Modified cumulative sum procedures for count data with application to early detection of morbidity in radio frequency-monitored animals

by Frederick Kweku Annan Holdbrook

A dissertation submitted in partial fulfillment , of the requirements for the degree of Doctor of Philosophy in Statistics

Montana State University

Abstract:

Radio frequency (RF) technology is used in electronic monitoring and data acquisition devices currently available to the commercial animal feeding industry for continuously monitoring the feeding and watering behaviors of feedlot animals.

There is therefore the need for statistical process control (SPC) procedures that can be used on-line, in conjunction with these electronic data collection systems, to achieve a cost effective system that can quickly detect animal morbidity.

In this dissertation, a modified cumulative sum (modified CUSUM) procedure is proposed based on modifying the traditional CUSUM scheme for count data. Additional CUSUM design parameters and additional out-of-control conditions are introduced to give the procedure several enhanced features, improve its detection capability, and reduce the average run length (ARL).

The method is evaluated using simulated Poisson data, and recommendations for the choice of the extra design parameters are discussed. An outline is provided to show how the modified procedure can be implemented for a generic cattle feedlot data, which can be acquired from a digital monitoring and data collection system based on radio frequency (RF) ear-tag technology.

Results demonstrate that the modified CUSUM scheme can indeed achieve higher sensitivity and performance than the traditional CUSUM scheme. The proposed modified CUSUM schemes can also be designed to be relatively robust to isolated outliers that may be present in the data. It is further demonstrated that an optimal design of this modified CUSUM scheme can indeed be very useful in the early detection of morbidity among group-fed animals.

MODIFIED CUMULATIVE SUM PROCEDURES FOR COUNT DATA WITH APPLICATION TO EARLY DETECTION OF MORBIDITY IN RADIO FREQUENCY-MONITORED ANIMALS

by

Frederick Kweku Annan Holdbrook

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

Doctor of Philosophy

in

Statistics

MONTANA STATE UNIVERSITY
Bozeman, Montana

February 2001

APPROVAL

of a dissertation submitted by

Frederick Kweku Annan Holdbrook

This dissertation has been read by each member of the dissertation committee and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the College of Graduate Studies.

William F. Quimby William F. Quimby 2/16/01
(Signature) Date

Approved for the Department of Statistics

John Lund John Lund 2/16/01
(Signature) Date

Approved for the College of Graduate Studies

Bruce McLeod Bruce R. McLeod 2-26-01
(Signature) Date

STATEMENT OF PERMISSION TO USE

In presenting this dissertation in partial fulfillment of the requirements for a doctoral degree at Montana State University, I agree that the Library shall make it available to borrowers under rules of the Library. I further agree that copying of this dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U. S. Copyright Law. Requests for extensive copying or reproduction of this dissertation should be referred to Bell & Howell Information and Learning, 300 North Zeeb Road, Ann Arbor, Michigan 48106, to whom I have granted "the exclusive right to reproduce and distribute my dissertation in and from microform along with the non-exclusive right to reproduce and distribute my abstract in any format in whole or in part."

Signature 

Date 2/16/04

I dedicate this manuscript to the loving memory of my dear mother Dinah. She had always encouraged and supported me wholeheartedly in my life decisions. Thank you Mom for having faith in my dreams.

Dinah Maame Gyamansa Akyempon
1929–1999

ACKNOWLEDGEMENTS

First of all, I thank my Lord Jesus Christ who has continually satisfied my deepest needs even when study fatigue and discouragement often set in.

I sincerely thank my advisor, Dr. William Quimby for his understanding, patience, guidance, assistance and suggestions during the preparation of this manuscript.

I am very grateful to Dr. John Borkowski who provided invaluable ideas, comments and suggestions in the fields of quality control and stochastic processes.

I especially thank my reading committee made up of professors W. Quimby, J. Borkowski, and Robert Boik as well as other members of my committee including professors James Robison-Cox, Warren Esty, and Lea Acord for their time and helpful comments and suggestions.

My thanks also go to James Jacklitch for helping me with LaTeX, and Lois Carbaugh for her immense help during my studies in Montana.

Most of all, I wish to express my gratitude to my entire family, brothers and sisters for their everlasting support and encouragement. I am most indebted to my brother Edward Delke Holdbrook without whose assistance none of these would have been possible. To my wonderful children Angela, Linda and Jesse Holdbrook, I say thank you very much for being understanding, loving, and cheerful all these times.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Radio frequency (RF) technology is used in electronic monitoring and data acquisition devices currently available to the commercial animal feeding industry for continuously monitoring the feeding and watering behaviors of feedlot animals.

There is therefore the need for statistical process control (SPC) procedures that can be used on-line, in conjunction with these electronic data collection systems, to achieve a cost effective system that can quickly detect animal morbidity.

In this dissertation, a modified cumulative sum (modified CUSUM) procedure is proposed based on modifying the traditional CUSUM scheme for count data. Additional CUSUM design parameters and additional out-of-control conditions are introduced to give the procedure several enhanced features, improve its detection capability, and reduce the average run length (ARL).

The method is evaluated using simulated Poisson data, and recommendations for the choice of the extra design parameters are discussed. An outline is provided to show how the modified procedure can be implemented for a generic cattle feedlot data, which can be acquired from a digital monitoring and data collection system based on radio frequency (RF) ear-tag technology.

Results demonstrate that the modified CUSUM scheme can indeed achieve higher sensitivity and performance than the traditional CUSUM scheme. The proposed modified CUSUM schemes can also be designed to be relatively robust to isolated outliers that may be present in the data. It is further demonstrated that an optimal design of this modified CUSUM scheme can indeed be very useful in the early detection of morbidity among group-fed animals.

CHAPTER 1

INTRODUCTION AND BACKGROUND INFORMATION

The rapid evolution of electronic monitoring and control systems has stimulated the search for more advanced statistical quality control (SQC) procedures that can be linked on-line to operate in conjunction with these automatic data acquisition and monitoring systems. There appears to be two main reasons for this rapid development. First are the tremendous advances in computer and information technology, which have made digital monitoring and data collection systems cheaper and more effective (in terms of their capability to continuously monitor a process and obtain measurements on several variables for every item produced). The second reason is the theoretical interest and the inherent challenge posed by the usually non-normal, autocorrelated nature of the data generated by these electronic data collection systems.

The incredible potential of automated systems to improve productivity and quality as well as their other related issues, have recently been discussed by Keats [60], among several others. Rather than a system that analyzes the sample data after the product is produced, many manufacturing and process industries are looking for a cost effective system that can continuously monitor the process during production and that identifies and reports trouble spots before bad products are made (Keats

[60]). Many industries, such as the automobile, chemical, semi-conductor, and food and beverage industries are already implementing sophisticated versions of such digital control systems (Palm, Rodriguez, Spiring, and Wheeler [85]), and many more are set to follow, including the commercial cattle-feeding industry. SPC related problems in the cattle-feeding industry discussed in this dissertation can be summarized as follows: we have discrete data from a cattle feedlot that are potentially overdispersed, and could be autocorrelated because they were obtained from a high-speed data acquisition system. A simple easily-implemented statistical process control procedure is desired, that can be used to predict morbidity in cows by monitoring the mean of this process.

