8-2015

Survival Analysis, Recidivism, and Booking Data from the Stearns County Jail

Thomas Erdahl
St. Cloud State University

Survival Analysis, Recidivism, and Booking Data from the Stearns County Jail

by

Thomas Erdahl

A Thesis

Submitted to the Graduate Faculty of

St. Cloud State University

in Partial Fulfillment of the Requirements

for the Degree

Master of Science in

Applied Statistics

August, 2015

David Robinson, Chairperson
Shiju Zhang
Mario Hesse

Abstract

This Master's thesis examines booking data from the Stearns County Jail covering the period of January 1, 2003, to January 31, 2015. The examination of the data will make use of Microsoft Excel, JMP, and SAS. The analysis will focus on recidivism, where recidivism is defined as being re-booked into the Stearns County Jail. The analysis will show a lognormal distribution best approximates the process at work behind the recidivism of those individuals going through multiple bookings in the Stearns County Jail. Differing recidivism for males and females are also revealed by the application of survival analysis, and gender is used as an explanatory variable in the lognormal distribution used as a model.

Acknowledgements

Kimberley Scotti

Dr. David Robinson

Joseph Kustritz

Table of Contents

Chapter Page

List of Tables

Table Page

List of Figures

Figure Page

Chapter 1: Introduction

The intent with this thesis is the examination of recidivism. To that end, data was obtained from Steans County that had bookings at the Stearns County Jail over the period 1/1/2003 to 1/31/2015. The definition of recidivism that could be supported by the data was simple: once a person was booked at the Stearns County Jail, recidivism was defined as being booked again at the Stearns County Jail.

The data begins with two Excel spreadsheets provided by Stearns County that lists bookings at the Stearns County Jail. The first of these spreadsheets lists bookings at the Stearns County Jail covering the period, January, 1, 2003, to October 30, 2014. This spreadsheet had 114,215 records. The second and smaller of the two Excel spreadsheets lists bookings at the Stearns County Jail over the period October 31, 2014, to January 31, 2015. This second spreadsheet had 2,139 records. The variables, or Excel columns, were the same and were presented in the same order. The two Excel spreadsheets were easily combined into one Excel spreadsheet of 116,354 records.

A unique booking number is assigned with every booking, and the construction of this booking number, in conjunction with the date and time of booking, enabled the use of survival analysis in the study of recidivism. The booking numbers uniquely identified individuals and the number of bookings each person has undergone. The booking number always includes an extension, a hyphen followed by three digits, for example, -001 or -002, that tells the number of bookings an individual has undergone at the Stearns County Jail.

The combination of this booking number, together with the date and time of the booking, is powerful information. There is much to examine from these simple pieces of information.

In the parlance of survival analysis, a "lifetime" could be defined as beginning when an individual was "released" from the Jail following a booking. This "lifetime" would be ended by an event defining a "failure"—that event would be the next booking of that individual into the Stearns County Jail.

As simple as this idea sounds, the data was far from ideal. Many problems and obstacles had to be overcome in order to use the data in the manner described. The raw data needed much "cleaning" which will be described in detail later. The final result of this work was a set of data in Excel that listed 88,768 unique booking events that occurred at the Stearns County Jail over this period of 1/1/2003 to 1/31/2015.

This cleaning and culling of the data in its Excel version, as well as the definition and creation of new variables, are described in detail. Many new dichotomous variables (or to use other words, dummy variables) were created that allowed many details to be revealed about the data. An example of this would be discerning which bookings were first bookings for a person, as opposed to bookings that are first shown for that person in the data (when there were earlier bookings that took place before 1/1/2003).

The only explanatory variable that could be used turned out to be that of gender. The age, race, birthdate, or other demographic information was not a part of the data. Two variables regarding the nature of the offense and its text description turned out not to be consistent enough upon which to base analysis. With this data, survival analysis rested on

the booking number and the timing of the booking events, with the additional dimension of gender.

The data was examined as a whole, as well as by booking number and gender. Descriptive and non-parametric examination of the data is performed. The three fundamental graphs of survival analysis—survival, probability density function, and hazard—are obtained from the data in a non-parametric manner. There is extensive use made of Kaplan-Meier plots of the data. Such KM plots are made of the entire set of data as well as in examining the gender and booking level.

Following the descriptive and non-parametric examinations of the data, a parametric description of the data is attempted. When JMP was used, it was found that a Frechet description of the data seemed to fit best, although a lognormal model is explored using SAS. SAS is not able to use a Frechet distribution. However, an issue becomes apparent when viewing the data, and it is a problem that has long been remarked upon not only in recidivism data, but in other applications of survival analysis

The issue is simply that much survival analysis assumes that failure events are inevitable and that all subjects under study will eventually succumb to the failure event in question, in this case recidivism. The problem is revealed when comparing some descriptive measures of the data obtained non-parametrically with descriptive measures obtained when assuming a lognormal distribution of the data, or in the case of assuming a Frechet distribution through the use of JMP.

There are many statistical issues of interests with this data, one of which is possible folly of assuming that all subjects will fail and using a statistical machinery that operates under that assumption. If there is any recidivism data where 100% of the people involved recidivated, it is certainly unusual.

Another issue is the possible artificiality of choosing a particular distribution to fit data, in particular, social science data. With this data, some changes were made in how recidivism time and follow time for censored observations were calculated. According to JMP analysis, the "best" distribution changed from lognormal to Frechet. What does this mean? Recidivism is a phenomenon that is dependent on data and definition, and is subject to policy decisions as well as decisions in how the data is processed. There is a fundamental question in such cases as to whether there is an underlying phenomenon that is independent of these relatively superficial decisions and attributes.

It is not meant to imply that it is useless, pointless, and uninformative to apply a particular distribution, whether it be Frechet, lognormal, or some other kind of distribution. What is meant is that, especially with social science data, there may be an underlying distribution, but it may very difficult to satisfactorily apply an accurate one to any actual data given these dependencies on data, definition, and assumptions. The search for a distribution should be regarded as searching for something that fits the data at hand for a particular purpose, with an awareness that this may be a fragile fit that depends on the details of data, definitions, and assumptions. Humility on the part of the analyst is essential.

Chapter 2: Literature Review

Overview

This literature review is regarding a statistics Master's degree that examines statistical analysis of the topic of recidivism in criminal justice. There is a large amount of literature on this issue. Even confined to the subtopic of statistical analysis, the volume is truly humbling. More specifically, the interest here is the application of survival analysis to recidivism.

The Bureau of Justice Statistics released a report on recidivism in September, 2014. According to this report on the numbers of incarcerated people in the US, as of December 31, 2013, there were an estimated figure of 1,574,741 prisoners incarcerated in both federal and state prisons in the US (Carson, 2014, p. 2). This also represented an increase of approximately 4,344 over the figure as of December 31, 2012 (p. 2). If one focuses on only inmates who are incarcerated for more than one year in federal or state prisons, the raw total number of incarcerated people increased by 5,400 from December 31, 2012, to December 31, 2013. However, because of the increase in total US population, the actual rate of incarceration such prisoners incarcerated for more than one year, actually decreased from 480 to 478 people per 100,000 (p. 1).

Whether measured by cumulative numbers or percentage of population, these are the largest figures in the world for any developed nation (International Centre for Prison Studies, 2014).

The Bureau of Justice Statistics had released another report on recidivism, this one in April, 2014. There were 404,638 prisoners in 30 states released from state prisons in 2005 that met BJS criteria for inclusion. Multistage random sampling yielded a sample of 69,279 individuals. These individuals then had their five year history following their 2005 release examined (Durose, Cooper, & Snyder, 2014, pp. 1, 17-18). Of the sample, 67.8% were arrested for a new crime within three years of release and 76.6% were arrested for a new crime within five years of release (p. 1). Also, 16.1% of those tracked were responsible for 48.4% of the arrests, although the report does not specify if these are arrests for new crimes or simply arrests for any reason (p. 8).

An examination of the issue of recidivism is important not only to society and to recidivists themselves, but also to those who operate jail facilities. One small study of the Stearns County Jail in Minnesota found that, for February and March of 2008, those who have been booked three times or fewer had an average stay of seven days while those who have been booked more than ten times had an average stay of 30 days (Cunniff, 2008, p. 9).

