

**CONSTRUCTING CONFIDENCE INTERVALS FOR  
POLARIZATION MODE DISPERSION**

by

Warrick Erlank

Submitted in fulfilment of the requirements for the degree of

**Masters in Mathematical Statistics**

at the Nelson Mandela Metropolitan University

December 2008

Supervisor: Mr D Friskin

Co-Supervisor: Prof A Leitch

## **Abstract**

Polarization mode dispersion (PMD) causes significant impairment in high bit-rate optical telecommunications systems. Knowledge of a fibre's PMD mean value, and the relevant confidence interval, is essential for determining a fibre's maximum allowable bit-rate. Various methods of confidence interval construction for time series data were tested in this dissertation using simulation. These included the autocovariance-matrix methods as suggested by Box and Jenkins, as well as the more practical and simpler batch means methods. Some of these methods were shown to be significantly better than the standard method of calculating confidence intervals for non time series data. The best of the tested methods were used on actual PMD data. The effect of using polarization scramblers was also tested.

## **Keywords**

Polarization mode dispersion

Confidence interval construction methods

Batch means

## Acknowledgements

I would like to express my gratitude to the following:

- My mother, father and grandmother for their support. Without you, I would not have completed this dissertation.
- My supervisor, Mr David Friskin. Thank you for your great guidance, encouragement and patience.
- My co-supervisor, Prof Andrew Leitch, and the NMMU Physics department.
- Lawrence Viljoen and Timothy Gibbon for their help in providing me with the data used.
- This project was financially supported by the NRF, Abdare Fibre Optic Cables, Corning Optical Fiber, Telkom (Pty) Ltd and THRIP, through the Telkom CoE at the NMMU.

## **DECLARATION BY STUDENT**

**FULL NAME:**           Warrick Erlank

**STUDENT NUMBER:** 201318989

**QUALIFICATION:**     **Masters in Mathematical Statistics**

**DECLARATION:**

In accordance with Rule G4.6.3, I hereby declare that the above-mentioned dissertation is my own work and that it has not previously been submitted for assessment to another University or for another qualification.

**SIGNATURE:**       .....

**DATE:**             .....

# Table of Contents

1	Introduction.....	1
2	Polarization Mode Dispersion .....	4
2.1	Introduction .....	4
2.2	PMD coefficient .....	5
2.3	Birefringence .....	7
2.4	Mode Coupling .....	8
2.5	DGD & PMD Distribution .....	10
2.6	Time and Wavelength Ensembles.....	11
2.7	Polarization Scramblers .....	12
3	Statistical Concepts .....	13
3.1	Basic Definitions.....	13
3.1.1	Stationarity .....	13
3.1.2	Autocovariance .....	14
3.1.3	Correlation.....	15
3.1.4	Autocorrelation.....	15
3.2	ARMA Models .....	16
3.2.1	AR Models.....	17
3.2.2	MA Models .....	17
3.2.3	ARMA Models .....	18
3.3	Confidence Interval Construction Methods.....	18
3.3.1	Autocovariance Methods.....	18
3.3.1.1	Autocovariance Matrix Method .....	18
3.3.1.2	Autocovariance Quarter Matrix Method .....	20
3.3.1.3	Autocorrelation-length Matrix Method.....	21
3.3.2	Batch Means Methods .....	22
3.3.2.1	Batch Means – 10 batches (batch size = 50).....	23
3.3.2.2	Batch Means – 50 batches (batch size = 10).....	24
3.3.2.3	Batch Means – Autocorrelation-length batch size.....	24
3.4	Test for Stationarity .....	25

4	Results .....	26
4.1	Overview .....	26
4.2	Simulation Example.....	26
4.2.1	Autocovariance Methods.....	27
4.2.1.1	Autocovariance Matrix Method .....	29
4.2.1.2	Autocovariance Quarter Matrix Method .....	30
4.2.1.3	Autocorrelation-length Matrix Method.....	31
4.2.2	Batch Means Methods .....	32
4.2.2.1	Batch Means (10 batches) Method.....	33
4.2.2.2	Batch Means (50 batches) Method.....	33
4.2.2.3	Batch Means (Autocorrelation-length batch size) Method .....	34
4.3	Model Results .....	35
4.3.1	AR(1) Model.....	36
4.3.2	AR(2) Model.....	37
4.3.3	MA(1) Model.....	39
4.3.4	MA(2) Model.....	40
4.3.5	ARMA(1,1) Model.....	42
4.3.6	Normal(5, 0.1) Model .....	44
4.3.7	Model Results Summary .....	45
4.4	Standard Error Histograms.....	47
4.4.1	True standard errors .....	47
4.4.2	MA(2) Histograms .....	49
4.5	Actual PMD data .....	56
4.5.1	Hopetown to Britstown .....	56
4.5.2	Britstown to Victoria West .....	60
4.6	Effects of Polarization Scramblers.....	64
5	Conclusion.....	67
6	References.....	68
7	Appendix .....	70

## Table of Figures

Figure 2.1 – 3D pulse traveling down a fibre .....	4
Figure 2.2 - Birefringence caused by manufacturing process .....	7
Figure 2.3 - Physical model of mode coupling .....	8
Figure 2.4 - Mode coupling and pulse spread.....	9
Figure 2.5 - Maxwellian distribution example.....	10
Figure 3.1 – Batch Means method example .....	22
Figure 4.1 - AR(1) 95% interval simulation results.....	36
Figure 4.2 - AR(1) 50% interval simulation results.....	37
Figure 4.3 - AR(2) 95% interval simulation results.....	38
Figure 4.4 - AR(2) 50% interval simulation results.....	38
Figure 4.5 - MA(1) 95% interval simulation results .....	39
Figure 4.6 - MA(1) 50% interval simulation results .....	40
Figure 4.7 - MA(2) 95% interval simulation results .....	41
Figure 4.8 - MA(2) 50% interval simulation results .....	41
Figure 4.9 - ARMA(1,1) 95% interval simulation results.....	42
Figure 4.10 - ARMA(1,1) 50% interval simulation results.....	43
Figure 4.11 - N(5, 0.1) 95% interval simulation results .....	44
Figure 4.12 - N(5, 0.1) 50% interval simulation results .....	45
Figure 4.13 - MA(2) SEs - Standard Method .....	49
Figure 4.14 - MA(2) SEs – Autocovariance Matrix Method.....	50
Figure 4.15 - MA(2) SEs – Autocovariance Quarter Matrix Method .....	51
Figure 4.16 - MA(2) SEs – Autocorrelation-length Matrix Method.....	52
Figure 4.17 - MA(2) SEs – Batch Means (10 batches) Method.....	53
Figure 4.18 - MA(2) SEs – Batch Means (50 batches) Method.....	54
Figure 4.19 - MA(2) SEs – Batch Means (Autocorrelation-length) Method .....	55
Figure 4.20 - PMD plot: Hopetown to Britstown.....	56
Figure 4.21 - Autocorrelation plot: Hopetown to Britstown .....	57
Figure 4.22 - 95% interval - Standard method: Hopetown to Britstown.....	58
Figure 4.23 - 95% interval - Batch Means (10) method: Hopetown to Britstown .....	59
Figure 4.24 - 95% interval - Autocorrelation-length Matrix method: Hopetown to Britstown ....	59
Figure 4.25 - PMD plot: Britstown to Victoria West.....	60
Figure 4.26 - Autocorrelation plot: Britstown to Victoria West .....	61
Figure 4.27 - 95% interval - Standard method: Britstown to Vic West.....	62
Figure 4.28 - 95% interval - Batch Means (10) method: Britstown to Vic West .....	63

Figure 4.29 - 95% interval - Autocorrelation-length Matrix method: Britstown to Vic West .....	63
Figure 4.30 - Unscrambled data .....	64
Figure 4.31 - Scrambled data .....	65



# 1 Introduction

Since its invention in the early 1970's, fibre optics has become a data communications technology that has revolutionized data transfer. The applications of the technology are widespread, ranging from global networks to local telephone exchanges to subscribers' homes to desktop computers. Fibre is an integral part of modern day communication infrastructure and can be found along roads, in buildings, hospitals and machinery.

Fibre optics, similar to copper wire, is used for moving information from one place to another. The medium of transmission, however, is light as opposed to electricity, and that light is carried within an optical fibre via the process of total internal reflection. Some of the many advantages of fibre over copper include:

- Higher data transmission rates
- Cheaper and lighter cables
- Safety in factories due to lack of sparks
- Data security

With the advent of the internet, the need for fibre optics has become especially important. The current boom in fibre optics companies is a direct result of the high bandwidth required by internet data, especially graphical data. One of the fastest growing fibre optic markets is transmitting information for so-called intelligent transportation systems: "smart" highways and streets with traffic lights that respond to changing traffic patterns, automated toll booths, and changeable message signs that give motorists information about delays and emergencies [1].

Recently fibre optic technology has moved relatively easily from speeds of 90 Mbit/s (Mega-bits per second) to faster and faster speeds and eventually to 2.5 Gbit/s (Giga-bits per second). Then a noticeable problem became apparent in the 10 Gbit/s systems and threatened to cause major disturbances at 40 Gbit/s networking as well as on multi-channel cable television (CATV) transmission systems [2]. For the first time the companies which produced or used fibre optics were faced with a major network killer. This problem, which was only discovered as recently as the early 1990's, is called polarization mode dispersion, or PMD.

The physics and causes behind PMD will be described in Chapter 2. In short, when light travels down a fibre, it has two modes or planes, one fast and one slow. PMD is the time delay between these two modes. The greater this delay, the more bit errors will be caused, as the receivers have difficulty in distinguishing unique data bits.

As the demand for higher data transmission increases, telecoms companies are forced to install new optical fibre or upgrade old fibres which have already been deployed. In order to determine if an old fibre has the capacity to tolerate higher transmission rates, engineers can sample the fibre's PMD values. If the fibre's mean PMD is below different threshold levels, with *a specific confidence level*, its data transmission rate can be increased safely, without loss of data packets. If the mean PMD on these fibres is too high however, then the fibres have to be replaced.

This concept of determining a fibre's mean PMD, with a specific confidence level, is the core concept of this project. Because PMD data are serially-correlated (each sample reading is not independent of the preceding readings), conventional methods of calculating these confidence limits do not perform well. Various methods of calculating confidence intervals for time series data like this have been suggested.

The first question which will be investigated is which of these methods performs the best in theory, using simulation. It will then be established whether or not PMD data are stationary – i.e. whether or not the data stay “constant” and do not change drastically over time. If PMD data are in fact stationary, then it will be possible to apply some of the methods tested during the simulation to actual PMD data to obtain accurate confidence intervals. These results will then be compared to the conventional methods of confidence interval construction.

A device, called a PMD scrambler, can be applied to a fibre to make the individual data points less serially correlated, i.e. more independent from one another. If the data are made to be independent, then the above time series methods of interval construction will not be necessary when determining mean PMD levels and the standard methods of confidence interval construction may be used. The final part of the results section will deal with this possibility.

To summarise, the three research questions which will be investigated are:

1. Simulate time series data to determine the best of several methods of confidence interval construction for the mean of time series data.

2. Test to see whether actual PMD data are stationary.
3. Test to see whether using a PMD scrambler on a fibre makes the data independent.

Chapter 2 contains the detailed physics behind PMD. Chapter 3 is a background of all the statistical concepts used in the project, including a review of different interval construction methods which have been tried on time series data. The major questions and results of the project are explored in chapter 4, and chapter 5 will be the conclusion with main findings.

## 2 Polarization Mode Dispersion

### 2.1 Introduction

In fibre optics, light travels down a fibre in two polarization modes, which are perpendicular to one another. In an *ideal* optical fibre, the core has a perfectly circular cross-section. In this case, the light would travel down the two orthogonal axes at exactly the same velocity, resulting in zero PMD. In a *realistic* fibre, however, there are random imperfections that break the circular symmetry, causing the light traveling along one polarization axis to move slower or faster than the light polarized along the other axis [3].

As a result, the two perpendicular pulses arrive at the end of the fibre at different times. Detectors capture the combined signal, which appears as a spread pulse. In a digital system, as the pulse spreads and overlaps (because of PMD) it becomes increasingly difficult to differentiate between 1's and 0's, which may result in bit errors [4].

Figure 2.1 [5] shows how an input pulse consists of two perpendicular axes; a slow polarization mode and a fast polarization mode. It also shows how the pulse will appear to expand.

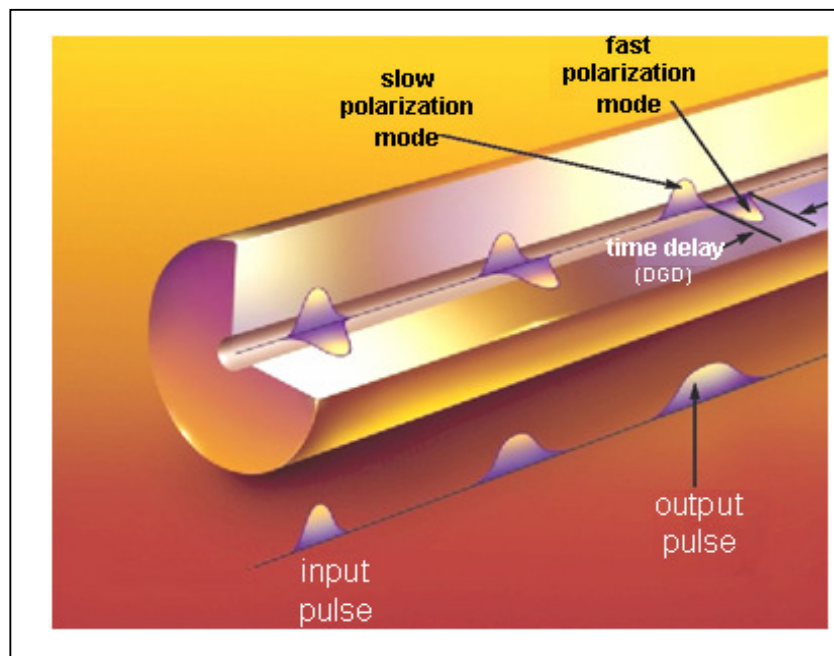


Figure 2.1 – 3D pulse traveling down a fibre

As can be seen in Figure 2.1, as the pulse moves down the fibre (direction is from left to right) PMD begins to affect the pulse and the polarized energy is separated by some time interval. This time interval is known as the differential group delay (DGD). A DGD value is measured and stored for every frequency over a wide range of bandwidth - in the experiments and calculations in subsequent chapters, a range of 100 nanometres of bandwidth is typically used. The average of these DGD values is then calculated and this value is referred to as the PMD, which is measured in picoseconds or ps ( $10^{-12}$  seconds). Specialists differ on its definition, but PMD is defined either as the mean DGD or its root-mean-square value [6]. For all calculations of PMD in this project, the mean DGD definition will be used.

If the bit errors are too numerous, then the transmitted information will be too corrupt to recover and the transmission link should be considered out of service. The quantity of bit errors encountered at the receiver is directly influenced by the amount of PMD in the fibre. Because of this, PMD limits transmission rates as well as the lengths of deployed fibre.

## 2.2 PMD coefficient

Because longer fibres will tend to exhibit larger PMD values, the PMD *coefficient* is defined as PMD per  $\sqrt{km}$ . This is then used as a standard measure so different fibre lengths are comparable. The maximum allowable length,  $L_{max}$ , for a given total average PMD coefficient, can be calculated from the following expression, assuming the PMD coefficient is less than or equal to 10% of the bit-rate, BR:

$$L_{max} = \frac{10000}{(PMD \cdot BR)^2}$$

where the PMD coefficient is measured in units of  $ps / \sqrt{km}$  and the bit-rate is in Gbit/s [7]. Table 2.1 contains a comparison of the different maximum allowable lengths of fibre for different PMD coefficient values:

<b>PMD coefficient</b>	<b>40 Gbit/s</b>	<b>10 Gbit/s</b>	<b>2.5 Gbit/s</b>
3.0	< 1km	11 km	180 km
1.0	6 km	100 km	1 600 km
0.5	25 km	400 km	6 400 km
0.1	625 km	10 000 km	160 000 km

**Table 2.1 - Transmission length as a function of PMD and transmission bit-rate.**

It is clear in Table 2.1, that the allowable length of fibre is very sensitive to different PMD coefficient values. As the PMD coefficient is calculated from the PMD itself, it is critical that researchers and manufacturers can estimate a particular fibre's mean PMD value, as well as a chosen confidence interval around that mean.

Many fibres installed before the mid 1990's, including some fibres which will be looked at, have PMD coefficients of about 1-2ps per  $\sqrt{km}$ . In South Africa where most fibres transmit data at 2.5 Gbit/s, these length issues are not yet a problem, but if telecommunication companies would like to upgrade their networks to 10 Gbit/s or faster in the future, then they will need to be very confident about the mean PMD values for each fibre.

The PMD coefficient can also be used to determine allowable bit-rates. Suppose one has a fibre lying between Cape Town and Port Elizabeth (approximately 750 km) which has a PMD coefficient between 1-2 ps, then allowable PMD bit-rates can be calculated as:

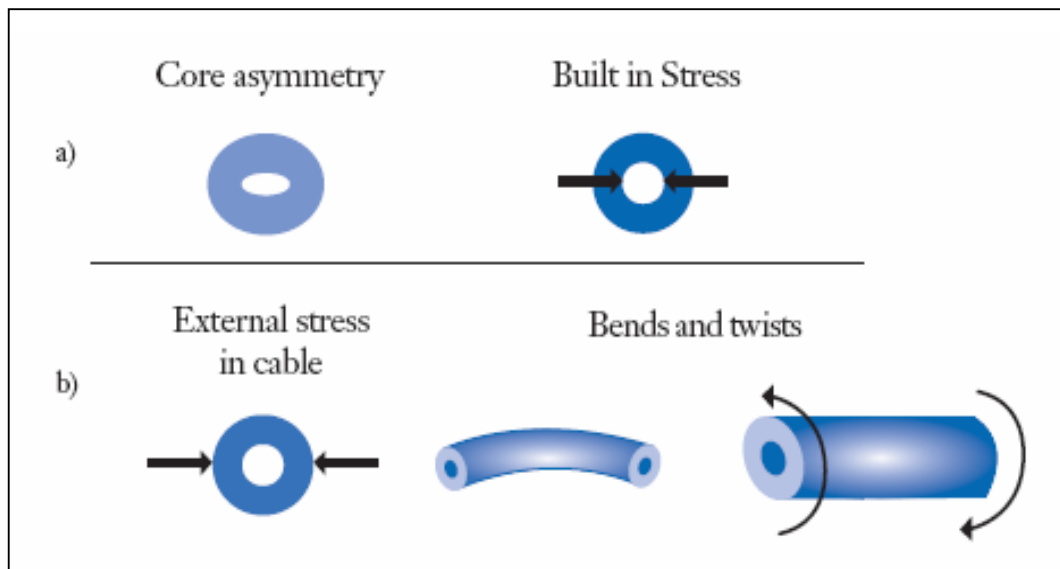
$$\begin{aligned}
 PMD &= (PMD\ coefficient) \times (\sqrt{km}) \\
 &= (1 - 2\ ps) / (\sqrt{km}) \times (\sqrt{750km}) \\
 &= 27 - 55\ ps
 \end{aligned}$$

For a 10 Gbit/s signal, the bits have a 100ps duration, therefore a PMD value of this size can be expected to result in a bit-error rate that is severe enough to cause service problems. A typical PMD limit for 10 Gbit/s signals is 10ps or 10% of the bit duration mentioned above. Similarly, a 40 Gbit/s link would only tolerate a 2.5ps PMD [8].

Two factors influencing PMD are fibre birefringence and mode coupling.

## 2.3 Birefringence

Birefringence is the phenomenon whereby light travels down the two perpendicular axes in a fibre at different speeds. There are many ways in which a fibre can become birefringent. Birefringence can arise due to internal stresses during the manufacturing process or through external stresses during cabling and installation, which can lead to a non-circular core. Even if the fibre is manufactured with an ideal circular core, birefringence can result during deployment or operation. Any non-uniform loading of the fibre cross-section, or bends or twists that are introduced to the fibre by sub-optimal cabling or installation will result in an asymmetric external stress being placed on the fibre [9]. Figure 2.2 shows different sources of birefringence: a) internal sources such as core asymmetry and built in stress, b) external sources such as bend twists and external stress applied to the fibre.



**Figure 2.2 - Birefringence caused by manufacturing process**

The core imperfections, and their effect on the PMD, remain constant over time. However, the mechanical stress on the fibre resulting from movement or environmental issues can vary, resulting in a dynamic aspect to PMD. This mechanical stress can be caused by a number of sources.

One of the major causes is temperature, i.e. the day-night or seasonal heating and cooling of the fibre. Viljoen et al showed that “while the environmental temperature is changing the PMD also changes, while for constant temperature stages, the PMD is more constant. This can be attributed to the thermal expansion/contraction of the fibre.” [10]. Although much fibre

is installed under ground and very often within conduits, it still experiences temperature variations and the corresponding physical stress.

A second source of mechanical stress can result from vibrations. Much fibre is installed alongside railway tracks because of the ease of installation. The vibrations from passing trains, however, can apply stress to the fibre. When fibre is not buried next to railways or highways, it is usually installed aerially, in which case the swaying of the fibre from the wind will affect the PMD.

## 2.4 Mode Coupling

Birefringence varies along a fibre's length due to its different sections experiencing different levels of birefringence in terms of the reasons given in section 2.3. The relative orientation of the slow and fast axes of the fibre also varies along different sections of a fibre's length. Because of these reasons, a fibre's PMD can not be approximated to the PMD of one section of the fibre. In order to understand the nature of PMD, one should represent a fibre as a series of concatenated birefringent sections of fibre separated by *mode coupling sites*. Mode coupling sites are locations where the birefringent axes of one section are rotated with respect to the other. Figure 2.3 shows a) an example of a concatenated series of birefringent fibres, while b) shows the energy interchanges which occur between each set of birefringent sections.

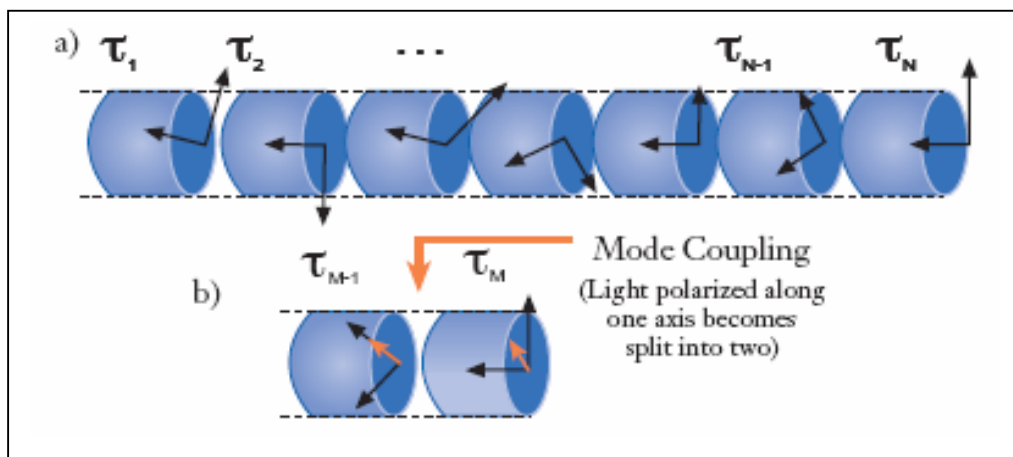
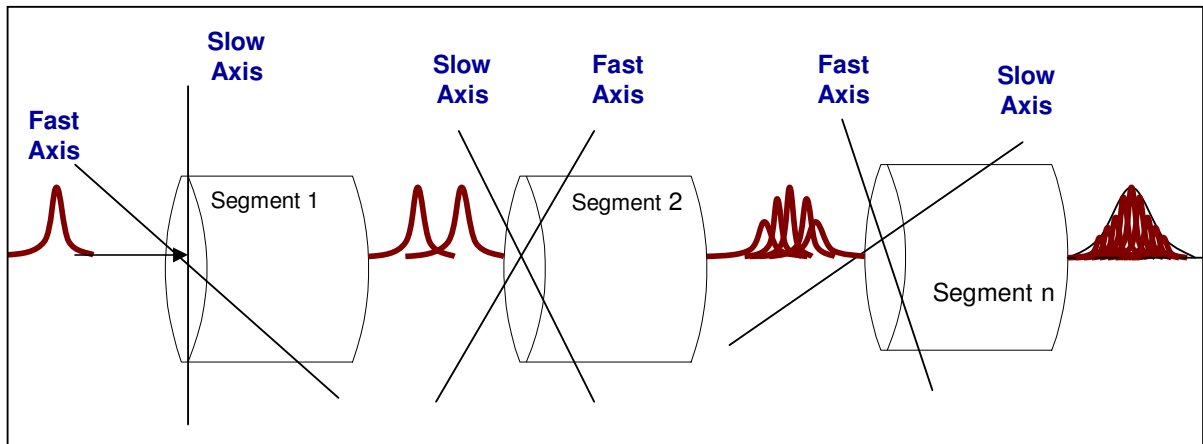


Figure 2.3 - Physical model of mode coupling



Each fibre section is randomly rotated relative to each other due to the manufacturing process or environmental effects. At each mode coupling site the relative alignment of axes of the birefringent sections can drift with respect to each other over time depending on external factors (e.g. temperature changes or external vibrations of the fibre) [9]. The mode coupling sites are also randomly distributed along each fibre [10].

As a result of mode coupling, the PMD of one section of a fibre may add or subtract from the PMD of the following section. The fast and the slow axes of the fibre segments along the path are aligned with each other differently for each wavelength. When the alignment is higher, the segments tend to add together, causing the DGD at each wavelength to increase. When the axes are not aligned the DGD decreases – the pulse travelling along the fast axis will, in this case, slow down at the mode coupling site and the pulse travelling along the slow axis will speed up, thus reducing the pulse gap between the fast and slow axes. The DGD grows roughly proportionally to the square root of the length of the fibre [11].



**Figure 2.4 - Mode coupling and pulse spread**

Figure 2.4 shows that as a single pulse (left of diagram) travels along the fibre length it will expand and contract between segments, depending on axis alignment, until it reaches the end of the fibre where it will appear to be “stretched out” (right of diagram). Mode coupling has been shown to decrease the mean PMD of an optical fibre [12] but increase the variability of the PMD [13].

## 2.5 DGD & PMD Distribution

If one were to measure the instantaneous value of DGD at the output of a fibre repetitively, the observed DGD would be a random value [9]. The DGD of an optical fibre usually varies with wavelength. It is generally accepted in the literature that at a given wavelength, the distribution of DGD values can be approximated by a Maxwellian probability density. Figure 2.5 contains an example of a histogram of repetitive DGD measurements as well as the curve corresponding to the Maxwellian distribution given by the formula above the curve, where  $x$  is the DGD and  $y$  is the average DGD or PMD.

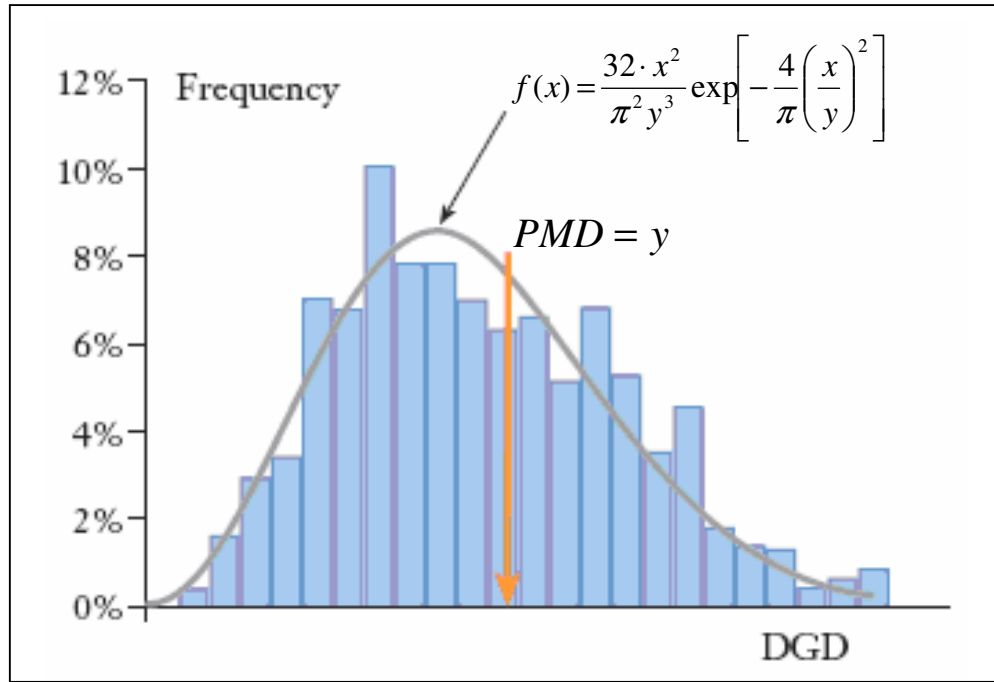


Figure 2.5 - Maxwellian distribution example

In Figure 2.5, The X-axis represents the instantaneous measured value of DGD and the Y-axis shows the frequency of the measured value.

The above probability distribution can be used to calculate the probability of a fibre experiencing a DGD value higher than a specific level, also known as an outage probability. The probability  $\alpha$  to obtain a DGD higher than a level called  $DGD_{\max}$  for a given wavelength is equal to:

$$\alpha = \int_{DGD_{\max}}^{\infty} \frac{32 x^2}{\pi^2 y^3} \exp\left[-\frac{4}{\pi} \left(\frac{x}{y}\right)^2\right] dx$$

The fact that knowledge of the PMD value of a fibre is required in order to calculate probabilities of the DGD exceeding specific levels shows how important PMD estimation is. PMD for each fibre is distributed normally [14], a property used in section 3.3.1. The histogram of PMD values can be plotted during data collection and used as a visual estimate in order to decide whether enough data has been collected.

## **2.6 Time and Wavelength Ensembles**

As the DGD of a fibre changes with time, because of environmental effects like temperature and vibrations, the DGD spectrum measured on one day may not be the same as that measured at a later time. Generally, any measured DGD value could occur at any other wavelength at another time [8].

The importance of a statistical analysis of PMD is underlined by the observed large fluctuations of the DGD with small wavelength drifts. A good statistical analysis of a fibre requires long wavelength scans, or long time averages [15]. However, the statistical nature of the DGD and the fact that any measurement involves only a finite sample set implies that there is an inherent limitation to the accuracy of PMD measurements. This limitation is independent of any measurement setup implementation, and is unavoidable, similarly to the well-known example of the calculation of the probability of a person obtaining a head from a coin toss with only a finite number of trials.

In practice, one is interested in the range of all possible values of the PMD for a fibre in its various physical states. This can either be estimated by collecting wavelength ensembles (averaging the DGD over a bandwidth range) or time ensembles (averaging the DGD over time). However, wavelength ensembles are much easier and faster to produce. Consequently, several major PMD measurement techniques are established on a wavelength ensemble. Note that large wavelength ranges correspond to long time fluctuations [6].

It has been found that the necessary wavelength ensemble over which to measure depends on the PMD value itself [8]. Since the required range is difficult to know before measuring, it is helpful to use a wide range.

When a fibre under test is exposed to a time-varying environment (such as in the case of installed cables), the measured PMD may also change in time as it samples its ensemble space or possible DGD values. Averaging the measured DGD over bandwidth as well as over time, in order to allow one to measure more of the fibre's distinct states, is called a "long term" measurement" [13]. These "long term" measurements were collected for the analysis of actual PMD data in sections 4.5 and 4.6.

## **2.7 Polarization Scramblers**

This time-averaging is usually impractical because the DGD may only change on a time-scale of days or slower. This means that a tolerable measurement duration usually does not give a sufficient sample of the possible DGD behaviour. So in practice, time averaging means measuring the test fibre once, and then disturbing the mode coupling geometry of the fibre, thereby increasing the number of possible states of the fibre, and then the PMD is measured again. Apart from physically disturbing the fibre, one way in which to 'randomise' the state of polarization of the fibre is through the use of a polarization scrambler.

Polarization scramblers, which are usually attached to either the transmitter or receiver part of a fibre, actively change the state of polarization of each pulse travelling through a fibre. This allows one to measure the various possible PMD values of a fibre much more quickly, and therefore allowing for a more efficient and accurate PMD estimation.

Chapter 4 will show to what extent a scrambler helps in generating PMD values which are not serially correlated.

