

مقدراً ملماً (الدرج 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 الواحد من اكثراً مقدرات دالة الكثافة الاحتمالية شيوعاً وهو ما يسمى بالدرج التكراري، وقد تم استخدام اسلوب المحاكاة لمقارنة تلك المقدرات اذا ثبتت النتائج افضلية اسلوب قاعدة الابهام لاكثر التجارب المقامة.

Nonparametric Estimator (Histogram) For Estimating Probability Density Function

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.

1. المقدمة

يعد اختيار دالة كثافة احتمالية ملائمة او دالة توزيع تجميعية واحداً من المسائل الضرورية لتمثيل المجتمع بشكل صحيح، لكن في الواقع العملي غالباً ما تظهر مشكلة تمثل بعدم معرفة تلك الدوال مما يتطلب اللجوء الى وسائل و طرائق تعمل على التقريب للدالة الحقيقة.
ان احدى الطرائق التي تستخدلم لغرض عمل تخمين جيد يتمثل من خلال مشاهدة بضعة قيم للمتغير العشوائي ومن ثم رسم شكل الدالة الملائم لتلك البيانات ويعود هذا الاسلوب احد الأساليب المتبعه لتقدير دالة الكثافة الاحتمالية او دالة التوزيع التجميعية.

ويعد الدرج التكراري احد اقدم المقدرات لدالة الكثافة الاحتمالية واوسعها استخداماً [5] ويعود استخدامه الى عام (1662)،اذ تم استخدامه في دراسة حالات الوفيات Mortality والتي قام بها John Grant [3]. اما تسميته فتعود الى عام 1895 والتي اطلق عليه هذه التسمية هو العالم Karl Pearson [7]. ان هدف هذا البحث يتمثل بمقارنة عدد من المقدرات المستخدمة في تقدير عرض الصندوق للدرج التكراري في حالة كون المقدر المستخدم هو درج 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] :

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

فإن هذا يقود إلى تقدير الدرج التكراري لدالة الكثافة الاحتمالية مع الإشارة إلى وجود صيغتان للدرج التكراري تمثل الاولى باستخدام عرض صندوق ثابت (متساوي لجميع الصناديق) كما في اعلاه ومن ثم يصبح المقدر [3][8] :



$$\hat{f}(x) = \frac{\#\{X_i \mid b_j < X_i \leq b_{j+1}\}}{nh} \quad \dots (2)$$

$$= \frac{\#\{X_i \leq b_{j+1}\} - \#\{X_i < b_j\}}{nh}, \quad x \in (b_j, b_{j+1}]$$

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

$$\hat{f}(x) = \frac{n_j}{nh}, \quad x \in (b_j, b_{j+1}] \quad \dots (3)$$

$$\equiv x \in (x_0 + jh, x_0 + (j+1)h], \quad j = 0, 1, \dots, k-1$$

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

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

2. خصائص المدرج التكراري

لتحقيق خصائص المدرج التكراري يجب تحقق الشروط الآتية:

- $\lim_{n \rightarrow \infty} h = 0, \quad \lim_{n \rightarrow \infty} nh = \infty$

• دالة الكثافة تكون تمهيدية وتحقق شرط Lipschitz.

• $f'(x)$ تكون مستمرة بشكل مطلق وقابلة للتكامل بشكل تربيعي.

وبتحقيق الشروط المذكورة انفا فان التحيز للمقدر سوف يكون:

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

اما التباين فيكون:

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

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:

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

$$= AMISE(\hat{f}(x)) + o(n^{-1}) + o(h^3)$$



و عند تقليل $AMISE$ بالنسبة لـ h نحصل على قيمة h المثلثى :

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

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

و عند تعويض قيمة h_{opt} في $AMISE$ نحصل على :

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

من الجدير باللحظة 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 تأثيرا معنوايا على المدرج التكراري و غالبا ما يتم اختيار نقطة البداية عند قيمة الصفر.

