1-1-2006

Statistical modeling via empirical recurrence rate

Hui Wang
University of Nevada, Las Vegas

STATISTICAL MODELING VIA EMPIRICAL RECURRENCE RATE

by

Hui Wang

Bachelor of Science
Beijing Institute of Machinery, Beijing, China
July 1994

A thesis submitted in partial fulfillment
of the requirements for the

Master of Science in Mathematical Sciences
Department of Mathematical Sciences
College of Sciences

Graduate College
University of Nevada, Las Vegas
December 2006

UNLV
UNIVERSITY OF NEVADA LAS VEGAS

Thesis Approval

The Graduate College
University of Nevada, Las Vegas

November 17, 2006

The Thesis prepared by

Hui Wang

Entitled

Statistical Modeling Via Empirical Recurrence Rate

is approved in partial fulfillment of the requirements for the degree of

Master of Science in Mathematical Sciences

Examination Committee Chair

Dean of the Graduate College

Examination Committee Member

Examination Committee Member

Graduate College Faculty Representative

ABSTRACT

Statistical Modeling via Empirical Recurrence Rate

by

Hui Wang

Dr. Chih-Hsiang Ho, Examination Committee Chair
Professor of Mathematical Sciences
University of Nevada, Las Vegas

A key parameter, most sought after by the modelers of reliability, is the failure rate of a targeted repairable system. Popular modeling techniques based on a point process such as the Power-law process often are handicapped by the requirement of a monotonic failure rate. In this thesis, we show the potential of building a linking bridge between the traditional homogeneous and nonhomogeneous Poisson processes and the classical time series via a sequence of the empirical recurrence rates, calculated at equally spaced intervals of time. The distinctive signature, marking the unique failure pattern of a repairable system, is displayed with an empirical recurrence rate time-plot, referred as the "fingerprint" or an "ERR-plot" of a targeted system. A major strength of our approach is that we present an interesting extension of advanced time series analysis techniques into the domain of data exploration of point processes, including but not limited to the events associated with repairable systems or natural phenomena (earthquakes and volcanic eruptions), and make new and innovative use of the well-known ARIMA method

possible for modeling the recurrence rate of such events ranging from constant recurrence rate to those show cyclic trends.

ARIMA time series modeling techniques are well developed. Therefore, the scope of our study is to investigate the merits of the transformation in terms of the diagnostics on the basic plots and some tests of goodness-of-fit via pseudo and real data.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to express my sincerely gratitude to my advisor, Dr. Ho, for his warm encouragement, patience, time, and dedication to keep me on the right track throughout this undertaking. His excellence in both research and teaching will always be a great example to me.

My deeply indebtedness also give to respectable committee members, Dr. Ananda, Dr. Catlin and Dr. Qian, for their positive inputs and mentoring during my graduate studies.

I would also like to thank my wife, Sue Ho, for her care and mental support that make me to achieve my goal. Last but not least, I would like to thank my families for their love and support throughout my education.

CHAPTER 1

INTRODUCTION

1.1 Reliability and the Weibull Distribution

Reliability plays a key role in developing quality products and in enhancing competitiveness. For most products, customers see reliability as one of the most important quality characteristics. In the last several decades, there has been much research on the theory and applications of reliability. Most of this literature covers the reliability of repairable systems. A repairable system is a system that, when a failure occurs, can be restored to an operating condition by some repair process other than replacement of the entire system. Many real world systems, such as automobiles, airplanes, computers, and air conditioners, are repairable systems.

The lifetime of a unit such as a component or system can be represented as nonnegative random variable T. Unless otherwise indicated, we will also assume that T has a continuous distribution. The distribution of such a random variable is called life-testing model, and such models are considered in the area of reliability. The probability that a unit survives beyond time t_0 is called the reliability at time t_0, and the reliability function, is defined as

$$R(t_0) = P[T > t_0] = 1 - F(t_0)$$

In biomedical applications the term "survival function" is also used. Life –testing model can be characterized in terms of a number of different concepts. The hazard function (HF) is defined by

$$h(t) = \frac{f(t)}{1 - F(t)}$$

In actuarial science $h(t)$ is known as the "force of mortality," and in extreme-value theory $h(t)$ is called the "intensity function." This concept is often referred to as the "failure rate." The Weibull distribution is related to the power law process, a commonly used model for repairable systems.

Definition 1 The Weibull distribution has survival function

$$S(t) = \exp\left[-\left(\frac{t}{\alpha}\right)^{\beta}\right], \quad t > 0$$

If T is a random variable with this c.d.f., Then we will write $T \sim WEI(\theta, \beta)$. The c.d.f., p.d.f., and hazard functions are therefore given as follows:

$$F(t) = 1 - S(t) = 1 - \exp\left[-\left(\frac{t}{\theta}\right)^{\beta}\right], \quad t > 0$$

$$f(t) = F'(t) = \frac{\beta}{\theta}\left(\frac{t}{\theta}\right)^{\beta-1} \exp\left[-\left(\frac{t}{\theta}\right)^{\beta}\right], \quad t > 0$$

$$h(t) = \frac{f(t)}{S(t)} = \frac{\frac{\beta}{\theta}\left(\frac{t}{\theta}\right)^{\beta-1} \exp\left[-\left(\frac{t}{\theta}\right)^{\beta}\right]}{\exp\left[-\left(\frac{t}{\theta}\right)^{\beta}\right]} = \frac{\beta}{\theta}\left(\frac{t}{\theta}\right)^{\beta-1}, \quad t > 0$$

The hazard function h is increasing when $\beta > 1$, and decreasing when $\beta < 1$. When $\beta = 1$, the hazard function is the constant function $h(t) = 1/\theta$. Thus, the exponential distribution is a special case of the Weibull distribution that occurs when $\beta = 1$.

Definition 2 Intensity Function (Rigdon and Basu, 2000)

Let $X(t)$ denote the number of occurrences in the interval $[0,t]$, and $P_n(t) = P[$n occurrences in an interval $(0,t)]$. The intensity function of a point process is

$$\lambda(t) = \lim_{h \to 0} \frac{P[X(t+h) - X(t) \geq 1]}{h}.$$

Roughly speaking, the intensity function is the probability of failure in a small interval divided by the length of the interval. Thus, there will be many failures over intervals on which $\lambda(t)$ is large, and fewer failure over intervals on which $\lambda(t)$ is small. It is instructive to compare the definitions of the hazard function and intensity function. The hazard function is the limit of a conditional probability that the one and only one failure will occur in a small interval, divided by the length of the interval. This probability is conditioned on survival to the beginning of the interval. The intensity function is the unconditional probability of a failure in a small interval divided by the length of the interval (Rigdon and Basu, 2000).