### 3 Statistical Concepts

The analysis of experimental data that have been observed at different points in time leads to new and unique problems in statistical modelling and inference. The obvious correlation introduced by the sampling of adjacent points in time can severely restrict the applicability of the many conventional statistical methods traditionally dependent on the assumption that the observations are independent and identically distributed. The systematic approach by which one goes about answering the mathematical and statistical questions posed by these time correlations is commonly referred to as time series analysis.

In this chapter, all statistical methods used in subsequent chapters will be described and defined.

#### 3.1 Basic Definitions

##### 3.1.1 Stationarity

The foundation of time series analysis is stationarity. A time series,  $X$ , is said to be *strictly stationary* if the probabilistic behaviour of every collection of values

$$\{x_1, x_2, \dots, x_k\}$$

is identical to that of the time shifted set

$$\{x_{1+h}, x_{2+h}, \dots, x_{k+h}\}$$

That is,

$$P\{x_1 \leq c_1, \dots, x_k \leq c_k\} = P\{x_{1+h} \leq c_1, \dots, x_{k+h} \leq c_k\}$$

for all  $k = 1, 2, \dots$ , all numbers  $c_1, c_2, \dots, c_k$  and all time shifts  $h = 0, \pm 1, \pm 2, \dots$ . This is a very strong condition that is hard to verify empirically. Moreover, it is difficult to assess strict stationarity from a single data set. A weaker version of stationarity, called *covariance stationarity*, is often assumed.

More specifically  $X$  is weakly stationary if:

- (i)  $E(X) = \mu$ , which is constant, and
- (ii)  $Cov(x_t, x_{t-k}) = \gamma_k$  depends only on  $k$ .

In practice, suppose the dataset has  $T$  data points  $\{x_t | t = 1, \dots, T\}$ . The weak stationarity implies that the time plot of the data would show that the  $T$  values fluctuate with constant variation around a constant level. For the remainder of this dissertation, unless otherwise stated, all references to “stationary” will mean “weakly stationary”.

Implicitly in the condition of weak stationarity, it is assumed that the first two moments of  $X$  are finite. From the definitions, if  $X$  is strictly stationary and its first two moments are finite, then  $X$  is also weakly stationary. The converse is not true in general. However, if the time series  $X$  is normally distributed, then weak stationarity is equivalent to strict stationarity. In chapter 4, the actual PMD data sets will be tested for weak stationarity.

### 3.1.2 Autocovariance

The covariance  $\gamma_k = Cov(x_t, x_{t-k}) = E[(x_t - \mu_t)(x_{t-k} - \mu_{t-k})]$  is called the lag  $k$  autocovariance of  $x_t$ . The autocovariance measures the linear dependence between two points on the same series observed at different times. Very smooth series exhibit autocovariance functions that stay large even when  $k$  is large, whereas choppy series tend to have autocovariance functions that are nearly zero for large separations.

In practice, the theoretical autocovariance is estimated from the sample autocovariance defined as:

$$\hat{\gamma}_k = \frac{1}{n} \sum_{t=1}^{n-k} (x_{t+k} - \bar{x})(x_t - \bar{x}), \quad (3.1.2.1)$$

It has two important properties:

(a)  $\gamma_0 = Var(X)$  and

(b)  $\gamma_k = \gamma_{-k}$

It is clear that, for  $k=0$ , the autocovariance reduces to the (assumed finite) variance, because  $\gamma_0 = E[(x_t - \mu_t)^2]$ . The second property holds because:

$$\begin{aligned} \gamma_k &= \gamma_{t+k-t} \\ &= E[(x_{t+k} - \mu)(x_t - \mu)] \\ &= E[(x_t - \mu)(x_{t+k} - \mu)] \\ &= \gamma_{t-(t+k)} \\ &= \gamma_{-k} \end{aligned}$$

The set of autocovariance values  $\gamma_0$  to  $\gamma_k$  is called the autocovariance function (ACVF) at lag  $k$ .

### 3.1.3 Correlation

The correlation coefficient between two random variables  $X$  and  $Y$  is defined as:

$$\rho_{x,y} = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{E[(X - \mu_x)(Y - \mu_y)]}{\sqrt{E(X - \mu_x)^2 E(Y - \mu_y)^2}}$$

Where  $\mu_x$  and  $\mu_y$  are the means of  $X$  and  $Y$  respectively, and it is assumed that the variances exist. It can be shown that  $-1 \leq \rho_{x,y} \leq 1$  and  $\rho_{x,y} = \rho_{y,x}$ . The two random variables are uncorrelated if  $\rho_{x,y} = 0$ . In addition, if both  $X$  and  $Y$  are normal random variables, then  $\rho_{x,y} = 0$  if and only if  $X$  and  $Y$  are independent. When the sample  $\{(x_t, y_t)\}_{t=1}^T$  is available, the correlation can be consistently estimated by its sample counterpart.

$$\hat{\rho}_{x,y} = \frac{\sum_{t=1}^T (x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^T (x_t - \bar{x})^2 \sum_{t=1}^T (y_t - \bar{y})^2}}$$

where  $\bar{x} = \sum_{t=1}^T x_t / T$  and  $\bar{y} = \sum_{t=1}^T y_t / T$  are the sample means of  $X$  and  $Y$ , respectively.

### 3.1.4 Autocorrelation

Consider a weakly stationary time series  $X$ . The correlation coefficient between  $x_t$  and  $x_{t-k}$  is called the lag  $k$  autocorrelation of  $x_t$  and is commonly denoted by  $\rho_k$ , which under the weak stationarity assumption is a function of  $k$  only. Specifically, it is defined:

$$\rho_k = \frac{\text{Cov}(x_t, x_{t-k})}{\sqrt{\text{Var}(x_t)\text{Var}(x_{t-k})}} = \frac{\text{Cov}(x_t, x_{t-k})}{\text{Var}(x_t)} = \frac{\gamma_k}{\gamma_0}$$

where the property  $Var(x_t) = Var(x_{t-k})$  for a weakly stationary series is used. From the definition,  $\rho_0 = 1$ ,  $\rho_k = \rho_{-k}$  and  $-1 \leq \rho_k \leq 1$ . In addition, a weakly stationary series  $x_t$  is not serially correlated if and only if  $\rho_k = 0$  for all  $k > 0$ .

The set of standardized autocovariance coefficients, or autocorrelation coefficients, constitutes the autocorrelation function (ACF). The set of autocorrelation coefficients  $\rho_0$  to  $\rho_k$  is called the autocorrelation function at lag  $k$ . To be able to detect autocorrelation in a time series, the ACF can be plotted against the lag variable  $k$ . The graph of the ACF is called a correlogram, which will be used to investigate the structures of the data sets.

In Chapter 4, various methods for constructing confidence intervals for the means of time series data will be tested and compared. For this, various time series data sets based on different models will be generated. These models will now be introduced.

## 3.2 ARMA Models

The time series models in sections 3.2.1 – 3.2.3 make use of a concept called Gaussian white noise. White noise is a (univariate or multivariate) discrete-time stochastic process whose terms are independent and identically distributed (i.i.d.), all with zero mean. The i.i.d. requirement is often too restrictive for applications and is typically replaced with a requirement that terms have constant second moments, zero autocorrelations and zero means. This is:

A one-dimensional stochastic process  $\dots, w_{t-2}, w_{t-1}, w_t, w_{t+1}, w_{t+2}, \dots$  is said to be white noise if:

$$\begin{aligned} E(w_t) &= 0 \\ V(w_t) &= \sigma^2 \\ \rho_n &= 0 \end{aligned}$$

for some constant  $\sigma$  and any integer  $n$ . An independent white noise whose terms are all normally distributed is called a Gaussian white noise.



### 3.2.1 AR Models

Autoregressive models are based on the idea that the current value of the time series,  $x_t$ , can be explained as a function of  $p$  past values,  $x_{t-1}, x_{t-2}, \dots, x_{t-p}$ , where  $p$  determines the number of steps into the past needed to forecast the current value. An autoregressive model of order  $p$ , abbreviated AR( $p$ ), is of the form:

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + w_t \quad (3.2.1.1)$$

Where  $x_t$  is stationary,  $\phi_1, \phi_2, \dots, \phi_p$  are parameters, ( $\phi_p \neq 0$ ) and  $w_t$  is a Gaussian white noise series with mean zero and variance  $\sigma_w^2$ . The mean of  $x_t$  in (3.2.1.1) is zero. If the mean,  $\mu$ , of  $x_t$  is not zero, replace  $x_t$  by  $x_t - \mu$ , i.e.,

$$x_t - \mu = \phi_1 (x_{t-1} - \mu) + \phi_2 (x_{t-2} - \mu) + \dots + \phi_p (x_{t-p} - \mu) + w_t$$

or write

$$x_t = \alpha + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + w_t$$

where  $\alpha = \mu(1 - \phi_1 - \dots - \phi_p)$

### 3.2.2 MA Models

The moving average model of order  $q$ , or MA( $q$ ) model, is defined to be:

$$x_t = \alpha + w_t - \theta_1 w_{t-1} - \theta_2 w_{t-2} - \dots - \theta_q w_{t-q}$$

where there are  $q$  lags in the moving average,  $\alpha$  is a constant,  $\theta_1, \theta_2, \dots, \theta_q$  are parameters, ( $\theta_p \neq 0$ ) and again  $w_t$  is assumed to be a Gaussian white noise.

### 3.2.3 ARMA Models

A time series is ARMA( $p, q$ ) if it is stationary and

$$x_t = \phi_1 x_{t-1} + \dots + \phi_p x_{t-p} + w_t - \theta_1 w_{t-1} - \dots - \theta_q w_{t-q}$$

with  $\phi_p \neq 0$ ,  $\theta_q \neq 0$ . The parameters  $p$  and  $q$  are called the autoregressive and the moving average orders, respectively. If  $x_t$  has a non-zero mean,  $\mu$ , set  $\alpha = \mu(1 - \phi_1 - \dots - \phi_p)$  and write the model as

$$x_t = \alpha + \phi_1 x_{t-1} + \dots + \phi_p x_{t-p} + w_t - \theta_1 w_{t-1} - \dots - \theta_q w_{t-q}$$

When  $q = 0$  and  $p \neq 0$ , the model is called an autoregressive model of order  $p$ , AR( $p$ ), and when  $p = 0$  and  $q \neq 0$ , the model is called a moving average model or order  $q$ , MA( $q$ ).

## 3.3 Confidence Interval Construction Methods

In Chapter 4, two methods of constructing confidence intervals will be tested against each other in order to determine the best method for constructing confidence intervals for certain types of time series data. Three versions of each of the methods will be tested. Once the best method has been established, it will then be applied to the PMD data sets to determine the precision of the estimates.

### 3.3.1 Autocovariance Methods

The first two variations of this method were suggested in Box and Jenkins [16]. The third variation was employed in an attempt to improve upon the Autocovariance Quarter-matrix method (3.3.1.2).

#### 3.3.1.1 Autocovariance Matrix Method

Let the time series,  $X$ , have a variance-covariance matrix of the form:

$$V = \begin{pmatrix} \text{var}(X_1) & \text{cov}(X_1, X_2) & \cdots & \text{cov}(X_1, X_n) \\ \text{cov}(X_2, X_1) & \text{var}(X_2) & \cdots & \text{cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(X_n, X_1) & \text{cov}(X_n, X_2) & \cdots & \text{var}(X_n) \end{pmatrix}$$

Then if  $X$  is a stationary process then the above simplifies to:

$$V = \begin{pmatrix} \gamma_0 & \gamma_{-1} & \cdots & \gamma_{-n+1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{-n+2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{n-1} & \gamma_{n-2} & \cdots & \gamma_0 \end{pmatrix}$$

The moment estimator of the mean  $\mu$  of a stationary process is the sample mean:

$$\bar{X} = \underline{b}^T \underline{X}$$

where

$$\underline{b} = \left( \frac{1}{n}, \dots, \frac{1}{n} \right)^T,$$

$$\underline{X} = (X_1, \dots, X_n)^T$$

The mean square error of  $\bar{X}$  is

$$E(\bar{X} - \mu)^2 = \text{var}(\bar{X})$$

$$= \text{var}(\underline{b}^T \underline{X})$$

$$= \underline{b}^T V \underline{b}$$

$$= \frac{1}{n^2} (1, \dots, 1) \begin{pmatrix} \gamma_0 & \gamma_{-1} & \cdots & \gamma_{-n+1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{-n+2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{n-1} & \gamma_{n-2} & \cdots & \gamma_0 \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

All  $\gamma_k$  in the above matrix sum will be estimated by  $\hat{\gamma}_k$  as defined in (3.1.2.1). If  $\underline{X}$  is

Gaussian, then  $\bar{X}$  is normal:  $\bar{X} \sim N(\mu, \text{var}(\bar{X}))$ .

Then a confidence interval for  $\mu$  can be calculated as:

$$P(-\mu_\alpha < \frac{\bar{X} - \mu}{\sqrt{\text{Var}(\bar{X})}} < \mu_\alpha) = 1 - \alpha$$

which can be rearranged to

$$P(\bar{X} - \mu_\alpha \sqrt{\text{Var}(\bar{X})} < \mu < \bar{X} + \mu_\alpha \sqrt{\text{Var}(\bar{X})}) = 1 - \alpha$$

where  $\mu_\alpha$  is such that  $P(|U| < \mu_\alpha) = 1 - \alpha$  and  $U = \frac{\bar{X}_n - \mu}{\sqrt{\text{Var}(\bar{X}_n)}} \sim N(0,1)$ . Therefore in order

to construct say a 95% ( $\alpha = 0.05$ ) confidence interval for the mean, the limits are

$$(\bar{X}_n - 1.96\sqrt{\text{Var}(\bar{X}_n)}; \bar{X}_n + 1.96\sqrt{\text{Var}(\bar{X}_n)})$$

### 3.3.1.2 Autocovariance Quarter Matrix Method

Estimates of  $\gamma_k$  are good if  $k$  is much smaller than  $n$ , where  $n$  is the total number of observations. For  $k$  close to  $n$ , there are too few pairs  $(X_t, X_{t+k})$  for the estimate to be reliable. Box and Jenkins suggest that the dataset should have at least 50 observations and  $k \leq n/4$  [16]. In order to incorporate this idea, when calculating the sum of the variance-covariance matrix, one should sum only those values of the matrix where  $k \leq n/4$ . It is then assumed the rest of the values of the matrix are equal to 0. The shaded area in the matrix below demonstrates the above idea:

$$\begin{pmatrix} \hat{\gamma}_0 & \hat{\gamma}_1 & \hat{\gamma}_2 & \cdots & \hat{\gamma}_{n/4-1} & 0 & \cdots & 0 \\ \hat{\gamma}_1 & \hat{\gamma}_0 & \hat{\gamma}_1 & \ddots & \hat{\gamma}_{n/4-2} & \ddots & \ddots & 0 \\ \hat{\gamma}_2 & \hat{\gamma}_1 & \ddots & \ddots & \vdots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \hat{\gamma}_0 & \hat{\gamma}_1 & \ddots & \ddots & \hat{\gamma}_{n/4-1} \\ \hat{\gamma}_{n/4-1} & \hat{\gamma}_{n/4-2} & \cdots & \hat{\gamma}_1 & \hat{\gamma}_0 & \ddots & \ddots & \hat{\gamma}_{n/4-2} \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \hat{\gamma}_1 \\ 0 & 0 & 0 & \hat{\gamma}_{n/4-1} & \hat{\gamma}_{n/4-2} & \cdots & \hat{\gamma}_1 & \hat{\gamma}_0 \end{pmatrix}$$

The  $\gamma_k = \gamma_{-k}$  property mentioned in 3.1.2 is used in the matrix above. The new variance estimate therefore becomes

$$\text{var}(\bar{X}_n) = \frac{1}{n^2} \begin{pmatrix} \hat{\gamma}_0 & \hat{\gamma}_1 & \hat{\gamma}_2 & \cdots & \hat{\gamma}_{n/4-1} & 0 & \cdots & 0 \\ \hat{\gamma}_1 & \hat{\gamma}_0 & \hat{\gamma}_1 & \ddots & \hat{\gamma}_{n/4-2} & \ddots & \ddots & 0 \\ \hat{\gamma}_2 & \hat{\gamma}_1 & \ddots & \ddots & \vdots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \hat{\gamma}_0 & \hat{\gamma}_1 & \ddots & \ddots & \hat{\gamma}_{n/4-1} \\ \hat{\gamma}_{n/4-1} & \hat{\gamma}_{n/4-2} & \cdots & \hat{\gamma}_1 & \hat{\gamma}_0 & \ddots & \ddots & \hat{\gamma}_{n/4-2} \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \hat{\gamma}_1 \\ 0 & 0 & 0 & \hat{\gamma}_{n/4-1} & \hat{\gamma}_{n/4-2} & \cdots & \hat{\gamma}_1 & \hat{\gamma}_0 \end{pmatrix}$$

The confidence interval is then constructed similarly as in (3.3.1.1).

### 3.3.1.3 Autocorrelation-length Matrix Method

The ACF of a stationary time series should approach 0 as the lag increases. The standard deviation of the estimate  $\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0}$  is approximately  $\frac{1}{\sqrt{n}}$  [16].

Now for any particular value  $k$ , there is the Bartlett test of the null hypothesis

$$H_0: \rho_k = 0$$

$$H_1: \rho_k \neq 0$$

The null hypothesis is rejected at the 5% significance level if  $|\hat{\rho}_k| > \frac{1.96}{\sqrt{n}}$  [17]. This test will

be used to determine, for each time series which is simulated, at which lag,  $k$ , the serial correlation becomes insignificant. This value  $k$  will then replace  $n/4$ , suggested in section (3.3.1.2), as the cut off number of rows and columns which will be summed in the autocovariance matrix. The variance estimate therefore becomes

$$\text{var}(\bar{X}_n) = \frac{1}{n^2} \begin{pmatrix} \hat{\gamma}_0 & \hat{\gamma}_1 & \hat{\gamma}_2 & \cdots & \hat{\gamma}_{k-1} & 0 & \cdots & 0 \\ \hat{\gamma}_1 & \hat{\gamma}_0 & \hat{\gamma}_1 & \ddots & \hat{\gamma}_{k-2} & \ddots & \ddots & 0 \\ \hat{\gamma}_2 & \hat{\gamma}_1 & \ddots & \ddots & \vdots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \hat{\gamma}_0 & \hat{\gamma}_1 & \ddots & \ddots & \hat{\gamma}_{k-1} \\ \hat{\gamma}_{k-1} & \hat{\gamma}_{k-2} & \cdots & \hat{\gamma}_1 & \hat{\gamma}_0 & \ddots & \ddots & \hat{\gamma}_{k-2} \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \hat{\gamma}_1 \\ 0 & 0 & 0 & \hat{\gamma}_{k-1} & \hat{\gamma}_{k-2} & \cdots & \hat{\gamma}_1 & \hat{\gamma}_0 \end{pmatrix}$$

where  $|\hat{\rho}_k| < \frac{1.96}{\sqrt{n}}$ . The confidence interval is then constructed similarly as in (3.3.1.1).

### 3.3.2 Batch Means Methods

The method of non-overlapping batch means is frequently used to estimate the steady-state mean,  $\mu$ , of a time series,  $X_1, X_2, \dots, X_n$ , or the  $Var(\bar{X})$  and owes its popularity to its simplicity and effectiveness. The classic approach divides the output  $X_1, X_2, \dots, X_n$  of a time series into a number of batches of equal size and treats the sample means of these batches (a.k.a. batch means) as i.i.d. normal data points to produce point and interval estimates.

Assume that the data has been split into  $b$  batches, each consisting of  $h$  observations (for now assume  $n = bh$ ). Then the  $i$ -th batch consists of the observations

$$X_{(i-1)h+1}, X_{(i-1)h+2}, \dots, X_{ih} \quad \text{for } i = 1, 2, \dots, b$$

The  $i$ -th batch mean is denoted as

$$\bar{X}_i(h) = \frac{1}{h} \sum_{j=1}^h X_{h(i-1)+j}$$

If  $n \neq bh$ , then the first  $m$  observations are considered a “warm-up period” and are rejected.

Each batch will then contain  $h = \frac{n-m}{b}$  observations. Figure 3.1 shows how the warm-up period is treated as well as how to split the sample up.

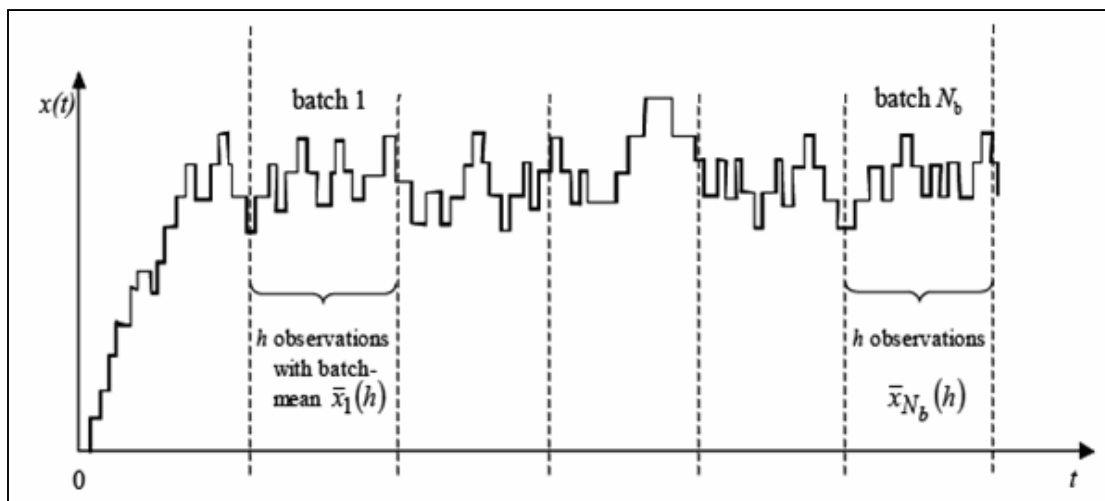


Figure 3.1 – Batch Means method example

The simulation in Figure 3.1 contains  $(hxb) + m$  observations.

The final estimator for  $\mu$  is then given by:

$$\bar{X}_b = \frac{1}{b} \sum_{i=1}^b \bar{X}_i(h) = \frac{1}{bh} \sum_{i=1}^b \sum_{j=1}^h X_{h(i-1)+j}$$

This is simply the sample average of the whole run (after the warm-up period). The variance,  $\sigma^2$ , is estimated by the variance of the  $b$  batch means:

$$\hat{V}_b(h) = \frac{1}{b-1} \sum_{i=1}^b (\bar{X}_i(h) - \bar{X}_b)^2$$

An approximate  $1 - \alpha$  confidence interval for  $\mu$  is thus computed as:

$$\bar{X}_b \pm t_{b-1} \sqrt{\hat{V}_b(h)/b}$$

The main setback with the application of the batch means method is the choice of the batch size  $h$ . If  $h$  is too small, the batch means will very likely be highly correlated and the resulting confidence interval will frequently have coverage below the desired  $1 - \alpha$ . On the contrary, a large batch size will probably result in very few batches being produced which will create potential problems with the application of the central limit theorem in order to use the confidence interval above.

Optimally, one would like as many batches as possible, to produce an accurate confidence interval, whilst still ensuring that the batch means are not highly correlated. For the scope of this project, two fixed number of batches were tested, namely 10 and 50, and the author's own suggested number of batches (see 3.3.2.3).

### 3.3.2.1 Batch Means – 10 batches (batch size = 50)

If the number of batches is fixed at 10, the batch sizes become large very quickly for long time series, but this will ensure that the batch means generated will not be highly correlated. The downside to this method is the potential problems associated with the application of the central limit theorem.

An approximate 95% confidence interval for  $\mu$  is thus computed as:

$$\bar{X}_{10} \pm t_9 \sqrt{\hat{V}_{10}(h)/10}$$

### 3.3.2.2 Batch Means – 50 batches (batch size = 10)

If the number of batches is fixed at 50, the batch size remains small until the time series becomes very long (e.g. the batch size will be 10 if there are 500 observations). As mentioned in section 3.3.2, the potential problem with this method is that the batch means will be highly correlated if the number of observations is not large enough.

An approximate 95% confidence interval for  $\mu$  for this method is thus computed as:

$$\bar{X}_{50} \pm t_{49} \sqrt{\hat{V}_{50}(h)/50}$$

### 3.3.2.3 Batch Means – Autocorrelation-length batch size

This method was created in an attempt to minimize the problems of the previous two fixed number of batches approaches. Similar to the method described in section 3.3.1.3., for each time series generated, Bartlett's Test was used to determine at which lag,  $k$ , the serial correlation becomes insignificant. This value  $k$  was then used as the new batch size, implying that there were  $n/k$  batches.

An approximate 95% confidence interval for  $\mu$  for this method is thus computed as:

$$\bar{X}_{n/k} \pm t_{n/k-1} \sqrt{\hat{V}_{n/k}(h)/(n/k)}$$

This method should generate a large enough number of batches in order to apply the central limit theorem, whilst still keeping the batch means themselves as independent as possible.



### 3.4 Test for Stationarity

In time series models, a linear stochastic process is said to have a *unit root* if 1 is a root of the process's characteristic equation. Consider a discrete time stochastic process  $X$  and suppose it can be written as an AR( $p$ ) process:

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + w_t$$

For convenience, assume  $x_0 = 0$ . If  $m=1$  is a root of the characteristic equation:

$$m^p - m^{p-1}\phi_1 - m^{p-2}\phi_2 - \dots - \phi_p = 0$$

then the stochastic process has a unit root. If the process has a unit root, then it is a non-stationary time series.

The *Dickey-Fuller test* tests whether a unit root exists in an autoregressive model. It is named after the statisticians D. A. Dickey and W. A. Fuller. Suppose one has an AR(1) model:

$$x_t = \phi_1 x_{t-1} + w_t.$$

A unit root is present if  $|\phi_1| = 1$ . The model would be non-stationary in this case. Moreover, it would be even more non-stationary if  $|\phi_1| > 1$ . The model can be rewritten as:

$$\Delta x_t = (\phi_1 - 1)x_{t-1} + w_t = \delta x_{t-1} + w_t$$

where  $\Delta$  is the first difference operator. Testing for a unit root is equivalent to testing  $\delta = \phi_1 - 1 = 0$  (null hypothesis). Since this test is performed over the residual term instead of the raw data, one cannot use the t-distribution to find critical values. Therefore this statistic,  $\tau$ , has a specific distribution simply known as the Dickey-Fuller table.

Dickey & Fuller also created an extension to their test called the *augmented Dickey-Fuller test (ADF)* which removes all autocorrelation in the time series and then tests using the same procedure as above. The ADF statistic used in the test is a negative number. The more negative the statistic, the stronger the rejection of the null hypothesis (i.e. that there exists a unit root) at some level of confidence.

## 4 Results

### 4.1 Overview

The first part of this chapter will contain the results from testing various methods of constructing confidence intervals for the means of time series data, which were simulated. The methods employed to simulate the time series data will be shown. An example of how each of the confidence interval construction methods was calculated will be demonstrated, along with the results of the simulations.

In the second part of the chapter, the most accurate confidence interval construction methods will be applied to actual time series data, i.e. PMD runs taken from the field. These will be compared to “standard” methods of confidence interval construction.

In the final part of this chapter, the effectiveness of using a PMD scrambler, when constructing confidence intervals, will be investigated.

### 4.2 Simulation Example

In order to determine which of the tested confidence interval construction methods works best for time series data, five time series models were simulated, namely: AR(1), AR(2), MA(1), MA(2), ARMA(1,1). All of these models were defined in Chapter 3. A dataset taken from the normal distribution was also simulated in order to see how the methods would work on independent data, i.e. data which are not serially correlated.

For each of the above six time series models, seven different methods of calculating the datasets' standard errors were utilized. These were:

1. “Standard” Standard Error Method
2. Autocovariance Matrix Method
3. Autocovariance Quarter Matrix Method
4. Autocorrelation-Length Matrix Method
5. Batch Means (10 batches of size 50) Method
6. Batch Means (50 batches of size 10) Method
7. Batch Means (Autocorrelation-length batch size) Method

The details of these methods were described in Chapter 3. One run of the simulation will now be shown to explain how the confidence intervals from the above methods were calculated.

To keep all the models comparable, each model was modelled to have  $E(X) = 5$  (true mean) with a  $V(X) = 0.1$ . This is what a typical PMD dataset would look like. For the simulation example an AR(1) time series model was used to generate the dataset.

For this model,  $\phi_1 = 0.8$ . This value was set arbitrarily for purposes of testing these methods, but was chosen as the dataset simulated from this value looks the most like a PMD dataset. Similarly, all parameters for the other models were set in order to most closely replicate a PMD dataset.

Using the formula below, from Chapter 3,

$$E[x_t] = \frac{\alpha}{1 - \phi_1}$$

one can solve for  $\alpha$  as  $\alpha = E[x_t](1 - \phi_1) = 5(1 - 0.8) = 1$ . So the AR(1) model will look as follows:

$$x_t = 0.8x_{t-1} + 1 + w_t$$

where  $w_t$  is a Gaussian white noise series.

#### 4.2.1 Autocovariance Methods

The  $\gamma_k = \gamma_{-k}$  property of autocovariance simplifies the summation of the autocovariance matrix for all of the autocovariance matrix methods. As discussed in Chapter 3, the matrix looks as follows:

$$V = \begin{pmatrix} \gamma_0 & \gamma_{-1} & \cdots & \gamma_{-n+1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{-n+2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{n-1} & \gamma_{n-2} & \cdots & \gamma_0 \end{pmatrix}$$

Because of the symmetry property, the matrix above can be rewritten as follows:

$$V = \begin{pmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{n-1} \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{n-2} \\ \gamma_2 & \gamma_1 & \gamma_0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \gamma_1 \\ \gamma_{n-1} & \gamma_{n-2} & \cdots & \gamma_1 & \gamma_0 \end{pmatrix}$$

As the matrix is perfectly symmetrical (around both diagonal axes), it is necessary to obtain only the first column (or row) of the matrix in order to calculate its sum. The only remaining information required is to calculate the number of times each autocovariance appears in the matrix. It is then simply a matter of multiplying each autocovariance by the number of times it appears in the matrix, and summing these terms, to obtain the matrix sum.

Table 4.1 contains the information which was used in the Excel simulation for one run (number of observations = 500) - a randomly generated AR(1) series, together with its autocovariance and autocorrelation values at each respective lag. The “# Times” columns refer to how many times each covariance appears in that respective method’s autocovariance matrix.

The three autocovariance methods will be investigated, followed by the three batch means methods.

AR(1)	Lag	Autocovariance	Autocorrelation	# Times in Matrix	# Times in ¼ Matrix	# Times in Auto-Matrix
5.000	0	0.21698507	1.0000	500	500	500
5.297	1	0.16151078	0.7443	998	998	998
5.027	2	0.11426503	0.5266	996	996	996
4.920	3	0.08591486	0.3959	994	994	994
5.198	4	0.05912268	0.2725	992	992	992
5.334	5	0.04135956	0.1906	990	990	990
5.114	6	0.03275974	0.1510	988	988	988
4.915	7	0.02931672	0.1351	986	986	986
4.333	8	0.02640783	0.1217	984	984	984
4.279	9	0.02202475	0.1015	982	982	982
4.188	10	0.02288795	0.1055	980	980	980
4.338	11	0.02132505	0.0983	978	978	978
4.299	12	0.01671380	0.0770	976	976	976
...	...	...	...	...	...	0
...	...	...	...	...	...	...
4.294	120	-0.03244482	-0.1495	760	760	0
4.279	121	-0.03082052	-0.1420	758	758	0
4.182	122	-0.02183413	-0.1006	756	756	0
4.439	123	-0.01304580	-0.0601	754	754	0
4.433	124	-0.01246330	-0.0574	752	752	0
5.056	125	-0.01392339	-0.0642	750	0	0
...	...	...	...	...	...	...
...	...	...	...	...	...	...
4.818	495	0.00041289	0.0019	10	0	0
5.109	496	0.00047568	0.0022	8	0	0
5.271	497	0.00064921	0.0030	6	0	0
5.572	498	0.00035834	0.0017	4	0	0
5.220	499	0.00005555	0.0003	2	0	0

Table 4.1: Autocovariance matrix methods calculations

#### 4.2.1.1 Autocovariance Matrix Method

To illustrate the above sum product idea, the variance for this method would be calculated as:

$$\begin{aligned}
 \text{Var}(\bar{X}_n) &= \frac{1}{n^2} \begin{pmatrix} \hat{\gamma}_0 & \hat{\gamma}_1 & \cdots & \hat{\gamma}_{n-1} \\ \hat{\gamma}_1 & \hat{\gamma}_0 & \cdots & \hat{\gamma}_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\gamma}_{n-1} & \hat{\gamma}_{n-2} & \cdots & \hat{\gamma}_0 \end{pmatrix} \\
 &= \frac{1}{n^2} (0.21699 \times 500 + 0.16151 \times 998 + \dots + 0.000358 \times 4 + 0.0000556 \times 2) \\
 &= \frac{1}{500^2} (278.81) \\
 &= 0.001115
 \end{aligned}$$

Therefore a 95% confidence interval for the mean would be:

$$\begin{aligned} & [\bar{X} - t_{\alpha,499} \sqrt{\text{Var}(\bar{X})}; \bar{X} + t_{\alpha,499} \sqrt{\text{Var}(\bar{X})}] \\ & [4.910 - 1.9647\sqrt{0.001115}; 4.910 + 1.9647\sqrt{0.001115}] \\ & [4.844; 4.976] \end{aligned}$$

During the simulation, after each confidence interval is constructed for each run, a check is done to ascertain whether the true mean of the series – in this case, five – falls within the constructed confidence interval or not. If this procedure is repeated a large number of times – 10,000 runs for each time series model – and the ratio of “hits” (when true mean falls within interval) is compared to the actual desired confidence level (e.g. 50% and 95% were tested), it will be possible to see which of the various methods work best for constructing confidence intervals for each time series model.