This same study, referring to a larger database with which to study jail bed usage, asks with an unintentional poignancy: "There is, for example, the ability to examine those persons who endure multiple bookings into the jail. What brings these people back to the jail over and over again? What factors drive up their jail stays?" (Cunniff, 2008, p. 15).

It is not difficult to establish that recidivism is a problem. Recidivism represents a terrible problem and a great expense for a society, a nuisance and a danger to the people in that society, and a wasted life for the recidivist himself. Everybody, including the recidivist,

has a stake in examining the issue of recidivism. Everyone has a stake in finding out what works and what does not in reducing the rates of recidivism. Hence, the importance of statistical analysis of recidivism.

Issues and Difficulties with Crime Statistics

Mosher, Miethe, and Phillips (2002) described many problems with gathering reliable crime statistics in their book, *The Mismeasure of Crime*. This book can leave a person in despair of ever seeing crime statistics without "issues" of one kind or another. There are three ways of gathering crime statistics: official crime statistics, such as the FBI's Uniform Crime Report (p. 59); self-reports of criminal activity from survey respondents (p. 101); and victim self-reports from survey respondents, referred to as "victimization surveys" (p. 135).

Official data can have many problems, as outlined by Mosher et al. (2002). Some issues are different jurisdictions defining and recording crimes differently (pp. 62-63), different record-keeping policies (pp. 65-66), and various incentives and interests in not recording crime numbers accurately (pp. 91-93). There is also an inherent ambiguity in classifying complex real incidents, as criminal acts can be, into neat categories (pp. 65-69).

Mosher et al. (2002) describe self-report surveys of parts of the general population and also of criminals regarding crimes they have committed (pp. 101-133). The information gathered would be direct and would not have to be obtained through the filter of official crime statistics. These are surveys and are subject to all the obligations imposed by survey methodology (pp. 102-105), and can have limitations. In particular, there are weaknesses as might be expected of a survey that relies on individuals to self-report illegal activities (p.

102). The problem with this data, as far as survival analysis of recidivism is concerned, is that it relies on human memory without any solid connections between crimes and when they were committed (pp. 101-102).

Victimization surveys would have similar issues regarding survey methodology (pp. 135-169) in addition to unique problems (p. 136). Both kinds of surveys have particular challenges with reliability and validity of responses (pp. 105-106, 159-168). In as far as the current thesis is concerned, both kinds of surveys, as a means of gathering data on criminal activity, have severe problems with time in that there are few clear and reliable markers in time as to when crimes occurred because of the reliance of human memory as to when events occurred (pp. 101-102, 161-163).

These issues would be an obvious problem for any statistical analysis that relies on clear markers in time, such as survival analysis. Again, this brings up the need for data that has clear and reliable divisions in time, and hence, one is led logically to data that records legal events in criminal history, such as arrests, charges, convictions, incarceration, and release from incarceration (Maltz, 1984, p.22).

Let it suffice to say that data regarding criminal justice tends to be messy, complicated, and can vary depending on jurisdiction source or record-keeping approach (Mosher et al., 2002, pp. 59-60). In this light, it is interesting to note that with a BJS study on recidivism from 2014, there is a warning about comparisons with some recidivism statistics found in BJS studies done on recidivism in the 1990's: "…direct comparisons between the published recidivism statistics should not be made" (Durose et al., 2014, p. 2). The problem

was one of different kinds of data used in the respective studies: Different attributes and demographics of the prisoners who were tracked seemed to be the main issue (Durose et al., 2014, p. 2).

Recall from the beginning of this literature review where a report from the Bureau of Justice Statistics (BJS) provided an estimate for the number of people incarcerated on December 31, 2013. This total was approximately 1,574,741 and was derived from all prisoners under jurisdiction of the federal government and the state governments, where "[j]urisdiction refers to the legal authority of state or federal correctional officials over a prisoner, regardless of where a prisoner is held" (Carson, 2014, p. 2). For example, it includes federal prisoners who are being "held in nonsecure privately operated community corrections facilities and juveniles held in contract facilities" (p. 2).

It is unclear from this document how all of the District of Columbia prisoners are defined, therefore one has to infer. There is a statement appearing under several tables in the report that reads, "[a]s of December 31, 2001, sentenced felons from the District of Columbia were the responsibility of the Federal Bureau of Prisons" (Carson, 2014, p. 11) which would raise the question of where District of Columbia non-felons can be found in the various tables that list federal and state prisoners. This is especially relevant for the state prisoner tables which list state-by-state totals, but does not show the District of Columbia (p. 3).

This question is also raised by another section of the report that states, "[t]he District of Columbia has not operated a prison system since yearend 2001" (Carson, 2014, p. 28).

However, "[j]ail inmates in the District of Columbia are included in the Annual Survey of Jails" (p. 28). There is a distinction to be made between jail and prison, and perhaps the report is calling attention to this distinction. Generally speaking, jail is considered a holding place for those who have been arrested, waiting for charges to be filed having been arrested, waiting for trial having been charged, or for those on trial if bail had not obtained. It could also be used for people are incarcerated for less serious crimes with a relatively short incarceration time. Prison is where people are incarcerated after being sentenced as punishment for committing a crimes, especially more serious ones, and where the incarceration time is relatively long. The thing to be noted, however, that the reader has to infer all this from the report. It is not stated.

If a reader puts these statements together from this report about the District of Colombia, it would appear the District of Columbia has jails but no prisons since 2001 and District of Columbia felons are incarcerated at federal prisons and are included in the report's population estimate, since federal prisoners are included. It has to be inferred that non-felons are housed in District of Columbia jails, but are not included in the totals of this particular document, since, as previously referenced, the District of Columbia is not listed on the tables which give the total population figures for each state. This digression into issues regarding the District of Columbia was done as a brief example of the complexities of criminal justice statistics. As a final detail to illustrate this, under that table that shows the state prison populations, there is a statement that says it "[i]ncludes imputed counts for

Nevada" (Carson, 2014, p. 3). One might have the expectation that it would be a very easy thing to simply count the people who are incarcerated, but not so.

Such is the nature of criminal justice statistics. It may appear to be an obvious and safe statement to make, but in statistical studies of any kind the nature of the data, along with its qualities and subtleties must be part noted as part of the analysis. It is the responsibility of the investigator to notice and address issues related to the integrity, reliability, consistency, and appropriateness of the data. This statement can be made regarding any statistical analysis of any sort of data in the social sciences, but is particularly true regarding analyses using criminal justice data. With criminal justice data the expectation should be that it will be messier, and require greater effort to obtain and understand, and clean, than data from other areas of study.

Brief History of Crime Statistics

How far back do crime statistics go? Not statistics about recidivism, but simple crime statistics? Mosher et al. (2002) are rather vague on precise terminology, but the following information can be garnered from their book: First, there was simple "judicial data," which are cumulative counts of legal proceedings (such as convictions and executions) kept by separate jurisdictions within nations. This was haphazard and variable in practice. The next development were "judicial statistics," which are more organized data collected from different jurisdictions (pp. 27-32). As an example, judicial statistics were not collected in England until 1805 (p. 32).

The “first national crime statistics” were done in France in the 1820s (pp. 27-28). Mosher et al. (2002) imply without saying directly that what set this apart from the “judicial statistics” mentioned previously was the attempt to gather data regarding the entire nation, describe the crime situation in that entire nation, and perform something we would recognize as statistical analysis (pp. 27-32). For example, the French analyst found there was “one accused person for every 4,463 inhabitants” for the period of 1826-1829 in France (p. 28). From our perspective, it is almost ironic that even with these early analyses there was a connection noticed between crime and alcohol, crime and poverty, crime and gender, and crime and age (pp. 27-30).

Other things were soon apparent even in these 19th century attempts to analyze crime. One is that measures of crime depend on how crime is defined and recorded by that society. A change in law can create crime that did not exist before, such as when Britain began requiring parents to send their children to school in 1870 (Mosher et al., 2002, p. 32) or when the New York City Police Department changed crime record practices in 1950 that resulted in large apparent increases in crimes such as robbery and larceny over a 1-year period (p. 39).

