EFFECTIVE SAMPLE SIZE IN ORDER STATISTICS OF CORRELATED DATA

by

Neill McGrath

A thesis

submitted in partial fulfillment

of the requirements for the degree of

Master of Science in Mathematics

Boise State University

May 2009

The thesis presented by *Neill McGrath* entitled *Effective Sample Size in Order Statistics of Correlated Data* is hereby approved.

Dr. Kyungduk Ko, Advisor	Date
Dr. Jaechoul Lee, Committee Member	Date
Dr. Jodi Mead, Committee Member	Date
John R. Pelton, Graduate Dean	Date

Dedicated to Nana

ACKNOWLEDGEMENTS

Many thanks to Dr. Kyungduk Ko

ABSTRACT

In this study we address the problem of using effective sample size (ESS) to approximate the probability distributions of order statistics from correlated data. We present these approximations and determine their accuracy through simulation studies. More often than not, correlation exists between data points in a set of data. When we use the original sample size of the data in a derivation of a model of the data, we automatically assume that each data point contains one data point's worth of information. If the data are correlated, then each data point contains less than one data point's worth of information making our assumption false. This is especially true in the case of data with a very high level of correlation. Effective sample size represents essentially how many pieces of uncorrelated information the sample would compare to and this is often much smaller than the original sample size. Here we calculate effective sample size which we then use in place of the original sample size. We use a method discussed by Thiebaux and Zwiers [9] for the calculation of effective sample size and show its usefulness using an application to the approximation of the probability distributions of order statistics in correlated data, and finally, we compare our results with simulated data.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

This study focuses on an application of effective sample size to the order statistics for correlated data. Effective sample size in correlated data provides a method to determine how many independent observations exist within a sample. Each independent observation is known to contain a preset amount of information. Determining how many observations in the sample represents one independent observation is the job of effective sample size [5].

When data are uncorrelated, it is seen that effective sample size is just the size of the sample. However, when data are correlated, effective sample size gets smaller than the sample size. We specifically consider the maximum and minimum order statistics in correlated data and approximate their distribution through the idea of effective sample size.

Thiebaux and Zwires have given multiple techniques for calculating effective sample size but do not show specific applications of these techniques. Work done by Laurmann and Gates has stressed the importance of effective sample size when working with correlated data, but do not clearly define how to calculate it. The importance of effective sample size has also been noted in [6], [7], [8], and [1], but they did not use

it in their calculations. Effective sample size was used by Ko and Lee [4] to calculate confidence intervals for long memory regression. Here we use effective sample size to approximate the distributions of order statistics in time series.

Effective sample size has been recognized as an important issue in atmospheric circulation models. Experiments involving atmospheric measurements of a process whose distribution does not depend on time, known as a stationary process, can be taken successively with the passage of very small amounts of time. Such a stationary time series is a process that is assumed to have no correlation between observations. The atmosphere does not have this property because such measurements have some dependency. Effective sample size can be used to take into account the dependency in atmospheric measurments [5].

The remainder of this thesis is organized as follows: In Chapter 2 we introduce the basics: the definition of order statistics and the concept of effective sample size, which are needed to understand the study. In Chapter 3 we present the autoregressive and moving average models, the calculations of effective samples sizes in the maximum and minimum order statistics for the models. We show how to apply effective sample size to the probability distributions of maximum and minimum order statistics from AR(1) and MA(1) models. In Chapter 4 simulation studies are given. Concluding remarks and further studies are given in Chapter 5.

CHAPTER 2

PRELIMINARIES

2.1 Order Statistics

Suppose that $X_1, X_2, ...X_n$ are independent and identically distributed with a probability density function $f(x)$ and a cumulative distribution function $F(x)$. Order statistics are obtained from sorting the data from the least to the greatest, $X_{(1)} \leq X_{(2)} \leq .... \leq X_{(n)}$. The probability distribution of order statistic, $X_{(k)}$ [3] is , Suppose that $X_1, X_2, ...X_n$ are independent and identically distributed with a probability density function $f(x)$ and a cumulative distribution function $F(x)$. Order statistics are obtained from sorting the data from the least to the greatest, $X_{(1)} \leq X_{(2)} \leq .... \leq X_{(n)}$. The probability distribution of order statistic, $X_{(k)}$ [3] is ,

$$g_k(X_{(k)}) = \frac{n!}{(k-1)!(n-k)!}[F(X_{(k)})]^{k-1}[1-F(X_{(k)})]^{n-k}f(X_{(k)}). \qquad (2.1)$$