In this dissertation, a modified cumulative sum (modified CUSUM) control chart technique is proposed for use in the cattle-feeding industry. The method involves modifying the traditional CUSUM scheme by introducing additional parameters and additional out-of-control conditions to give the procedure several enhanced features. Determination of the parameter settings and the average run length are discussed, and simulated Poisson data are used to evaluate the procedure. Several examples are provided to illustrate the method. Finally, an outline is provided for implementing the modified procedure for a generic cattle feedlot data.

In the remainder of this chapter, I will first provide and discuss several general definitions and terminology from the field of statistical quality control that are used

frequently in this dissertation. Next will be a brief discussion on background information and overview of the relevant statistical process control literature. This is followed by a review of general control charting procedures, and a discussion on the main motivation behind this research (which also outlines the main problem being considered in this dissertation).

Definitions and Terminology

This section outlines and discusses several quality control definitions and terminology conventions that are used throughout this dissertation. The definitions have been adopted mainly from texts by Besterfield [7]; Burr [19] and Montgomery [77].

1. A product refers to all categories of manufactured goods such as computers and clothing, processed goods including meat products and other food items, and services such as banking and health care. Every product is considered to be the result of a process.

2. Quality is defined formally as the totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs (*International Standards Organizations*, ISO 8402).

   Implicit in the above definition of quality are the following meanings discussed in Montgomery [77, page 4]:

   (a) quality means fitness for use.

(b) quality is inversely proportional to variability. Thus, a decrease in the variability of an on-aim process will, by implication, increase the quality of the output (or product) from that process.

3. Quality improvement is the reduction of variability in products and processes.

4. A quality product is a good or excellent product that fulfills customers' expectation based on the intended use and selling price.

5. Control is the process of regulating or directing an activity to verify its conformance to a standard and to take corrective action if required (Besterfield [7, page 1]).

6. Quality control is the regulatory process for those activities which measure a product's performance, compare the performance with established standards, and pursue corrective action regardless of where those activities occur (Besterfield [7]).

7. Statistical quality control (SQC) is the branch of quality control that uses statistical methods to collect data in a sound, unbiased manner and to analyze and interpret the results appropriately so as to obtain satisfactory, dependable and economic quality (Besterfield [7]).

8. Statistical process control (SPC) is the part of SQC consisting of a collection

of powerful problem-solving tools useful in achieving process stability and improving capability through the reduction of variability.

9. Data collected from any service or production process have one thing in common: they vary. They can vary from time to time, customer to customer, operator to operator, piece to piece, and sample to sample. Burr [19] puts it more succinctly: "whenever we have variation we have a statistical problem, whether we like it or not, whether we know it or not." For data collected by on-line data acquisition systems, the sampling intervals tend to be very short, and large amount of data are generated within that short sample interval. Consequently, data obtained from such automatic data collection systems often tends to be correlated. The two main sources of variation discussed in the SPC environment are defined below.

   (a) Chance causes (or common causes) of variation refer to the natural, purely random variability or background noise that are an inherent part of the process. This represents the "stable" pattern of variation in the process.

   (b) Assignable causes (or special causes) of variation are the systematic non-random patterns of variation that are not part of the chance cause pattern, but which may occasionally be present in the output of a process. Variations of this type are generally larger than the background noise and can result in an unacceptable level of process performance. Removing special

causes usually improves the process (and the quality of the product).

In the above definitions, it should be mentioned that the terms chance and assignable causes were introduced by Shewhart, whereas W. Edwards Deming [25], who pioneered the philosophy of total and continuous quality improvements, suggested the corresponding alternative terms common and special causes.

10. The process is said to be in a state of statistical control when the assignable causes have been eliminated and the process is operating with only chance causes of variation present. At this stage, Besterfield [7, page 73] says that "no higher degree of uniformity can be attained with the existing process, ... except through a change in the basic process itself."

11. An out-of-control process is a process that is operating in the presence of special or assignable causes of variation.

12. The control chart is a very powerful SPC tool that can be used on-line to monitor and quickly detect the presence of assignable causes of process shifts so that investigative and corrective actions can be taken.

13. Shewhart charts are the conventional (or classical) control charts that follow the general construction principles developed by Dr. Walter Shewhart [101]. Examples of these are the X-chart (individuals chart), R-chart, $\overline{X}$-chart and s-chart for variables data, and p- chart, np-chart, c-chart and u-charts for attribute

data.

14. A CUSUM control chart (or CUSUM) is an abbreviation for the cumulative sums control chart.

15. Traditional (or standard or regular) CUSUM refers to the classical CUSUM design without any of the recent modifications or enhancements.

16. EWMA control chart (or EWMA) is an abbreviation for the exponentially weighted moving average control chart.

17. ARL is the abbreviation for the average run length (which is the mean of the run length distribution). In other words, the ARL is the average number of observations (or samples) before the process signals out-of-control. An essential distinction is made between the in-control ARL and the out-of-control ARL. The in-control ARL is the expected time until the process yields a false out-of-control signal, while it is in control. A signal occurring while the process is deemed to be in-control, is also referred to as a false alarm. The out-of-control ARL is the expected time until the process signals an alarm once it shifts to an out-of-control state. Ideally, it is desired that the in-control ARL will be high (meaning low false alarm rate) while the out-of-control ARL will be low.

18. The optimality property of a control scheme can be stated as follows: among all procedures with the same in-control ARL, the optimal procedure is the one

that has the smallest (or quickest) expected time until it signals a change, once the process shifts to the out-of-control state (Moustakides [81]).

19. An item is said to be nonconforming if it fails to conform to product specifications in some respect. Otherwise, it is said to be conforming. Specific instances of failure in a particular item are called nonconformances (or nonconformities).

20. Discrete data refer to observations made on a discrete variable. For example, the result of counting nonconformities or nonconforming pieces in a sample so that only whole numbers or integers may occur is discrete. The term "attribute" data originally meant counts of the number of conforming and nonconforming items in a sample (Hawkins and Olwell [47, page 105]). In recent usage however, attribute data is typically used when the emphasis is on the classification of items into conforming and nonconforming. This is an attempt to distinguish attribute data from "count data", a term often used when the emphasis is on the frequency of occurrence of nonconformances or other discrete events. Most authors think that this technical distinction is not essential, since both count and attribute data are discrete. Subsequently, the term count data is used for all discrete data because it has wider appeal (Lucas [68]), and this convention will be adopted here in this dissertation.