1.2 Homogeneous Poisson Process (HPP)

Let $X(t)$ denote the number of occurrences in the interval $[0,t]$, and $P_n(t) = P[$n occurrences in an interval $(0,t)]$. Consider the following properties:

1. $X(0) = 0$

2. $P[X(t+h) - X(t) = n | X(s) = m] = P[X(t+h) - X(t) = n]$ for all $0 \leq s \leq t,\ h > 0$

3. $P[X(t+h)-X(t)=1]=\lambda h+o(h)$ for some constant $\lambda > 0$

4. $P[X(t+h)-X(t)\geq 2]=o(h)$

Based on the above properties, we have $X(t) \sim POI(\lambda t)$, where $\mu = E(X(t)) = \lambda t$. The proportionality constant λ reflects the rate of occurrence or intensity of the Poisson process. Because λ is assumed constant over t, the process is referred to as a homogeneous Poisson process (HPP), and the model $X \sim POI(\lambda)$ is applicable for any interval of length t, $[s, s+t]$, with $\mu = \lambda t$. In terms of a repairable system, this implies that the system is neither improving nor wearing out with age, but rather is maintaining a constant intensity of failure.

1.3 Nonhomogeneous Poisson Process (NHPP)

$\{X(t),\ t \geq 0\}$ is said to be a nonhomogeneous Poisson process with intensity function $\lambda(t)$ if:

1. $X(0)=0$

2. $\{X(t), t \geq 0\}$ has independent increments.

3. $P[X(t+h)-X(t)=1]=\lambda(t)h+o(h)$

4. $P[X(t+h)-X(t)\geq 2]=o(h)$

Then $P(X(t)=n)=\dfrac{e^{-\lambda(t)}(\lambda(t))^n}{n!}$, $n \geq 0$, where $\lambda(t)=\int_0^t \lambda(s)ds$.

Then cumulative distribution function for the time to first occurrence, T_1, now becomes

$$F_1(t)=1-\exp[-\lambda(t)]$$

$$X(t) \sim POI[\lambda(t)]$$

Unlike the homogeneous Poisson process failure probability, the intensity, $\lambda(t)$, may be depend on the age t of the system. $\lambda(t)$ would be decreasing during debugging, $\lambda(t)$ would be constant over the system useful life, and would be increasing during the wear-out phase of the system. In some case when the intensity function, $\lambda(t)$, is constant for all t, the nonhomogeneous Poisson process reduced to the homogeneous Poisson process.

A nonhomogeneous Poisson process with intensity $\lambda(t) = (\beta/\theta)(t/\theta)^{\beta-1}$ for $\theta, \beta > 0$, called a Weibull process or power-law process. The name Weibull process derives primarily from the resemblance of the intensity function of the process to the hazard function of a Weibull distribution. In a Weibull process, the time to first occurrence T_1, follows a Weibull density $WEI(\theta, \beta)$. A Weibull process is appropriate for three types of systems: increasing recurrence rate ($\beta > 1$), decreasing recurrence rate ($\beta < 1$), and constant recurrence rate ($\beta = 1$).

CHAPTER 2

STATISTICAL INFERENCE FOR A POWER-LAW PROCESS

2.1 Preliminaries

A nonhomogeneous Poisson process is often suggested as an appropriate model when a system whose rate varies over time. If the process is waning or developing, the rate λ should be a monotonically decreasing or increasing function of t. The nohomogeneous Poisson process (NHPP) has a mean value function denoted by $\mu(t|\Theta)$, where Θ is a vector of parameters. The intensity function $\lambda(t|\Theta)$ is described as follows:

$$\lambda(t|\Theta) = \frac{d}{dt}\mu(t|\Theta).$$

Arguments are presented in Bain and Engelhard (1980, 1991), Crow (1974, 1982), and Ho (1993, 1998) for the choice of $\Theta = (\theta, \beta)$ and

$$\lambda(t|\theta, \beta) = (\beta/\theta)(t/\theta)^{\beta-1}.$$

The underlying point process is called a power-law process, which has proved versatile in the reliability studies of repairable systems. Note that

$$\mu(t|\theta, \beta) = (t/\theta)^{\beta}.$$

Therefore, the β parameter affects how the system deteriorates or improves over time. If $\beta > 1$, then the intensity function $\lambda(t)$ is increasing, and the failure tend to occur more frequently. If $\beta < 1$, then $\lambda(t)$ is decreasing, and the system is improving. Finally, if

$\beta=1$, then the power law process reduces to the simpler homogeneous Poisson process with intensity $1/\theta$. The θ parameter is a scale parameter. There are several reasons why the power-law process is a widely used model for repairable systems. The key reason for the popularity of the power-law process is that statistical inference procedures are well developed.

2.2 Statistical Inferences

Suppose that a repairable system is observed until n failures occur, so we observe the failure time $0 < t_1 < t_2 < \ldots < t_n$, so the joint p.d.f. of a failure truncated NHPP as

$$f(t_1,t_2,\ldots,t_n)=\left(\prod_{i=1}^{n}\lambda(t_i)\right)\exp\left[-\int_0^{t_n}\lambda(x)dx\right]$$

$$=\left(\prod_{i=1}^{n}\frac{\beta}{\theta}\left(\frac{t_i}{\theta}\right)^{\beta-1}\right)\exp\left[-\int_0^{t_n}\frac{\beta}{\theta}\left(\frac{x}{\theta}\right)^{\beta-1}dx\right]$$

$$=\frac{\beta^n}{\theta^{n\beta}}\left(\prod_{i}^{n}t_i\right)^{\beta-1}\exp\left[-\left(\frac{t_n}{\theta}\right)^{\beta}\right]$$

To get the MLE's, we take the logarithm of this joint density and set the first partial derivatives (with respect to θ and β) equal to zero. The log-likelihood function is

$$l(\theta,\beta|t)=n\ln\beta-n\beta\ln\theta+(\beta-1)\sum_{i=1}^{n}\ln t_i-\left(\frac{t_n}{\theta}\right)^{\beta}\quad\text{and}$$

$$0=\frac{\partial l}{\partial\theta}=-\frac{n\beta}{\theta}+\frac{\beta}{\theta}\left(\frac{t_n}{\theta}\right)^{\beta}$$

$$0=\frac{\partial l}{\partial\beta}=\frac{n}{\beta}-n\ln\theta+\sum_{i=1}^{n}\ln t_i-\left(\frac{t_n}{\theta}\right)^{\beta}\ln\left(\frac{t_n}{\theta}\right)$$

The first equation simplifies to $0 = -n + \left(\frac{t_n}{\theta}\right)^{\beta}$

which can be solved for θ (in terms of β) to obtain

$$\hat{\theta} = t_n / n^{1/\hat{\beta}}$$

Substituting back into the first equation yields

$$0 = \frac{\partial l}{\partial \beta} = \frac{n}{\beta} - n \ln \frac{t_n}{n^{1/\beta}} + \sum_{i=1}^{n} \ln t_i - \left(\frac{t_n n^{1/\beta}}{t_n}\right)^{\beta} \ln\left(\frac{t_n n^{1/\beta}}{t_n}\right)$$

Solving for β yields

$$\hat{\beta} = \frac{n}{\sum_{i=1}^{n-1} \ln\left(t_n / t_i\right)}.$$

Furthermore, $\frac{2n\beta_0}{\hat{\beta}} = 2n\beta_0 \left(\frac{n}{\sum_{i=1}^{n-1} \ln(t_n / t_i)}\right)^{-1} = 2\beta_0 \sum_{i=1}^{n-1} \ln(t_n / t_i) = -2\beta_0 \sum_{i=1}^{n-1} \ln(t_i / t_n)$ has a chi-squared distribution with $2n-2$ degrees of freedom (e.g., Crow, 1974, 1982; Rigdon and Basu, 2000).

