Journal of Economics and Administrative Sciences (JEAS)

Available online at http://jeasiq.uobaghdad.edu.iq

تقدير دالة البقاء بأستعمال الخوارزمية الجينية

الباحث/ انور طاهر عبدالهادي
anwertaher17@gmail.com

أ.م.د. صباح منفي رضا
كلية الادارة والاقتصاد ـ جامعة بغداد
drsabah@coadec.uobaghdad.edu.iq

Received:11/3/2020 **Accepted :16/8/2020** **Published :October / 2020**

مستخلص البحث :

ان تحليل البقاء هو عبارة عن تحليل البيانات التي تكون في شكل اوقات من اصل الوقت حتى حدوث حدث النهاية ، وفي البحوث الطبية يكون اصل الوقت هو تاريخ تسجيل المفردة او المريض في دراسة ما مثل التجارب السريرية لمقارنة نوعين من الدواء او اكثر اذا كانت نقطة النهاية هي وفاة المريض او اختفاء المفردة فالبيانات الناتجة من هذه العملية تسمى اوقات البقاء اما اذا كانت النهاية هي ليست الوفاة فالبيانات الناتجة تسمى بيانات الوقت حتى الحدث ، اي ان تحليل البقاء هو من الخطوات والاجراءات الاحصائية لتحليل البيانات عندما يكون المتغير المعتمد هو الوقت حتى الحدث والوقت قد يكون ايام او اسابيع او اشهر او سنوات منذ بداية تسجيل المفردة حتى الحدث .

يهتم هذا البحث بمسألة تقدير دالة البقاء لبيانات مراقبة بأستعمال احد اهم خوارزميات الذكاء الاصطناعي وهي الخوارزمية الجينية وذلك بهدف الحصول على تقديرات مثلى لمعلمات توزيع ويبل وهذا بدوره ينعكس على تقديرات داله البقاء حيث يتم توظيف الخوارزمية الجينية في طريقة الامكان الاعظم وطريقة العزوم ، طريقة المربعات الصغرى و طريقة المربعات الصغرى الموزونة المعدلة والحصول على مقدرات اكثر كفاءة من الطرق التقليدية , ثم بعد ذلك سوف يتم اجراء مقارنة بين الطرق بالاعتماد على الجانب التجريبي ويتم تقييم الطريقة الافضل بالاعتماد على معيار متوسط مربعات الخطأ لدالة البقاء , كذلك سوف يتم تطبيق الطرق على بيانات حقيقة لمرضى سرطان الرئة والقصبات .

وتوصلت الدراسة الى ان افضل طريقة لتقدير معلمات توزيع ويبل ودالة البقاء التي انتجها الجانب التجريبي هي طريقة المربعات الصغرى المسندة الى الخوارزمية الجينية

المصطلحات الرئيسة للبحث : دالة البقاء , توزيع ويبل , الخوارزمية الجينية , طريقة الامكان الاعظم , طريقة العزوم , طريقة المربعات الصغرى , بيانات المراقبة

1 - المقدمة Introduction

لغرض دراسة وقت البقاء عند الاصابة بمرض من الامراض الخطرة لابد من تحديد الانموذج المناسب او التوزيع المناسب الذي يتبعه الوقت المراد دراسته وبعد ان نحدد الانموذج المناسب يتم تقدير هذا الانموذج وان بعض المقدرات تكون غير كفوءة بسبب وجود معادلات غير خطية في النموذج الذي تم تحديده عند استعمال الطرائق الكلاسيكية الاعتيادية كطريقة الامكان الاعظم (MLE) وطريقة العزوم (MOM) وطريقة المربعات الصغرى (LSM) مما قد يؤدي الى الحصول على تقدير غير دقيق لدالة البقاء .

ولأهمية موضوع زمن البقاء وتأثره بعوامل متعددة فقد ظهرت الحاجة الماسة لتطوير الاساليب والوسائل الاحصائية لزيادة الدقة والمعرفة الشاملة لدراسة وقت البقاء ، لذا تعددت نماذج البقاء فمنها ما يدرس متغير الوقت فقط ومنها ما يتوسع ليشمل الوقت وبعض المتغيرات الاساسية مثل العمر الجنس وغيرها عند القيام بعملية التقدير, ولغرض الحصول على مقدرات جيدة يمكن الاعتماد عليها للوصول الى نتائج اكثر دقة لذلك يجب اختيار الطريقة المناسبة للتقدير .

ان الهدف من هذا البحث هو توظيف الخوارزمية الجينية في تقدير معلمات توزيع ويبل ذو معلمتين للحصول على تقدير امثل للمعلمات من خلال استعمال طرائق التقدير الكلاسيكية الاعتيادية (MLE , MOM , LSM) , و مقارنتها مع تقديرات الطرائق الإعتيادية لتقدير دالة البقاء ، باستعمال المقياس الإحصائي أقل متوسط لمربعات الخطأ (MSE) بإستعمال المحاكاة لغرض المقارنة بغية التوصل إلى أفضل طريقة في التقدير لقد تم التطرق الى دوال البقاء وتحليلها منذ زمن اذ بدأت الدراسات الجدية لها منذ خمسينات القرن الماضي .

وفي عام (2011) قام الباحثان [Gasmi & Berzig] [5] بدراسة " تقدير معلمات توزيع ويبل المعدل بالاعتماد على معاينة من النوع الاول" حيث تضمن البحث استعمال طريقة الامكان الاعظم ومن ثم استعمال مصفوفة المعلومات (fisher) للحصول على التقدير الفترة للمعلمات وبعد ذلك تقدير دالة الوفاة ودالة البقاء ودالة المخاطرة وتم تطبيقها على بيانات حقيقية وفي عام (2015) قام الباحثون (Akma et al) [1] بدراسة "بعض طرائق تحليل البقاء بواسطة نموذج ويبل بوجود بيانات المراقبة " حيث تضمنت الدراسة طريقة تقدير معلمات توزيع ويبل هي طريقة الامكان الاعظم وكذلك طريقة الرسم الاحتمالي وثم تقدير دالة البقاء بالاعتماد على بيانات المراقبة . في عام (2017) قام [Chang-Jun واخرون] [12] بدراسة " تقييم معلمات توزيع ويبل ذو ثلاث معلمات بالاعتماد على الخوارزمية الجينية المعدلة " حيث تم تقدير معلمات توزيع ويبل ذو ثلاث معلمات باستعمال الخوارزمية الجينية من خلال توظيف الخوارزمية الجينية بمقدرات الامكان الاعظم وتمت مقارنة الخورزمية الجينية التقليدية مع الخورزمية المعدلة في تقدير معلمات توزيع ويبل ذو ثلاث معلمات وتوصلوا ان الخوارزمية الجينية المطورة او المعدلة اعطت تقدير افضل من الخوارزمية الجينية التقليدية بالاعتماد على دراسة المحاكاة .