Accordingly, for the above run of the simulation, a zero would be captured for the “Mean in Interval” variable, as the true mean, five, does not fall into the interval [4.844; 4.976].

#### 4.2.1.2 Autocovariance Quarter Matrix Method

The variance for this method is calculated as:

$$\begin{aligned} \text{var}(\bar{X}) &= \frac{1}{n^2} \begin{pmatrix} \hat{\gamma}_0 & \hat{\gamma}_1 & \hat{\gamma}_2 & \dots & \hat{\gamma}_{n/4-1} & 0 & \dots & 0 \\ \hat{\gamma}_1 & \hat{\gamma}_0 & \hat{\gamma}_1 & \ddots & \hat{\gamma}_{n/4-2} & \ddots & \ddots & 0 \\ \hat{\gamma}_2 & \hat{\gamma}_1 & \ddots & \ddots & \vdots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \hat{\gamma}_0 & \hat{\gamma}_1 & \ddots & \ddots & \hat{\gamma}_{n/4-1} \\ \hat{\gamma}_{n/4-1} & \hat{\gamma}_{n/4-2} & \dots & \hat{\gamma}_1 & \hat{\gamma}_0 & \ddots & \ddots & \hat{\gamma}_{n/4-2} \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \hat{\gamma}_1 \\ 0 & 0 & 0 & \hat{\gamma}_{n/4-1} & \hat{\gamma}_{n/4-2} & \dots & \hat{\gamma}_1 & \hat{\gamma}_0 \end{pmatrix} \\ &= \frac{1}{500^2} (0.21699 \times 500 + 0.16151 \times 998 + \dots + (-0.01305 \times 754) + (-0.01246 \times 752)) \\ &= \frac{1}{500^2} (1091.05) \\ &= 0.00436 \end{aligned}$$

Notice the differences in the matrix summation. The 95% confidence interval for the mean would be as follows:

$$\begin{aligned}
 & [\bar{X} - t_{\alpha, 499} \sqrt{\text{Var}(\bar{X})}; \bar{X} + t_{\alpha, 499} \sqrt{\text{Var}(\bar{X})}] \\
 & [4.910 - 1.9647 \sqrt{0.00436}; 4.910 + 1.9647 \sqrt{0.00436}] \\
 & [4.780; 5.040]
 \end{aligned}$$

Here a one would be recorded for “Mean in Interval” as the interval contains five.

Unfortunately this method seems to produce negative terms during some of the matrix summation steps. These runs were excluded from the simulation and only the valid runs were used. Practically, this means that this method cannot be used when it returns a negative value for the variance calculation (i.e. the matrix summation). In section 4.3, the number of runs excluded will be stated for each simulation / method combination.

#### 4.2.1.3 Autocorrelation-length Matrix Method

As can be seen from Table 4.1, the autocorrelation drops below  $\frac{1.96}{\sqrt{n}} = 0.0877$  at lag 12

(See Bartlett’s test from section 3.3.1.3), so  $k = 12$  in the following variance formula:

$$\begin{aligned}
 \text{var}(\bar{X}) &= \frac{1}{n^2} \begin{pmatrix} \hat{\gamma}_0 & \hat{\gamma}_1 & \hat{\gamma}_2 & \cdots & \hat{\gamma}_{k-1} & 0 & \cdots & 0 \\ \hat{\gamma}_1 & \hat{\gamma}_0 & \hat{\gamma}_1 & \ddots & \hat{\gamma}_{k-2} & \ddots & \ddots & 0 \\ \hat{\gamma}_2 & \hat{\gamma}_1 & \ddots & \ddots & \vdots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \hat{\gamma}_0 & \hat{\gamma}_1 & \ddots & \ddots & \hat{\gamma}_{k-1} \\ \hat{\gamma}_{k-1} & \hat{\gamma}_{k-2} & \cdots & \hat{\gamma}_1 & \hat{\gamma}_0 & \ddots & \ddots & \hat{\gamma}_{k-2} \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \hat{\gamma}_1 \\ 0 & 0 & 0 & \hat{\gamma}_{k-1} & \hat{\gamma}_{k-2} & \cdots & \hat{\gamma}_1 & \hat{\gamma}_0 \end{pmatrix} \\
 &= \frac{1}{500^2} (0.21699 \times 500 + 0.16151 \times 998 + \dots + 0.021325 \times 978 + 0.01671 \times 976) \\
 &= \frac{1}{500^2} (736.97) \\
 &= 0.00295
 \end{aligned}$$

The 95% confidence interval for the mean would be as follows:

$$[\bar{X} - t_{\alpha; 499} \sqrt{\text{Var}(\bar{X})}; \bar{X} + t_{\alpha; 499} \sqrt{\text{Var}(\bar{X})}]$$

$$[4.910 - 1.9647\sqrt{0.00295}; 4.910 + 1.9647\sqrt{0.00295}]$$

$$[4.803; 5.017]$$

A one would be captured for “Mean in Interval” here too, as the interval contains five.

## 4.2.2 Batch Means Methods

Table 4.2 contains the same AR(1) run values as in Table 4.1, and shows the means of the three batch means methods used. This information is all that is necessary for computing the confidence intervals using the Batch Means methods.

Data Point	AR(1)	Batch Mean (10 batches)	Batch Mean (50 batches)	Batch Mean (autocorrelation-length size)
1	5.000	4.896	4.942	Warm-up Period
2	5.297			
3	5.027			
4	4.920			
5	5.198			
6	5.334			
7	5.114			
8	4.915			
9	4.333			
10	4.279			4.466
11	4.188		4.498	
12	4.338			
13	4.299			
14	4.865			
15	4.489			
16	4.469			
17	4.878			
18	4.583			
19	3.997			
20	4.872			
21	5.130		5.328	5.317
...	...	...	...	...
481	4.692		4.253	
482	4.285			
483	3.853			
484	4.527			
485	4.377			
486	3.972			
487	3.751			
488	4.256			5.113
489	4.427			
490	4.390			
491	4.857		4.927	
492	4.672			
493	4.507			
494	4.752			



495	4.488			
496	4.818			
497	5.109			
498	5.271			
499	5.572			
500	5.220			

**Table 4.2: Batch Means methods calculations**

#### 4.2.2.1 Batch Means (10 batches) Method

The batch means (from batches of size 50) are computed using the following formula from Chapter 3:

$$\bar{X}_i(50) = \frac{1}{50} \sum_{j=1}^{50} X_{50(i-1)+j}$$

And the variance is calculated from:

$$\hat{V}_{10}(50) = \frac{1}{9} \sum_{i=1}^{10} (\bar{X}_i(50) - \bar{X}_{10})^2$$

The 95% confidence interval for this method is calculated as:

$$\begin{aligned} & [\bar{X}_{10} - t_{10-1} \sqrt{\hat{V}_{10}(50)/10}; \bar{X}_{10} + t_{10-1} \sqrt{\hat{V}_{10}(50)/10}] \\ & [4.910 - 2.262\sqrt{0.042/10}; 4.910 + 2.262\sqrt{0.042/10}] \\ & [4.76388; 5.05676] \end{aligned}$$

This interval does contain five, so a one would be captured under “Mean in Interval” variable.

#### 4.2.2.2 Batch Means (50 batches) Method

The batch means are computed using the following formula from Chapter 3:

$$\bar{X}_i(10) = \frac{1}{10} \sum_{j=1}^{10} X_{10(i-1)+j}$$

And the variance is calculated from:

$$\hat{V}_{50}(10) = \frac{1}{49} \sum_{i=1}^{50} (\bar{X}_i(10) - \bar{X}_{50})^2$$

Using this method, the 95% confidence interval for the mean would be calculated as follows:

$$\begin{aligned}
& [\bar{X}_{50} - t_{50-1} \sqrt{\hat{V}_{50}(10)/50}; \bar{X}_{50} + t_{50-1} \sqrt{\hat{V}_{50}(10)/50}] \\
& [4.910 - 2.010\sqrt{0.100/50}; 4.910 + 2.010\sqrt{0.100/50}] \\
& [4.82067; 4.99997]
\end{aligned}$$

As this interval does not contain five, a zero would be captured under “Mean in Interval” variable.

#### 4.2.2.3 Batch Means (Autocorrelation-length batch size) Method

As the autocorrelation length is 12 (see 4.2.1.3) which is not a divisor of 500, a “warm-up” period of eight observations is used. The first batch, of size 12, therefore starts at observation nine. This batch size yields 41 batches for this sample.

The 95% confidence interval is constructed as:

$$\begin{aligned}
& [\bar{X}_{41} - t_{41-1} \sqrt{\hat{V}_{41}(12)/41}; \bar{X}_{41} + t_{41-1} \sqrt{\hat{V}_{41}(12)/41}] \\
& [4.907 - 2.021\sqrt{0.016/41}; 4.907 + 2.021\sqrt{0.0168/41}] \\
& [4.86730; 4.94715]
\end{aligned}$$

This interval does not contain five, so a zero would be captured under “Mean in Interval” variable.

### 4.3 Model Results

For each of the five time series models and one normal distribution model which were tested, the above simulation was run 10,000 times each using 500 observations. When generating each time series, it was assumed that each series had an expected value of 5 and a variance of 0.1. This is what a typical PMD series would look like. Using these values, the necessary parameters for each model could be calculated, as will be shown. After each simulation run, a check was made to see if the two confidence intervals constructed (50% and 95% intervals were tested) contained the true mean, 5. The number of times the interval actually did contain this mean was recorded for each time series model and each method used. A simple check could then be performed which will show the effectiveness of each method.

The reason that two intervals, 95% and 50%, were tested was because although 95% confidence intervals are commonly used in practice, if some of the methods over-estimated the size of the interval in certain time series models, it would be difficult to quantify by how much they over-estimated and compare that number to how much other models under-estimated the confidence interval size. To rectify this problem, 50% confidence intervals were also tested. Now, if a method produced intervals which contained the true mean 40% of the time, this number would be comparable to and equally as bad as a method which contained the mean 60% of the time, although for practical purposes the method which contained the mean 60% of the time should be preferred in order to remain on the conservative side.

Apart from the six confidence interval construction methods outlined above, the “standard” method used for independent data was also included:

$$[\bar{X} \pm t_{\alpha/2} \frac{S}{\sqrt{n}}]$$

in order to show its drawbacks when it comes to time series (serially correlated) datasets. For all of the models, all  $w_t$  were generated to follow a  $N(0, 0.3^2)$  distribution. This produced series which were most like an actual PMD dataset.

### 4.3.1 AR(1) Model

As mentioned in (3.2) the model for this series looks as follows:

$$x_t = 0.8x_{t-1} + 1 + w_t$$

After generating this series 10,000 times, and summing the number of times the constructed intervals contained the true mean for each method, the following results were obtained:

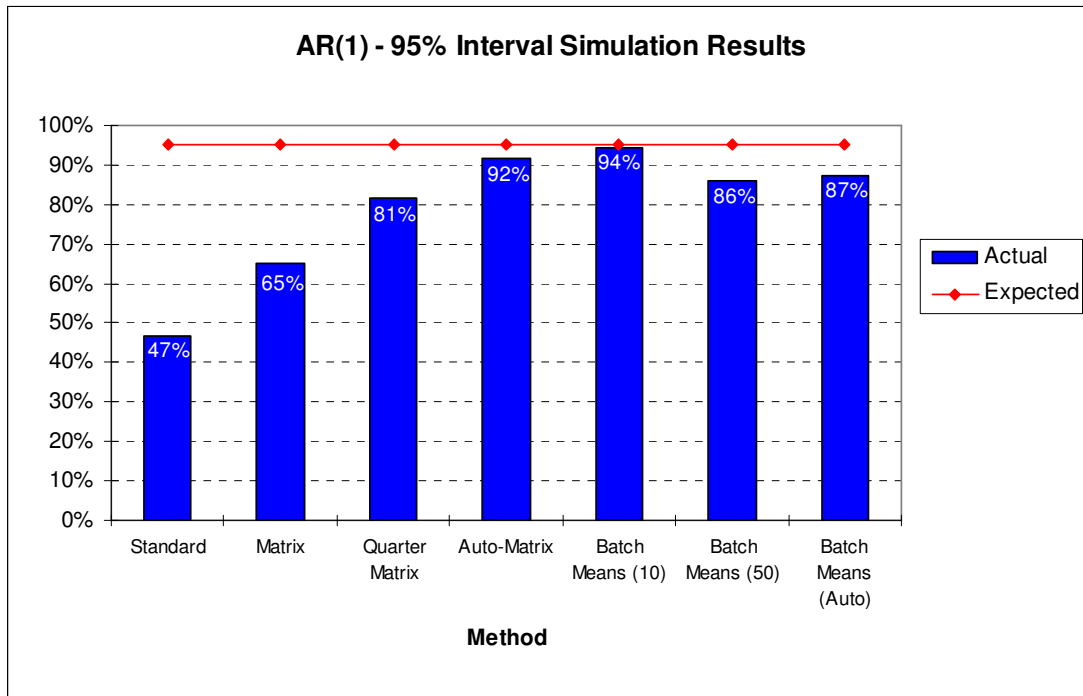


Figure 4.1 - AR(1) 95% interval simulation results

As the constructed confidence interval was supposed to be a 95% interval, one can see that some of the above methods did not perform well. The Standard method employed for independent data did particularly poorly, where only 47% of the time the interval contained the true mean. The best of the methods was the Batch Means (10 batches of size 50) method which constructed intervals containing the mean 94% of the time, followed closely by the Autocorrelation length matrix method at 92%. 737 of the 10,000 observations (7.4%) produced negative matrix sums for the Quarter-matrix method and were excluded, so the 81% recorded for this method is interpreted as 81% of the valid runs contained the true mean. This same number of invalid runs would apply for the 50% intervals in each method as well.

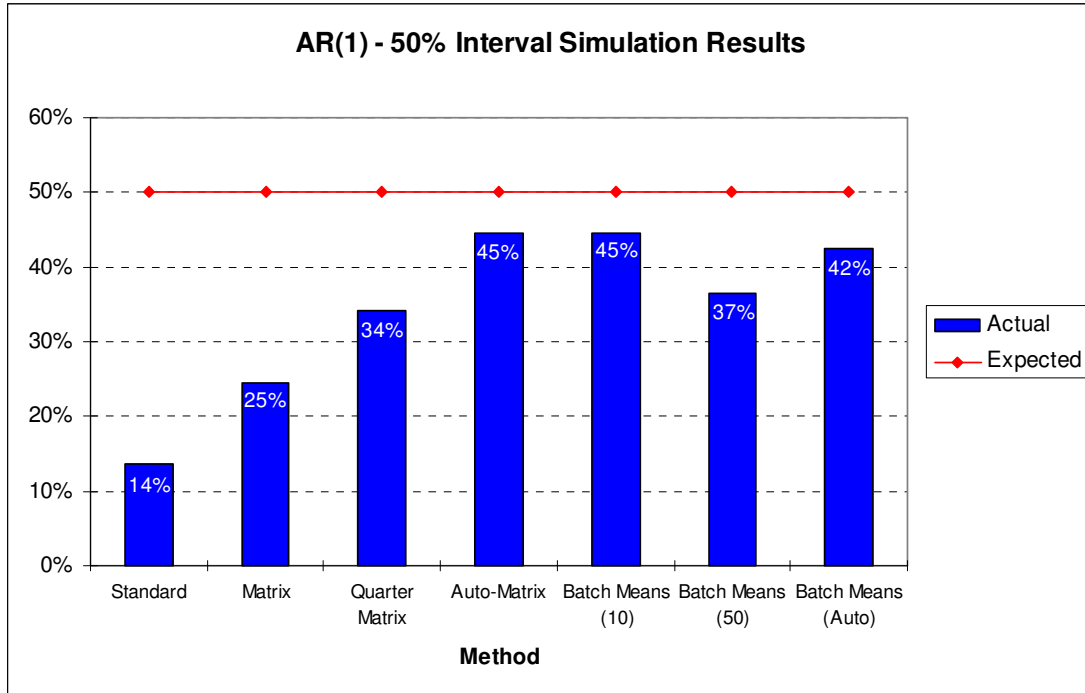


Figure 4.2 - AR(1) 50% interval simulation results

Figure 4.2 contains the 50% interval simulation results for the same time series data. As with the 95% simulation results, all of the confidence interval construction methods underestimated the correct sizes of the confidence intervals. Again, the Standard method performed badly at 14%, while the Batch Means (10 batches) method and the Autocorrelation length matrix method performed the best, coming closest to the expected 50%.

### 4.3.2 AR(2) Model

For this model,

$$E[x_t] = \frac{\alpha}{1 - \phi_1 - \phi_2}$$

Here it is assumed that  $\phi_1 = 0.6$  and  $\phi_2 = 0.2$ , then one can again solve for  $\alpha = 1$ . The time series now looks like:

$$x_t = 0.6x_{t-1} + 0.2x_{t-2} + 1 + w_t$$

Here 703 runs were invalid for the Quarter Matrix method (7.0%). The results for this model are as follows:

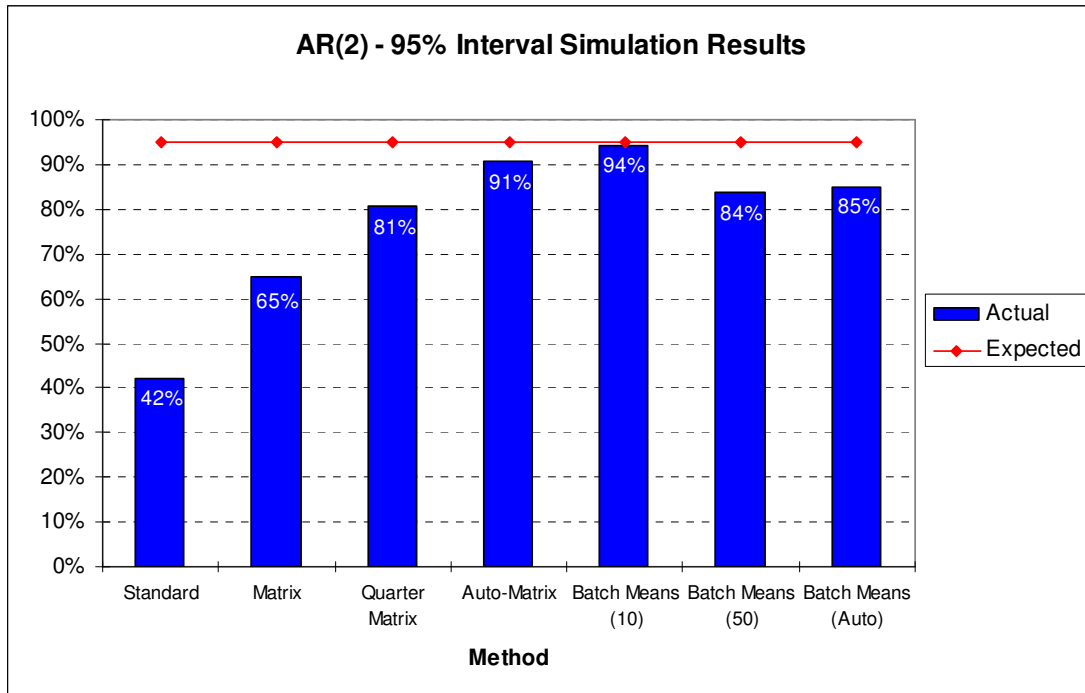


Figure 4.3 - AR(2) 95% interval simulation results

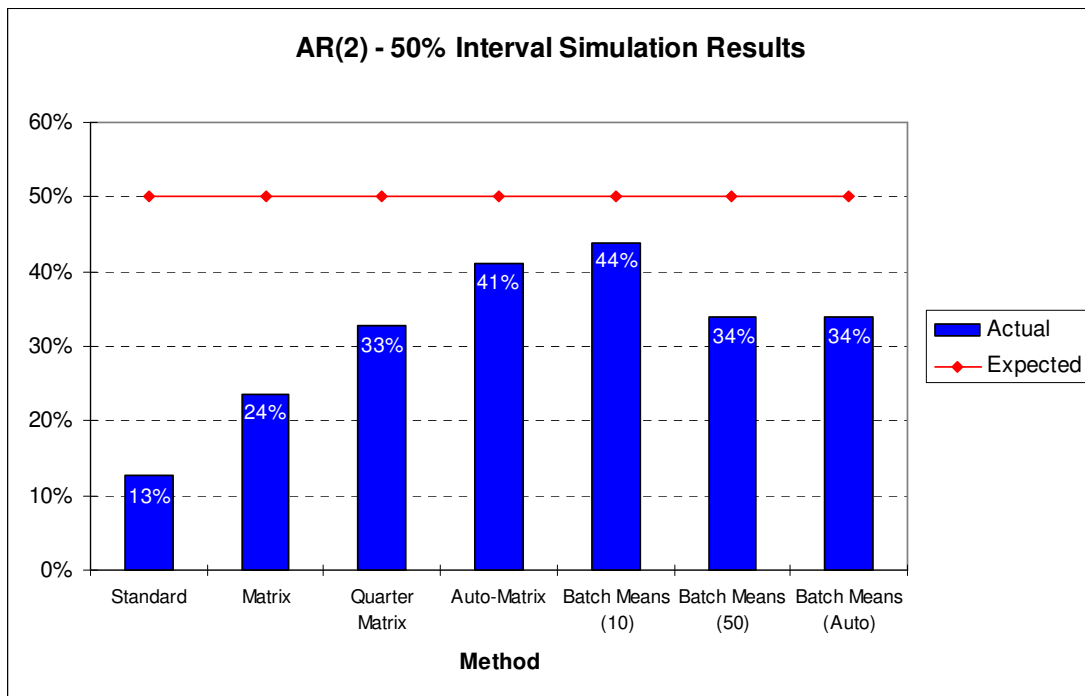


Figure 4.4 - AR(2) 50% interval simulation results

In both the 95% and 50% confidence interval simulations which were run, the Batch Means (10 batches) generated the confidence intervals which contained the true mean the closest number of times compared to the expected, followed again by the Autocorrelation matrix method.

### 4.3.3 MA(1) Model

For this model it is assumed that  $\theta_1 = 0.8$ , so the model will look as follows:

$$x_t = 5 + w_t + 0.8w_{t-1}$$

Here 772 runs were invalid for the Quarter Matrix method (7.7%). The results for this model are as follows:

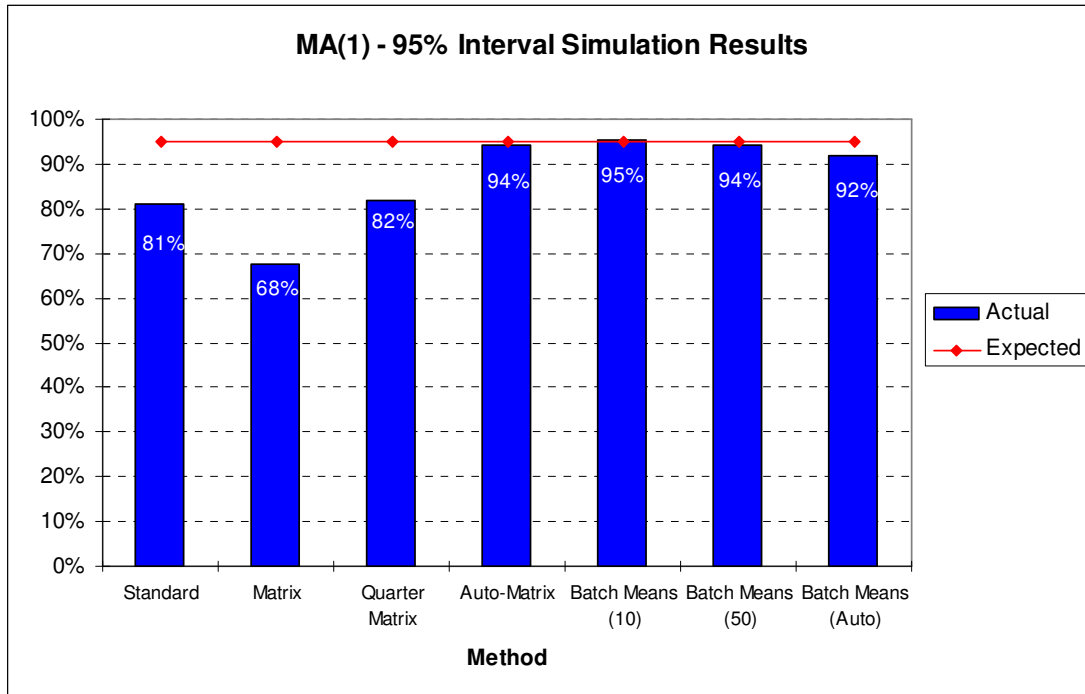


Figure 4.5 - MA(1) 95% interval simulation results

Here one can see that auto-correlation matrix and all of the batch means methods produced the best confidence intervals.

Figure 4.6 shows that the same methods performed well for the 50% confidence intervals, all producing intervals containing the true mean ranging from 42% to 48% of the time. The Matrix method performed the poorest in both the 95% and 50% results.

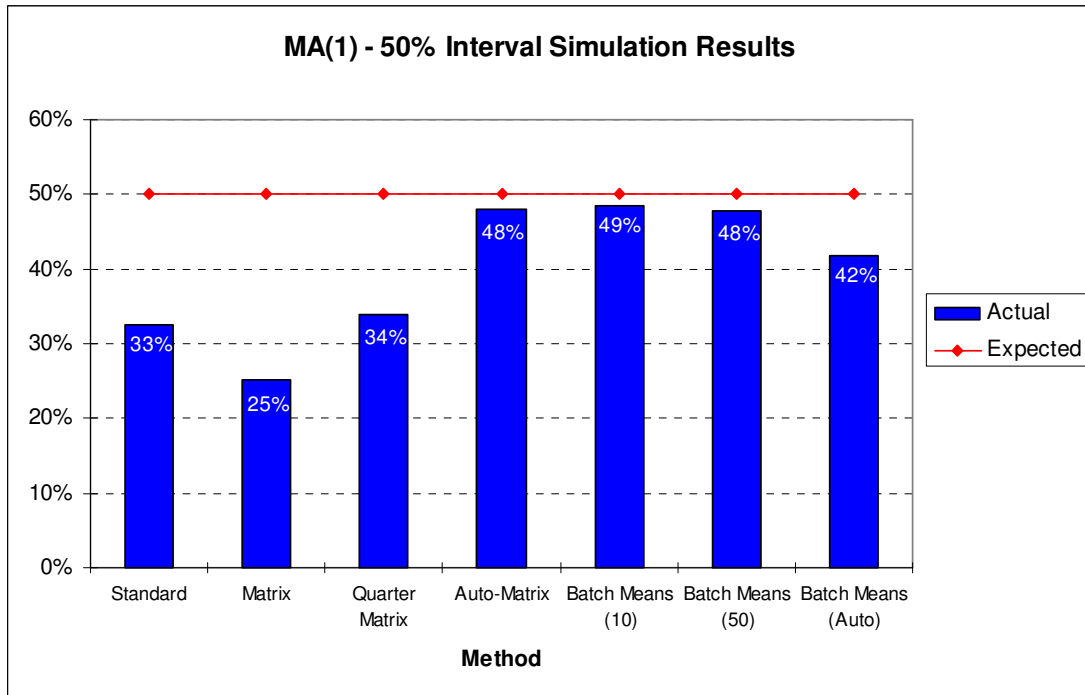


Figure 4.6 - MA(1) 50% interval simulation results

#### 4.3.4 MA(2) Model

For this model it is assumed that  $\theta_1 = 0.8$  and  $\theta_2 = 1$ , so the model will look as follows:

$$x_t = 5 + w_t + 0.8w_{t-1} + 0.1w_{t-2}$$

Here 738 runs (7.4%) were excluded from our Quarter Matrix calculations below. The results for this model are as follows:



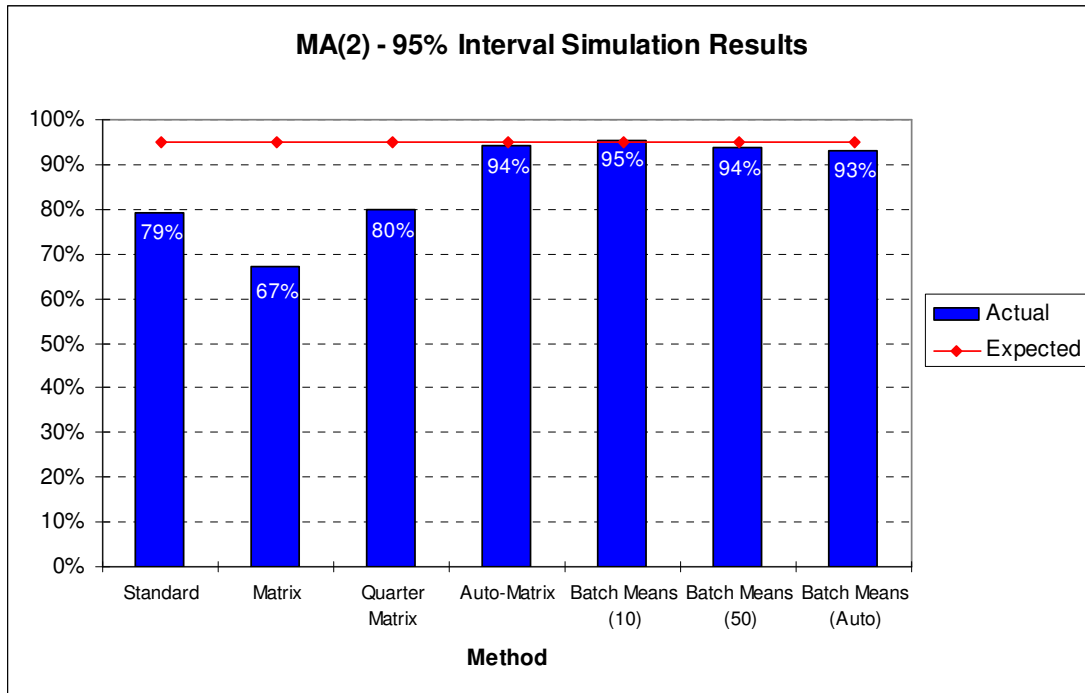


Figure 4.7 - MA(2) 95% interval simulation results

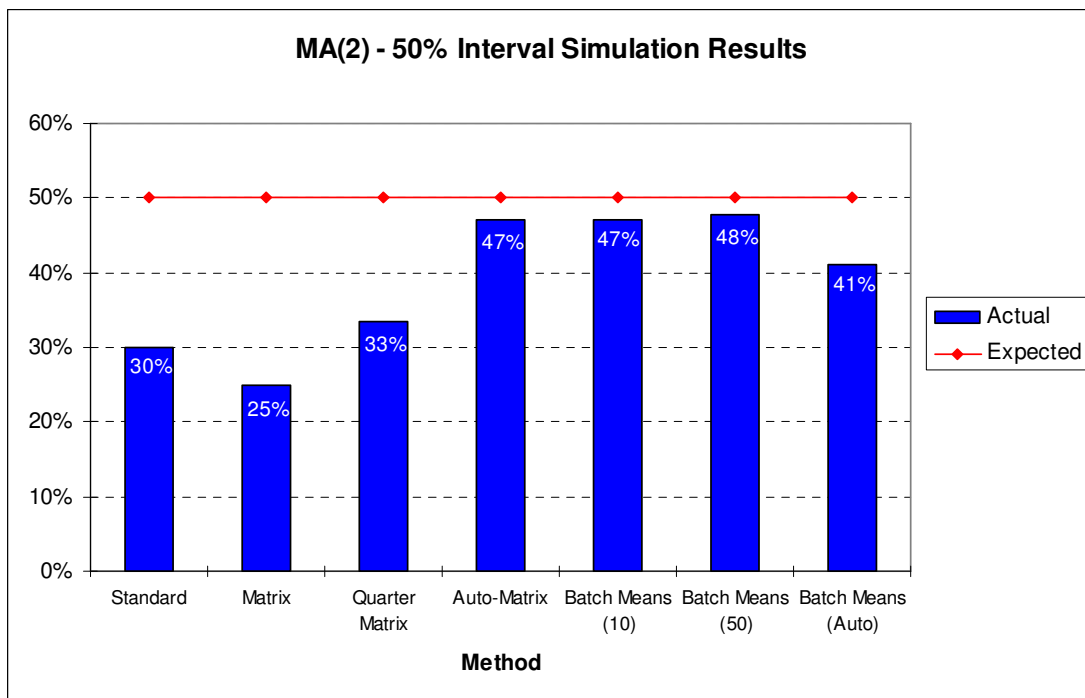


Figure 4.8 - MA(2) 50% interval simulation results

Figure 4.7 and 4.8 show that the auto-correlation matrix method and all the batch means approaches performed very well here again.