For the maximum order statistic $X_{(n)}$ its probability distribution is simplified to

$$g_n(X_{(n)}) = n[F(X_{(n)})]^{n-1}f(X_{(n)}).$$

Similarly, we have the following simplification for the probability distribution of the minimum order statistic $X_{(1)}$:

$$g_1(X_{(1)}) = n[1 - F(X_{(1)})]^{n-1} f(X_{(1)}).$$

The maximum and minimum order statistics are important order statistics which are used in many disciplines. In time series, they represent the distribution of the highest reading recorded and the lowest reading recorded, respectively. For example, the amount of water flowing into a reservoir at a given point in time may determine if the dam will fail. Then one might be interested in the distribution of the maximum order statistic in this case because it will determine the probability the dam will withstand yearly river fluctuations into the distant future.

As can be seen, the distributions of order statistics depend on the sample size n. They are only accurate when the data are independent and identically distributed. However, the distribution (2.1) cannot be applied to correlated data. Our goal is to use the distribution (2.1) of the maximum and minimum order statistics for correlated data by replacing the original sample size n with an effective sample size n_e. To the best of our knowledge, this work has not been done yet in literature.

2.2 Effective Sample Size

Rarely in the sampling process do we obtain data that are made up of completely uncorrelated observations. In time series there will always be some sort of correlation between data points and a sample may not contain as much information as it would first appear to contain. For example, if we have two temperature readings taken an hour apart from each other, there is going to be some dependence between the two readings. This means that knowing information about one of the observations tells us some of the information about the other observation. The question then arises: how does this independence issue affect the information obtained in the sample? If data points are correlated, then we have effectively less information about the total population than the sample size indicates. That is, the size of the sample is not the effective size given the amount of information the sample actually contains.

We consider a method to obtain this effective sample size, n_e, discussed by Thiebaux and Zwiers [9], which depends on the variance of the sample mean of the independent data, $Var(\bar{Y}_{IID})$ and the variance of the sample mean of the correlated data, $Var(\bar{Y}_{CORR})$. The effective sample size can be calculated by the relation [4]:

$$\frac{Var(\bar{Y}_{IID})}{Var(\bar{Y}_{CORR})} = \frac{n_e}{n}, \tag{2.2}$$

where n is the original sample size. Here $\bar{Y}_{IID}$ is the sample mean of assumed uncorrelated data and $\bar{Y}_{CORR}$ is the sample mean of assumed correlated data. A part of

this thesis is dedicated to calculating the variance of the correlated data.

For data that are correlated, the original sample size becomes less informative in measuring the amount of information and effective sample size may be a better choice for further statistical analysis. This is best illustrated in an example with 1000 data points that have a correlation coefficient of one because, in this case, even though we have 1000 points, one data point has all the information. Sample size can be replaced by effective sample size in order to create a more accurate model of data that is correlated.

The information we need in calculating effective sample size n_e is the original sample size and the amount of dependence within the data. A popular way to measure this dependence is by using the autocovariance function, $\gamma(h)$, which compares the variance of the data with a time-shifted version of itself and by using the autocorrelation function, $\rho(h)$. The size of the time shift h is called lag which denotes distance between observations.

The relationship between the autocovariance function and the autocorrelation function is [2],

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}, \quad h = 0, 1, 2, ..., n-1. \tag{2.3}$$

In time series, if a process has autocovariance and autocorrelation functions that depend only on the time lag h, the process is called stationary. In this study we assume a stationary time series.

CHAPTER 3

APPROXIMATION OF PROBABILITY DISTRIBUTION OF MAXIMUM AND MINIMUM ORDER STATISTICS

Here we focus on the approximate distributions of the maximum and minimum order statistics in the samples from the first order autogregressive model (AR(1)) and the first order moving average model (MA(1)). That is, we use these processes as a means for generating correlated data. Then the probability distributions of the maximum and minimum order statistics of those data are approximated.

In Section 3.1 autoregressive and moving average models are introduced with their autocovariance and autocorrelation functions. In Section 3.2 the computational formulas of effective sample sizes in the maximum and minimum order statistics of AR(1) and MA(1) models are derived. Also, their applications to the probability distributions of those order statistics are presented.

