Educational Measurement
https://jem.atu.ac.ir/

ATU PRESS

Determining the Cut-off Score of Criterion-Referenced Tests Using Non-Parametric Estimation Methods of the Youden Index

Maryam Parsaeian — PhD. Candidate, Faculty of Psychology and Education, University of Tehran, Tehran, Iran.E-mail: maryam.parsaeian@ut.ac.ir

Ebrahim Khodaie * — *Corresponding Author,*Associate Professor, Faculty of Psychology and Education, University of Tehran, Tehran, Iran.Email : khodaie@ut.ac.ir

Balal Izanloo — Assistant Professor, University of Kharazmi, Faculty of Psychology and Education , Tehran, Iran.E-mail: b.ezanloo@gmail.com

Keyvan Salehi — Associate Professor, Faculty of Psychology and Education, University of Tehran, Tehran, Iran. E-mail : keyvansalehi@ut.ac.ir

Sima NaghiZadeh — Assistant Professor, National Organization of Educational Testing (NOET), Tehran.E-mail : s_naghizadeh@yahoo.com

Abstract

The Youden index is a commonly used summary measure for the Receiver Operator Characteristic (ROC) curve that both measures the performance of a criterion-referenced test and specifies the cutoff score value for the test. This research aims to compare and evaluate three empirical non-parametric estimation methods, kernel with Silverman's bandwidth method and kernel with maximum likelihood cross-validation bandwidth method to calculate the value of Youden's index. In this research, bootstrap standard error (BSE), root mean square error (RMSE), square integrated error (ISE) and mean square integrated error (MISE) indices were used to evaluate the performance. The dataset utilized in this study comprises the scores in various sections (grammar, listening, reading and writing) of 461 participants who were involved in the Tolimo language proficiency examination, which was formulated and implemented by the national organization of educational testing of Iran. The results showed that the kernel method with maximum likelihood cross-validation had a higher Youden index value. The obtained cutoff scores were 479 for the kernel methods and 465 for the empirical method. According to the acceptable results of the evaluation indices, kernel methods especially with the optimal bandwidth of the maximum likelihood cross-validation lead to more reliable estimates of the Youden index and the cutoff score.

Keywords: Cut score, Youden index, Empirical nonparametric estimation, Kernel nonparametric estimation, Bootstrap standard error, ISE and MISE

Cite this Article: Parsaeian, M., Khodaie, E., Izanloo, B., Salehi, K. & NaghiZadeh, S. (2024). Determining the Cut-off Score of Criterion-Referenced Tests Using Non-Parametric Estimation Methods of the Youden Index Subject of Study: Tolimo Test. *Educational Measurement, 14*(56), 35-61. https://doi.org/10.22054/jem.2024.77090.3508

Publisher: Allameh Tabataba'i University Press
DOI: https://doi.org/10.22054/jem.2024.77090.3508

14(56), Spring 2024, 35-61
eISSN: 2470-6240 ISSN: 2252-004x
Article Type: Reserch Article
Received in revised form December 08, 2023
Accepted December 20, 2024
Received September 14, 2023
Published Online June 03, 2024

1. Introduction

Youden's index is a commonly utilized metric within the realm of medical diagnosis and patient classification. However, there is a noticeable dearth of research within the educational domain aimed at establishing the optimal threshold for criterion-referenced assessments. The utilization of non-parametric estimation techniques for calculating the Youden index is prevalent, prompting the exploration of three distinct non-parametric estimation methods (accompanied by two bandwidth approaches) in this study to derive the Youden index for determining the cutoff score of the Tolimo test. The word Tolimo is an abbreviation of “The test of language by the Iranian measurement organization”. Tolimo test is one of the standard and valid tests to determine the level of English language knowledge of students in Iran, which is designed and conducted by the national organization of educational testing of Iran. This study aims to assess and compare three empirical nonparametric estimation techniques for computing the Youden index, a key metric derived from the Receiver Operating Characteristic (ROC) curve. The primary objective is to pinpoint the most dependable method for ascertaining the cutoff score for the Tolimo test, a determination that will significantly impact the test's interpretation and decision-making processes within educational environments.

2. Literature Review

The literature review segment of the manuscript concentrates on the methodologies employed in ascertaining the cut score for criterion-referenced assessments. The article recognizes the necessity for an efficient approach in determining the cut score and underscores the significance of the Youden index as a succinct metric derived from the receiver operator characteristic (ROC) curve. The literature review establishes the framework for the investigation by emphasizing the gap in existing research and the importance of pinpointing a dependable technique for establishing cut scores in educational environments, particularly focusing on studies utilizing nonparametric estimation methodologies. As per the theoretical underpinnings and historical context of various studies such as Fluss et al. (2005), Ewald (2006), Leeflang et al. (2008), Hirschfeld and do Brasil (2014) showed that when the experimental optimization method is used to estimate the

Youden index It leads to highly variable results because the method is sensitive to randomness in the sample. Furthermore, in cases where the data exhibit normal distribution or approximate normality, employing the kernel density method with Silverman's bandwidth generates more consistent estimates of the Youden index in contrast to the experimental non-parametric estimation technique.

3. Methodology

The section of the article outlining the research methodology details the approach taken to compare and assess non-parametric estimation techniques. Specifically, the investigation focuses on three methods: the kernel method utilizing Silverman's bandwidth, the kernel method with maximum likelihood cross-validation bandwidth, and an empirical method. Various evaluation metrics such as bootstrap standard error (BSE), root mean square error (RMSE), square integrated error (ISE), and mean square integrated error (MISE) are employed by the researchers to gauge the effectiveness of these methods. The significance of obtaining dependable estimates for the Youden index and the cut-off score in criterion-referenced assessments is underscored in this section, emphasizing the necessity of a sturdy methodology to ensure precise outcomes. The methodology adopted in the research endeavors to tackle these issues and offer insights on the optimal approach for determining the cut-off score.

4. Results

The findings indicated that the kernel technique utilizing Silverman's bandwidth yielded a Youden index of 0.63 with a threshold of 479. The root mean square error for this approach was 0.715, alongside a bootstrap standard error of 0. Consequently, the computation of Youden's index through this approach demonstrated a considerable level of consistency and precision. Conversely, the kernel method employing MLCV generated a superior Youden index of 0.75 using the same threshold of 479. Therefore, the kernel method with MLCV provides a more precise evaluation of the Youden index in contrast to utilizing Silverman's bandwidth, signifying a heightened capability in distinguishing between qualified and unqualified individuals. Nonetheless, its bootstrap standard error stood at 0.039, indicating a minor level of variability in the approximation. The empirical technique yielded a Youden index of 0.31 with a threshold of 465. The root mean

square error for this technique was 0.68, with an equivalent bootstrap standard error of 0.017.

5. Conclusion

The findings of this study suggest that the choice of bandwidth determination method in relation to the kernel method plays a crucial role in estimating the Youden index. In cases where the data exhibit normality, employing Silverman's bandwidth in the kernel method yields consistent outcomes; whereas, for non-normal data, the use of maximum likelihood cross-validation bandwidth is recommended for computing and assessing the Youden index in criterion-referenced assessments and establishing the cut score. Nevertheless, further investigation is essential to validate these findings and assess the efficacy of these approaches across various domains. Such inquiries can enhance the accuracy of person classification and diminish the occurrence of erroneous test outcomes. Additionally, it is prudent to analyze the distinct impacts of setting the cut score on individual and institutional ramifications to facilitate informed decision-making in cut score determination. These endeavors are instrumental in refining assessment and decision-making frameworks within educational and institutional settings.

Acknowledgments

The current article is taken from the doctoral thesis of Tehran University's measurement and measurement field, entitled "Comparison of benchmark test scores in deep learning algorithms and the selected methods studied: Tolimo test" with the support of Tehran University and the country's education assessment organization. We would like to thank the Education Measurement Organization of the country for their cooperation in the implementation of this research and the Educational Measurement Quarterly for reviewing the article.

اندازه‌گیری تربیتی
https://jem.atu.ac.ir/

دوره ۱۴، شماره ۵۵، بهار ۱۴۰۳، ۳۵-۶۱
ISSN: 2252-004x
eISSN: 2470-6240

نوع مقاله: مقاله پژوهشی

تاریخ دریافت: ۱۴۰۲/۰۶/۲۳؛ تاریخ بازنگری: ۱۴۰۲/۰۹/۱۷
تاریخ پذیرش: ۱۴۰۲/۰۹/۲۹؛ تاریخ انتشار: ۱۴۰۳/۰۴/۱۵

تعیین نمره برش آزمون‌های ملاک‌مرجع با استفاده از روش‌های برآورد ناپارامتری شاخص یودن مورد مطالعه: آزمون تولیمو

مریم پارساییان
دانشجوی دکتری رشته سنجش واندازه‌گیری، دانشگاه تهران، تهران، ایران.
رایانامه: maryam.parsaeian@ut.ac.ir

ابراهیم خدایی *
نویسنده مسئول، دانشیار، دانشگاه تهران، تهران، ایران. رایانامه:
khodaie@ut.ac.ir

بلال ایزانلو
استادیار گروه برنامه ریزی درسی، دانشگاه خوارزمی، تهران، ایران. رایانامه:
b.ezanloo@gmail.com

کیوان صالحی
دانشیار، دانشگاه تهران، تهران، ایران. رایانامه: keyvansalehi@ut.ac.ir

سیما نقی‌زاده
استادیار، سازمان سنجش آموزش کشور، تهران، ایران. رایانامه:
s_naghizadeh@yahoo.com