Thus, a size α test of $H_0 : \beta = \beta_0$ against $H_a : \beta \neq \beta_0$ is to reject H_0 if

$$2n\beta_0 / \hat{\beta} \leq \chi^2_{\alpha/2}(2n-2) \text{ or } 2n\beta_0 / \hat{\beta} \geq \chi^2_{1-\alpha/2}(2n-2),$$

where $\chi^2_{\alpha/2}(2n-2)$ is the $100\alpha/2$ percentile of chi-squared distribution with $2n-2$ degrees of freedom. For $H_0 : \beta = 1$, the test statistic $\frac{2n\beta_0}{\hat{\beta}}$ reduces to

$$Z = 2\sum_{i=1}^{n-1} \ln(t_n / t_i) = -2\sum_{i=1}^{n-1} \ln(t_i / t_n).$$

2.3 Empirical Example

Three sets of pseudo-data based on five numbers (14, 34, 42, 72, and 244; Ascher, 1983) are used for the following analysis. Dot-plots (Figure 2-1) display the system activities as waning, random, and developing, respectively. Estimates of β and p-values (Table 2-1) confirm the claimed temporal trends,

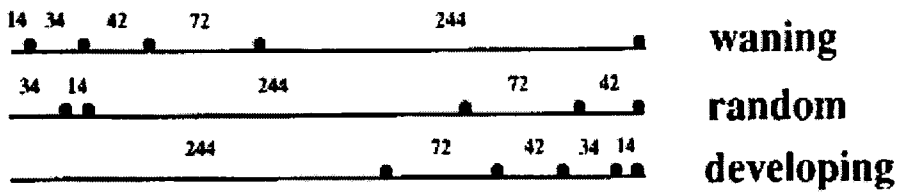

Figure 2-1 Dot-plots of pseudo-data in their original chronological orders

Table 2-1 Summary statistics of the pseudo-data

	Waning	Random	Developing
$\hat{\beta}$	0.6287716	0.9863077	5.414036
$\hat{\theta}$	31.39658	79.40588	301.594
$2n/\hat{\beta}$	15.90403	10.13882	1.847051
One-sided p-Value	0.043775	0.2554175	0.0146529

CHAPTER 3

EMPIRICAL RECURRENCE RATES TIME SERIES

3.1 The Empirical Recurrence Rates

Let $t_1,....,t_n$ be the n ordered failures during an observation period, $(0,T)$, from the first occurrence to the last occurrence. Then a discrete time series $\{z_l\}$ is generated sequentially at equidistant time intervals $h, 2h,..., lh,..., Nh\ (=T)$. If 0 is adopted as the time-origin and h as the time–step, then we regard z_l as the observation at time $t = lh$. Therefore, we propose a time series of the empirical recurrence rates as follows:

$$z_l = n_l/lh = \text{Total number of failures in}\,(0,lh)/lh,$$

where $l = 1,2,..., N$. Note that z_l evolves over time and it is simply the MLE for the mean rate of a simple Poisson process observed in $(0,lh)$. The time plot of the empirical recurrence rate (ERR-plot) offers the possibility of further insights into the data.

3.2 ERR-plot of pseudo-data

ERR-plots for the observation period, (0, T), are produced respectively for the data sets presented in Ch. 2. Consistent with the previous notation, we use h = 10, 20, 40, 50, 60, and 70. Because the sample total of these five numbers is 406, we recommend

$T = h\left\{\left[\frac{406}{h}\right]+1\right\}$, where $\left[\frac{406}{h}\right]$ is the largest integer less than or equal to $\frac{406}{h}$ for each h = 10, 20, 40, 50, 60, and 70.

The distinctive signature, marking the unique failure pattern of each repairable system (data set), is displayed in Figures 3-1, 2, and 3. Clearly, the overall pattern of the recurrence rates is affected but preserved by the choice of h. The sensitivity of this parameter will be addressed later.