3.1 Autoregressive and Moving Average Processes

For the formulas and notations in this section, we follow the presentation of [2]. The pth order autoregressive model, AR(p) is defined by

$$\sum_{i=0}^{p} \phi_i X_{t-i} = e_t, \quad t = 1, 2, ..., n,$$

where the ϕ_is are the autoregressive coefficients with $\phi_0 = 1$ and e_t is white noise with mean zero and innovation variance σ^2. We assume the white noise term e_t has the normal distibution.

In this study we use a special case of the AR(p) model which is the AR(1) model. This model is represented by

$$(1 - \phi B)X_t = e_t, \quad t = 1, 2, ..., n,$$

where ϕ is the autoregressive coefficient and B is the backshift operator, $BX_t = X_{t-1}$.

The autocorrelation function $\rho(h)$ and autocovariance function $\gamma(h)$ of the AR(1) model are, respectively,

$$\gamma(h) = \frac{\phi^h \sigma^2}{1 - \phi^2}, \quad h = 0, 1, 2, ..., n - 1$$

and

$$\begin{aligned} \rho(h) &= \frac{(\phi^h \sigma^2)/(1 - \phi^2)}{\sigma^2/(1 - \phi^2)} \\ &= \phi^h, \quad h = 0, 1, 2, ..., n - 1. \end{aligned}$$

On the other hand, the pth order moving-average model for a time series $\{X_t\}_{t=1}^n$

is defined by

$$X_t = \sum_{j=0}^{p} \theta_j e_{t-j}, \quad t = 1, 2, ..., n,$$

where θ_j are the moving-average coefficients with $\theta_0 = 1$ and e_t are the white noise terms which will be assumed to be normally distributed throughout the study. In this study we specifically use the MA(1) model which has the form

$$X_t = e_t + \theta e_{t-1}, \quad t = 1, 2, ..., n,$$

where θ is the first order moving average coefficient.

The moving-average model coefficient, θ, is related to the autocovariance $\gamma(h)$ via the function,

$$\gamma(h) \quad = \quad E(X_t X_{t+h})$$

so

$$\gamma(h) = \begin{cases} \sum_{i=0}^{p-|h|} \theta_i \theta_{i+|h|} \sigma^2, & |h| \leq p \\ 0, & \text{otherwise.} \end{cases}$$

The autocovariance at lag zero is

$$\gamma(0) = (1 + \theta^2)\sigma^2,$$

and at lag one

$$\gamma(1) = \theta\sigma^2.$$

Note that the autocovariance at higher lag in MA(1) process is zero. Thus, from (2.3), the autocorrelation at lag one is

$$\rho(1) = \frac{\theta}{1+\theta^2}.$$

Note that the AR and MA coefficients determine the structures of the autocovariance and autocorrelation functions in time series, and in turn, the autocovariances and autocorrelations are essential in calculateing effective sample size, which will be seen in the next section.

3.2 Effective Sample Size in AR(1) and MA(1)

A rearrangement of (2.2) yields

$$n_e = n\frac{Var(\bar{Y}_{IID})}{Var(\bar{Y}_{CORR})}.$$

Thiebaux and Zwiers [9] show that the variance of the sample means from the correlated data, $Var(\bar{Y}_{CORR})$, which is needed for the calculation of effective sample size can be expressed in terms of the variance and autocorrelation function $\rho(h)$, where h is the lag between observations. The autocorrelations of all possible lags are used

in the calculation of the effective sample size for AR models, and only the autocorrelation at lag one is used for the MA model. The variance of the sample means from the correlated data is expressed as

$$\begin{aligned} Var(\bar{Y}_{CORR}) &= E[(\bar{Y}-\mu)^2] \\ &= E[(\sum_{i=1}^{n}(Y_i-\mu)/n)^2] \\ &= \sum_{i,j=1}^{n} E[((Y_i-\mu)(Y_j-\mu))]/n^2 \\ &= \sum_{i,j=1}^{n} \gamma(i-j)/n^2 \\ &= \sum_{h=-(n-1)}^{(n-1)} (n-|h|)\gamma(h)/n^2 \\ &= \sigma^2 \sum_{h=-(n-1)}^{(n-1)} (1-\frac{|h|}{n})\rho(h)/n. \end{aligned} \tag{3.1}$$

Here $\gamma(h)$ is the autocovariance between observations with lag h, n is the original sample size, $\rho(h)$ is the autocorrelation, and σ^2 is the innovation variance of the stationary time series process. The innovation variance is actually not needed, as you will see due to cancellation, so $\rho(h)$ is the last piece of the puzzle needed in order to calculate n_e.