چکیده

شاخص یودن یک معیار خلاصه متداول برای منحنی ویژگی عملکرد (ROC) است که هم کارایی یک آزمون ملاک‌مرجع را می‌سنجد و هم مقدار نمره برش را برای آزمون مشخص می‌کند. این پژوهش با هدف مقایسه و ارزیابی سه روش برآورد ناپارامتری تجربی، هسته با روش پهنای باند سیلورمن و هسته با روش پهنای باند اعتبارسنجی متقابل ماکسیمم درستنمایی برای محاسبه مقدار شاخص یودن انجام شده است. در این پژوهش برای ارزیابی عملکرد از شاخص‌های خطای استاندارد بوت استرپ (BSE)، ریشه میانگین مربعات خطا (RMSE)، مربع خطای یکپارچه (ISE) و میانگین مربعات خطای یکپارچه (MISE)استفاده شده است. مجموعه داده‌های مورد استفاده در این مطالعه شامل نمرات بخش‌های مختلف (گرامر، شنیداری، درک‌مطلب و نوشتاری) ۴۶۱ شرکت‌کننده بود که در آزمون مهارت زبان تولیمو شرکت داشتند که توسط سازمان سنجش آموزش کشور تدوین و اجرا شد. نتایج نشان داد که روش هسته با اعتبارسنجی متقابل ماکسیمم درستنمایی دارای مقدار شاخص یودن بالاتری بود. نمرات برش به دست آمده برای روش‌های هسته ۴۷۹ و برای روش تجربی ۴۶۵ به دست آمد. با توجه به نتایج قابل قبول شاخص‌های ارزیابی، روش‌های هسته به‌ویژه با پهنای باند بهینه اعتبارسنجی متقابل ماکسیمم درستنمایی منجر به برآوردهای قابل اعتمادتری از شاخص یودن و نمره برش شد.

کلیدواژه‌ها: نمره برش، شاخص یودن، ، برآورد ناپارامتریک تجربی، برآورد ناپارامتریک هسته، خطای استاندارد بوت‌استرپ، ISE و MISE

استناد به این مقاله: پارساییان، مریم؛ خدایی، ابراهیم؛ ایزانلو، بلال؛ صالحی، کیوان و نقی‌زاده، سیما. (۱۴۰۳). تعیین نمره برش آزمون‌های ملاک‌مرجع با استفاده از روش‌های برآورد ناپارامتری شاخص یودن، مورد مطالعه: آزمون تولیمو. *فصلنامه اندازه‌گیری تربیتی*، ۱۴(۵۵)، ۳۵-۶۱. https://doi.org/10.22054/jem.2024.77090.3508

ناشر: دانشگاه علامه طباطبائی

مقدمه

آزمون‌های ملاک‌مرجع اغلب نیاز به ایجاد یک مقدار برش برای طبقه‌بندی افراد به عنوان قبول یا مردود دارند. انتخاب این نمره برش به هزینه‌های مرتبط و پیامدهای طبقه‌بندی نادرست افراد بستگی دارد. علیرغم اینکه تفسیر یک آزمون با نتایج باینری سرراست و ساده است، تفسیر یک آزمون با نتایج پیوسته به صورت باینری به این سادگی نیست. زیرا در آزمونی با نتایج پیوسته (یا چندگانه)، هر مقدار آزمون را می‌توان به عنوان نمره برش در نظر گرفت. به عنوان مثال حداقل نمره قبولی در دروس کارشناسی، کارشناسی ارشد و دکتری در ایران به ترتیب برابر ۱۰، ۱۲ و ۱۴ است. حداقل نمره قبولی برای فرصت مطالعاتی دانشجویان دکتری در آزمون‌های $MSRT$، تولیمو و تافل به ترتیب برابر ۵۰، ۴۸۰ و ۶۰ است. در واقع با در نظر گرفتن این حداقل نمرات که همان نمرات برش هستند می‌توان دانشجویان را به دو دسته قبول و رد طبقه‌بندی کرد (Thiele & Hirschfeld, 2020).

اکثر آزمون‌ها دارای مقیاس پیوسته‌ای از مقادیر هستند که یکی از این مقدارها را می‌توان به عنوان نقطه برش برای جداسازی افراد لایق و نالایق انتخاب کرد. بدیهی است نتایج نادرست آزمون‌ها مخصوصا آزمون‌های سرنوشت‌ساز می‌تواند پیامدهای منفی زیادی هم برای داوطلبان و هم برای سازمان برگزارکننده داشته باشد (Dardick & Weiss, 2019). از جمله پیامدهای داوطلبانی که به اشتباه رد شده‌اند (خطای منفی کاذب) می‌توان به از دست دادن فرصت شغلی، افسردگی، اضطراب و همچنین مشکلات مالی از جمله ثبت‌نام برای آزمون مجدد و یا از دست دادن درآمد ناشی از پیشنهاد شغلی اشاره کرد. از سوی دیگر پذیرش داوطلبانی که صلاحیت قبولی در آزمون را ندارند (خطای مثبت کاذب) دارای عواقب منفی مانند کاهش کیفیت آموزش می‌شود چرا که نامزدهای فاقد صلاحیت که در آزمون پذیرفته می‌شوند، ممکن است نتوانند با دوره آموزشی هماهنگ شوند یا استانداردهای برنامه را برآورده کنند. این امر می‌تواند منجر به پایین آمدن استانداردهای تحصیلی و کاهش کیفیت آموزش شود که منجر به ایجاد فضای رقابتی کمتری می‌شود و موفقیت را برای نامزدهای واجد شرایط دشوارتر می‌کند. در نهایت پذیرش نامزدهای فاقد صلاحیت منجر به هدر رفتن منابع مالی و زمانی می‌شود. به عنوان مثال، نامزدهای فاقد صلاحیت ممکن است نیاز به گذراندن دوره‌های تقویتی یا آزمون مجدد داشته باشند که منجر به صرف هزینه و زمان می‌شود. به غیر از داوطلبان، نتایج اشتباه آزمون برای سازمان

برگزارکننده نیز پیامدهایی مانند از دست دادن اعتماد عمومی و هزینه‌های مالی از جمله هزینه آزمون مجدد را در پی دارد. لذا لازم است نمره برش به درستی و با دقت تعیین شود.

شاخص یودن معیاری را برای انتخاب مقدار آستانه "بهینه" (c^*) ارائه می‌دهد که مقداری بین ۰ و ۱ می‌گیرد و به صورت زیر تعریف می‌شود:

$$J = \max_{c}\{Se(c) + Sp(c) - 1\}\ ,\forall c \qquad (1)$$

طبق روش یودن، نقطه‌ای به عنوان نقطه برش انتخاب می‌شود که تابع یودن را ماکسیمم کند (Youden, 1950) که حداکثر فاصله یا اختلاف عمودی بین منحنی ROC و خط مورب یا شانس است (Schisterman et al., 2005). بنابراین طبق رابطه (۱) در روش یودن، نقطه برش بهینه c^* عبارت از نقطه c ی است که منجر به ماکسیمم مقدار تابع یودن به ازای همه مقادیر نقطه برش ممکن c موجود در مقادیر داده‌ها می‌شود.

مقدار شاخص یودن بین ۰ و ۱ است که جداسازی کامل توزیع مقادیر نشانگر برای جمعیت‌های قبول و مردود منجر به $J = 1$ می‌شود در حالی که همپوشانی کامل منجر به $J = 0$ می‌شود (Fluss et al., 2005). به عبارت دیگر مقدار شاخص یودن ۱ نشان‌دهنده یک آزمون کامل است، در حالی که مقدار شاخص یودن ۰ نشان دهنده یک آزمون بی فایده است که هیچگونه اطلاعاتی در مورد داوطلبان نمی‌دهد. بنابراین در صورتی که مقدار شاخص یودن برابر صفر باشد می‌توان نتیجه گرفت که طبقه‌بندی‌کننده بهتر از شانس عمل نکرده است.

برای برآورد منحنی ROC از رویکردهای پارامتریک، نیمه‌پارامتریک و ناپارامتریک استفاده شده است که با توجه به رابطه شاخص یودن با منحنی ROC ، این رویکردها با شاخص یودن نیز مرتبط هستند. در برآورد پارامتری، به این فرض نیاز هست که داده‌ها از کدام خانواده چگالی تولید می‌شوند. اغلب این فرض بر اساس شواهد کم یا بدون شواهد است و اگر فرض موردنظر اشتباه باشد، نتیجه نادرستی حاصل می‌شود. ولی در برآورد چگالی ناپارامتریک نیازی به چنین فرضی نیست. منظور از اصطلاح ناپارامتریک این نیست که چنین مدل هایی کاملا فاقد پارامتر هستند، بلکه تعداد و ماهیت پارامترها انعطاف‌پذیر هستند و از قبل ثابت نیستند. برآورد چگالی ناپارامتریک می‌تواند به عنوان یک تحلیل اولیه یا خود به عنوان یک ابزار تجزیه و تحلیل استفاده شود. ایده پشت برآورد چگالی ناپارامتریک قدیمی است، اما بسیاری از روش‌ها کاملا محاسباتی هستند. از آنجایی که