2 – الطرائق والادوات Methods and Tools

1-2 بيانات المراقبة Censoring Data

تعد بيانات المراقبة مصدراً للصعوبة في تحليل بيانات البقاء وتكمن صعوبتها في ان بعض المفردات لا يمكن مشاهدتها او ملاحظتها خلال وقت الدراسة حتى الحدث اذ يقال إن وقت بقاء الفرد يخضع للمراقبة عندما لا يتم مشاهدة او ملاحظة نقطة النهاية لذلك الفرد. قد يكون هذا بسبب بيانات الدراسة حللت عند نقطة ما في وقت لا يزال فيه بعض الأفراد على قيد الحياة. كذلك قد يكون حالة البقاء في وقت التحليل غير معلومة وذلك لان بعض المفردات قد فقدت المتابعة , على سبيل المثال افترض أنه بعد تحديد المريض في تجربة سريرية ، ينتقل المريض إلى جزء آخر من البلاد ، أو إلى بلد آخر ، ولا يمكن تتبعه بعد الآن. المعلومات الوحيدة المتوفرة عن تجربة البقاء على قيد الحياة لهذا المريض هي آخر تاريخ عرف فيه أنه على قيد الحياة.[3]

2-2 دالة البقاء The Survival Function

إن المقدار الأساسي المستعمل لوصف ظواهر وقت الحدث هي دالة البقاء ، وهي احتمال بقاء شخص ما على قيد الحياة بعد الوقت t يتم تعريفها على أنها [6] :

$$S(t) = p(T > t) \qquad \dots (1)$$

$S(t)$: تمثل دالة البقاء

T : متغير عشوائي مستمر الذي يمثل وقت البقاء على قيد الحياة .

بذلك تكون دالة البقاء بالاعتماد على دالة التوزيع كما في الصيغة الاتية :

$$S(t) = 1 - F(t) \qquad \dots (2)$$

حيث ان :

$F(t)$: تمثل دالة التوزيع التراكمية (The Cumulative Distribution Function)

$$F(t) = pr(T \leq t) \quad \dots (3)$$

يتم رسم منحني دالة البقاء $S(t)$ ليمثل تمثيلا بيانيا لاحتمال بقاء الفرد في نقاط زمنية معينة , وان منحنيات البقاء تتبع الخصائص الاتية:

دالة رتيبة (Monotone Function).

دالة متناقصة (Decreasing Function).

تكون دالة البقاء مساوية للواحد $S(t) = 1$ عندما يكون الوقت مساويا الى الصفر اي $(t = 0)$

في اغلب الاحيان كلما زاد عمر الشخص فان احتمال البقاء يقترب من الصفر اي انه:

$$S(t) \to 0 \ \ as \ t \to \infty$$

تتخذ منحنيات البقاء مجموعة متنوعة من الاشكال اعتمادا على توزيع البيانات وان جميع هذه المنحنيات تتمز بهذه الخصائص [13]

2- 3 توزيع ويبل Weibull Distribution

يعد توزيع ويبل من التوزيعات المهمة في دراسة أوقات الحياة للإنسان أي مدى بقاء الانسان على قيد الحياة عند اصابته بمرض خطير ويعتبر توزيع ويبل من التوزيعات المستمرة وجده لأول مرة (Walodd Weibull) في سنة 1951م حيث انه من الممكن ان يعتمد على عدة معالم لذلك فان توزيع ويبل يأخذ الاهمية القصوى في الدراسات العلمية التي تعتمد تحديد فترة البقاء وان المعالم الموجودة في التوزيع سواء كانت معلمتين او اكثر تشير الى ان التوزيع له مجال $0 \leq x < \infty$ وان معلمة القياس α تكون اكبر من الصفر وان معلمة الشكل β يجب ان تكون اكبر من الصفر وان دالة الكثافة الاحتمالية لتوزيع ويبل ذي معلمتين تأخذ الصيغة آلاتية [10] :

$$f(t, \alpha, \beta) = \frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1} \exp\left(-\left(\frac{t}{\alpha}\right)^{\beta}\right) \quad \alpha, \beta > 0, I_{(0,\alpha)} \quad \dots (4)$$

حيث ان :

