روش درست‌نمایی نیمرخ در تحلیل شواهدی داده‌های ژنتیک

مهدی عمادی [۱*]، انیس سنچولی[۲] مرتضی محمدی[۳]

[۱] دانشگاه فردوسی مشهد

[۲] دانشگاه فردوسی مشهد

[۳] دانشگاه زابل

۱۴۰۳/۲/۳۱

چکیده: در علم آمار سه مکتب به نام‌های مکتب فراوانگرایان، مکتب بیزین‌ها و مکتب شواهدی وجود دارد. در مکتب فراوانگرایان تنها به مشاهدات و بسامد رخدادها استناد می‌شود و بر حسب آن می‌توان مسائل را حل کرد، در حالی که در مکتب بیزین‌ها علاوه بر مشاهدات، اطلاعات و باورهای اولیه محقق نیز مهم است و در حل مسئله و نتیجه‌گیری مورد نظر قرار می‌گیرند و در مکتب شواهدی برای اندازه گیری میزان پشتیبانی داده‌ها از فرضیه مفروض، دو معیار، یکی برای اندازه گیری کفایت شواهد (شامل داده‌ها و اطلاعات مسئله) و دیگری اندازه گیری ارتباط شواهد (شامل فرضیه‌های مسئله) وجود دارد، که این دو معیار از هم جدا هستند. بنابراین، تنها معیاری که هر دو را اندازه گیری می‌کند، شاخص نسبت درست‌نمایی است.

*نویسنده مسئول:جناب آقای دکتر مهدی عمادی emadi@um.ac.ir

واژه‌های کلیدی: رهیافت شواهدی، نسبت درست‌نمایی، درست‌نمایی نیمرخ، رگرسیون لجستیک.

۱ مقدمه

از اواخر سده بیستم با چاپ و انتشار مقالات و کتاب ریچارد رویال [۱] (۱) رهیافت فکری جدیدی تحت عنوان رهیافت شواهدی که مبتنی بر مفهوم درست‌نمایی است شکل گرفت و رقیبی برای دو رهیافت دیگر که در مقالات افران [۲] (۲) و گوستافسون [۳] (۳) وجود دارد شد. تابع درست‌نمایی باید برحسب پارامتر مورد علاقه باشد، ساده‌ترین راه حذف پارامترهای مزاحم از تابع درست‌نمایی است که روش‌های گوناگون در رویال (۱) وجود دارد، در این مقاله روش درست‌نمایی نیمرخ برای تجزیه و تحلیل داده‌های ژنتیک در رابطه با انتقال بیماری سیستیک فیبروزیست بررسی شده و داده‌ها از مقاله استرانگ (۴) و از سایت [۴]انجمن ژنتیک بین المللی مدیفایر [۵] گرفته شده.

۱.۱ تابع درست‌نمایی نیمرخ

در این روش ماکسیمم تابع درستنمایی را زمانی که مقدار پارامتر مورد علاقه ثابت در نظر گرفته شود، بدست می‌آید. در این روش می‌توان پارامتر γرا با روش جایگزینی با یک برآورد مقدار واقعی آن، γ_0 به ازای هر مقدار θ، از مقدار $\hat{\gamma}(\theta)$، که $L(\theta, \gamma)$ را بیشینه می‌سازد، استفاده می‌شود. تابعی که از این طریق به صورت

---

[۱] Royall Richard

[۲] Efran

[۳] Gustafson

[۴] http://locuszoom.sph.umich.edu

[۵] شامل گروه‌هایی از کشورهای فرانسه، کانادا و دانشگاه‌های جانزهاپکینز وسترن است.

زیر بدست می‌آید تابع درستنمایی نیم‌رخ نامیده می‌شود.

$$L_P(\theta) = L(\theta, \hat{\gamma}(\theta)) = max_\gamma(L(\theta, \gamma))$$

مثال ۱. *فرض کنید متغیر تصادفی X دارای توزیع نرمال با میانگین مجهول μ و واریانس مجهول δ^2 باشد. برای محاسبه ی استنباط شواهدی درباره پارامتر مورد علاقه‌ی μ بر اساس نمونه تصادفی به حجم n از توزیع فوق داریم:*

$$L(\mu, \delta^2) = (\frac{1}{\sqrt{2\pi}\delta})^n exp(\frac{-1}{2\delta^2}\sum_{i=1}^{n}(x_i - \mu)^2)$$

ماکسیمم تابع درستنمایی فوق را زمانی‌که μ ثابت فرض شود برحسب δ^2 بدست می‌آوریم:

$$L_P(\mu) = max_{\delta^2}[(\frac{1}{\sqrt{2\pi}\delta})^n exp(\frac{-1}{2\delta^2}\sum_{i=1}^{n}(x_i - \mu)^2)]$$

ماکسیمم فوق به ازای $\delta^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$ رخ می‌دهد، پس داریم:

$$L_P(\mu) \propto (\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2})^{\frac{-n}{2}}$$

حال می‌توان استنباط شواهدی را بر حسب تابع درستنمایی نیم‌رخ فوق بدست‌آورد.

