Some NONPARAMETRIC ESTIMATORS FOR RIGHT CENSORED SURVIVAL DATA

بعض المقدرات اللامعلمية لبيانات البقاء المراقبة من النوع الاول

أ.م .د. انتصار عريبي فدعم الدوري/ كلية الادارة والاقتصاد / جامعة بغداد
entsar_arebe_aldoori@yahoo.com
الباحث/ ماجد عبد العزيز عبد الرضا الربيعي/ كلية الادارة والاقتصاد / جامعة بغداد
majid_alrubi@yahoo.com

OPEN ACCESS
P - ISSN 2518 - 5764
E - ISSN 2227 - 703X

Received:28/11/2018
Accepted :29/1/2019

المستخلص Abstract

أن استعمال الطرائق المعلمية وما يتبعها من أساليب تقدير يتطلب وجود العديد من الشروط الأولية الواجب توافرها كي تمثل تلك النماذج المجتمع تحت الدراسة تمثيلا ملائما الأمر الذي دفع الباحثون إلى البحث عن طرائق أكثر مرونة من الطرائق المعلمية وتمثلت هذه الطرائق بالطرائق اللامعلمية ، وكذلك أهتم الكثير من الباحثين في حقل البقاء واختبارات الحياة بدراسة دالة البقاء وطرائق تقديرها ومنها الطرائق اللامعلمية.
ولغرض عمل استدلال احصائي حول معلمات التوزيع الاحصائي لأوقات الحياة التي تتعامل مع بيانات مراقبة (Censored data) ومن الجانب التجريبي من هذا البحث فقد تمت المقارنة بين طرائق التقدير اللامعلمية لدالة البقاء بوجود بيانات مراقبة من النوع الاول (Typy - I censored data) وكانت هذه الطرائق هي (طريقة مقدرKaplan-Meier وطريقة مقدر Kernel وطريقة مقدر Nelson-Aalen وطريقة مقدر Thompson وطريقة مقدر Pandey) ، و تكون هذه الطرائق أكثر مرونة في تحليل البيانات والتي لا تشترط بدورها معرفة التوزيع الذي تسلكه البيانات للحصول على المقدر، ولغرض اجراء عملية المقارنة بين افضلية المقدرات تم توظيف اسلوب المحاكاة باستعمال المقياسيين الاحصائيين متوسط مربعات الخطأ التكاملي (IMSE) و متوسط الخطأ النسبي المطلق (MAPE) ولاحجام عينات مختلفة (n = 15, 30, 50 ,100) ، وقد تم التوصل الى افضلية طريقة هي (Kaplan-Meier) على باقي الطرائق اللامعلمية ، وقد بينت النتائج أن قيم دالة البقاء تبدأ بالتناقص مع إزدياد الزمن بالنسبة للمقدر اللامعلمي .

المصطلحات الرئيسة للبحث : طريقة كابلن ـ ماير ، طريقة كرنل ، طريقة نيلسون ـالن ،طريقة ثامسون ، طريقة باندي ، متوسط مربعات الخطأ التكاملي ، متوسط الخطأ النسبي المطلق .

Journal of Economics and Administrative Sciences
2019; Vol. 25, No.113
Pages: 475- 498

* بحث مستل من رسالة ماجستير

1. المقدمة (Introduction)

تحليل البقاء هو فرع من فروع الأحصاء يتعامل مع حدوث حالة الموت للمصاب بمرض معين في التجارب الطبية ويتعامل مع حدوث الفشل في الأنظمة الميكانيكية ، فتحليل البقاء له تسميات مختلفة ففي الحقول الطبية والعلوم الصحية يطلق عليه بتحليل دوال البقاء بسبب إن الحدث الحرج هو الموت وفي الدراسات الهندسية يطلق عليه مصطلح المعولية او الموثوقية والعلوم الاجتماعية يشير إلى دراسات تاريخ الحدث . وبشكل أوضح فتحليل دوال البقاء يتضمن نمذجة الوقت ،أي بدراسة الوقت للمصاب منذ تشخيص الإصابة لحين حدوث الحدث .

فعند معرفة التوزيع فمن السهولة حساب دالة البقاء لها بالاعتماد على الطرائق المعلمية الأعتيادية المعروفة، لكن صعوبة تحديد التوزيع الأحتمالي ، دفعت بعض الباحثين الى استعمال طرائق أكثر مرونةً من الطرائق المعلمية لتحليل البيانات، وهي الطرائق اللامعلمية (Nonparametric methods) وهي من الطرائق الاحصائية الاستدلالية التي يمكن استعمالها للتوصل الى استنتاجات بشأن المجتمع قيد الدراسة في ضوء العينة بغض النظر عن نوع التوزيع النظري لذلك المجتمع، إذ لاتتطلب هذه الطرائق أية افتراضات أو معلومات حول خصائص التوزيع للمجتمع، فضلاً عن الوقت المستغرق لتحليل البيانات بهذه الطرائق يكون أقل مما في الطرائق المعلمية، وبالتالي يؤدي الى الاسراع في الحصول على النتائج والافادة منها تطبيقياً وقد ازداد الاهتمام في الآونة الأخيرة بالطرائق اللامعلمية لأهميتها في تقدير دالة البقاء .

2. مشكلة البحث (Problem of Search)

تتمثل مشكلة البحث ان اختلاف بيانات البقاء عن غيرها من البيانات والتي تستعمل فيها الاجرءات الاحصائية القياسية ، وكذلك صعوبة الطرائق المعلمية في حساب دالة البقاء تأتي من عدم إلمام غير المختصين بالاحصاء والرياضيات وبعض فروع الهندسة بالتوزيعات الاحصائيه وكذلك عدم توفر معلومات وافية وجيدة للباحثين لحساب دالة البقاء ، كل هذا مما يدفعنا للبحث عن طرائق بديلة في حساب دالة البقاء منها الطرائق اللامعلمية ولما لهذه الطرائق من اهمية في تقدير دالة البقاء حيث يتم الوصول الى النتائج من دون وجود افتراضات حول التوزيع الإحصائي للبيانات وهذه الميزة التي جعلت التعامل مع البيانات يكون أسهل وأكثر مرونة .

3. هدف البحث (Purpose of Search)

يتمحور هدف البحث الى مقارنة بعض مقدرات دالة البقاء باستعمال الطرائق اللامعلمية لبيانات مراقبة من النوع الاول منها طريقة مقدر(Kaplan- Meier) و طريقة مقدر (Kernel) وطريقة مقدر (Nelson-Aalen) وطريقة مقدر (Thompson) و طريقة مقدر (Pandey) لبيان مدى كفاءتها بإستعمال المؤشرين الإحصائيين متوسط مربعات الخطأ التكاملي (IMSE) و متوسط الخطأ النسبي المطلق (MAPE) للوصول الى افضل طريقة لبيانات مراقبة من النوع الاول .

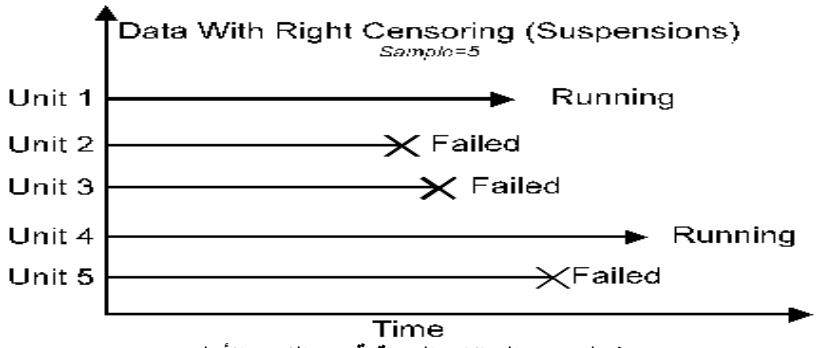
4. تحليل البقاء (Survival analysis)

" تحليل البقاء يتضمن نمذجة الوقت ،أي بدراسة الوقت للمصاب منذ تشخيص الإصابة لحين حدوث الحدث (فالحدث يمثل الموت في أدبيات تحليل البقاء في التجارب الطبية) أو المراقبة (وتشمل التعافي، الانسحاب من المستشفى دون معرفة حالته الصحية أو الموت بسبب آخر غير سبب الدراسة) ، فهو الأسلوب الإحصائي الوحيد الذي يتعامل مع البيانات المراقبة وغير المراقبة" (Kartsonaki, 2016, p.263) .

1-5 أنواع البيانات المراقبة

1-البيانات المراقبة من النوع الاول (Censored Type I Data)

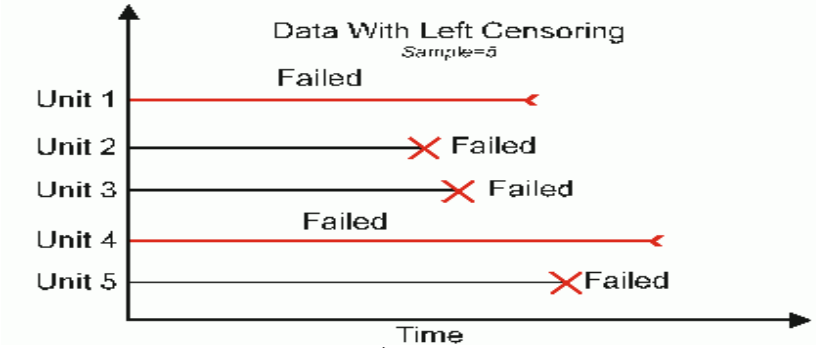
في هذا النوع تتم مراقبة البيانات خلال فترة زمنية معينة أي يكون زمن المراقبة ثابت لجميع وحدات العينة أي انه اذا كان لدينا عينة بحجم n من الوحدات تخضع لاختبار الحياة عند الزمن صفر بحيث نراقب عمل وحدات هذه العينة لحين الوصول الى فترة زمنية محددة لتكن t_0 وبعد الوصول الى هذا الوقت نتوقف عن المراقبة أي ينتهي الاختبار أي ان عدد الحالات التي تتم مراقبتها هو m من الحالات وان m متغير عشوائي لا يمكن تحديده الا بعد انتهاء الزمن t_0 ، واوقات الفشل هي $t_1, t_2, ..., t_r$ وهناك (n-r) لم تجرب الحدث (لم تموت) ، وان الزمن يكون مقدار ثابت ويطلق على هذا النوع بالمراقبة من جهة اليمين (Right censoring) لان اوقات الفشل من جهة اليمين تعتبر مفقودة، من خلال شكل (1) نلاحظ إن المصاب الاول والرابع لم يموتوا أثناء فترة البحث .



