مقدر لا معلمي (المدرج التكراري) لتقدير دالة كثافة احتمالية

د. مناف يوسف

جامعة بغداد– كلية الادارة والاقتصاد

قسم الاحصاء

المستخلص

في هذا البحث تم تقديم عدد من المقدرات الخاصة بتقدير معلمة عرض الصندوق لواحد من اكثر مقدرات دالة الكثافة الاحتمالية شيوعا وهو مايسمى بالمدرج التكراري، وقد تم استخدام اسلوب المحاكاة لمقارنة تلك المقدرات اذ اثبتت النتائج افضلية اسلوب قاعدة الابهام لاكثر التجارب المقامة.

Abstract

In this paper we introduce several estimators for Binwidth of histogram estimators' .We use simulation technique to compare these estimators .In most cases, the results proved that the rule of thumb estimator is better than other estimators.

١. المقدمة

يعد اختيار دالة كثافة احتمالية ملائمة أو دالة توزيع تجميعية واحدا من المسائل الضرورية لتمثيل المجتمع بشكل صحيح، لكن في الواقع العملي غالبا ما تظهر مشكلة تتمثل بعدم معرفة تلك الدوال مما يتطلب اللجوء الى وسائل و طرائق تعمل على التقريب للدالة الحقيقة.

ان احدى الطرائق التي تستخدم لغرض عمل تخمين جيد يتمثل من خلال مشاهدة بضعة قيم للمتغير العشوائي ومن ثم رسم شكل الدالة الملائم لتلك البيانات ويعد هذا الأسلوب احد الأساليب المتبعة لتقدير دالة الكثافة الاحتمالية أو دالة التوزيع التجميعية.

ويُعد المدرج التكراري احد اقدم المقدرات لدالة الكثافة الاحتمالية واوسعها استخداما [5] ويعود استخدامه الى عام (١٦٦٢)،اذ تم استخدامه في دراسة حالات الوفيات **Mortality** والتي قام بها **John Grant** [3]. اما تسميته فتعود الى عام ١٨٩٥ والتي اطلق عليه هذه التسمية هو العالم **Karl Pearson** [7] .

ان هدف هذا البحث يتمثل بمقارنة عدد من المقدرات المستخدمة في تقدير عرض الصندوق للمدرج التكراري في حالة كون المقدر المستخدم هو مدرج تكراري ذو عرض صندوق ثابت وبيان افضل المقدرات التي تعطي الوصف الاكثر ملائمة للبيانات المعطاة.

يتم صياغة هذا المقدر عادة من خلال تقسيم الفترة الكلية الى مجموعة مكونة من **K** من الفترات الجزئية متساوية الابعاد (المسافات) وتدعى بالصناديق **Bins** مع عرض صندوق ثابت **h** [2] .

ومع استبدال دالة التوزيع التجميعية $F(x)$ بدالة التوزيع التجريبية [1][3][6] :

$$F_n(x) = \frac{\#\{X_i \le x\}}{n} \qquad \dots (1)$$

فان هذا يقود الى تقدير المدرج التكراري لدالة الكثافة الاحتمالية مع الاشارة الى وجود صيغتان للمدرج التكراري تتمثل الاولى باستخدام عرض صندوق ثابت (متساوي لجميع الصناديق) كما في اعلاه ومن ثم يصبح المقدر [3] [8] :

$$\hat{f}(x) = \frac{\#\{X_i \,|\, b_j < x_j \le b_{j+1}\}}{nh} = \frac{\#\{X_i \le b_{j+1}\} - \#\{X_i < b_j\}}{nh}, \quad x \in (b_j, b_{j+1}] \qquad \dots (2)$$

اذ ان $(b_j, b_{j+1}]$ يعرف الحدود للصندوق j وان

$$\hat{f}(x) = \frac{n_j}{nh}, \quad x \in (b_j, b_{j+1}] \equiv x \in (x_0 + jh, x_0 + (j+1)h], \; j = 0,1,\dots,k-1 \qquad \dots (3)$$