تعریف ۱. *در بیماری سیستیک فیبروزیست (CF)یک آنوزمال را در نظر بگیرید که $CFTR$ یک تنظیم کننده هدایت غشایی است. فرض کنید یک شرکت دارویی علاقمند به برآورد نسبت θ برای افراد مبتلا به CF در یک کشور خاص است که حاملان ژنوتیپ هستند. شرکت فکر می‌کند این نسبت حدود ۵۰ درصد است. اما اگر بیشتر از این مقدار باشد، شرکت دارویی به دنبال توسعه درمانی هدفمند برای بیماری مورد نظر است. برای برآورد نسبت θ، شرکت دارویی می‌تواند از یک نمونه تصادفی از افراد جامعه استفاده کند و آنها را برای وجود جهش عامل ژنوتیپ آزمایش کند، سپس می‌تواند از روش‌های آماری مانند: نسبت درست‌نمایی یا عامل بیز برای برآورد نسبت θ استفاده کند. بمنظور تعیین این که آیا نسبت افراد مبتلا به CF با نسبت حاملان ژنوتیپ*

به‌طور قابل توجهی بیشتر از ۵۰ درصد است یا خیر، شرکت دارویی می‌تواند از یک آزمون فرضیه استفاده کند بطوریکه:

فرضیه صفر ($H_0 : \theta \leq ۵۰٪$) و فرضیه جایگزین ($H_1 : \theta > ۵۰٪$) است. با تحلیل داده‌های جمع‌آوری شده و استفاده از آزمون‌های آماری می‌توان فرضیه صفر را پذیرفت یا رد کرد و بررسی کرد که آیا نسبت θ به طور قابل توجهی بیشتر از ۵۰ درصد است یا خیر. در صورتی‌که نسبت θ بیشتر از ۵۰ درصد باشد، شرکت دارویی می‌تواند به دنبال توسعه درمانی هدفمند برای جهش مورد نظر باشد.

۲.۱ تحلیل داده‌های ژنتیک در حضور پارامتر مزاحم توسط تابع درستنمایی نیمرخ

به عنوان مثال در نظر گرفتن مدل چند پارامتره $L_n(\theta, \lambda)$ که پارامتر مورد علاقه θ است و λ جز پارامترهای مزاحم محسوب می‌شود. درستنمایی شرطی و حاشیه‌ای همیشه در دسترس نیستند بنابراین همیشه می‌توان درستنمایی نیمرخ را محاسبه کرد. درستنمایی نیمرخ راه حل کلی برای نشان دادن قوت شواهد آماری برای پارامتر مورد علاقه در حضور پارامترهای مزاحم ارائه می‌دهد و در برنامه‌های کاربردی ژنتیک قابل استفاده است.

۱ در نمونه‌های بزرگ درستنمایی نیمرخ دو خصوصیت مهم درستنمایی را دارد بطوریکه در آن احتمال شواهد گمراه کننده توسط تابع کوهانی توصیف می‌شود.

۲ در نرم افزار R بسته $PLikelihood$ برای محاسبه نسبت درستنمایی نیمرخ در بسیاری از مدل‌های آماری رایج: (مدل‌های خطی، خطی عمومی، ترکیبی و ترتیبی نسبتی) همراه با رسم نمودار تابع درستنمایی پارامتر مورد علاقه در دسترس است.

۳ و همچنین از بسته $EVIAN$ نیز برای تجزیه و تحلیل رگرسیون خطی و لجستیکی مناسب برای مطالعات انجمن‌های ژنتیکی ارائه شده است.

به مثال شرکت دارویی بازمی‌گردیم، شرکت دارویی از نسبت درستنمایی نیمرخ استفاده کرده است تا قدرت شواهد آماری را برای ارتباط بین ایلئوس مکونیوم و لوکوس $SLC۲۶A۹$ در بیماری CF ارزیابی می‌کند. جمعیت مورد پژوهش شامل ۹۰۱ خواهر و برادر مبتلا به CF است که توسط انجمن بین المللی مدیفایر جمع آوری شده است. در اینجا از مدل رگرسیون لجستیک به عنوان مدل آماری استفاده شده است.

$$\log \frac{p_i}{۱ - p_i} = \beta_۰ + \beta_۱ G_{i۱} + \gamma_۲ Z_{i۲} + \gamma_۳ Z_{i۳}$$

۱. که در آن $p_i = E(y_i) = p(y_i = ۱)$ و $G_{i۱} = ۰, ۱, ۲$ است.