قدرت محاسباتی در چند دهه پیش به سرعت افزایش یافته است، روش‌های ناپارامتریک بیشتر مورد توجه قرار گرفته و پیشرفت‌های زیادی صورت گرفته است (Kile, 2010). Hsieh and Turnbull (۱۹۹۲) بدون مفروضات پارامتریک برای توزیع‌های اساسی، برآوردهای نقطه‌ای ناپارامتری را برای شاخص یودن بر اساس برآوردهای تجربی و هسته‌ای برای توزیع‌های زیربنایی مطالعه کردند. لذا در این مقاله نمره برش آزمون تولیمو با استفاده از شاخص یودن و بر اساس دو روش برآورد ناپارامتری تجربی و هسته به دست آورده شد و نتایج دو روش بر اساس دو شاخص خطای استاندارد بوت‌استرپ و ریشه میانگین مربعات خطا ارزیابی و مقایسه شدند. از آنجایی که در روش هسته، مقدار پهنای باند تاثیر زیادی در نتیجه دارد لذا از دو روش تعیین پهنای باند در روش هسته استفاده شد و در نهایت نتایج با استفاده از معیار میانگین خطای مجذور یکپارچه[1]($MISE$) و خطای مجذور یکپارچه[2] (ISE) مقایسه و تحلیل شدند. لازم به ذکر است برای دسترسی به نتایج دقیق‌تر برای تعیین نمره برش بهینه متقاضیان فرصت مطالعاتی تمام نمره‌های صحیح از ۴۷۰ تا ۵۰۰ در نظر گرفته شدند. روش‌های مختلف تخمین ناپارامتریک از جمله روش هسته با پهنای باند سیلورمن، روش هسته با حداکثر احتمال پهنای باند اعتبار سنجی متقاطع و روش تجربی در تعیین نمره برش آزمون تولیمو با استفاده از روش مقایسه ای چگونه است و کدام روش تخمین‌های قابل اعتمادتری را ارائه می‌دهد؟

پیشینه پژوهش

منحنی ویژگی عملکرد [3] (ROC) یک روش گرافیکی محبوب است که از دهه ۱۹۷۰ باارزش‌ترین ابزار آماری برای توصیف و مقایسه دقت آزمون‌های تشخیصی برای تمایز بین دو جمعیت شناخته شده است (Zhou et al., 2009) به طوری که در بسیاری از زمینه‌های علمی از قبیل رادیولوژی (Metz, 1989)، روانپزشکی (Hsiao et al., 1989)، اپیدمیولوژی (Aoki et al., 1997) و سیستم‌های بازرسی ساخت (Somoza et al., 1990) استفاده شد. در اواخر دهه ۱۹۸۰، محققان شروع به استفاده از روش منحنی‌های ROC برای ارزیابی آزمون‌های تشخیصی پزشکی کردند (Hanley & McNeil, 1982). از منحنی ROC برای مشکلات زیست‌پزشکی به منظور بررسی اثربخشی نشانگرهای

1 .Mean integrated squared error
2. Integrated square error
3. receiver operating characteristic

تشخیصی پیوسته در تشخیص افراد بیمار و سالم بسیار استفاده شده است (Shapiro, 1999; Greiner et al., 2000). اگر مقدار نشانگر آزمایش‌شده بیشتر از مقدار آستانه معین باشد، فرد بیمار (مثبت) ارزیابی می‌شود، در غیر این صورت آزمودنی سالم (منفی) تشخیص داده می‌شود. دقت هر مقدار آستانه معین را می‌توان با احتمال مثبت واقعی (حساسیت[1]) و احتمال منفی واقعی (وضوح[2]) اندازه‌گیری کرد (Fluss et al., 2005).

منحنی ROC نمودار حساسیت (Se(c)) را در مقابل (وضوح – ۱) (1-Sp(c)) روی تمام مقادیر آستانه ممکن (c)نشانگر رسم می‌کند. برای ارزیابی توانایی تمایز یک نشانگر، خلاصه کردن اطلاعات منحنی ROC در یک مقدار یا شاخص کلی رایج است. چندین شاخص از جمله مساحت زیر منحنی ROC (AUC) و شاخص یودن در ادبیات یافت می‌شود که در کاربردهای مختلف مورد استفاده قرار گرفته شده‌اند (Shapiro, 1999; Aoki et al., 1997; Greiner et al., 2000; Grmec & Gašparovic, 2000).

با توجه به اهمیت تعیین نمره برش، تحقیقات زیادی بر روی تعیین تجربی نقطه‌های برش بهینه متمرکز شده است که در ادامه به برخی از تحقیق‌هایی که با روش یودن انجام شده است اشاره شده است.

Eckes (۲۰۱۷) به منظور تعیین نمرات برش در آزمون تعیین سطح زبان انگلیسی از ترکیب دو رویکرد منحنی ویژگی عملکرد و روش گروهی نمونه اولیه[3]استفاده کرد. روش گروهی نمونه اولیه، یک رویکرد اندازه‌گیری راش[4] را برای تجزیه و تحلیل مهارت آزمون شونده با مفهوم نمونه‌های اولیه برگرفته از تحقیقات در مورد قضاوت و طبقه‌بندی انسان ترکیب می‌کند. کارشناسان ابتدا زبان‌آموزانی را شناسایی می‌کنند که در هر یک از پنج سطح مهارت زبانی مشخص شده توسط چارچوب مرجع مشترک اروپایی[5] برای زبان‌ها مشخص شده است. به عبارت دیگر برای این مجموعه از نمونه‌های اولیه، برآوردهای مهارت زبان حاصل از مدل راش به عنوان ورودی برای تحلیل منحنی ویژگی عملکرد استفاده شد. در نهایت نمره‌های برش با استفاده از شاخص یودن به دست آمد که نرخ کلی طبقه‌بندی صحیح را به حداکثر می‌رساند و نرخ کلی طبقه‌بندی اشتباه را به حداقل می‌رساند. وی نتیجه

1. sensitivity
2. specificity
3. Prototype group method (PGM)
4 . Rasch
5. Common european framework of reference (CEFR)

گرفت که این روش امکان تنظیم نمرات برش را فراهم می‌کند که سطح بالایی از دقت طبقه‌بندی را از نظر مطابقت با طبقه‌بندی‌های متخصص نمونه‌های اولیه آزمون‌شونده نشان می‌دهد. علاوه بر این، نمرات برش مبتنی بر منحنی ویژگی عملکرد با دقت طبقه‌بندی بالاتری نسبت به نمره‌های برش حاصل از تجزیه و تحلیل رگرسیون لوجستیک از همان داده‌ها مرتبط بود.

Nakas و همکاران (2010) تعمیمی از شاخص یودن به نام J3 را برای مشکل طبقه‌بندی داده‌ها به سه کلاس برای انتخاب یک نقطه برش بهینه پیشنهاد کردند. آنها هر دو روش پارامتری و ناپارامتریک را برای تخمین و آزمایش J3 برای داده‌های مربوط به بیماران با ویروس نقص ایمنی بررسی کردند که منجر به نتایج امیدوارکننده‌ای برای طبقه‌بندی دقیق بیماران در گروه‌های مختلف شد.

Luo and Xiong (2013) تعمیمی از شاخص یودن را برای طبقه‌بندی داده‌ها به سه گروه تشخیصی ترتیبی پیشنهاد کردند و از روش‌های پارامتریک و ناپارامتریک برای تخمین شاخص یودن بهینه و نقاط برش مربوطه و فواصل اطمینان متناظر استفاده کردند.

Carvalho and Branscum (2018) یک رویکرد ناپارامتریک بیزی برای تخمین شاخص یودن سه کلاسه و مقادیر برش بهینه متناظر آن برای ارزیابی اختلال شناختی در بیماران مبتلا به پارکینسون ارائه کردند. آنها از روش مبتنی بر مخلوط‌های فرآیند دیریکله[1] برای تخمین شاخص یودن سه کلاسه و مقادیر برش بهینه استفاده کردند.

نتایج حاصل از این مطالعه اهمیت انتخاب یک رویکرد قابل اعتماد برای تعیین نمره برش را نشان می‌دهد که پیامدهای مهمی برای تجزیه و تحلیل آزمون و تصمیم‌گیری در محیط های آموزشی دارد. شاخص یودن یک معیار بسیار مورد استفاده در حوزه تشخیص پزشکی است، جایی که کارایی یک نشانگر زیستی (نشانگر غربالگری یا پیش‌بینی کننده) در طبقه بندی وضعیت بیماری مورد بررسی قرار می‌گیرد؛ تحقیقات محدودی کاربرد این تکنیک را در زمینه‌های آموزشی بررسی کرده است. در نتیجه مقایسه و ارزیابی تکنیک‌های مختلف تخمین ناپارامتریک برای محاسبه شاخص یودن و سنجش قابلیت اطمینان آنها در تقریب نمره برش در محیط‌های آموزشی می‌تواند در تعیین نمره برش آزمون‌های ملاک‌مرجع علی‌الخصوص از نوع سرنوشت‌ساز مؤثر باشد.

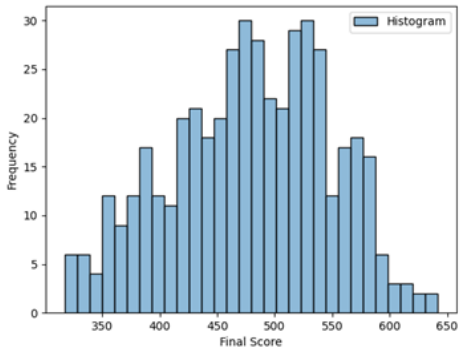
روش‌های برآورد ناپارامتری

1 . Dirichlet process mixtures (DPM)

اکثر روش‌ها برای برآورد نقاط برش بهینه با استفاده از شاخص یودن به عنوان متریک توسعه داده شده‌اند (Fluss et al., 2005; Leeflang et al., 2008). از بین روش‌های ناپارامتری می‌توان گفت روش‌های هیستوگرام و ناپارامتری تجربی قابل استفاده برای معیارهای دیگر غیر از شاخص یودن نیز هستند و روش هموارسازی مبتنی بر هسته مختص برآورد شاخص یودن است (Thiele & Hirschfeld, 2020).