شكل (1) البيانات المراقبة من النوع الأول

2-بيانات المراقبة من النوع الثاني (Censored Type II Data)

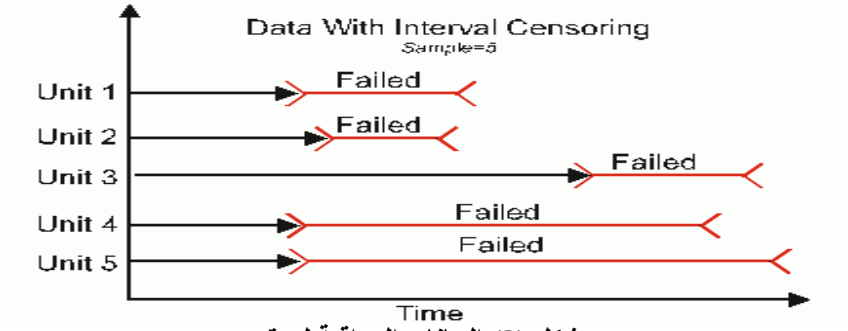
هذا النوع من البيانات تتم مراقبة عدد محدد من الحالات فاذا كان لدينا عينة بحجم n من الحالات توضع على اختبار الحياة عند الزمن صفر ونراقب عمل r من الوحدات اذ ان $r < n$ لذلك فأن الزمن يكون متغير عشوائي لايمكن تحديده ويتوقف الاختبار لحين الوصول على r ، واذا كانت هناك $t_1, t_2, ..., t_r$ من اوقات الفشل ،اذ (r) هو عدد حالات الفشل محدد مسبقا اذ ان الدراسة تنتهي عند الوصول إلى الفشل $T = t_r$، وهى تعني ببساطة ان الحدث وقع عند بداية الدراسة ،وتدعى ايضا بيانات المراقبة من جهة اليسار (Left censoring)



شكل (2) البيانات المراقبة من النوع الثاني

3- **بيانات المراقبة لفترة** (Interval censoring)

إن النوعَ الثالث للمراقبة يدعى فترة البيانات المراقبة، وتعني إن الحالات فشلت ضمن فترةٍ، حيث ان اوقات الفشل لا تعرف بشكل محدد، وبعني بان الحالات سيحدث لها الحدث خلال فترة من الوقت ،ويأتي هذا النوع من البياناتِ كثيراً في الاختبارات الرياضية كسباقات الركض و الجري اذ إن المشاهداتِ لَم تراقب .



شكل (3) البيانات المراقبة لمدة

. (Hoyland & Rausand, 1994, pp 468-469)

2-5 **الطرائق اللامعلمية** (Nonparametric Methods)

اولاً : **طريقة مقدر** (Kaplan-Meier)

(Kaplan-Meier Estimator Method) (KM.E.M)

إن مقدرات **(Kaplan-Meier)** لبيانات البقاء أصبحت مجالاً جيداً للتعامل مع أوقات بقاء مختلفة ، خصوصاً عندما لا تكون جميعها خاضعة للاستمرارية في الدراسة وتعد هذه الطريقة من الطرق الشائعة جداً ،ففي المجال الطبي فان مقدر كابلن ـ مايرمن يستعمل لقياس جزء من حياة المرضى ولفترة زمنية معينة بعد العلاج ، وعندما تكون البيانات تحت المراقبة فإن مقدر **(Kaplan-Meier)** يكون بالصيغة الآتية :

$$\hat{S}(t_i)=\prod_{i=1}^{n}\left(1-\frac{1}{n_i}\right) \quad \text{... (1)}$$

نفرض أن لدينا n من الوحدات في تجربة ورتبت الأوقات المشاهدة لهذه الوحدات n $(t_1, t_2, \dots, t_n)$ فان مقدر **(Kaplan Meier)** لدالة البقاء على قيد الحياة لحالة البيانات المراقبة من اليمين **(right censored data)** يكون بالصيغة الاتية :

$$\hat{S}(t_i) = \prod_{i:Ti\le t}\left(\frac{n-i}{n-i+1}\right)^{d_{(i)}} \qquad for \;\; t \le T_{(n)} \qquad \text{... (2)}$$

وان $di = I_{\{ti=Ti\}}$:مؤشر الى المشاهدات المستقلة المراقبة .

وان رتبة $d_{(i)}$ تشير الى رتبة ذات صلة بـ t_i **(Peterson,1977, p. 854)** .

تم الاعتماد في التقدير في البحث على المعادلة (2) .

ثانياً : **طريقة مقدر** (Kernel Estimator Method) (K.E.M) (Kernel)

تعد هذه الطريقة **(Kernel estimator)** من الطرائق اللامعلمية لتقدير دوال الكثافة والدوال التجميعية ودالة المعّولية ودوال والانحدار وغيرها تم اقتراحه من الباحثان **(Parzen, 1962) and (Rosenblatt, 1956)**

والهدف من استعمال هذا المقدّر هو لغرض تعديل البيانات بالشكل الذي يجعل الحصول على مقدرات ذات صفات تتقارب مع خواص المعلمات الحقيقية ، وهناك أنواع مختلفة من دوال (Kernel) ، وسيتم استعمال دوال (Kernel) من نوع (Gaussian) والمعّرفة بالصيغة الآتية:

$$K(u) = \frac{1}{\sqrt{2\pi}} exp\left(\frac{-u^2}{2}\right) \quad , \; 0 \le u \le 1 \qquad \dots (3)$$

وان استعمال طريقة مقدر (Kernel) يتطلب تحديد المعلمة التمهيدية (h)، أن المعلمة التمهيدية لطريقة مقدر (Kernel) يمكن أن تكون ثابتة .

❖ طريقة مقدر (Kernel) ذو المعلمة التمهيدية الثابتة

(Fixed Bandwidth Kernel estimator Method) (K.F)

تعني الثابتة هي استعمال المعلمة التمهيدية الواحدة على طول الخط الحقيقي المستعمل لتقدير دالة البقاء ويتم تقديرها، نفترض ان $(t_1, t_2, \dots, t_n)$ عينة من المشاهدات المستقلة والمتماثلة التوزيع مع دالة الكثافة الأحتمالية وتعّرف كالآتي:

$$\hat{f}_{(K.F)}(t) = \frac{1}{n\,h}\sum_{i=1}^{n} K\left(\frac{t - T_i}{h}\right) \quad , \quad i = 1, 2, \dots . n \qquad \dots (4)$$

إذ أن :

$K(u)$: تمثل دالة (Kernel) وتحسب من معادلة رقم (3).

t: يمثل زمن تقدير دالة البقاء.

n: حجم العينة المدروسة .

h: تمثل المعلمة التمهيدية (bandwidth)، الثابتة وتقدر حسب الخطوات الآتية :

$$\hat{h} = 1.06\,\hat{\sigma}\,n^{-1/5} \qquad \dots (5)$$

إذ أن :

$\hat{h}$: تقدير المعلمة التمهيدية .

$$\hat{\sigma} = min\left[S\,, \frac{\hat{D}}{1.349}\right] \qquad \dots (6)$$

وان :

$$\hat{D} = X_{(0.75n)} - X_{(0.25n)} \qquad \dots (7)$$

S: الانحراف المعياري للعينة.

$\hat{D}$: الفرق بين القيم الطبيعية المعيارية .

وعليه نحصل على تقدير دالة الكثافة لـ (Kernel) ومنها يمكن أن نحصل على تقدير الدالة التجميعية لـ (Kernel-) وكالاتي:

$$\hat{F}_{(K.F)}(t) = \frac{1}{n\,h}\sum_{i=1}^{n}\int_{-\infty}^{t} K\left(\frac{y - T_i}{h}\right) dy \qquad \dots (8)$$

ومن ثم يمكن الحصول على تقدير دالة البقاء لـ (Kernel) وحسب الصيغة الآتية:

$$\widehat{S}_{(K.F)}(t) = 1 - \left[\frac{1}{n\,h}\sum_{i=1}^{n}\int_{-\infty}^{t} K\left(\frac{y-T_i}{h}\right)\,dy\right] \qquad \dots(9)$$

إذ أن :

$\widehat{S}_{(K.F)}(t)$: تقدير دالة البقاء لـ (Kernel) ذو المعلمة التمهيدية الثابتة.

(حمود ، 2005 ; Miladinovic, 2008).

ثالثاً : طريقة مقدر (Nelson-Aalen)

(Nelson-Aalen Estimator Method) (N.E.M)

تعد هذه الطريقة من الطرائق اللامعلمية لتقدير معدل الخطر التراكمي من بيانات البقاء المراقبة ونظرا لعدم وجود افتراضات التوزيع المطلوب ، استنادا إلى أوقات الحدث الفردية ومعدل الخطر التراكمي H(t) في الوقت t فان :

$$\hat{H}(t) = \sum_{t_i \le t} \frac{d_{(i)}}{n-i+1} \quad , \text{for } t>o \qquad \dots(10)$$

حيث ان $di = I_{\{ti=Ti\}}$: مؤشر الى المشاهدات المستقلة المراقبة .

وان رتبة $d_{(i)}$ تشير الى رتبة ذات صلة بـ t_i .

وعليه فان مقدر Nelson-Aalen يمكن ايجاده بالصيغة الاتية :

$$\widehat{S}\,(t)_{NE} = \exp(-\hat{H}(t)) \qquad \dots(11)$$

(Nelson, 1972, pp 945-966).

رابعاً: طريقة مقدر (Thompson)

(Thompson –Type Estimator Method) (T.E.M)

أقترح الباحث (Thompson) في عام (1968)، طريقة لتقدير معلمات التوزيعات ودالة البقاء بالطرائق المعلمية، في حالة توفر معلومات أولية عن المعلمات بشكل قيمة أولية معلومة مستعملة في طريقة التقلص وهي كالآتي :

$$\widehat{S}_{(T.)}(t) = S_0(t) + \frac{\left[\widehat{S}(t) - S_0(t)\right]^3}{\left[\widehat{S}(t) - S_0(t)\right]^2 + \widehat{var}\left[\widehat{S}(t)\right]} \qquad \dots(12)$$

إذ أن :

$S_0(t)$: قيمة أولية معلومة لدالة البقاء.