۲. برای مقدار ایلئوس مکونیوم یک نتیجه دو دویی در نظر گرفته شده است، به این معنی که اگر فرد i ام با ایلئوس مکونیوم بدنیا آمده باشد $y_i = ۱$ در غیر این صورت $y_i = ۰$ است.

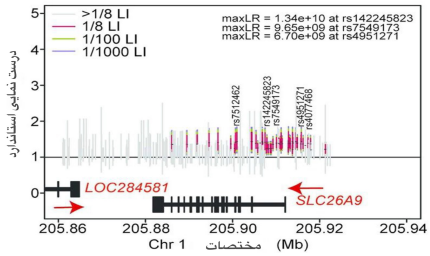
۳. متغییرهای $\gamma_۳$ ، $\beta_۰$، $\gamma_۲$جزء پرامترهای مزاحم هستند و $\beta_۱$پارامتر مورد علاقه است.

۴. متغییرهای $Z_{i۲}$، $G_{i۱}$، $Z_{i۳}$ به ترتیب به معنای تعداد آلل‌های جایگزین در یک نکلئوتید خاص در i-امین فرد بیمار شرکت کننده در اجرای تست ژنتیکی بر روی DNA آن‌ها است.

فرع ۱. *طبق تعاریف گفته شده نمودار و جدول تحلیلی زیر را داریم:*

۱. نمودار مقابل بازه شواهدی ارتباط بین ناحیه کروموزمی با ایلئوس مکونیوم را نشان می‌دهد، که هر بازه خلاصه تابع درست‌نمایی برای هر SNP درج شده است.

۲. بازه‌های درست‌نمایی که با رنگ سبز و آبی مشخص شده‌اند ماکسیمم نسبت درستنمایی نیمرخ را نشان می‌دهند، که نماینده‌های محدوده با داده‌های جمعیتی rs هستند که بازه درستنمایی باریکی را نشان می‌دهند که دور از خط $OR = ۱$ قرار دارند و دلیل قوی درباره ارتباط این منطقه ارائه می‌کند.



شکل ۱: تجزیه و تحلیل EP از لوکوس کروموزوم $SLC۲۶A۹$۱ با ایلئوس مکونیوم در CF.

جدول ۱: خلاصه شبیه سازی داده های آماری برای تجزیه و تحلیل EP ساده در مقایسه با تجزیه و تحلیل EP ساده برای نمونه‌های نامرتبط و نسبتاً مرتبط CF

$Robust factor(\frac{a}{b})$	$\frac{۱}{۱۰۰۰}LI$	$\frac{۱}{۱۰۰}LI$	$\frac{۱}{۸}LI$	MLE	maxLR	MAF	
							$a.\ Unrelated\ n = ۵,۸۶۹$
NA	۱٫۱۳۵۳, ۱٫۶۵۳۰	۱٫۱۷۶۳, ۱٫۵۹۵۳	۱٫۲۳۴۵, ۱٫۵۱۶۳	۱٫۳۶۹۹	۳۱۷, ۴۶۲, ۲۴۷	۰٫۴۰۸۹	$rs۷۵۱۲۴۶۲$
NA	۱٫۱۶۴۵, ۱٫۶۸۲۷	۱٫۲۰۶۶, ۱٫۶۲۳۹	۱٫۲۶۶۲, ۱٫۵۴۷۴	۱٫۳۹۸۰	۱۳, ۴۲۳, ۰۰۰, ۰۰۰	۰٫۴۶۵۰	$rs۱۴۲۲۴۵۸۲۳$
NA	۱٫۱۶۴۶, ۱٫۶۷۴۴	۱٫۲۰۳۷, ۱٫۶۲۰۱	۱٫۲۶۳۲, ۱٫۵۴۳۸	۱٫۳۹۸۲	۹, ۶۵۲, ۱۳۷, ۲۱۸	۰٫۳۹۵۱	$rs۷۵۴۹۱۷۳$
NA	۱٫۱۶۱۵, ۱٫۶۸۷۰	۱٫۲۰۳۵, ۱٫۶۳۲۲	۱٫۲۶۳۰, ۱٫۵۵۱۴	۱٫۳۹۸۰	۶, ۶۹۵, ۶۵۸, ۴۷۸	۰٫۴۱۹۴	$rs۴۹۵۱۲۷۱$
NA	۱٫۱۵۸۶, ۱٫۶۸۷۰	۱٫۲۰۰۵, ۱٫۶۲۸۰	۱٫۲۵۹۸, ۱٫۵۴۷۴	۱٫۳۹۴۵	۴, ۶۷۹, ۱۱۷, ۸۶۵	۰٫۴۱۲۹	$rs۴۰۷۷۴۶۸$
							$b.\ Related\ n = ۶,۷۷۰$
۰٫۹۷۵	۱٫۱۰۴۰, ۱٫۵۸۷۲	۱٫۱۴۱۰, ۱٫۵۳۱۸	۱٫۱۹۷۴, ۱٫۴۵۹۷	۱٫۳۲۲۱	۱۹, ۶۵۶, ۱۹۶	۰٫۴۰۹۱	$rs۷۵۱۲۴۶۲$
۰٫۹۶۶	۱٫۱۳۲۴, ۱٫۶۱۵۷	۱٫۱۷۰۴, ۱٫۵۶۳۲	۱٫۲۲۸۲, ۱٫۴۸۹۶	۱٫۳۵۲۶	۵۰۲, ۳۱۸, ۲۸۲	۰٫۴۶۵۲	$rs۱۴۲۲۴۵۸۲۳$
۰٫۹۷۰	۱٫۱۳۲۶, ۱٫۶۰۷۸	۱٫۱۷۰۶, ۰٫۵۵۵۶	۱٫۲۲۵۳, ۱٫۴۸۶۱	۱٫۳۴۹۴	۳۷۶, ۸۹۵, ۰۰۱	۰٫۳۹۳۷	$rs۷۵۴۹۱۷۳$
۱٫۰۰۴	۱٫۱۲۶۶, ۱٫۶۰۳۴	۱٫۱۶۴۵, ۱٫۵۵۱۴	۱٫۲۲۲۰, ۱٫۴۷۸۳	۱٫۳۴۲۴	۲۷۸, ۹۵۸, ۸۰۸	۰٫۴۱۹۳	$rs۴۹۵۱۲۷۱$
۰٫۹۹۵	۱٫۱۲۶۶, ۱٫۶۰۷۵	۱٫۱۶۴۵, ۱٫۵۵۵۳	۱٫۲۲۲۰, ۱٫۴۸۲۱	۱٫۳۴۵۸	۲۳۷, ۶۵۳, ۲۷۴	۰٫۴۱۲۹	$rs۴۰۷۷۴۶۸$