21. Variables data refers to the numerical value of a quality characteristic that is measured on a continuous scale.

Background and Overview of Relevant SPC Literature

The purpose of this section is to provide background information and discuss relevant statistical quality control literature pertaining to the aims of this dissertation.

Quality control is a very broad field that deals with the regulatory process for those activities which measure a product's performance, compare the performance with established standards, and pursue corrective action regardless of where those activities occur. The formal beginning of statistical quality control (SQC) can be traced back to 1924, when Dr. Walter A. Shewhart of Bell Telephone Laboratories developed the first control chart. Later in 1931, Shewhart's [101] classic book *Economic Control of Quality of Manufactured Product* was published. It contained a complete exposition of the theory, practical application, and economics of the control charts. Burr [19, page 2] remarks in his 1979 book *Elementary Statistical Quality Control* that "seldom has a whole field of knowledge been so well explored and its application so well pointed in the first publication."

Statistical quality control charts made a great impact on manufacturing processes during World War II. The American Society for Quality Control which was subsequently formed in 1946, has been instrumental in promoting the use of quality control for all types of production and service. Today, control charts have become a popular and powerful statistical process control (SPC) tool, and they have found useful application in many fields including the manufacturing industry, management, government, crime control, health care and environmental surveillance. Volumes have

been written about the construction, applications, and properties of statistical control procedures. Texts such as the 1988 book by Grant and Leavenworth [39] titled *Statistical Quality Control* and Besterfield's [7] 1979 book *Quality Control* together with Burr [19], provide insightful and straight forward introduction to the Shewhart quality control charts. *Introduction To Statistical Quality Control* by Montgomery [77] is a more accessible textbook, and it provides excellent discussions on Shewhart charts, as well as the CUSUM and EWMA control schemes with interesting applications.

While Shewhart charts are widely used, easy to construct and interpret, and very effective for detecting large shifts in the parameter of the process, they are not so useful for detecting smaller shifts. In addition, there are violations of the underlying assumptions and the occurrence of trends and unusual patterns that needed to be studied. Goldsmith and Whitfield [38] had this to say about the Shewhart charts in 1961:

> The Shewhart quality control chart with fixed control lines suffers from the disadvantage that the observations are viewed independently, no account being taken of runs of observations all higher or all lower than the long-term mean of the process. This results in a relative insensitivity to moderate changes in the short-term mean value.

In response to these deficiencies, several adaptations and modifications of the Shewhart charts have been considered. These include schemes devised with both warning lines and action limits based on zone rules. Many of these modified Shewhart charts are discussed in Hill [49], Page [87, 88], and Mandel [75] who examined control charts that exhibit certain types of trends. The Western Electric [119] *Statistical Quality*

Control Handbook of 1956 contains several of these supplementary runs tests. The use of these supplementary runs tests (or *sensitizing rules* as they are often called) were advocated in an effort to improve the sensitivity of the Shewhart charts in detecting shifts of relatively smaller magnitude. Because the implementation of Shewhart control charts with these supplementary runs rules is often cumbersome and difficult to maintain, the search for other improvements and alternatives continued.

E. S. Page [86], in 1954, introduced the cumulative sum (CUSUM) control chart, and Roberts [95] proposed the exponentially weighted moving average (EWMA) control scheme in 1959, both as alternatives to the Shewhart charts. Duncan [26] also proposed models in 1956 that deal with the economic aspects in the design of control charts.

In what follows, a greater emphasis will be placed on the CUSUM procedure because it will be used to address the aims of this dissertation. Early papers that deal with the mathematical foundations of CUSUMs include Page [86], Ewan and Kemp [27], Kemp [62] and a series of three illuminating articles by Johnson and Leone [55, 56, 57] who also developed optimal CUSUM schemes for several distributions in the exponential family. Of particular interest is the realization that the CUSUM is based on the likelihood ratio strategy, and it can be considered as a sequence of Wald sequential probability ratio tests (SPRTs).

Two forms of the CUSUMs are generally discussed in the literature: tabular (or decision interval) CUSUM and the graphical V-Mask. The tabular form of the

CUSUM is most preferred because it has decision intervals which serve as control limits. The original CUSUM scheme proposed by Page plots the cumulative score

$$S_n = \sum_{i=1}^{n} x_i$$

against n, where $S_0 = 0$, and x_i is a score assigned to the ith observation. Based on Page's procedure, a one-sided control scheme would take action after the nth observation if

$$S_n - \min_{0 \le i < n} S_i \ge h,$$

where h is a positive constant called the decision interval (Goldsmith and Whitfield [38]). This initial procedure, however, looked more complicated when applied to the two-sided control case. G. A. Barnard [4], in a very stimulating article published in 1959, proposed a simpler method by shifting the origin of measurements to a target value μ, and plotting the cumulative sums of deviations

$$S_n = \sum_{i=1}^{n} (x_i - \mu).$$

In this way the CUSUM can be considered as a random walk model. The symmetric V-shaped mask was advocated for a two-sided control scheme based on this cumulative deviation chart. In 1960, Ewan and Kemp [27] also conducted a detailed study of the CUSUM procedure, and they showed that instead of plotting the cumulative sum of deviations, it is better to plot

$$S_n = \sum_{i=1}^{n} (x_i - k)$$

where k is a reference value. They also showed that for a k value near $(\mu_0 + \mu_1)/2$, the ARL at the acceptable mean level μ_0 is near its maximum for a given ARL at an unacceptable process level μ_1. In a CUSUM design therefore, the values of k and h need to be determined to achieve desirable ARL properties.

An extensive body of literature now exists on the design of CUSUM control charts. Bissell [9], Lucas [69], and Woodall [124] (on general designs of CUSUM techniques); Hawkins and Olwell [46] (on CUSUMs for location and shape parameters based on the inverse Gaussian distribution); Bourke [14] and Lucas [70] (on CUSUMs for low count-level processes); and Taylor [110] (on economic design of CUSUM charts) are very informative. Van Dobben de Bruyn's [113] 1968 monograph titled *Cumulative Sum Tests: Theory and Practice*, is one of the earliest texts published about CUSUMs. A more recent text is the *Cumulative Sum Charts and Charting for Quality Improvement* by Hawkins and Olwell [47] which is an extremely useful, compact book devoted entirely to CUSUMs. Hawkins and Olwell discuss virtually all the essential aspects and techniques in the design and application of CUSUMs, and they also provide an extensive bibliographical index. In Chapter 6, they provide concise summary of the theoretical foundations of the CUSUM, and the derivation of the optimal CUSUM designs for several distributions belonging to the exponential family (several of them originally discussed by Johnson and Leone [55, 56, 57]). Hawkins and Olwell [47] also provide interesting discussions on the various methods currently available for computing the ARLs of the CUSUM. What is remarkable is that the CUSUMs for count

distributions are generally defined in exactly the same way as CUSUMs for continuous variables, and that the CUSUM can be easily designed for other distributional parameters apart from the mean.