قدیمی‌ترین و پرکاربردترین برآوردگر چگالی احتمال ناپارامتری، هیستوگرام است که به اوایل قرن نوزدهم برمی‌گردد. با توجه به مبدا x_0 و a بلوک به عرض h ، بلوک‌های هیستوگرام به صورت بازه‌های $[x_0 + mh\,, x_0 + (m+1)h]$ به ازای اعداد صحیح مثبت و منفی m تعریف می‌شوند (Silverman, 2018). به عبارت دیگر در این روش، با تقسیم داده‌ها به بازه‌های یکنواخت و شمارش تعداد مشاهدات در هر بازه، برآوردی از چگالی احتمال به دست آورده می‌شود. بنابراین روش هیستوگرام تنها یک برآورد تقریبی از چگالی احتمال است و وابسته به تعداد و انتخاب بازه‌ها است. نمودار مربوط به داده‌های این مقاله با ۳۰ بازه در شکل ۱ نشان داده شده است.

شکل ۱. برآورد هیستوگرام نمره کل آزمون تولیمو با ۳۰ بازه



به منظور شرح روش‌های تجربی و هسته، فرض کنید که نتایج یک آزمون تشخیصی (نشانگر) $x_1, \ldots, x_m$ و $y_1, \ldots, y_n$ از دو نمونه تصادفی در جمعیت‌های قبول (P) و مردود (R) با توابع توزیع تجمعی F_P و G_R در دسترس هستند. برای هر آستانه معین c می‌توان حساسیت و وضوح نمره برش را بر حسب توابع توزیع تجمعی به صورت زیر تعریف کرد:

$$Se(c) = 1 - G_R(c) \qquad , \qquad Sp(c) = F_P(c)$$

بنابراین شاخص یودن عبارت می‌شود از:

$$J = Se(c) + Sp(c) - 1 = F_P(c) - G_R(c) \qquad (2)$$

مقدار c که منجر به مقدار ماکسیمم J می‌شود به عنوان آستانه بهینه c^* در نظر گرفته می شود. برآورد J با برآورد G_R و F_P (یعنی $(\hat{G}_R, \hat{F}_P)$) و جایگزینی این برآوردها در معادله (۲) یعنی $\hat{J} = \max_c\{\hat{F}_P(c) - \hat{G}_R(c)\}$ حاصل می‌شود. در این مقاله برای برآورد G_R و F_P و تعیین مقدار شاخص یودن و تعیین نمره برش آزمون تولیمو از روش‌های برآورد ناپارامتری تجربی (EMP) و هسته (KDE) استفاده می‌شود.

روش تجربی (EMP)

ساده‌ترین روش برای بهینه‌سازی یک متریک، روش ناپارامتری تجربی است که نقطه برشی را برمی‌گزیند که مقدار متریک بهینه را در نمونه به دست می‌دهد بدین‌صورت که در تمام نقاط برش جستجو می‌کند و در نهایت نقطه برشی را انتخاب می‌کند که منجر به مقدار متریک بهینه در نمونه می‌شود. بنابراین تعیین نمره برش با استفاده از روش برآورد تجربی معادل انتخاب یک نمره برش بر اساس تجزیه و تحلیل منحنی ROC (هموار نشده) است (Thiele & Hirschfeld, 2020).

شاخص یودن در روش EMP مستقیما از داده‌های مشاهده شده بدون هیچ فرض توزیعی محاسبه می‌شود. در روش تجربی از تابع توزیع تجربی به منظور برآورد تابع توزیع تجمعی احتمالات پیش‌بینی شده از یک مدل طبقه‌بندی باینری و بر اساس داده‌های مشاهده‌شده استفاده می‌شود. به عبارت دیگر در روش تجربی، تابع توزیع تجمعی نشانگر با تابع توزیع تجمعی تجربی نمونه برآورد می‌شود (Fluss et al., 2005). تابع توزیع تجمعی تجربی به صورت زیر محاسبه می‌شوند:

$$\hat{F}_P(c) = \frac{1}{n}\sum_{i=1}^{n} I(y_i \le c) \qquad , \qquad \hat{G}_R(c) = \frac{1}{m}\sum_{i=1}^{m} I(x_i \le c)$$

که در آن؛

$$I(u \le c) = \begin{cases} 1 & u \le c \\ 0 & u > c \end{cases}$$

پس از محاسبه تابع توزیع تجمعی تجربی هر یک گروه‌های قبول و مردود، J توسط رابطه (۲) به دست می‌آید و در نهایت مقدار c که مقدار $\hat{J}$ را بیشینه می‌کند به عنوان c^* تعیین می شود (Ruopp et al., 2008).

روش هسته (KDE)

روش ناپارامتری هسته، برای اولین بار توسط Parzen (۱۹۶۲) معرفی شد که برای برآورد تابع چگالی استفاده می‌شود که یکی از روش‌های متداول برای هموارسازی است. این روش نسبت به روش تابع توزیع تجربی دارای دقت بسیار بیشتری است (Silverman, 2018). یکی از مزایای این روش این است که تخمین هموارای از تابع توزیع تجمعی را ارائه می‌دهد. (Fluss et al., 2005). فرض کنید متغیرهای تصادفی $X_1, X_2, \ldots, X_n$ مستقل و هم‌توزیع با تابع چگالی نامعلوم f باشند. یک برآوردگر چگالی هسته[1] تک متغیره (KDE) به صورت رابطه (۳) تعریف می‌شود.

$$\hat{f}_n(t) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{t - X_i}{h}\right), \quad t \in R \ , \ h > 0 \qquad (3)$$

که در آن $K(.)$ تابع هسته نامیده می‌شود که به طور کلی یک تابع هموار و متقارن مانند مثلثی، مستطیلی، گاوسی، اپاچنیکوف[2] و ... است و $h > 0$ پهنای باند یا ضریب هموارسازی نامیده می‌شود که میزان هموارسازی را کنترل می‌کند و برای عملکرد برآوردگر بسیار مهم است، t یک مقدار آستانه[3] و n، حجم نمونه است. بنابراین هدف از برآورد چگالی هسته، برآورد چگالی احتمال $f(t)$ با استفاده از نمونه تصادفی $X_1, X_2, \ldots, X_n$ است (Park & Marron, 1990). انتخاب پهنای باند چالش برانگیزترین مرحله در ساخت برآوردگر چگالی با استفاده از روش هسته است. زیرا اگر حجم نمونه بزرگ باشد و داده‌ها به طور فشرده[4] توزیع شده باشند، بهتر است مقدار بزرگی برای h انتخاب شود تا چگالی هسته بیشترین احتمال را به هر نقطه بدهد. اما وقتی حجم نمونه کوچک است و داده‌ها پراکنده[5] هستند، بهتر است مقدار کوچکی برای h انتخاب شود تا چگالی هسته احتمالات بیشتری را در مقادیر همسایه یک نقطه پخش کند. به عبارت دیگر پهنای کم هسته به این معناست که میانگین گیری در هر نقطه بر روی تعداد کمی از مشاهدات انجام می‌شود و لذا این برآورد، توجه زیادی به نمونه‌ها داشته و به نمونه اجازه تغییرپذیری نمی‌دهد. چنین برآوردی، کم‌برآورد[6] نامیده می‌شود. بزرگ‌تر شدن پهنای هسته، منجر به برآورد هموارتری می‌شود

1. Kernel Density Estimator
2 . Epanechnikov
3 . threshold value
4 . tightly
5 . sparse
6 . underestimation

به طوری که اگر به میزان کافی، پهنای باند بزرگ شود و در واقع مقدار مناسبی برای پهنای باند انتخاب شود ساختار اساسی چگالی داده‌ها قابل شناسایی است (Wand & Jones, 1994). از سوی دیگر کارایی مجانبی برآوردگر چگالی برای توابع مختلف هسته توسط Silverman (۱۹۸٦) بررسی شده است (Silverman, 2018). او نشان داد که با فرض انتخاب بهینه برای پارامتر هموارسازی، بازده مجانبی توابع مختلف هسته نزدیک به یک است و تفاوت چندانی بین آنها وجود ندارد. بنابراین در عمل انتخاب نوع تابع هسته تاثیر چندانی ندارد و انتخاب بهینه پارامتر هموارسازی اهمیت بسیار بیشتری در دقت برآورد چگالی خواهد داشت. از این رو، در این مقاله از تابع هسته گاوسی که دارای خواص مطلوبی از جمله تقارن است و نسبت به دیگر توابع هسته متداول‌تر است استفاده شده است (Trosset, 2009). با انتخاب تابع هسته گاوسی در رابطه (۳)، $K(t) = \Phi(t)$ می‌شود که نشانگر یک توزیع نرمال استاندارد $N(0,1)$ است که بایستی برای هر دو گروه قبول و مردود با حجم نمونه n و m نوشته شود. طبق رابطه (۳) و با فرض تابع هسته گاوسی، مشاهده می‌شود که پارامتر h در مخرج توابع توزیع بالا نقش انحراف استاندارد توزیع نرمال را دارد. یعنی h نقش مهمی به عنوان عامل مقیاس دارد که پراکندگی تابع هسته را مشخص می‌کند. در نهایت بعد از محاسبه توابع هسته مربوط به گروه‌های قبول و مردود طبق رابطه (۳)، مقدار c که مقدار $\hat{f}$ موجود در رابطه (۲) را بیشینه می‌کند به عنوان c^* تعیین می‌شود (Zou et al., 1998).