3. اختيار عرض الصندوق Bin width

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

3.1 اساليب الاختيار الشخصي

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

- اختيار او تجزئة مدى العينة من 5 الى 20 صندوق.[4]
- استخدام الصيغة $1 + 2.2 \log_{10}^n$ المقترنة من قبل الباحث Larson عام 1975 [3].
- استخدام على الاقل $(2n)^{1/3}$ من الصناديق.[8]
- استخدام صيغة Sturges (1926) المساوية تقريبا الى $1 + \log_2^n$ [8].
- استخدام صيغة Cencov (1962) الذي لاحظ ان عدد الصناديق تكون متناسبة الى الجذر التكعيبي للعينة $\sqrt[n]{n}$
- قاعدة الإبهام Rule of thumb والمتمثلة بالصيغة: $\frac{Range(x)}{2(1 + \log_2^n)}$.

3.2 قاعدة المصدر الطبيعي:[6][5][3]

تسمى هذه القاعدة ايضا بقاعدة القياس Scale Rule (Rule of thumb) ولاختيار عرض الصندوق بالاعتماد على هذه القاعدة تعتمد على الصيغة في المعادلة (9) المذكورة افرا:

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



إذ تتضمن هذه القاعدة دالة الكثافة f التي غالباً ما تكون مجهولة، لذلك فإن الطريقة الأكثر بساطة لاختيار عرض الصندوق تتمثل باختيار دالة كثافة معينة مثل دالة Gaussian ومن ثم تعويضها في المعادلة (9) فنحصل على قيمة h ومن ثم فإن: [6][3]

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

الصيغة المذكورة انفا تتطلب تقدير σ ، وهناك عدة اختيارات منها استخدام الانحراف المعياري.

$$h_{opt} = 3.49 S n^{-1/3} \quad \dots (12)$$

في حين اقترح الباحثان Freedman and Diaconis عام 1981 استخدام مقياس Interquartile اذ استخدما القاعدة الآتية: [6]

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

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

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

اذا ان :

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

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

$$h = 3.49 \hat{\sigma} n^{-1/3} \quad \text{اذا ان} \quad \dots$$

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

اذا يمثل:

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

3.3 قاعدة اختيار عرض الصندوق فوق التمهيدي

تعتمد هذه القاعدة على تقدير الحد الادنى (f') في الصيغة (9) بالاعتماد على البيانات، اذ استخدم

الباحثان Terrel and Scott عام (1985) [8] المدى للبيانات كتقدير لمعلمة القياس، اذ ان العدد الامثل

للصناديق يجب ان لا يقل عن $(2n)^{1/3}$ وهذا يتطابق مع الصيغة ادناء:

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

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

اما الباحثان Freedman and Diaconis [8]:

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

اما المقدر المقترح فيكون:

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



في حين القاعدة المقترحة لهذه الصيغة فتتضمن استخدام الصيغة الآتية:

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

اذا ان :

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

4- الجانب التجربى

تم في هذا البحث تم استخدام الاسلوب التجربى (المحاكاة) في مقارنة مقدرات عرض الصندوق لمقدار دالة الكثافة الاحتمالية والمسمى بالمدرج التكراري لفرض بيان افضل الاساليب او الطرائق المتبعة لتمثيل البيانات تمثيلا سليما.

وقد استخدم لغرض المقارنة ثلاثة توزيعات هي:

- التوزيع الطبيعي:

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

لكن بمتوسط صفر وبيانات 5 ، 10 ، 15 لمعرفة اداء تلك المقدرات في حالة كون البيانات متجانسة وغير متجانسة.

- توزيع Student's t ذو درجة n :

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

مع الإشارة إلى استخدام التحويل الآتى ، بافتراض ان x يتوزع طبيعيا بمتوسط μ وبيان 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 الآتية:

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



وقد تم استخدام التحويل الآتي للحصول على متغير مربع كاي :
بافتراض ان x يتوزع طبيعياً فان:

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

وان قيم التباين المستخدمة هي 5 ، 10 ، 15 ، كذلك تم استخدام حجوم عينات مختلفة هي (100,50,20)، والرموز الآتية تشير الى مقدرات عرض الصندوق Binwidth المستخدمة في مقدر المدرج التكراري :

h_1 : وتشير الى المقدر الاول المذكور في المعادلة (9).

h_2 : وتشير الى المقدر الذي يشير الى تقسيم الفترة من 5-20 صندوق.

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

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 . $\frac{\text{Range}(x)}{2(1 + \log_2^n)}$

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 . $\hat{\sigma} = \min\left(S, \frac{IQR}{1.349}, \frac{MedAD}{0.6745}\right)$ اذ ان $h = 3.49 \hat{\sigma} n^{-1/3}$

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

اما معيار المقارنة المستخدم فهو معيار MISE ، والاشكال الآتية (3-1)،(4-6)،(6-7)،(9-7) 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 500 مرة لكل حالة من الحالات المعطاة.