DATA SET: 244, 72,42,34,14

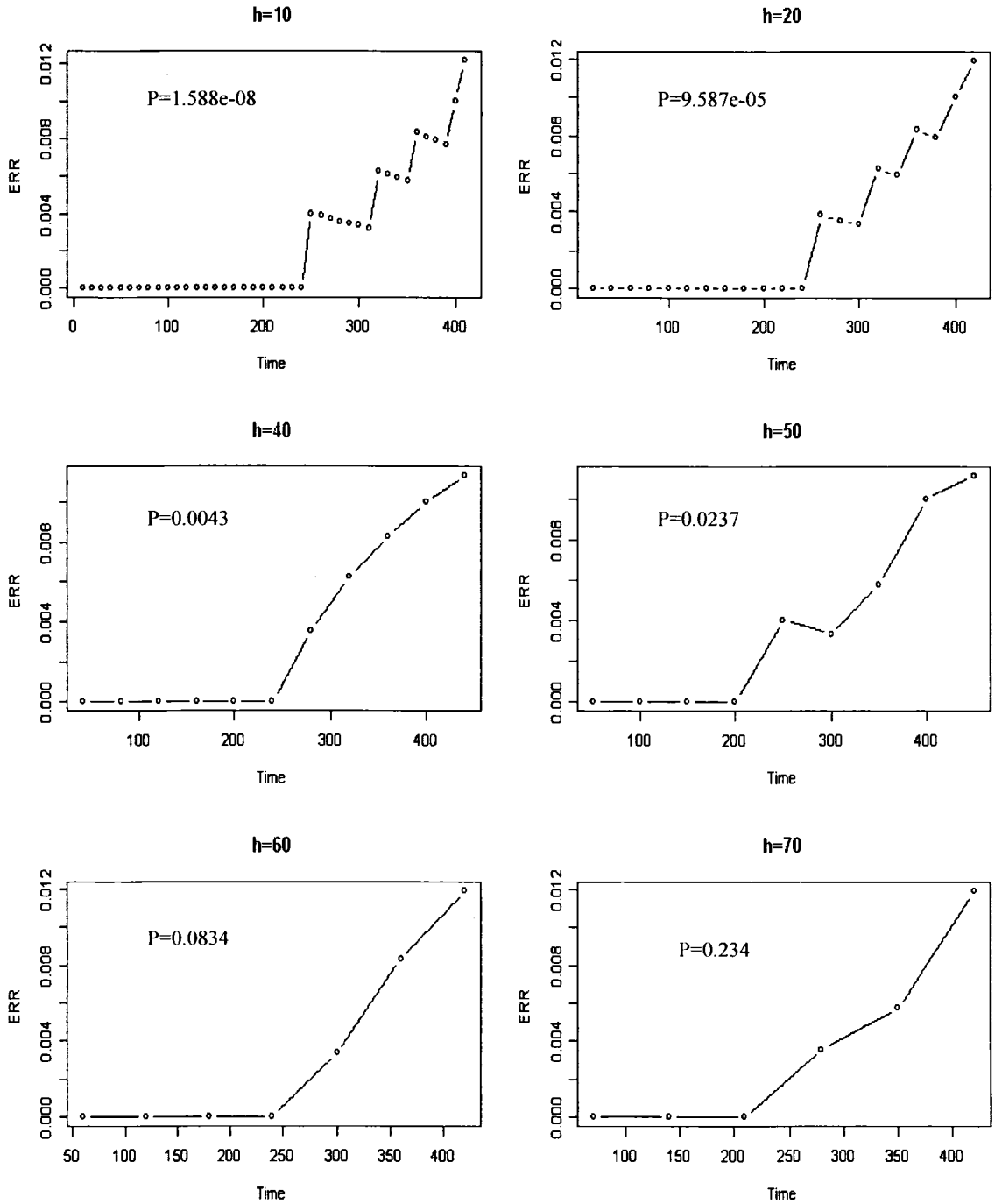

Figure 3-1 ERR-Plots with different time step (h) for the developing data set

DATA SET: 14,34,42,72,244

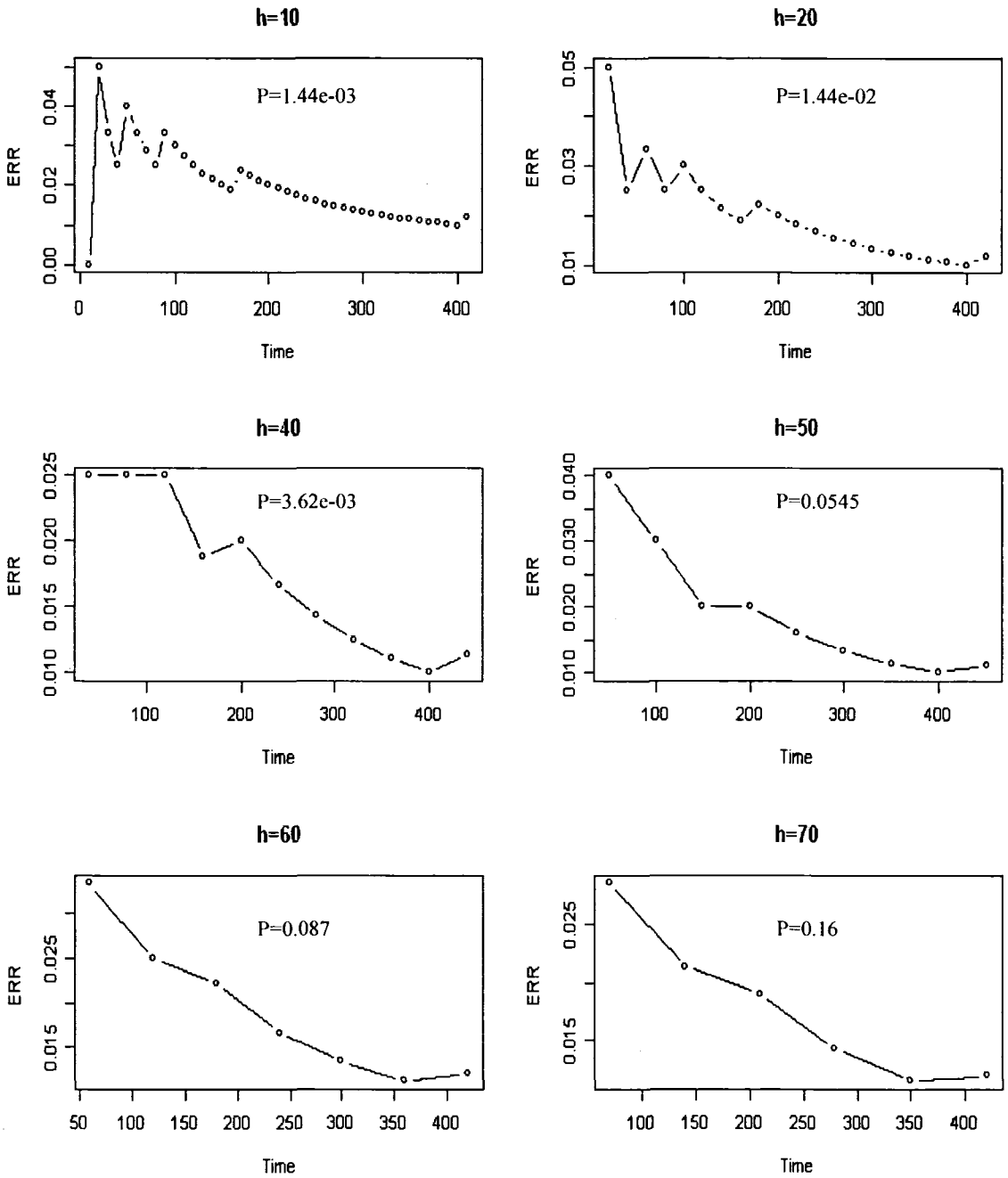

Figure 3-2 ERR-Plots with different time step (h) for the waning data set

DATA SET: 34,14,244,72,42

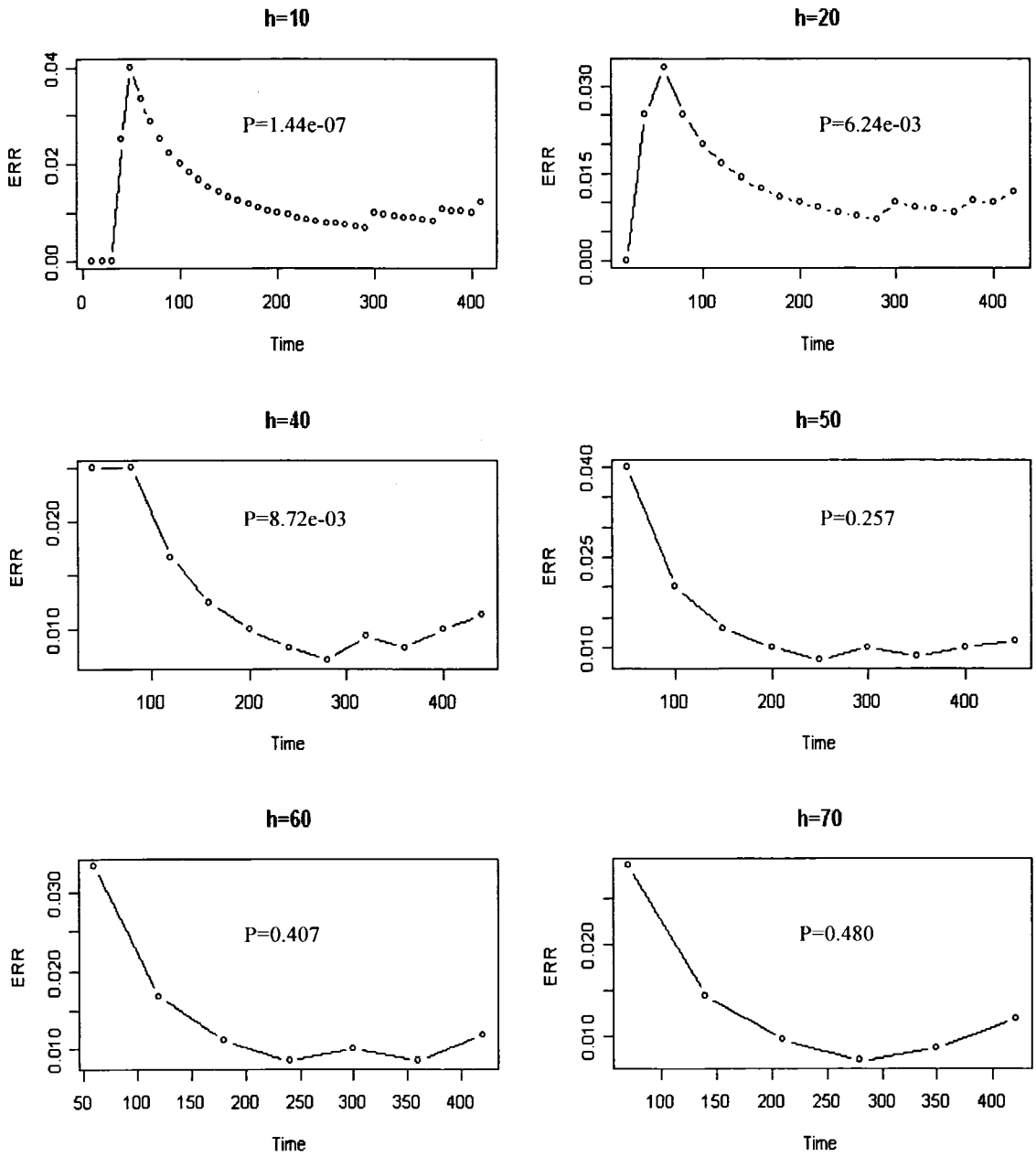

Figure 3-3 ERR-Plots with different time step (h) for the random data set

3.3 Basic Theory

3.3.1 Autocorrelation Function

Theorem 1

Given $Y_1, Y_2, ..., Y_n \overset{iid}{\sim} Y$, where $hY \sim POI(h\lambda)$, $h > 0$, $\lambda > 0$ then,

The autocovariance function of $\{\bar{Y}_j\}$ at lag k is

$$\gamma_j(k) = Cov(\bar{Y}_j, \bar{Y}_{j+k}) = \frac{\lambda}{h(j+k)}, \quad \text{for} \quad j = 1,2,.....,n-1, \quad k = 0,1,...,n-j.$$

The autocorrelation function of $\{\bar{Y}_j\}$ at lag k is

$$\rho_j(k) = \gamma_j(k)/[\gamma_j(0)\gamma_{j+k}(0)]^{1/2} = Cor(\bar{Y}_j, \bar{Y}_{j+k}) = [j/(j+k)]^{1/2},$$

$$\text{for } j = 1,2,.....,n-1, \quad k = 0,1,...,n-j.$$

Proof:

$$Cov(\bar{Y}_j, \bar{Y}_{j+k}) = Cov\left(\frac{1}{j}\sum_{i=1}^{j} Y_i, \frac{1}{j+k}\sum_{l=1}^{j+k} Y_l\right)$$

$$= \frac{1}{j(j+k)}\sum_{i=1}^{j}\sum_{l=1}^{j+k} Cov(Y_i, Y_l)$$

$$= \frac{1}{j(j+k)}\sum_{i=1}^{j} Cov(Y_i, Y_i) \quad (\because Cov(Y_i, Y_l) = 0, \quad if \quad i \neq l)$$

$$= \frac{1}{j(j+k)}\sum_{i=1}^{j} \frac{\lambda}{h} \quad (\because Cov(Y_i, Y_i) = \frac{\lambda}{h})$$

$$= \frac{1}{j(j+k)} \times \frac{j\lambda}{h} = \frac{\lambda}{h(j+k)}$$