Both CUSUM and EWMA schemes are superior to the Shewhart charts in many ways. Both make use of past data, and are able to react quickly to small persistent shifts in the parameter of the process. Although the performance of the Shewhart control scheme with supplementary runs rules improves greatly, it does not equal the performance level of the CUSUM (Champ and Woodall [20]). Page [88] showed that many of the amended Shewhart rules are in fact equivalent to a restricted type of the CUSUM scheme. In particular, the classical Shewhart chart with fixed control limits is known to be equivalent to the a CUSUM scheme where the parameters $h = 0$, and k is equal to the control limit (see for example, Lucas [74]). It has also been shown by Bissell [10] that the CUSUM performs much better than the Shewhart chart when the process has a linear trend rather than an abrupt shift. Further discussions on control charts under a linear trend can be found in Gan [32], among others.

Optimality properties of the CUSUM design have also been examined by several authors. Lorden [65] studied the asymptotic optimality of CUSUM procedures for detecting a step change in distribution. Moustakides [81] used Lorden's criteria and showed that among all tests with the same false alarm rate, the CUSUM has the smallest reaction time for detecting the shift from an in-control distribution to a single specific out-of-control distribution. Ritov [94] (using a decision theory approach),

Yashchin [133], Gan [31] and Buckley [18] have also studied optimality properties of the CUSUM. Optimal CUSUM designs have been discussed for binomial counts (Gan [33]), and Gan [34] has discussed optimal design of exponential CUSUM control charts. The EWMA control chart has also been studied extensively and its optimality properties and ARL have been evaluated (see for example, Crowder [22, 23]; Lucas and Saccucci [73]; Srivastava and Wu [107]).

Concerning CUSUM and EWMA, most of the literature I have seen so far seems to suggest that, while they may be nearly close in performance and efficiency, the CUSUM is more widely used and preferable. Even when forecast errors from correlated data are being monitored, properly designed CUSUMs can outperform the EWMAs (Vander Wiel [114]). Most implementation of statistical control charts are concerned with the monitoring of the mean of a process data, often under Gaussian assumptions. While both the CUSUM and EWMA schemes are efficient for detecting small shifts in the mean, they perform poorly against large shifts which, on the contrary, can be promptly detected by the Shewhart individuals charts.

To address this and other inadequacies, several adaptations, extensions and modifications of the CUSUM and EWMA schemes have been proposed. This includes the combined Shewhart-CUSUM schemes for the process mean (Lucas [67]; Yashchin [129]) which utilizes the nice features of both procedures to rapidly detect small as well as large shifts in the process mean. CUSUM procedures for variability and scale parameters (Hawkins [41]; Srivastava [106]) and combined Shewhart-CUSUM schemes

for the process variance (Gan [30]) have also been discussed. In fact as far back as 1961, Goldsmith and Whitfield [38] had suggested that: "it may sometimes be better to plot a Shewhart chart and a cumulative deviation chart simultaneously to obtain a picture of the process variation."

Hawkins [42] described another interesting extension of the CUSUM, called the self-starting CUSUM scheme, which seeks to avoid the problem of parameter estimation by standardizing individual, successive observations using the mean and standard deviation of all observations accumulated to date. Implementing this scheme for a count data CUSUM can, however, be complicated, and the results can be greatly impacted by the presence of outliers in the data. Adaptive and Bayesian procedures have also been discussed for the CUSUM (e.g. Healy [48]; Joseph and Bowen [59]) and for the EWMA (e.g. Hubele and Chang [50]). CUSUM schemes based on variable sample size have also been studied (e.g. Hughes, Reynolds and Arnold [51]). Other authors like Bissell [12], and Yashchin [130] have studied weighted CUSUMs, which can be used to monitor process data when they are paired with measurements on another variable. Hawkins and Olwell [47, chapter 3] provides a detailed discussion of these procedures.

Two exceptionally useful enhancements to the CUSUM control schemes are the Fast Initial Response (FIR) CUSUMs (Lucas and Crosier [71]), where the CUSUM is set to an initial positive head start value to give a faster signal if the process is out of control, and the robust CUSUMs of Lucas and Crosier [72] and Hawkins [45]. Lucas

and Crosier [72] use a method based on a two-in-a-row outlier rule, while Hawkins [45] uses a windsorization procedure whereby an observation exceeding a specified threshold value would be replaced by that threshold. This "windsorized" value would then be used in updating the CUSUM. Lucas and Saccucci [73] also discussed EWMA control schemes with the fast initial response feature, combined Shewhart-EWMA and robust EWMA control charts.

The performance of a control chart is usually measured by the average run length (ARL) which, for the CUSUM and EWMA, is generally very difficult to calculate except through the use of numerical and analytical approximations (e.g. Page [86, 88]; Ewan and Kemp [27]; Goel and Wu [37], and Bissel [9]). In 1972, Brook and Evans [17] developed a Markov chain model for analyzing CUSUMs, which allowed simpler calculations of the run length distribution and its moments. Their method has since been extended by many, including Woodall [122], Lucas [67], Lucas and Crosier [71, 72], and Lucas and Saccucci [73] who used the Markov chain technique to compute the ARLs of several modified CUSUM schemes. The Markov chain approach will be discussed further in Chapter 3.

Although most of the earlier work dealing with CUSUMs involved continuous variable data, there seem to be a growing number of recent applications involving CUSUMs for count data. A comprehensive bibliography and review of control charts based on attributes can be found in Woodall [125]. Brook and Evans [17] were the first to consider CUSUMs based on the Poisson distribution, and they demonstrated how

the ARL and other moments of the run length distribution can be obtained via the Markov chain approach. Lucas [68] provided a more detailed discussion of count data CUSUMs, with particular emphasis on Poisson CUSUMs and time-between-events CUSUMs.

One very important fact about the design of CUSUMs is that a CUSUM scheme requires the precise statistical distribution for the data to be specified together with the exact parameter values of the model. This can be very problematic since in practice the exact model is never known, and parameter values must be estimated. Munford [82], in 1980, developed the cumulative score (CUSCORE) technique which discretizes the observations by assigning a score of -1, +1, or 0 to the sample values according to whether they are "extreme negative", "extreme positive" or otherwise. Even though his procedure is attractive in the sense that it avoids the distributional problems with the data and provides simple and compact expressions for the calculation of the ARL, it is not very efficient for detecting larger deviations in the process mean. Munford's procedure was later refined by Ncube and Woodall [84] who used a combined Shewhart-CUSCORE scheme to achieve better reaction to both large and small shifts. More recently, Radaelli [92] extended Munford's procedure to the CUSUM surveillance of rare health events, and Radaelli [93] considered an extension to the Poisson and negative binomial dynamics. Related procedures for grouped-data, based on gauging, have been discussed (e.g. Xiao [128]; Steiner, Geyer and

Wesolowsky [109]; Steiner [108]) where an observable continuous variable is measured, but the data values are simply classified into one of several classes or intervals of values of the variable.

