر $f_0 \sim N(0,1)$ و $f_1 \sim N(1,1)$ باشد، آنگاه p—مقدار عبارت است از

$$\begin{aligned} \hat{p}(x) &= P_{H_0}\left(\exp\left\{-X + \frac{1}{2}\right\} \leq b\right) \\ &= 1 - \Phi\left(\frac{1}{2} - \ln b\right). \end{aligned}$$

ولی اگر $f_0 \sim \exp(1)$ و $f_1 \sim \exp(2)$ باشد، آنگاه

$$\begin{aligned} \hat{p}(x) &= P_{H_0}\left(\frac{1}{2}\exp\{X\} \leq b\right) \\ &= P_{H_0}(X \leq \ln(2b)) = 1 - \frac{1}{2b}. \end{aligned}$$

ملاحظه می‌شود که در دو مسئله فوق $\hat{p}$ توابعی متفاوت از b است؛ یعنی، ممکن است میزان پشتیبانی داده‌ها از فرضیه $H_0$ در مقابل $H_1$ بر اساس $b$ یکسان باشد اما p—مقدارهای آنها متفاوت باشد و بالعکس، ممکن است که مقادیر p—مقدارها در دو آزمون یکسان باشد اما میزان پشتیبانی واقعی داده‌ها از فرضیه $H_0$ در مقابل $H_1$ یکسان نباشد؛ برای مثال، p—مقدار ۰/۰۳ در آزمون فرضیه اول به معنی $b \approx 0/25$ در حمایت از $H_0$ در مقابل $H_1$ است؛ در حالی‌که، همین p—مقدار در آزمون فرضیه دوم به معنی $b \approx 0/52$ در حمایت از $H_0$ در مقابل $H_1$ است.

[15] Bayes factor

# ۵ قانون درستنمایی

همان‌طور که در مقاله اول[۱] به آن اشاره شد، از نسبت درستنمایی دو فرضیه[۱۶] می‌توان به‌عنوان معیاری برای اندزه‌گیری میزان پشتیبانی داده‌ها از یک فرضیه در مقابل فرضیه دیگر استفاده کرد. که از آن با نام "قانون درستنمایی" یاد می‌شود.

# ۶ توجیه قانون درستنمایی

۱) برابری نسبت درستنمایی با عامل بیز

زمانی‌که فرضیه‌های $H_0$ و $H_1$ هر دو ساده هستند داریم

$$\begin{aligned} b &= \frac{P(H_0|X=x)/P(H_1|X=x)}{P(H_0)/P(H_1)} \\ &= \frac{\frac{P(X=x|H_0)\ P(H_0)}{P(X=x|H_1)\ P(H_1)}}{\frac{P(H_0)}{P(H_1)}} \\ &= \frac{P(X=x|H_0)}{P(X=x|H_1)} = r \end{aligned}$$

که در آن r نسبت درستنمایی $H_0$ به $H_1$ است.

به‌عبارت دیگر، مشاهده $X=x$ نسبت بخت‌ها را با ضریبی برابر r افزایش (یا کاهش) می‌دهد. اگر $r > 1$ باشد در این‌صورت می‌توان نتیجه گرفت که داده‌ها از فرضیه $H_0$ بیشتر از فرضیه $H_1$ پشتیبانی می‌کنند و هر چه r بزرگ‌تر از یک باشد میزان پشتیبانی از فرضیه‌ی $H_0$ نیز بیشتر است؛ و بالعکس، اگر $r < 1$ باشد در این‌صورت می‌توان نتیجه گرفت که داده‌ها از فرضیه $H_1$ بیشتر از فرضیه $H_0$ پشتیبانی می‌کنند و هر چه r کوچک‌تر از یک باشد میزان پشتیبانی از فرضیه‌ی $H_1$ نیز بیشتر است. بنابراین می‌توان نسبت درستنمایی را به‌عنوان معیاری برای پشتیبانی داده‌ها از فرضیه صفر در مقابل فرضیه جانشین درنظر گرفت.

۲) تعمیم درست اصل عکس نقیض است.

قانون درستنمایی می‌گوید: از دو فرضیه، فرضیه‌ای که احتمال بیشتری به مشاهده $x$ نسبت می‌دهد از جانب $x$ بیشتر پشتیبانی می‌شود و هر چه نسبت این دو احتمال بیشتر باشد، پشتیبانی $x$ از آن فرضیه بیشتر است.

علاوه بر این، اگر $r = 0$ آنگاه، $P(X=x|H_0) = 0$ و $P(X=x|H_1) > 0$. و در نتیجه $D \to \sim H_0$. یعنی قانون درستنمایی در حالت خاص همان اصل "عکس نقیض" است (که در مقاله اول به آن اشاره شد).

۳) سازگار است

به این معنی که هنگامی‌که $H_0$ درست است وقتی $n \to \infty$ آنگاه، $r \xrightarrow{p} \infty$ و هنگامی که $H_1$ درست است وقتی $n \to \infty$ آنگاه، $r \xrightarrow{p} 0$ [۴].