Mosher et al. (2002) mention one particularly important detail about crime statistics that is something that has always bedeviled analysis and continues to be a problem today. It involves how different jurisdictions will define and/or count crimes differently and have very different raw numbers as a result (pp. 37-38). Strictly speaking, they are referring to American crime statistics, but it is not difficult to see how the same problem would occur in

any country. A related point was made by Nagin and Pepper (2012) in their article as one of the reasons to be skeptical of the analyses done on the issue of deterrence and the death penalty. Different states have different sanction regimes regarding the death penalty, and sanctions other than the death penalty, and previous analyses have not taken that into account (Nagin & Pepper, 2012, pp. 36-37).

The Study of Recidivism

In the field of criminal justice studies, it is very common to examine a group of offenders to sort out those who have had a "successful outcome" from those who have not. The point of such comparisons and analyses is to try to find out what factors led to a successful outcome or find out the factors that led to an unsuccessful outcome, or to evaluate some program or some approach in its effect on recidivism. The obvious hope is for the criminal justice apparatus do more of the factors that lead to successful outcomes while lessening, changing, or eliminating the factors that lead to unsuccessful outcomes.

The study of recidivism is now a long-standing scholarly discipline wherein much study and examination is made of the process by which criminals are treated and, it is hoped, rehabilitated. The application of survival analysis to the study of recidivism has been around since the early 1970s (Schmidt & Witte, 1988, p. 16).

An organization called The Sentencing Project has an on-line database on studies of recidivism made between 1995 and 2009, covering all 50 states and the District of Columbia. The organization does not claim the list is comprehensive for the US. In fact, the study by Duwe and Kerschner (2007) referred to later is not listed, so there are undoubtedly more

studies that can be found for that time period. There are 99 studies on that database (The Sentencing Project, 2010).

One of the first things a reader of the literature can note is that terminology is critical. Terms such as “jail,” “prison,” “sentence,” and even “recidivism” are often used without clear definitions. As a more specific example, Nagin and Pepper (2012) expressed concern in their report over the lack of clarity and the confusion over the terms such as “homicide” and “murder” (p. 20). Other examples of this will be seen as our discussion progresses.

An article in a periodical was one original source for the interest in this topic but, oddly enough, it was not about recidivism. It was about the statistical analysis involved in finding a connection between the death penalty and crime deterrence. This article was in the statistical magazine, *Significance*, and it stated something astounding. The issue was whether or not the death penalty was a deterrent to murder. It was written by Daniel Nagin, who had been the co-editor of a report by the National Research Council entitled, “Deterrence and the Death Penalty”, completed in April of 2012. The article in *Significance* was an explanation for the conclusion of that report by the committee (Nagin, 2014, p. 9).

Since the Supreme Court lifted the moratorium on the death penalty in 1976 many studies and analyses had been done regarding the issue of the death penalty being a deterrent to murder. The conclusion described by Nagin in his article that was so astonishing was this: all of the research done since the moratorium was lifted, now covering 38 years, on the issue of whether or not the death penalty has a deterrence effect on

murder, should be ignored. That is, ignored for purpose of informing public policy. The following conclusion comes from that original report:

> CONCLUSION AND RECOMMENDATION: The committee concludes that research to date on the effect of capital punishment on homicide is not informative about whether capital punishment decreases, increases, or has no effect on homicide rates. Therefore, the committee recommends that these studies not be used to inform deliberations requiring judgments about the effect of the death penalty on homicide. Consequently, claims that research demonstrates that capital punishment decreases or increases the homicide rate by a specified amount or has no effect on the homicide rate should not influence policy judgments about capital punishment. (Nagin & Pepper, 2012, p. 2)

The reader is invited to ponder this for moment. Since 1976, all of the research, all the statistical analyses, all the reports and studies that have been written, are for naught and can be disregarded.

The reasons for this conclusion are related to the data used and the statistical analyses that have been done. There are problems with data available (Nagin & Pepper, 2012, pp. 36-37), model uncertainty upon which statistical analysis is based ("strong and unverified assumptions") (pp. 115-119), among other issues. In fact, some studies can have results changed greatly, even to the point of reaching the opposite conclusion, if small changes are made in the model used (p. 115). Thus, in such cases, the results were "fragile" (p. 10).

The report by Nagin and Pepper (2012) and the article by Nagin (2014) are about problems with the statistical analysis of deterrence and the death penalty, which is a different subject than the statistical analysis of recidivism. However, some of the issues seemed reminiscent of the problems in analyzing recidivism. This is especially true regarding

the data availability, quality of the data, comparison of data from different jurisdictions, and the definitions of terms. There is also the inherent problems with taking a large, prickly, and complicated such as recidivism, and finding a way to perform a satisfactory statistical analysis and attempt to draw some valid and reliable conclusions. The common thread to both areas of study are criminal justice statistics, which will be discussed here.

The importance of the data and definition issues was further emphasized by a conversation with a Stearns County corrections official who was asked about what she thought was lacking in studies of recidivism that she had seen. Her response was that recidivism was treated too simply. She thought of it as a complicated issue, and for it to be treated as one unitary concept without differentiating between different levels or kinds of recidivism was a disservice (Becky Bales-Cramlett, personal communication, 10/20/2014).

Indeed, it is a complicated issue. For the purpose of statistical analysis, and to begin to understand the point being made by the Stearns County official, was one must first look at the term “recidivism” since the concept must be defined in order to analyze it adequately. Issues that have come from previous attempts to define recidivism will be discussed.

Definitions of Recidivism

Most people would roughly define recidivism as the event of an individual committing a new criminal offense after committing previous offenses. More specifically, supposed there is a person who has committed one or more criminal offenses and has been caught and punished for at least one of these offenses. In the aftermath of suffering the punitive consequences for those offenses. He commits a new crime when he has freedom

and opportunity. Most discussion or studies of recidivism do not go much beyond this. Here is a definition of recidivism from an on-line dictionary that is extremely superficial, but would encompass much colloquial thinking on a definition: “The behavior of a repeat or habitual criminal. A measurement of the rate at which offenders commit other crimes, either by arrest or conviction baselines, after being released from incarceration (TheFreeDictionary. com, 2014).

This definition has problems, but is included here to provide an example of a popular view of recidivism. One problem with this definition is that recidivism is not an attempt at “measurement of the rate at which offenders commit other crimes.” For one thing, many crimes go unreported (Mosher et al., 2002, pp. 54-55). When examining the topic of crime and how it is dealt with by society, a very common analogy one encounters is that of a funnel.

To understand the funnel analogy, suppose one were able to look at all offenses committed and all the offenders who committed those offenses, for example, in one year. The number of people eventually going to prison for the commission of some of those crimes would be much smaller than the real number of offenders. Even when one looks at reported crimes, “from police statistics to imprisonment data” the number of offenders dwindles as the legal process advances through its stages (Mosher et al., 2002, pp. 59-60). Rate of arrest or rate of incarceration would seem to be a poor way to find out the rate of crime commission by a particular offender, unless one could verify with certainty that any crimes committed by an individual were all punished by incarceration.

As an example, an article in *the New York Times* dated June 12, 1994, was entitled, "The Crime Funnel". According to this article, if we define the most serious crimes as "homicides, rapes, robberies, aggravated assaults, burglaries, larcenies and auto thefts", as of that time, about 15 million such crimes were reported to the police in a year and 21% of these crimes resulted in an arrest made (*New York Times*, June 12, 1994).

As another example of a recidivism definition, the following is from the National Institute of Justice:

> Recidivism is one of the most fundamental concepts in criminal justice. It refers to a person's relapse into criminal behavior, often after the person receives sanctions or undergoes intervention for a previous crime. Recidivism is measured by criminal acts that resulted in rearrest, reconviction or return to prison with or without a new sentence during a three-year period following the prisoner's release. (National Institute of Justice, 2014)

This definition has a number of aspects to note. First, note the need for the experience of being caught, convicted, and punished for a crime at least once, since there must have been "sanctions" or "interventions" for a "previous crime". It is not stated directly, but is certainly implied by the use of the term "prisoner's release," that incarceration is involved in the "sanctions" or "interventions." Not all crime results in incarceration. Some offenses are punished without any incarceration, of which traffic tickets are a trivial example. Without stating it directly, this definition narrows the focus of recidivism to offenses serious enough to have warranted incarceration as a penalty and then evaluating how well the offender responds once there is a release from incarceration.

Based on this definition, we can view recidivism as a linked process where the first item in this process is to be incarcerated, the second item is to be released from custody,

and the third item is “rearrest, reconviction or return to prison with or without a new sentence,” as stated in the previous block quote. If this third item occurs, it can be said that the offender is a recidivist. By using the fact of an initial incarceration, offenses that did not result in an initial incarceration are left out of the discussion.