Q.E.D.

All the terms are correlated; the further apart they are, the less is the correlation between them. Therefore, time series models (e.g., Box and Jenkins, 1976) should be used for predicting future recurrence rate.

3.3.2 Ljung and Box Test (LB-test)

In time series data analysis, after treating the original data by eliminating and subtracting the trend and seasonal components, we need to check if the residuals are observed values of independent and identically distributed random variable. A popular test, formulated by Ljung and Box (1978), uses the following test statistic:

$$\hat{Q}_{LB}(\hat{\rho}) = n(n+2)\sum_{k=1}^{m} \hat{\rho}_k^2 \ /(n-k),$$

where $\hat{\rho}_k = \sum_{l=k+1}^{n} \hat{a}_l \hat{a}_{l-k} \Big/ \sum_{l=1}^{n} \hat{a}_l^2$, the estimated autocorrelation at lag k

n , the sample size

m , number of lags being tested (As a rule of thumb, the sample ACF and PACF are good estimates of the ACF and PACF of a stationary process for lags up to about a third of the sample size, Brockwell and Davis, 2003).

$\hat{a}_1,...,\hat{a}_n$, residuals after a model has been fitted to a series $y_1,...,y_n$; if no model is being fitted, then $\hat{a}_1,...,\hat{a}_n$ are the "mean corrected" series of $y_1,...,y_n$.

We reject the iid hypothesis at level α if $\hat{Q}_{LB}(\rho) > \chi^2_{1-\alpha,m-p-q}$, where $\chi^2_{1-\alpha,m-p-q}$ is the $1-\alpha$ quantile of the chi-squared distribution with $m-p-q$ degrees of freedom, $p+q$ is the number of parameters of the fitted model.