# ۷ ویژگی‌های قانون درستنمایی

۱) در اصل درستنمایی صدق می‌کند.

بدیهی است که قانون درستنمایی در این اصل صدق می‌کند، زیرا نسبت دو تابع درستنمایی است.

[۱۶] در ادامه هر کجا سخن از فرضیه به میان آید منظور فرضیه ساده است.

۲) به فضای نمونه بستگی ندارد.

از آنجا که تابع درستنمایی تنها به مقدار مشاهده شده $x$ بستگی دارد؛ بنابراین بدیهی است که قانون درستنمایی به فضای نمونه بستگی نداشته باشد.

۳) به قاعده توقف نمونه‌گیری بستگی ندارد.

فرض کنید دنباله متغیرهای تصادفی مستقل $X_1, X_2, ...$ را مشاهده کنیم تا براساس مقادیر مشاهده شده $x_1, ..., x_n$ نمونه‌گیری متوقف شود. در این‌صورت، تابع درستنمایی (در حالت گسسته) عبارت است از

$$\begin{aligned} & f(x_1, ..., x_n, n; \theta) \\ &= P(X_1 = x_1, ..., X_n = x_n, N = n; \theta) \\ &= P(X_1 = x_1, ..., X_n = x_n; \theta) \\ &= f_1(x_1; \theta) ... f_n(x_n; \theta) \end{aligned}$$

که در آن $f_i(x; \theta)$ تابع چگالی متغیر تصادفی $X_i$ است. تساوی دوم از آن جهت برقرار است که حجم نمونه براساس مقادیر مشاهده شده $x_1, ..., x_n$ تعیین شده است. یعنی،

$$\{N = n\} \subset \{X_1 = x_1, ..., X_n = x_n\}.$$

اثبات در حالت پیوسته به‌طور مشابه است.

شایان ذکر است، از آنجا که قانون درستنمایی در اصل درستنمایی صدق می‌کند؛ لذا، به طرح نمونه‌گیری (فضای نمونه و قاعده توقف) نباید بستگی داشته باشد.

۴) احتمال گمراه‌کنندگی قانون درستنمایی کم است

قضیه زیر این مطلب را تایید می‌کند [۸، ۹].

**قضیه ۱** فرض کنید متغیر تصادفی $X$ تحت فرضیه‌های $H_0$ و $H_1$ به‌ترتیب دارای توزیع‌های $f_0$ و $f_1$ باشد. در این‌صورت برای هر ثابت $k > 0$ داریم

$$P_{H_1}(R \geq k) \leq 1/k$$

و

$$P_{H_0}(R \leq 1/k) \leq 1/k$$

که در آن

$$R = f_0(X)/f_1(X)$$

اثبات:

$$\begin{aligned} P_{H_1}(R \geq k) &= \int_{r \geq k} f_1(x) \ dx \\ &= \int_{f_0(x)/f_1(x) \geq k} f_1(x) \ dx \\ &\leq \frac{1}{k} \int_{f_0(x)/f_1(x) \geq k} f_0(x) \ dx \\ &\leq \frac{1}{k}. \end{aligned}$$

در حالتی که $X$ گسسته باشد کافی است در مراحل اثبات فوق نماد انتگرال را با نماد مجموع جایگزین نمائیم. اثبات نامساوی دیگر به‌طور مشابه می‌باشد.

روابط فوق، وقتی $k$ بزرگ باشد، بیانگر این مطلب هستند که احتمال پشتیبانی قوی از فرضیه نادرست، بسیار کوچک است. حائز اهمیت است

که توزیع‌های متغیر تصادفی $X$ تحت فرضیه‌های $H_0$ و $H_1$ هر چه باشند، روابط فوق برقرار هستند. به همین دلیل روابط فوق را "کران‌های جهانی"[۱۷] می‌نامند (برای شرح بیشتر در این مورد مرجع [۸، ۹] ملاحظه شود).

همچنین ثابت می‌شود که اگر کاربری عمداً بخواهد نمونه‌گیری را آنقدر ادامه دهد تا نتیجه $r \leq 1/k$ حاصل شود؛ اگر $H_0$ درست باشد، احتمال اینکه وی موفق شود بیشتر از $1/k$ نیست [۷]!

۵) قانون درستنمایی به فرضیه جانشین بستگی دارد.

در قانون درستنمایی برخلاف p—مقدار، فرضیه جانشین نه تنها در تعیین جهت، بلکه در میزان شواهد موجود در داده‌ها به‌نفع (یا بر علیه) یک فرضیه نیز موثر است.

٦) قانون درستنمایی نسبت به $H_0$ و $H_1$ متقارن است.