$\widehat{S}(t)$: تقدير دالة البقاء بأحدى الطرائق المعلمية .

$\widehat{S}_{(T.)}(t)$: تقدير (Thompson) لدالة البقاء .

$\widehat{var}\left[\widehat{S}(t)\right]$: تقدير تباين دالة البقاء باحدى الطرائق المعلمية.

(Thompson,1968, pp 113-115).

وتم تطوير (صيغة معدلة) لهذه الطريقة باستعمال الاسلوب اللامعلمي وكالآتي :

$$\widehat{S}(t) = \widehat{S}_{(K.F)}(t) + \frac{\left[\widehat{S}_{(K.M)}(t) - \widehat{S}_{(K.F)}(t)\right]^3}{\left[\widehat{S}_{(K.M)}(t) - \widehat{S}_{(K.F)}(t)\right]^2 + \widehat{var}\left[\widehat{S}_{(K.M)}(t)\right]} \qquad \dots(13)$$

إذ أن :

$\widehat{S}(t)$: تقدير دالة البقاء اللامعلمي .

$\widehat{S}_{(K.F)}(t)$: تقدير دالة البقاء بطريقة (Kernel) ذو المعلمة التمهيدية الثابتة.

$\widehat{S}_{(K.M)}(t)$:تقدير دالة البقاء بطريقة (Kaplan-Meier) .

$\widehat{var}[\widehat{S}_{(K.M)}(t)]$: تقدير تباين دالة البقاء المقدرة بطريقة (Kaplan-Meier) .

أن تقدير تباين دالة البقاء وفق طريقة كابلن - ماير (Kaplan-Meier) ، تم باستعمال صيغة تعرف بـ (Greenwood)، وتحسب كالاتي :

$$\widehat{var}[\widehat{S}(t)_{(K.M)}] = \widehat{S}^2(t) \sum_{t_i<t} \frac{1}{n_i(n_i-1)} \qquad \dots(14)$$

إذ أن :

$\widehat{S}^2(t)$: : تقدير دالة البقاء بطريقة كابلن- ماير.

$\widehat{var}[\widehat{S}(t)_{(K.M)}]$: تقدير تباين دالة البقاء .

. (Qabaha,2007, pp 107-109)

خامساً : طريقة مقدر (Pandey)

(Pandey Estimator Method) (P.E.M)

اقترح الباحث (Pandey) في عام (1979) طريقة لتقدير معلمات التوزيعات ودالة البقاء وبالطرائق المعلمية في حالة توفر معلومات أولية عن المعلمات بشكل قيمة أولية معلومة يستعملها في طريقة التقلص وكالآتي :

$$\widehat{S}_{(P)}(t) = \frac{\widehat{d}^2[\widehat{S}(t)]^3}{\widehat{d}^2[\widehat{S}(t)]^2 + K^2\widehat{var}[\widehat{S}(t)]} \qquad \dots(15)$$

إذ أن :

$\widehat{S}_{(P)}(t)$: تقدير (Pandey) لدالة البقاء.

K : قيمة ثابتة , $0<K<1$.

وان:

$$\widehat{d} = K + \frac{(1-K)S_0(t)}{\widehat{S}(t)} \qquad \dots(16)$$

وتم تطوير (صيغة معدلة) لهذه الطريقة باستعمال أسلوب لامعلمي وكالآتي :

$$\widehat{S}_{(P)}(t) = \frac{\widehat{d}^2[\widehat{S}_{(K.M)}(t)]^3}{\widehat{d}^2[\widehat{S}_{(K.M)}(t)]^2 + K^2\widehat{var}[\widehat{S}_{(K.M)}(t)]} \qquad \dots \quad (17)$$

إذ أن :

$$\widehat{d} = K + \frac{(1-K)\widehat{S}_{(K.F)}(t)}{\widehat{S}_{(K.M)}} \qquad \dots(18)$$

$\widehat{S}_{(P)}(t)$: تقدير دالة البقاء اللامعلمي بطريقة (Pandey) .

.(Qabaha,2007, pp 107-109)

5. المحاكاة

تم استعمال أسلوب المحاكاة للطرائق اللامعلمية لتقدير دالة البقاء ، وتعدّ عملية المحاكاة من أفضل الأساليب المستعملة في توليد بيانات بأحجام عينات ، وتمتاز عملية المحاكاة بالمرونة إذ تعطي القدرة على التجريب والاختيار من خلال تكرار العملية، كتب برنامج المحاكاة بإستعمال تطبيق (MATLAP2017b) ونفذ على الحاسبة الألكترونية. وتم توليد أرقام عشوائية U_i تتبع التوزيع المنتظم .

$$U_{i\sim}U_{(0,1)} \quad ,i=1,2,\dots,n$$

U_i: يمثل متغير عشوائي مستمر يتم توليده بإستعمال الحاسبة الألكترونية على وفق الصيغة الآتية :-

$$U = RND$$

... (19)

وجرى تحويل البيانات المولدة التي تتبع التوزيع المنتظم الى التوزيع الطبيعي اللوغاريتمي بمعلمتين بإستعمال دالة التوزيع التجميعية وحسب طريقة التحويل المعكوس الآتية :-

$$f(t,\mu,\sigma)=\frac{1}{t\sigma\sqrt{2\pi}}exp\left\{-\frac{[\ln t-\mu]^2}{2\sigma^2}\right\};\ t>0,\sigma>0\ ,-\infty<\mu<\infty \quad \dots (20)$$

$$F(t,\mu,\sigma)=\Phi\left(\frac{ln(t)-\mu}{\sigma}\right)$$

$$u=F(t,\mu,\sigma)$$

$$u=\Phi\left(\frac{ln(t)-\mu}{\sigma}\right) \quad \dots (21)$$

$$t=F^{-1}(u)$$

$$t=exp[\sigma\emptyset^{-1}(u)+\mu] \quad \dots (22)$$

تم أختيار قيم أفتراضية أولية مختلفة للمعلمتين (معلمة القياس μ و معلمة الشكل σ) للتوزيع الطبيعي اللوغاريتمي وجرى تشكيل ست حالات مبينة في الجدول الآتي:-

جدول (1) يبين القيم الأفتراضية للمعلمات والحالات المقترحة

Cases الحالات	μ معلمة القياس	σ معلمة الشكل
.I	1	0.5

.II	1	1
.III	1.5	0.5
.IV	1.5	1
.V	2	3
.VI	2.5	3

وتم توليد بيانات التوزيع الطبيعي اللوغاريتمي الملوث وفقا للصيغة الاتية :

$$t = \left[(1-\tau)exp[\sigma_1 \emptyset^{-1}(u)+\mu_1]\right] + [\tau exp[\sigma_2 \emptyset^{-1}(u)+\mu_2]\,] \quad \ldots (23)$$

بنسب تلويث (10% , 20% , 30%) = τ، تم أختيار حالتين ملوثة (حمود ، واخرون ، 2013).

وتم اختيار حجوم مختلفة للعينة بشكل يتناسب مع معرفة مدى تأثير حجم العينة على دقة وكفاءة النتائج المستحصلة من طرائق التقدير المستعملة في البحث، وتم استعمال أربعة حجوم للعينات هي: n = (15, 30, 50 ,100)، واختيار التكرار لاحجام العينات والمساوي الى L= 1000 لكل تجربة من أجل الحصول على دقة وتجانس للمقدرات، ولغرض الوصول الى المقدر الأفضل لدالة البقاء من خلال المفاضلة بين طرائق التقدير المستعملة، فقد جرى الاعتماد بشكل عام على المعيارين الاحصائيين كأساس للمقارنة وهما :

1. <u>متوسط مربعات الخطأ التكاملي (IMSE):</u>

لكون (MSE) يحسب لكل (t_i) من الزمن فأن (IMSE) يمثل بمثابة التكامل للمساحة الكلية لـ (t_i) واختزالها بقيمة واحدة تعتبر عامة للزمن ،
او معبرة عن الزمن الكلي وصيغة هذا المقياس تكون كما يأتي :-

$$IMSE[S(T)] = \frac{1}{L}\sum_{i=1}^{L}\{\frac{1}{n_t}\sum_{j=1}^{n_t}[\widehat{S}_i(t_j) - S(t_j)]^2\}$$

$$= \frac{1}{L}\sum_{i=1}^{L} MSE\,[\widehat{S}\,(t_j)] \quad \ldots (24)$$

إذ أن:
L : يمثل عدد مرات تكرار التجربة .

2. <u>متوسط الخطأ النسبي المطلق (MAPE):</u>

وللسبب المذكور في الفقرة السابقة أستُعمل هذا المقياس الذي يحسب على وفق الصيغة الآتية:

$$MAPE = \left[\frac{\sum_{i=1}^{L}|\widehat{S}_i(t) - S(t)|/S(t)}{L}\right] \quad \ldots (25)$$

جدول (2) تفسير الرموز في الجداول

التفاصيل	
K.M	طريقة مقدر كابلن- ماير (Kaplan – Meier Estimator Method)
KEM	طريقة مقدر (Kernel) (Kernel Estimator Method)
N.E	طريقة مقدر (Nelson-Aalen)
T.E	طريقة مقدر (Thompson)
P.E	طريقة مقدر (Pandey)
IMSE	متوسط مربعات الخطأ التكاملي (Integral Mean Squares Error)
MAPE	متوسط الخطأ النسبي المطلق (Mean Absolute Percentages Error)

6. مناقشة نتائج المحاكاة :

سيتم عرض نتائج تجارب المحاكاة وتحليلها لمقدرات دالة البقاء بحسب الطرائق اللامعلمية المبينة في الجانب النظري من هذا البحث وكما يأتي :

جدول (3) يبين متوسط مربعات الخطأ التكاملي (IMSE) و متوسط الخطأ النسبي المطلق (MAPE) لمقدرات دالة البقاء