دو روش تعیین پهنای باند

در برآورد چگالی هسته، انتخاب پارامتر پهنای باند بسیار مهم است زیرا همواری و دقت چگالی برآورد شده را تعیین می‌کند. روش سرانگشتی سیلورمن متداول‌ترین روش تعیین پهنای باند است که زمانی که تابع هسته و توزیع واقعی نرمال فرض شود بر اساس میانگین مجانبی خطای مربع یکپارچه ($AMISE$) عمل می‌کند (Węglarczyk, 2018). این روش با فرض وجود یک تابع هسته در هر نقطه مشاهداتی و تجمیع آنها به همراه مقدار مناسب پهنای باند، تابع چگالی احتمال را به روش ناپارامتری محاسبه می‌کند (Adamowski, 1987). پهنای باند بهینه در این روش بر اساس انحراف استاندارد نمونه و حجم نمونه برآورد می‌شود. طبق این روش پهنای باند با فرض هسته گاوسی برای نمونه‌ای به حجم n از رابطه (۴) به دست می‌آید.

$$h = 1/06\, s\, n^{-0/2} \qquad (4)$$

رابطه (۴) نسبت به داده پرت حساس است چرا که داده پرت منجر به بزرگ شدن مقدار s می‌شود و در نتیجه برآورد بزرگ‌تری از پهنای باند در برآورد تابع چگالی حاصل می‌شود. لذا از دامنه میان چارکی داده‌ها استفاده کرده و رابطه زیر به عنوان یک برآوردگر نیرومند برای تعیین h بهینه با استفاده از روش سرانگشتی سیلورمن حاصل می‌شود:

$$h_{opt} = 1/06n^{-0.2} \times min\left\{s, \frac{iqr}{1.34}\right\} \qquad (5)$$

که در آن s انحراف معیار نمونه و iqr دامنه بین چارکی است که از تفاضل چارک سوم از چارک اول نمونه داده‌ها به دست می‌آید (Thiele & Hirschfeld, 2020; Zucchini et al., 2003).

رویکرد دیگر برای تعیین پهنای باند، روش اعتبارسنجی متقابل ماکسیمم درستنمایی[1] است که توسط Habbema و همکاران (۱۹۷٤) و Duin (1976) پیشنهاد شد. طبق این روش، برای انتخاب h بایستی شبه درستنمایی[2] $\prod_{i=1}^{n} \hat{f}_h(X_i)$ ماکسیمم شود. در این روش تابع هسته بر روی زیرمجموعه‌ای از X_j بر اساس رویکرد اعتبارسنجی متقابل leave-one-out[3] برآورد می‌شود. بدین‌صورت که به تعداد حجم نمونه، داده‌ها تقسیم می‌شوند وهر بار به طورسیستماتیک یک مشاهده از یک مجموعه داده برای ارزیابی و آزمون کنار گذاشته می‌شود. به عنوان مثال مجموعه داده‌ای به حجم n طوری تقسیم می‌شود که هر بار تنها یک داده برای ارزیابی وآزمون نگه داشته می‌شود و از سایر داده‌ها برای آموزش مدل استفاده می‌شود و این کار n بار تکرار می‌شود. تابع هدف که $MLCV$ را به حداکثر می رساند به صورت زیر بیان می‌شود:

$$MLCV(h) = \frac{1}{n}\sum_{i=1}^{n} log\left[\sum_{j} K\left(\frac{X_j - x_i}{h}\right)\right] - log\left[(n-1)h\right]$$

که در آن، $h_{mlcv} = \underset{h>0}{\operatorname{argmax}}\, MLCV(h)$.

خوانندگان به منظور بررسی عمیق روش‌های انتخاب پهنای باند می‌توانند به Hall (1982)، Park and Marron (1990)، Van Es (1991)، Jones (1996) و همکاران، Loader (1999)، Heidenreich (2013) و همکاران، Barbeito and Cao (۲۰۲۰) مراجعه کنند.

---

1. Maximum likelihood cross-validation
2. pseudo-likelihood
3 . Leave-one-out cross validation ($LOOCV$)

معیارهای ارزیابی پهنای باند

همه روش‌های تعیین باند بر اساس حداقل خطا عمل می‌کنند بدین‌معنی که پهنای باندی برآورد می‌شود که دارای حداقل خطا باشد. در این مقاله فرض می‌شود که پارامتر هموارسازی در تمام مقادیر x ثابت است بنابراین منطقی است که از یک معیار خطای کلی برای انتخاب پهنای باند استفاده شود (Kile, 2010). بدین‌منظور از معیار میانگین خطای مجذور یکپارچه ($MISE$) استفاده شد که به صورت رابطه (۶) تعریف می‌شود:

$$MISE(h) = E\left[\int \left(\hat{f}_h(x) - f(x)\right)^2 dx\right] \qquad (6)$$

هدف ما یافتن h ی است که حتی برای نمونه‌های خیلی کوچک، $MISE$ را کمینه کند. پهنای باندی که منجر به حداقل $MISE(h)$ می‌شود با h_{MISE} نشان داده می‌شود و به عنوان پهنای باند $MISE$ داده‌ها در نظر گرفته می‌شود. برای یافتن h بهینه از رابطه (۶) برای برآوردگر هسته لازم است رابطه مجانبی برای $MISE$ پیدا شود که سرراست‌تر باشد که بدین‌منظور از بسط تیلور استفاده می‌شود. لذا رابطه (۶) به صورت زیر نوشته می‌شود:

$$\begin{aligned} MISE(h) &= \int E\left[\left(\hat{f}_h(x) - f(x)\right)^2\right] dx \\ &= \int \left[E\left(\hat{f}_h(x) - f(x)\right)\right]^2 dx + \int Var\left(\hat{f}_h(x)\right) dx \\ &= \int \left[Bias\left(\hat{f}_h(x)\right)\right]^2 + \int Var\left(\hat{f}_h(x)\right) dx \end{aligned}$$

که در آن؛

$$\begin{aligned} Bias\left(\hat{f}_h(x)\right) &= E\left(\hat{f}_h(x)\right) - f(x) \\ &= E[K_h(x - X)] - f(x) \\ &= \int K_h(x - X) f(x) dx - f(x) \end{aligned}$$

و طبق رابطه $Var(X) = E(X^2) - [E(X)]^2$ خواهیم داشت:

$$Var\left(\hat{f}_h(x)\right) = \frac{1}{nh^2} Var\left(K\left(\frac{x - X}{h}\right)\right)$$

$$= \frac{1}{n}\int \frac{1}{h^2} K^2\left(\frac{x - X}{h}\right) f(x) dx$$
$$- \frac{1}{n}\left[\frac{1}{h}\int K\left(\frac{x - X}{h}\right) f(x) dx\right]^2$$

معیار میانی بین $MISE$ و MSE ، خطای مجذور یکپارچه (ISE) است که یک اندازه گیری ناهمخوانی[1]است که برای برآورد مقدار پارامتر هموارسازی استفاده می‌شود و به صورت زیر تعریف می‌شود:

$$ISE(h) = \int_{-\infty}^{+\infty} \left(\hat{f}_h(x) - f(x)\right)^2 dx$$

داده‌ها : آزمون تولیمو

آزمون تولیمو[2] یکی از آزمون‌های استاندارد و معتبر تعیین سطح دانش زبان انگلیسی دانشجویان داخل ایران است که توسط سازمان سنجش طراحی و برگزار می‌شود. دانشجویان برای تحصیل در مقاطع کارشناسی‌ارشد و دکترا در رشته‌های مختلف دانشگاهی در ایران و یا برای اخذ بورس تحصیلی و نیز افرادی که در جستجوی کار هستند به منظور استخدام در برخی موسسات و ارگان‌های دولتی باید در یکی از آزمون‌های معتبر زبان مانند آزمون تولیمو، آزمون تافل، آزمون $MCHE$ و یا هر آزمون دیگری نمره قابل قبولی را کسب کنند که این نمره مورد قبول در هر یک از این آزمون‌ها را هر یک از موسسات و دانشگاه‌ها بر اساس انتظارات خود تعیین می‌کنند. به عنوان نمونه در حال حاضر برای فرصت مطالعاتی دانشجویان دکتری، حداقل نمره قبولی موردنیاز برای آزمون تولیمو ۴۸۰ تعیین شده است.

در آزمون تولیمو هر چهار مهارت گرامر، شنیداری، درک مطلب و نوشتاری مورد آزمون قرار می‌گیرد. بدین‌صورت که در جلسه آزمون تولیمو دو دفترچه در اختیار داوطلبان قرار داده می‌شود که بخش نوشتاری در یک دفترچه و شامل یک سوال تشریحی و سه بخش دیگر در دفترچه دیگر و به صورت ۱۰۵ سوال در اختیار داوطلبان قرار می‌گیرد که هر بخش شامل ۳۵ سوال چهار گزینه‌ای است. بخش تشریحی آزمون تولیمو به بخش نوشتاری آزمون آیلتس و آزمون جی‌آرای (GRE) شباهت دارد و نشان‌دهنده مهارت داوطلبان در نوشتن متن و مقاله به زبان انگلیسی است. از آنجایی که در این آزمون نمره منفی وجود ندارد لذا معمولا داوطلبان به همه سوال‌های چهارگزینه‌ای پاسخ می‌دهند. نمره کل داوطلبان در بازه ۲۰۰ تا ۶۷۷ قرار می‌گیرد.