The following discussion concerns the treatment of autocorrelated data in SPC, and it is very important because it has direct bearing on the justification of assumptions that are made in the last section of chapter 2 of this dissertation. Conventional control charts were initially introduced as monitoring tools for detecting the presence of out-of-control situations, which are typically caused by systematic nonrandom patterns of variation in the process. Identifying and eliminating (or drastically reducing) those causes of variation is very essential. The control chart relies on the assumption that if the process is operating in the state of statistical control, then the observations from the process can be regarded as independent and identically distributed random variables. Within the environment of SQC, the assessment of the state of a process is usually considered in terms of hypothesis testing, where the null hypothesis is often stated as

H_0 : The process is in the state of statistical control

and the alternative hypothesis is

H_1 : The process is out of control.

The emphasis on hypothesis testing when the data are serially correlated and SPC is used to monitor the residuals from a time series model, has generated some vigorous debate among statisticians. The presence of autocorrelation makes the assumption of

independence inappropriate, and can invalidate the Markov property of the control schemes (Yashchin [132]) and lead to poor control chart performance. Goldsmith and Whitfield [38], Johnson and Bagshaw [54], and Bagshaw and Johnson [5] were among the first to carry out extensive study about the effects of autocorrelation on the CUSUM and other control schemes. They showed that positive autocorrelation can reduce the in-control ARL of the process while negative correlation leads to higher ARLs. However for larger shifts or deviations in the current mean, the serial correlation has little effect on the ARL.

Two general approaches to dealing with inherently autocorrelated process data have been discussed in the literature (e.g. Adams, Woodall and Superville [1]; Vander Wiel [114]). One approach calls for widening the control limits (usually for positively correlated observations) to achieve an acceptable rate of false alarms. This method, while capable of reducing the number of false alarms, can also render the control chart useless for the purpose of signaling unusual behavior that could result in process improvement. For negatively correlated observations the control limits are usually narrowed to achieve similar effect. The second approach involves modeling the autocorrelation and it appears to be more appealing to researchers, partly because of the academic challenge and interest posed by the nature of the data. Most of these procedures still emphasize hypothesis testing. Barnard [4] first touched on this issue in his 1959 classic article where he suggested that:

> Now that the theory of stochastic processes has grown into a well-rounded

> body of theory, it would seem appropriate to consider industrial processes whose natural, intrinsic variability is best described as a run of dependent random variables, that is, as a stochastic process.

Barnard's proposal that the object of control charts should be to model and estimate the process mean drew both widespread support and concern from many statisticians at the time.

Recently these views have resurfaced in an article by Alwan and Roberts [2] who believe that "if a process can be modeled, then the traditional objectives of quality control–or surveillance–can be better served." Expressing concern that the concentration on hypotheses testing runs the danger of narrowing the perspectives of SPC procedures, Alwan and Roberts argued that the dichotomy of a "state of statistical control" versus "out of control" is too sharp, and they wrote that " a state of statistical control is often hard to attain; indeed, in many applications it appears that this state is never achieved, except possibly as a crude approximation." Also, in a four-panel discussion on the subject of "Process Control and Statistical Inference" (Crowder, Hawkins, Reynolds, and Yashchin [24]), three of the discussants affirmed the dominant view that any problem involving detection of changes in a process is inherently related to hypothesis testing, while Crowder thinks there should be more emphasis on estimation and engineering control, arguing that the emphasis on hypothesis represents a major deficiency in most control charts with autocorrelation.

Indeed, several authors, including Alwan and Roberts [2], Montgomery and Friedman [78], Montgomery and Mastrangelo [79], Box and Kramer [15], Harris and Ross

[40], Wardell, Moskowitz, and Plante [116], and have also recently suggested that autocorrelation should be modeled by fitting an appropriate time series effects, and then applying standard control charting procedures to the residuals (or the one-step forecast errors). However, most of these papers still emphasize hypothesis testing. The Box-Jenkins family of time series models are often employed in this endeavor. Montgomery and Mastrangelo and Alwan and Roberts have also recommended that the EWMA be used as an approximation to the underlying time series model. Others authors such as Fellner [29] and Lucas [74] have expressed great concern about this modeling approach. During the discussion of Wardell et al. [116], Lucas [74] argued that "it is essential for the data to be correlated (or for there to be a restriction on sample size) for a CUSUM or EWMA control scheme to work effectively", and he went further to suggest variance-component modeling as an alternative approach for dealing with correlated process data.

The above discussions also relate directly to data obtained from on-line electronic control systems of which the generic feedlot data, discussed later in the next section, is an obvious example. Runger and Willemain [97] and Keats [60] have discussed the prevalent trends in many industries to employ digital control and on-line data collection systems and their impact on SPC. Two direct impacts on SPC are mentioned in the literature. First, the measurement interval is so small, often leading to serial correlation among the observations. Second is the tremendous amount of data generated by these high-speed data collection systems for analysis. It is strongly desired to have

a SPC scheme to work in conjunction with these on-line systems during production rather than analyzing the sample data after production.

The lack of consensus in these discussions is evident, suggesting that much less guidance is available for choosing and designing monitoring schemes for potentially autocorrelated data.

The final extensions and adaptations I would like to mention regarding statistical control procedures is in the area of multivariate quality control. This presents far more challenging problems of detection and diagnosis of persistent shifts. Excellent discussion on this subject can be found in Hawkins and Olwell [47, chapter 8] who remark in page 190 of their book that "multivariate control may be much more sensitive to shifts than is the collection of univariate control", and that it "can be more specific in diagnosing causes than is a collection of univariate charts."

General Control Charting Procedures

In this section, the three general types of control charts commonly discussed in the quality control literature will be reviewed. These are the Shewhart, EWMA and the CUSUM control charts. The same general design principles for these schemes can be applicable for both attribute and variable data types.

1. **Shewhart Charts:**

Suppose some quality characteristic θ (such as the process mean) is to be monitored, and let $\widehat{\theta}$ be an estimate of θ based on a random sample of n units. Typically this will be a sample drawn from a process that is operating in statistical control, where the output or the observations are assumed to be independent identically distributed random variables. The values of $\widehat{\theta}$ for each successive sample are then plotted on the control chart. Let $\mu_{\widehat{\theta}}$ and $\sigma_{\widehat{\theta}}$ be the mean and the standard deviation of the sampling distribution of $\widehat{\theta}$. Then the center line (CL), upper control limit (UCL) and lower control limit (LCL) of the Shewhart control chart are given by the general model

$$\begin{aligned} \text{UCL} &= \mu_{\widehat{\theta}} + k_1\sigma_{\widehat{\theta}} \\ \text{CL} &= \mu_{\widehat{\theta}} \\ \text{LCL} &= \mu_{\widehat{\theta}} - k_1\sigma_{\widehat{\theta}} \end{aligned}$$

where k_1 is a constant representing the number of standard deviations a particular value of $\widehat{\theta}$ is allowed to vary from $\mu_{\widehat{\theta}}$ without triggering an alarm (or an out-of-control process). The value of k_1 is based on the distribution of $\widehat{\theta}$; but in practice, it is customary to choose $k_1 = 3$ (that is, "3-sigma" limits).