		متوسط مربعات الخطأ التكاملي (IMSE)						متوسط الخطأ النسبي المطلق (MAPE)					
cases	n	KM	KEM	NE	TE	PE	BEST	KM	KEM	NE	TE	PE	BEST
I	15	0.0000011	0.0173511	0.0028075	0.0034529	0.0127221	KM	0.0012091	0.0887891	0.0425251	0.0970571	0.0815796	KM
	30	0.0000009	0.0062551	0.0013319	0.0006999	0.0055051	KM	0.0010774	0.0382283	0.0205832	0.0355563	0.0388187	KM
	50	0.0000007	0.0047851	0.0007773	0.0005923	0.0035252	KM	0.0007822	0.0255132	0.0120033	0.0214164	0.0236896	KM
	100	0.0000005	0.0034592	0.0002937	0.0003632	0.0000081	KM	0.0009792	0.0145626	0.0051107	0.0141742	0.0119025	KM
II	15	0.0000009	0.0223833	0.0028223	0.0019513	0.0140031	KM	0.0010546	0.0966684	0.0425054	0.1313121	0.0835444	KM
	30	0.0000008	0.0097723	0.0013675	0.0005371	0.0063812	KM	0.0009403	0.0454216	0.0208255	0.0546293	0.0405314	KM
	50	0.0000007	0.0054328	0.0007417	0.0001979	0.0036003	KM	0.0008106	0.0264083	0.0172435	0.0268485	0.0237643	KM
	100	0.0000005	0.0032702	0.0002951	0.0000021	0.0017799	KM	0.0000465	0.0142973	0.0051166	0.0133865	0.0118385	KM
III	15	0.0000013	0.0172972	0.0028058	0.0033767	0.0126711	KM	0.0017115	0.0888361	0.0425033	0.0999243	0.0815715	KM
	30	0.0000011	0.0062575	0.0013617	0.0007101	0.0055266	KM	0.0009113	0.0381772	0.0207782	0.0357564	0.0388213	KM
	50	0.0000007	0.0047498	0.0007303	0.0005175	0.0034255	KM	0.0002614	0.0256914	0.0117975	0.0218842	0.0236722	KM
	100	0.0000006	0.0034918	0.0003162	0.0000037	0.0018608	KM	0.0001695	0.0145355	0.0052448	0.0140985	0.0119163	KM
.I	15	0.0000	0.022	0.002	0.002	0.014	K	0.001	0.096	0.042	0.131	0.083	K

V		014	6093	8663	0575	1791	M	7225	7551	6661	8353	6386	M
	30	0.0000 013	0.009 7085	0.001 3441	0.000 4974	0.006 3349	K M	0.001 0225	0.045 3726	0.020 6632	0.053 8173	0.040 4934	K M
	50	0.0000 009	0.005 3893	0.000 7629	0.000 1868	0.003 5631	K M	0.000 8278	0.026 3957	0.011 9383	0.027 5355	0.023 7456	K M
	10 0	0.0000 007	0.003 3455	0.000 3166	0.000 1357	0.000 0028	K M	0.000 6018	0.014 3214	0.005 2338	0.013 1135	0.011 8642	K M

تتمة جدول(3)

		متوسط مربعات الخطأ التكاملي (IMSE)						متوسط الخطأ النسبي المطلق (MAPE)					
cases	n	KM	KEM	NE	TE	PE	BEST	KM	KEM	NE	TE	PE	BEST
.V	15	0.0000 009	0.024 9842	0.002 8564	0.000 0011	0.013 8531	K M	0.000 9300	0.099 9911	0.042 7381	0.189 2233	0.082 8831	K M
	30	0.0000 008	0.012 5863	0.001 4048	0.000 0009	0.006 9187	K M	0.000 6448	0.049 9692	0.020 9492	0.076 7483	0.041 2623	K M
	50	0.0000 007	0.007 4622	0.000 7566	0.000 0004	0.004 0148	TE	0.000 7692	0.029 9334	0.011 8854	0.040 2134	0.024 5354	K M
	10 0	0.0000 005	0.003 6628	0.000 3176	0.000 0001	0.001 8761	TE	0.000 8900	0.014 8473	0.005 2713	0.015 1175	0.011 9625	K M
.VI	15	0.0000 011	0.025 0592	0.002 8543	0.000 0009	0.013 8941	T E	0.000 9748	0.099 9952	0.042 7075	0.191 0424	0.082 8746	K M
	30	0.0000 009	0.012 5471	0.001 3732	0.000 0007	0.006 8653	T E	0.000 7826	0.049 9816	0.020 7926	0.077 6622	0.041 2177	K M
	50	0.0000 008	0.007 4326	0.000 7251	0.000 0004	0.003 9692	T E	0.000 6592	0.029 9611	0.011 6861	0.040 5741	0.024 5027	K M
	10 0	0.0000 005	0.003 7139	0.000 3018	0.000 0002	0.001 8827	T E	0.000 4281	0.014 9125	0.005 1517	0.015 7824	0.011 9674	K M

جدول (4) قيم المعيارين (IMSE) و (MAPE) لمقدرات دالة البقاء ولجميع مستويات التلوث الحالة الاولى(I)

Sample Size حجوم العينات	الطرائق اللامعلمية	مستويات التلوث الثلاث			مستويات التلوث الثلاث		
		10%	20%	30%	10%	20%	30%
n=15	KM	0.0000006	0.0000006	0.0000004	0.0011069	0.0011592	0.0012593
	KEM	0.0172823	0.0173041	0.0172893	0.0887286	0.0889224	0.0887085
	NE	0.0027564	0.0027837	0.0028344	0.0421876	0.0422764	0.0426583
	TE	0.0034822	0.0034174	0.0034327	0.0963752	0.0965112	0.0956772
	PE	0.0127242	0.0127281	0.0127183	0.0815722	0.0815581	0.0815683
n=30	KM	0.0000004	0.0000003	0.0000002	0.0010205	0.0010568	0.0010995
	KEM	0.0063112	0.0063011	0.0062812	0.0382657	0.0382324	0.0381951
	NE	0.0013633	0.0014057	0.0013081	0.0207554	0.0210345	0.0203873
	TE	0.0007249	0.0007478	0.0007075	0.0361722	0.0358111	0.0358323
	PE	0.0055556	0.0055829	0.0055396	0.0388553	0.0388464	0.0388291
n=50	KM	0.0000004	0.0000003	0.0000003	0.0009939	0.0009382	0.0008841
	KEM	0.0047627	0.0047578	0.0047639	0.0255945	0.0256034	0.0255235
	NE	0.0007446	0.0007517	0.0007618	0.0118336	0.0118852	0.0119255
	TE	0.0005408	0.0005360	0.0005630	0.0219687	0.0213785	0.0214813
	PE	0.0034728	0.0034618	0.0034988	0.0236745	0.0236693	0.0236695
n=100	KM	0.0000003	0.0000002	0.0000003	0.0006147	0.0007084	0.0006283
	KEM	0.0034417	0.0034554	0.0034029	0.0145556	0.0145616	0.0145255
	NE	0.0002936	0.0003071	0.0003061	0.0050935	0.0052069	0.0052453
	TE	0.0000002	0.0000001	0.0000001	0.0147035	0.0143357	0.0145964
	PE	0.0018147	0.0018231	0.0017972	0.0119024	0.0119015	0.0118853

جدول (5) قيم المعيارين (IMSE) و (MAPE) لمقدرات دالة البقاء ولجميع مستويات التلوث الحالة الثانية (II)

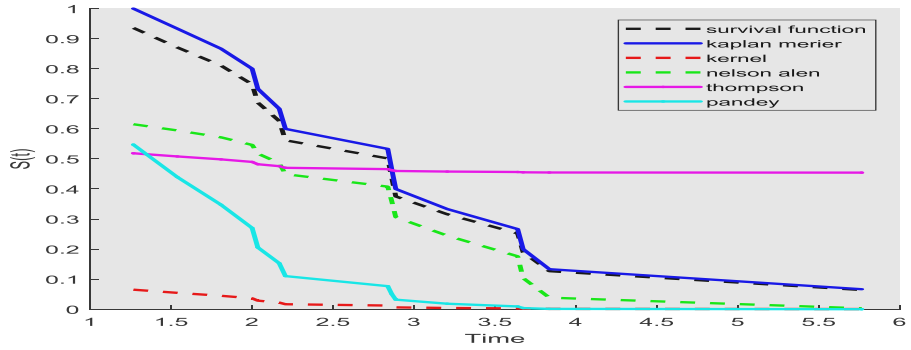
Sample Size حجوم العينات	الطرائق اللامعلمية	مستويات التلوث الثلاث			مستويات التلوث الثلاث		
		10%	20%	30%	10%	20%	30%
n=15	KM	0.0000009	0.0000006	0.0000004	0.0012239	0.0010109	0.0011116
	KEM	0.0225734	0.0223773	0.0223676	0.0966656	0.0967014	0.0966964
	NE	0.0028725	0.0028555	0.0027815	0.0426976	0.0427623	0.0423345
	TE	0.0020639	0.0019701	0.0019702	0.1285454	0.1320577	0.1329565
	PE	0.0141745	0.0139984	0.0139927	0.0836045	0.0835525	0.0835484
n=30	KM	0.0000008	0.0000006	0.0000003	0.0009094	0.0009117	0.0009894
	KEM	0.0097887	0.0097718	0.0097215	0.0455096	0.0454462	0.0453675
	NE	0.0013713	0.0013774	0.0012998	0.0208335	0.0208663	0.0203443
	TE	0.0005197	0.0005236	0.00049336	0.0572846	0.0538234	0.0537733
	PE	0.0063701	0.0063892	0.0063441	0.0405494	0.0405145	0.0404926
n=50	KM	0.0000006	0.0000005	0.0000003	0.0008981	0.0008605	0.0008006
	KEM	0.0054302	0.0054105	0.0053688	0.0263954	0.0264746	0.0264031
	NE	0.0007529	0.0007618	0.0007505	0.0118235	0.0119383	0.0118693
	TE	0.0002004	0.0001768	0.0001781	0.0276846	0.0268153	0.0270124
	PE	0.0036019	0.0035555	0.0035405	0.0237614	0.0237524	0.0237283
n=100	KM	0.0000005	0.0000004	0.0000002	0.0007196	0.0006701	0.0007004
	KEM	0.0032547	0.0033146	0.0033138	0.0142975	0.0142675	0.0143022
	NE	0.0003155	0.0003142	0.0002990	0.0053171	0.0052112	0.0051221
	TE	0.0000019	0.0000015	0.0000011	0.0135753	0.0130986	0.0131323
	PE	0.0017667	0.0018243	0.0018095	0.0118316	0.0118534	0.0118492