This may be an unsatisfactory result to some. Just because an offense did not result in incarceration should not mean the offense was not “serious” or someone was not harmed. In fact, in an ideal situation, all offenses above the level of traffic offenses would be considered, but the problem may well be one of what data is obtainable. It would extremely difficult to obtain a comprehensive list of ALL offenses in a jurisdiction (above the level of traffic infractions) together with the history of the offender after their offense was assessed and a punishment exacted. A running theme throughout this narrative relates to the data that is available for analysis. Sometimes the data that is desired, such as described above, is simply not available or is very limited.

To return to the definition provided by the National Institute of Justice, notice that there is no place for crimes committed during incarceration, such as crimes committed on computer, assaults on other prisoners, or assisting in criminal activities on the outside. This would be troublesome for definitions of recidivism. For example, if an incarcerated person assaults another prisoner, would we say that the offender has recidivated before they were released from custody?

Suppose an incarcerated person commits identity theft while in prison, but it is not found out until after he is released from prison. He is then arrested, charged, convicted, and

returned to prison. The complexities of definition when it comes to recidivism are apparent in the description of situations such as this. One can wonder if this would count as a recidivism when this "new" offense was committed before he was released from prison. We usually think of recidivism as a return to crime, while this incident would indicate that the offender had never left it. Furthermore, suppose an incarcerated person commits the same identity theft offense while in prison, but in this case, the crime is detected while he is still in custody. He is charged, convicted, and incarcerated—but in this case, it means he is not released from prison. As soon as the current sentence ends, a new sentence begins for the newer offense, and he is not released from incarceration. Do we view this as recidivism? A new offense was committed, but there was never a release from custody.

The details of each of these two scenarios would muddy the application of survival analysis. If survival analysis, or more specifically, survival time is "[t]he time until the occurrence of a particular event such as death or the failure of a component" (Upton & Cook, 2006, p. 414), then these two scenarios would pose problems if the event is recidivism and the time measured is the interval from release from incarceration until whatever event defines the recidivism.

Still in reference to the previously given definition of recidivism, note something terribly obvious: that a first time offender is not a recidivist. From the perspective of survival analysis, for a first time offender the initial event, release from incarceration, is not present. Furthermore, the same is true of a habitual criminal who may have committed several crimes but has never been "caught" (or at least, arrested). For such a person, committing a

new crime is not the act of a recidivist since there has been no legal consequence for any previous crime, no incarceration. If the person has not been "caught" for any previous crimes there will be no incarceration as a beginning event on which to base a recidivism. Also, there will be no record as part of any data to be analyzed for recidivism.

To summarize and build upon the previous discussion of recidivism, we start with an individual in incarceration for a criminal offense. It is possible to simply begin with a person being convicted of an offense with or without incarceration, but then we would need to have some standards of offense level. Otherwise, we would find ourselves in a mess of data definitions and problems with data integrity because we could view someone having two traffic tickets as being a recidivist. Getting data complete and reliable enough to include details of an offense, the legal repercussions, subsequent history, and enough details to have meaningful separations between "significant" and "insignificant" offenses is a tall order, or at least it is at the present time. Note that incarceration is a legal event that will leave reliable data records, and incarceration can be used as a rough indication of the seriousness of an offense. The crime and the subsequent incarceration are an indication of conduct detrimental to the larger society and therefore of great interest to all to determine if the individual returns to this mode of conduct, i.e., recidivates.

Furthermore, it can be conjectured that the necessity of incarcerating an individual is a great burden on the society which has to create and maintain the mechanism of incarceration, and therefore it is incarceration that is of concern. For an individual to return

to the status of being incarcerated is something that would be very undesirable for everybody concerned, not simply the person being returned to incarceration.

Because of this, it would certainly seem logical to view incarceration as the central point in the process of evaluating recidivism because of the relative cost to society as compared to committing offenses that do not result in incarceration. For the statistical analyst, incarceration events will leave relatively reliable and obtainable data records. The event of incarceration will also enable the analyst to minimize data evaluation issues such as finding a way to separate offenses based on seriousness.

There is a further issue to note regarding this definition of recidivism from the National Institute of Justice. It comes in the phrase "… return to prison with or without a new sentence…" The clear implication here is that an offender returning to prison without committing a new offense could be considered a recidivism. How could this happen? A parole violation or violating some term of a supervised release, such as failing a drug test or violating a restraining order, may result in a re-incarceration. Some people might offer the opinion that violating a restraining order is a new offense, but note that when we use the term "committing a new offense," it was meant to refer to a new crime with new charges and a new conviction. If the event happens where an offender is re-incarcerated without being charged for a new offense, we will refer to this as a "technical violation" in the sense he has violated some term or terms of a supervised release.

Undoubtedly, for some observers, viewing a technical violation as an incidence of recidivism would be an unexpected twist in the view of recidivism. To the uninitiated,

recidivism is about a criminal who is committing new crimes. The idea of someone being returned to jail without being charged with a new crime, but instead because of the ramifications of a previous offense, may seem like something other than recidivism. To others, it may seem very natural to consider this an extension of the idea of recidivism. After all, the person is freed, under supervised release, and should have a clear understanding of what needs to be done to stay out of prison. To such a view, a technical violation is an indication of a return to crime, thus an indication of recidivism.

A complete statistical analysis of the topic of recidivism would most likely prefer to look at the issue of recidivism both ways. Thus define recidivism as being the result of new offenses and also define it in a way that can view recidivism as resulting from technical violations. Looking at it in both ways is sure to be more informative than looking at each in isolation from the other.

The final topic to discuss from the recidivism definition given by the National Institute of Justice is that of time. This definition gave a time limit, or to use other words, a follow-up period of three years after an offender's release from prison. The implication here would be that an offender could be free from custody for four years, for example, and not commit any new crimes. In the fourth year, he commits a new crime, is caught and punished. By the definition of recidivism given above, this would not be considered recidivism because it occurred outside the time frame of three years specified by the recidivism definition. The problem, however, is that if the data is being analyzed for a particular study, when such an incident of recidivism occurs, the study is probably already completed. The same is true of

any study of recidivism: it will have to end, the report written, and published somewhere. If it is being read, the study has been completed, and any of the subjects followed who have not recidivated may have done so after completion of the study. Hence, the need for censoring some observations in survival analysis. The observation has to end, but the event has not yet happened to some of the subjects.

The point is to distinguish between a definition of recidivism in the abstract—being released from incarceration and committing a new crime—as opposed to a definition of recidivism used in a study. Studies must end at some point in time, so if recidivism over time is being studied, it follows that there must be a defined time interval, a follow-up period, during which occurrences of recidivism are observed or not observed. Because of this necessity of limiting the time of follow-up, it follows that there will be recidivism that occurs after the follow-up period ends.

Nuances of Recidivism Analysis

Obviously, any study has to end at some point in time, and if recidivism occurs after the follow-up period used in the study, it will not show up in the study as recidivism. All studies of recidivism over time will have an arbitrary time limit else the studies would never be published, statistical or otherwise. The only sure, but wildly impractical, way around this is for a study to follow a group of released offenders until re-incarceration or death.

This need for a follow-up period is not unique to survival analyses. It is necessitated by studying recidivism over time, using any sort of study, statistical or not. The recidivism definition given by The National Justice Institute used three years. This appears to be the

most common length of time. The Sentencing Project organization has compiled a database of 99 studies of recidivism completed over the years 1995-2009. Of these 99 studies, 42 used a follow-up period of 3 years (The Sentencing Project, 2010). Three years matches the definition given previously for the recidivism used by the National Institute of Justice.

No other time period was even close in frequency. The second most frequently occurring time interval for evaluating recidivism was the use of mixed time periods for follow-up. This occurred in 18 of the 99 studies (The Sentencing Project, 2010).