اذ يمثل n_j عدد المشاهدات في الصندوق j وان $h = b_{j+1} - b_j$. أما الصيغة الثانية لهذا المقدر في حالة كون عرض الصندوق متغير (اي ان قيمة h متغيرة من صندوق لاخر) فتكون (يسمى هذا المقدر بالمدرج التكراري ذو العرض المتغير):

$$\hat{f}(x) = \frac{n_j}{n(b_{j+1} - b_j)}, \quad x \in (b_j, b_{j+1}] \qquad \dots (4)$$

٢. خصائص المدرج التكراري

لتحقيق خصائص المدرج التكراري يجب تحقق الشروط الاتية:

- $Lim_{n\to\infty} nh = \infty$, $Lim_{n\to\infty} h = 0$.
- دالة الكثافة تكون تمهيدية وتحقق شرط Lipschitz .
- $f'(x)$ تكون مستمرة بشكل مطلق وقابلة للتكامل بشكل تربيعي.

وبتحقق الشروط المذكورة انفا فان التحيز للمقدر سوف يكون:

$$Bias(\hat{f}(x)) = \frac{f'(x)}{2}\{h - 2(x - b_j)\} + o(h^2), \quad x \in (b_j, b_{j+1}] \qquad \dots (5)$$

اما التباين فيكون:

$$Var(\hat{f}(x)) = \frac{f(x)}{nh} + o(n^{-1}) \qquad \dots (6)$$

وبدمج التباين مع مربع التحيز نحصل على:

$$MSE(\hat{f}(x)) = \frac{f(x)}{nh} + \frac{[f'(x)]^2}{4}\left[h - 2(x - b_j)\right]^2 + o(n^{-1}) + o(h^3) \dots (7)$$

في حين ان MISE يصبح:

$$MISE(\hat{f}(x)) = \frac{1}{nh} + \frac{h^2}{12}\int_{-\infty}^{\infty}[f'(x)]^2 dx + o(n^{-1}) + o(h^3) = AMISE(\hat{f}(x)) + o(n^{-1}) + o(h^3) \qquad \dots (8)$$

وعند تقليل AMISE بالنسبة لـ h نحصل على قيمة h المثلى:

$$h_{opt} = \left[\frac{6}{\int_{-\infty}^{\infty}[f'(x)]^2 dx}\right]^{1/3} n^{-1/3} \quad \text{... (9)}$$

$$= \left[\frac{6}{R(f')}\right]^{1/3} n^{-1/3}$$

وعند تعويض قيمة h_{opt} في AMISE نحصل على:

$$AMISE_{opt} = \left[\frac{9}{16} R(f')\right]^{1/3} n^{-2/3} \quad \text{... (10)}$$

من الجدير بالملاحظة ان للمدرج التكراري اختيارين مهمين له وهما عرض الصندوق وموقع حافات الصندوق (Bin-edge) او نقاط البداية للمدرج التكراري، اذ ان كلا هذين الاختيارين لهما تاثيرا معنويا على المدرج التكراري وغالبا ما يتم اختيار نقطة البداية عند قيمة الصفر.

٣. اختيار عرض الصندوق Bin width

يسمى عرض الصندوق للمدرج التكراري بالمعلمة التمهيدية كونه يسيطر على كمية التمهيد المراد تطبيقها على البيانات.هنالك عدد من الطرائق لاختيار عرض الصندوق سوف يتم التطرق الى بعض منها، وكالاتي:

٣.١ اساليب الاختيار الشخصي

تسمى هذه الاساليب ايضا بطرائق ad-hoc والتي نعني بها استخدام الباحث لخبرته الشخصية في اختيار معلمة عرض الصندوق ،اي انها عملية اختيار من قبل الباحث وليست طريقة مبنية على اساس رياضي بحت، وهنالك عدد من طرائق الاختيار الشخصي للصندوق او عرض الصندوق وكالاتي:

- اختيار او تجزئة مدى العينة من ٥ الى ٢٠ صندوق.[4]
- استخدام الصيغة $1+2.2Log_{10}^{n}$ المقترحة من قبل الباحث Larson عام ١٩٧٥.[3]
- استخدام على الاقل $(2n)^{1/3}$ من الصناديق.[8]
- استخدام صيغة Sturges (١٩٢٦) المساوية تقريبا الى $1+Log_2^n$.[8]
- استخدام صيغة Cencov (١٩٦٢) الذي لاحظ ان عدد الصناديق تكون متناسبة الى الجذر التكعيبي للعينة $h\alpha C\sqrt[3]{n}$.[8]
- قاعدة الإبهام Rule of thumb والمتمثلة بالصيغة: $\frac{Range(x)}{2(1+Log_2^n)}$.