In summary the effective sample size for the AR(1) process is

$$n_e = \frac{n}{\sum_{h=-(n-1)}^{n-1}(1-\frac{|h|}{n})\phi^{|h|}},$$

while for the MA(1) process it is

$$n_e = \frac{n}{1 + (1 - \frac{1}{n})\frac{2\theta}{1+\theta^2}}.$$

Here it is seen that the innovation variance cancels leaving n_e only as a function of the AR(1) and MA(1) coefficients.

We conclude that upon fitting the AR(1) or MA(1) models to time series, we are able to calculate n_e. This brings us back to the approximation of the probability distributions of the minimum and maximum order statistics in correlated data from AR(1) and MA(1) models. These can now be approximated as

$$g_n(X_{(n)}) \approx n_e[F(X_{(n)})]^{n_e-1} f(X_{(n)}) \tag{3.2}$$

for the maximum order statistic $X_{(n)}$ and

$$g_1(X_{(1)}) \approx n_e[1 - F(X_{(1)})]^{n_e-1} f(X_{(1)}) \tag{3.3}$$

for the minimum order statstic $X_{(1)}$. We use the approximate probability distributions, (3.2) and (3.3), for the maximum and minimum order statistics from AR(1) and MA(1) models in the simulation study in Chapter 4.

CHAPTER 4

SIMULATION

4.1 Simulation Scheme

In this section, we test the ideas previously discussed and simulate the distributions of the maximum and minimum order statistics $X_{(n)}$ and $X_{(1)}$ from AR(1) and MA(1) models. Since there is no exact distribution of the order statistic in correlated data, we will generate data sets of size 1000 and pull out the maximum and minimum order statistics. The number of datasets of size 1000 will be adjusted to $25, 50, 100, 500, 1000, 5000$ generating the respective number of data points. We will then find the 90th, 95th, and 99th percentile values of the distributions of the maximum and minimum order statistics. We will do this for AR and MA coefficient values of: $0, .2, .4, .6, .8$ assuming e_t in AR(1) and MA(1) processes has a normal distribution with mean zero and variance one.

We will calculate the effective sample size using (3.1). We will then use the inverse CDF method [3], as applied to the standard normal distribution, to find the $X_{(n)}$ value that defines the 90th, 95th, and 99th percentiles of the distribution according to effective sample size. For example the approximate cumulative distribution of $X_{(n)}$

is

$$G_n(X_{(n)}) \approx F(X_{(n)})^{ne}.$$

The $90th$ percentile is

$$X_{(n)} \approx F^{-1}(.9^{\frac{1}{ne}}),$$

where n_e is an effective sample size. F is the cumulative density function of the normal distribution. Upon finding this value, we will add in the case of the maximum order statistic, or subtract in the case of the minimum order statistic, the adjustment factor,

$$\frac{\rho(1)}{1-\rho(1)^2}. \tag{4.1}$$

This adjustment factor came from examination of the simulated data. The calculated quantile values were compared to the quantiles of the simulated data, and the difference between these two values was plotted as the AR and MA coefficeints increased. Upon inspection, the difference curve that resulted, we chose (4.1) because it closely modeled the difference between these quantile values. An adjustment function is needed because the standard normal distribution has a variance of one but, as the autocorrelation is increased, the inverse CDF method using the normal distribution

does not achieve the correct quantile value.

In summary, the calculation of the quantile for the maximum order statistic at the 90th percentile is

$$X_{(n)} \approx F^{-1}(.9^{\frac{1}{n_e}}) + \frac{\rho(1)}{1-\rho(1)^2} \tag{4.2}$$

while for the minimum order statistic it is

$$X_{(n)} \approx F^{-1}(.9^{\frac{1}{n_e}}) - \frac{\rho(1)}{1-\rho(1)^2}. \tag{4.3}$$

This adjustment factor will make the quantile values using n_e close to the ones of the simulated data.