SNPها با بیشترین $maxLRs$ (ماکسیمم نسبت درستنمایی نیمرخ) نمایش داده شده‌اند، و تحلیل‌های آماری برای واریانت‌های $rs۷۵۱۲۴۶۲$ و $rs۴۰۷۷۴۶۸$ که شواهد ارتباط قبلی با CF را نشان داده‌اند، نیز آمده است و فاکتور تنظیم قوی برای تجزیه و تحلیل نمونه مرتبط اعمال شده است.

MAF: بیانگر فراوانی آلل‌های کمتر که همان مقادیر $Z_{i۲}$، $G_{i۱}$، $Z_{i۳}$ در مدل رگرسونی تعریف شده هستند.

MLE: برآوردگر ماکسیمم درست‌نمایی است.

بحث و نتیجه‌گیری

رهیافت شواهدی یک روش آماری است که برای استخراج اطلاعات از داده‌های احتمالی استفاده می‌شود. این رهیافت به دلیل خواص و ویژگی‌های خود قابلیت استفاده در مسائل پزشکی را دارد. به طور مثال: در آزمایشات بالینی برای تایید داروهای ژنتیک، رهیافت شواهدی را می‌توان به عنوان یک روش برای ارزیابی یکسان بودن داروهای ژنتیک با داروهای اصلی استفاده کرد.

مراجع

[1] ROYALL, R. Statistical evidence a likelihood paradigm. CRC Press, New York, 2017.

[2] EFRON, B. Why isn't everyone a bayesian. The American Statistician 40, 1 (1986), 1–5.

[3] GUSTAFSON, P. Parameter restrictions for the sake of identification is there utility

in asserting that perhaps a restriction holds. Statistical Science 38, 3 (2023), 477–489.

[4] Strug, L.J., 2018. The evidential statistical paradigm in genetics. Genetic Epidemiology, 42(7), pp.590607.

Profile likelihood evidence analysis of genetic data

[1]mahdi Emadi, Ferdowsi University, Mashhad , Iran.

[2]Anis Sanchouli, Ferdowsi University, Mashhad , Iran.

[3]Morteza Mohammadi, Zabol University, Zabol , Iran.

Abstract: In the science of statistics, there are three schools of thought, the Bayesian school and the evidence school. In the frequentist school, only the observations and frequency of events are considered and problems can be solved according to that, while in the Bayesian school, in addition to the observations, information and creative beliefs of the researchers are important and they are involved in solving the problem and drawing conclusions.

Keywords: evidence paradigm, likelihood ratio, profile likelihood, logistic regression.