8. تحليل نتائج تجربة المحاكاة

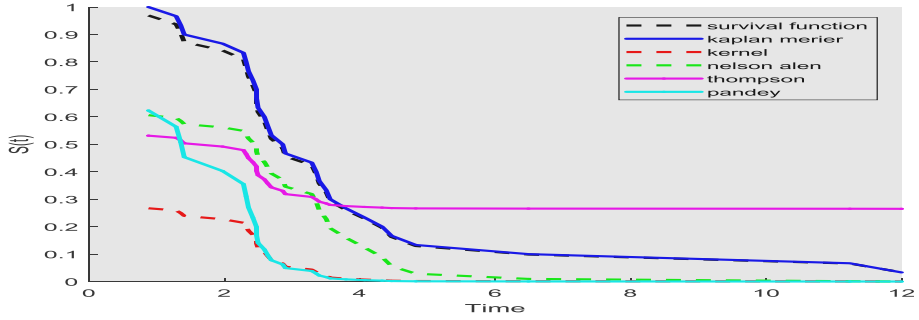
من خلال النتائج في جدول (5,4,3) تبين مايلي :

1. عند حجوم العينات المختلفة (n=(15, 30,50,100 يلاحظ وبصورة عامة افضلية طريقة (Kaplan-Meier) لامتلاكها اصغر قيمة من متوسط مربعات الخطأ التكاملي (IMSE) و متوسط الخطأ النسبي المطلق (MAPE)عن باقي الطرائق اللامعلمية .
2. عند متوسط الخطأ النسبي المطلق (IMSE) للطرائق اللامعلمية الحالة الاولى والثانية والثالثة والرابعة وعند حجم العينة n=15,30,50,100 يلاحظ أفضلية طريقة (Kaplan-Meier) عن باقي الطرائق اللامعلمية.
3. عند متوسط الخطأ النسبي المطلق (IMSE) للطرائق اللامعلمية الحالة الخامسة وعند حجم العينة n=15,30 يلاحظ أفضلية طريقة (Kaplan-Meier) وعند حجم العينة n=50,100 يلاحظ أفضلية طريقة (Thompson) عن باقي الطرائق اللامعلمية .
4. عند متوسط الخطأ النسبي المطلق (IMSE) للطرائق اللامعلمية الحالة السادسة وعند حجم العينة n=15,30,50,100 يلاحظ أفضلية طريقة (Thompson) عن باقي الطرائق اللامعلمية.
5. عند متوسط مربعات الخطأ التكاملي (MAPE) للطرائق اللامعلمية ولجميع الحالات وعند حجم العينة n=15,30,50,100 يلاحظ أفضلية طريقة (Kaplan-Meier) عن باقي الطرائق اللامعلمية .
6. اما عن الطرائق (Nelson-Aalen) و(Pandey) والتي اظهرت نتائجهما لمتوسط مربعات الخطا التكاملي (IMSE) و متوسط الخطأ النسبي المطلق (MAPE) وعند حجم العينة n=15,30,50,100 ، افضل من نتائج الطرائق اللامعلمية (طريقة مقدر Thompson وطريقة مقدر Kernel) بصورة عامة ، وذلك من خلال اقتراح معدل لمقدر كابلن-ماير بالطرق المعدلة في تلك الطرائق وكما مبينة في الجانب النظري وكذلك من خلال اقترابهما من قيم دالة البقاء الافتراضية الحقيقية .
7. عند متوسط الخطأ النسبي المطلق (IMSE) للطرائق اللامعلمية الحالة الاولى وعند حجم العينة n=15,30,50ولنسب التلوث (10% , 20% , 30%) يلاحظ أفضلية طريقة (Kaplan-Meier) وعند حجم العينة n=100 يلاحظ أفضلية طريقة (Thompson) .
8. عند متوسط الخطأ النسبي المطلق (IMSE) للطرائق اللامعلمية الحالة الثانية وعند حجم العينة n=15,30,50,100ولنسب التلوث (10% , 20% , 30%) يلاحظ أفضلية طريقة (Kaplan-Meier) عن باقي الطرائق اللامعلمية .
9. عند متوسط مربعات الخطأ التكاملي (MAPE) للطرائق اللامعلمية ولجميع الحالات وعند حجم العينة n=15,30,50,100 ولنسب التلوث (10% , 20% , 30%) يلاحظ أفضلية طريقة (Kaplan-Meier) عن باقي الطرائق اللامعلمية .
10. لجميع النماذج ولجميع حجوم العينات ونسب التلوث يلاحظ تناقص (IMSE) و (MAPE) عند زيادة حجم العينة.
11. الاشكال (4) و(5) و(6) و(7) و(8) و(9) و(10) و(11) و(12) و(13) توضح النتائج لأفضل طريقة من الطرائق اللامعلمية ولحجوم العينات المختلفة (n=15,30,50,100) واقتراب قيم دالة البقاء التقديرية من قيم دالة البقاء الحقيقية .

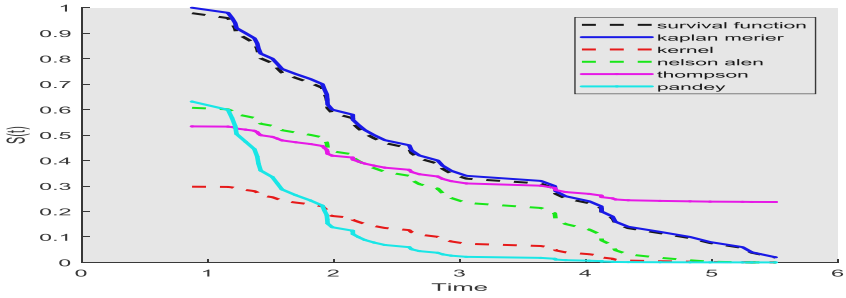
شكل (4) يوضح الطرائق اللامعلمية لدالة البقاء لحجوم العينة للحالة الاولى (I) لنسبة التلوث 10%



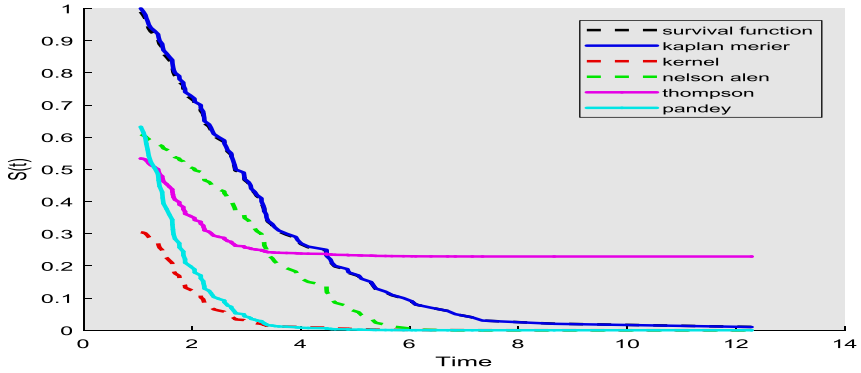
n=15



n=30



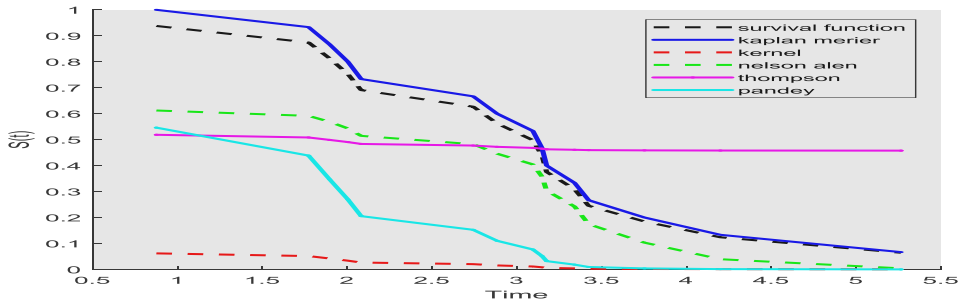
n=50



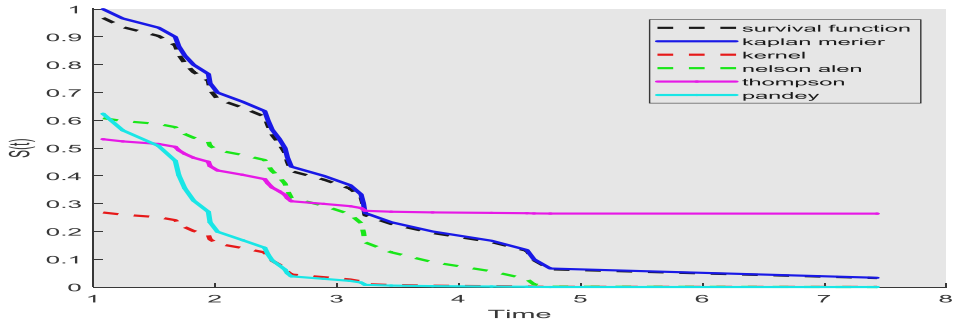
n=100

شكل (5) يوضح الطرائق اللامعلمية لدالة البقاء لحجوم العينة للحالة الاولى (I) لنسبة التلوث 20%