4.2 Simulation Results

Tables 4.1 through 4.4 compare the quantiles that were calculated using effective sample size (ESS) with the quantiles from the simulated data (SIM). As you move down a column the percentile and the AR/MA coefficient increases. As you move across a row the sample size increases. It is important to state that when either the AR or MA coefficient is zero, we are observing a case where $n_e = n$. This means that if we did not take into account autocorrelation in the data, the quantile values would always be the same as (ESS) in these zero correlation tables.

Tables 4.1 and 4.2 show the simulation results for probability distributions of the

maximum order statistic $X_{(n)}$ from AR(1) and MA(1) models, respectively.

TABLE 4.1 Percentile of $X_{(n)}$ in AR(1) Process

n		25	50	100	500	1000	5000
ϕ	Percentile	ESS SIM	ESS SIM	ESS SIM	ESS SIM	ESS SIM	ESS SIM
0	0.9	2.63 2.65	2.86 2.87	3.06 3.00	3.53 3.55	3.69 3.68	4.09 4.08
	0.95	2.86 2.93	3.01 3.07	3.29 3.22	3.71 3.75	3.89 3.87	4.25 4.25
	0.99	3.35 3.33	3.54 3.47	3.71 3.57	4.11 4.17	4.26 4.24	4.61 4.69
0.2	0.9	2.72 2.72	2.96 2.83	3.16 3.09	3.64 3.58	3.81 3.72	4.22 4.15
	0.95	2.96 2.99	3.18 3.06	3.37 3.28	3.83 3.72	3.99 3.94	4.38 4.32
	0.99	3.45 3.44	3.64 3.58	3.93 3.74	4.22 4.20	4.41 4.28	4.74 4.87
0.4	0.9	2.81 2.85	3.06 3.12	3.22 3.39	3.77 3.86	4.02 4.03	4.37 4.50
	0.95	3.06 3.11	3.35 3.35	3.57 3.65	4.02 4.03	4.20 4.22	4.59 4.66
	0.99	3.57 3.70	3.77 3.82	4.02 4.06	4.37 4.49	4.49 4.63	4.90 5.12
0.6	0.9	3.11 3.20	3.35 3.54	3.58 3.77	4.10 4.36	4.29 4.64	4.71 5.08
	0.95	3.38 3.45	3.59 3.80	3.82 4.07	4.29 4.58	4.48 4.84	4.88 5.28
	0.99	3.93 3.89	4.1 4.31	4.29 4.52	4.31 5.14	4.88 5.37	5.25 5.64
0.8	0.9	4.03 3.99	4.34 4.53	4.59 4.84	5.09 5.74	5.34 6.09	5.76 6.73
	0.95	4.34 4.41	4.58 5.06	4.87 5.25	5.34 6.07	5.51 6.37	5.97 7.05
	0.99	4.93 5.27	5.10 5.94	5.38 5.89	3.54 6.76	5.76 7.01	6.33 7.79

TABLE 4.2 Percentile of $X_{(n)}$ in MA(1) Process

n		25	50	100	500	1000	5000
θ	Percentile	ESS SIM	ESS SIM	ESS SIM	ESS SIM	ESS SIM	ESS SIM
0	0.9	2.64 2.56	2.86 2.86	3.07 3.06	3.53 3.57	3.71 3.70	4.10 4.12
	0.95	2.87 2.88	3.08 3.12	3.28 3.29	3.71 3.74	3.88 3.90	4.26 4.29
	0.99	3.35 3.31	3.54 3.54	3.72 3.62	4.11 4.06	4.26 4.26	4.61 4.54
0.2	0.9	2.78 2.73	2.95 2.90	3.18 3.08	3.64 3.60	3.82 3.77	4.22 4.18
	0.95	3.01 2.95	3.18 3.10	3.39 3.30	3.83 3.77	4.00 3.96	4.39 4.33
	0.99	3.50 3.37	3.65 3.55	3.83 3.72	4.23 4.10	4.39 4.34	4.74 4.71
0.4	0.9	2.84 2.80	3.07 3.12	3.29 3.30	3.77 3.78	3.95 4.04	4.35 4.39
	0.95	3.09 3.08	3.31 3.33	3.51 3.54	3.96 3.98	4.14 4.23	4.52 4.58
	0.99	3.58 3.58	3.78 3.99	3.96 3.96	4.36 4.37	4.53 4.54	4.88 4.87
0.6	0.9	2.98 2.99	3.21 3.40	3.43 3.57	3.90 4.07	4.09 4.32	4.49 4.81
	0.95	3.23 3.28	3.44 3.63	3.65 3.87	4.10 4.26	4.28 4.52	4.67 4.94
	0.99	3.74 3.87	3.92 4.16	4.10 4.19	4.51 4.74	4.67 4.88	5.03 5.26
0.8	0.9	3.05 3.30	3.28 3.63	3.51 3.91	3.98 4.53	4.17 4.75	4.57 5.25
	0.95	3.30 3.56	3.52 3.93	3.73 4.22	4.18 4.77	4.35 4.94	4.74 5.48
	0.99	3.81 4.07	4.00 4.42	4.18 4.74	4.59 4.41	4.75 5.50	5.11 5.95