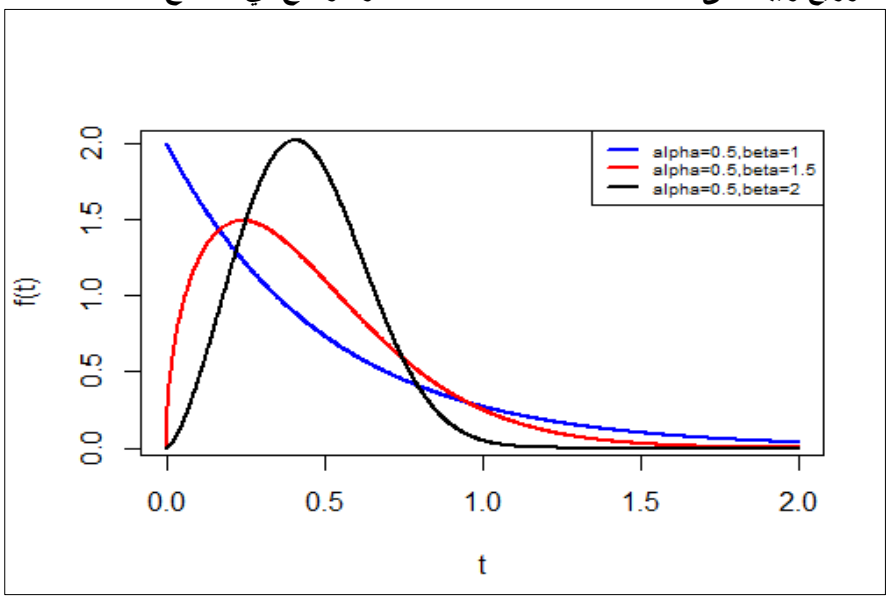
β : معلمة الشكل

α : معلمة القياس

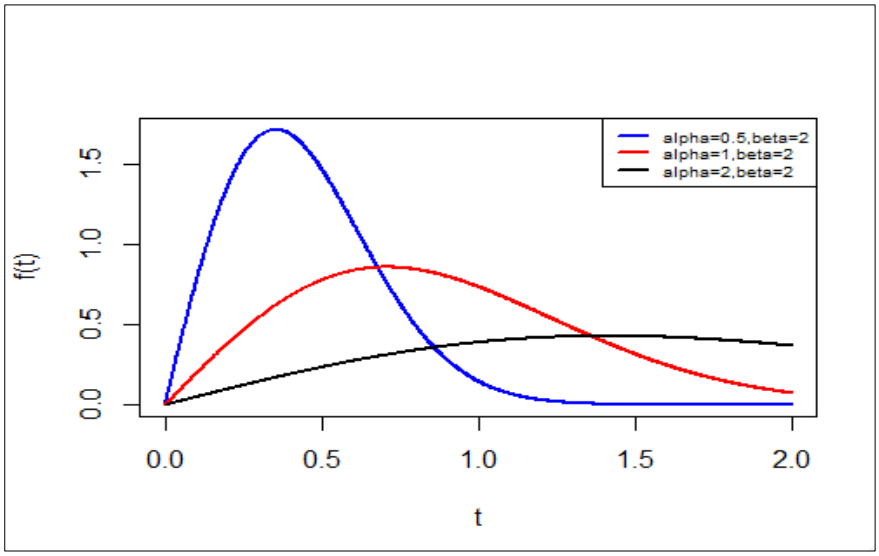
اما بالنسبة لدالة التوزيع التراكمية فتأخذ الصيغة الاتية

$$F(t) = 1 - \exp\left(-\left(\frac{t}{\alpha}\right)^{\beta}\right) \quad \dots (5)$$

تؤثر معلمات توزيع ويبل على شكل دالة الكثافة الاحتمال كما هو موضح في النماذج أدناه:



الشكل (1) تأثير معلمة الشكل في شكل دالة كثافة الاحتمال لتوزيع ويبل ذو معلمتين



الشكل (2) تأثير معلمة القياس في شكل دالة الكثافة الاحتمال لتوزيع ويبل ذو معلمتين

2-4دالة البقاء لتوزيع ويبل The Survival Function of Weibull Distribution

ويمكن ايجاد دالة البقاء لهذا التوزيع وكما يأتي [4] :

$$S(t) = 1 - F(t, \alpha, \beta)$$

$$S(t) = exp\left(-\left(\frac{t}{\alpha}\right)^{\beta}\right) \qquad \dots (6)$$

5-2 طريقة الامكان الاعظم Maximum Likelihood Estimation

تعد طريقة الامكان الاعظم من الطرق المهمة في التقدير لما تتميز به من خصائص وصفات كثيرة، فاذا افترضنا ان $t_1, t_2, t_3, \dots, t_n$ تمثل عينة عشوائية بحجم n مسحوبة من مجتمع له دالة كثافة احتمالية $f(t, \alpha, \beta)$ حيث ان α, β تمثل المعلمات المجهولة و t يمثل المتغير العشوائي اي ان [7] :

$$L = \prod_{i=1}^{n} f(t_i, \alpha, \beta) \quad \dots (7)$$

واذا كانت لدينا $F\left(t/\theta\right)$, $f\left(t/\theta\right)$ دالة الكثافة ودالة التوزيع على التوالي والبيانات من النوع الاول تكون دالة الامكان الاعظم بالشكل الاتي :

$$L\left(\theta, t\right) = C \prod_{i=1}^{m} f(t_i\ /\theta) \prod_{j=1}^{k} \left[1 - F\left(\frac{T_j}{\theta}\right)\right]^{S_j} \quad \dots (8)$$

حيث ان C تمثل توافيق ثابتة وبصورة عامة فأن $m \neq k$ ففي حاصل الضرب الاول يعمل المؤشر (i) على متابعة حالات الفشل أما في حاصل الضرب الثاني فهناك مؤشر اخر يختلف وهو (j) يعمل على المراقبة الثابتة بالاوقات T_j

وعند توظيف دالة الكثافة والتوزيع لتوزيع ويبل ذو معلمتين مع اهمال الثابت C واخذ اللوغارتم للحصول على دالة الامكان الاعظم وكالاتي :-

$$L(\alpha, \beta) = m[Ln(\beta) - \beta Ln(\alpha)] + (\beta - 1)\sum_{i=1}^{m} Ln(t_{(i)}) - \sum_{i=1}^{m}\left(\frac{t_{(i)}}{\alpha}\right)^{\beta} - \sum_{j=1}^{k} s_j \left(\frac{T_l}{\alpha}\right)^{\beta} \dots (9)$$

وبالاشتقاق الجزئي للمعلمتين β, α نحصل على الاتي

$$\frac{\partial L(\alpha, \beta)}{\partial \beta} = -\frac{m\beta}{\alpha} + \frac{\beta}{\alpha}\left[\sum_{i=1}^{m}\left(\frac{t_i}{\alpha}\right)^{\beta} + \sum_{j=1}^{k} s_j \left(\frac{T_j}{\alpha}\right)^{\beta}\right] = 0 \qquad \dots (10)$$

$$\frac{\partial L(\alpha,\beta)}{\partial \beta}=\frac{m}{\beta}-mLn(\alpha)+\sum_{i=1}^{m}Ln(t_i)-\left[\sum_{i=1}^{m}\left(\frac{t_i}{\alpha}\right)^{\beta}Ln\left(\frac{t_i}{\alpha}\right)+\sum_{j=1}^{k}s_j\left(\frac{T_j}{\alpha}\right)Ln\left(\frac{T_j}{\alpha}\right)\right]=0 \quad ...(11)$$

ومن المعادلة رقم (10) يمكن الحصول على المقدر $\hat{\alpha}$ وعلى النحو الاتي :-

$$\hat{\alpha}=\left[\frac{1}{m}\left[\sum_{i=1}^{m}t_i^{\hat{\beta}}+\sum_{j=1}^{k}s_j\,T_j^{\hat{\beta}}\right]\right]^{\frac{1}{\hat{\beta}}}...12$$

وبما ان المعادلة رقم (11) لاخطية يصعب حلها باستعمال الطرق الاعتيادية فسوف يتم حلها باستعمال الطرق العددية ومنها طريقة نيوتن رافسون للحصول على المقدر $\hat{\beta}$ وتعويضه في معادلة $\hat{\alpha}$ لاستخراج قيمتها [10] وبعد ايجاد المقدرات باستعمال الامكان الاعظم نقدر دالة البقاء من خلال الصيغة الاتية :

$$\hat{S}(t)=\exp\left(-\left(\frac{t}{\hat{\alpha}_{MLE}}\right)^{\hat{\beta}_{MLE}}\right) \quad ...(13)$$