Oftentimes, warning limits are also constructed between the center line and the control limits, according to the formulas

$$\begin{aligned} \text{UWL} &= \mu_{\widehat{\theta}} + k_2\sigma_{\widehat{\theta}} \\ \text{LWL} &= \mu_{\widehat{\theta}} - k_2\sigma_{\widehat{\theta}} \end{aligned}$$

where UWL and LWL are the upper and lower warning limits, and where k_2 ($k_2 < k_1$) is the number of standard deviations a particular value of $\widehat{\theta}$ is allowed to vary from $\mu_{\widehat{\theta}}$ without triggering a warning alarm. In practice, it is common to choose $k_2 = 2$.

2. **EWMA Charts:**

The EWMA control chart is based on the control statistic

$$Z_i = \lambda X_i + (1-\lambda)Z_{i-1}, \qquad 0 < \lambda \leq 1, \tag{1.1}$$

where λ is a specified constant weighting factor, and X_i are a sequence of quality measurements assumed to be independent and identically distributed random variables with mean μ and variance σ. Here again, the X_i can be individual observations or some empirical estimate of the process parameter. When the process is in control, the target value is $\mu = \mu_0$. The starting value Z_0 is usually taken to be $Z_0 = \mu_0$, otherwise it is computed as the average of the observations. A graphical display of the chart is obtained by plotting Z_i against the time order (or the sample number) i.

After successive substitutions, the recursion in (1.1) can be written out as

$$\begin{aligned} Z_i &= \lambda X_i + \lambda(1-\lambda)X_{i-1} + \lambda(1-\lambda)^2 X_{i-2} + \cdots + \lambda(1-\lambda)^{i-1} + (1-\lambda)^i Z_0 \\ &= \lambda \sum_{j=0}^{i-1} (1-\lambda)^j X_{i-j} + (1-\lambda)^i Z_0, \end{aligned}$$

which is an exponentially weighted average of all past observations, where the weights $\lambda \sum_{j=0}^{i-1}(1-\lambda)^j$ and $(1-\lambda)^i$ sum up to unity. It can also be shown that

the variance of the control statistic Z_i is

$$\begin{aligned} \sigma^2_{z_i} &= \text{Var}\left(\lambda\sum_{j=0}^{i-1}(1-\lambda)^j X_{i-j} + (1-\lambda)^i Z_0\right) \\ &= \frac{\lambda}{2-\lambda}[1-(1-\lambda)^{2i}]\sigma^2. \end{aligned} \tag{1.2}$$

The control limits for sample i are obtained by the formulas

$$\begin{aligned} \text{UCL}_i &= \mu_0 + L\sigma\sqrt{\frac{\lambda}{2-\lambda}[1-(1-\lambda)^{2i}]} \\ \text{CL}_i &= \mu_0 \\ \text{UCL}_i &= \mu_0 - L\sigma\sqrt{\frac{\lambda}{2-\lambda}[1-(1-\lambda)^{2i}]} \end{aligned}$$

where L is the multiple of standard deviations a particular value of Z_i is allowed to vary from μ_0 without triggering an alarm. The EWMA control chart triggers an alarm when $Z_i > \text{UCL}_i$ or $Z_i < \text{LCL}_i$. Appropriate values of the parameters λ and L have to be determined in the design. As $i \to \infty$, the variance in (1.2) $\sigma^2_{z_i} \to \lambda/(2-\lambda)$. Thus, the control limits can also be based on the asymptotic variance, using the formulas

$$\begin{aligned} \text{UCL} &= \mu_0 + L\sigma\sqrt{\frac{\lambda}{2-\lambda}} \\ \text{UCL} &= \mu_0 - L\sigma\sqrt{\frac{\lambda}{2-\lambda}}. \end{aligned}$$

This represents the situation where the process has been running continuously for a long time and has reached steady-state values.

The EWMA statistic in (1.1) can also be written in the form

$$Z_i = Z_{i-1} + \lambda(X_i - Z_{i-1}),$$

in which case the EWMA can be considered as a one-step ahead forecast for the process (Lucas and Saccucci [73]). It has recently been pointed out by Woodall [126] that because the EWMA control chart limits are determined based on the assumption of independence of the data over time, the EWMA chart is not more robust than other control charts, and that "the fact that the EWMA chart is based on a statistic which could be used (in another way) with autocorrelated data is of no help in this regard."

3. **CUSUM Charts:**

CUSUMs also work by directly accumulating information across successive observations to gain greater sensitivity towards small persistent shifts in the process parameter. The Shewhart and the EWMA control charts defined earlier are inherently two-sided (for monitoring both increases and decreases in the parameter of the process). The tabular CUSUM, on the other hand, is inherently a one-sided control procedure. Maintaining a two-sided control will therefore require two separate, one-sided tabular CUSUM schemes. We will describe in this section the most common CUSUM scheme, which is the normal CUSUM for monitoring the mean of a process (based on Gaussian assumptions). Within this framework, an upper (one-sided) tabular CUSUM scheme S_i^+ is used for detecting an increase in the process mean from an *acceptable level* or target μ_0 to an *unacceptable level* μ_1 ($\mu_1 > \mu_0$), where μ_1 is the shift for which maximum sensitivity is required. If the purpose is to detect a decrease in the mean

($\mu_1 < \mu_0$), then the lower (one-sided) CUSUM scheme S_i^- will be required. These two separate schemes can be used together to monitor both increases and decreases from the target value.

Suppose $X_1, X_2, \ldots$, is a sequence of measurements on a process characteristic. The X_i's are assumed to be independent and identically distributed. For monitoring the mean of a normal process, X_i are often sample averages, but they can also be individual observations. Let μ be the process mean that is to be monitored. Then the classical tabular normal CUSUM for a two-sided control of the process mean is defined by the recursions

$$S_0^+ = 0 \tag{1.3a}$$

$$S_0^- = 0 \tag{1.3b}$$

$$S_i^+ = \max[\,0,\ \ X_i - (\mu_0 + K) + S_{i-1}^+] \tag{1.3c}$$

$$S_i^- = \max[\,0,\ \ (\mu_0 - K) - X_i + S_{i-1}^-] \tag{1.3d}$$

where S_i^+ and S_i^- are respectively the upper and lower CUSUM statistics at sample i, and where $\max[a,\ b]$ denotes the maximum of a and b. The design parameter K is the reference level, and it represents a level of indifference that prevents the scheme from drifting toward the signal level H when the process is on target (Yashchin [132]). The signal level H is commonly called the decision interval. The process will be declared out-of-control if the CUSUM either signals an upward shift in the process mean ($S_i^+ > H$) or signals a lower shift in the

process mean ($S_i^- < H$). It is common in practice to standardize the design parameters K and H for detecting a maximum specified shift of Δ as multiples of the process standard deviation σ, where $K = k\sigma$, $H = h\sigma$, and $\Delta = \delta\sigma$. In applications, typical values for the design parameters are $h = 4$ or 5 and $k = 0.5$.