٣.٢ قاعدة المصدر الطبيعي:[3][5][6]

تسمى هذه القاعدة ايضا بقاعدة القياس (Scale Rule) ولاختيار عرض الصندوق بالاعتماد على هذه القاعدة تعتمد على الصيغة في المعادلة (٩) المذكورة انفا:

$$h_{opt} = \left[\frac{6}{R(f')}\right]^{1/3} n^{-1/3}$$

إذ تتضمن هذه القاعدة دالة الكثافة f التي غالبا ماتكون مجهولة،لذلك فان الطريقة الاكثر بساطة لاختيار عرض الصندوق تتمثل باختيار دالة كثافة معينة مثل دالة Gaussian ومن ثم تعويضها في المعادلة (٩) فنحصل على قيمة h ومن ثم فان:[3][6]

$$h_{opt} = 2\times 3^{1/3}\pi^{1/6}\sigma n^{-1/3} = 3.49\,\sigma n^{-1/3} \quad \dots (11)$$

الصيغة المذكورة انفا تتطلب تقدير σ ،وهنالك عدة اختيارات منها استخدام الانحراف المعياري.[3]

$$h_{opt} = 3.49\,S\,n^{-1/3} \quad \dots (1\,2)$$

في حين اقترح الباحثان Freedman and Diaconis عام ١٩٨١ استخدام مقياس Interquartile range اذ استخدما القاعدة الاتية:[6]

$$h = 2\,IQR\,n^{-1/3} \quad \dots (13)$$

اما الباحث Silverman [5] قام بدمج هاتين الفكرتين من خلال القاعدة الاتية:

$$h = 3.49\,\hat{\sigma}\,n^{-1/3} \quad \dots (14)$$

اذ ان :

$$\hat{\sigma} = Min\left(S, \frac{IQR}{1.349}\right) \quad \dots (15)$$

اما القاعدة المقترحة لهذه الصيغة فتتضمن استخدام الصيغة المعتمدة على (MedAD) Median Absolute Deviation:

$$h = 3.49\,\hat{\sigma}\,n^{-1/3}$$

إذ إن :

$$\hat{\sigma} = Min\left(S, \frac{IQR}{1.349}, \frac{MedAD}{0.6745}\right) \quad \dots (16)$$

اذ يمثل:

$$MedAD = Median(|x_i - Median(x_i)|) \quad \dots (17)$$

٣.٣ قاعدةاختيار عرض الصندوق فوق التمهيدي

تعتمد هذه القاعدة على تقدير الحد الادنى $R(f')$ في الصيغة (٩) بالاعتماد على البيانات،اذ استخدم الباحثان Terrel and Scott عام (١٩٨٥) [8] المدى للبيانات كتقدير لمعلمة القياس ،اذ ان العدد الامثل للصناديق يجب ان لايقل عن $(2n)^{1/3}$ وهذا يتطابق مع الصيغة ادناه:

$$h \le 3.55\,\sigma n^{-1/3} \quad \dots (18)$$

اذ يشير $\hat{\sigma}$ إلى الانحراف المعياري ويستخدم بدلا عنه الانحراف المعياري للعينة S.

اما الباحثان Freedman and Diaconis فاستخدما المقياس [8]:

$$h \le 2.6\,IQR\,n^{-1/3} \quad \dots (19)$$

اما المقدر المقترح فيكون:

$$h \le 5.26\,MedAD\,n^{-1/3} \quad \dots (20)$$

في حين القاعدة المقترحة لهذه الصيغة فتتضمن استخدام الصيغة الاتية:

$$h \le 3.55\,\hat{\sigma}\,n^{-1/3}$$

اذ ان :

$$\hat{\sigma} = Min\left(S, \frac{IQR}{1.349}, \frac{MedAD}{0.6745}\right) \quad \text{... (21)}$$

٤- الجانب التجريبي

تم في هذا المبحث تم استخدام الاسلوب التجريبي (المحاكاة) في مقارنة مقدرات عرض الصندوق لمقدر دالة الكثافة الاحتمالية والمسمى بالمدرج التكراري لغرض بيان افضل الاساليب او الطرائق المتبعة لتمثيل البيانات تمثيلا سليما.