6-2 طريقة العزوم Method of Moments

تتمثل هذه الطريقة جوهر المساواة بين عزم المجتمع وعزم العينة المناظره له وبالتالي سوف نحصل على عدد من المعادلات بالنسبة لمعلمات المجتمع وبحل هذه المعادلات نحصل على التقديرات المطلوبة كما تستعمل هذه الطريقة للحصول على التوزيع الاحتمالي لمتغير عشوائي ما من خلال الدالة المولدة للعزوم , و يتم الحصول على تقديرات المعلمات لتوزيع ويبل ذو المعلمتين بأستعمال متوسط العينة $\bar{t}$ وتباين العينة s^2 :

$$\bar{t}=\sum_{i=1}^{n}\frac{t_i}{n}$$

$$s^2=\sum_{i=1}^{n}\frac{(t_i-t)^2}{n-1}$$

وبمساواة عزم المجتمع مع عزم العينة وتباين المجتمع مع تباين العينة وعلى النحو الاتي :-

$$\mu=\bar{t}$$

$$\bar{t}=\alpha\Gamma\left(1+\frac{1}{\beta}\right)$$

$$s^2=\alpha^2\left[\Gamma\left(1+\frac{2}{\beta}\right)-\left(\Gamma\left(1+\frac{1}{\beta}\right)\right)^2\right]$$

ويمكن الحصول على المعلمة $\hat{\beta}$ من خلال قسمة التباين على مربع المتوسطات وكالاتي :-

$$=\frac{\alpha^2\left[\Gamma\left(1+\frac{2}{\beta}\right)-\left(\Gamma\left(1+\frac{1}{\beta}\right)\right)^2\right]}{\left[\alpha\Gamma\left(1+\frac{1}{\beta}\right)\right]^2}$$

وبتبسيط المعادلة اعلاه يتم الحصول على المقدرات كالاتي :-

$$\hat{\beta}_{mom}=\frac{\Gamma\left(1+\frac{2}{\hat{\beta}}\right)}{\Gamma^2\left(1+\frac{1}{\hat{\beta}}\right)}-1 \quad ...(14)$$

$$\hat{\alpha}_{mom}=\frac{\bar{t}}{\Gamma\left(1+\frac{1}{\hat{\beta}}\right)} \quad ...(15)$$

وبعد ايجاد المقدرات باستعمال طريقة العزوم نقدر دالة البقاء من خلال الصيغة الاتية [2] :

$$\hat{S}(t) = \exp\left(-\left(\frac{t}{\hat{\alpha}_{mom}}\right)^{\hat{\beta}_{mom}}\right) \quad \dots (16)$$

7-2 طريقة المربعات الصغرى Least Squares Method

تستعمل طريقة المربعات الصغرى (LSM) على نطاق واسع في العديد من المسائل العلمية الخاصة بعملية تقدير المعلمات للنماذج , فأذا كان لدينا النموذج الاتي [9] :

$$y_i = \beta_0 + \beta_1 X_1 \quad i = 1, 2, \dots, n \quad \dots (17)$$

فأن تقديرات المربعات الصغرى للمعلمات β_1, β_o يتم الحصول عليها بتصغير المقدار :-

$$Z(\beta_0, \beta_1) = \sum_{i=1}^{n} wi(y_i - \beta_0 - \beta_1 x_i)^2 \dots (18)$$

ولذلك فأن المقدرات الخاصة بـ β_1, β_o تأخذ الصيغة الاتية :-

$$\hat{\beta}_o = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i$$

وأن :

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \quad \dots (19)$$

ولتقدير معلمات توزيع ويبل ذو معلمتين β, α بأستعمال طريقة المربعات الصغرى في البدء يجب تحويل شكل التوزيع الى الشكل الخطي وعلى النحو الاتي :-

$$\ln\left(-\ln(1 - F(t))\right) = \beta \ln(t) - \beta \ln(\alpha)$$

وبجعل المعادلة اعلاه مشابهة لمعادلة الانحدار رقم $(2 - 21)$ وتكون كالأتي :-

$$y_i = \ln(-\ln(1 - F(t)))$$

وان قيم x_i تأخذ الصيغة الأتية :-

$$x_i = \ln(t_i)$$

وأن :-

$$\beta_0 = -\beta \ln(\alpha)$$

$$\beta_1 = \beta$$

ولهذا اذا كان لدينا عينة عشوائية $t_1, t_2, \dots \dots, t_n$ مسحوبة من توزيع ويبل بأحصاءات رتبية

$$T_{(1)} < T_{(2)} < \cdots < T_{(n)}$$

ونفرض :

$$t_{(1)} < t_{(2)} < \cdots < t_{(n)}$$

مشاهدات مرتبة فأن دالة التوزيع لها يتم تقديرها بأستعمال رتبة المتوسط وعلى النحو الاتي :-

$$\hat{F}(t_{(i)}) = \frac{i}{n+1}$$

حيث ان i يشير الى i^{th} اصغر قيمة لـ $T_{(1)} < T_{(2)} < \cdots < T_{(n)}$ و أن $(i = 1, 2, \dots n)$ ولذا تكون تقديرات β_1, β_o بالشكل التالي :

$$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} \ln\left(-\ln\left(1 - \hat{F}(t_{(i)})\right)\right) - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} \ln(t_{(i)})$$

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} \ln(t_{(i)}) \ln\left(-\ln\left(1-\hat{F}(t_{(i)})\right)\right) - \sum_{i=1}^{n} \ln(t_{(i)}) \sum_{i=1}^{n} \ln\left(-\ln\left(1-\hat{F}(t_{(i)})\right)\right)}{n\sum_{i=1}^{n} \ln^2(t_{(i)}) - \left(\sum_{i=1}^{n} (t_{(i)})\right)^2} \dots (20)$$

وبما أن

$$\beta_1 = \beta$$

لذلك تكون

$$\widehat{\beta} = \widehat{\beta}_1$$

وأن :

$$\beta_0 = -\beta \ln(\alpha)$$

يعني ذلك ان :

$$\alpha = \exp\left(-\left(\frac{\beta_0}{\beta}\right)\right)$$

لذلك تكون :

$$\widehat{\alpha} = \exp\left(-\left(\frac{\left(\sum_{i=1}^{n} \ln\left(-\ln(1-\widehat{F}(t_{(i)}))\right)\right) - \widehat{\beta}\sum_{i=1}^{n} \ln(t_{(i)})}{n\widehat{\beta}}\right)\right) \quad \dots (21)$$

وبعد ايجاد المقدرات باستعمال طريقة المربعات الصغرى نقدر دالة البقاء من خلال الصيغة الاتية :

$$\widehat{S}(t) = \exp\left(-\left(\frac{t}{\widehat{\alpha}_{LS}}\right)^{\widehat{\beta}_{LS}}\right) \quad \dots (22)$$