Motivating Example: Cattle Feedlot Study

Research for this dissertation was motivated by the challenges encountered in dealing with an extensive amount of information collected from a cattle feedlot study using the radio frequency (RF) technology. The radio frequency technology is based on the Growsafe® System, which is a new feeding behavior technology on the market for commercial feedlot (Sowell et al. [104]; Schwartzkopf-Genswein et al. [100]).

Description of a Generic Feedlot Data

The Growsafe® System typically consists of a black mat that contains an antenna located in the back of the feedbunk, a radio frequency transponder molded into a plastic ear-tag, and a personal computer for the data collection. Each animal in the feedlot is fitted with an ear-tag containing an identification number. Whenever an animal wearing the transponder is close enough to the feedbunk (usually very close, with its head almost lowered into the feedbunk), the system records data such as the identification number, location of the animal in the feedlot, time the visit was made, and the duration of visit. These electronic recordings are sent to a reader panel every 5-6 seconds (an average of about 5.25 sec), and the data collected are sent to a

personal computer for storage and future analysis. This recording of the presence of the animal at the feedbunk is termed a **hit** (or a **feeding hit**).

A generic feedlot data description might look like this: Over 100 cattle will be kept in a feedlot for some prolonged period of time (say, three months), and up to 18 hours of feedlot activity from 6 AM to 12 midnight will be recorded continuously for each animal.

A suitable aggregation interval (or sampling interval) is chosen for the analysis. Typically, a 3-hour aggregation interval will be adequate for this purpose. This aggregation or sampling interval will be selected such that it is large enough in time, to minimize the effects of autocorrelation in the data while retaining enough details about the feeding patterns of the animals in the feedlot.

If a 3-hour sampling interval is chosen, then the 18 hours of feedlot activity will be divided into six 3-hour time periods: 6 AM - 9 AM, 9 AM - 12 PM, 12 PM - 3 PM, 3 PM - 6 PM, 6 PM - 9 PM, and 9 PM - 12 AM. The total number of feeding hits recorded within each 3-hour interval will be taken as one observation, giving a total of six observations per day for each animal. The data for all animals, taken together, will therefore indicate a strong positive skewness and large over-dispersion.

Purpose and Initial Problems

The Growsafe® System was originally developed to improve the diagnosis of sickness and reduce chick mortality in ostriches. The current study however, is being conducted to evaluate the potential for using this technology for a similar purpose in

beef cattle.

Sickness among feedlot cattle continues to have a major economic consequence for commercial cattle feeding industry. A sick animal often experiences loss of appetite, depression and elevation of temperature. If not detected and treated early, a sickness can result in the death of the animal or lead to the sick animal infecting others, which ultimately results in unneeded mass medications. One of the most common sickness is the bovine respiratory disease (BRD), which can also cause detectable lung lesions that result in decreased growth performance of the cattle.

Because medical treatment is more effective the earlier it is administered in the disease, early identification of sick animals is desirable. The current method of identifying sick animals at the feedlot employs pen riders who set out to look for visual signs of morbidity, such as nasal discharge, soft cough and rapid shallow breathing. Obviously, this process is subjective and difficult to monitor.

Goals and Objectives of Current Research

A more objective criteria than the use of pen-riders, such as the amount of feed intake, is needed to more effectively assess the presence and severity of sickness within individual animals. Unfortunately, the Growsafe® System does not provide direct measurement of the feed intake or weight gain after bunk visits.

Animal science literature however, strongly suggests that feeding and watering patterns are directly related to the status of health of the animals in the feedlot

(Putnam et al. [90]); and (Basarab et al. [6]). In particular, feeding frequency (or the number of feeding hits recorded in the duration of bunk visits) is a good indicator of an animal's feed intake while in the stall (Putnam et al. [90]; Schwartzkopf-Genswein et al. [100]). Pijpers et al. [89] found that in general, feed intake decreases with the onset of sickness, and that healthy animals tend to feed more and spend longer time at the feedlot than morbid ones. Longer time at the feedlot translates into higher number of feeding hits.

The main goal therefore, is to develop a simple, on-line statistical process control (SPC) procedure that could be linked with the automatic acquisition of feeding behavior data, to objectively monitor and detect morbid individual animals earlier than conventional methods. I refer to this as the 'broad objectives' or 'our purpose'. More specifically, the application of a cumulative sum (CUSUM) procedure will be explored.

Outline of Related Problems

SPC procedures can also be used for modeling which will involve estimation, hypothesis testing, and prediction. In terms of hypothesis testing we may, for our present purpose, consider the null hypothesis associated with a specific animal

H_{0i}: Animal i is healthy

against the alternative hypothesis

H_{ai}: Animal i is sick.

One can envisage several problems and difficulties in a CUSUM analysis of this problem, such as those listed below:

1. The health of the animal, in this case, is not a variable that is directly observable or measurable by the system. It is latent.

2. Animal's *health* could lie within a continuum of values. Thus, there may not be a clear, distinct threshold or cut-off point for classifying an animal as either "healthy" or "sick". Such a threshold value, even if it exists, may be difficult to find.

3. Data on only one variable (*the number of feeding hits*) is available. A procedure that can best utilize the observations made on this single variable will be most desired.

4. We could not find a suitable transformation that would make these data approximately normally distributed. It is essential however, not to transform the count data, but to leave the data in the original units in order to ensure simplicity of use, evaluation, and interpretation of the results. Fortunately, the CUSUM procedure can be applied to count data.

5. Implementation of a SPC procedure such as the CUSUM scheme requires an explicit and precise statistical model for the observations; see for example Hawkins and Olwell ([47, page 14]). Specifying such a model may be difficult, and would

require extensive historical data. Additional problems may be encountered if the data are over-dispersed.

6. An important assumption underlying most SPC methods is that the observations are statistically independent and identically distributed. In practice however, the assumption of independence is not exactly satisfied by data collected over time.

7. Determination of the model parameters and optimal settings of the design parameters for the SPC or CUSUM procedure may be difficult.