وقد استخدم لغرض المقارنة ثلاث توزيعات هي:

- **التوزيع الطبيعي:**

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{\sigma^2}}, I^{(x)}_{(-\infty,\infty)}, -\infty < \mu < \infty, \sigma^2 > 0$$

لكن بمتوسط صفر وتباينات ٥ ، ١٠ ، ١٥ لمعرفة اداء تلك المقدرات في حالة كون البيانات متجانسة وغير متجانسة.

- **توزيع Student's t ذو درجة n :**

$$f(x) = \frac{\Gamma(\frac{v+1}{2})}{\Gamma(\frac{v}{2})} \cdot \frac{1}{\sqrt{v\pi}} \cdot \frac{1}{(1+\frac{x^2}{v})^{\frac{v+1}{2}}}, I^{(x)}_{(-\infty,\infty)}, where \quad v = n-1$$

مع الإشارة إلى استخدام التحويل الاتي ،بافتراض ان x يتوزع طبيعيا بمتوسط μ وتباين σ^2 فان:

$$t = \frac{x-\mu}{s} \sim t_{(n-1)}$$

إذ يشير (n-1) الى درجة الحرية، مع الاشارة الى ان v=n-1.

مع كون قيمة المتوسط المفترضة هي صفر وان s يشير الى الانحراف المعياري للعينة.

- **توزيع مربع كاي χ^2 ذو درجة حرية (n) :**

يقال للدالة بانها دالة كثافة احتمالية تتبع توزيع مربع كاي اذا حققت الدالة الاتية:

$$f(x) = \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}$$

وقد تم استخدام التحويل الاتي للحصول على متغير مربع كاي :
بافتراض ان x يتوزع طبيعيا فان:

$$\frac{(n-1)s^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{\sigma^2} \sim \chi^2_{(n-1)}$$

وان قيم التباين المستخدمة هي ٥ ، ١٠ ، ١٥، كذلك تم استخدام حجوم عينات مختلفة هي (١٠٠,٥٠,٢٠)، والرموز الاتية تشير الى مقدرات عرض الصندوق **Binwidth** المستخدمة في مقدر المدرج التكراري :

h_1 : وتشير الى المقدر الاول المذكور في المعادلة (٩).

h_2 : وتشير الى المقدر الذي يشير الى تقسيم الفترة من ٥-٢٠ صندوق.

h_3 : ويتمثل باستخدام صيغة **Larson** $1+2.2Log_{10}^{n}$.

h_4 : ويتمثل باستخدام على الأقل $(2n)^{1/3}$ من الصناديق.

h_5 : ويتمثل باستخدام صيغة **Sturges** $1+Log_2^n$.

h_6 : وتمثل قاعدة الابهام $\frac{Range(x)}{2(1+Log_2^n)}$.

h_7 : قاعدة المصدر الطبيعي والمساوية الى $h_{opt} = 3.49\, S\, n^{-1/3}$.

h_8 : ومساوية الى $h = 2\, IQR\, n^{-1/3}$.

h_9 : ومساوية الى $h = 3.49\, \hat{\sigma}\, n^{-1/3}$ اذ ان h_{10} $\hat{\sigma} = Min\left(S, \frac{IQR}{1.349}, \frac{MedAD}{0.6745}\right)$

ومساوية الى $h = 3.55\, \hat{\sigma}\, n^{-1/3}$ إذ أن $\hat{\sigma} = Min\left(S, \frac{IQR}{1.349}, \frac{MedAD}{0.6745}\right)$:

اما معيار المقارنة المستخدم فهو معيار **MISE** ، والاشكال الاتية (١-٣)،(٤-٦)،(٧-٩) تشير الى اشكال المناظرة الى قيم **MISE** لكل مقدر من مقدرات عرض الصندوق المستخدمة لكل توزيع من التوزيعات المذكورة انفا ولحجوم العينات وقيم التباينات المستخدمة. وقد تم تكرار تنفيذ التجارب ٥٠٠ مرة لكل حالة من الحالات المعطاة.