6-2 الخوارزمية الجينية Genetic Algorithm

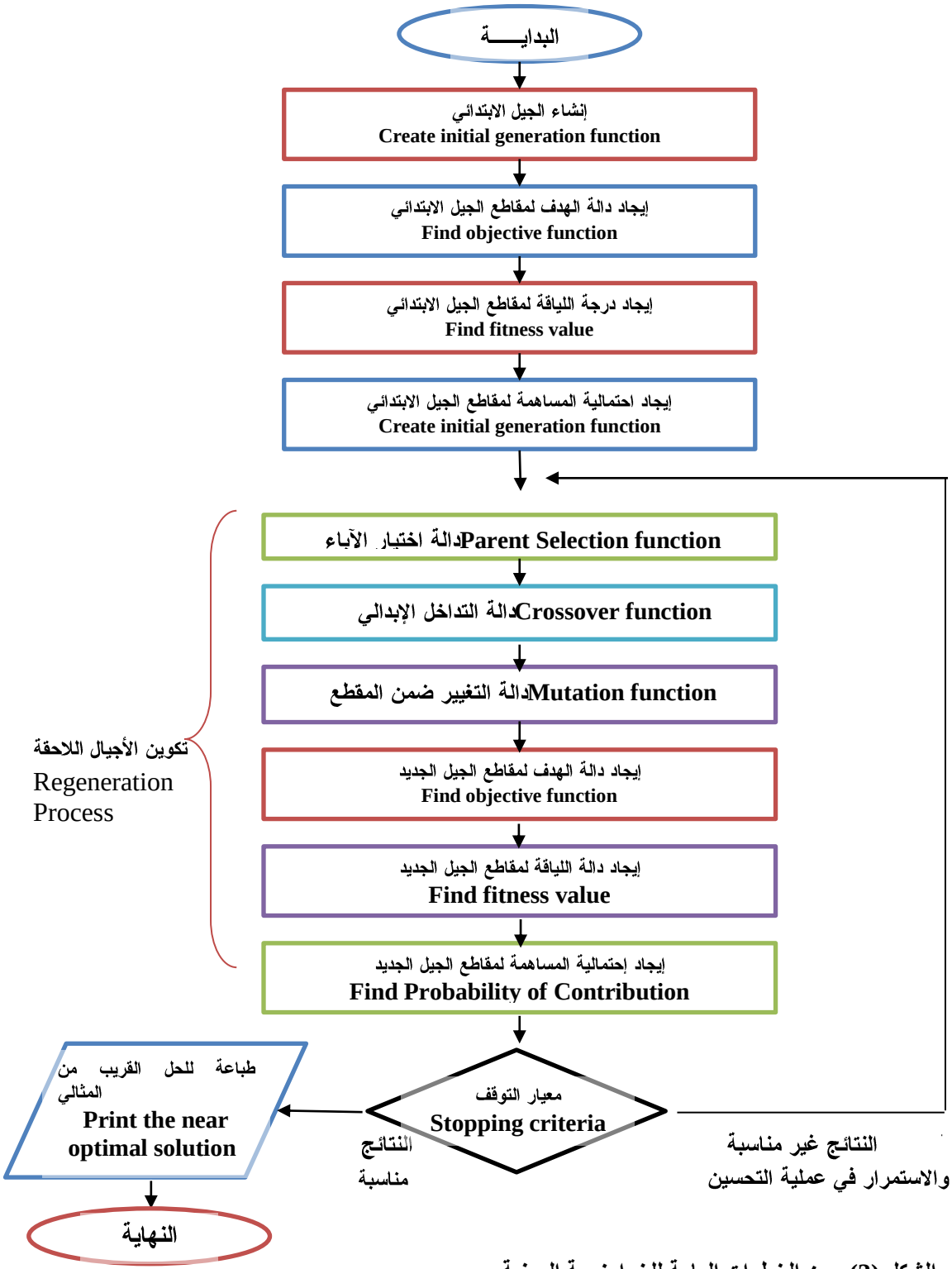
إن الكائنات الحية تتناسب مع الظروف المحيطة بها وتحاول أن تتكيف مع تغير الظروف عبر الأجيال وفي حالة عدم تمكنها من التكيف والتطور فإنها تنقرض. فالكائنات ذات الصفات القوية هي التي تسود، بينما تضمحل وتموت الكائنات ذات الصفات الضعيفة، كما تعد الطفرة الوراثية التي تحدث بنسبة قليلة جداً من العوامل المهمة التي تساعد على تطوير الصفات الوراثية المنقولة عبر الجينات، ومن هذه الأفكار تم اشتقاق طريقة طبقت على المسائل المتعلقة بالحاسوب وهي الخوارزمية الجينية (Genetic Algorithm GA).

اختصرت الخوارزمية الجينية الكثير من الجهد والزمن المطلوبين لدى مصممي البرامج والانظمة من خلال إيجادها خوارزمية عامة يُعتمد.عليها في حل مختلف أنواع المسائل، بدلاً من بناء خوارزمية خاصة لكل مسألة، مع مراعاة التغيرات.اللازمة التي تتناسب مع خصوصية كل مسألة من حيث حجم ونوع البيانات. المستخدمة وطبيعة دالة الهدف والقيود لكل مسألة.

تكمن فلسفة الخوارزمية الجينية على توليد عدد كبير من الحلول المتوفرة لمشكلة معينة ومن ثم يتم تقييم كل حل من هذه الحلول وصولا الى الحل الافضل وتكون فرصته اكبر لتوليد حل اخر وبتكرار هذه العملية تتطور نوعية الحلول الموجودة وتصل او تقترب من الحل الامثل[8] :

1-6-2 المخطط العام للخوارزمية الجينية General Diagram of the GA

تتضمن الخوارزمية الجينية عدداً من الخطوات الأساسية، هذه الخطوات تكون مترابطة بعضها مع البعض الآخر، ولا يمكن تطبيق هذه الخوارزمية على أية مسألة ما لم تطبق جميع هذه الخطوات وإلاّ تفقد الخوارزمية الجينية قيمتها وفائدتها في إيجاد أو تحسين الحل والشكل (3) يوضح المخطط العام للخوارزمية الجينية. [11]



الشكل (3) يبين الخطوات العامة للخوارزمية الجينية

2-6 -2 تطبيق الخوارزمية الجينية في تقدير معلمات توزيع ويبل

Application of Genetic Algorithm to Estimate Weibull Parameters Distribution

نطبق خطوات الخوارزمية الجينية في معادلة دالة الهدف لكل طريقة لإيجاد تقديرات معلمات توزيع ويبل ذو معلمتين ثم تقدير دالة البقاء وفقاً لما يأتي : [13]