### 4.3.5 ARMA(1,1) Model

For this model it is assumed that  $\phi_1 = 0.9$  and  $\theta_1 = -0.5$ , so the model will look as follows:

$$x_t = 5 + 0.9\phi_1 + w_t - 0.5w_{t-1}$$

Here 630 runs (6.3%) were excluded from our Quarter Matrix calculations below. The results for this model were as follows:

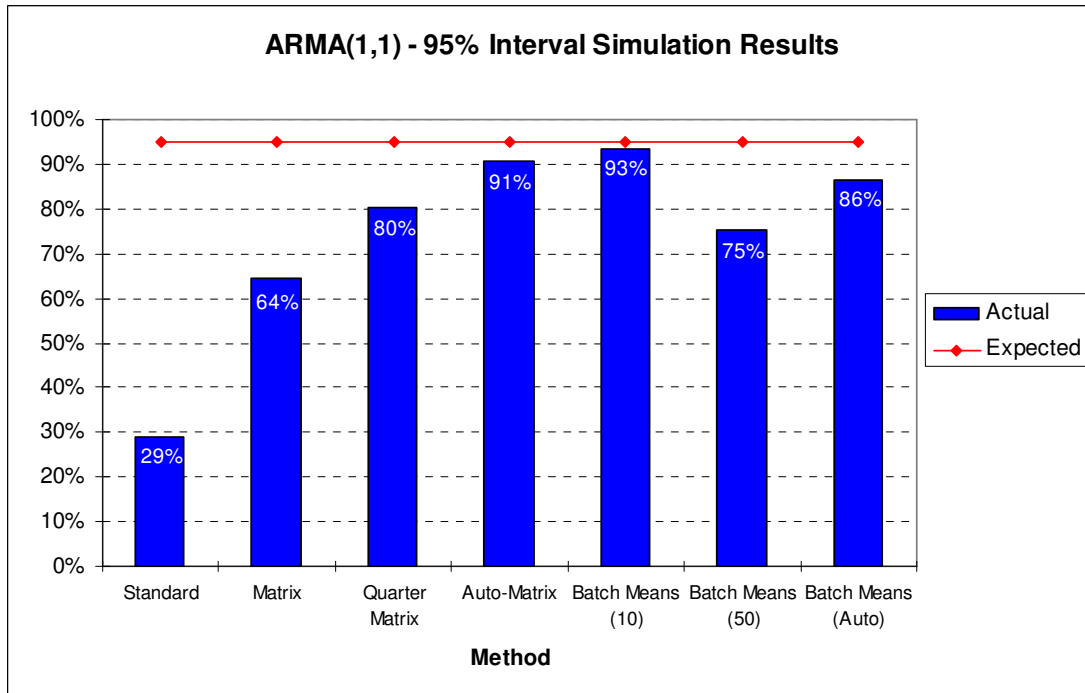
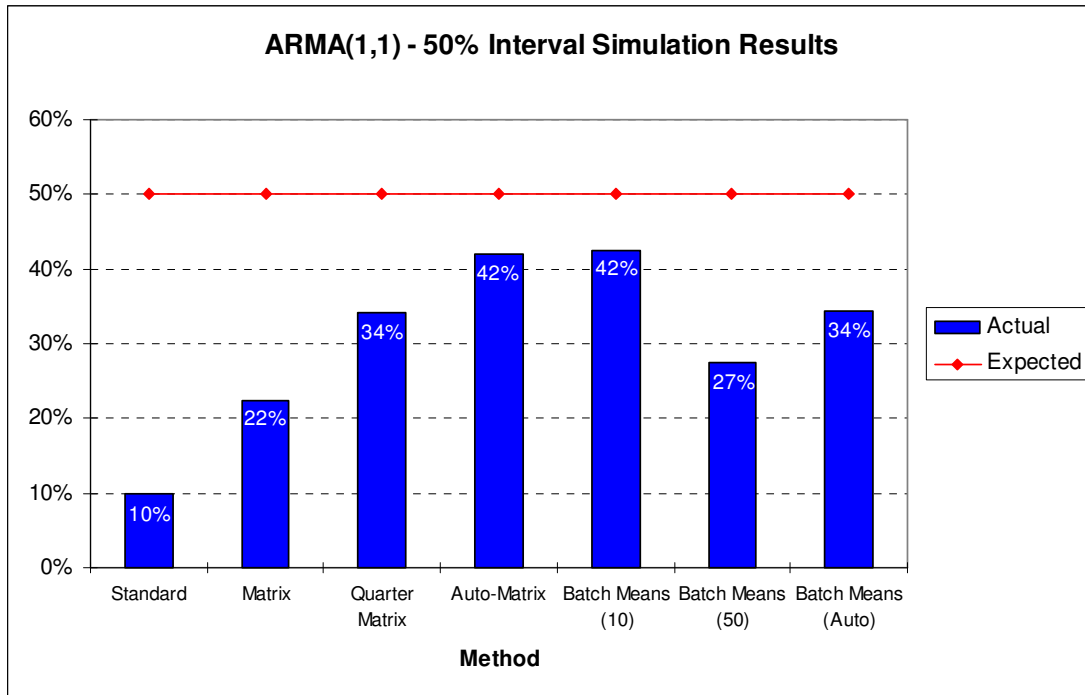


Figure 4.9 - ARMA(1,1) 95% interval simulation results

Once again, one can see how poorly the Standard method of interval construction performs with serially correlated data in both the 95% and 50% confidence intervals. The Autocorrelation-length matrix method and the Batch means (10 batches) again seem to outperform the other methods.



**Figure 4.10 - ARMA(1,1) 50% interval simulation results**

### 4.3.6 Normal(5, 0.1) Model

Here all the points in the series were generated using a  $N(5,0.1)$  distribution, therefore each point is independent of all others. Here 1150 runs (11.5%) were excluded from our Quarter Matrix calculations below. The results for this model were as follows:

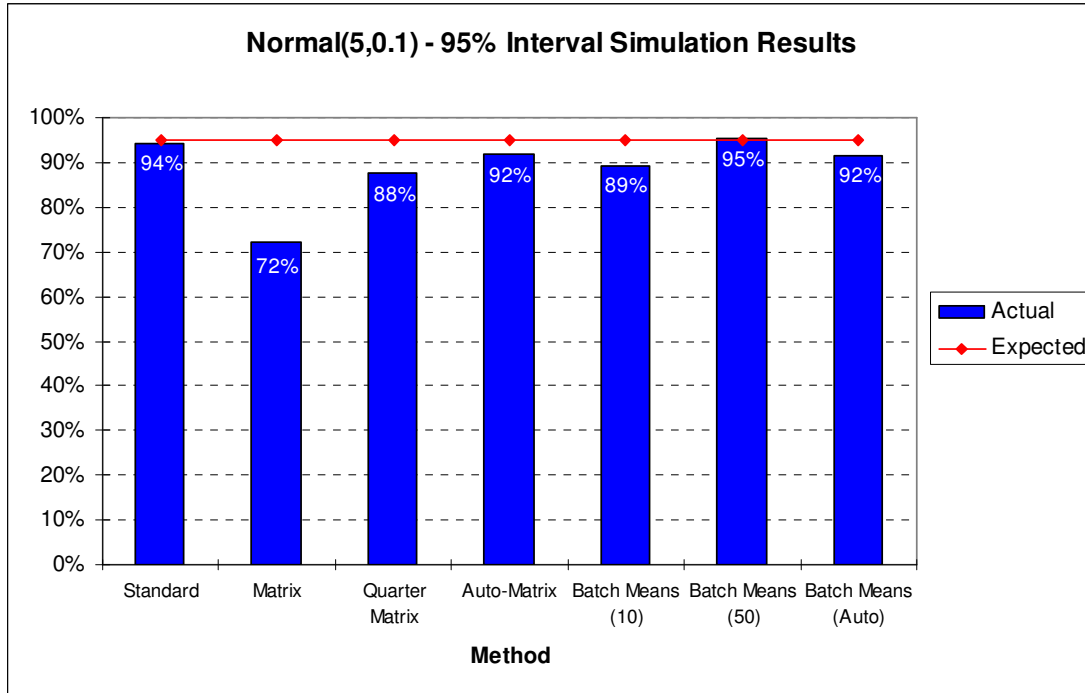


Figure 4.11 -  $N(5, 0.1)$  95% interval simulation results

Here one can see that the traditional method of calculating confidence intervals does very well with 94% of all constructed intervals containing the true mean. However, many of the other methods, which were primarily developed to deal with serially-correlated data, also did very well here.

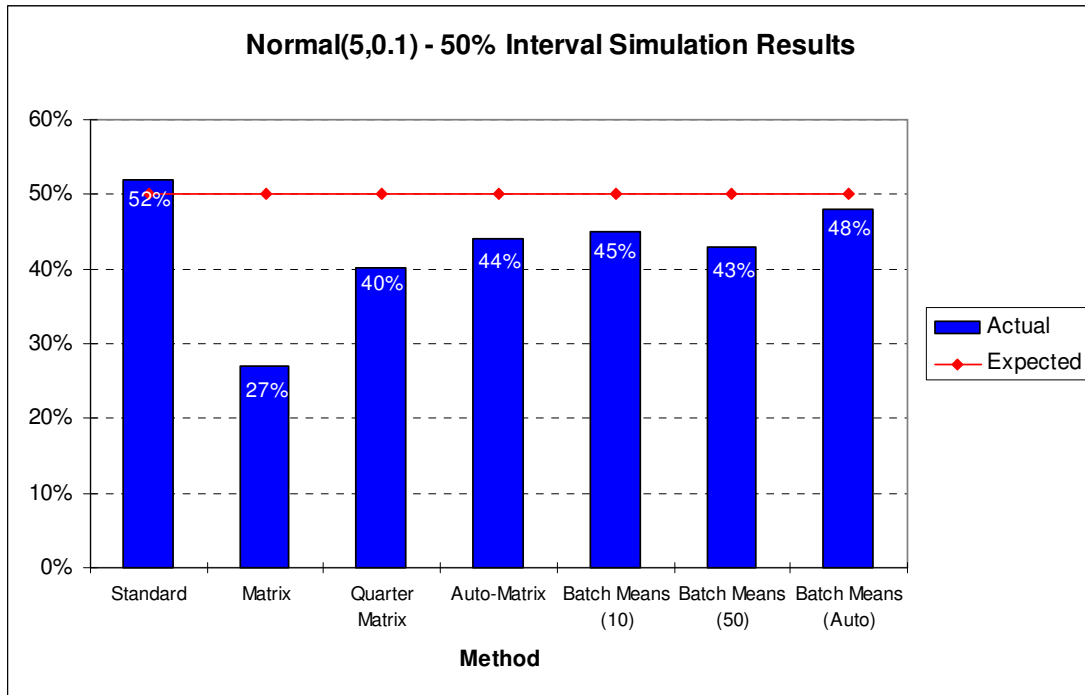


Figure 4.12 - N(5, 0.1) 50% interval simulation results

Figure 4.12 shows that the Standard method performed very well with a hit-rate of 52% when constructing 50% intervals, but several of the other models performed very well too.

### 4.3.7 Model Results Summary

Each of the summary tables 4.3 and 4.4 consist of the results from the bar charts shown in sections 4.3.1-4.3.6. For each time series model's column, the method which was closest to the expected result has been shaded and bolded.

#### 95% Confidence Interval summary

Method	AR(1)	AR(2)	MA(1)	MA(2)	ARMA(1,1)	Normal(5,0.1 <sup>2</sup> )	MAD
Standard	47%	42%	81%	79%	29%	94%	33%
Matrix	65%	65%	68%	67%	64%	72%	28%
Quarter Matrix	81%	81%	82%	80%	80%	88%	13%
Auto-Matrix	92%	91%	94%	94%	91%	92%	3%
Batch Means (10)	<b>94%</b>	<b>94%</b>	<b>95%</b>	<b>95%</b>	<b>93%</b>	89%	<b>2%</b>
Batch Means (50)	86%	84%	94%	94%	75%	<b>95%</b>	7%
Batch Means (Auto)	87%	85%	92%	93%	86%	92%	6%
MAD	16%	18%	8%	9%	21%	6%	

Table 4.3: 95% confidence interval result summary

Table 4.3 shows that the 95% confidence interval construction method with the lowest MAD (Mean Absolute Deviation) at 2%, across all time series models, is the Batch Means (10 batches) method. Poorest of all the methods employed was the Standard method of calculating confidence intervals for independent data with a MAD of 33%. This shows how important it is in practice to determine beforehand whether one's data is independent or serially correlated. Notice, however, how well the method performed for the normal model, with its intervals containing the true mean 94% of the time versus an expected 95%.

All of the models calculated intervals for the normal (independent) data well, with the average deviation of 6% across all models. With exception of the Batch Means (10) and Auto-Matrix methods, all the methods seemed to struggle calculating the correct size confidence intervals for the AR and ARMA models. The average deviation from the expected was 16% for the AR(1), 18% for the AR(2) and 21% for the ARMA(1,1) models.

#### 50% Confidence Interval summary

Method	AR(1)	AR(2)	MA(1)	MA(2)	ARMA(1,1)	Normal(5,0.1 <sup>2</sup> )	MAD
Standard	14%	13%	33%	30%	10%	52%	26%
Matrix	25%	24%	25%	25%	22%	27%	25%
Quarter Matrix	34%	33%	34%	33%	34%	40%	15%
Auto-Matrix	45%	41%	48%	47%	42%	44%	6%
Batch Means (10)	45%	44%	49%	47%	42%	45%	5%
Batch Means (50)	37%	34%	48%	48%	27%	43%	11%
Batch Means (Auto)	42%	34%	42%	41%	34%	48%	10%
MAD	16%	18%	10%	11%	20%	8%	

**Table 4.4: 50% confidence interval result summary**

Table 4.4 shows that the 50% confidence interval construction method with the lowest MAD (Mean Absolute Deviation) at 5%, is the Batch Means (10 batches) method. Both the Standard method and Matrix method performed poorly here, with MADs of 26% and 25% respectively. The methods performed best with the normal data again, with an average deviation of 8% across all methods. The methods again performed quite well on the MA(1) and MA(2) data as well.

The results in Table 4.3 and Table 4.4 show that overall, the best of the tested methods was the Batch Means (10 batches) method, followed by the Autocorrelation-length autocovariance method, while the poorest of all the methods was the Standard method of calculating confidence intervals.



For MA( $q$ ) processes, it can be shown that all  $\rho_k$  can be solved by using the ACF which has the following form:

$$\rho_k = \begin{cases} \frac{-\theta_k + \theta_1\theta_{k+1} + \theta_2\theta_{k+2} + \dots + \theta_{q-k}\theta_q}{1 + \theta_1^2 + \theta_2^2 + \dots + \theta_q^2} & k = 1, \dots, q \\ 0 & k > q \end{cases}$$

Once all  $\rho_k$  have been calculated, one can use the equality  $\rho_k = \frac{\gamma_k}{\gamma_0}$  to calculate each  $\gamma_k$ ,

where  $V[x_t] = \gamma_0 = (1 + \theta_1^2 + \theta_2^2 + \dots + \theta_q^2)\sigma_\varepsilon^2$  ( $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ )

For an ARMA(1,1) process all  $\gamma_k$  can be calculated using the following properties:

$$\begin{aligned} \gamma_0 &= \frac{1 + \theta_1^2 - 2\phi_1\theta_1}{1 - \phi_1^2} \sigma_\varepsilon^2 \quad \text{where } \varepsilon_t \sim N(0, \sigma_\varepsilon^2) \forall t \\ \gamma_1 &= \frac{(1 - \phi_1\theta_1)(\phi_1 - \theta_1)}{1 - \phi_1^2} \\ \gamma_k &= \phi_1\gamma_{k-1} = \phi_1^{k-1}\gamma_1 \quad \text{for } k \geq 2 \end{aligned}$$

Using the above properties and the constants used in section 4.3 ( $\phi_1, \phi_2, \theta_1, \theta_2$ ) the following theoretical standard errors were calculated for each model:

Model	True S.E.
AR(1)	0.07040
AR(2)	0.07031
MA(1)	0.02544
MA(2)	0.02685
ARMA(1,1)	0.25213
Normal(5,0.1)	0.01414

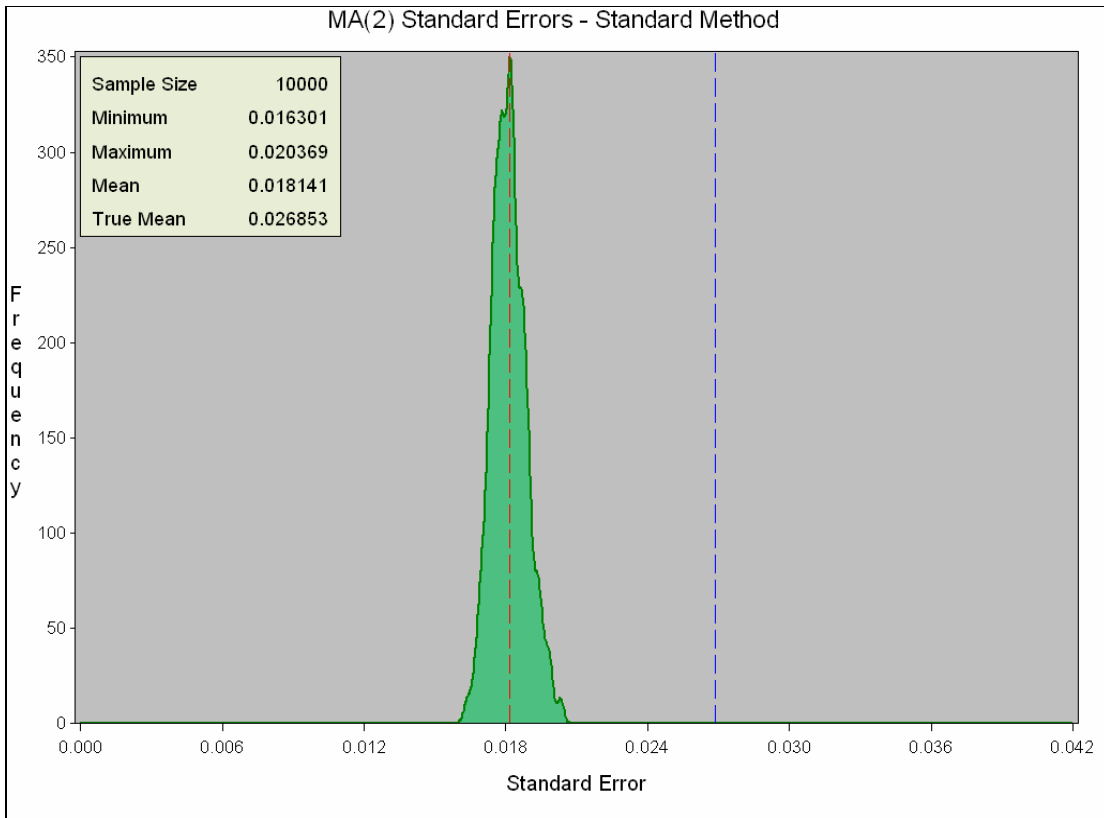
**Table 4.5 True standard errors**

The standard error for the normal data set was simply calculated as:  $SE = \sqrt{\frac{0.1}{500}}$



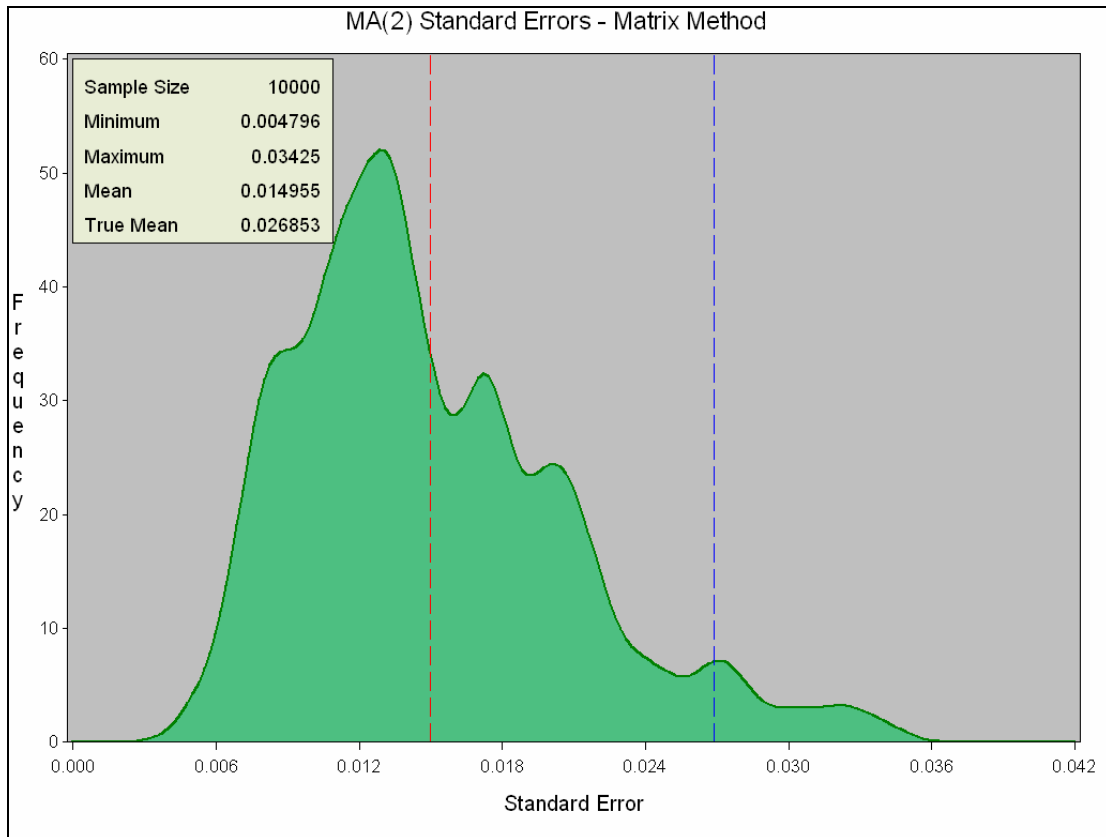
#### 4.4.2 MA(2) Histograms

This section contains examples of what the histograms of standard errors, obtained from the simulations, look like for a particular time series model, in this case the MA(2) model. All of the remaining time series models' histograms can be found in Appendix A. All of the histograms were drawn in SAS using the same scale (w.r.t the x-axis) across the different methods, so as to make comparisons easier.



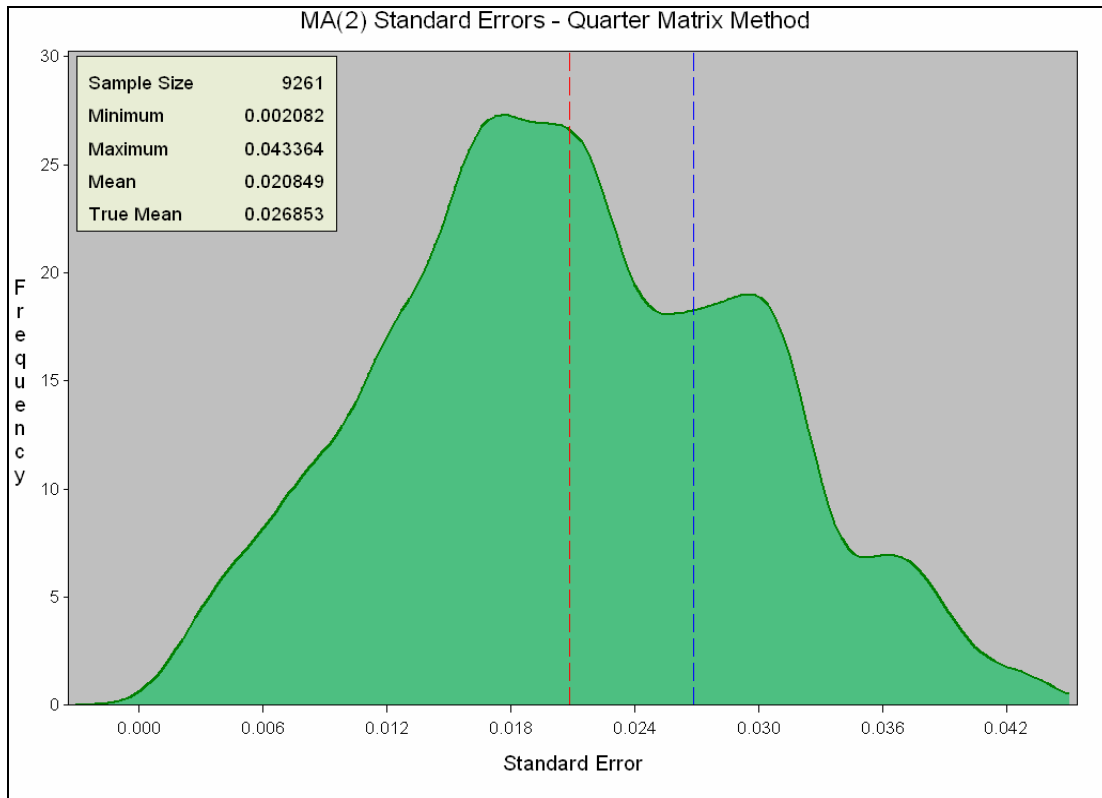
**Figure 4.13 - MA(2) SEs - Standard Method**

Figure 4.13 contains the histogram of simulated standard error values (green) for the standard method. A kernel density function was used to smooth the data. The red dotted line is the mean of all the standard errors generated, while the blue dotted line represents the true standard error for this time series (with the specified constants). The histogram shows that all of the simulated standard errors which were generated for the MA(2) model were too small. This would explain why this method of creating confidence intervals did poorly during the simulation (see section 4.3.4).



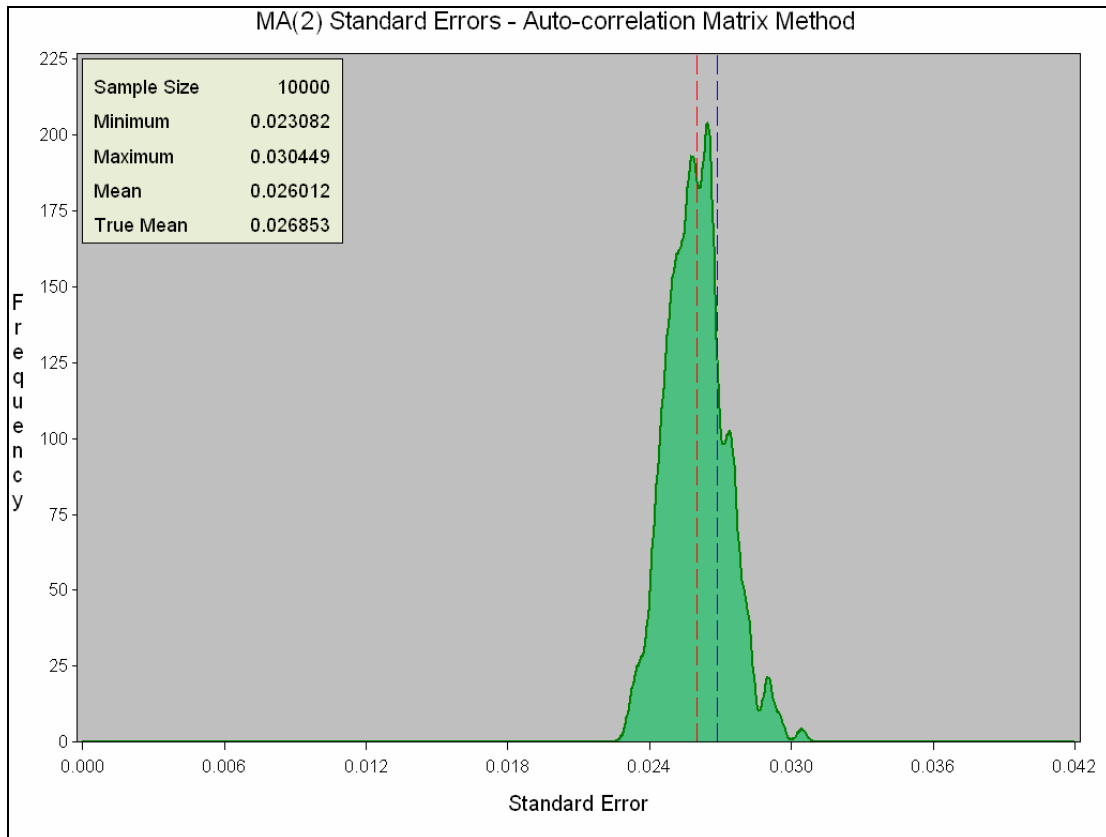
**Figure 4.14 - MA(2) SEs – Autocovariance Matrix Method**

Figure 4.14 shows that the autocovariance matrix method did even poorer than the standard method because, even although some of the generated standard errors were larger than the true standard error, the mean of the standard errors was 0.014955, whereas the true standard error was 0.026853.



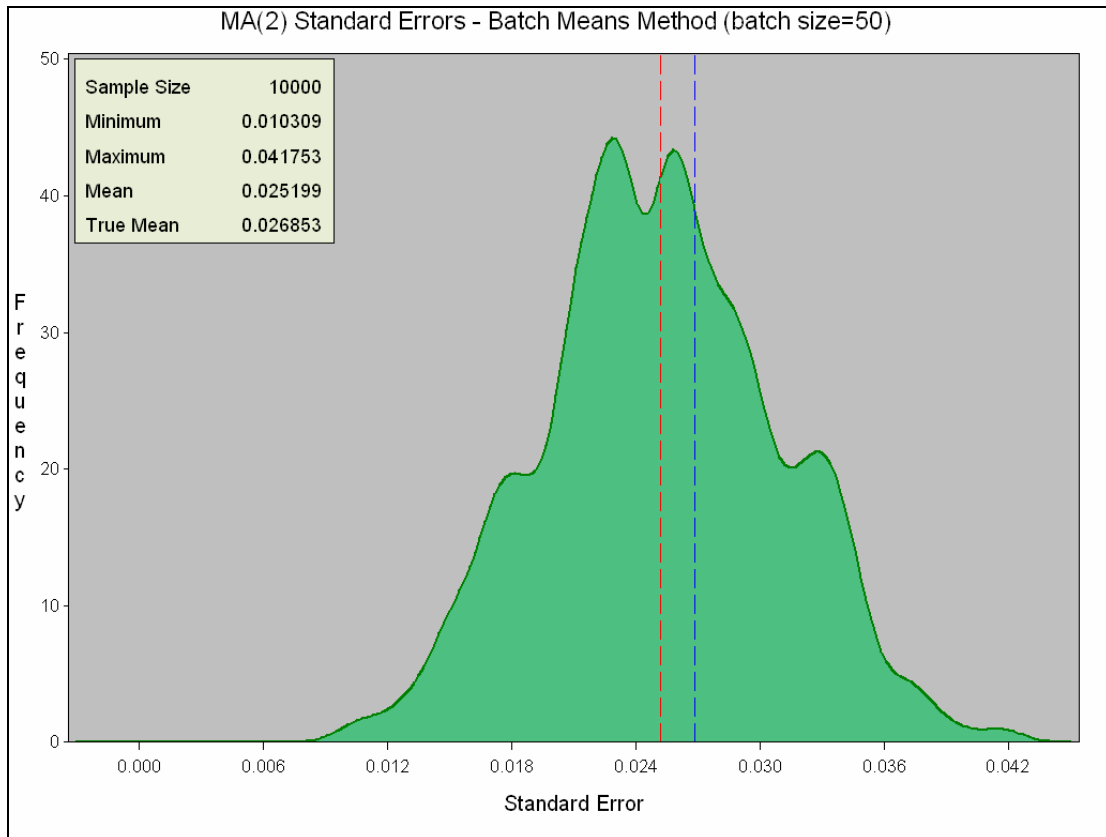
**Figure 4.15 - MA(2) SEs – Autocovariance Quarter Matrix Method**

Figure 4.15 shows that many of the standard errors generated using the autocovariance quarter matrix method were better than the previous two methods, but the mean was still lower than the true standard error. The sample size here was 9261, showing that more than 7% of simulation runs yielded invalid results which were discarded.