٤.١- تفسير النتائج

من الاشكال (١-٣) لحالة التوزيع الطبيعي يلاحظ:

- لجميع حجوم العينات والتباينات اوضحت النتائج ان اقل قيمة لمعيار MISE كانت عند استخدام مقدر قاعدة الابهام h_6 يليه المقدرين h_1 و h_9 بشكل متساوي، عدا في حالة حجم عينة ٢٠ وتباين ١٥ اشارت النتائج الى افضلية المقدر h_4 يليه المقدر h_6 .
- لجميع حجوم العينات والتباينات اوضحت النتائج تماثل قيم MISE للمقدر h_1 مع قيم MISE للمقدر h_9 .
- لحجوم عينة ٥٠ ، ١٠٠ اشارت النتائج الى ان قيمة $MISE(h_2)$ كانت اعظم مايمكن.
- لحجم عينة ١٠٠ تماثلت قيم MISE المقابلة للمقدرات h_7 و h_{10} .
- لحجم عينة ٥٠ تماثلت قيم MISE المقابلة للمقدرات h_3 و h_2 .
- لحجم عينة ٢٠ وتباين ٥ اشارت النتائج الى ان قيمة $MISE(h3)$ كانت اعظم ما يمكن، وعند تباينات ١٠،١٥ اظهرت النتائج ان قيمة $MISE(h_7)$ كانت اعظم مايمكن.
- تناقص قيم MISE مع تزايد قيم التباينات .
- تزايد قيم MISE مع تزايد حجوم العينات.

للاشكال (٤-٦) لحالة توزيع Student's t يلاحظ:

- افضلية الصيغة h_6 على بقية المقدرات لعرض الصندوق يليه المقدر h_8.
- في حين اظهرت النتائج تدني اداء مقدري h_3 h_2 على التوالي.
- تماثل قيم MISE المقابلة للمقدرات h_1 و h_9.
- تزايد قيم MISE مع تزايد قيم التباينات لجميع مقدرات عرض الحزمة عدا لبعض المقدرات التي كان لها سلوكا مغايرا.
- تزايد قيم MISE مع تزايد حجوم العينات.

للاشكال (٧-٩) لحالة توزيع Chi-2 يلاحظ:

- افضلية الصيغة h_{10} على بقية المقدرات لعرض الصندوق في حال حجوم العينات الكبيرة n=100 يليه مقدر h_6 ،اما في حجوم العينة الصغيرة والمتوسطة اثبتت النتائج افضلية مقدر h_6 على بقية المقدرات يليه المقدر h_{10} .
- اظهرت النتائج تدني اداء مقدري h_3 h_2 على التوالي.
- تماثل قيم MISE المقابلة للمقدرات h_1 و h_9 لجميع حجوم العينات والتباينات.
- تماثل قيم MISE المقابلة للمقدرات h_2 مع h_3 عند حجم عينة ٥٠.
- تزايد قيم MISE مع تزايد قيم التباينات عدا لبعض الحالات عند استخدام المقدرات h_6 ، h_7 ، h_8 و h_{10}.
- تزايد قيم MISE مع تزايد حجوم العينات.

إما الإشكال من ١٠- ١٥ فتمثل إشكال لبعض التجارب المنفذة لبيان تأثير استخدام كل من المقدرات h على الشكل النهائي للمدرج التكراري ومن ثم التمثيل السليم لبيانات المنحى الأصلي.

إذ يلاحظ من الشكل (١٠):

تشابه الإشكال الناتجة من استخدام h_1 ، h_4 و h_{10} ، كذلك الاشكال الناتجة من استخدام h_2 ، h_5 و h_7 على الرغم من اختلاف قيمهم لكن هذا الاختلاف كان ضئيلا بحيث لم يتاثر الشكل النهائي تاثرا كثيرا.

في حين من الشكل (١١) يلاحظ:

تشابه الاشكال الناتجة من استخدام h_2 و h_3 ، وكذلك الأشكال الناتجة من استخدام h_4 ، h_7 ، h_9 و h_{10} على الرغم من اختلاف قيم تلك المقدرات .