In both tables we see that the values given by the quantiles determined by effective sample size are much closer to the quantiles of the simulated data than if autocorrelation had not been considered. This can be seen when comparing the (ESS) values from Tables 4.1 and 4.2 where the AR(1) and MA(1) coefficients are zero to the (SIM)

values where these coefficeints are not zero. For example in Table 4.1 we see that the quantile value of the 90th percentile for a sample size of 100 is 3.06 with $\phi = 0$. This would be the quantile value always if correlation was not considered. However, if we increase to $\phi = 0.6$ we see that ESS gives us a quantile value of 3.58 and an SIM of 3.77. 3.58 is much better than 3.06 according to SIM.

As correlation increases, the differences between the quantiles become more erratic, especially for large sample sizes. For example when $\phi = 0.8$ and $n = 5000$ in Table 4.1 we see for the 90th percentile ESS gives 5.76 and SIM gives 6.73. Normally, as sample size increases, you would expect better results. However, the problem is that as sample size increases the distribution of the order statistic, in the case of the maximum order statistic, shifts to the right. A better adjustment factor could potentially remedy this problem.

We repeat the same procedure as above only for the minimum order statistic. Tables 4.3 and 4.4 demonstrate quantile values from the simulated data and the calculated quantiles using effective sample size for the minimum order statistic $X_{(1)}$ in the AR(1) and MA(1) processes. In this case, the same adjustment factor is subtracted out.

TABLE 4.3 Percentile of $X_{(1)}$ in AR(1) Process

n		25	50	100	500	1000	5000
ϕ	Percentile	ESS SIM	ESS SIM	ESS SIM	ESS SIM	ESS SIM	ESS SIM
0	0.9	-1.35 -1.33	-1.69 -1.68	-1.99 -2.01	-2.60 -2.60	-2.83 -2.85	-3.31 -3.31
	0.95	-1.21 -1.19	-1.57 -1.57	-1.89 -1.92	-2.51 -2.52	-2.74 -2.76	-3.24 -3.24
	0.99	-0.96 -0.90	-1.35 -1.38	-1.69 -1.72	-2.36 -2.39	-2.60 -2.64	-3.11 -3.12
0.2	0.9	-1.35 -1.33	-1.72 -1.71	-2.04 -2.02	-2.67 -2.61	-2.91 -2.91	-3.40 -3.39
	0.95	-1.20 -1.18	-1.59 -1.58	-1.92 -1.88	-2.58 -2.55	-2.82 -2.83	-3.33 -3.32
	0.99	-0.93 -0.94	-1.35 -1.38	-1.71 -1.65	-2.42 -2.41	-2.67 -2.71	-3.20 -3.18
0.4	0.9	-1.36 -1.31	-1.74 -1.76	-2.10 -2.10	-2.78 -2.79	-3.03 -3.07	-3.55 -3.61
	0.95	-1.19 -1.11	-1.59 -1.59	-1.97 -1.95	-2.68 -2.67	-2.94 -2.98	-3.47 -3.53
	0.99	-0.89 -0.81	-1.33 -1.30	-1.75 -1.73	-2.51 -2.51	-2.78 -2.86	-3.34 -3.39
0.6	0.9	-1.52 -1.32	-1.96 -1.87	-2.35 -2.29	-3.08 -3.19	-3.34 -3.48	-3.88 -4.11
	0.95	-1.33 -1.08	-1.81 -1.69	-2.21 -2.13	-2.97 -3.04	-3.24 -3.34	-3.80 -4.01
	0.99	-0.94 -0.65	-1.52 -1.32	-1.97 -1.91	-2.79 -2.82	-3.08 -3.19	-3.66 -3.88
0.8	0.9	-2.13 -1.14	-2.69 -1.87	-3.16 -2.63	-3.97 -3.95	-4.27 -4.45	-4.86 -5.34
	0.95	-1.88 -0.78	-2.49 -1.52	-2.99 -2.34	-3.84 -3.78	-4.16 -4.32	-4.77 -5.20
	0.99	-1.43 -0.04	-2.13 -0.90	-2.69 -1.97	-3.63 -3.43	-3.97 -4.07	-4.62 -5.01