This dissertation will be structured as follows. In Chapter 2, proposed underlying models will be discussed, beginning from the point of view of a stochastic process. Time series and Markov processes will be discussed as special types of stochastic processes. The Poisson process will be discussed briefly. The negative binomial distribution will also be discussed because it is one of the simplest and common models for overdispersed count data. The need for a modified CUSUM scheme will also be discussed in Chapter 2. In Chapter 3, the derivation of the general CUSUM for exponential family of distributions will be given. Count data CUSUMs will also be reviewed in some detail, and a summary of results on count data CUSUMs such as the computation of the ARL using the Markov chain approach of Brook and Evans [17] will be presented. The proposed modified CUSUM methods are presented in Chapter 4. In Chapter 5, numerical examples of the modified CUSUM scheme based

on the Poisson model are presented. Computational results involving summary and evaluation of the procedure, and design implementation are discussed in Chapter 6. An outline of how the procedure may be applied to the generic Feedlot data is given in Chapter 7, and finally conclusions, discussions and suggestions for future research are given in Chapter 8.

CHAPTER 2

PROPOSED UNDERLYING MODEL

The purpose of this chapter is to outline the basic assumptions and to suggest and discuss suitable underlying models (or class of models) for analyzing the type of generic feedlot data described in Chapter 1, keeping in view both the broad objectives and the specific goals that were also stated in Chapter 1. The unpredictable nature of future observations (for example, the number of feeding hits per 3-hr period) recorded on the animals in the generic feedlot, even when each animal is perfectly healthy, suggests a general stochastic model will be a good starting point in the analysis. After a brief general introduction, this discussion will move on to time series and Markov processes as special types of stochastic processes that could be reasonable and relevant models for this problem. The Poisson process will be discussed next because the Poisson distribution is the most commonly used model for count data. The negative binomial distribution will also be considered as one of the several generalizations of the Poisson distribution, and also as a useful, bona fide count data model in its own right. This chapter will also establish the need for a modified cumulative sum procedure, which will be presented in Chapter 4.

Basic Assumptions

Based on the discussions in the previous chapter, the following five basic assumptions are essential throughout this dissertation. Additional assumptions will be made as and when they become necessary.

1. Prior to being received in the feedlot, all cows are checked to make sure they are 'perfectly healthy'.

2. The health status of an animal in the feedlot at any time period t can be "directly assessed" by the cumulative amount of feed it has consumed up to that time.

3. The amount of feed consumed by the cow can be "estimated" by the number of feeding hits.

4. Using a "wide enough" sampling interval can greatly reduce potential serial correlation in the data, and diminish its impact on the analysis.

5. No animal in the feedlot has any influence on, or is influenced by, the feeding behavior of other animals in the feedlot.

Although assumptions 4 and 5 may not sound very reasonable, they are useful in this case for developing simple statistical ideas.

Stochastic Process

Stochastic processes cover a very broad field and have several classes, each with extensive literature base and numerous areas of application. The literature includes texts such as Bhat [8], Taylor and Karlin [111], Ross [96], and also Krishnan [63] which contains a more technical viewpoint. The following definition of a stochastic process is given in Krishnan [63, page 39].

DEFINITION 2.1. A stochastic process $\{X_t, t \in T\}$ is a familiy (or a collection) of random variables defined on the probability space $(\Omega, \mathcal{F}, P)$ and taking values in the measurable space $(\Re, \mathcal{R})$, where:

Ω = nonempty set called the sample space;

$\mathcal{F}$ = a σ-field of subsets of Ω;

P = a probability measure defined on the measureable space $(\Omega, \mathcal{F})$;

$\Re$ = the real line; and

$\mathcal{R}$ = the σ-field of Borel sets on the real line.

The probability space $(\Omega, \mathcal{F}, P)$ is called the *base space* and the random variable X_t is called the *state* of the process at time t (for each $t \in T$). The measurable space $(\Re, \mathcal{R})$ is called the *state space* of the stochastic process and it is the set of all possible values that X_t can assume (Ross [96, page 72]). The time set T is called the *index set* (or *parameter space*). If T is the countable set $\mathbb{Z}$, the stochastic process is often represented as $\{X_n, n \in \mathbb{Z}\}$ and is ascribed the name a *discrete-time process*. When

T is an interval of the real line ($T = \Re$), the stochastic process is called a *continuous-time process.* Thus, a stochastic process is simply a familiy of random variables that describes the evolution through time of some (physical) process (Ross [96, page 72]).

Consider a stochastic process $\{X_t, t \in T\}$ which satisfies $\mathbb{E}[|X_t|^2] < \infty$ for every $t \in T$, where the symbol $\mathbb{E}[\ .\ ]$ denotes the mathematical expectation. Assuming the components of the model are additive, a stochastic representation of the process (i.e. a stochastic model for the process) can be given as

$$X_t \ = \ \mu_t + \epsilon_t, \tag{2.1}$$

where X_t is a random variable representing the state of the process at time t, and where μ_t is a non-random component of the process, and $\{\epsilon_t\}$ are independent identically distributed random variables with a finite variance. The process ϵ_t is commonly called the *innovation* because ϵ_t need not have a zero mean; rather ϵ_t is new or innovative at time t (Joe [53, page 260]).

Analyzing the data via stochastic model (2.1) requires fitting a 'satisfactory' model, which involves parameter estimation and model checking or diagnostics. If an appropriate model can be obtained, it can be used to detect sick animals in the feedlot, and it can also be used to greatly enhance our understanding of the underlying process mechanism that generated the data. The immediate problem however is that the stochastic model given in (2.1) can involve a large number of parameters, and it can also exhibit some very complicated dependence structure such that even the mean function $\mathbb{E}[X_t]$ and the variance function $\sigma_t^2 = \mathbb{E}[X_t - \mathbb{E}(X_t)]^2$ can vary

with time. In view of this, certain "stationarity" conditions will be imposed. In this dissertation, a stochastic process $\{X_t, t \in \mathbb{Z}\}$ where $\{\mathbb{Z} = 0, \pm 1, \pm 2 \ldots\}$ will be called *stationary* (more specifically, weakly stationary or second-order stationary) if its mean and variance functions exist and are both constants (not dependent on the time parameter t), and if its covariance function $\gamma(t,s) = \mathbb{E}[(X_t - \mu_t)(X_s - \mu_s)]$, for all $s, t \in \mathbb{Z}$, is a function only of the time difference $|t - s|$. A stochastic process that does not satisfy these conditions is said to be *nonstationary*. Generally, depending on the nature of the data, it is possible to examine the dependence structure of model (2.1) and determine what kind of time series dependence (dependence decreasing with lag) is involved.

Time Series Process

Time series modeling and analysis have been around for a long time, and have found applications in a variety of areas in scientific, commercial, industrial, and socio-economic fields. Some of the more accessible texts are Wei [117], and Brockwell and Davis [16] both of which discuss in-depth theory and applications. The traditional analysis of time series assumes an underlying continuous variable (usually a Gaussian random variable). The series is then decomposed into different components such as the trend component, seasonal component, and random noise component. Box and Jenkins developed an effective iterative procedure for analyzing time series models. Their method also provides techniques for handling certain nonstationary time series