من الشكل (١٢) يلاحظ:

تشابه اشكال المدرج التكراري الناتجة من استخدام h_7 h_9 و h_{10} ، وكذلك الأشكال الناتجة من استخدام مقدري h_1 مع h_4 .

من الشكل (١٣) يلاحظ:

اشارت النتائج الى تشابه اداء المقدرات كما في الشكل (١٢)،اذ تشابهت الاشكال المرافقة للمقدر الناتج من استخدام المقدرات h_7 h_9 و h_{10} ،وكذلك عند استخدام المقدرات h_2 ، h_3 ، h_4 و h_5 على الرغم من اختلاف قيمهم.

الشكل (١٤) يشير:

الى تشابه الاشكال المرافقة للقيم h_9 و h_{10} ،وكذلك الاشكال المرافقة لـ h_2 ، h_3 ، h_4 و h_5 على الرغم من اختلاف قيم تلك المقدرات.

الشكل (١٥) يشير:

الى تشابه اشكال المدرج التكراري المرافقة للقيم h_2 ، h_3 ، h_4 و h_5 ،وكذلك الأشكال المرافقة لـ h_9 و h_{10} .

٥- الاستنتاجات

أثبتت النتائج الحالات الاتية:

<u>للتوزيع الطبيعي:</u>

- **افضلية استخدام مقدر قاعدة الابهام h_6 او المقدرات h_1 او h_9 كبدائل جيدة خاصة مع استعمال حجم عينة صغير.**
- **لايفضل استخدام المقدرات h_2 و h_3.**
- **من الأشكال نرى أن بعض المقدرات لمعلمة عرض الصندوق اظهرت تماثلا في الشكل النهائي للبيانات وهذا يعود الى اختلاف قيم تلك المقدرات اختلافا ضئيلا بحيث لم يتاثر الشكل النهائي تاثرا كثيرا.**
- **تأثر مقياس MISE بقيم التباينات وحجوم العينات اذ تأثر هذا المقياس تأثرا عكسيا مع تزايد قيم التباينات ،في حين تأثرا طرديا مع تزايد حجوم العينات.**

<u>لتوزيع Student's t:</u>

- **افضلية استخدام مقدر قاعدة الابهام h_6 او المقدر h_8.**
- **لا يفضل استخدام المقدرات h_2 و h_3.**
- **تأثر مقياس MISE بقيم التباينات وحجوم العينات اذ تزايد هذا المقياس مع تزايد قيم التباينات عدا لبعض المقدرات التي اظهرت فيه النتائج تناقص هذا المقياس مع تزايد قيم التباينات ، في حين تأثرا طرديا مع تزايد حجوم العينات.**

<u>لتوزيعChi-2 :</u>

- **يفضل استخدام المقدر h_{10} مع حجوم العينات الكبيرة، في حين يفضل استعمال المقدر h_6 مع حجوم العينات الصغيرة والمتوسطة.**
- **لايفضل استخدام المقدرات h_2 و h_3 ، مما ينتج عدم افضلية استعمال هذه المقدرات في جميع الحالات لما له من تأثير سلبي على اداء المقدرات.**

المصادر

1.Hardle, W. (1991)" Smoothing techniques with implementation in S"Springer-Verlage, New York.
2.Rao, P.B.L.S. (1983)" Nonparametric functional estimation" Academic press.
3.Scott, D.W. (1979)" On optimal and data based histogram "Biometrika, Vol.66, No.3, PP 605-610.
4.Scott, D.W. and Factor, L.E. (1981)" Monte Carlo study of three based nonparametric density estimation" JASA, Vol.76, No.373, PP 9-15.
5.Silverman, B.W. (1986)" Density estimation for statistics and data analysis" Chapman and Hall, London.
6.Siminoff, J. (1996)" Smoothing methods in statistics "Springer-Verlag, New York.
7.Tarter, M.E. and Kronmal R.A. (1976) "An introduction to the implementation and theory of nonparametric density estimation" The American stat., Vol.30, No.3, PP 105-112.
8.Terrel, G.R. & Scott, D.W. (1985)" Oversmoothed nonparametric density estimates" JASA, Vol.80, No.389, PP209-214.