3.4 LB-test for the pseudo data

$P-$values on the iid test using Ljung and Box method are recorded inside the box of each ERR-plot (Figures 3-1, 2, 3). The $p-$values increase with h for each data set (Table 3-1). A plausible explanation is that, by the Central Limit Theorem, large h produces sample means (i.e., empirical recurrence rates for our time series) which are closer to the true mean. Thus, it is harder for the LB-test to reject an iid test for a larger h. Apparently, the significance (p -value ≤ 0.05, say) of the LB-test doesn't resemble closely with that of the $Z-$test, presented in Table 2.1. Of course, the main goal of our approach is to build a workable bridge between the traditional homogeneous and nonhomogeneous Poisson process and the classical time series. Fortunately, time series modeling are well developed and are largely applied in many other fields, which will greatly facilitate the needs of researchers in finding the best model for the empirical recurrence rates proposed in this thesis.

Table 3-1 P-value for the iid hypothesis test using Ljung-Box test for pseudo data

Time step (h)	Data		
	Waning 14,34,42,72,244	Random 34,14,244,72,42	Developing 244,72,42,34,14
10	$p = 1.48e-03$	$p = 1.44e-07$	$p = 1.588e-08$
20	$p = 1.44e-02$	$p = 6.24e-03$	$p = 9.587e-05$
40	$p = 3.62e-03$	$p = 8.72e-03$	$p = 0.0043$
50	$p = 0.0545$	$p = 0.257$	$p = 0.0237$
60	$p = 0.087$	$p = 0.407$	$p = 0.0834$
70	$p = 0.16$	$p = 0.480$	$p = 0.234$

CHAPTER 4

APPLICATIONS

4.1 Dot-plot and Z - test for the Mining Data

The control of industrial accidents generally requires, from time to time, new safety equipment, safety regulations, improved machinery, etc.; hence, one may expect that the occurrence of accidents would tend to decrease with time. Because of serious injuries or, perhaps, deaths that may occur as a result of an industrial accidents, it is usually important to know whether or not the safety action are resulting in a significant decrease of accidents. The nonhomogeneous Poisson process with Weibull intensity function may possibly be useful in measuring this decrease (Crow, 1974).

The data in Table 4-1 (Maguire et al, 1952, Table 1) represent days between explosions in mines in Great Britain involving more than 10 men killed. The data cover the period from December 6, 1875 to May 29, 1951. A dot-plot is presented as Figure 4.1, which suggests the applicability of a nonhomogeneous Poisson process with Weibull intensity. Table 4-2 summarizes the results, which further confirms a significant decreasing trend in mining accidents during the observation period.

Table 4-1 Time intervals in days between explosions in mines, involving 10 men killed, from 6 December 1875 to 29 May 1951

378	59	54	498	217	156
36	61	217	49	120	47
15	1	113	131	275	129
31	13	32	182	20	1630
215	189	23	255	66	29
11	345	151	195	291	217
137	20	361	224	4	7
4	81	312	566	369	18
15	286	354	390	338	1357
72	114	58	72	336	
96	108	275	228	19	
124	188	78	271	329	
50	233	17	208	330	
120	28	1205	517	312	
203	22	644	1613	171	
176	61	467	54	145	
55	78	871	326	75	
93	99	48	1312	364	
59	326	123	348	37	
315	275	457	745	19	

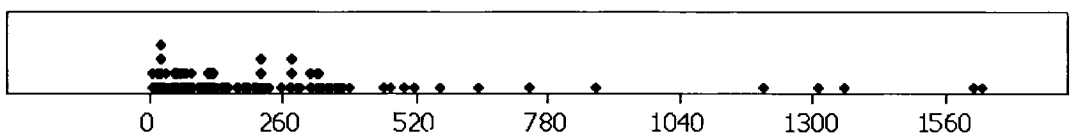

Figure 4-1 Dot-plots of Mining Data

Table 4.2 Summary Statistics of the Mining Data

$\hat{\beta}$	0.711759
$\hat{\theta}$	36.04305
$2n/\hat{\beta}$	258.597
One-sided $p-value$	0.025

4.2 ERR-plotting and LB-test

ERR-plots for the observation period, (0, T), are produced respectively for the mining data (Table 4.1). Consistent with the previous notation, we use h = 200k, for k = 1, ..., 6. Because the sample total of the 109 successive mine accidents was 26,263, we use

$T = h\left\{\left[\frac{26263}{h}\right]+1\right\}$, where $\left[\frac{26263}{h}\right]$ is the largest integer less than or equal to $\frac{26263}{h}$

for each of the above h values. The ERR-plots are displayed in Figure 4.2. Clearly, there is a similarity in their patterns. In contrast to the dot-plot, the proposed graphing technique is extremely valuable for such a large data set. Results on the iid test using Ljung and Box method are recorded inside the box of each plot. All the p-values are

approximately zero. They are slightly increasing with h, the length of the time-step, which is consistent with those of the pseudo data.