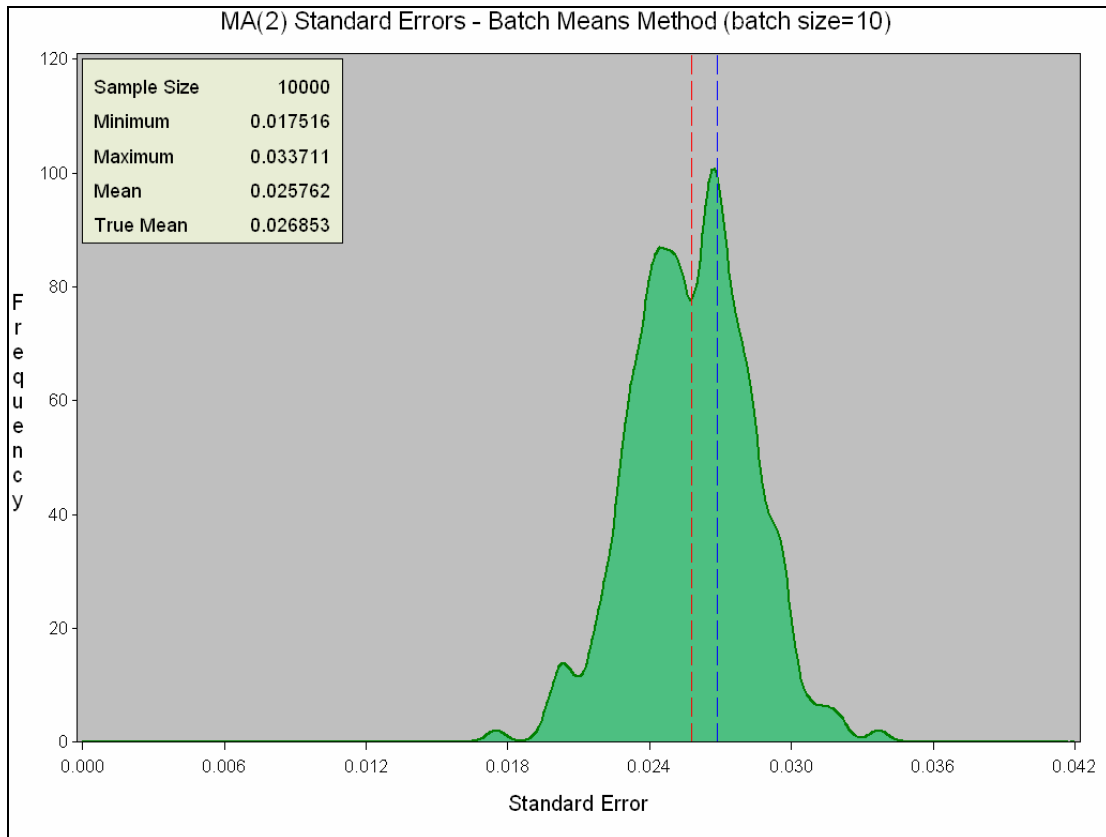
**Figure 4.16 - MA(2) SEs – Autocorrelation-length Matrix Method**

The Autocorrelation-length matrix method produced standard errors which were very close to the true mean, with a mean of 0.026012. This differs from the true mean by 0.0084 or 3.14%. This explains why this method performed very well during the simulation. The method created confidence intervals which contained the true mean 94% of the time when it was creating 95% intervals, and 47% of the time when it was creating 50% intervals.



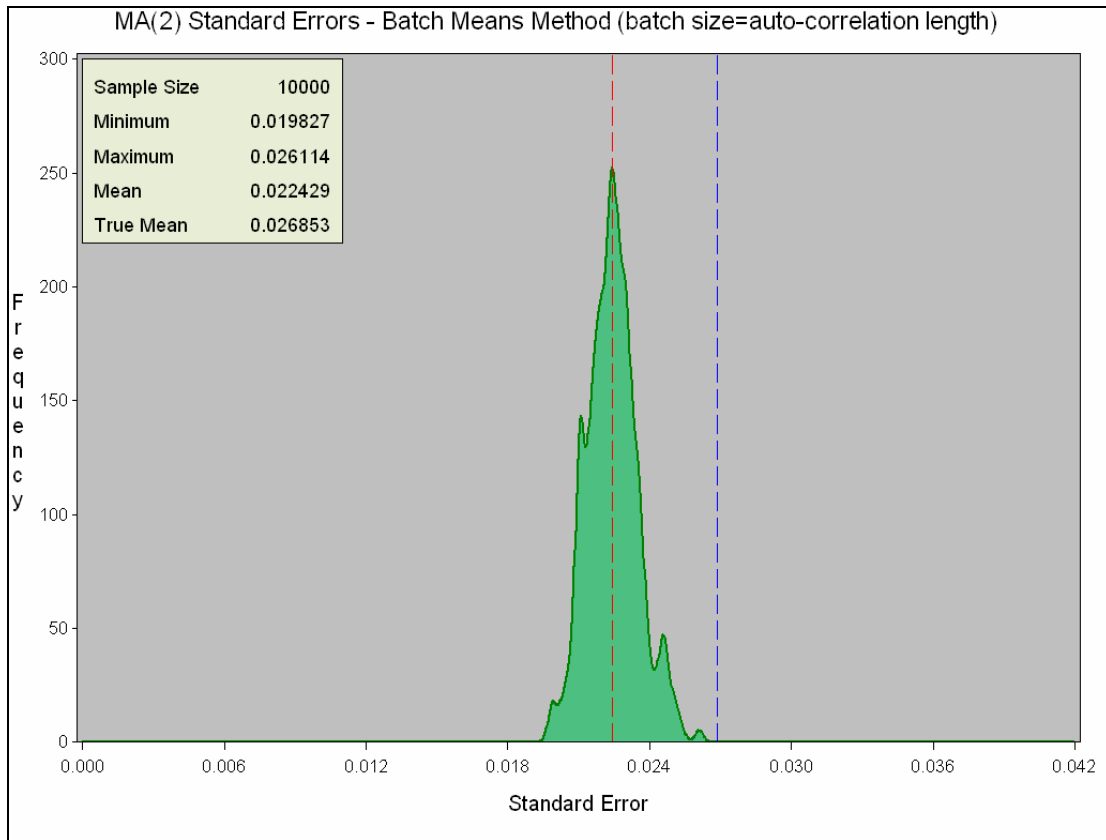
**Figure 4.17 - MA(2) SEs – Batch Means (10 batches) Method**

The first of the batch means methods which was tested was the 10 batches of size 50 method. It performed very well with a mean standard error of 0.025199, differing from the true mean by 6.16%



**Figure 4.18 - MA(2) SEs – Batch Means (50 batches) Method**

Figure 4.18 shows that the Batch Means method with 50 batches of size 10 also performed very well with a mean standard error of 0.025762.



**Figure 4.19 - MA(2) SEs – Batch Means (Autocorrelation-length) Method**

Although Figure 4.19 shows that all of the standard errors generated by the auto-correlation length batch means method were below the true mean (the maximum standard error generated was 0.027114), it performed quite well with a mean standard error of 0.022429.

## 4.5 Actual PMD data

Now that the best of the methods tested in section 4.3 has been established, those methods were applied to actual PMD data which were recorded in the field. For each dataset, the following steps will be performed:

- The PMD-plot will be shown.
- An autocorrelation plot will be examined in order to detect whether the data are serially correlated.
- The data will be tested for stationarity.
- The best methods, determined in section 4.3, will be used to construct confidence intervals.

### 4.5.1 Hopetown to Britstown

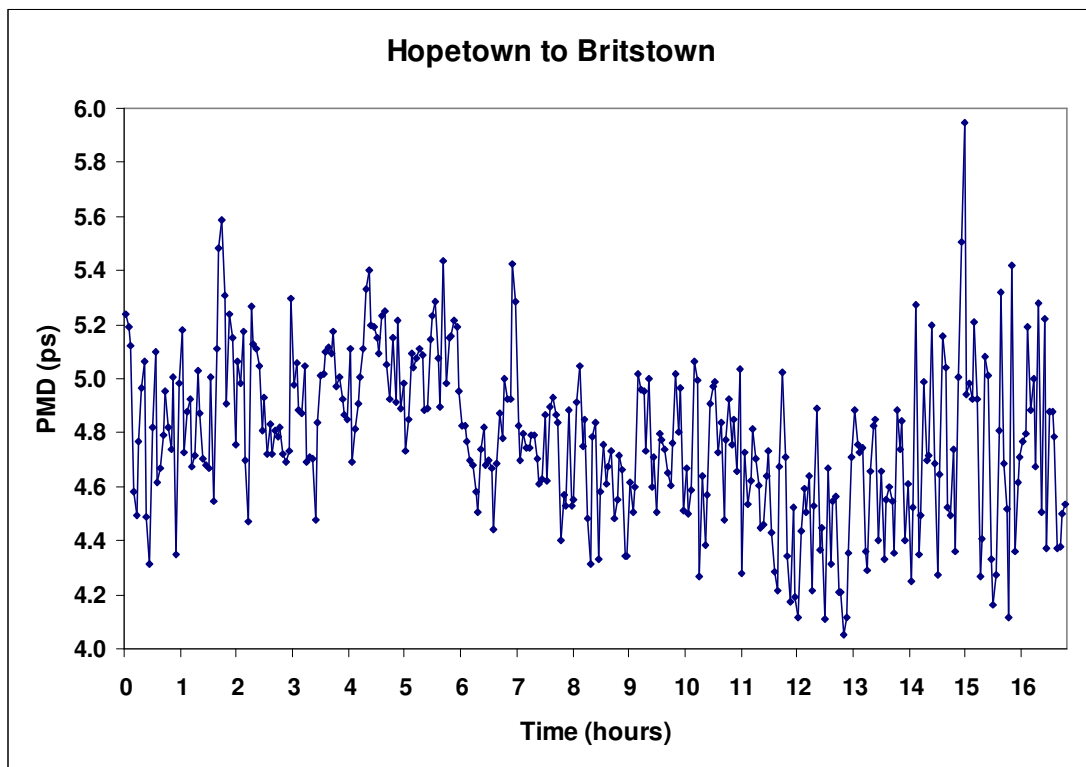


Figure 4.20 - PMD plot: Hopetown to Britstown

The data in figure 4.20 were recorded over a 17 hour period between the towns of Hopetown and Britstown, which are situated in Northern Cape Province of South Africa. Apart from the PMD spike near the end of the recording, the dataset appears to be stationary.



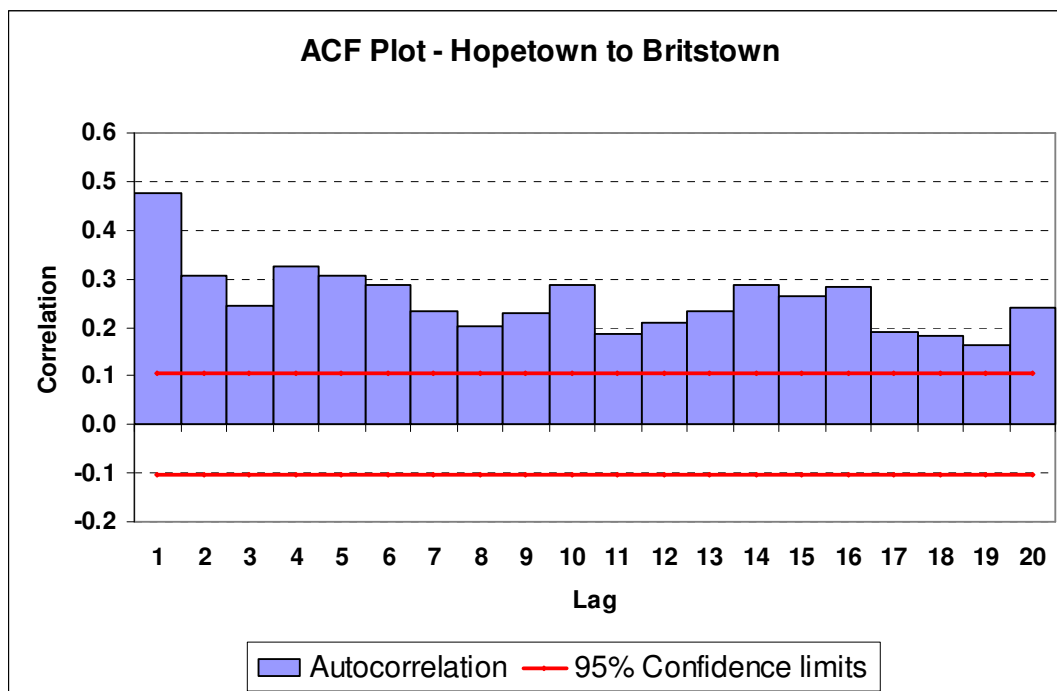


Figure 4.21 - Autocorrelation plot: Hopetown to Britstown

A quick look at the autocorrelation function plot shows that the data have significant autocorrelations up to at least lag 20. This means that the standard method of calculating confidence intervals is very likely insufficient and our other methods should be utilized. If the blue autocorrelation bars had dropped off below the 95% confidence limit lines from lag 1 onwards, the standard, autocovariance and batch means methods would be expected to perform similarly. Table 4.6 contains output from the program EViews 3.1 which shows the statistics required to test the data for stationarity.

Null Hypothesis: Hopetown to Britstown data has a unit root		
Exogenous: Constant		
Lag Length: 3 (Automatic based on AIC, MAXLAG=25)		
	t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic	-5.222724	0.0000
Test critical values:		
1% level	-3.448889	
5% level	-2.869605	
10% level	-2.571135	
*MacKinnon (1996) one-sided p-values.		

Table 4.6 – Augmented Dickey-Fuller stationarity results for Hopetown to Britstown

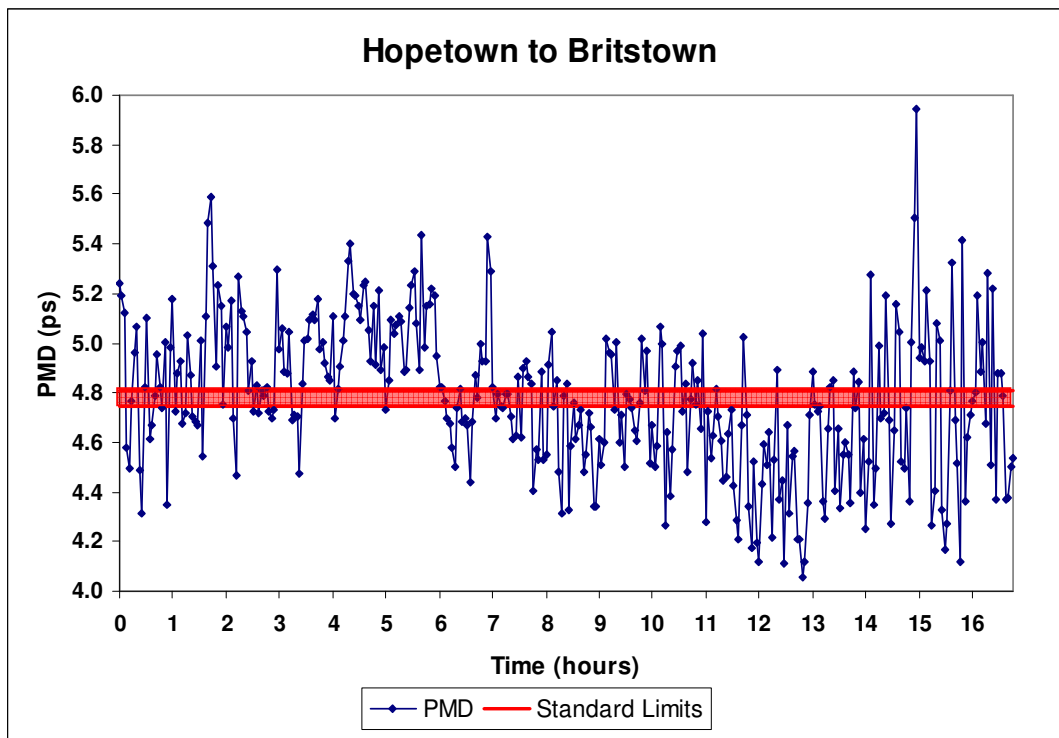
Recall from section 3.4, that the more negative the ADF statistic, the stronger the rejection of the null hypothesis that there is a unit root (i.e. that the data are non-stationary). Here one can see the critical value for a 5% level of significance is -2.87. A test statistic lower than -2.87 therefore implies stationarity, whereas a value greater than -2.87 does not imply stationarity. The  $\tau$ -statistic for this dataset is -5.22, therefore the dataset is stationary. Knowing that the data are stationary, one can proceed to utilizing the interval construction methods tested in section 4.3.

The following standard errors were calculated, using the standard and two best methods from section 4.3, for the Hopetown to Britstown dataset:

Method	Standard Error	95% Lower	95% Upper
Standard	0.0159	4.7479	4.8104
Auto-Matrix	0.0616	4.6579	4.9003
Batch Means (10)	0.0281	4.7225	4.8355

**Table 4.7 – Standard Errors and 95% confidence limits for Hopetown to Britstown data.**

The mean of the Hopetown to Britstown data was 4.779. Combining this with the standard errors calculated by the different methods produces lower and upper confidence limits as shown in Table 4.7.



**Figure 4.22 - 95% interval - Standard method: Hopetown to Britstown**

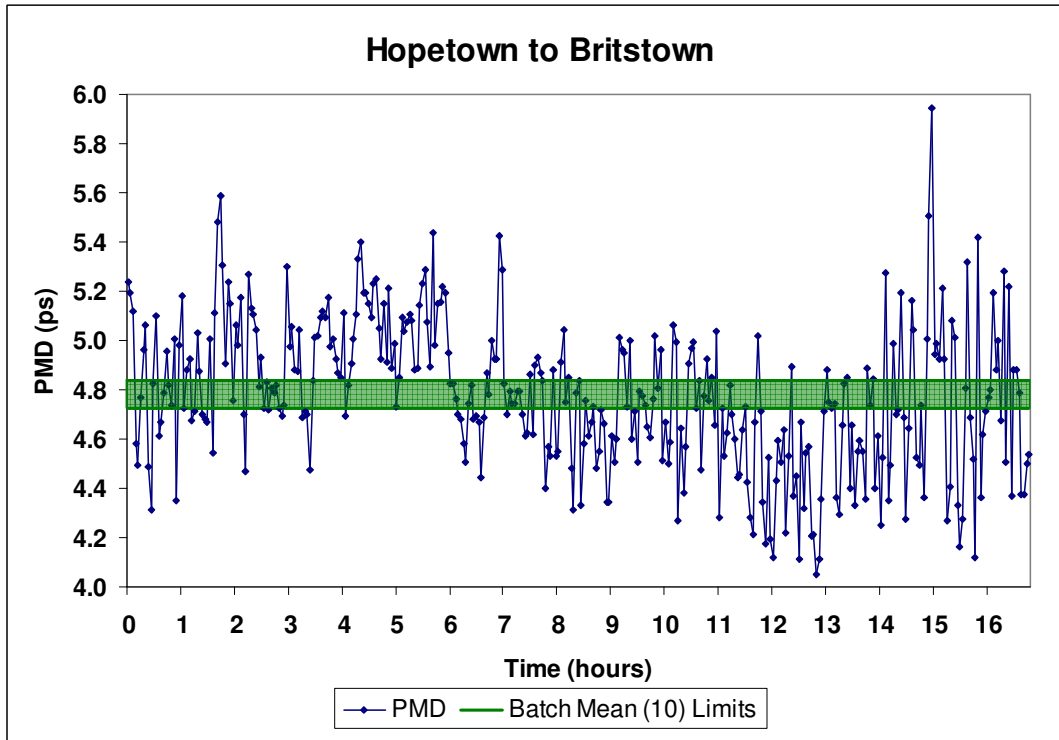


Figure 4.23 - 95% interval - Batch Means (10) method: Hopetown to Britstown

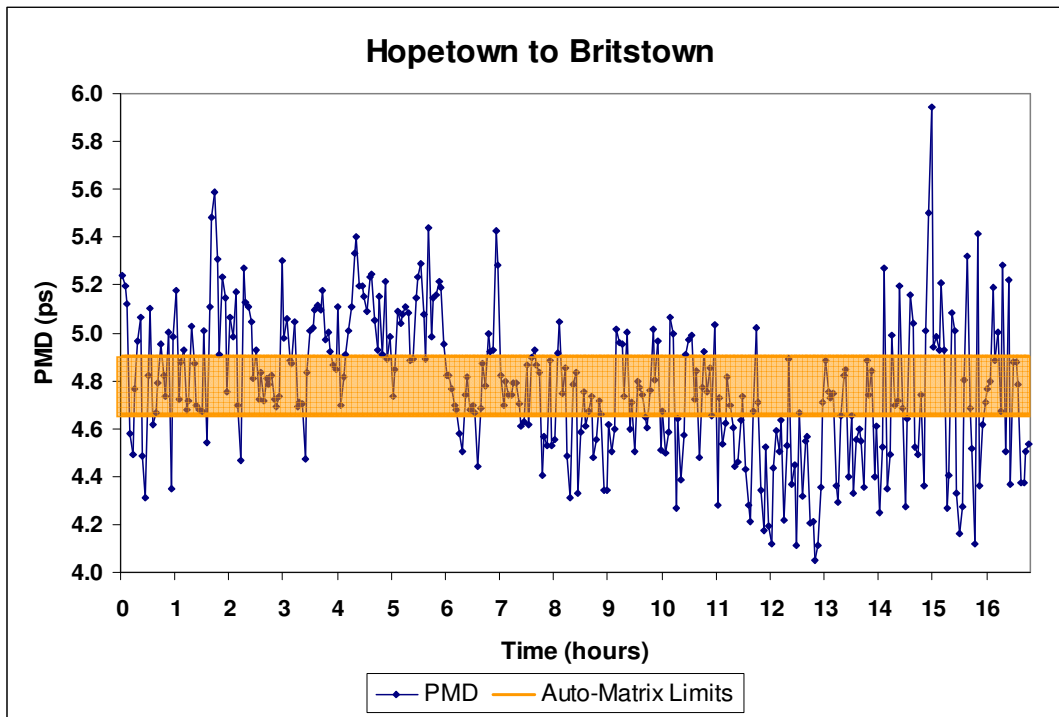


Figure 4.24 - 95% interval - Autocorrelation-length Matrix method: Hopetown to Britstown

Figures 4.22 – 4.24 show the different confidence intervals for the mean of the PMD dataset using the various interval construction methods. It is clear that for serially-correlated data,

the methods tested in section 4.3 produce much wider and more conservative confidence intervals than the standard method.

Using the above data, a telecoms company could conclude that the intervals constructed would contain the true mean of the fibre 95% of the time. They could then use this piece of information to decide whether the fibre's mean PMD level is too high (in which case the fibre might need replacing) and whether or not they could increase the data transfer rate being utilized.

#### 4.5.2 Britstown to Victoria West

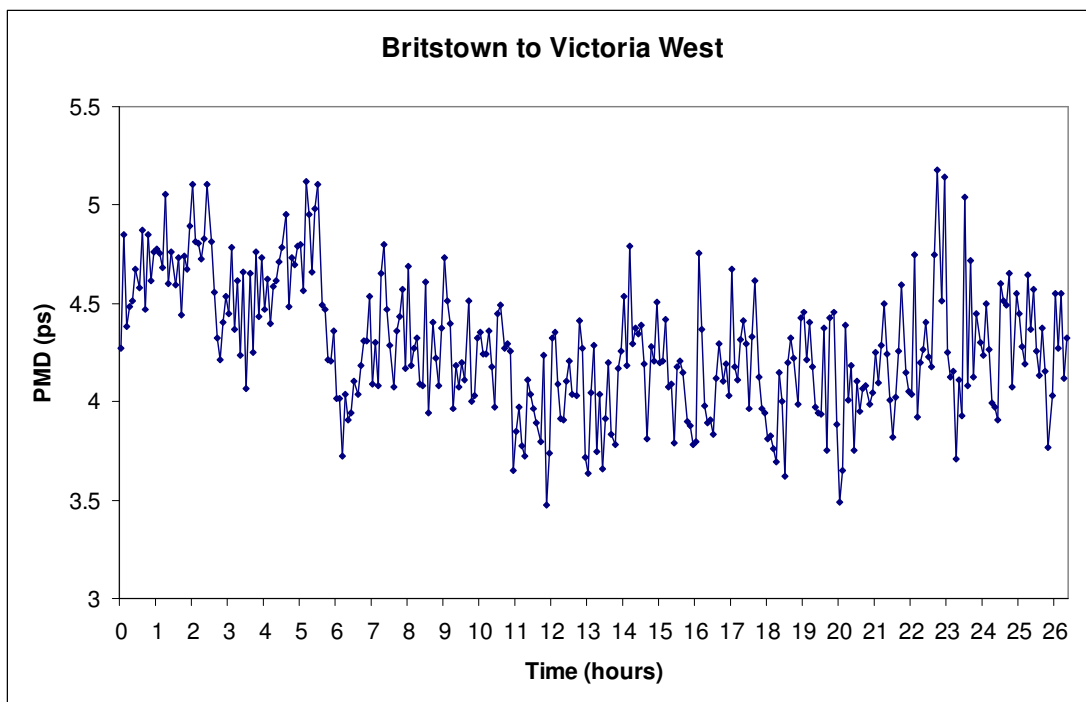
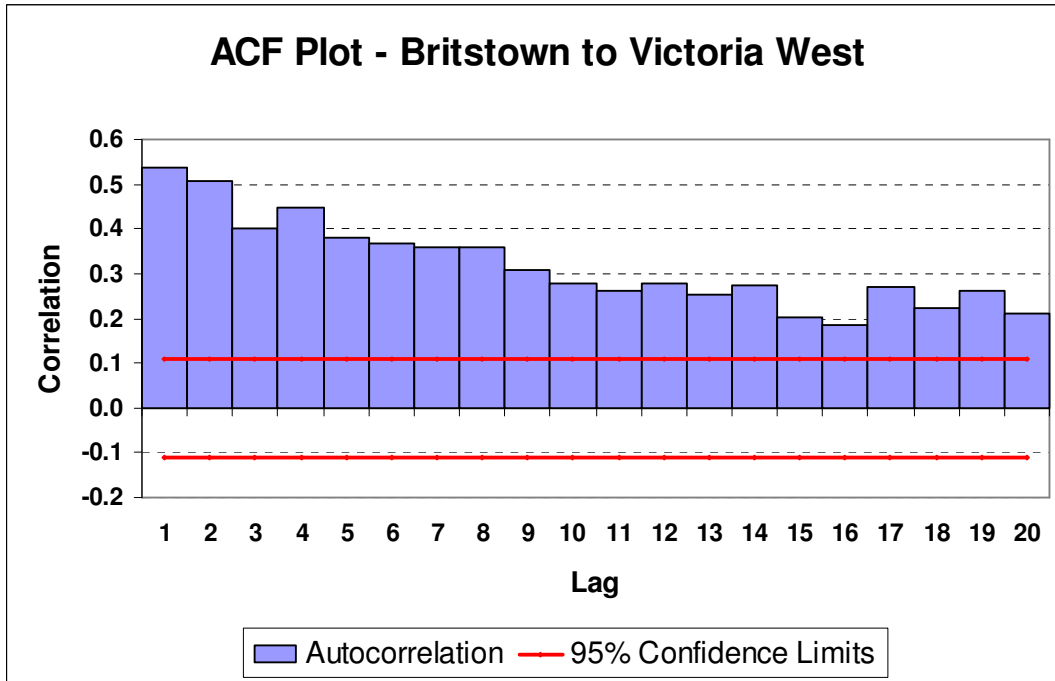


Figure 4.25 - PMD plot: Britstown to Victoria West

The data in figure 4.25 were recorded over a 27 hour period between the towns of Britstown and Victoria West, which are situated in Northern Cape Province of South Africa. Here too the dataset appears to be stationary.



**Figure 4.26 - Autocorrelation plot: Britstown to Victoria West**

The autocorrelation function plot of the dataset shows significant serial correlation between the data points.

Null Hypothesis: Britstown_VicWest has a unit root		
Exogenous: Constant		
Lag Length: 3 (Automatic based on AIC, MAXLAG=25)		
	t-Statistic	Prob.*
<hr/>		
Augmented Dickey-Fuller test statistic	-4.249485	0.0006
Test critical values:		
1% level	-3.451078	
5% level	-2.870561	
10% level	-2.571647	
<hr/>		
*MacKinnon (1996) one-sided p-values.		

**Table 4.7: Augmented Dickey-Fuller stationarity results Britstown to Victoria West data**

As the  $\tau$ -statistic of this dataset is -4.249 is below the critical value of -2.87, one can conclude that the dataset is stationary and proceed to the next step.

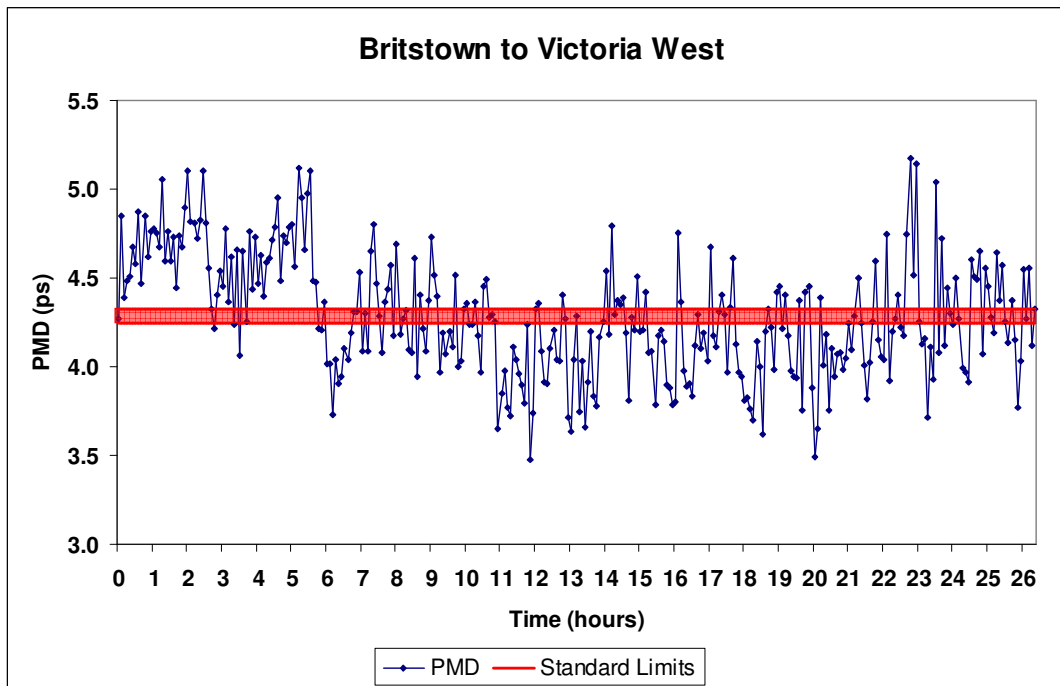
The following standard errors were calculated for this dataset:

Method	Standard Error	95% Lower	95% Upper
Standard	0.0188	4.2483	4.3218
Auto-Matrix	0.0902	4.1082	4.4619
Batch Means (10)	0.0351	4.2145	4.3556

**Tables 4.8: Standard Errors and 95% confidence limits for Britstown to Victoria West data**

The mean of the Britstown to Victoria West data was 4.285. Combining this with the standard errors calculated by the different methods produces lower and upper confidence limits as shown in Table 4.8.