TABLE 4.4 Percentile of $X_{(1)}$ in MA(1) Process

n		25	50	100	500	1000	5000
θ	Percentile	ESS SIM	ESS SIM	ESS SIM	ESS SIM	ESS SIM	ESS SIM
0	0.9	-1.35 -1.34	-1.69 -1.72	-2.00 -1.99	-2.60 -2.59	-2.83 -2.81	-3.31 -3.31
	0.95	-1.21 -1.16	-1.57 -1.59	-1.89 -1.89	-2.51 -2.51	-2.75 -2.74	-3.24 -3.25
	0.99	-0.96 -0.89	-1.35 -1.36	-1.69 -1.64	-2.36 -2.34	-2.60 -2.58	-3.11 -3.14
0.2	0.9	-1.37 -1.32	-1.74 -1.72	-2.06 -2.01	-2.69 -2.66	-2.93 -2.86	-3.42 -3.36
	0.95	-1.22 -1.13	-1.61 -1.58	-1.94 -1.89	-2.59 -2.55	-2.84 -2.79	-3.34 -3.28
	0.99	-0.95 -0.89	-1.37 -1.31	-1.74 -1.76	-2.43 -2.43	-2.69 -2.63	-3.22 -3.17
0.4	0.9	-1.45 -1.36	-1.83 -1.77	-2.15 -2.12	-2.80 -2.77	-2.98 -3.05	-3.54 -3.57
	0.95	-1.29 -1.20	-1.69 -1.66	-2.03 -1.99	-2.70 -2.67	-2.95 -2.96	-3.46 -3.45
	0.99	-1.01 -0.82	-1.45 -1.39	-1.82 -1.78	-2.54 -2.52	-2.80 -2.79	-3.34 -3.34
0.6	0.9	-1.58 -1.46	-1.94 -1.88	-2.27 -2.31	-2.93 -3.01	-3.17 -3.31	-3.68 -3.85
	0.95	-1.42 -1.31	-1.80 -1.72	-2.15 -2.15	-2.83 -2.90	-3.08 -3.21	-3.60 -3.76
	0.99	-1.13 -1.00	-1.56 -1.49	-1.93 -1.96	-2.67 -2.71	-2.93 -3.04	-3.47 -3.64
0.8	0.9	-1.62 -1.49	-2.01 -2.08	-2.34 -2.50	-3.00 -3.32	-3.25 -3.65	-3.76 -4.24
	0.95	-1.46 -1.27	-1.87 -1.91	-2.22 -2.31	-2.90 -3.21	-3.16 -3.52	-3.68 -4.14
	0.99	-1.17 -0.87	-1.62 -1.60	-2.00 -2.08	-2.73 -2.99	-3.00 -3.31	-3.55 -4.02

The same patterns appear in these tables as appear in the tables for the maximum order statisic. As the sample size becomes large at the same time the correlation increases, we see that a gap develops between effective sample size quantile(ESS) and simulated quantile(SIM). In this case, as sample size increases, the distribution of the minimum order statistic shifts to the left. For the most part it should be noticed that

the values given by ESS are far superior to the values you would get if you did not take into account the correlation. For example in Table 4.3, if we do not take into account correlation, the quantile value for the 90th percentile of a sample of size 100 would be -1.99. However if we do have correlation, say $\phi = 0.6$, ESS gives us a value of -2.35 while SIM gives us a value of -2.29. As you can see -2.35 is much closer to -2.29 than -1.99.