1 - البداية في البدء يتم تكوين الكروموسوم من خلال قيم المعلمات α , β التي تكون جينات الكروموسوم

2 - انشاء جيل اولي عن طريق ايجاد قيم اولية للجينات يتم توليدها بشكل عشوائي

3 - في دالة الهدف يتم تقييم الكروموسوم واختياره الذي يمتلك دالة هدف صغيرة التي تقابل اكبر احتمال ثم ايجاد دالة التقييم اله من خلال المعادلة الاتية :-

$$\text{fitness function} = \frac{1}{1 + \text{objective function}} \quad \dots(23)$$

fitness function: تمثل دالة التقييم

$\text{objective function}$: تمثل دالة الهدف

ويمكن حساب احتمالية دالة التقييم (افضل تقييم) من خلال الصيغة الاتية :-

$$C_{(i)} = \frac{f_{(i)}}{\sum_{i=1}^{n} f_{(i)}} \quad \dots(24)$$

$C_{(i)}$: احتمالية الكروموسوم i

$f_{(i)}$: دالة التقيم للكروموسوم i ، n: حجم المشاهدات

وباستعمال احد معايير الاختيار عجلة الروليت بتوليد رقم عشوائي $r_{(c)}$ بالفترة [0,1] ومقارنته مع الكروموسوم الاول $c_{(1)}$ فاذا كان $r_{(c)} < c_{(1)}$ سوف يتم اختيار الكروموسوم الاول وهكذا يتم في كل مرة تحديد كروموسوم واحد للمجتمع الجديد بالاعتماد على دالة التقييم

4 - تهجين الكروموسومات المختارة عن طريق التزاوج بيم كروموسومين وبتطبيق احد معايير التهجين وهو التهجين المنظم بالاعتماد على احتمالية التهجين P_c وهذه القيمة الاحتمالية يحددها الباحث حيث تكون بالمدى $P_c \geq 0.25$ وتقارن هذه القيمة مع قيمة الجينات للكروموسومين (الاباء) لتكوين الجيل الجديد (الابناء) ويحدث التبادل عندما تكون قيمة الجين اكبر او تساوي P_c المحددة .

5 - عملية الطفرة التي تعتمد على القيمة الاحتمالية P_m للمعلمات وهذه القيمة الاحتمالية يتم حسابها من خلال الصيغة الاتية :

$$P_m = \begin{cases} 0.09 - \dfrac{\text{Fitvalue} - f_{mean}}{f_{max}} & \text{if Fitvalue} > f_{mean} \quad \dots(25) \\ 0.09 & \text{Otherwise} \end{cases}$$

حيث ان :

Fitvalue : تمثل قيمة دالة التقييم

f_{mean}: تمثل متوسط المجتمع

f_{max}: تمثل اكبر قيمة للمجتمع

وباستبدال جينات اختيرت عشوائياً مع قيمة جديدة نحصل عليها بشكل عشوائي ايضاً من خلال الصيغة الاتية :

مجموع الجينات = (عدد الجينات في الكروموسوم) × (مجموع المجتمع)

6 - يتم الرجوع الى الخطوة الثالثة الى ان يتحقق معيار التوقف الإنجازية المنفصلة

7 - تقيم المعلمات بالاعتماد على قيمة دالة الهدف لحساب معلمات توزيع ويبل

3 - المحاكاة Simulation

1-3 المـقـدمـة Introduction

سوف نتناول في هذه الفقرة المقارنة بين طرائق التقدير الكلاسيكية وطرائق التقدير للخوارزمية الجينية لتقدير معلمات توزيع ويبل ودالة البقاء ويتم ذلك من خلال تصميم تجربة المحاكاة حيث سيتم محاكاة عدد كبير جدا من الحالات التي يمكن مواجهتها في الواقع العملي بهدف الوصول لنتائج أكثر شمولية.

ولبيان أفضلية طرائق معينة يؤدي إلى اللجوء إلى تجارب المحاكاة اذ أنها تتيح للباحث اختيار حجوم عينات مختلفة مع حالات متنوعة للمعلمات الخاصة بالنموذج وتكرار التجربة مرات عدة ولجميع الطرائق المستعملة بهدف التوصل إلى الطريقة المثلى. وكان العامل الأهم للاستعمال الواسع لأساليب المحاكاة هو تطور الحاسوب في العقود الأخيرة.

2-3 توليد الأعداد العشوائية Generate Random Number

سوف يتم استعمال طريقة التحويل المعكوس لتوليد قيم متغير تتبع توزيع ويبل ذي معلمتين وذلك على النحو الاتي:

توليد الأعداد العشوائية التي تتبع التوزيع المنتظم (U_i) على المدة $(0,1)$

يتم مساواة قيم U_i بدالة التوزيع وكما في المعادلة الاتية :

$$U_i = F(t_i) \quad \dots (26)$$

تحويل العدد العشوائي المنتظم بطريقة المعكوس وكما مبين في المعادلة الآتية :

$$t_i = F^{-1}(U_i) \quad \dots (27)$$

للحصول على متغير عشوائي يصف الأنموذج تحت التجربة .

ولتحويل الاعداد العشوائية الى بيانات تتبع توزيع ويبل ذي المعلمتين بأسلوب رياضي احصائي وكالآتي:

$$U_i = F(t_i)$$

$$U_i = 1 - \exp\left(-\left(\frac{t_i}{\alpha}\right)^{\beta}\right)$$

$$1 - U_i = \exp\left(-\left(\frac{t_i}{\alpha}\right)^{\beta}\right)$$

وبأخذ اللوغارتم الطبيعي للطرفين للمعادلة اعلاه نحصل على ما يلي:

$$\log(1 - U_i) = -\left(\frac{t_i}{\alpha}\right)^{\beta}$$

$$\log(1 - U_i) = -\frac{t_i^{\beta}}{\alpha^{\beta}}$$

$$t_i^{\beta} = \log(1 - U_i)\,(-\alpha^{\beta})$$

وبضرب المعادلة اعلاه بالمقدار $\left(1^{\frac{1}{\beta}}\right)$ نحصل على المعادلة الاتية :

$$t_i = -\alpha\,(\log(1 - U_i))^{\frac{1}{\beta}} \quad \dots (28)$$

3 - 3 مراحل تطبيق تجارب المحاكاة Stages of application of simulation experiments

لقد تضمنت تجارب المحاكاة في تطبيقات اساليب تقدير دالة البقاء لهذه الدراسة المراحل الأتية:

بالنسبة للمعلمات والنماذج المفترضة فكانت كالأتي:

وقد أُختيرت قيم افتراضية لمعلمة الشكل ومعلمة القياس وقد تم استعمال هذه القيم من البيانات الحقيقية بعد تقديرها باستعمال طريقة الامكان الاعظم وهذه القيم هي كما مبينة بالجدول (1)

جدول (1) يبين القيم الافتراضية للنماذج المستعملة في المحاكاة

النموذج	β	α
الاول	1.2302	92.938
الثاني	1.50	90.20
الثالث	1.20	95.20

اختيار حجم العينة n :

فقد أختيرت اربعة حجوم مختلفة للعينة بشكل يتناسب مع معرفة مدى تأثير حجم العينة لكل مستوى من مستويات التأثير العشوائي (t_i) والمستعملة في هذه الدراسة هي (60,100، 40، 20) إذ يمثل (n=20) حجم العينة الصغيرة و(n=40,60) حجم العينة المتوسطة و(n=100) حجم العينة الكبيرة.

مرحلة المقارنة :

وهنا المرحلة الأخيرة ، وهي المقارنة بين طرائق التقدير ، اذ تم استعمال متوسط مربعات الخطأ(MSE) حيث تم تكرار التجربة 1000 مرة وصيغته كما يلي :

$$MSE\left(\hat{S}(t)\right) = \frac{1}{R}\sum_{i=1}^{R}(\hat{S}(t) - S(t))^2 \quad \dots (29)$$

R . . . . ، $i = 1,$ حيث أن R : تمثل عدد المكررات (Replications) لكل تجربة .

4-3 نتائج المحاكاة Simulation results

في هذه الفقرة سيتم عرض وتحليل نتائج المحاكاة لتقدير دالة البقاء بالطرائق التقلدية والجينية وفيما يأتي النتائج الموضحة في الجداول التي سيتم تحليلها حسب تسلسل الجداول وكما يأتي :

جدول (2) يوضح قيم متوسط مربعات الخطأ للانموذج الاول

Sample Size	Methods	Classic	Genetic	Best
n=20	MLE	0.8564987113299	0.0491282939827502	MLE-GA
	MOM	0.0421233282350875	0.000694552069589193	MOM-GA
	LS	0.00643403240772745	0.000257816579562108	LS-GA
Best		LS	LS-GA	LS-GA
n=40	MLE	0.763966381821285	0.0249082213903305	MLE-GA
	MOM	0.0218450429992282	0.000151679530006394	MOM-GA
	LS	0.00333087537401691	6.27764098640677e-05	LS-GA
Best		LS	LS-GA	LS-GA
n=60	MLE	0.616329629320668	0.0164538279113592	MLE-GA
	MOM	0.0143894973408813	5.00471182607555e-05	MOM-GA
	LS	0.00214311978482119	1.83833946889059e-05	LS-GA
Best		LS	LS-GA	LS-GA
n=100	MLE	0.465117711934094	0.00991452695601720	MLE-GA
	MOM	0.00630697384572508	1.15105459826550e-05	MOM-GA
	LS	0.00129425130979425	2.80147606868204e-06	LS-GA
Best		LS	LS-GA	LS-GA

جدول (3) يوضح قيم متوسط مربعات الخطأ للانموذج الثاني

Sample Size	Methods	Classic	Genetic	Best
n=20	MLE	1.11198195203440	0.000503278094996516	MLE-GA
	MOM	0.0400027498393681	0.000777219892194066	MOM-GA
	LS	0.00643373028193891	0.000264447372241147	LS-GA
Best		LS	LS-GA	LS-GA
n=40	MLE	1.02513933226456	0.000489014535777666	MLE-GA
	MOM	0.0190036899403078	0.000176351397725658	MOM-GA
	LS	0.00333030408681597	7.09128420033473e-05	LS-GA
Best		LS	LS-GA	LS-GA
n=60	MLE	0.867637539229784	0.000465346364603931	MLE-GA
	MOM	0.0123573174397349	6.17436919871592e-05	MOM-GA
	LS	0.00214418982538107	2.27591652183501e-05	LS-GA
Best		LS	LS-GA	LS-GA
n=100	MLE	0.679128742394182	0.000432785128287496	MLE-GA
	MOM	0.00431793394771540	1.46545267553690e-05	MOM-GA
	LS	0.00129482645139626	3.97802704681917e-06	LS-GA
Best		LS	LS-GA	LS-GA

جدول (4) يوضح قيم متوسط مربعات الخطأ للانموذج الثالث

Sample Size	Methods	Classic	Genetic	Best
n=20	MLE	0.829010800207235	0.0491324324328685	MLE-GA
	MOM	0.0424932486150453	0.000690860894816389	MOM-GA
	LS	0.00643479836656889	0.000256876298775211	LS-GA
Best		LS	LS-GA	LS-GA
n=40	MLE	0.737928177002768	0.0249084954206752	MLE-GA
	MOM	0.0220240001889099	0.000150969304654875	MOM-GA
	LS	0.00333183020509627	6.20970576256515e-05	LS-GA
Best		LS	LS-GA	LS-GA
n=60	MLE	0.594290548425976	0.0164538331366322	MLE-GA
	MOM	0.0145030159057076	4.98870825618807e-05	MOM-GA
	LS	0.00214464987424608	1.81554889833584e-05	LS-GA
Best		LS	LS-GA	LS-GA
n=100	MLE	0.449614447398950	0.00991473815878032	MLE-GA
	MOM	0.00636170908480336	1.16137283595613e-05	MOM-GA
	LS	0.00129561778581450	2.79721208147540e-06	LS-GA
Best		LS	GA-LS	LS_GA

5-3 مناقشة النتائج Discuss The Results

نلاحظ من الجداول اعلاه :

1 - عند احجام العينة (n= 20 , 40 , 60 ,100) بالنسبة للقيم الافتراضية للمعلمات ولكل النماذج ان طريقة الامكان الاعظم المسندة الى الخوارزمية الجينية MLE-GA على طريقة الامكان الاعظم الكلاسيكية MLE وطريقة العزوم المسندة الى الخوارزمية الجينية MOM-GA على طريقة العزوم الكلاسيكية MOM و طريقة المربعات الصغرى المسندة الى الخوارزمية الجينية LS-GA على طريقة المربعات الصغرى الكلاسيكية LS

2 – كانت طريقة المربعات الصغرى المسندة الى الخوارزمية الجينية LS – GA هي الافضل في المرتبة الاولى على كافة طرائق التقدير المسندة للخوارزمية الجينية من حيث عدد مرات امتلاكها اقل (MSE) لمقدرات انموذج توزيع ويبل ذو معلمتين وذلك لمختلف حجوم العينات للنماذج المفترضة والقيم الافتراضية للمعلمات