Graphically these confidence intervals are shown as in Figures 4.27 – 4.29:



**Figure 4.27 - 95% interval - Standard method: Britstown to Vic West**

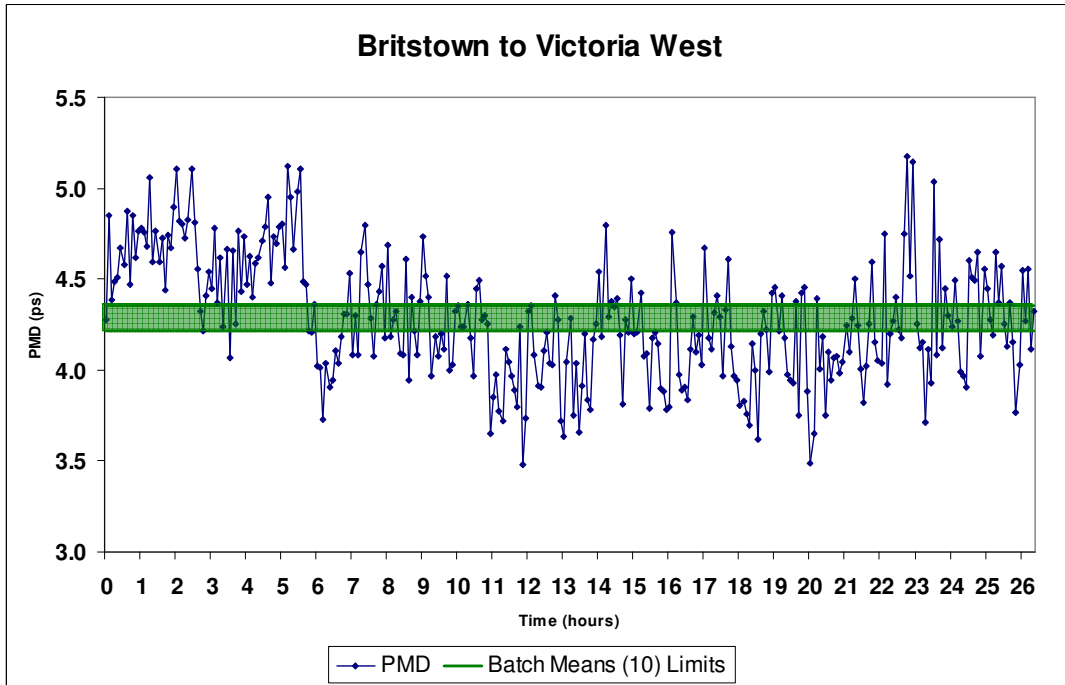


Figure 4.28 - 95% interval - Batch Means (10) method: Britstown to Vic West

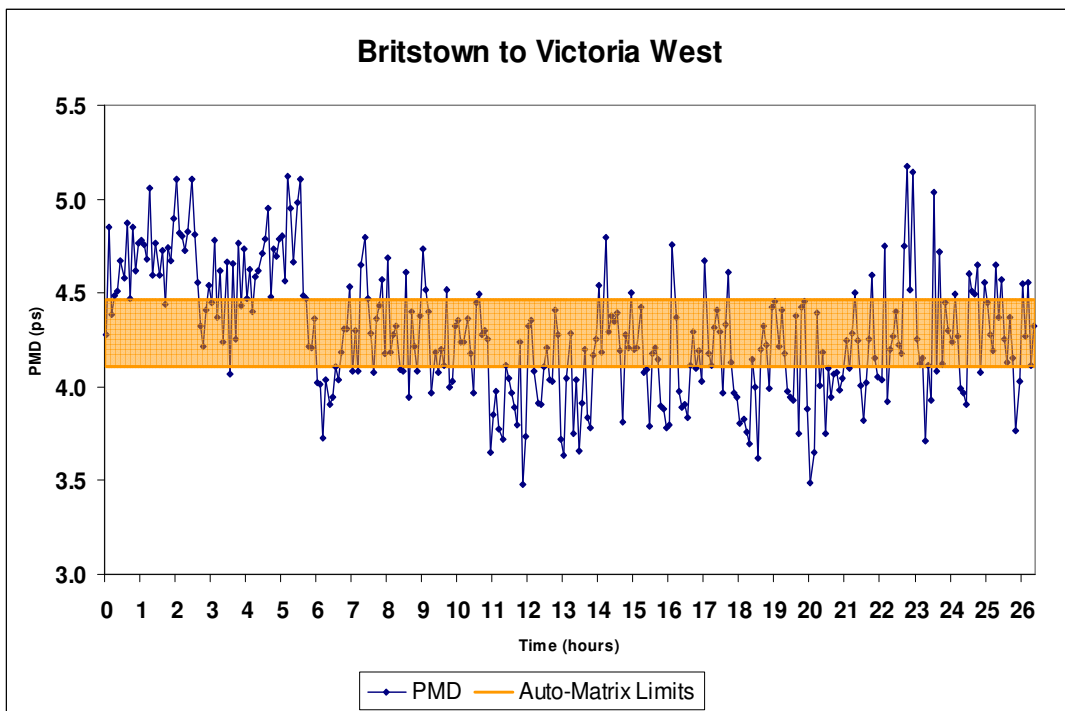


Figure 4.29 - 95% interval - Autocorrelation-length Matrix method: Britstown to Vic West

### 4.6 Effects of Polarization Scramblers

As mentioned in Chapter 2, fibres which are buried, as opposed to aerial (or a combination of buried/aerial), tend to show a lower variation in the different PMD values measured. This is because these fibres are not exposed to environmental shocks such as trains or trucks riding over the ground where the fibres are laid, as well as temperature and wind extremes. Stated another way, buried fibres take a longer time period to fully sample through each possible state of the fibre.

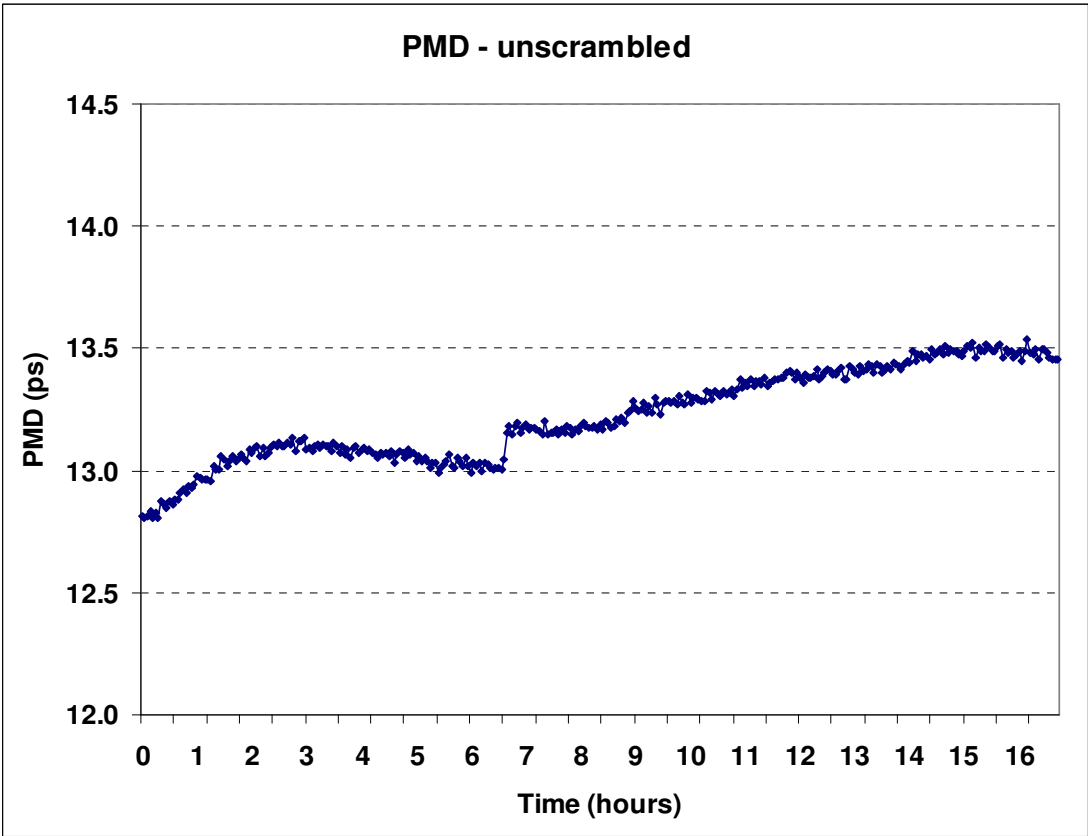


Figure 4.30 - Unscrambled data

Consider Figure 4.30 which contains PMD data generated in a laboratory. The series does not look very “random” or independent, with its steady upwards trend. This is one of the problems with measuring PMD. If the data are non-stationary – as one would expect the above dataset to be – then the interval construction methods already described cannot be used.

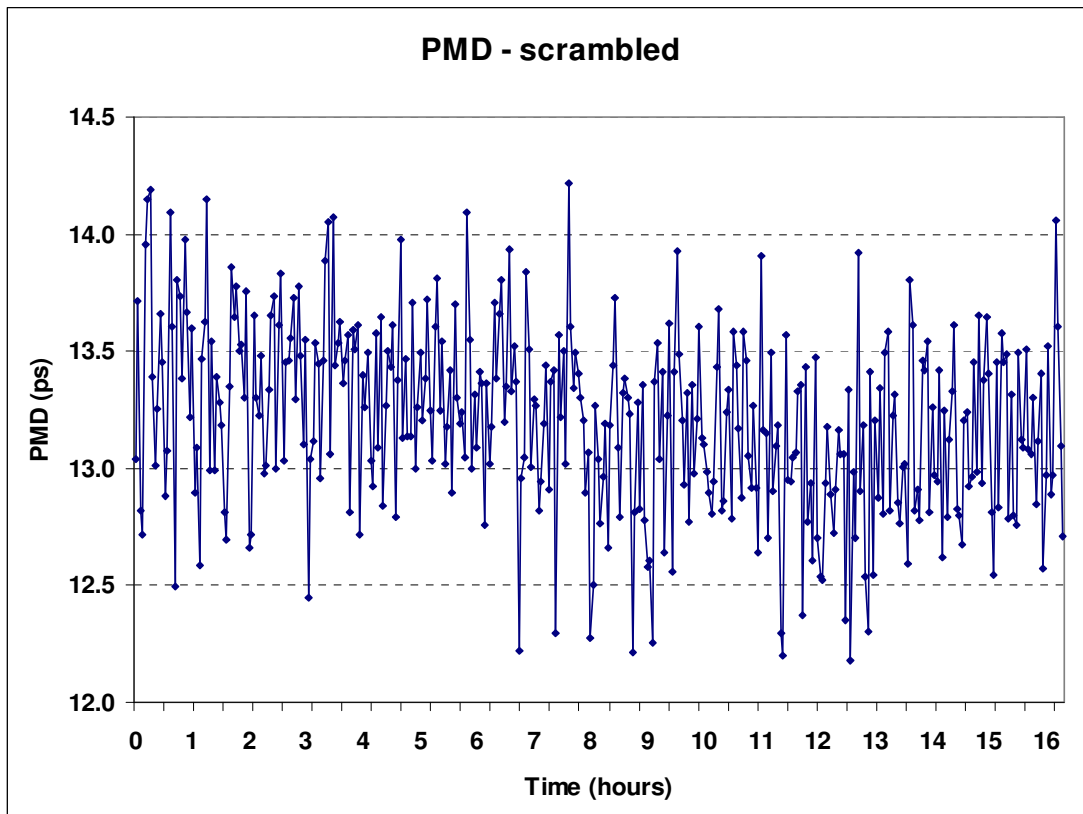


Null Hypothesis: Unscrambled has a unit root		
Exogenous: Constant		
Lag Length: 2 (Automatic based on AIC, MAXLAG=25)		
	t-Statistic	Prob.*
<hr/>		
Augmented Dickey-Fuller test statistic	-2.002448	0.2858
Test critical values:	1% level	-3.449977
	5% level	-2.870084
	10% level	-2.571391
<hr/>		
*MacKinnon (1996) one-sided p-values.		

**Table 4.9: Augmented Dickey-Fuller stationarity results - unscrambled data**

As expected the Augmented Dickey-Fuller  $\tau$ -statistic of -2.00 is above the 5% critical value of -2.87, implying that this dataset is not stationary.

Polarization scramblers were fitted to the start and end sections of this fibre, and the same measurements were carried out. Figure 4.31 shows the output.



**Figure 4.31 - Scrambled data**

It is immediately clear that the use of scramblers makes the data more independent and less serially correlated. The scrambler literally scrambles the states of the fibre and makes each

observation from scratch. Measurements made on an unscrambled fibre (especially if buried) are recorded when the fibre's orthogonal axes have not been reset after each recording. The stationarity test for the scrambled data, which confirms the stationarity, can be found in Table 4.10.

Null Hypothesis Scrambled has a unit root		
Exogenous: Constant		
Lag Length: 5 (Automatic based on AIC, MAXLAG=25)		
	t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic	-5.756122	0.0000
Test critical values:		
1% level	-3.447675	
5% level	-2.869071	
10% level	-2.570849	

\*MacKinnon (1996) one-sided p-values.

**Table 4.10: Augmented Dickey-Fuller stationarity results - Scrambled data**

Apart from allowing one to quickly get a picture of the range of possible PMD values, scrambling also makes it more likely that a fibre's PMD data are stationary. This then allows one to use the methods tested in section 4.3.

## 5 Conclusion

Polarization mode dispersion is a statistical effect created by minute random deviations, and an effect found to some extent in all fibres. Understanding PMD, its contributing factors and its effect on system performance becomes critical when planning to build a new optical network or when considering an existing installation.

The following conclusions, which answer the research questions mentioned in Chapter 1, and contributions to current knowledge were reached during this project:

- The Standard method of calculating standard errors and confidence intervals is insufficient for PMD data, as well as other forms of time series data, as it produces confidence intervals which are too narrow. The confidence intervals which are constructed with the other methods are wider and therefore more reliable.
- There is a significant problem with the application of the Quarter-matrix method as suggested by Box and Jenkins [16], namely that some of the variances, and therefore standard errors, calculated are negative. The Autocorrelation-length Matrix method is a significant improvement upon this method.
- The Batch Means method works well especially when there is a relatively large batch size. In the simulation, the method with 10 batches of size 50 performed the best of all the methods tested.
- The field data which were tested was found to be stationary.
- The use of PMD scramblers greatly reduces serial correlation between data points. The resultant datasets are also very likely to be stationary.

Potential future areas of research include:

- Methods for determining optimal batch size when using the Batch Means method.
- Methods for determining sample sizes when a given precision is required.
- Whether the use of polarization scramblers increase/decrease the confidence intervals produced.

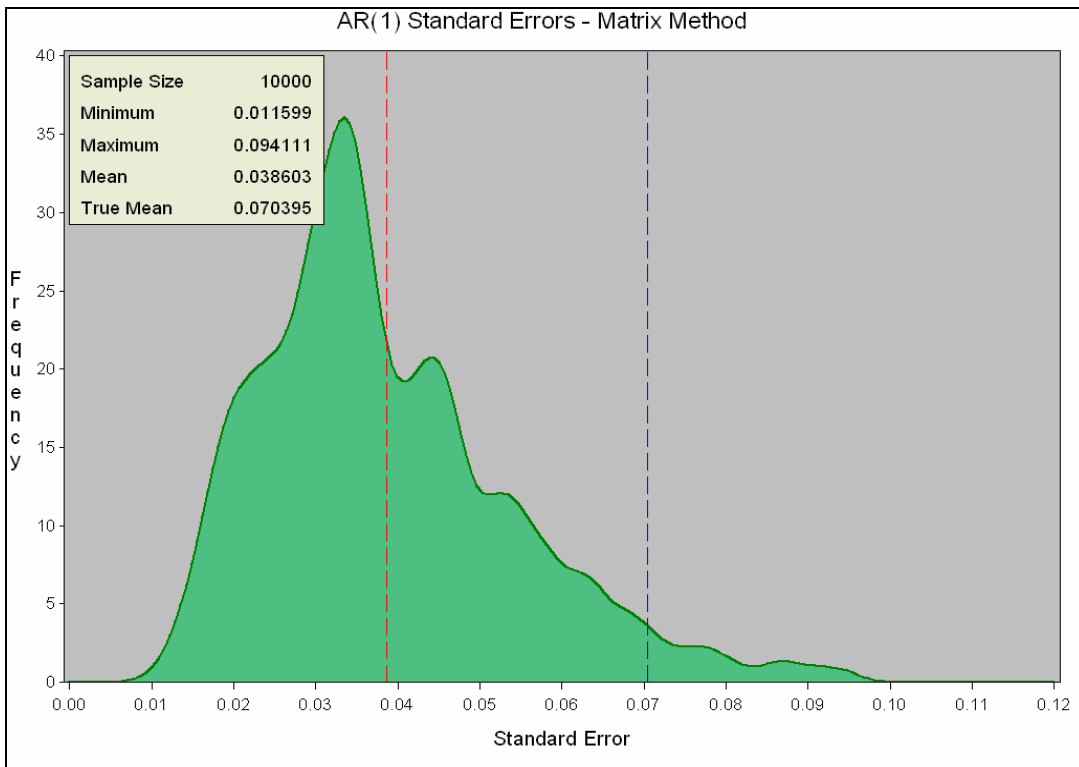
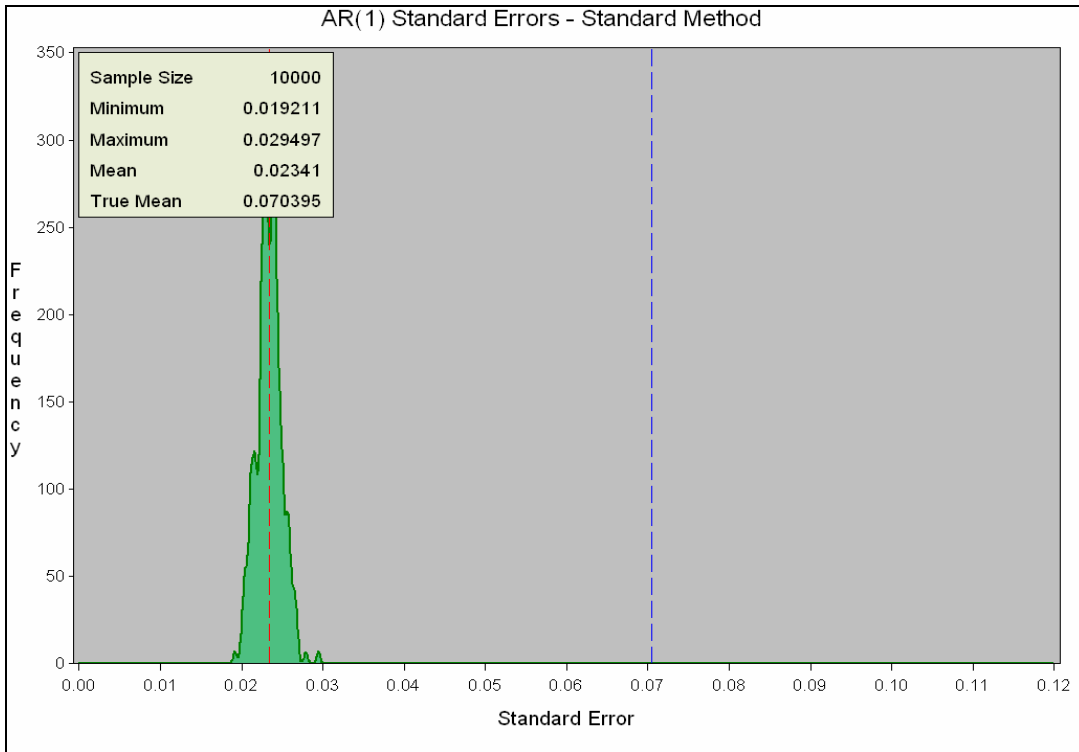
## 6 References

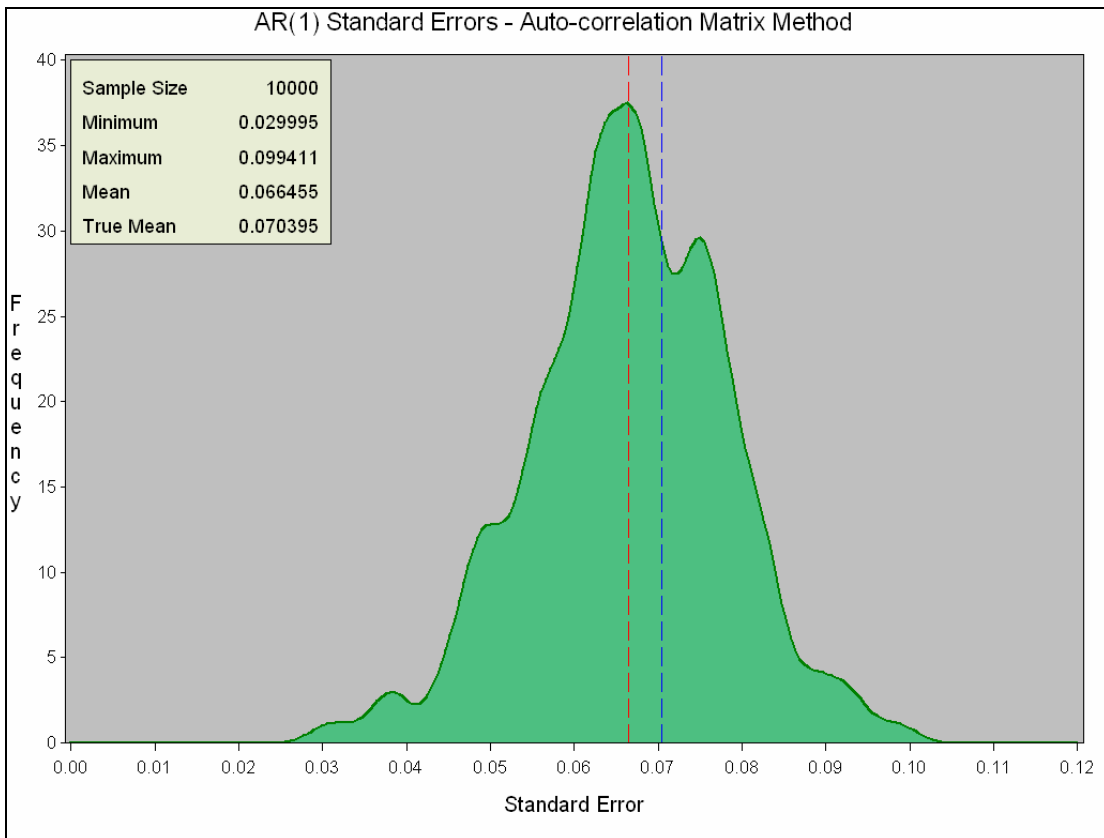
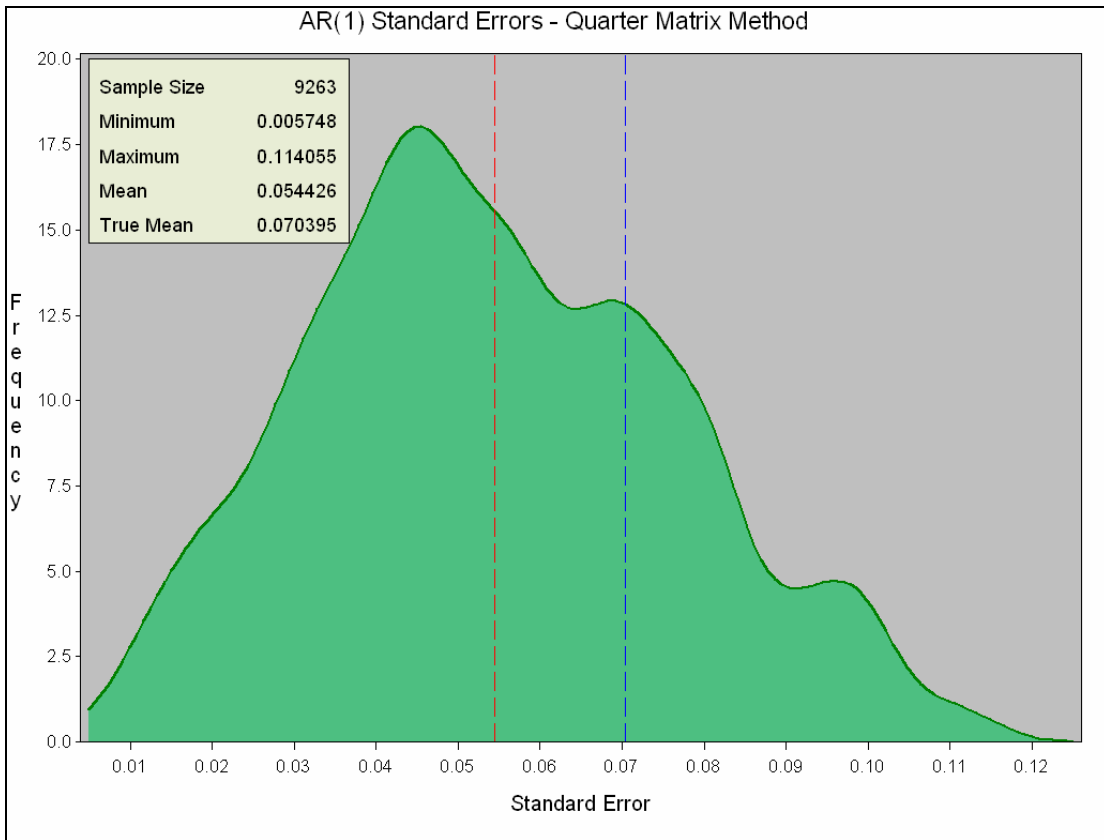
- [1] Microsoft Encarta Online Encyclopedia 2008. "Fiber Optics". <<http://encarta.msn.com>> (Accessed 29 Dec 2008).
- [2] Dal Forno A. O, Paradisi A, Passy R, von der Weid J. P. Experimental and Theoretical Modeling of Polarization-Mode Dispersion in Single-Mode Fibers. IEEE Photonics Technology Letters. Vol 12, No. 3, March. 2000.
- [3] Wikipedia. "Polarization Mode Dispersion" <[http://en.wikipedia.org/wiki/Polarization\\_mode\\_dispersion](http://en.wikipedia.org/wiki/Polarization_mode_dispersion)>. (Accessed 30 Oct 2008).
- [4] Whitman J R. Polarization Mode Dispersion (PMD): What is the statistical method for determining Link PMD and why is it important. Corning Optical Fibre White Paper. Issue WP3229. 1999.
- [5] Guidelines magazine, Summer 2001 edition. "Welcome the Light". <[www.corning.com/WorkArea/downloadasset.aspx?id=7109](http://www.corning.com/WorkArea/downloadasset.aspx?id=7109)>. (Accessed 30 Oct 2008).
- [6] Gisin N, Gisin B, Von der Weid, J P, Passy R. How accurately can one measure a statistical quantity like Polarization Mode Dispersion? Photonics Technology Letters, IEEE, Vol 8, No. 12, December 1996.
- [7] Viljoen, P E. Polarization Mode Dispersion: impact on Optical Fibre Systems. Corning Ltd. 1999.
- [8] Kelly M, Wunderling M. PMD measurement of fibre links: How much wavelength range is needed? Agilent Technologies Application Note. 2004.
- [9] Ten S, Edwards M. An Introduction to the Fundamentals of PMD in Fibers (White Paper). Corning. 2006.
- [10] Viljoen L, Conibear A B, Leitch A W R. Temperature Dependence of Polarization Mode Dispersion. Department of Physics, University of Port Elizabeth. 2004.
- [11] Gordon, J P, and Kogelnik, H. PMD fundamentals: Polarization mode dispersion in optical fibres. PNAS, Vol 97, No 9. 2000.

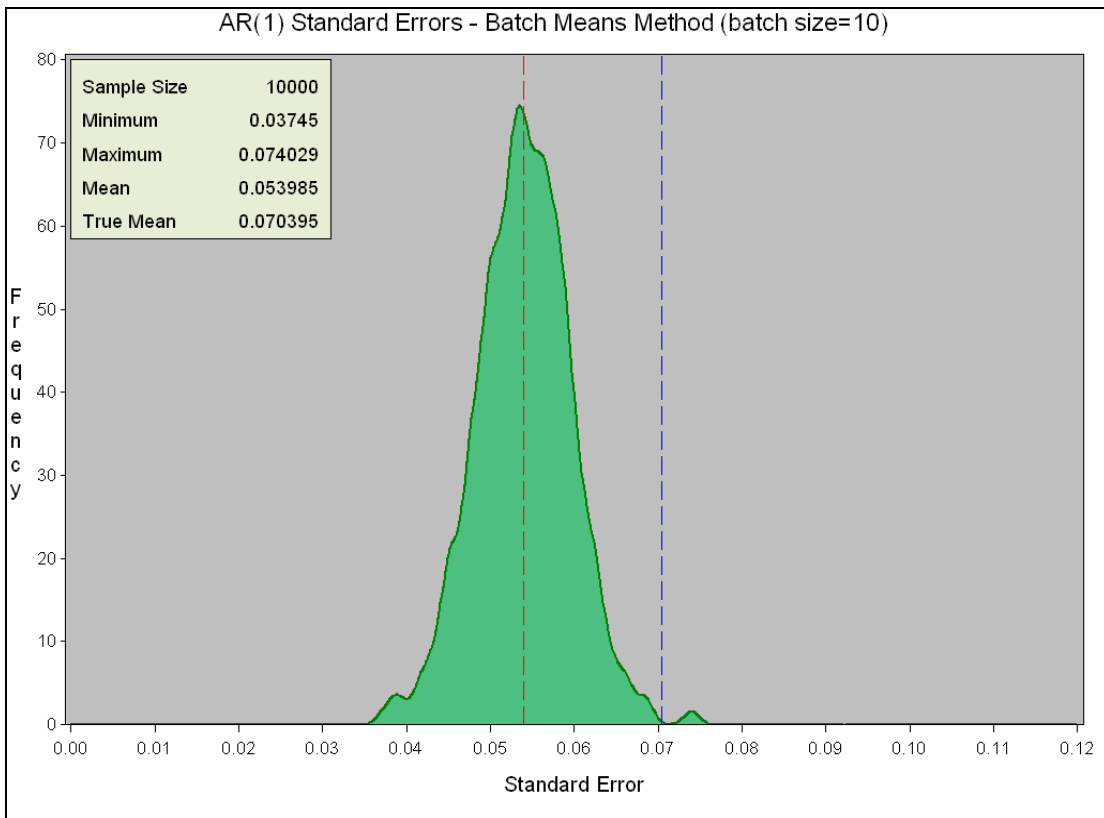
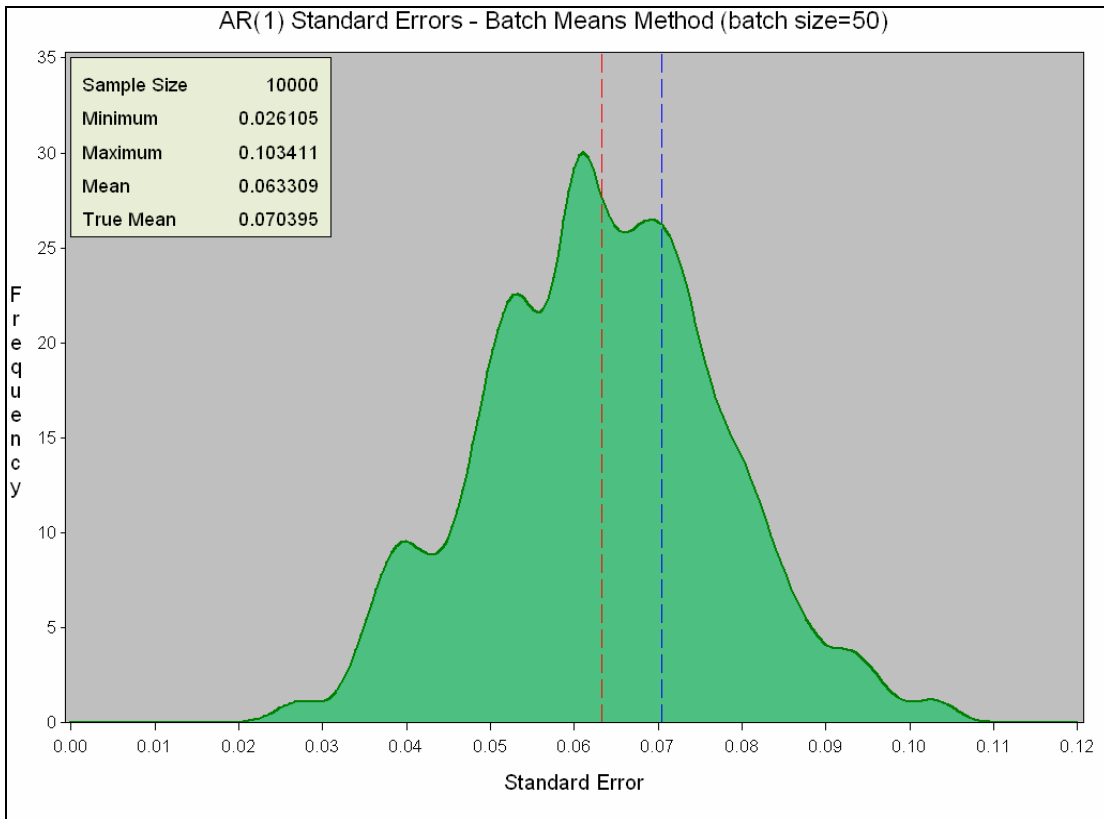
- [12] Gisin, N, Von der Weid, J. P, Pellaux, J. P. Polarization Mode Dispersion of short and long single mode fibres. *Journal of Lightwave Technology*. Vol 9, No 7. 1991.
- [13] Nolan D A, Chen X and Li M-J. Fibres with low Polarization Mode Dispersion. *Journal of Lightwave Technology*. Vol 22, No 4. 2004.
- [14] Gisin, N. Polarization Mode Dispersion: Definitions, Measurements and Statistics. SOFM'94, NIST, Boulder 1994.
- [15] Gisin, N., Passy, R, Bishoff, J. C, Perny, B. Experimental Investigations of the Statistical Properties of Polarization mode Dispersion In Single Mode Fibres. *IEEE Photonics Technology Letters*, Vol. 5, No. 7, July. 1993.
- [16] Box, G E, Jenkins, G M. *Time Series Analysis: Forecasting and Control*. Holden-Day. 1976.
- [17] Bartlett, M S. On the Theoretical Specification of Sampling Properties of Autocorrelated Time Series. *Journal of the Royal Statistical Society, ser. B8*, Vol. 27. 1946.