As far as could be determined, there was no statistical reason for using one length of time over another, other than the obvious reason that the longer period of observation can capture more accurate assessments of recidivism. The nature of recidivism does create some considerations for the time period chosen. In their book, *Predicting Recidivism Using Survival Models*, Peter Schmidt and Ann Dryden Witte (1988) observed that a great deal of recidivism occurs within the first year or so. For one cohort, the peak hazard rate occurred at 15 months, while the other cohort had a peak hazard rate at 6 months after release (p. 29). Recidivism data generally speaking shows a hazard rate that increases and reaches a peak shortly after the release date. The hazard rate then begins to decrease, and become quite small by the end of the observation period. In fact, using follow-up times of 70 months and 46 months for their cohorts, Schmidt and Witte (1988), had some difficulty in finding a statistical distribution for a model where the hazard rate would drop as rapidly as shown in the actual data (pp. 81-82).

With this behavior seen in recidivism data and in the absence of statistical or data reasons, follow-up periods of a year or less would certainly seem suspect. Two years would seem to be a minimum length of follow-up period. Three years, the most common follow-up length of time seen, is understandable as the most common because it will capture the bulk of the recidivism while being a manageable length of time for those who are studying the data. After all, studies (and their observation periods) must come to an end.

In survival analysis, there will be a finite follow-up period. This would leave us with an obvious scenario. Suppose there is a follow-up period of three years and we have two released offenders who are released from incarceration on the same day, but one commits another offense and is "caught" (whether that means he was arrested, convicted, or incarcerated) just one day inside of that three year limit. The other person commits an offense and is "caught" just one day outside of that three year limit. The first would count as recidivism and the second would not, at least as far as that study is concerned.

In fact, it would appear this would be true if the two individuals committed their respective offenses on the same day within the three year limit, but were "caught" on different sides of that 3-year limit. One situation would be considered recidivism and one would not. Unfortunately, this sort of event would appear to be unavoidable. In an analysis of recidivism, whether statistical or not, the analysis will have to end, or at least define an ending point in time for the data, there will always be the possibility of subjects who recidivate after that defined interval of time. There will also be the possibility that an individual committed a crime but was not "caught", therefore leaving a recidivism

undetectable to the record-keepers and researchers. This would appear to an interested reader as recidivism in the abstract, in the sense that a released offender has committed a new offense, but for the purposes of those performing the study, it cannot count as such.

In the most fundamental application of survival analysis, for example, evaluating the lifetimes of electric components such as light bulbs, the "lifespans" (time to failure) are analyzed and compared. The most well-known extension of this idea is to medical treatments, such as for cancer. Treatments are given, and the lifespans of the patients are analyzed (Singh & Mukhopadhyay, 2011, p. 145). The further extension of this idea to recidivism is that an incarcerated prisoner is freed from custody, and the length of time he "stays out of trouble" and "goes straight" is analyzed. Getting "caught" terminates that length of time and is akin to the electrical component failing or the patient dying.

However, in recidivism, there are complications that arise in the application of survival analysis, complications that do not pose problems in the other situations that are most popularly associated with survival analysis. In the previous paragraph, quotes were placed around terms that are much more complicated than might appear. For electrical components, unit failure is a straight-forward unambiguous term. For patients undergoing treatment for a medical condition, death or otherwise not responding to treatment is a relatively clear situation to observe. Recidivism, however, is not so clear. Recidivism can be subject to interpretation, can have different definitions, and is woefully dependent on good record-keeping.

Consider the terms “staying out of trouble” and “getting caught.” In discussing criminal justice issues, a person may use these terms freely without realizing these terms are much more nebulous than he or she may think. If an incarcerated person is freed from custody and commits a crime, if the crime is not noticed or goes unreported, there is no recidivism that will be part of any data for later analysis. The crime, and the resulting recidivism, have to be defined in data by a legal event. The only legal events that can possibly fulfill this role are being arrested, being charged, being convicted, or being re-incarcerated.

Hence, in survival analysis, “staying out of trouble” would mean avoiding whatever legal event is being defined to indicate recidivism. If we define re-incarceration as what creates an incidence of recidivism, then simply being arrested would not be enough to consider the person a recidivist. Bear in mind that for the people who know the person who was arrested, being arrested would certainly indicate the person was getting into trouble. However, if recidivism is defined as being re-incarcerated, for statistical data to be analyzed, a dummy variable indicating recidivism would still have a value of zero for the person who is arrested. Similarly, “getting caught” would mean that the person underwent the legal event used to define a recidivism. If we are using re-incarceration as the legal event defining recidivism, getting arrested without re-incarceration would not constitute “getting caught,” as strange as that may appear.

The heart of the concept of recidivism is the issue of a criminal committing new offenses or refraining from criminal acts. But as we have discussed, it is difficult to support

the idea of recidivism without the use of a legal event to define the recidivism, such as being arrested, being charged, being convicted, or being re-incarcerated, things that can be borne out by data.

In view of this, recall our previous mention of "technical violations," where an offender is on parole or other supervised release, violates the terms of that release, and is re-incarcerated because of that technical violation. If we are basing a definition of recidivism simply on re-incarceration, this would therefore be a case of recidivism, but perhaps at the expense of some ideal that recidivism should be about new offenses being committed. We could use the definition of recidivism as being re-incarceration for a new offense, but we would need to have data that would allow for this distinction to be made. If our data simply notes a re-incarceration without noting whether it resulted from a technical violation or resulted from a new charge, we are stuck. We could go through each re-incarceration and determine the basis for the re-incarceration, but if the data has 2,000+ separate records to go through, that is a tall order. The researcher may not have access to that information to even go through that procedure.

Also, consider a situation where an offender has been freed from custody, but is found to have committed a crime before the crime that resulted in his initial incarceration, a "cold case", and is caught for that previous offense. One could hold the opinion this would not be considered recidivism because it is not a new offense, but the discussion made previously about re-incarceration because of technical violations applies here, but with new emphasis.

The most obvious event to use for defining recidivism is committing a new offense and the date that it occurred. We have already discussed some of the problems with using a criminal offense as a basis for recidivism. There are a few problems with using this standard. First of all, not all offenses are known and not all known offenses are reported to the police. If the offense is reported, the offense date is not always known with any precision, and given that an offense occurred and is reported to the police, it will not necessarily be solved. Finally, given an offense being observed, reported to police, and solved by the police, it may not result in someone being punished. Recall the funnel comparison mentioned earlier.

The final item to add to this is the need for an event defining recidivism to appear in data, unlike an unreported offense. Given that a legal event needs to define the occurrence of recidivism, what should it be? The choices would appear to be being arrested, being charged, being convicted, or being re-incarceration. The caveat is that it is possible to be arrested and re-incarcerated for technical violations, meaning these could happen without a new crime being committed.

In his widely cited examination of recidivism, which was entitled *Recidivism*, Maltz (1984) examined several studies to see the various definitions of recidivism that have been used. There were 177 studies total. Two of those categories were: "offense data" (which had four sub-categories, one of which was "arrest") accounted for 50 of the studies and "parole-probation infractions" (which had five sub-categories) accounted for 55 of the studies. Then, without sub-categories, were "court appearance" (3 studies), "reconviction" (22 studies), "sentencing" (8 studies), and "return to prison" (39 studies) (Maltz, 1984, p. 63).

A later version of this was done with the database of the 99 recidivism studies compiled by The Sentencing Project organization. This database was mentioned earlier. Recall that this was composed of 99 studies of recidivism done between 1995 and 2009. Only 10 used arrest, 24 used re-incarceration, 29 used conviction, and 30 used some combination of arrest, conviction, and re-incarceration. Five of the remaining six studies dealt with juveniles and had standards for recidivism appropriate for juvenile justice administration, such as a "delinquency complaint" or an "adjudication" in juvenile court (The Sentencing Project, 2010). The remaining study was of "homicide offenders" and the standard of recidivism was "committing another homicide within 5 years" which would have to most likely mean conviction (The Sentencing Project, 2010, p. 6).

It is interesting to note that, aside from these six studies just mentioned, none of the 99 studies used a definition that could be construed as "offense related" aside from that sub-category of arrest or "parole-probation infractions." Since Maltz's paper was done in 1984 and would have used studies done prior to 1984, it can be conjectured that the definitions based on offense (aside from arrest) and "parole-probation infractions" used in the earlier set of studies described by Maltz (1984) proved to be more difficult with which to work with than using the clearly defined legal events that will leave a clear digital record. It would be interesting to see the size of the data used in the earlier studies because it could be speculated that some of those categories would have required a person going through paper files.