4.1- تفسير النتائج

من الاشكال (3-1) 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 20 وتبين 15 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 h_9 .
- لحجوم عينة 50 ، 100 اشارت النتائج الى ان قيمة $MISE(h_2)$ كانت اعظم ما يمكن.
- لحجم عينة 100 تمثلت قيم MISE المقابلة للمقدرات h_7 و h_{10} .
- لحجم عينة 50 تمثلت قيم MISE المقابلة للمقدرات h_3 و h_2 .
- لحجم عينة 20 وتبين 5 اشارت النتائج الى ان قيمة $MISE(h_3)$ كانت اعظم ما يمكن، وعند تباينات 10، 15 اظهرت النتائج ان قيمة $MISE(h_7)$ كانت اعظم ما يمكن.
- تناقص قيم MISE مع تزايد قيم التباينات .
- تزايد قيم MISE مع تزايد حجوم العينات.

للأشكال (6-4) 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_2 و h_3 على التوالي.
- تمثل 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.

للأشكال (9-7) 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_2 و h_3 على التوالي.

- تمثل قيم MISE المقابلة للمقدرات h_1 و h_9 لجميع حجوم العينات والتباينات.
- تمثل قيم MISE المقابلة للمقدرات h_2 مع h_3 عند حجم عينة 50.

- تزايد قيم MISE مع تزايد قيم التباينات عدا لبعض الحالات عند استخدام المقدرات h_7 ، h_6 ، h_8 ، h_9 و h_{10} .

• تزايد قيم MISE مع تزايد حجوم العينات.

اما الاشكال من 10 - 15 فتمثل إشكال بعض التجارب المنفذة لبيان تأثير استخدام كل من المقدرات على الشكل النهائي للمدرج التكراري ومن ثم التمثيل السليم لبيانات المنحى الأصلي.

اذ يلاحظ من الشكل (10):

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

**فى حين من الشكل (11) يلاحظ:**

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

من الشكل (12) يلاحظ:

تشابه اشكال المدرج التكراري الناتجة من استخدام h_7 و h_9 ، وكذلك الاشكال الناتجة من استخدام مقدري h_1 مع h_4 .

من الشكل (13) يلاحظ:

اشارت النتائج الى تشابه اداء المقدرات كما في الشكل (12)، اذ تشابهت الاشكال المرافقة للمقدار الناتج من استخدام المقدرات h_7 و h_9 ، وكذلك عند استخدام المقدرات h_2 ، h_3 ، h_4 و h_5 على الرغم من اختلاف قيمهم.

الشكل (14) يشير:

الى تشابه الاشكال المرافقة للفي 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 (15) يشير:

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



5- الاستنتاجات

أثبتت النتائج الحالات الآتية:
للتوزيع الطبيعي:

- افضلية استخدام مقدر قاعدة الابهام h_6 او المقدرات h_1 او h_9 كبدائل جيدة خاصة مع استعمال حجم عينة صغير.
- لا يفضل استخدام المقدرات h_2 و h_3 .
- من الأشكال نرى أن بعض المقدرات لمعلمته عرض الصندوق اظهرت تماثلاً في الشكل النهائي للبيانات وهذا يعود إلى اختلاف قيم تلك المقدرات اختلافاً ضئيلاً بحيث لم يتاثر الشكل النهائي تائراً كثيراً.
- تأثر مقياس MISE بقيم التباينات وحجوم العينات إذ تأثر هذا المقياس تائراً عكسياً مع تزايد قيمة التباينات، في حين تأثر طردياً مع تزايد حجوم العينات.

للتوزيع t :Student's t

- 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 إذ تزداد هذا المقياس مع تزايد قيمة التباينات عدا لبعض المقدرات التي اظهرت في النتائج تناقض هذا المقياس مع تزايد قيمة التباينات ، في حين تأثر طردياً مع تزايد حجوم العينات.

للتوزيع Chi-2 :

- يفضل استخدام المقدر h_{10} مع حجوم العينات الكبيرة، في حين يفضل استعمال المقدر h_6 مع حجوم العينات الصغيرة والمتوسطة.
- لا يفضل 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.