# 7 Appendix

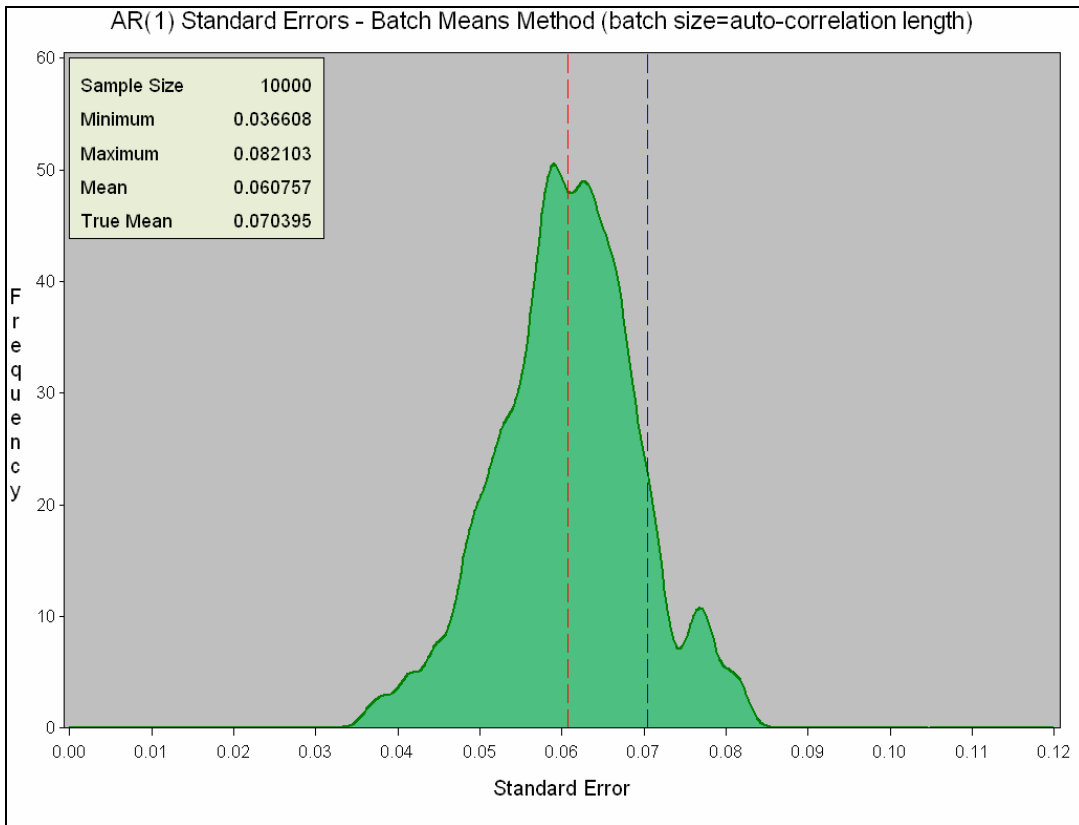
## AR(1) data – Various methods



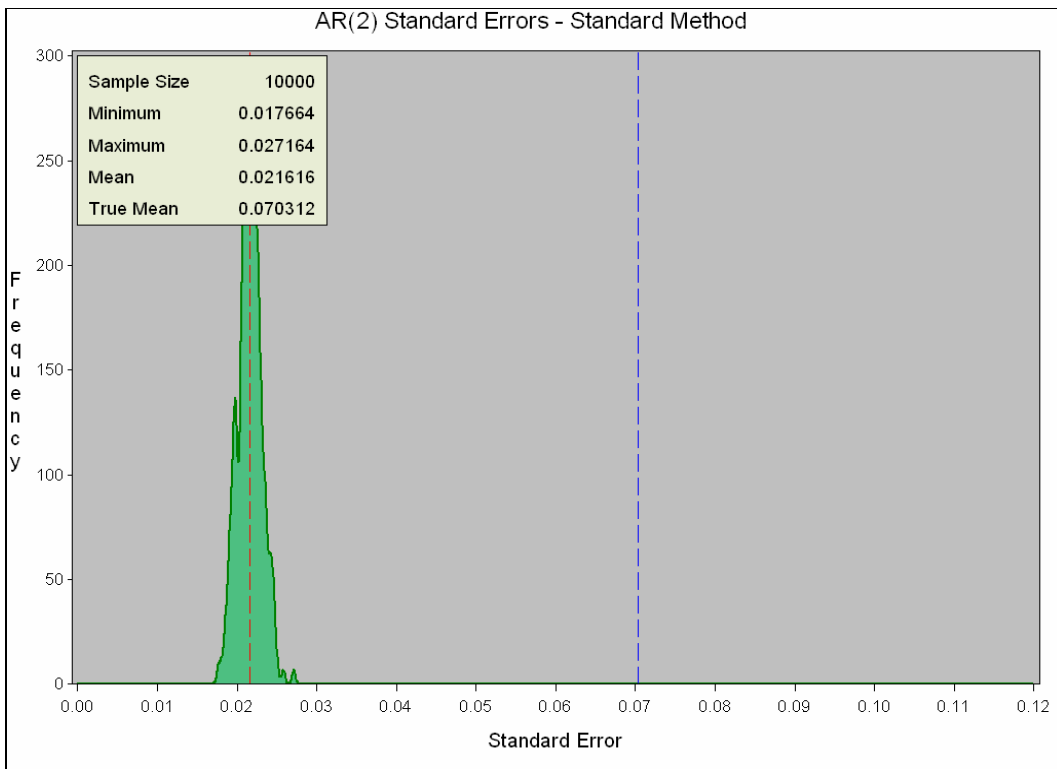


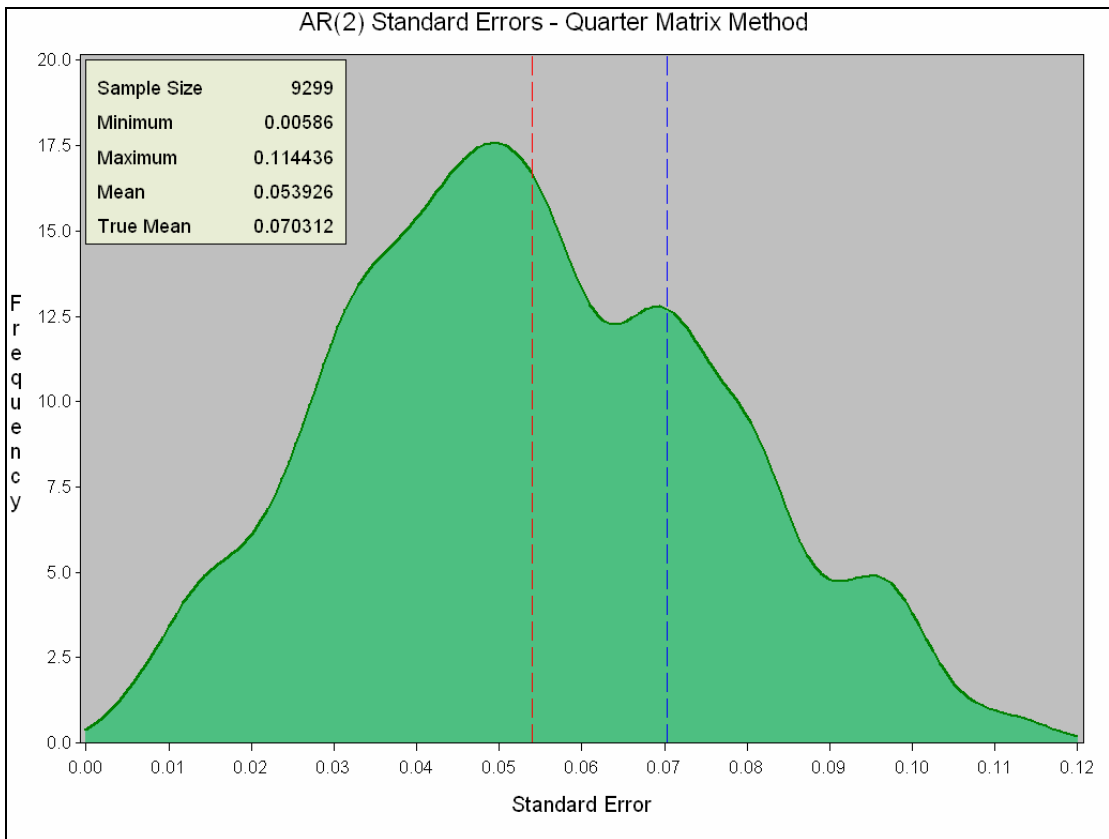
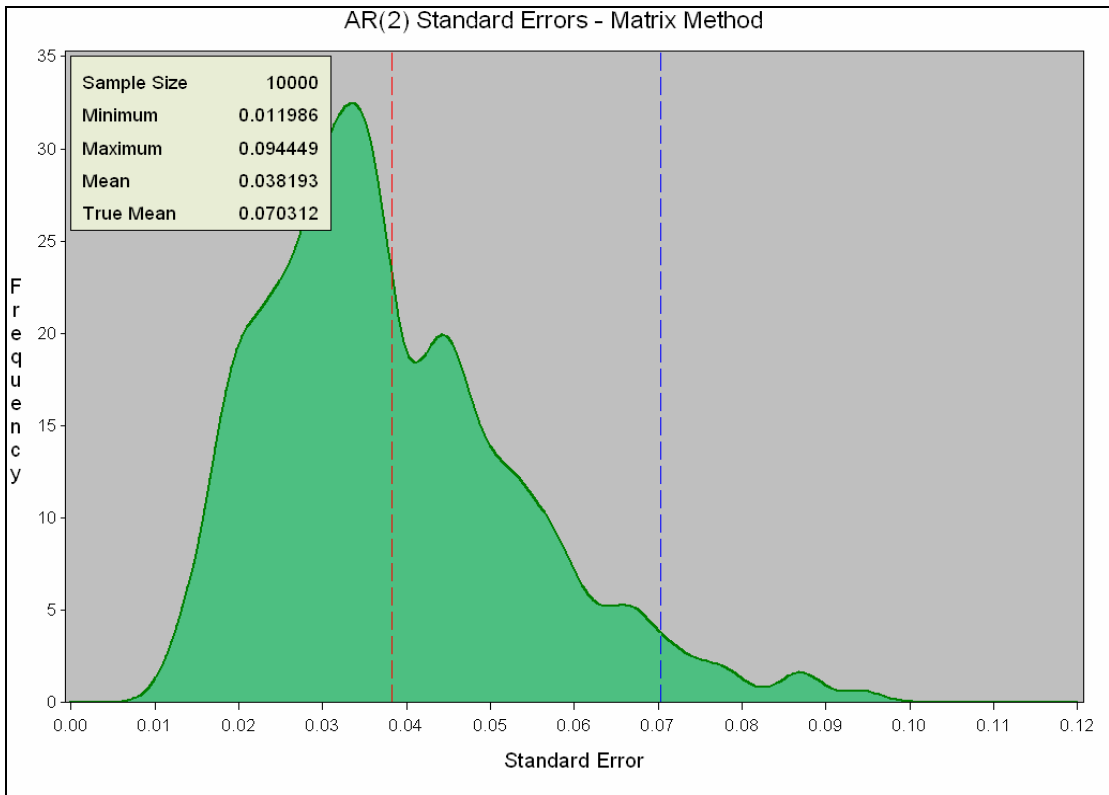


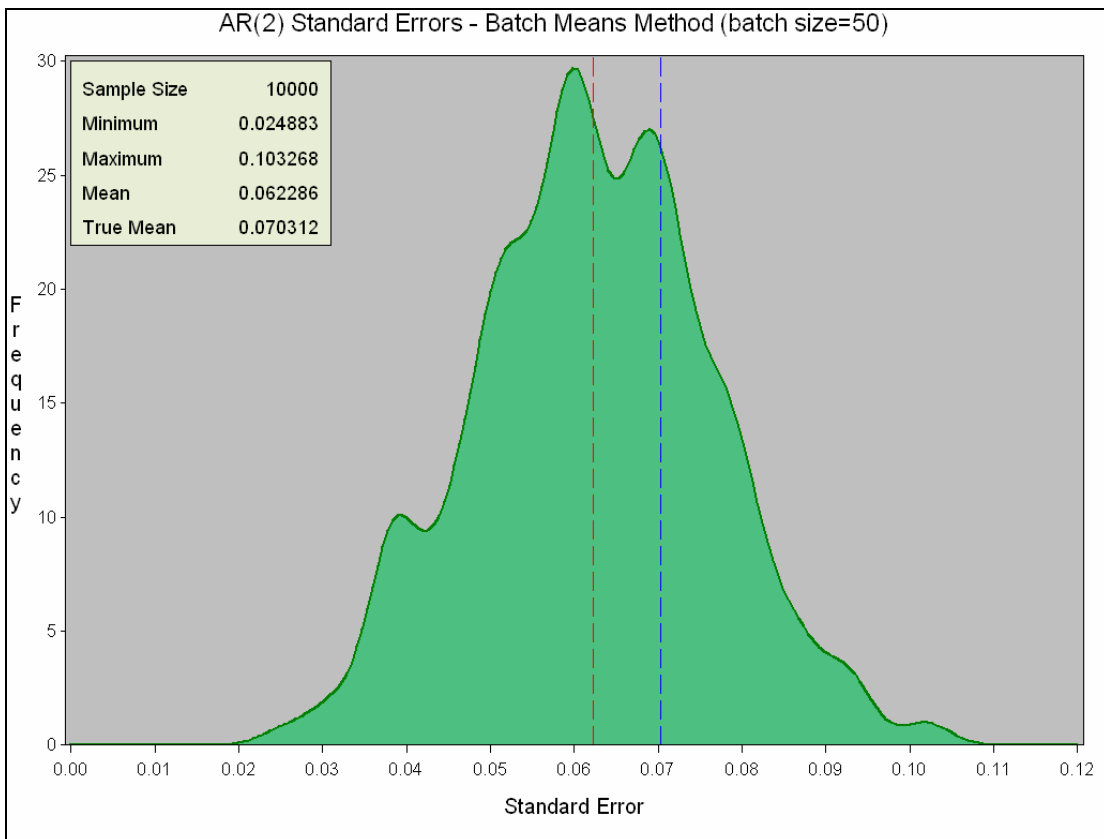
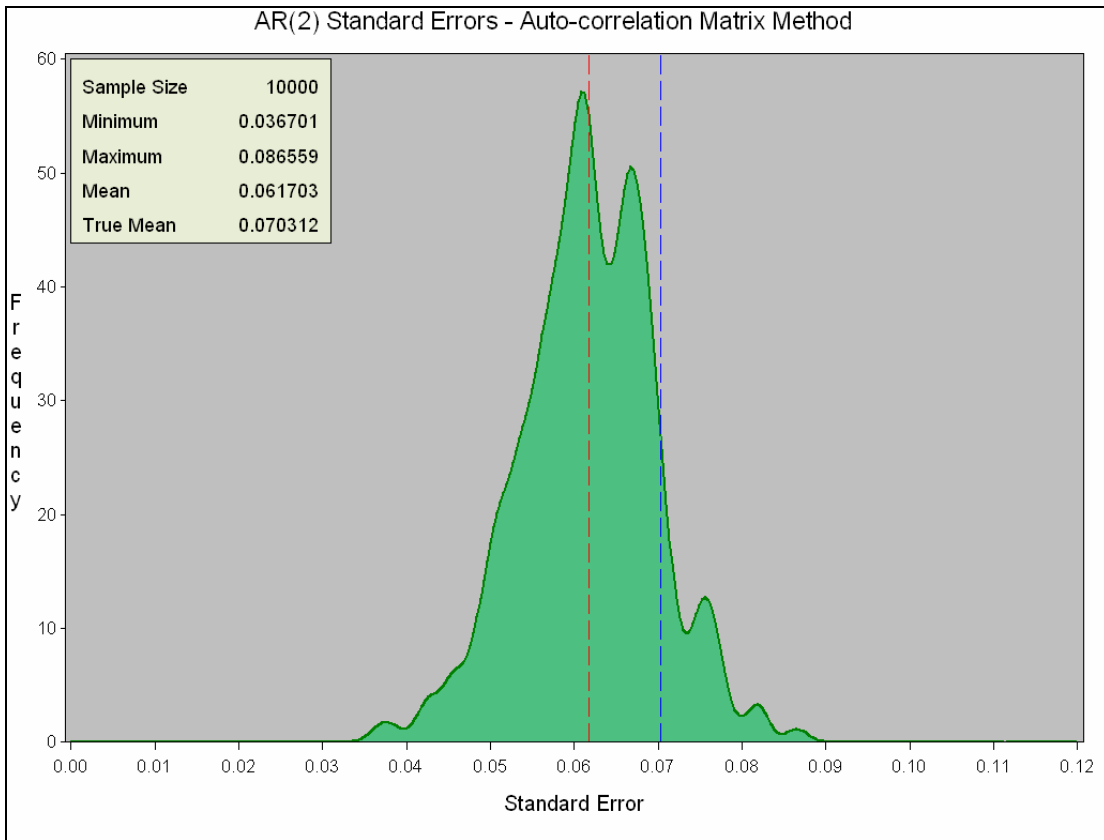


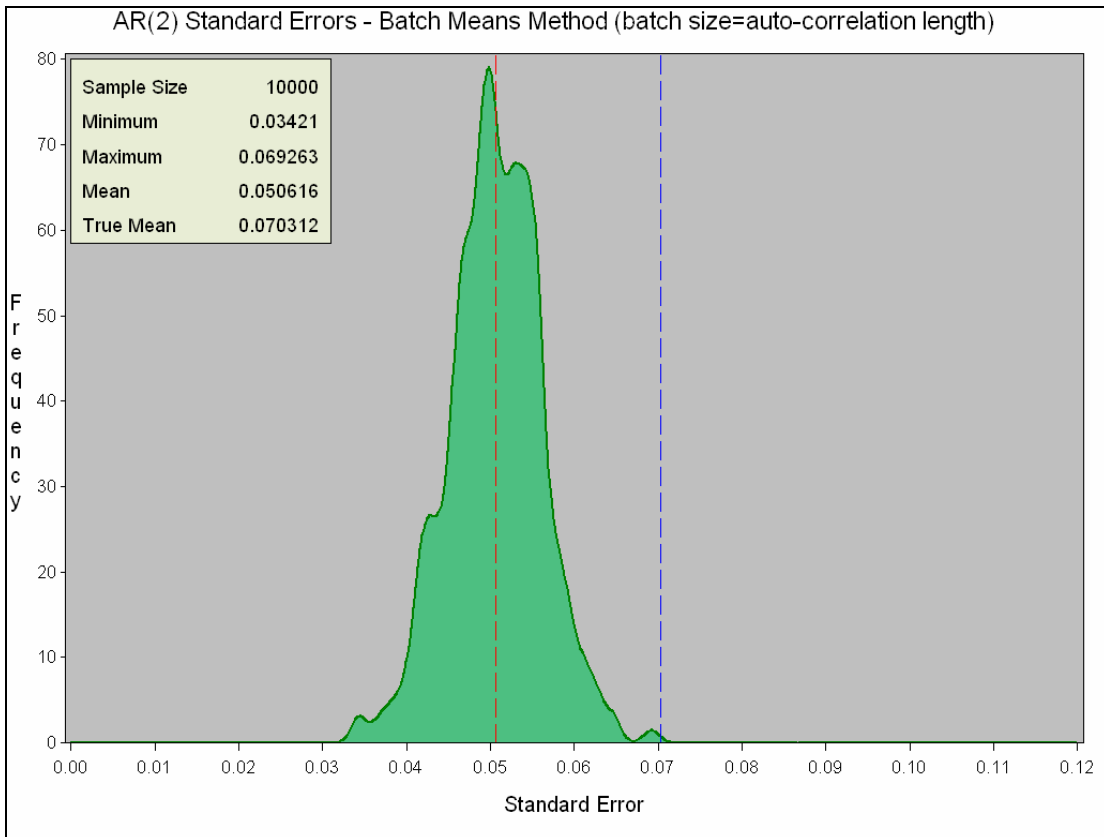
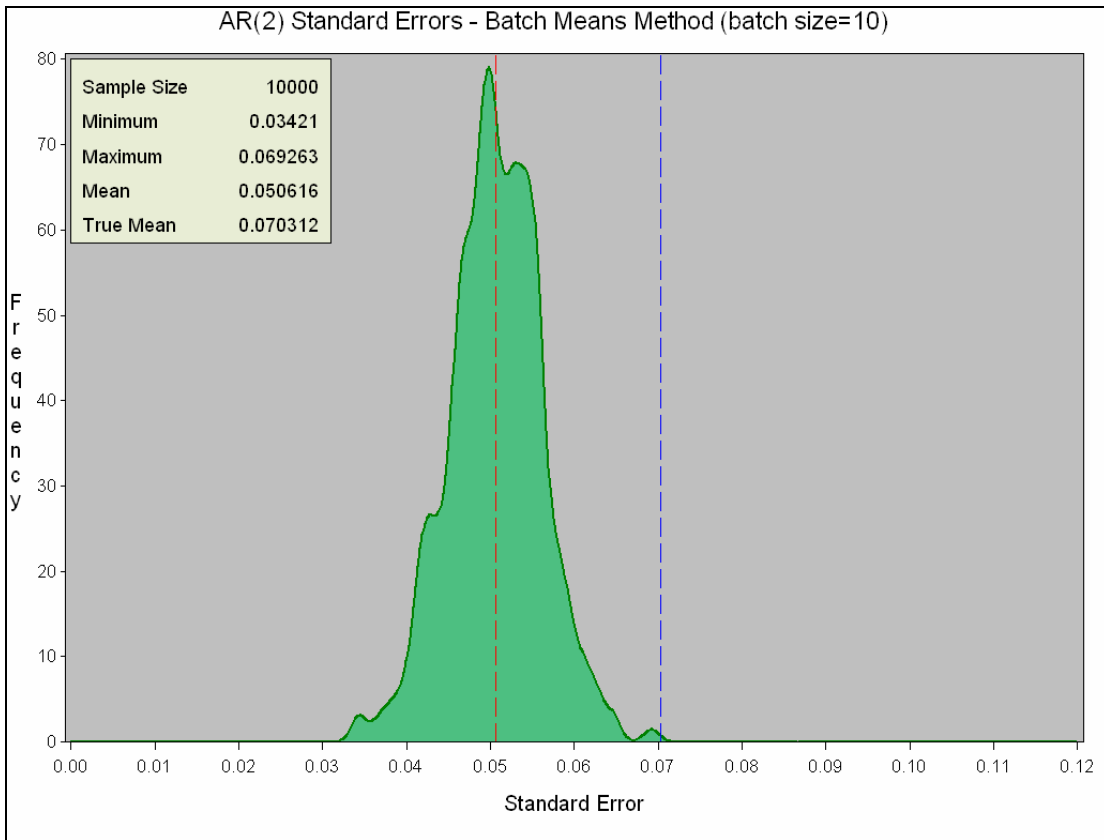


### AR(2) – Various methods

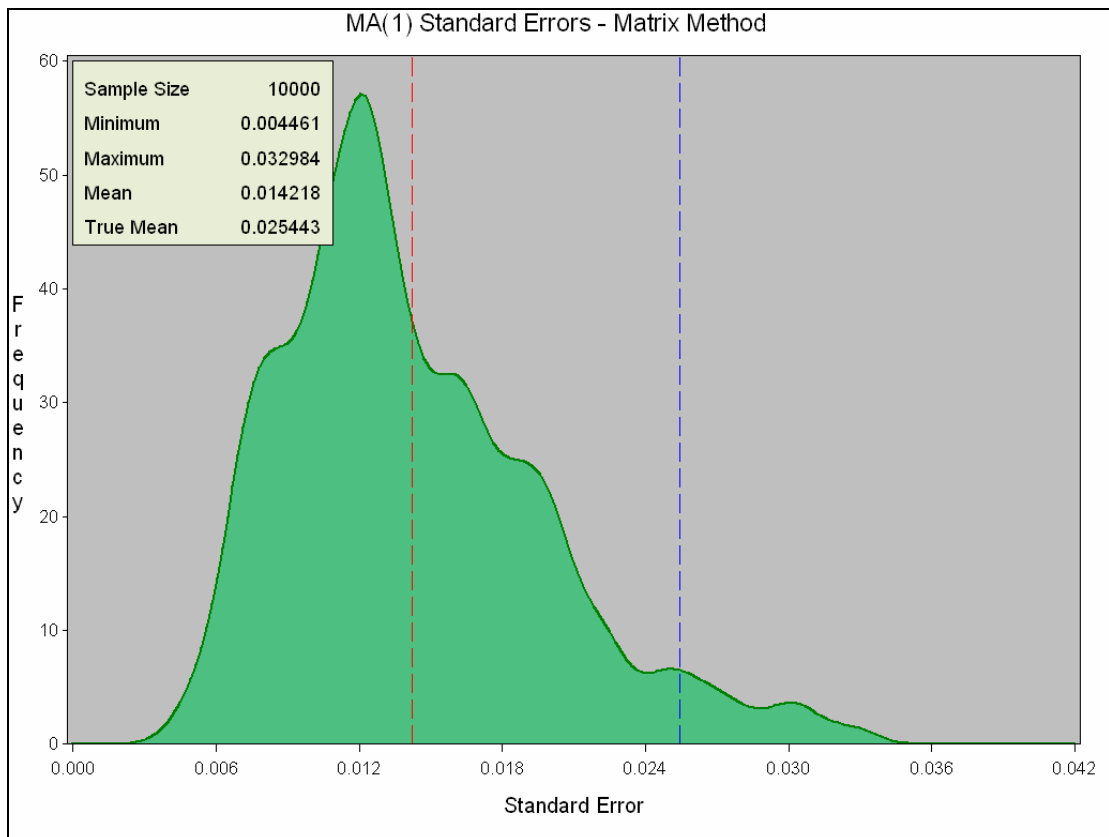
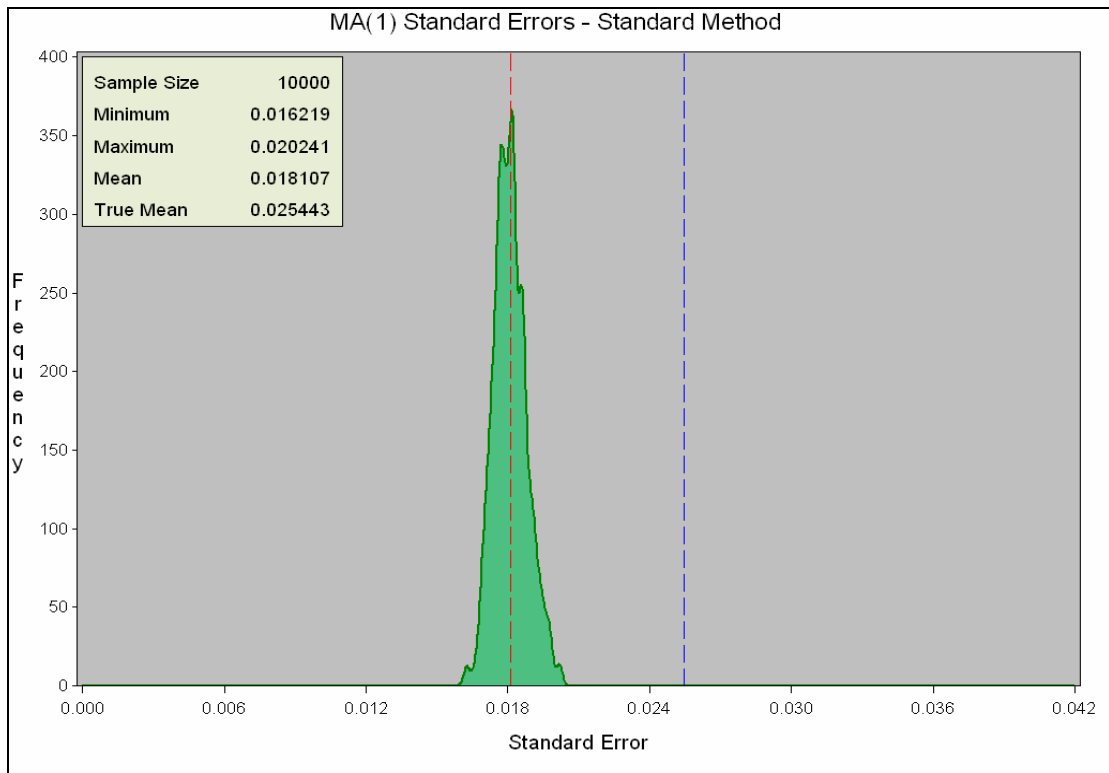