In reference to this database of 99 studies, of the 30 studies that used a combination of events to define recidivism, only one used being charged as part of the definition (The Sentencing Project, 2010, p. 7). It is possible to be charged without being arrested. One can speculate that if one is charged, the charges may be dropped, or the legal process may find the person "not guilty." This would mean an awkward situation for a recidivism definition based on being charged in that a person may be charged, found to be not guilty, but still considered to be a recidivist. One can take the view that being charged means the person is not "staying out of trouble," but it may be understandable why the study practitioners may have preferred to use being convicted instead of being charged. Using conviction as a standard of recidivism would mean the person had been charged and the legal process found the person guilty. This would avoid the problems of using the standard of simply being charged as the definition of recidivism.

Survival analysis will need a working definition of recidivism for the purpose of statistical analysis. If we use the results of the previous discussions, it is understandable that one must start with an incarcerated person who is being released from incarceration. This would obviously require that the individual in question has been found guilty of an offense, pleaded guilty, or pleaded no-contest (when that option is available), and is therefore, incarcerated. The fact that the person is incarcerated can be taken to mean the offense committed was very serious, and also the fact of incarceration represents the maximum burden on society, meaning it is very desirable for everybody that the person should not re-offend.

It should be pointed out again that not all offenses that are detected, and have accused perpetrators caught and determined to be guilty, result in an incarceration. It is possible that someone could be punished with home confinement, a fine, or probation. Also, when an offense was not punished with incarceration, it does not mean that it was not a serious matter to somebody. Nor does it mean that nobody was hurt or damaged in some way by the commission of that crime. The trouble comes with obtainable data. If data is available that would allow an analyst to look at all offenses above the level of traffic violations (as an example), whether or not incarceration was involved as an ultimate penalty, that would be ideal. A singular event is needed to be a starting point, one that is easy to record and observe later, and it makes sense that incarceration should be that event.

If our survival analysis starts off with the release from incarceration, whether or not there is parole or some other supervised release. The time spent "staying out of trouble" is period of time of interest, the "lifetime." The event that signals that lifetime has ended is a legal event that is relatively easy to record and become a part of a set of data to be observed later. If we follow what has been seen in 99 studies of recidivism (The Sentencing Project, 2010), we can leave out from consideration the legal event of being charged, and use the one or more of the three remaining legal events: being arrested, being convicted, and being re-incarcerated.

The problem will then be which one, or which combination of these legal events to use. Used separately, each has advantages and disadvantages. In the case of arrest, if someone is re-incarcerated, it is difficult to see how that process would have started without

an arrest, but recall the analogy of the funnel. Not everyone who was arrested ends up being incarcerated. As said previously, it is possible to be charged without being arrested. Not everyone who is arrested is necessarily guilty of a significant offense, yet at the same time, one can say that being arrested is a sign the person is not "staying out of trouble." It could be conjectured that if one uses arrest as a standard of recidivism, the price may one of spurious results.

If one is considering the use of arrest for a crime, it should be pointed out that because a person was arrested after being released from prison, it does not necessarily mean a new crime had been committed by that person. The arrest may be something related to the original crime the prisoner had been incarcerated, and released from custody for, such as a parole violation. It may also be related to an old crime committed before the crime that lead to the original jail sentence. Also, being arrested did not mean the person did anything wrong, and it does not necessarily mean they were re-incarcerated or even charged with a new crime. The individual may have been arrested and found to have no connection to the crime in question or there may have been insufficient evidence.

In the case of using conviction, this would certainly appear to be the best option. To be convicted of a crime is quite a bit further down the funnel than arrest, and is a reflection of a legal process where there was enough in the way of evidence and seriousness of offense to indicate the person had not left a life of crime. This would fulfill the definition of recidivism that most people would inherently understand—a criminal who has re-offended. The objections to using this as a standard of recidivism might start with the idea that a study

of recidivism is most useful if it deals with the very worst individuals and crimes. If the standard of conviction is used, it can lead to the question of conviction of what offense? The conviction may be for a lesser offense that is not considered as serious as other offenses, and it may happen that a researcher happens to be more concerned about recidivists committing more serious offenses.

Not all convictions lead to an incarceration, even for a repeat offender. Again, we are lead to the bottom of the funnel: incarceration. As discussed previously, incarceration presents the largest burden on society by an offender and is also a proxy indicator to the seriousness of the offense (Maltz, 1984, p. 62). Incarceration is also likely to lead to the most easily accessible and accurate data since much of it is public. Whatever legal event is used to define a recidivism is best thought of as dependent on the purpose of the researchers and the availability of data. There is no best definition of recidivism based on one of these three legal events. Indeed, of the 99 recidivism studies mentioned previously, 30 used a composite of one sort or another to indicate whether a recidivism had occurred (The Sentencing Project, 2010). This can be taken as a sign that the situation at hand can indicate the best definition of recidivism to use. Factors such as the purpose of the researcher, the recidivism issue to be examined, the effect of a program or intervention, and the data that is available.

In their book, *Predicting Recidivism Using Survival Models*, Schmidt and Witte (1988) define recidivism obliquely: "We seek to explain and predict the length of time that elapses between an individual's release from incarceration and his or her return to prison, using data

on two cohorts of releases from North Carolina's state prisons" (p. 3). This has a deceptive clarity that leaves important aspects without discussion. It leaves the impression that recidivism was defined as being released from prison, being convicted of another crime, and subsequently returned to prison as punishment for that crime, with the "survival time" being the length of time between release from prison and being sent back as punishment for the next conviction.

One issue one can have with their recidivism definition is that it would appear that if an offender were on parole, violated some term of that parole, and was sent back to prison, it would be considered recidivism even though no new offense had occurred. As mentioned previously, people may disagree about whether or not this constitutes recidivism. The standard used is simply "return to prison."

Also, suppose a person commits a first offense, was charged and convicted, but there was no prison time as part of the punishment. Perhaps there was some served in jail, but not in prison. Based on the definition used by Schmidt and Witte (1988), if that person commits another crime after being convicted of the first crime, it would not appear to be a case of recidivism since the person had never been released from prison as a result of the first offense. They had never been sent to prison in the first place.

We can think of this the other way around. Suppose an offender was released from prison, committed a new offense, was charged and convicted of this offense, but did not get a prison sentence for the second offense. There may even have been some time served in

jail, but not prison. Based on the definition used by Schmidt and Witte, this would also not appear to be a case of recidivism since there was no "return to prison."

Finally, in reference to the study by Schmidt and Witte, notice that the data only involves North Carolina state prisons. The initial incarcerations were to North Carolina prisons and the re-incarcerations were to North Carolina prisons. Many of those initial incarcerations may have themselves been an incidence of recidivism in a broader sense because the offenders were incarcerated in, and released from, another prison system. Some of those released from a North Carolina prison, who did not recidivate as far as North Carolina was concerned, may have actually recidivated into another prison system. The data only showed an event of recidivism if the person was released from a North Carolina prison and was re-incarcerated into a North Carolina prison.

So, the issue of recidivism is release from prison in North Carolina to being re-incarcerated in North Carolina. A person may have been in prison in Minnesota, released, gone to North Carolina, committed a crime there and ultimately re-incarcerated. This would not be recidivism because the person was not in the data as being released from a North Carolina prison. The reverse may have happened: the person may have been released from a North Carolina prison, come to Minnesota, committed a new crime and been re-incarcerated in Minnesota. As before, this would not have been considered recidivism because it would not have been in the data. Their survival analysis uses the time from release from prison in North Carolina to being re-incarcerated in North Carolina.

This issue faced by Schmidt and Witte (1988) in being limited by their data to North Carolina is an example of a more general problem that will be faced by most attempts to study recidivism. Given that one must define recidivism by certain legal events such as arrest, conviction, and incarceration, jurisdictions that keep records of such legal events to define recidivism will tend to keep records for their own jurisdiction and not others, so if the legal event happens in another jurisdiction, it may not be noted as an incidence of recidivism.

Schmidt and Witte (1988) were certainly aware of these issues, but were limited by the data available. Schmidt and Witte (1988) wrote "We analyzed timing of return to prison in North Carolina because this was the only definition of recidivism that our data would support" (p. 9). It is easy to point out such problems and weaknesses with the definitions of recidivism used by various analysts. It is important to remember that constraints on how recidivism is defined are provided by the data that is obtainable. Recidivism definitions can be created that are solid, consistent, and free from contradictions and obvious weaknesses, such as described here, but the problem will be finding data that can sustain that definition.