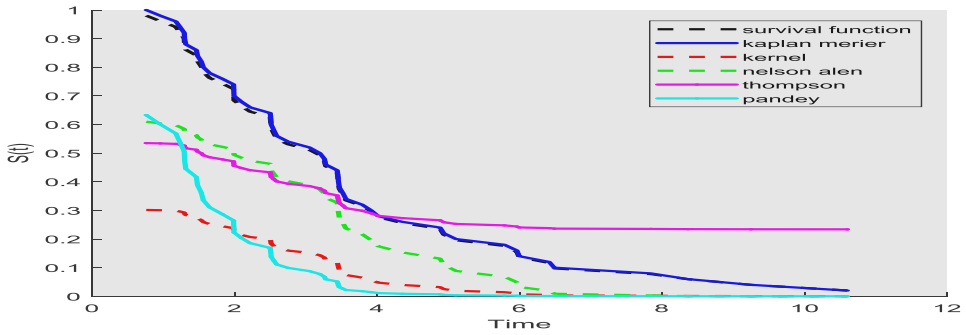
n=15



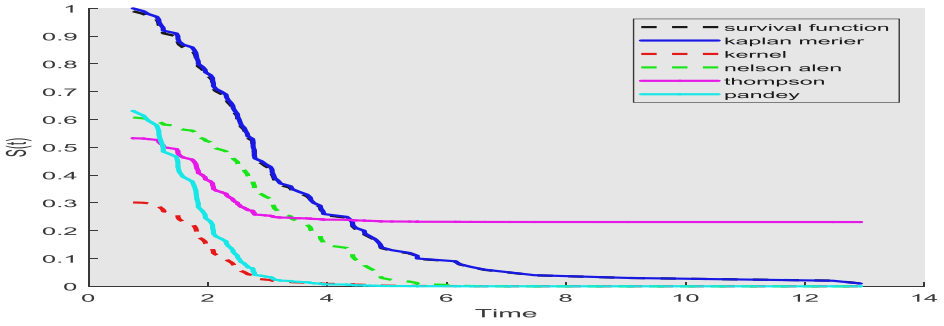
n=30



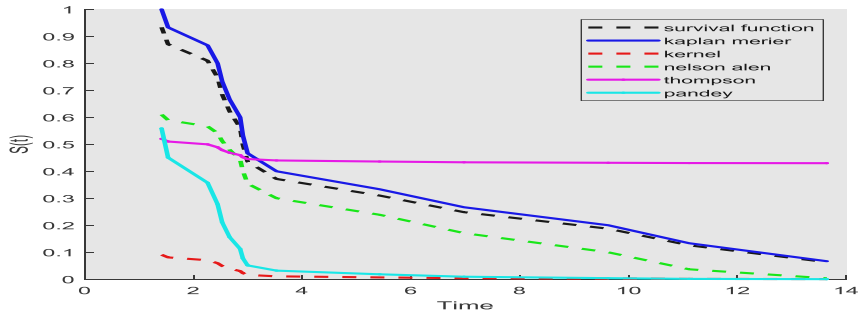
n=50



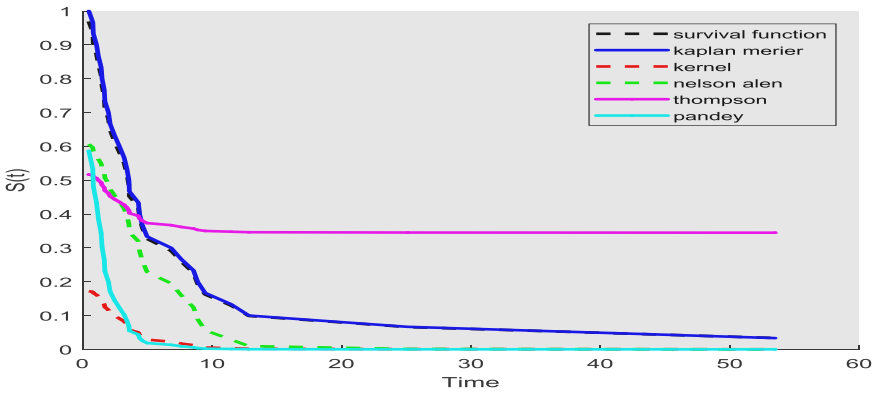
n=100



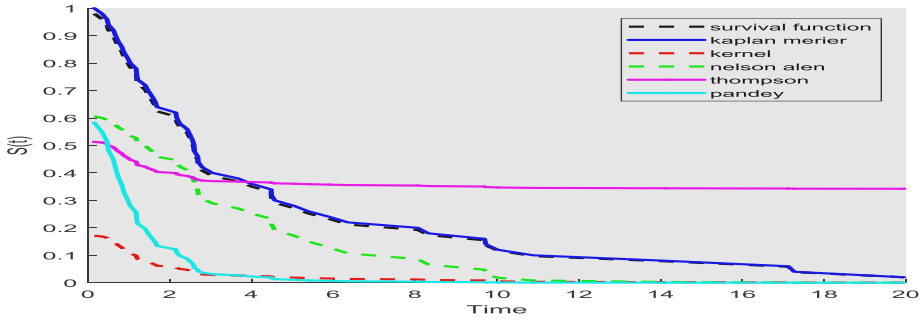
شكل (6) يوضح الطرائق اللامعلمية لدالة البقاء لحجوم العينة للحالة الاولى (I) لنسبة التلوث 30%



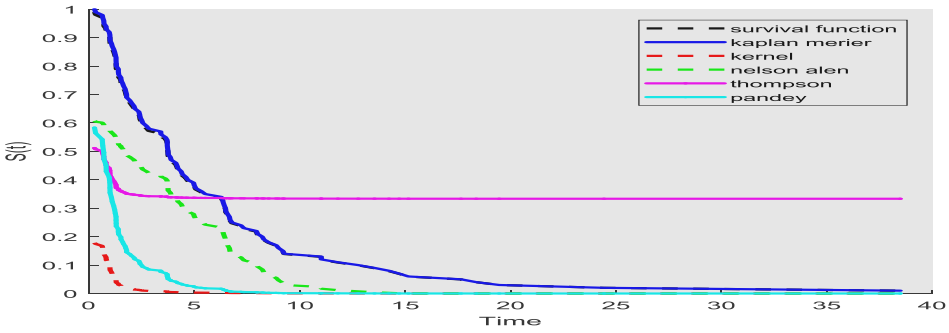
n=15



n=30



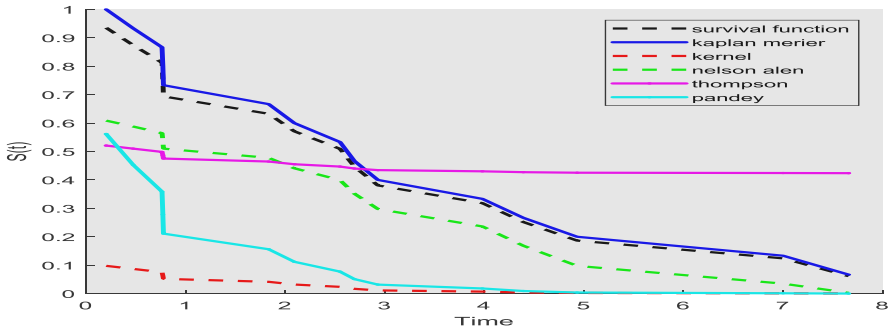
n=50



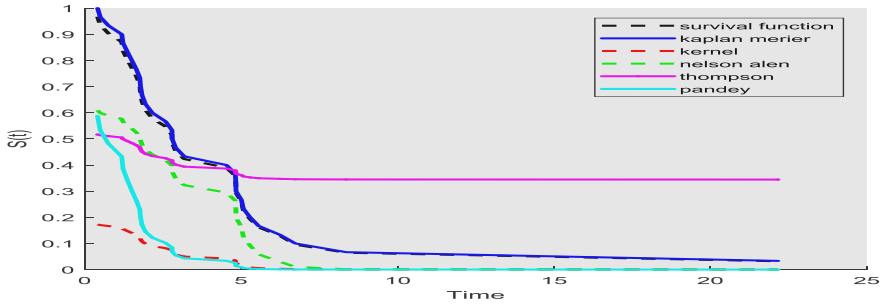
n=100

شكل (7) يوضح الطرائق اللامعلمية لدالة البقاء لحجوم العينة للحالة الثانية (II) لنسبة التلوث 10%

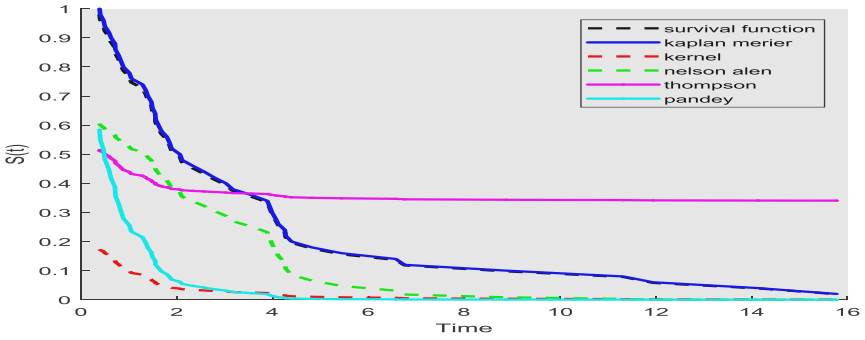
n=15



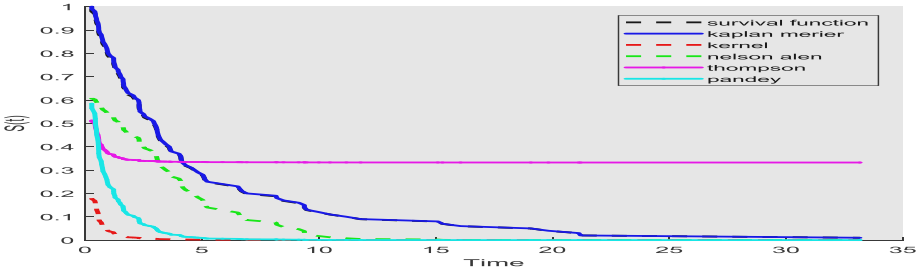
n=30



n=50

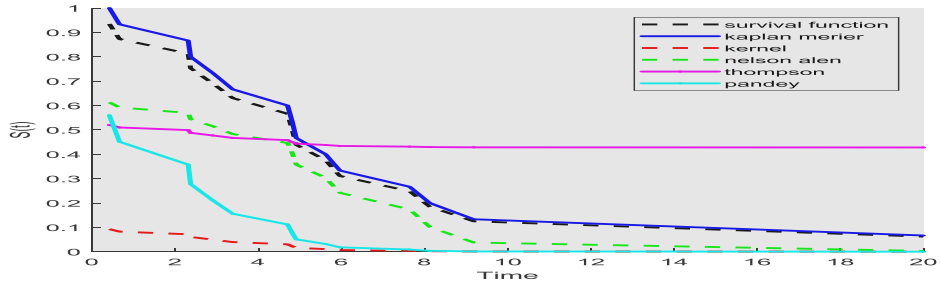


n=100

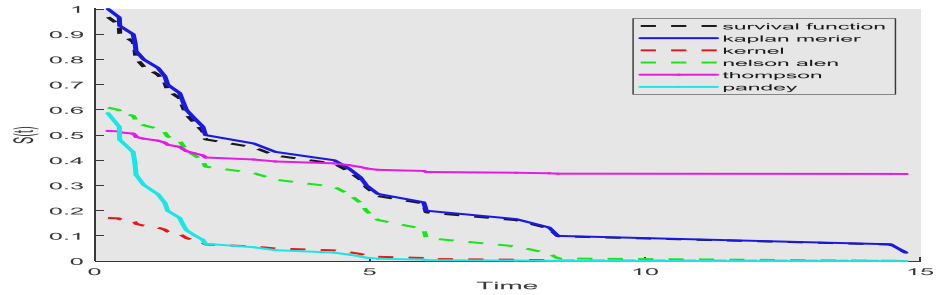


شكل (8) يوضح الطرائق اللامعلمية لدالة البقاء لحجوم العينة للحالة الثانية (II) لنسبة التلوث 20%