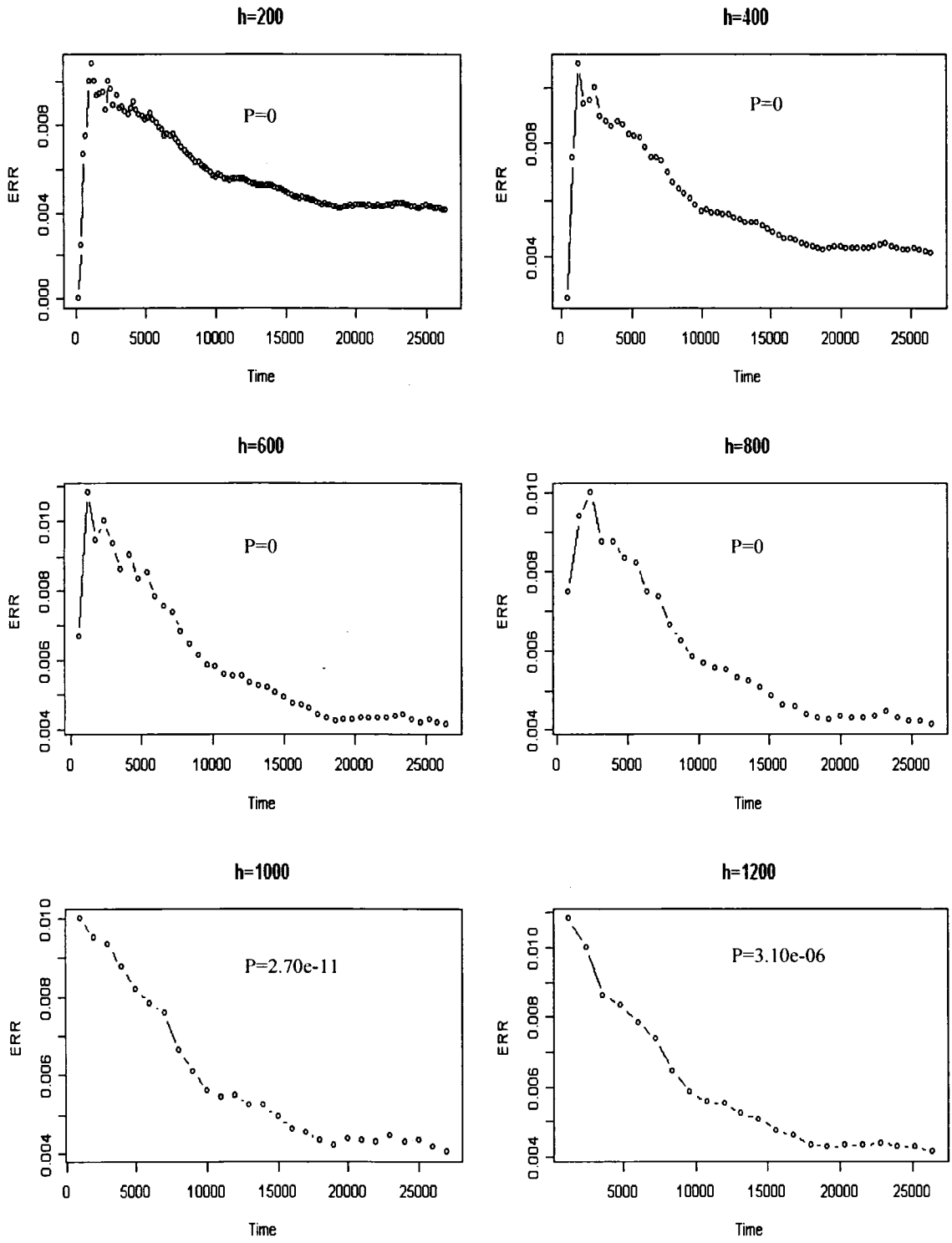

Figure 4-2 ERR-plots with different time-step (h) for data of mine accidents

CHAPTER 5

CONCLUSIONS

The main and long-term goal of this thesis is to characterize the recurrence rate presented by a point process with a discrete time series. The proposed empirical recurrence rate plot shows tremendous potential in serving as a workable bridge between two of the most powerful tools in the literature of statistics: Stochastic processes and time series. Plotting a data set in an intelligent way often lays the groundwork for a rigorous model fitting procedure that follows. The merits of transforming a dot plot to an ERR-plot are clearly demonstrated in our presentations. The proposed graphing technique is extremely valuable for a large data set such as the mine accidents data. Although recurrence rates of most repairable systems show simple patterns, some are more complicated than the example demonstrated in this thesis. Fortunately, time series modeling are well developed and are largely applied in many other fields. Statistical software packages are abundant, which will greatly facilitate the needs of researchers using the proposed methods.

CHAPTER 6

R-PROGRAM

Program 1: ERR-plotting of the developing pseudo data

```
h=c(10,20,40,50,60,70)
ERR=function(h){

Y=c(244,72,42,34,14)
Z=cumsum(Y)
M=cut(Z,seq(0,(as.integer(Z[5]/h)+1)*h,h))
N=table(M)
R=cumsum(N)
Q=h*seq(1,length(R),1)
ERR=R/Q     }

Q=function(h){
Y=c(244,72,42,34,14)
Z=cumsum(Y)
M=cut(Z,seq(0,(as.integer(Z[5]/h)+1)*h,h))
N=table(M)
R=cumsum(N)
Q=h*seq(1,length(R),1)
    }

 par (mfrow = c(3,2))
 plot (Q(10),ERR(10),type ='b', xlab = 'Time', ylab = 'ERR', main = "h=10")
 plot (Q(20),ERR(20),type ='b', xlab = 'Time', ylab = 'ERR', main = "h=20")
 plot (Q(40),ERR(40),type ='b', xlab = 'Time', ylab = 'ERR', main = "h=40")
 plot (Q(50),ERR(50),type ='b', xlab = 'Time', ylab = 'ERR', main = "h=50")
 plot (Q(60),ERR(60),type ='b', xlab = 'Time', ylab = 'ERR', main = "h=60")
 plot (Q(70),ERR(70),type ='b', xlab = 'Time', ylab = 'ERR', main = "h=70")
```

Program 2: ERR-plotting of the waning pseudo data

```
h=c(10,20,40,50,60,70)
 ERR=function(h){
 Y=c(14,34,42,72,244)
    Z=cumsum(Y)
    M=cut(Z,seq(0,(as.integer(Z[5]/h)+1)*h,h))
    N=table(M)
    R=cumsum(N)
    Q=h*seq(1,length(R),1)
    ERR=R/Q     }

 Q=function(h){
 Y=c(14,34,42,72,244)
    Z=cumsum(Y)
    M=cut(Z,seq(0,(as.integer(Z[5]/h)+1)*h,h))
    N=table(M)
    R=cumsum(N)
    Q=h*seq(1,length(R),1)
      }

 par (mfrow = c(3,2))
 plot (Q(10),ERR(10),type ='b', xlab = 'Time', ylab = 'ERR', main = "h=10")
 plot (Q(20),ERR(20),type ='b', xlab = 'Time', ylab = 'ERR', main = "h=20")
 plot (Q(40),ERR(40),type ='b', xlab = 'Time', ylab = 'ERR', main = "h=40")
 plot (Q(50),ERR(50),type ='b', xlab = 'Time', ylab = 'ERR', main = "h=50")
 plot (Q(60),ERR(60),type ='b', xlab = 'Time', ylab = 'ERR', main = "h=60")
 plot (Q(70),ERR(70),type ='b', xlab = 'Time', ylab = 'ERR', main = "h=70")
```

Program 3: ERR-plotting of the random pseudo data

```
h=c(10,20,40,50,60,70)
ERR=function(h){

Y=c(34,14,244,72,42)
   Z=cumsum(Y)
   M=cut(Z,seq(0,(as.integer(Z[5]/h)+1)*h,h))
   N=table(M)
   R=cumsum(N)
   Q=h*seq(1,length(R),1)
```

```
   ERR=R/Q     }

Q=function(h){
Y=c(34,14,244,72,42)
   Z=cumsum(Y)
   M=cut(Z,seq(0,(as.integer(Z[5]/h)+1)*h,h))
   N=table(M)
   R=cumsum(N)
   Q=h*seq(1,length(R),1)
     }
par (mfrow = c(3,2))
plot (Q(10),ERR(10),type ='b', xlab = 'Time', ylab = 'ERR', main = "h=10")
plot (Q(20),ERR(20),type ='b', xlab = 'Time', ylab = 'ERR', main = "h=20")
plot (Q(40),ERR(40),type ='b', xlab = 'Time', ylab = 'ERR', main = "h=40")
plot (Q(50),ERR(50),type ='b', xlab = 'Time', ylab = 'ERR', main = "h=50")
plot (Q(60),ERR(60),type ='b', xlab = 'Time', ylab = 'ERR', main = "h=60")
plot (Q(70),ERR(70),type ='b', xlab = 'Time', ylab = 'ERR', main = "h=70")
```