A study by Duwe and Kerschner (2007) evaluated the effect Minnesota's "Challenge Incarceration Program" (CIP) on recidivism. The CIP program details are not relevant here, but what is of interest is the different ways in recidivism was defined for use in survival analysis. It used four different standards to define recidivism: re-arrest, felony re-conviction, re-incarceration for new offense, and re-incarceration for any reason. The difference between the last two categories is that the last category combines re-incarceration for new

offenses with re-incarceration for “supervised release violations” (p. 623). This would be an example of using a mixed definition of recidivism that was seen on the database of 99 recidivism studies done by The Sentencing Project referred to earlier.

Maltz (1984) had an entire chapter on the intricacies of possible recidivism definitions (pp. 54-67). He proposed eight possible recidivism definitions which could be triggered by “reactivating event[s]” (pp. 65-66). Maltz had survival analysis in mind when speculating on recidivism definitions (pp. 68, 115), and the “reactivating event[s]” are all legal events that would leave a record in data somewhere.

In the study referenced earlier by the Bureau of Justice Statistics, the primary definition used for recidivism was being arrested for a new offense (Durose et al., 2014, p. 1). Although, this work has an annoying habit of referring to this simply as “arrest” at times instead of “arrest for a new offense.” This study also used five other definitions to give a total of six definitions of recidivism. The second definition was “adjudication or disposition” which was also referred to simply as “adjudication” where if an arrest lead to “…a subsequent court adjudication or disposition (e.g., convictions, dismissals, acquittals, or deferred adjudications).” It is important to note that the event of adjudication defined the event of recidivism, but the date used was the arrest date that lead to the adjudication, not the date of the adjudication. The third event is conviction for a new offense. As with their use of adjudication, conviction defined the event of recidivism, but the date used was the arrest date, not the date of conviction (p. 21). In the case of adjudication and conviction, it

would appear to be common sense that these events would mean a new offense had been committed, but the authors do not say so explicitly.

The fourth event used to define recidivism by Durose et al. (2014) is "incarceration" and the term appears to refer to a "…prison or jail sentence" where it is important to remember the distinction between jail and prison. Again, the date used for the recidivism is the date of arrest that lead to the incarceration. The fifth event was called "imprisonment" and appears to refer to receiving a prison sentence, not a jail sentence. Once again, the event, in this case, imprisonment, defines the recidivism, but the date used is the arrest date (p. 21). The difference in this study between incarceration and imprisonment is worth stating again. For this particular report, incarceration meant to be put in prison or jail, imprisonment meant to be put in prison. Interestingly, Durose et al. (2014) report that sometimes the data was not specific enough in whether jail or prison was used, and so in those cases "a sentence of a year or more was defined as imprisonment" (p. 21).

The last definition of recidivism used by Durose et al. (2014) is called "return to prison" (p. 21). They describe this category as follows:

> Classifies persons as a recidivist when an arrest resulted in a conviction with a disposition of a prison sentence or when the offender was returned to prison without a new conviction because of a technical violation of his or her release, such as failing a drug test or missing an appointment with a parole officer. (p. 15)

These last three categories used by Durose et al. (2014)—incarceration, imprisonment, and return to prison—are easy to confuse. Incarceration and imprisonment appear to imply that they are the result of a new offense, but the authors do not state this clearly, which is unfortunate. In view of the last category specifically including a return to

prison because of “technical violations,” it would be logical if incarceration and imprisonment would mean that only new offenses would result in these particular recidivism events. This last category of “return to prison” would appear to exclude confinement to jails (p. 15).

As far as the data used in the study, there are signs that getting proper data was an issue for this study. The categories of adjudication, conviction, incarceration, and imprisonment could use data from 29 of the 30 states, while the category of “return to prison” relied on data from 23 of the 30 states (Durose et al., 2014, p. 14).

Durose et al. (2014) never mentioned survival analysis, and their report is inferential in that a sample is used to determine something of the larger group of prisoners released in 2005. However, the definitions used certainly appear to be sound. Note again how arrest for new offense was the primary definition of recidivism, which would conform to the expectation of most people that released offenders should “stay out of trouble.” As discussed previously, various definitions of recidivism are desirable, and most likely will have to be defined by legal events that can be found in data records. In the case of this report by Durose et al. (2014) other categories are used, but with each of those categories, the date of arrest is used as the date of recidivism. In the case of the category of incarceration, the incarceration has to occur in order to be considered to be recidivism, but the date of arrest is used as the date of the recidivism (p. 21).

In this way, the lag that can occur in the use of legal events to define recidivism is avoided. The purest data that can be used in survival analysis of recidivism would be the

date on which an offense was committed. The problem is that this is information that is unknown with any precision. Legal events are much more likely to be in recorded in data for later use, and if the offense date is not known with precision or reliability, the arrest date would be the next one to use, but this would present problems that are discussed elsewhere in this review. If other legal events are used, especially the use of re-incarceration, there will be time lag since the legal process may take some time. In survival analysis, the aim is to explore the time to recidivate. If the offense date is not practical, and a legal event date is used that occurs well after the offense was committed and well after the person was arrested, the potential distortion in the survival analysis is obvious.

The study by Duwe and Kerschner (2007) also provides an illustration of the importance of the definition of recidivism and how complicated details can be in the study of recidivism. Among many results, in the comparison between CIP participants and a control group, lower recidivism was found for the CIP group in the first three categories of recidivism (re-arrest, felony re-conviction, and re-incarceration for new offense). But when the recidivism definition was re-incarceration for any reason, the recidivism rate was very similar because the two groups of offenders "…returned to prison…at virtually the same rate." That is, even though the recidivism for committing a new offense was lower, being returned to prison because of a "technical violation" was higher for the CIP group (p. 627).

This means differential results depending on the definition of recidivism. If re-incarceration for new offense had been the only standard used, the program would show

favorable results. If re-incarceration for any reason had been the only standard used, the program apparently made no difference.

If a crime is committed, there can be a considerable length of time between the commission of the crime and being convicted of that crime and being re-incarcerated. The arrest, investigation, legal process, and other steps can take a considerable length of time. It would certainly appear that defining recidivism as the time from release from an incarceration to re-incarceration would distort the length of time it would take to "re-offend," meaning the time to committing another crime. The time to re-offend would appear to be much shorter than the time to re-incarceration and that needs to be remembered when analyzing time to recidivism: it is not the time to re-offend that is being studied. It is the time to a legal event such as arrest, conviction, or re-incarceration. The reason is that these are known events with a record that will appear in data. Not so for committing an offense.

Statistical Issues in the Evaluation of Recidivism

Among their two cohorts, Schmidt and Witte (1988) did have some offenders who returned to prison within one to three months after release (pp.26-27). The authors did note this very early return to prison on the part of some and discussed an inevitable time lag between committing an offense and being re-incarcerated (p. 29). Schmidt and Witte did not mention an effect of technical violations, which we discussed earlier, but technical violations could certainly explain how a few individuals were returned to prison so quickly. Unfortunately, there appears to be no way to tell how these offenders got back to prison so

quickly. Again, in the case of technical violations, we are back to a consideration of recidivism without a new offense.

Because of the use of legal events to defined recidivism combined with the time lag involved for the legal process to run its course, there is a potential for distorting the survival analysis. This would be even more distortion introduced if two cases for similar crimes take markedly different times to go through the legal process. For example, an offender released from jail commits a criminal act 4 months after release. Ultimately, he is re-incarcerated 11 months after his release (and 7 months after the offense was committed). If we had defined recidivism as being re-incarcerated, his survival time would be 11 months.

This 11 month survival time may be a bit unsatisfying to a person who viewed the actual survival time as being more legitimately 4 months and not 11, because that is when the new offense was committed. The problem is that to use the 4 month survival time figure would mean that the offense date was known, something that is not always realistic, as discussed earlier. Also, someone else who committed a similar offense 4 months after release from incarceration may have details to the case that mean the legal process takes 5 months longer than for the first offender, so this individual would be incarcerated 16 months after release and 12 months after committing the offense, leaving a survival time of 16 months. So, we have similar cases with similar offenses committed 4 months after release, but with different survival times of 11 and 16 months.

This example, where recidivism was defined as re-incarceration, will be especially confusing to someone who mistakes recidivism with committing a new criminal offense.