در این پژوهش، از نمره کل داوطلبان یک دوره از آزمون تولیمو استفاده شده است که در دوره مزبور، ۴۶۱ داوطلب از رشته‌ها و دانشگاه‌های مختلف شرکت‌کرده‌اند. دامنه نمره‌های کسب شده در این دوره از آزمون تولیمو از ۳۱۷ تا ۶۴۲ بوده است و جمعا ۱۵۳

1 . discrepancy
2 . The test of language by the Iranian measurement organization

نمره به دست آورده شده است که با رسم جدول فراوانی مشخص شد مد داده‌ها نمره ۴۷۰ است که ۹ نفر این نمره را کسب کرده‌اند. میانگین و میانه نمره‌های کسب شده به ترتیب برابر ۴۷۸ و ۴۸۲ است که بیانگر این مطلب است که عملکرد معمولی داوطلبان در حدود محدوده متوسط نمرات است. از سوی دیگر انحراف استاندارد نمره‌ها برابر ۶۹ است که می‌توان اینگونه تفسیر کرد که تنوع متوسطی در نمرات وجود دارد و برخی از داوطلبان به طور قابل توجهی بالاتر از آن نمره یا کمتر از مقدار میانگین (یعنی ۴۷۸) را گرفته‌اند. مقادیر آماره‌های توصیفی داده‌های استفاده شده در این مطالعه در جدول ۱ آورده شده است.

جدول ۱. آماره‌های توصیفی داده‌ها

میانگین	میانه	مد	چارک اول	چارک سوم	انحراف‌استاندارد
۴۷۸	۴۸۲	۴۷۰	۴۳۲	۵۲۹	۶۹

روش

داده‌ها با استفاده از نرم‌افزار اکسل و محاسبات با استفاده از نرم‌افزار پایتون و با به‌کارگیری کتابخانه‌های *pandas*، *numpy*، *sklearn*، *seaborn* و *matplotlib* انجام شدند.

داده‌های این مقاله در ۳۱ فایل اکسل بدین‌صورت ذخیره شدند که از دستور *if* اکسل برای تعیین وضعیت پذیرش یا رد هر داوطلب بر اساس نمره نهایی آنها در رابطه با هر یک از اعداد طبیعی در بازه ۴۷۰ تا ۵۰۰ مثلا ۴۷۰ استفاده شد. داوطلبانی که نمرات نهایی بالاتر از ۴۷۰ داشتند پذیرفته در حالی که آنهایی که نمرات نهایی کمتر از ۴۷۰ داشتند رد شدند. این نتایج به ترتیب با ۱ و ۰ نمایش داده شدند. بنابراین هر فایل شامل دو ستون نمره نهایی و برچسب قبول و رد بر حسب نمره برش انتخابی شد که برای هر فایل به طور مجزا مقدار شاخص یودن و به منظور برآورد پایایی و میزان تغییرپذیری روش‌ها، خطای استاندارد بوت استرپ و ریشه میانگین مربعات خطا برای هر سه روش تجربی، روش هسته با روش پهنای باند سیلورمن و روش هسته با تعیین پهنای باند به روش اعتبارسنجی متقابل ماکسیمم درستنمایی به دست آورده شد. خطاهای استاندارد بوت استرپ برآوردی از تغییرپذیری برای مقادیر شاخص یودن را ارائه می‌دهد. به منظور محاسبه خطای استاندارد بوت استرپ از مجموعه داده خود، k نمونه مکرر و با جایگذاری گرفته و برای هر نمونه، خطای استاندارد از طریق رابطه $\frac{s}{\sqrt{n}}$ محاسبه شد. میانگین k خطای استاندارد برابر با خطای استاندارد بوت‌استرپ در نظر گرفته شد. از آنجایی که مبنای اساسی ریشه میانگین مربعات خطا در

اختلاف بین مقدار پیش‌بینی شده حاصل از مدل یا برآوردگر آماری و مقدار واقعی است لذا ستون برچسب به عنوان متغیر وابسته واقعی در نظر گرفته شد. در مورد پیش‌بینی متغیر وابسته، روش زیر را اجرا کردیم: در مرحله اول، مدل مربوطه را برای هر یک از گروه‌های قبول و مردود برازش دادیم. با توجه به اینکه انتظار می‌رود گروه پذیرفته‌شدگان نمره بالاتری نسبت به گروه مردودی‌ها کسب کرده باشند، به این افراد مقدار ۱ در متغیر وابسته پیش‌بینی شده و در غیر این صورت مقدار ۰ اختصاص داده شد. در ادامه مراحل اجرای هر یک از روش‌ها (EMP و KDE) که برای هر یک از ۳۱ فایل داده به طور جداگانه اجرا شد شرح داده می‌شود.

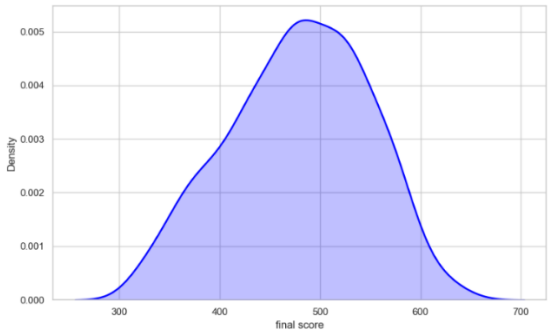
برای برآورد شاخص یودن با استفاده از روش EMP، تابع توزیع تجمعی تجربی برای هر دو گروه مردود (۰) و قبول (۱)، نرخ مثبت کاذب (FPR) و نرخ مثبت واقعی (TPR) محاسبه شد. مقادیر ایجاد شده در دو ستون با عنوان‌های FPR و TPR در یک فایل اکسل ذخیره شدند. در مرحله بعد ماکسیمم قدرمطلق تفاضل این دو ستون به دست آورده شد. به عنوان مثال ۲۰۳ امین مقدار دارای ماکسیمم $|FPR - TPR|$ شد. در نهایت با صعودی کردن ستون مربوط به نمره کل و پیدا کردن ۲۰۳ امین عدد که در فایل ما ۴۷۰ بود، نمره برش برای فایل مربوطه تعیین شد.

برای برآورد شاخص یودن با استفاده از روش هسته و با به کارگیری روش سرانگشتی سیلورمن ابتدا پیش‌فرض نرمال بودن داده‌ها بررسی شد. در واقع توزیع و نقاط پرت در داده ها به منظور برآورد شاخص یودن با استفاده از رویکرد ناپارامتریک هسته مورد بررسی قرار گرفت. برای شناسایی نقاط پرت بالقوه در مجموعه داده‌ها از روش دامنه بین چارکی (IQR) استفاده شد. در این مقاله داده‌ای به عنوان نقطه پرت در نظر گرفته شد که خارج از محدوده مقادیر بین $Q_1 - 1.5IQR$ و $Q_3 + 1.5IQR$ قرار بگیرد. از شکل ۲ مشخص است که توزیع داده‌ها تقریبا نرمال است. در نهایت مشخص شد که داده‌های مربوط به نمره کل حاوی داده پرتی نیستند. لذا از روش سرانگشتی سیلورمن برای تعیین پهنای باند روش ناپارامتری هسته استفاده شد.

برای هر نمره برش بین ۴۷۰ تا ۵۰۰، مقدار هسته بر اساس تابع هسته گاوسی و با به کارگیری روش پهنای باند مرتبط برای هر دو گروه قبول و مردود برآورد شد و در نهایت ماکسیمم تفاضل مقادیر حاصل به عنوان مقدار شاخص یودن در نظر گرفته شد.

لازم به ذکر است که رویکرد $MLCV$ با استفاده از تابع $GridSearchCV$ موجود در کتابخانه $sklearn$ پیاده‌سازی شد. $GridSearchCV$ فرآیند تنظیم فراپارامتر را به منظور تعیین مقادیر بهینه برای یک مدل معین انجام می‌دهد. از آنجایی که عملکرد مدل هسته به طور قابل توجهی به مقدار پهنای باند بستگی دارد لذا با استفاده از تابع $GridSearchCV$، ۱۰۰۰ مقدار در بازه ۱۰ به توان اعداد بین ۱- و ۱ امتحان و مقدار بهینه پهنای باند بهینه انتخاب شد.

شکل ۲. نمودار تابع چگالی نمره کل آزمون تولیمو



یافته‌ها

داده استفاده شده در این پژوهش شامل نمره کل ۴۶۱ داوطلب آزمون تولیمو است که در این آزمون شرکت کرده‌اند. به منظور اجرای پژوهش، ستونی با مقادیر واقعی قبول و رد بر اساس نمره‌های ۴۷۰ تا ۵۰۰ تشکیل شد که دارای مقادیر ۰ (مردود) و ۱ (قبول) بودند. در مجموع ۳۱ فایل عددی حاصل شد که هر فایل داده برای هر روش استفاده شد و در نهایت بهترین مقدار برای هر روش بر اساس ماکسیمم مقدار شاخص یودن انتخاب گردید. نتایج برآورد و خطاهای استاندارد بوت‌استرپ (BSE) و ریشه میانگین مربعات خطا $(RMSE)$ برای این ۳۱ فایل داده که برگرفته از خروجی ۹۳ کد است در جدول ۱ آورده شده است.