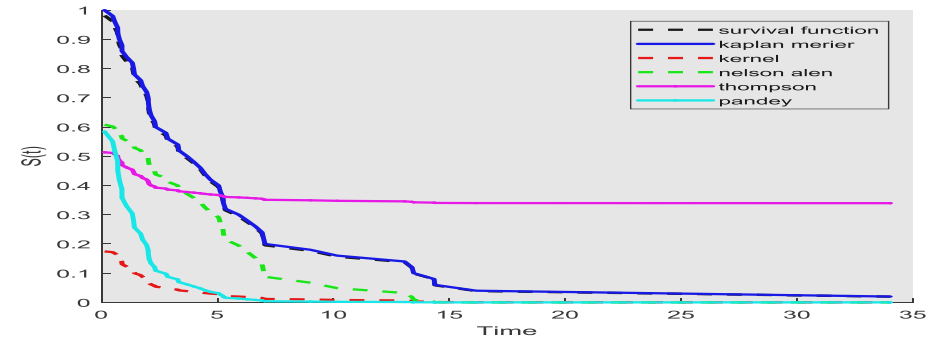
n=15



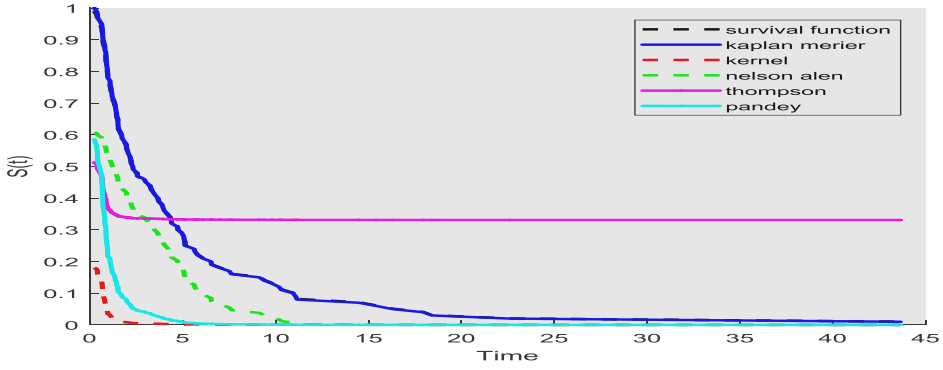
n=30



n=50

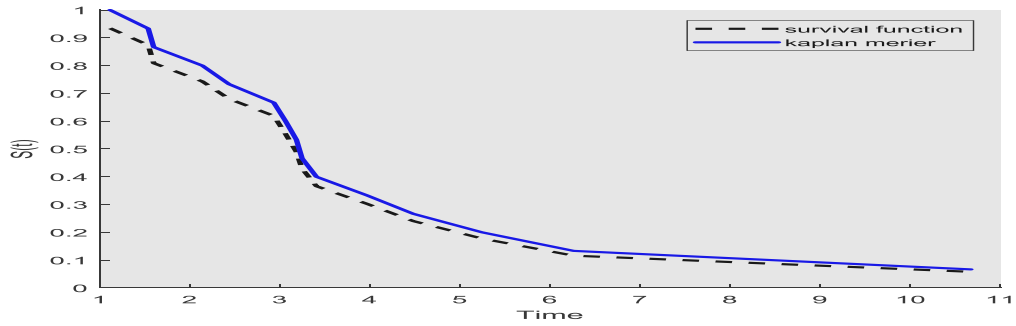


n=100

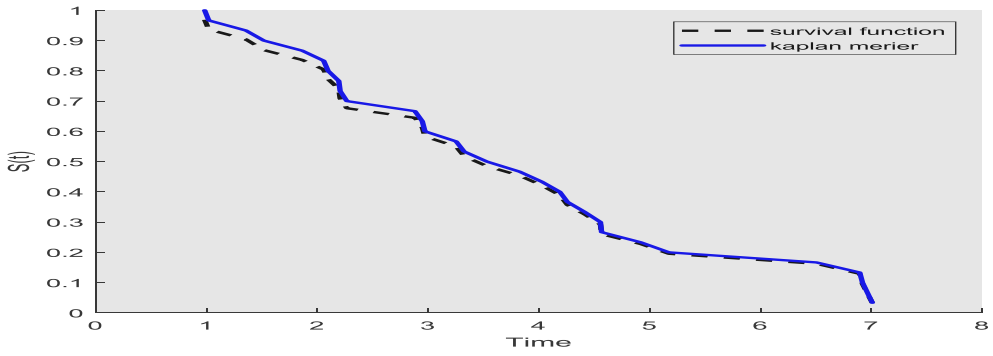


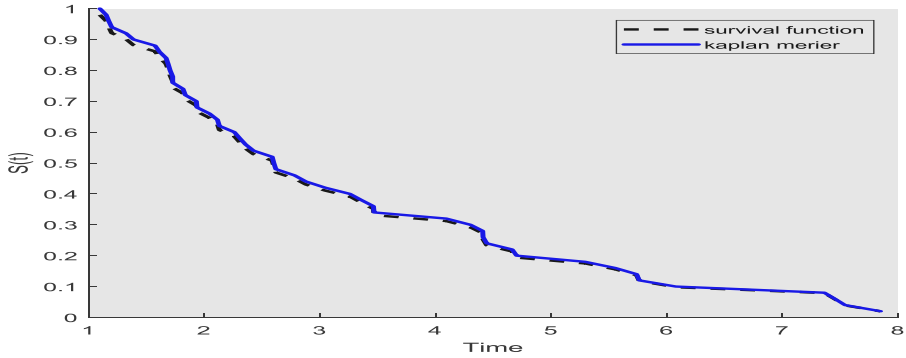
شكل (9) يوضح الطرائق اللامعلمية لدالة البقاء لحجوم العينة للحالة الثانية (II) لنسبة التلوث 30%

n=15

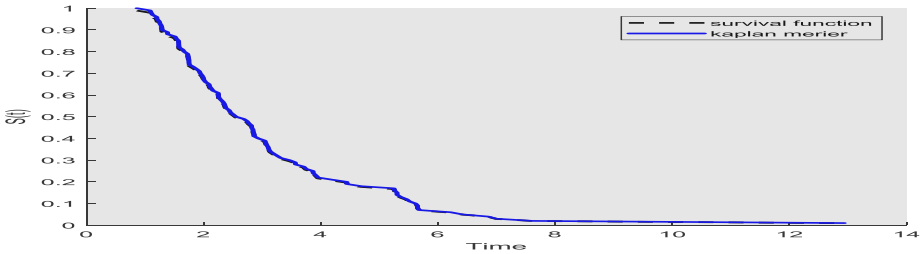


n=30





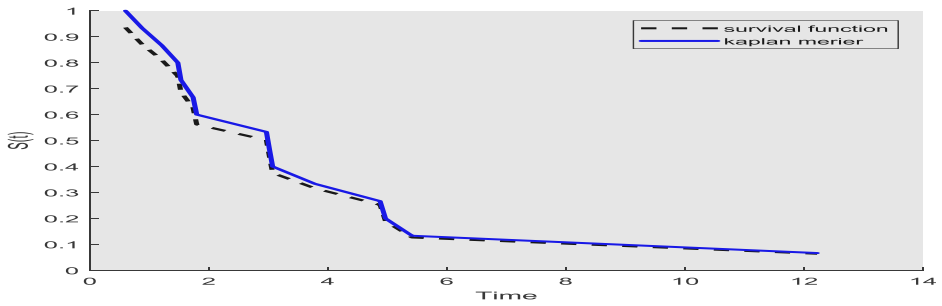
n=50



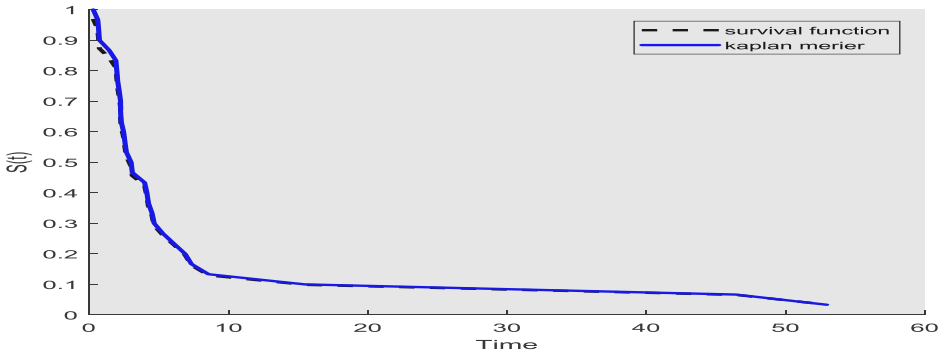
n=100

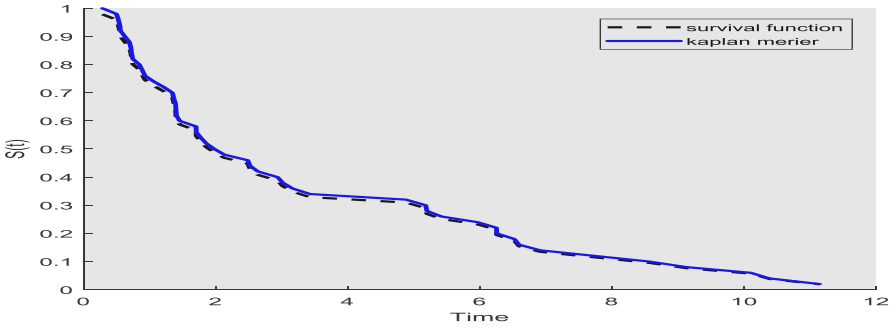
شكل (10) يوضح طريقة كابلن ـ ماير ودالة البقاء الحقيقية لحجوم العينة للحالة الاولى (I)