اگر $r_{01}$، نسبت درستنمایی فرضیه $H_0$ در مقابل $H_1$، و $r_{10}$، نسبت درستنمایی فرضیه $H_1$ در مقابل $H_0$ باشد، آنگاه $r_{10} = 1/r_{01}$. در نتیجه اگر جای $H_0$ و $H_1$ را عوض کنیم، میزان پشتیبانی داده‌ها از فرضیه‌ها تغییر نمی‌کند.

۷) وقتی $H_1$ به $H_0$ میل می‌کند باید معیار پشتیبانی به سمت مقداری میل کند که نشان‌دهنده پشتیبانی یکسان از هر دو فرضیه است.

نسبت درستنمایی دارای این ویژگی هست. به عبارت دیگر، فرض کنید $H_0 : \theta = \theta_0$ و $H_1 : \theta = \theta_1$. در این‌صورت

$$\theta_1 \to \theta_0 \quad \Rightarrow \quad r \to 1$$

۸) نسبت درستنمایی کالیبره است.

از آن‌جا که نسبت درستنمایی برابر عامل بیز است، لذا کالیبره نیز می‌باشد.

## ۸ نتیجه‌گیری

۱) باید توجه داشت، علیرغم این‌که p—مقدار یک معیار پشتیبانی داده‌ها از فرضیه صفر می‌باشد اما معیار دقیقی نیست.

۲) منطقی به‌نظر می‌رسد که برای اندازه‌گیری میزان شواهد موجود در داده‌ها در حمایت از یک فرضیه ساده در مقابل فرضیه ساده دیگر از نسبت درستنمایی آن دو فرضیه استفاده کنیم.

در مقاله سوم اندزه‌گیری میزان شواهد موجود در داده‌ها را در حالت فرضیه‌های مرکب مورد بررسی قرارخواهیم داد.

## ۹ قدردانی

بدین‌وسیله از داور محترم مقاله که با نکاتی که مطرح نمودند بر کیفیت این مقاله افزودند تشکر و قدردانی می‌گردد.

[۱۷] Universal bounds

جدول ۱. جدول پارادوکس لیندلی (بر گرفته از [۲])

| | | | n | | | | |
|---|---|---|---|---|---|---|---|
| $\hat{p}$ | ۱ | ۱۰ | ۲۰ | ۵۰ | ۱۰۰ | ۱۰۰۰ | ۱۰۰۰۰ |
| ۰/۱ | ۰/۴۲ | ۰/۴۷ | ۰/۵۶ | ۰/۶۵ | ۰/۷۲ | ۰/۸۹ | ۰/۹۶ |
| ۰/۰۵ | ۰/۳۵ | ۰/۳۷ | ۰/۴۲ | ۰/۵۲ | ۰/۶۰ | ۰/۸۲ | ۰/۹۴ |
| ۰/۰۱ | ۰/۲۱ | ۰/۱۴ | ۰/۱۶ | ۰/۲۲ | ۰/۲۷ | ۰/۵۳ | ۰/۷۸ |
| ۰/۰۰۱ | ۰/۰۸۶ | ۰/۰۲۴ | ۰/۰۲۶ | ۰/۰۳۴ | ۰/۰۴۵ | ۰/۱۲۴ | ۰/۳۰۹ |

# مراجع

[۱] ارقامی، ن.ر.، زمان زاده، ا. و راحتی، س. (۱۳۸۷)، شواهد آماری، قسمت اول: تفسیر نتایج آزمون‌های آماری کلاسیک، اندیشه آماری، سال ۱۳، شماره ۱، ۳—۱۰.

[2] Berger, J.O. and Sellke, T. (1987), Testing a point null hypothesis: The irreconcilability of P values and evidence, *Jornal of the American Statistical Association*, 82(397), 112-122.

[3] Casella, G. and Berger, R.L. (2002), *Statistical Inference*, 2nd Ed., Duxbury Press, the University of Michigan.

[4] Lehmann, E.L. and Casella, G. (1998), *Theory of Point Estimation*, 2nd Ed. Springer, New York.

[5] Lehmann, E.L. and Romano, J.P. (2005), *Testing Statistical Hypotheses*, 3rd Ed. Springer, New York.

[6] Lindley, D.V. (1975), A statistical paradox, *Biometrika*, 44, 187-192.

[7] Robbins, H. (1970), Statistical methods related to the law of the iterated logarithm, *Annals of Mathematical Statistics*, 41(5), 1397-1409.

[8] Royall, R. (1997), *Statistical Evidence: A Likelihood Paradigm*, Chapman and Hall/CRC, Florida.

[9] Royall, R. (2000), On probability of observing misleading statistical evidence, *Jornal of the American Statistical Association* , 95(451), 760-768.

[10] Schervish, M.J. (1996), P-values: What they are and what they are not, *The American Statistician*, 50(3), 203-206.

[11] Sellke, T. and Bayarri, M.J. and Berger, G. (2001), Calibration of P-values for testing precise null hypotheses, *The American Statistician*, 55(1), 62-71.