جدول ۲. برآورد J و c^* برای آزمون تولیمو

	$KDE-Silverman$	$KDE-MLCV$	EMP
J	۰/۶۳	۰/۷۵	۰/۳۱
$RMSE$	۰/۷۱۵	۰/۷۲	۰/۶۸
BSE	۰/۰۰	۰/۰۳۹	۰/۰۱۷
c^*	۴۷۹	۴۷۹	۴۶۵

بر اساس نتایج نوشته شده در جدول ۲ مشاهده می‌شود که مقدار شاخص یودن برای روش هسته با پهنای باند $MLCV$ از دو روش دیگر بیشتر است ولی نمره برش به دست آمده در هر دو روش هسته ۴۷۹ به دست آمده است که تنها یک نمره با میانگین نمره‌ها تفاوت دارد و در روش تجربی برابر ۴۶۵ شد که از مد نمره‌ها یعنی ۴۷۰ تفاوت کمتری دارد. شاخص یودن یک معیار عملکردی است که معمولا در آزمایش‌های تشخیصی از جمله آزمون‌های ملاک‌مرجع مورد استفاده قرار می‌گیرد که نشان‌دهنده توانایی یک آزمون برای طبقه‌بندی صحیح افراد به دو گروه (قبول و مردود) است. مقدار شاخص یودن بالاتر نشان دهنده توانایی تشخیص بهتر افراد قبول و مردود است بنابراین بر اساس شاخص یودن، روش هسته با پهنای باند $MLCV$ روش بهتری برای طبقه‌بندی داوطلبان به دوگروه قبول و رد است. از سوی دیگر مقدار ریشه میانگین مربعات خطا و خطای استاندارد بوت استرپ روش هسته با پهنای باند $MLCV$ از دو روش دیگر کمی بیشتر است. خطای استاندارد بوت‌استرپ که برآوردی از تغییرپذیری نمونه‌برداری هر روش است در روش هسته با پهنای باند سیلورمن صفر شده است بدین معنی روش سیلورمن منجر به برآوردهای پایدارتر و دقیق‌تری شده است. لذا شاخص‌های ارزیابی هسته ISE و $MISE$ برای این دو روش محاسبه شد که نتایج به همراه پهنای باند بهینه برآورد شده این دو روش در جدول ۳ آورده شده است.

جدول ۳. ارزیابی عملکرد روش‌های هسته با نمره برش ۴۷۹ و پهنای باند بهینه

شاخص‌های ارزیابی	$KDE-Silverman$	$KDE-MLCV$
h_{opt}	۲۱/۴۶	۶
ISE	۰/۲۶	۰/۱۸
$MISE$	۰/۱۳	۰/۱۹

$h_{opt}=$ پهنای باند بهینه

ISE اختلاف کلی بین شاخص یودن برآورد شده و مقدار واقعی را اندازه‌گیری می‌کند بنابراین مقدار کمتر ISE بیانگر برازش بهتر با شاخص واقعی یودن است. $MISE$ یک اندازه

گیری متوسط از اختلاف در کل محدوده مقادیر ممکن را ارائه می‌دهد به طوری که $MISE$ کمتر نشان‌دهنده این است که عملکرد کلی مدل بهتر است. از سوی دیگر پهنای باند کمتر می‌تواند جزئیات محلی بیشتری را ثبت کند، در حالی که پهنای باند بیشتر ممکن است داده‌ها را بیش از حد هموار کند و حتی باعث محو شدن ویژگی‌های مهم شود. با توجه به این توضیحات به نظر می‌رسد که روش هسته با $MLCV$ در مقایسه با روش هسته با پهنای باند سیلورمن عملکرد بهتری در برآورد شاخص یودن دارد چرا که دارای پهنای باند کمتر، ISE کمتر و $MISE$ کمی بالاتر است به علاوه بر اساس جدول ۲ ، مقدار شاخص یودن که معیار مهمی برای ارزیابی عملکرد طبقه‌بندی است در روش هسته با $MLCV$ بیشتر است.

بحث و نتیجه‌گیری

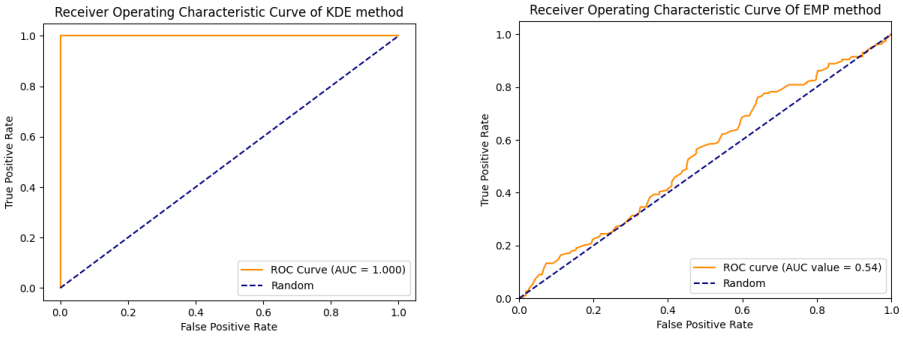
تعیین دقیق نمره برش بهینه در آزمون‌های ملاک‌مرجع برای به حداکثر رساندن اثربخشی آنها ضروری است. شاخص یودن یکی از شاخص‌هایی است که هم کارایی یک نشانگر تشخیصی را می‌سنجد و هم مقدار نمره برش را برای نشانگر انتخاب می‌کند. از آنجایی که یک روش واحدی برای تعیین دقیق نمره برش وجود ندارد لذا برآورد دقیق شاخص یودن چالش‌برانگیز است. لذا در این مطالعه، مقدار شاخص یودن با استفاده از سه روش ناپارامتری تجربی، برآورد چگالی هسته با پهنای باند سیلورمن و برآورد چگالی هسته با پهنای باند اعتبارسنجی متقابل ماکسیمم درستنمایی برآورد شد. همچنین از شاخص‌های خطای استاندارد بوت‌استرپ (BSE)، ریشه میانگین مربعات خطا ($RMSE$)، مربع خطای یکپارچه (ISE) و میانگین مربعات خطای یکپارچه ($MISE$) برای ارزیابی عملکرد استفاده شد.

در روش هسته بایستی سه مورد تابع هسته، پهنای باند و حجم نمونه مشخص شود. با توجه به تحقیقات انجام شده توسط Silverman (1976) مشخص شد که تابع هسته نقش چندانی در برآورد چگالی ندارد و مقدار دقیق پهنای باند در نتیجه بسیار مهم است. به طوری که مقدار کم و مقدار زیاد h می‌تواند منجر به اطلاعات نادرستی شود. بنابراین برآورد مقدار بهینه h برای ایجاد معنی‌دارترین و دقیق‌ترین چگالی بسیار مهم است. روش‌های مختلفی برای تعیین پهنای باند وجود دارد از جمله می‌توان به روش جایگذاری سیلورمن، اعتبارسنجی متقابل، اعتبارسنجی متقابل ماکسیمم درستنمایی ($MLCV$) و اعتبارسنجی متقابل کمترین توان‌های دوم[1] ($LSCV$) اشاره کرد (Silverman, 2018).

1. Least squares cross validation

طبق مبانی نظری و پیشینه پژوهش‌های مختلف از جمله Fluss و همکاران (2005)، Ewald (2006)، Leeflang و همکاران (2008)، Hirschfeld and do Brasil (2014) نشان دادند که هنگامی که از روش بهینه‌سازی تجربی برای برآورد شاخص یودن استفاده می‌شود به نتایج بسیار متغیری منجر می‌شود زیرا این روش به تصادفی بودن در نمونه حساس است. به علاوه در صورتی که داده‌ها نرمال و یا نزدیک به نرمال باشند روش چگالی هسته با پهنای باند سیلورمن منجر به برآوردهای پایدارتری از شاخص یودن در مقایسه با روش برآورد ناپارامتری تجربی می‌شود. برتری روش هسته نسبت به روش برآورد تجربی را می‌توان با رسم منحنی ROC و محاسبه سطح زیر نمودار (AUC) نیز نشان داد که در شکل ۳ آورده شده است.

شکل ۳. منحنی‌های ویژگی عملکرد و مقدار مساحت زیر منحنی نمره برش‌های بهینه به دست آمده از روش‌های هسته و تجربی



به طور کلی روش تعیین پهنای باند نقش مهمی در پایایی روش هسته دارد که بهترین کار، استفاده از چند روش و ارزیابی آنهاست. پایایی روش‌ها با استفاده از خطای استاندارد بوت‌استرپ ارزیابی شد.

نتایج نشان داد که روش هسته با پهنای باند سیلورمن منجر مقدار شاخص یودن ۰/۶۳ با نمره برش ۴۷۹ را به دست آورد. ریشه میانگین مربعات خطای این روش ۰/۷۱۵ با خطای استاندارد بوت‌استرپ آن ۰ به دست آمد. بنابراین برآورد شاخص یودن با استفاده از این روش نسبتا پایدار و دقیق بود. از سوی دیگر، روش هسته با $MLCV$ دارای مقدار شاخص یودن بالاتر یعنی ۰/۷۵ با همان نمره برش ۴۷۹ را تولید کرد. یعنی روش هسته با $MLCV$ برآورد دقیق‌تری از شاخص یودن را در مقایسه با روش هسته با پهنای باند سیلورمن می‌دهد

به معنای این است که قدرت تمایز بین افراد لایق و نالایق در این روش بیشتر است. ولی خطای استاندارد بوت استرپ آن برابر ۰/۰۳۹ است که نشان‌دهنده درجه کمی از تغییرپذیری در برآورد است. روش تجربی منجر به مقدار شاخص یودن ۰/۳۱ با نمره برش ۴۶۵ شد. ریشه میانگین مربعات خطای این روش ۰/۶۸ بود و خطای استاندارد بوت‌استرپ آن برابر ۰/۰۱۷ شد.