Program 4: ERR-plotting of the mine accidents data

```
h=c(seq(200,1200,by=200))
ERR=function(h){

Y=c(378,36,15,31,215,11,137,4,15,72,96,124,50,120,203,176,55,93,59,315,59,61,1,13,1
89,345,20,81,286,114,108,188,233,28,22,61,78,99,326,275,54,217,113,32,23,151,361,31
2,354,58,275,78,17,1205,644,467,871,48,123,457,498,49,131,182,255,195,224,566,390,7
2,228,271,208,517,1613,54,326,1312,348,745,217,120,275,20,66,291,4,369,338,336,19,3
29,330,312,171,145,75,364,37,19,156,47,129,1630,29,217,7,18,1357)
   Z=cumsum(Y)
   M=cut(Z,seq(0,(as.integer(Z[109]/h)+1)*h,h))
   N=table(M)
   R=cumsum(N)
   Q=h*seq(1,length(R),1)
   ERR=R/Q     }

Q=function(h)     {
Y=c(378,36,15,31,215,11,137,4,15,72,96,124,50,120,203,176,55,93,59,315,59,61,1,13,1
89,345,20,81,286,114,108,188,233,28,22,61,78,99,326,275,54,217,113,32,23,151,361,31
2,354,58,275,78,17,1205,644,467,871,48,123,457,498,49,131,182,255,195,224,566,390,7
2,228,271,208,517,1613,54,326,1312,348,745,217,120,275,20,66,291,4,369,338,336,19,3
29,330,312,171,145,75,364,37,19,156,47,129,1630,29,217,7,18,1357)
   Z=cumsum(Y)
```

```
   M=cut(Z,seq(0,(as.integer(Z[109]/h)+1)*h,h))
   N=table(M)
   R=cumsum(N)
   Q=h*seq(1,length(R),1)
            }

par (mfrow = c(3,2))
plot (Q(200),ERR(200),type ='b', xlab = 'Time', ylab = 'ERR', main = "h=200")
plot (Q(400),ERR(400),type ='b', xlab = 'Time', ylab = 'ERR', main = "h=400")
plot (Q(600),ERR(600),type ='b', xlab = 'Time', ylab = 'ERR', main = "h=600")
plot (Q(800),ERR(800),type ='b', xlab = 'Time', ylab = 'ERR', main = "h=800")
plot (Q(1000),ERR(1000),type ='b', xlab = 'Time', ylab = 'ERR', main = "h=1000")
plot (Q(1200),ERR(1200),type ='b', xlab = 'Time', ylab = 'ERR', main = "h=1200")

for (j in c(200, 400, 600, 800, 1000, 1200)) {

Y=c(378,36,15,31,215,11,137,4,15,72,96,124,50,120,203,176,55,93,59,315,59,61,1,13,1
89,345,20,81,286,114,108,188,233,28,22,61,78,99,326,275,54,217,113,32,23,151,361,31
2,354,58,275,78,17,1205,644,467,871,48,123,457,498,49,131,182,255,195,224,566,390,7
2,228,271,208,517,1613,54,326,1312,348,745,217,120,275,20,66,291,4,369,338,336,19,3
29,330,312,171,145,75,364,37,19,156,47,129,1630,29,217,7,18,1357)
   Z=cumsum(Y)
   M=cut(Z,seq(0,(as.integer(Z[109]/j)+1)*j,j))
   N=table(M)
   R=cumsum(N)
   Q=j*seq(1,length(R),1)
   ERR=R/Q
  A=Box.test (ERR, lag=as.integer ((5/12)*(Z [109]/j)), type="Ljung")$p.value
   print(A)      }
```

REFERENCES

1. Ascher H., Discussion on statistical Methods in Reliability, by J. F. Lawless: Technometrics, v25, p.305-335, 1983.
2. Bain Lee J., and Engelhardt Max, Inference on the Parameters and current System Reliability for a Time Truncated Weibull Process. Technometrics v. 22, No .4, p.305-335. 1980.
3. Bain Lee J., and Engelhardt Max, Statistical Analysis of Reliability and Life-Testing Models : Theory and Methods, Second Edition, Marcel-Dekker: New York. 1991.
4. Box G.E.P., and Jenkins G.M., Time Series Analysis Forecasting and Control Holden Day. 1976.
5. Brockwell Peter J., Davis Richard A., Introduction to Time Series and Forecasting. Springer Texts in Statistics, 2003.
6. Crow Larry H., Reliability Analysis for Complex, Repairable Systems. Reliability and Biometry, p. 379-410. 1974.
7. Crow Larry H., Confidence Interval Procedures for the Weibull Process with Applications to Reliability Growth, Technometrics, v.24, No. 3, p.67-72. 1982.
8. Ho Chih-Hsiang, Forward and Backward Tests for an Abrupt Change in the Intensity of a Poisson Process: J. Statist. Comput. Simul. v.48, No.2, p. 245-252. 1993.

9. Ho Chih-Hsiang, Repeated Significance Tests on Accumulating Data of Repairable Systems, Commun. Statist., 27(5), p.1181-1200. 1998.

10. Ljung G.M., Box G.E. P., On a measure of lack of fit in time series models. Biometrika, 65,2, pp.297-303.1978.

11. Maguire, B.A., Pearson E.S., Wynn A.H.A., The Time Intervals between Industrial Accidents, Biometrika, 39, p.168-180.1952.

12. Rigdon Steven. E., Basu Asit. P., Statistical Methods for the Reliability of Repairable Systems. Wiley. 2000.

VITA

Graduate College
University of Nevada, Las Vegas

Hui Wang

Local Address:
1600 E University Ave. Apt 209
Las Vegas, Nevada, 89119

Degrees:
Bachelor of Engineering, Mechanical Engineering, 1994
Beijing Institute of Machinery, Beijing

Master of Science in Engineering
University of Nevada Las Vegas, 2003

Thesis Title:
Statistical Modeling via Empirical Recurrence Rate

Thesis Examination Committee:
Chairperson, Dr. Chih-Hsiang Ho, Ph. D.
Committee Member, Dr Malwane Ananda, Ph. D.
Committee Member, Dr Sandra Catlin, Ph. D.
Graduate Faculty Representative, Dr. Shizhi Qian, Ph. D.