6 - الاستنتاجات Conclusions

من خلال ما تم عرضه في الجانب التجريبي تم التوصل الى ان طرائق التقدير بالاعتماد على الخوارزمية الجينية افضل من طرائق التقدير التقليدية لتقدير دالة البقاء وبالتالي تعطي احتمال بقاء ادق وذلك لحجوم العينات كافة ولكل النماذج و عندما تم توظيف الخوارزمية الجينية في طرائق التقدير الكلاسيكية فتبين ان افضل طريقة لتقدير دالة البقاء هي طريقة المربعات الصغرى المسندة على الخوارزمية الجينية LS – GA عند حجم العينة ($n = 20, 40, 60, 100$) ولكل النماذج و اظهرت نتائج تجارب المحاكاة ان قيم المقياس الاحصائي متوسط مربعات الخطأ (MSE) تتناقص بزيادة حجم العينة وهذا ينسجم من النظرية الاحصائية .

7 - التوصيات Recommendations

من خلال الاستنتاجات واجراءات البحث تم التوصل الى استعمال طريقة المربعات الصغرى بتوظيف الخوارزمية الجينية لتقدير دالة البقاء لمرضى بهدف الحصول على احتمال بقاء على قيد الحياة دقيق واخذ نماذج مختلفة او توزيعات مختلفة عن توزيع ويبل او استعمال الخوارزمية الجينية في تقدير دالة البقاء وخصوصا في حالة الامراض ذات الخطورة العالية لتزويد الادارة الصحية بالتنبؤ بالفترة الزمينة لبقاء المرضى على قيد الحياة والتي بدورها تقوم بأخذ التدابير لذلك كما نوصي باستعمال اسلوب اخر مثل الشبكات العصبية الاصطناعية لتقدير دالة البقاء

8 - المصادر References

[1]- **Akma, N., Fauzi, M., Elfaki, F. A. M., & Ali, Y. (2015). " Some Method On Survival Analysis Via Weibull Model In the Present of Partly Interval Censored " .A Short Review. International Journal of Computer Science and Network Security (IJCSNS), 15(4), 48**

[2]- **Al-wakeel, A. A. S. Razali, AM, & Mahdi, AA, (2016) "Estimation accuracy of Weibull distribution parameters" , Journal of Applied Sciences Research, vol. 5, no. 7, pp. 790-795.2009**

[3]-**Collett, D. (2015). "Modelling survival data in medical research", third edition. CRC press , Taylor & Francis Group , International Standard Book Number-13: 978-1-4987-3169-0**

[4]-**Ellis, W. C., & Rao Tummala, V. M. (1986). " Minimum Expected Loss Estimators of the Shape & Scale Parameters of the Weibull Distribution " . IEEE Transactions on Reliability,35(2),212–215.**

[5]-**Gasmi, S., & Berzig, M. (2011). " Parameters estimation of the modified Weibull distribution based on type I censored samples ". Applied Mathematical Sciences, 5(57–60), 2899–2917.**

[6]-**Klein, J. P., & Moeschberger, M. L. (2003). " Survival analysis: Techniques for censored and truncated data " . In Pharmaceutical Statistics.**

[7]-**Ng, H. K. T., Luo, L., Hu, Y., & Duan, F. (2012). "Parameter estimation of three-parameter Weibull distribution based on progressively Type-II censored samples " . Journal of Statistical Computation and Simulation, 82(11), 1661–1678.**

[8]-**Mahdavi, I., Paydar, M. M., Solimanpur, M., & Heidarzade, A. (2009). "Genetic algorithm approach for solving a cell formation problem in cellular manufacturing". Expert Systems with Applications, 36(3), 6598-6604.**

[9]-**Osatohanmwen, P., Oyegue, F. O., Osemwenkhae, J. E., & Ekhosuehi, N. (2017). "An appraisal on some methods for estimating the 2-parameter weibull distribution with application to wind speeds sample " . Sri Lankan Journal of Applied Statistics, 18(3).**

[10]- **Rinne, H. (2009) , " The Weibull distribution " . Justus-Liebig-University, Giessen, Germany, published by Chapman & Hall/CRC,ISBN: 13: 978-1-4200-8743-7, 2009.**

[11]- **S.N.Sivanandam, S.N.D. 2008. " Introduction to genetic algorithms " (pp. 15-37). Springer, Berlin, Heidelberg**

[12]- **Wen, C.-J., Liu, X. and Cheng, X. 2017. " Parameter Evaluation of 3-parameter Weibull Distribution based on Adaptive Genetic Algorithm " . 138: 426–431.**

[13]- **ZHAO, GUOLIN, M. A. (2008). " Nonparametric and Parametric Survival Analysis of Censored Data with Possible Violation of Method Assumptions " . Greensboro, Vol 49 , 73-69.**

Estimate The Survival Function By Using The Genetic Algorithm

Anwar Taher Abdel Hadia **Sabah Manfi Redha**

anwertaher17@gmail.com drsabah@coadec.uobaghdad.edu.iq

Received:11/3/2020 Accepted :16/8/2020 Published :October / 2020

Abstract :

Survival analysis is the analysis of data that are in the form of times from the origin of time until the occurrence of the end event, and in medical research the origin of time is the date of registration of the individual or the patient in a study such as clinical trials to compare two types of medicine or more if the end point It is the death of the patient or the disappearance of the individual. The data resulting from this process is called survival times. But if the end is not death, the resulting data is called time data until the event. That is, survival analysis is one of the statistical steps and procedures for analyzing data when the adopted variable is time to event and time. It could be days, weeks, months, or years from the start of the term registration until the event.

This research is concerned with the question of estimating the survival function of observational data using one of the most important artificial intelligence algorithms which is the genetic algorithm and that In order to obtain optimum estimates for weibull distribution parameters, this in turn is reflected in the estimation of survival function, whereby the genetic algorithm is employed in the maximum likelihood method , moment method , the least squares method and the modified weighted least squares method. And for the capabilities of more efficient than traditional methods, and then will be a comparison between the roads depending on the experimental side is evaluated the best way depending on mean square error criterion of survival function, it will also be applied methods on the fact that data for patients with lung cancer and bronchitis.

The study found that the best way to estimate the weibull distribution parameters and the survival function produced by the experimental side is the hybrid method of the least squares using the genetic algorithm.

keywords : Survival function , Weibull distribution , Genetic algorithm , Maximum Likelihood Estimation , Method of moment , Least squares method , Censoring Data