این یافته‌ها نشان می‌دهند که انتخاب روش تعیین پهنای باند مربوط به روش هسته می تواند بر برآورد شاخص یودن تأثیر بسزایی داشته باشد. با توجه به نتایج حاصل شده در این مقاله و پیشینه روش‌ها می‌توان اینگونه نتیجه گرفت که در صورتی که داده‌ها نرمال باشند روش هسته با پهنای باند سیلورمن می‌تواند منجر به نتایج پایداری شود و در غیر این صورت روش هسته با پهنای باند اعتبارسنجی متقابل ماکسیمم درستنمایی برای محاسبه و ارزیابی شاخص یودن در آزمون‌های ملاک‌مرجع و تعیین نمره برش ارجحیت دارد.

با این حال تحقیقات بیشتری برای تایید این نتایج و بررسی عملکرد این روش‌ها در زمینه های مختلف مورد نیاز است. این تحقیقات می‌توانند به بهبود فرآیند طبقه‌بندی افراد و کاهش خطاهای منفی و مثبت کاذب در نتایج آزمون‌ها کمک کنند. همچنین مفید است اثرات مختلف تعیین نمره برش بر عواقب فردی و سازمانی نیز مورد توجه قرار گیرد تا تصمیم‌گیری های بهتری در تعیین نمره برش گرفته شود. این اقدامات می‌توانند به بهبود سیستم‌های ارزیابی و تصمیم‌گیری در محیط‌های آموزشی و سازمانی کمک کنند.

تعارض منافع

هیچ تعارض منافعی وجود ندارد.

سپاسگزاری

مقاله حاضر برگرفته از رساله دکتری رشته سنجش و اندازه‌گیری دانشگاه تهران با عنوان« مقایسه نمرات برش آزمون‌های ملاک مرجع در الگوریتم‌های یادگیری عمیق و روش‌های منتخب مورد مطالعه : آزمون تولیمو » با حمایت دانشگاه تهران و سازمان سنجش آموزش کشور است. بدینوسیله از سازمان سنجش آموزش کشور به خاطر همکاری در اجرای پژوهش حاضر و از فصلنامه اندازه‌گیری تربیتی برای داوری مقاله سپاسگزاری می‌شود.

References

ADAMOWSKI, K. (1987). *Nonparametric techniques for analysis of hydrological events.* Paper presented at the Water for the Future: Hydrology in Perspective (Proceedings of the Rome Symposium).

Aoki, K., Misumi, J., Kimura, T., Zhao, W., & Xie, T. (1997). *Evaluation of cutoff levels for screening of gastric cancer using serum pepsinogens and distributions of levels of serum pepsinogen I, II and of PG I/PG II ratios in a gastric cancer case-control study*. *Journal of Epidemiology, 7*(3), 143-151.

Barbeito, I., & Cao, R. (2020). *Nonparametric curve estimation and bootstrap bandwidth selection*. *Wiley Interdisciplinary Reviews: Computational Statistics, 12*(3), e1488.

Carvalho, V. I. d., & Branscum, A. J. (2018). *Bayesian nonparametric inference for the three-class Youden index and its associated optimal cutoff points. Statistical methods in medical research, 27*(3), 689-700.

Dardick, W. R., & Weiss, B. A. (2019). *An Investigation of Chi-Square and Entropy Based Methods of Item-Fit Using Item Level Contamination in Item Response Theory. Journal of Modern Applied Statistical Methods, 18*.

Duin. (1976). *On the choice of smoothing parameters for Parzen estimators of probability density functions*. *IEEE Transactions on computers, 100*(11), 1175-1179.

Eckes, T. (2017). *Setting cut scores on an EFL placement test using the prototype group method: A receiver operating characteristic (ROC) analysis.* Language Testing, *34*(3), 383-411.

Ewald, B. (2006). *Post hoc choice of cut points introduced bias to diagnostic research.* Journal of clinical epidemiology, *59*(8), 798-801.

Fluss, R., Faraggi, D., & Reiser, B. (2005). *Estimation of the Youden index and its associated cutoff point*. Biometrical Journal, *47*(4), 458-472. doi:10.1002/bimj.200410135

Greiner, M., Pfeiffer, D., & Smith, R. D. (2000). *Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests*. Preventive veterinary medicine, *45*(1-2), 23-41.

Grmec, Š., & Gašparovic, V. (2000). *Comparison of APACHE II, MEES and Glasgow Coma Scale in patients with nontraumatic coma for prediction of mortality.* Critical Care, *5*(1), 1-5.

Habbema, J., Hermans, J., & Van den Broek, K. (1974). *A stepwise discriminant analysis program using density estimation.*

Hall, P. (1982). *Cross-validation in density estimation.* Biometrika, *69*(2), 383-390.

Hanley, J. A., & McNeil, B. J. (1982). *The meaning and use of the area under a receiver operating characteristic (ROC) curve*. Radiology, *143*(1), 29-36.

Heidenreich, N.-B., Schindler, A., & Sperlich, S. (2013). *Bandwidth selection for kernel density estimation: a review of fully automatic selectors.* AStA Advances in Statistical Analysis, *97*, 403-433.

Hirschfeld, G., & do Brasil, P. E. A. A. (2014). *A simulation study into the performance of "optimal" diagnostic thresholds in the population:"Large" effect sizes are not enough.* Journal of clinical epidemiology, *67*(4), 449-453.

Hsiao, J. K., Bartko, J. J., & Potter, W. Z. (1989). *Diagnosing diagnoses: Receiver operating characteristic methods and psychiatry.* Archives of General Psychiatry, *46*(7), 664-667.

Hsieh, F., & Turnbull, B. (1992). *Nonparametric methods for evaluating diagnostic tests.* Retrieved from

Jones, M. C., Marron, J. S., & Sheather, S. J. (1996). *A brief survey of bandwidth selection for density estimation.* Journal of the American statistical association, *91*(433), 401-407.

Kile, H. (2010). *Bandwidth Selection in Kernel Density Estimation.* (Master). University of Science and Technology, Norwegian.

Leeflang, M. M., Moons, K. G., Reitsma, J. B., & Zwinderman, A. H. (2008). *Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions.* Clinical chemistry, *54*(4), 729-737.

Loader, C. R. (1999). *Bandwidth selection: classical or plug-in?* The Annals of Statistics, *27*(2), 415-438.

Luo, J., & Xiong, C. (2013). *Youden index and associated cut-points for three ordinal diagnostic groups*. Communications in Statistics-Simulation and Computation, *42*(6), 1213-1234.

Metz, C. E. (1989). *Some practical issues of experimental design and data analysis in radiological ROC studies.* Investigative radiology, *24*(3), 234-245.

Nakas, C. T., Alonzo, T. A., & Yiannoutsos, C. T. (2010). *Accuracy and cut-off point selection in three-class classification problems using a generalization of the Youden index.* Statistics in Medicine, *29*(28), 2946-2955.

Park, B. U., & Marron, J. S. (1990). *Comparison of data-driven bandwidth selectors.* Journal of the American statistical association, *85*(409), 66-72.

Parzen, E. (1962). *On estimation of a probability density function and mode.* The annals of mathematical statistics, *33*(3), 1065-1076.

Ruopp, M. D., Perkins, N. J., Whitcomb, B. W., & Schisterman, E. F. (2008). *Youden Index and optimal cut-point estimated from observations affected by a lower limit of detection.* Biometrical Journal: Journal of Mathematical Methods in Biosciences, *50*(3), 419-430.

Schisterman, E. F., Perkins, N. J., Liu, A., & Bondell, H. (2005). *Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples.* Epidemiology, 73-81.

Shapiro, D. E. (1999). *The interpretation of diagnostic tests.* Statistical methods in medical research, *8*(2), 113-134.

Silverman, B. W. (2018). Density estimation for statistics and data analysis: Routledge.

Somoza, E., Mossman, D., & McFeeters, L. (1990). *The INFO-ROC technique: a method for comparing and optimizing inspection systems.* Review of progress in quantitative nondestructive evaluation, 601-608.

Thiele, C., & Hirschfeld, G. (2020). *cutpointr: Improved estimation and validation of optimal cutpoints in R*. arXiv preprint arXiv*:2002.09209*.

Trosset, M. W. (2009). *An introduction to statistical inference and its applications with R:* CRC Press.

Van Es, B. (1991). *Likelihood cross-validation bandwidth selection for nonparametric kernel density estimators †*. Journal of Nonparametric Statistics, *1*(1-2), 83-110. doi:10.1080/10485259108832513

Wand, M. P., & Jones, M. C. (1994). *Kernel smoothing:* CRC press.

Węglarczyk, S. (2018). *Kernel density estimation and its application.* Paper presented at the ITM web of conferences.

Youden, W. J. (1950). *Index for rating diagnostic tests.* Cancer, 3(1), 32-35.

Zhou, X.-H., McClish, D. K., & Obuchowski, N. A. (2009). *Statistical methods in diagnostic medicine*: John Wiley & Sons.

Zou, K. H., Tempany, C. M., Fielding, J. R., & Silverman, S. G. (1998). *Original smooth receiver operating characteristic curve estimation from continuous*

data: statistical methods for analyzing the predictive value of spiral CT of ureteral stones. Academic radiology, 5(10), 680-687.

Zucchini, W., Berzel, A., & Nenadic, O. (2003). *Applied smoothing techniques. Part I: Kernel Density Estimation, 15*, 1-20.