n=15

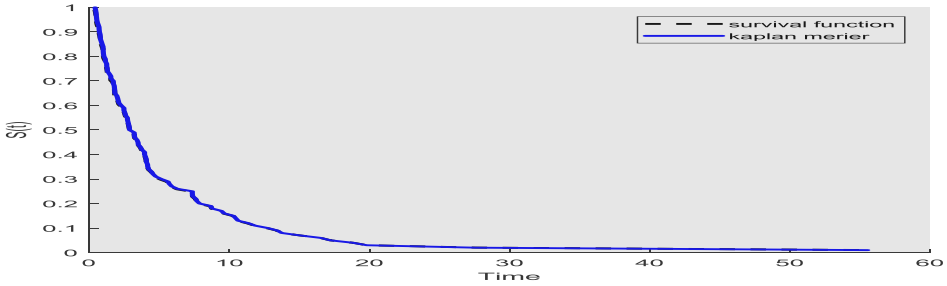


n=30





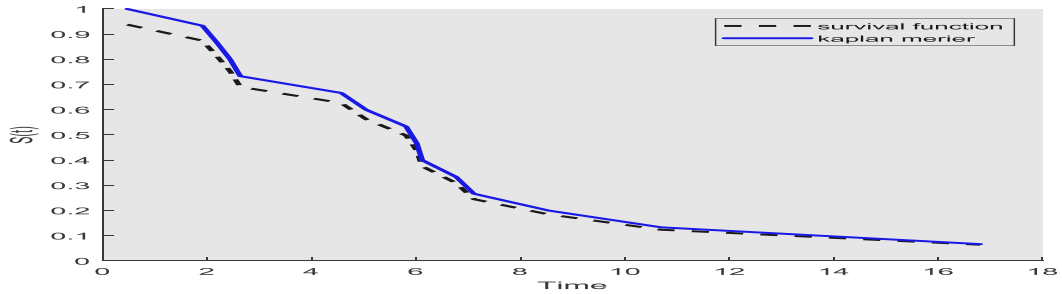
n=50



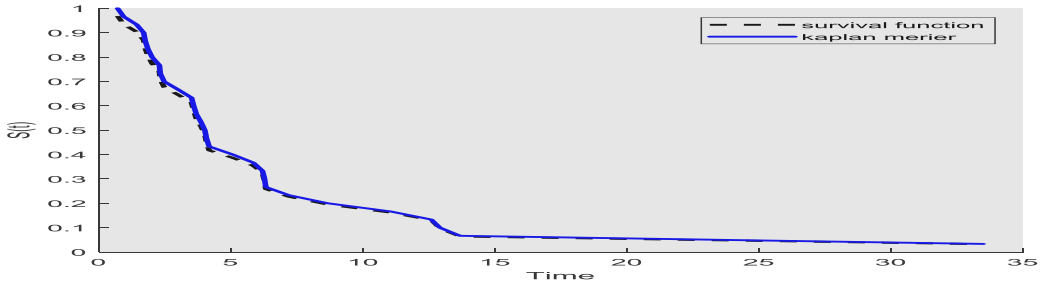
n=100

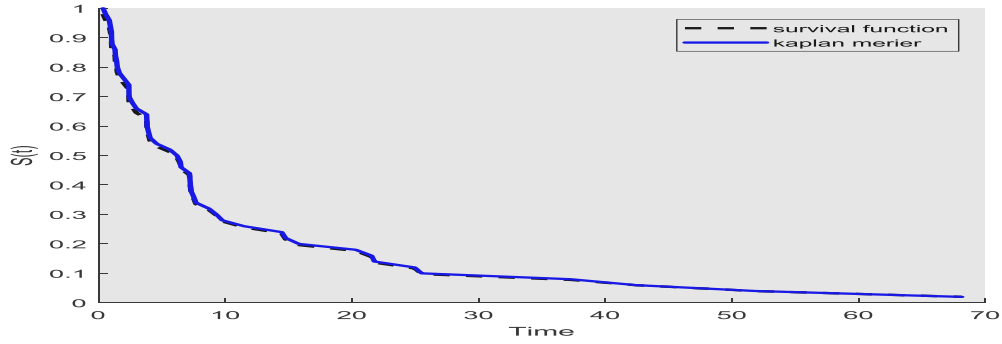
شكل (11) يوضح طريقة كابلن ـ ماير ودالة البقاء الحقيقية لحجوم العينة للحالة الثانية(II)

n=15

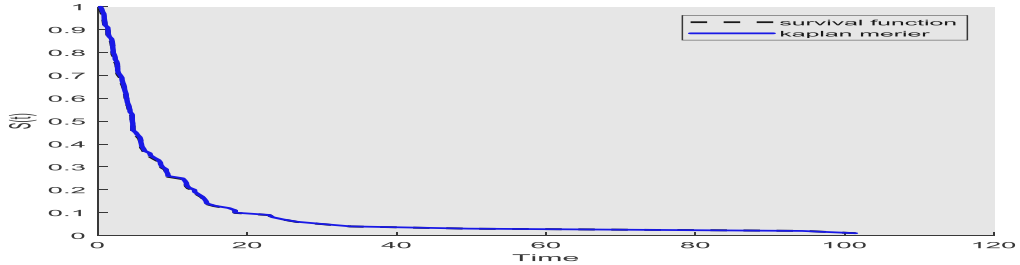


n=30





n=50



n=100

شكل (13) يوضح طريقة كابلن ـ ماير ودالة البقاء الحقيقية لحجوم العينة للحالة الرابعة(IV.)

9. الاستنتاجات

1. اظهرت نتائج الجانب التجريبي وبالاعتماد على المعيارين الاحصائيين متوسط مربعات الخطأ التكاملي (IMSE) ومتوسط الخطأ النسبي المطلق (MAPE)، ان افضل الطرائق اللامعلمية هي طريقة كابلن –ماير لتقدير دالة البقاء التي تمتلك اكثر عدد مرات أفضلية من الطرائق اللامعلمية الاخرى المستعملة في هذا البحث ، وان قيم دالة البقاء لطريقة كابلن - ماير تكون قريبة من قيم دالة البقاء الحقيقية في اغلب الحالات ولحجوم العينات المختلفة .

2. اظهرت الطرائق اللامعلمية حصانة للمقدرات (دالة البقاء) ومقارنتها للتلوث المختلفة وباختلاف حجوم العينات المشار اليها .

3. كلما زادت نسبة التلوث زادت دقة كفاءة المقدر وقلت قيمة (IMSE) و(MAPE) ، اي ان المقدر حصين وله امكانية في مقاومة التلوث ، اي ان جزء منها يتبع توزيع محدد وجزء يتبع توزيع مختلف .

10. التوصيات

من خلال اجراءات البحث واستنتاجاته تم التوصل الى مجموعة من التوصيات اهمها:

1-ضرورة اعتمـاد طريقة (كابلن- ماير (KM)) للطرائق اللامعلمية لايجاد دالة البقاء للبيانات المراقبة من النوع الاول.

2-ضرورة اعتماد وزارة الصحة لمقدرات دالة البقاء لامراض السرطان وامراض اخرى والافادة في وضع الخطط العلاجية المستقبلية ومتابعة المرضى .

11. المصادر

أولاً : المصادر العربية : (Arabic Reference)

1. حمود، مناف يوسف (2005) "مقارنة المقدرات اللامعلمية لتقدير دوال الكثافة الاحتمالية"، أطروحة دكتوراه فلسفة في علوم الاحصاء مقدمة الى مجلس كلية الادارة والاقتصاد، جامعة بغداد.

2. حمود، مناف يوسف ،نايف، قتيبة نبيل .،محمد، محمد جاسم (2013) "بعض المقدرات اللامعلمية لتقدير دوال البقاء او المعولية لتوزيع ويبل الملوث "، المؤتمر العلمي الرابع لاتحاد الاحصائيين العرب .

ثانياً : المصادر الاجنبية : (English Reference)

3. Christiana Kartsonaki. (2016) " Survival analysis", Diagnostic Histopathology Volume 22, Issue 7, July 2016, Pages 263-270.

4. Hoyland, A. and Rausand, M. (1994) "System reliability theory: models and statistical methods", Wiley, New York.

5. Miladinovic, B. (2008) "Kernel density estimation of reliability with applications to extreme value distribution", university of south Florida, Scholar commons, theses and dissertations.

6. Nelson, W. (1972)" Theory and application of hazard plotting for censored failure Data ", Techno metrics 14:945-966.

7. Peterson, Arthur V., Jr. (1977) " Expressing the Kaplan-Meier estimator as a function of empirical sub survival functions",Journal of the American Statistical Association, Vol. 72, No. 360, pp. 854- 858.

8. Qabaha, M. (2007) "Ordinary and Bayesian shrinkage estimation", An-Najah university, J. Res. (N. Sc), vol.21.

9. Thompson, J.R. (1968)" Some Shrinkage techniques for estimating the mean", Journal of the American Statistical Association, 63:321, 113-122.

Some NONPARAMETRIC ESTIMATORS FOR RIGHT CENSORED SURVIVAL DATA

Abstract

The using of the parametric models and the subsequent estimation methods require the presence of many of the primary conditions to be met by those models to represent the population under study adequately, these prompting researchers to search for more flexible parametric models and these models were nonparametric, many researchers, are interested in the study of the function of permanence and its estimation methods, one of these non-parametric methods.

For work of purpose statistical inference parameters around the statistical distribution for life times which censored data , on the experimental section of this thesis has been the comparison of non-parametric methods of permanence function, the existence of surveillance (Type I- censored data) employing simulation style using (Kaplan-Meier estimation method, Kernel estimation method, Nelson-Aalen estimation method, Thompson–Type estimation method and Pandey estimation method) these methods have the most flexibility in data analysis the statements with no knowledge of the distribution who inserts data for the estimator, to get the best way to assess the permanence function using the simulation method of two of the statisticians measure, (IMSE) Integral Mean Squares Error and Mean Absolute Percentages Error (MAPE) for different sample size like (n = 15, 30, 50 ,100), has been reached to a preference of the best way to estimate is Kaplan-Meier method from the remainder of the nonparametric methods, the results show that permanence function values start to decrease with increasing of time in relation to nonparametric estimation.

Key Word: Kaplan-Meier Method, Kernel Method, Nelson-Aalen Method, Thompson –Type Method, Pandey Method, Integral Mean Squares Error, Mean Absolute Percentages Error .