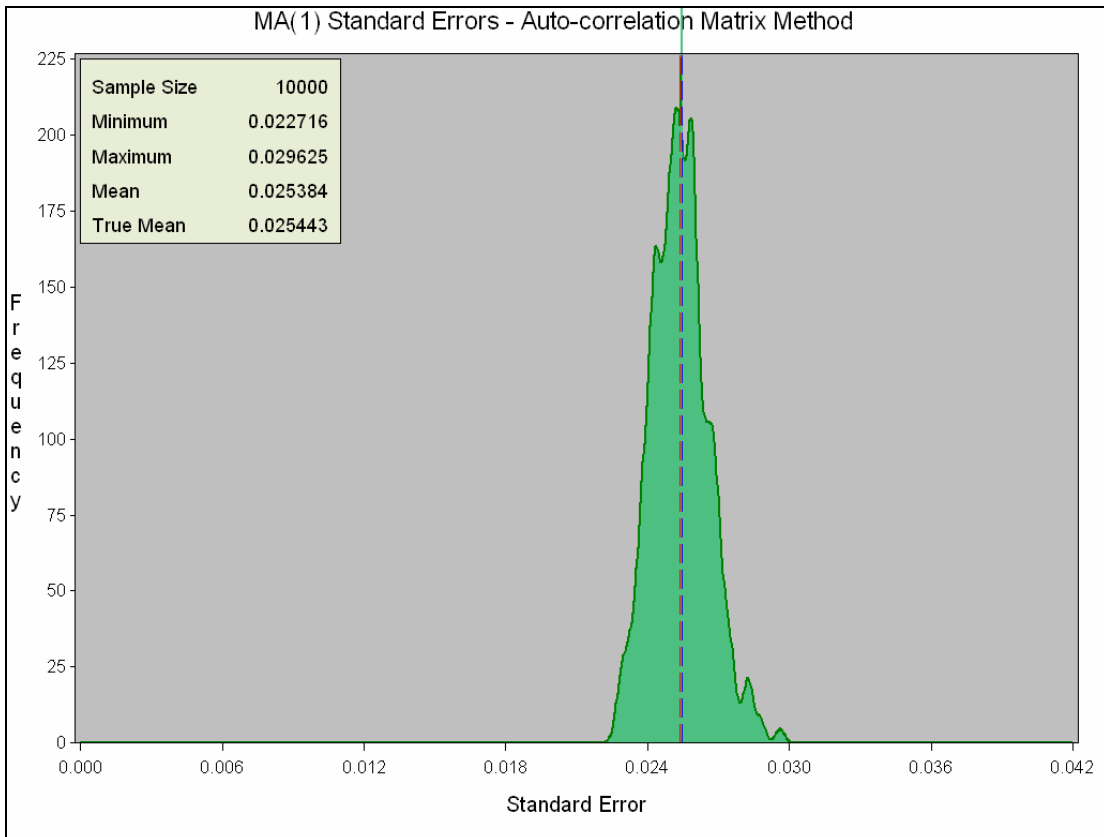
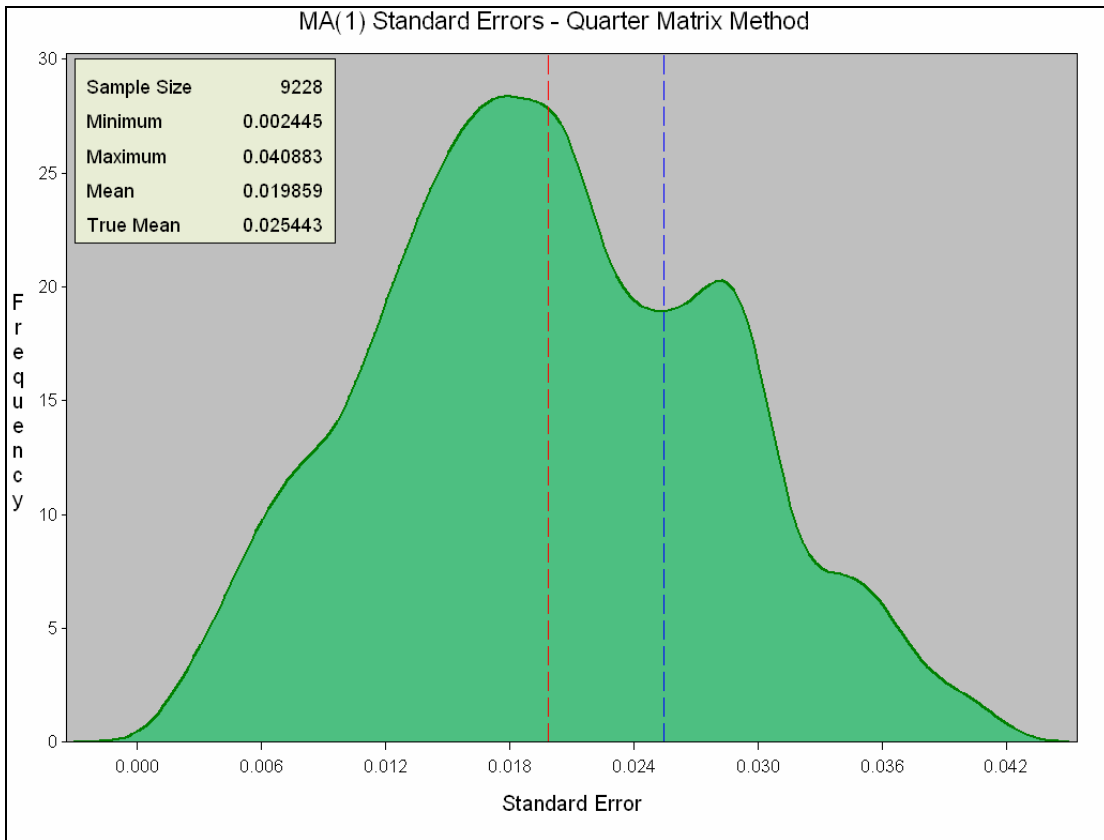


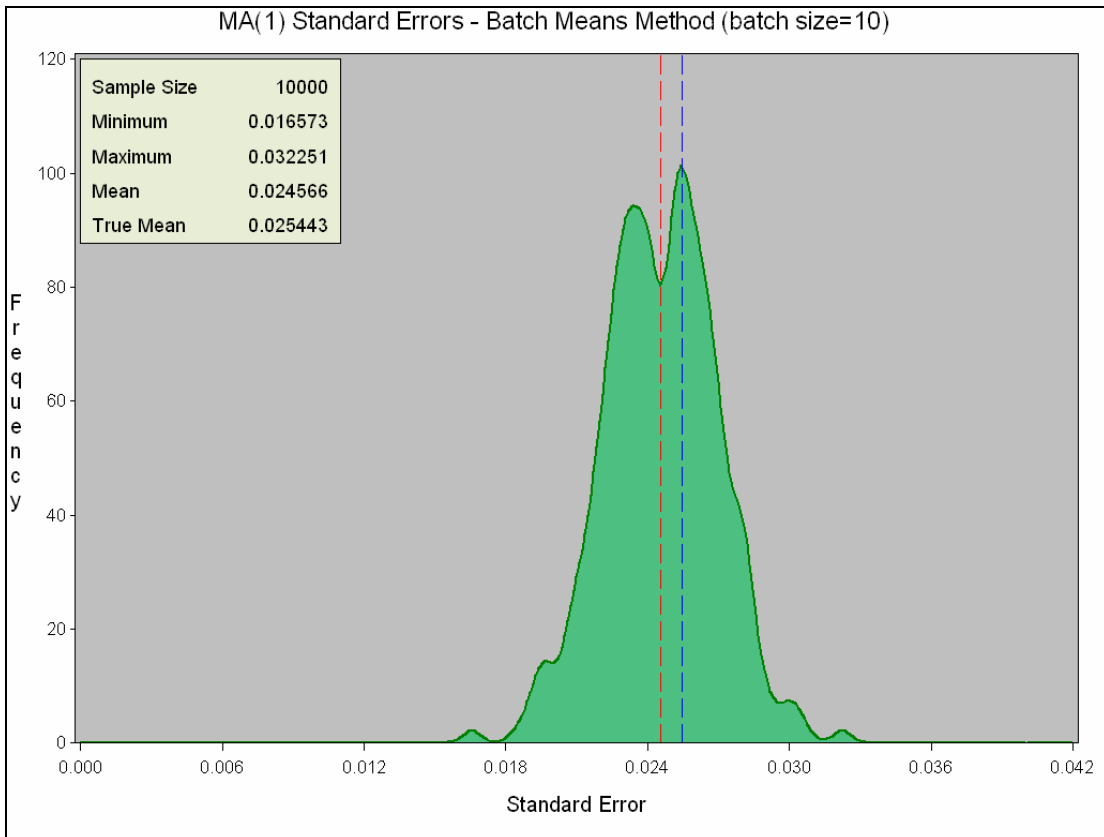
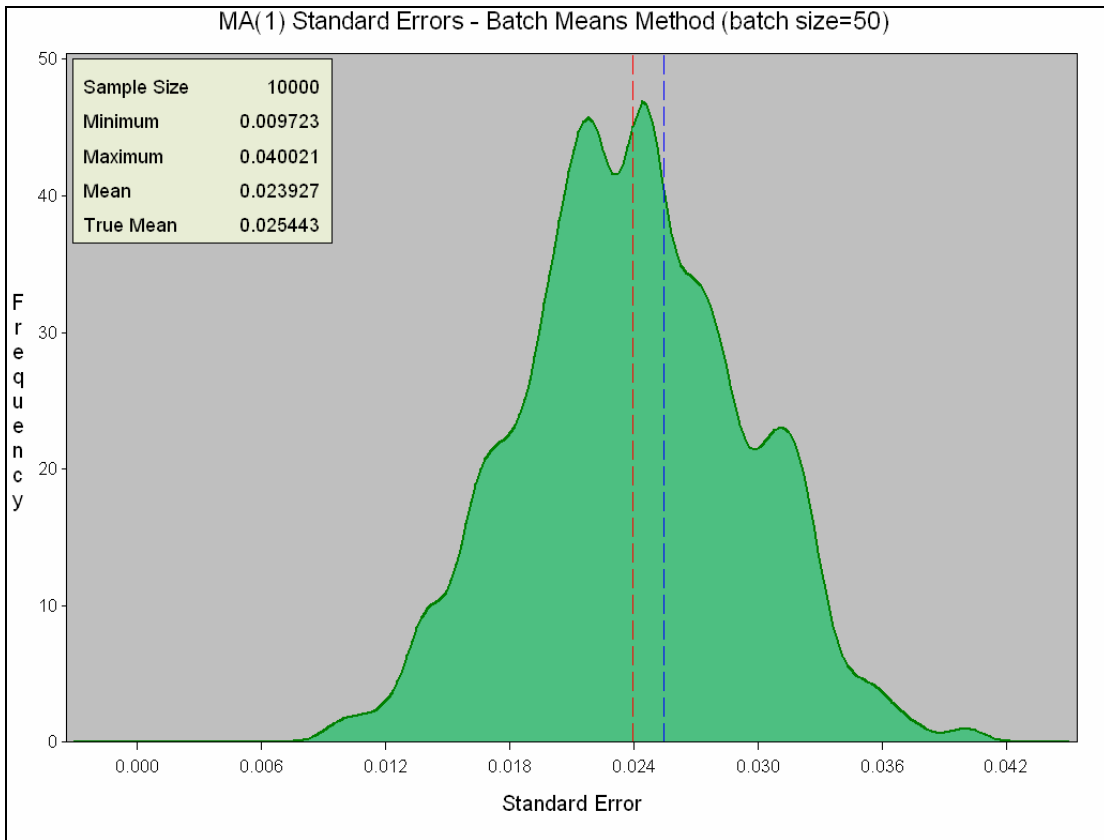


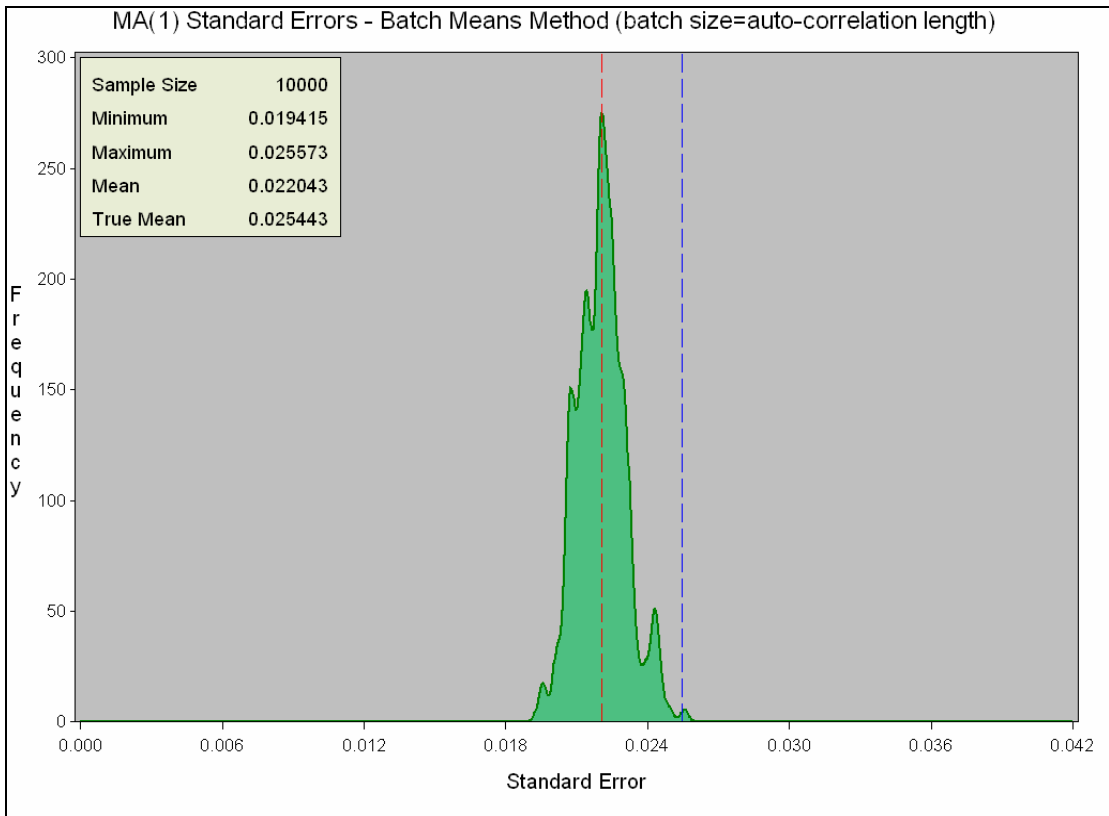


## MA(1) data – Various methods

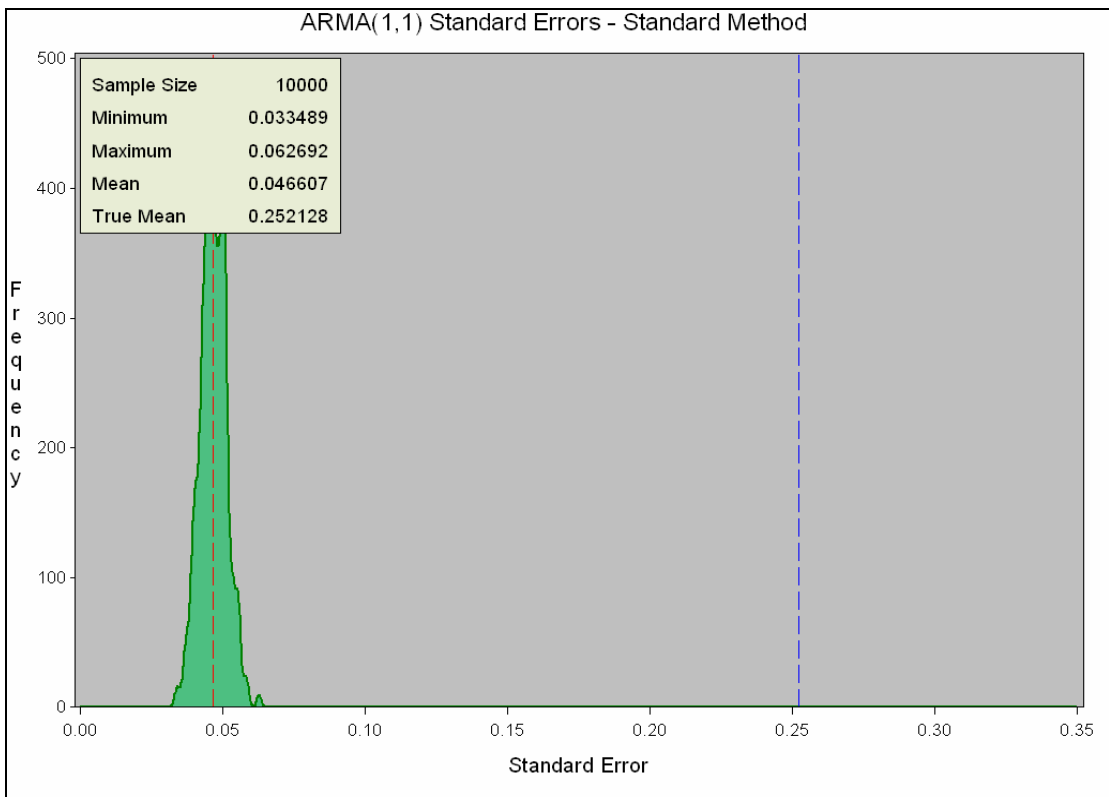




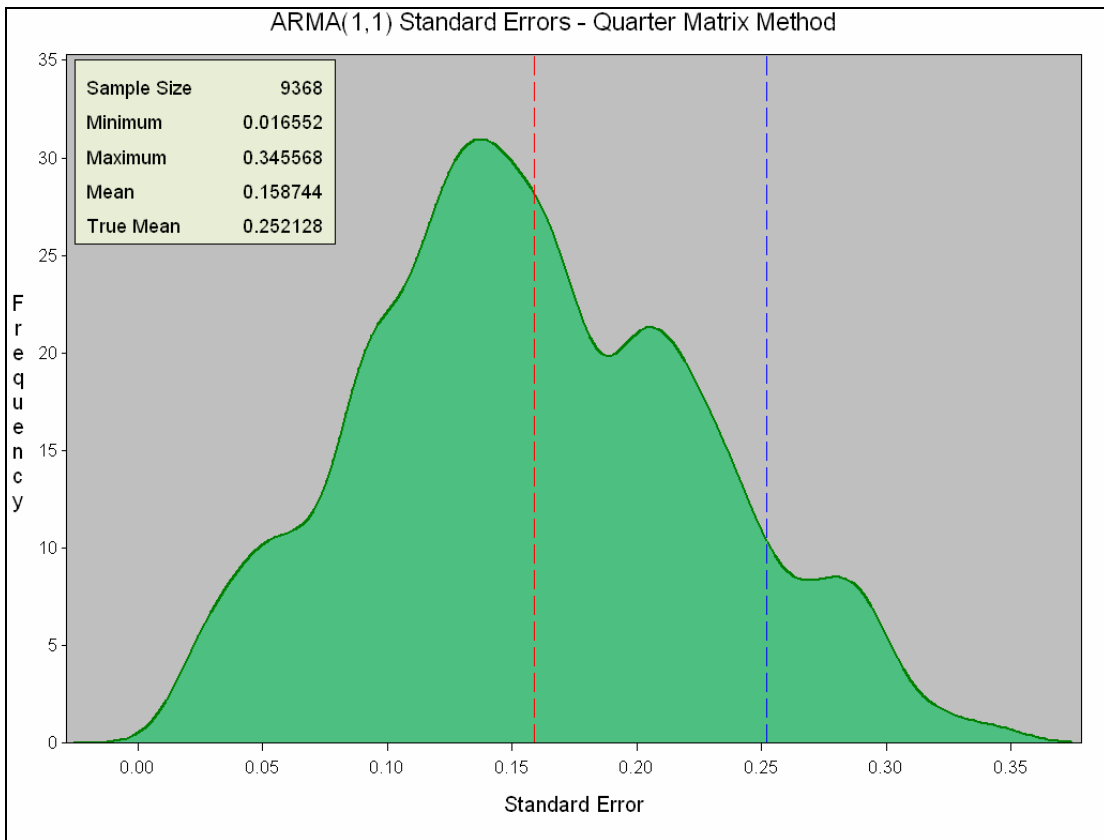
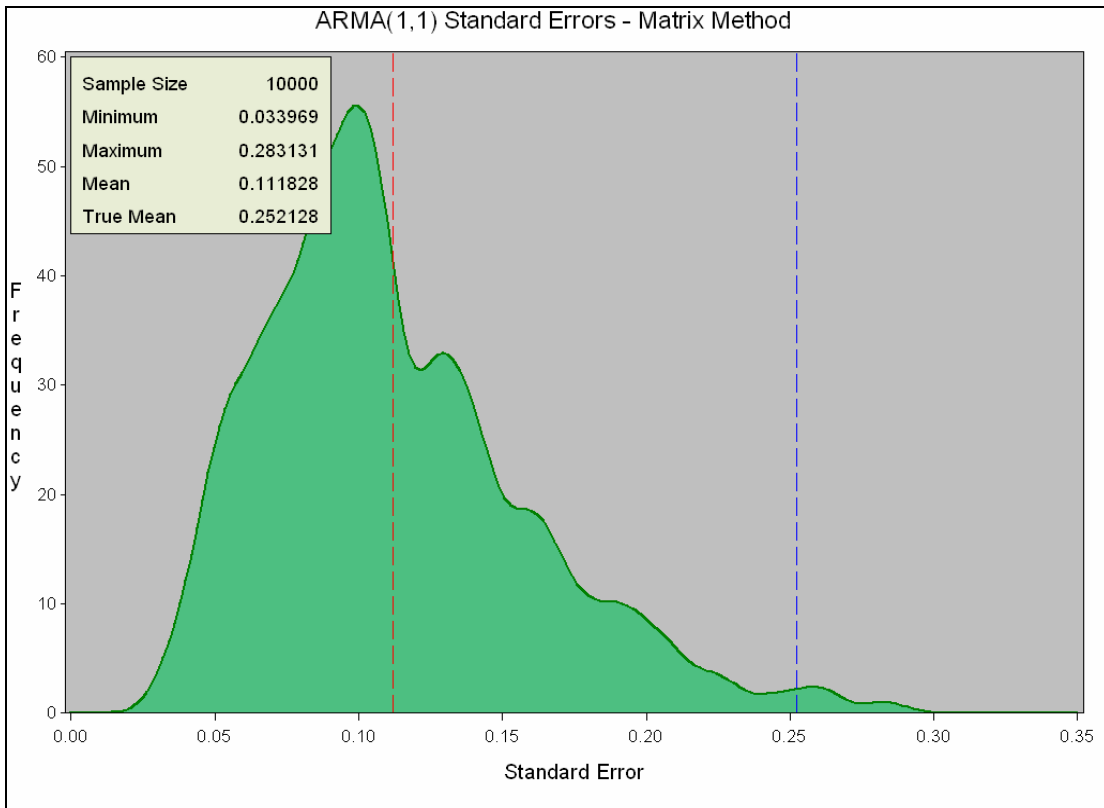


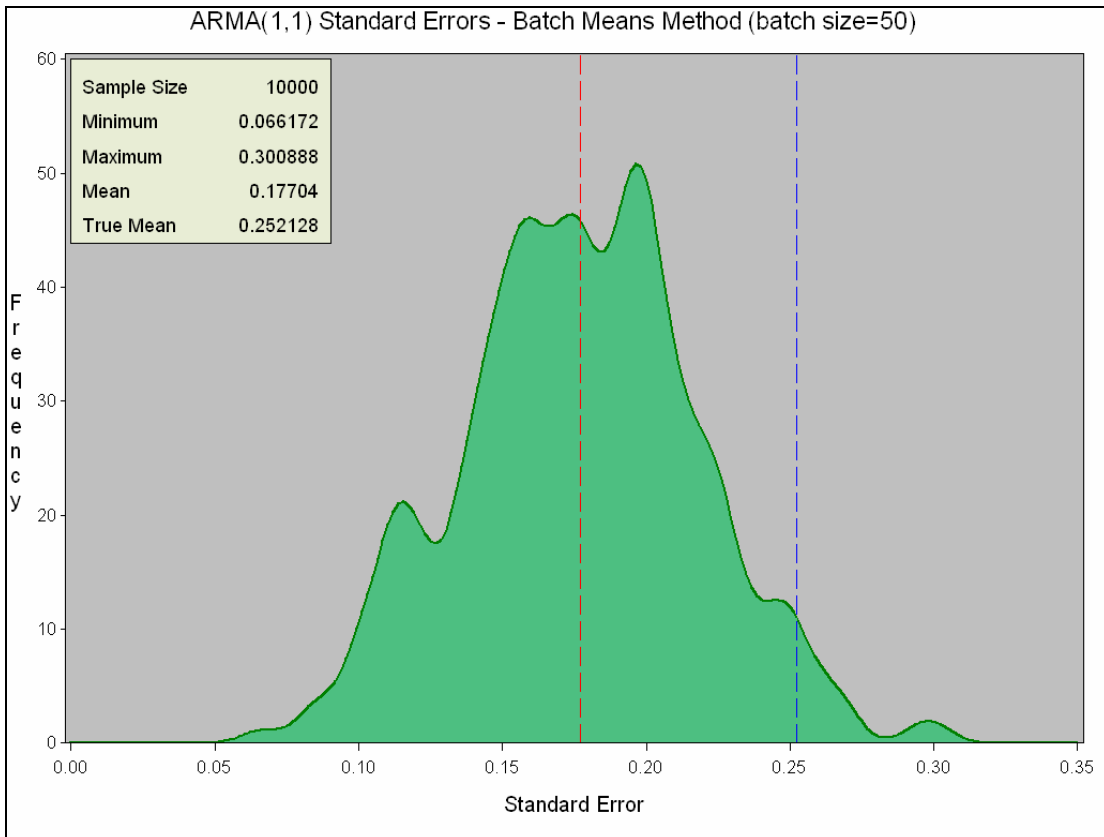
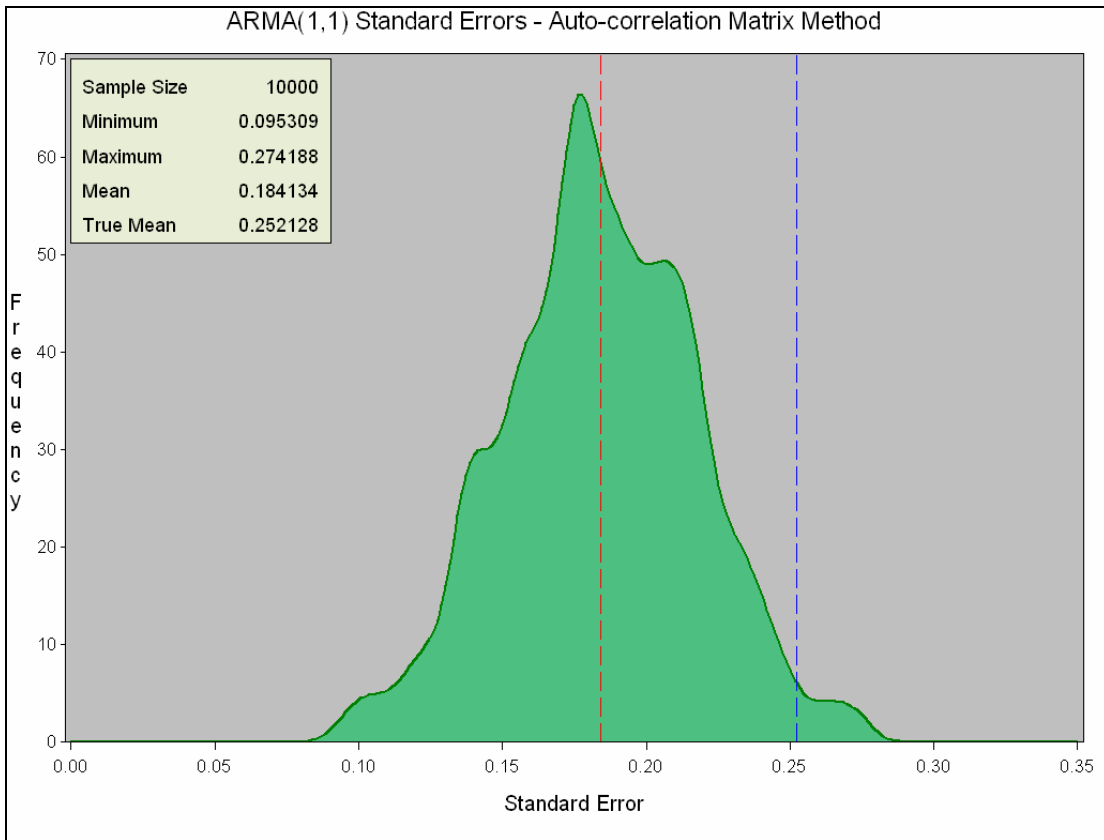


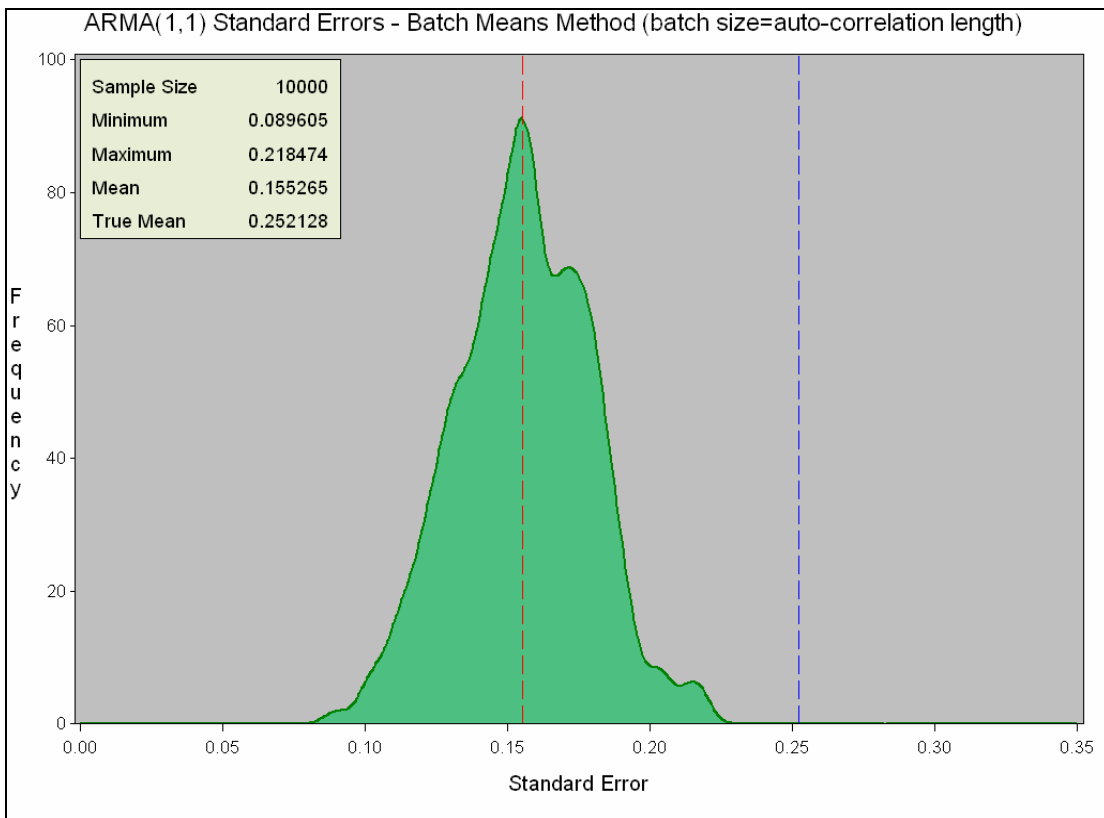
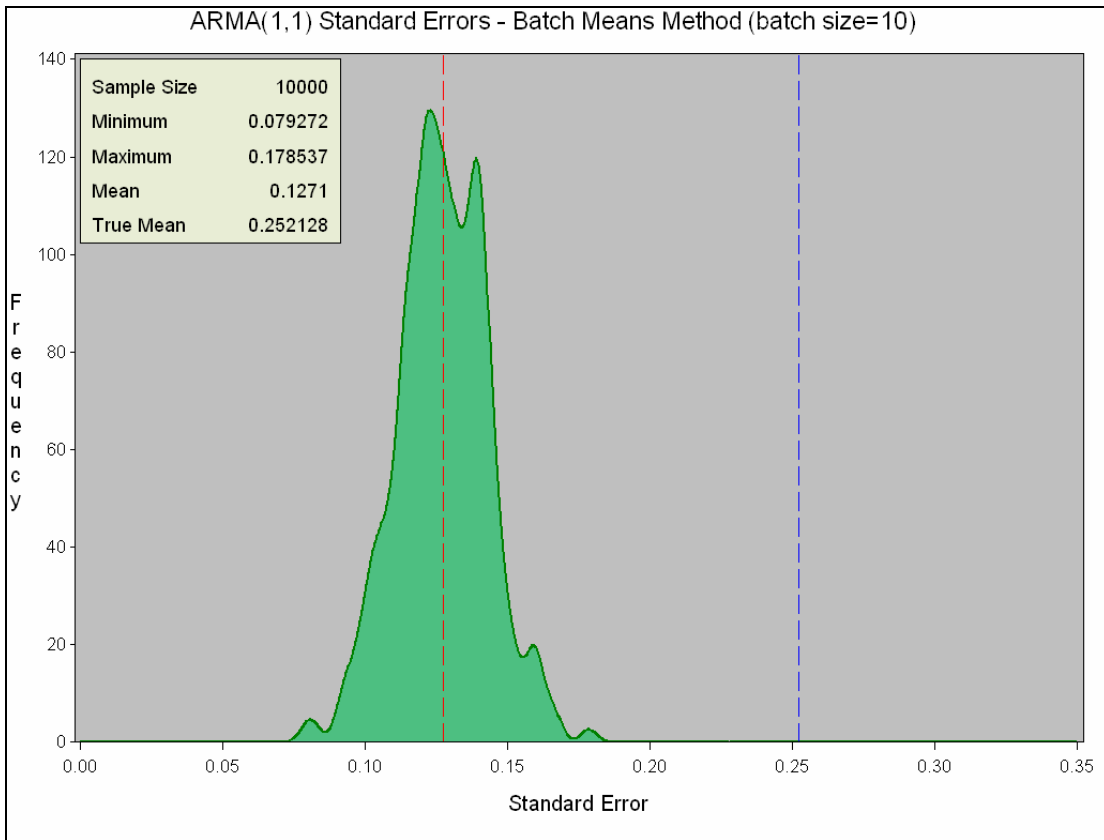
**ARMA(1,1) data – Various methods**











### Normal (5,0.1) data – Various methods