CHAPTER 5

DISCUSSION OF SIMULATION RESULTS

It is possible to use order statistics of correlated data as an application of effective sample size. This application makes it possible to approximate the distributions of the order statistics from correlated data. The ESS approximation is accurate in low correlation settings but not quite accurate for strong correlation settings. The fact that the fit of the results is quite accurate for low correlation suggests that effective sample size, coupled with the proposed autocorrelation function adjustment, produces an accurate approximation. We can also see that correlation cannot be ignored because it makes a large difference in the distributions of the minimum and maximum order statistics.

It would be of great interest to try to develope this study further by obtaining a method that works for negative autocorrelation. Theoretically, this should be quite possible. Also, it would be interesting to look at other order statistics besides the minimum and the maximum and see if the adjustment factor can be applied in a linear combination in order to make a better adjustment. It would also be of interest to improve on the adjustment factor and make it a matter of multiplication instead of addition. Other adjustment factors might work better than the one presented here.

REFERENCES

[1] Anderson, T.W. 1958. *An Introduction to Multivariate Statistical Analysis.* Chicago, Illinois: Wiley.

[2] Fuller, Wayne A. 1996. *Introduction to Statistical Time Series.* Chicago, Illinois: Wiley.

[3] Hogg, R., Craig, A., and McKean, J. 2005. *Introduction to Mathematical Statistics.* New Jersey: Prentice Hall.

[4] Ko, K., Lee, J., and Lund, R.B. 2008. "Confidence Intervals for Long Memory Regression." *Statistics and Probability Letters*, vol. 78, pp. 1894-1902.

[5] Laurmann, John A. and Gates, Lawrence. 1977. "Statistical Considerations in the Evaluation of Climactic Experiments with Atmospheric General Circulation Models." *Journal of Atmospheric Science*, vol. 34, pp. 1187-1199.

[6] Leith, C.E. 1973. "The Standard Error of Time-Average Estimates of Climate Means." *Journal of Applied Meteorology*, vol. 12, pp. 1066-1069.

[7] Lorenz, E.N. 1973. "On the Existence of Extended Range Predictability." *Journal of Applied Meteorology*, vol. 12, pp. 543-546.

[8] Madden, R.A. 1976. "Estimates of the Natural Variability of Time-averaged Sea-level Pressure." *Monthly Weather Review*, vol. 104, pp. 942-952.

[9] Thiebaux, H.J. and Zwiers, F.W. 1984. "The Interpretation and Estimation of Effective Sample Size." *Journal of Applied Meteorology*, vol. 23. pp. 800-811.

Appendix A

PROGRAMMING

Sample code to generate the data set with AR coefficient 0.9 and a sample size of 5000. The for loop generates 1000 $X_{(n)}$ values.

```
ordermax<-numeric()


for (i in 1:1000){
error.model=function(n){rnorm(n,mean=0, sd=1)}



data<-arima.sim(model=list(order = c(1, 0, 0),  ar=c(0.9)), n=5000,
n.start=2,
rand.gen=error.model )

ordermax[i]<-max(data)}

ordermax

quantile(ordermax, prob=c(.9,.95,.99))
```

A short program to calculate effective sample size with an AR coefficient of 0.9 and a sample size of 5000.

```
 part<-numeric()

for (i in 1:4999){

part[i]<-(1-(i/5000))*(((.9)^i))
}

ne<-5000/(1+2*(sum(part)))

ne
```

The program used for calculating the percentiles for the effective sample size method is very simple. To calculate the $90th$ percentile value of a dataset with effective sample size 225:

```
qnorm(.9^(1/225))
```

Here is the code that simulates data using an MA coefficient of 0.8 and draws the highest statistic from a sample of five thousand, repeating this one thousand times.

```
ordermax<-numeric()


for (i in 1:1000){
error.model=function(n){rnorm(n,mean=0, sd=1)}



data<-arima.sim(model=list(order = c(0, 0, 1),  ma=c(0.8)), n=5000,
n.start=2,
rand.gen=error.model )

ordermax[i]<-max(data)}

ordermax

quantile(ordermax, prob=c(.9,.95,.99))
```

Here is the code to calculate effective sample size.

```
k<-0

for (i in 1:10){

part<-(1-(1/500))

ne[i]<-500/(1+(2*(part)*(k/(1+k^2))))
k<-k+.1}

ne
```