Imagine a situation where analysts are attempting an evaluation of a rehabilitation program, it is natural to ask how much a distortion would be provided by these issues and how this could interfere with objectively identifying whether the program is effective or not. There could be two assumptions made to minimize this concern: it could be viewed as rare enough not to significantly affect results or it could be viewed as occurring more or less equally for control and treatment groups. One can even make the assumption that some combination of the two is at work. One could also use data from only one jurisdiction so the same entity will, it would be hoped, be behaving consistently with all offenders.

For another issue of time lag, it should be pointed out the law may apprehend criminals at very different times. As another example, suppose we use re-incarceration as the defining event of recidivism and have two criminals who have been released from incarceration. If these two may separately commit new crimes on the same date, but suppose one is "caught" very quickly and the other takes longer to solve. The ultimate re-incarceration for one person occurs in eight months after release while the other re-incarceration takes three years. This second person therefore had a longer time to recidivate according to the definition of recidivism that was used, and the two survival times are 8 and 24 months. Furthermore, if we have created a model of time to recidivism that uses explanatory variables to explain recidivism time, the time difference will be attributed to the explanatory variables rather than vagaries of crime circumstances and the legal process.

In the case of the work by Schmidt and Witte (1988), the total size of the two cohorts together used in their study is very large, some 19,136 individuals (p. 21). When the size of the data is large, one can make the assumption that problems such as described above are small in relation to the total. One can also make the assumption that such distortions can have some "cancelling out".

One example of a problem with measuring time to recidivate and a way to deal with this problem is provided by Duwe and Kerschner (2007). As mentioned previously, they used survival analysis to study recidivism and evaluate the effect of a program on recidivism. Recall that four standards for defining recidivism were made: re-arrest, felony re-conviction, re-incarceration for new offense, and re-incarceration for any reason. For the first three categories, it is possible that the person could violate a term of a supervised release and be held in custody without triggering the recidivism definition (apparently, this could happen without being considered an arrest, indicating the term "arrest" may have nuances).

If such an event happened, the offender was in custody yet the legal event that would end their survival time had not yet occurred, meaning survival time included time in custody. To address this, the time spent in prison during the resulting legal process was subtracted from survival time: "…deducting the amount of time spent in prison from the total at-risk period, or 'street time.'" Otherwise, in terms of survival analysis, the time spent in prison under these circumstances would count as time spent being free and able to choose a criminal lifestyle, and for the time being, choosing not to (p. 623).

There is a term that is encountered in literature on recidivism. It is "incapacitation" and refers to the fact that if a criminal is incarcerated, he is unable to commit crimes "on the street" and crimes are therefore prevented that would have otherwise occurred if that offender were not incarcerated. "Selective incapacitation" was an idea of great interest in the 1970s and 1980s. This was based on the observation that a small number of criminals committed an inordinate number of crimes, and that if these individuals were put into prison for long periods of time, a large number of crimes would be prevented. The "selective" portion of the term meant being able to determine who these people were, presumably through some statistical evaluation.

Thus, "selective incapacitation" meant somehow evaluating criminals for the potential to commit crimes and incarcerating some of them for longer periods of time if they were found to be a particularly high risk for committing crime, i.e., recidivism, than other criminals who had committed similar offenses yet found not to be as likely to commit further crimes. (Schmidt & Witte, 1988, pp. 1-2). The backdrop of this interest in selective incapacitation was the increasing cynicism of the public of the idea of rehabilitation of the incarcerated from approximately 1968 on (Auerhahn, 1999, pp. 704-705).

There were also models that described the commission of crime that can broadly be described as "criminal career" models. Very simply, these models assumed offenders are simply the way they are, a "state of nature." An offender starts a criminal career early, commonly during the teen-age years (age c) and commit crimes at a regular rate until he "retires" (age r). This rate is could be referenced as α, a mean rate of arrest. There is no

rehabilitation or reform, and the only function of prison is that it means that the incarcerated person cannot commit crimes during the time of imprisonment, at least not against the public. In this view, prison will have no impact on α or on r aside from incapacitation (Barnett, 1987, p. 21).

Other versions of this line of thought may think of the rate differently. It is very common that instead of measuring the rate of arrest (α), we can also think of a yearly rate of committing crimes (λ). The total number of crimes committed by an individual in his "criminal career" would therefore be λT, if we define T to be the total length of time of a criminal career. With this model, a criminal committing crimes can be thought of as a Poisson process (Shinnar & Shinnar, 1975, p. 586). If a criminal career is indeed a "state of nature" for a criminal, then recidivism would be the normal for such individuals. Interruptions in that criminal career are provided by incarceration, but recidivism represents the return of that individual to his state of nature once that incarceration is completed.

Schmidt and Witte (1988) discuss the use of statistical models to describe and model recidivism, and also to predict recidivism for random and non-random samples and for individuals. Schmidt and Witte found models that described the recidivism behavior of large numbers of released offenders fairly well, but were simply not good enough to make predictions of random subsamples, non-random subsamples, and individual predictions (pp. 156-157). In terms of individual prediction, with one cohort they had a false positive rate 47% and a false negative rate of 28%, a result they felt was encouraging. They obtained R^2 values of .10 for one cohort and .12 for the other, and found these to be solid results for

recidivism analysis. Schmidt and Witte also expressed optimistic hope that explanatory power could be doubled from what was achieved in their work, such a model would be able to explain 20% to 25% of the variation in a dependent variable describing recidivism (p. 14-15).

Schmidt and Witte (1988) make it clear that predicting recidivism for individuals can and should be attempted for purposes of study and model-testing, but it should not be used for actual decision-making in the real world, at least not based on their experience (pp. 5-6). Because it would necessitate differential treatment of offenders who may have committed similar crimes, especially for 1980s style "selective incapacitation," it is patently obvious that recidivism prediction on the individual level would have to excellent. Schmidt and Witte examined the state of prediction of recidivism on an individual level, and attempted to do it themselves, but found the results are simply not good enough, as described previously. This was as of 1988, but there is no indication that any prediction models today are accurate and reliable enough to do this.

There are echoes of this issue are around today with the on-going issue of the civil confinement of sex offenders after their prison sentence has been completed. Murderers are put into prison for long periods of time, frequently for the remainder of the murderer's life, so sex offender recidivism may be the most volatile recidivism topic of all. Nobody wants to be responsible for releasing a sex offender who recidivates.

The discussion of prediction in the area of recidivism is a bit reminiscent of the Foundation series written by Isaac Asimov, which he started in the 1950s. In this series, a

group of scientists perfect the mathematical discipline of “psychohistory,” which allows accurate prediction and manipulation of the historical course of humankind on a galactic scale over the course of hundreds of years.

Although Asimov never says what psychohistory is, beyond vague references to being highly advanced mathematics, to a reader it can seem clear that psychohistory is chiefly statistics. One idea that is stated many times in this series, is that the behavior of an individual is unpredictable, but group behavior can be understood and predicted (Asimov, 1982, p. iv). The idea of prediction in the area of recidivism, and especially with the selective incapacitation, is reminiscent of the ideas in this fictional work.

With issues of crime, punishment, and rehabilitation, the data and the contexts from which that data originate are complex and varied. There will problems with data. Data of almost any kind, especially in the social sciences, will most likely need considerable “cleaning.” Data that does not need cleaning is likely data that was well-thought out before its collection: in effect, pre-cleaning. However, that is much more likely to happen in controlled scientific experiments than with data involving the social sciences. Recall that the origins of survival analysis were with the evaluation of the durability of electrical components.

There is an issue that must be dealt with when one is applying statistical modeling to observational data: to what extent are fundamental assumptions being fulfilled? Can one regard fundamental assumptions as ideals that can never be completely achieved, but can approached asymptotically. Is it possible to regard a statistical technique as being valid if

some but not all assumptions hold, or if there is theoretical numerical scale where it can be said that the model is OK to use if the fundamental assumptions are 90% fulfilled?

The statistician David Freedman (2010) has pointed out that we have gone astray on some issues of statistical analysis, especially in observational studies and in the social sciences. He was an advocate of so-called "shoe leather methods" involving sound "subject-matter knowledge" (pp. xiii-xiv, 53). Statistical techniques cannot make up for bad data, incorrect statistical approaches, or other shortcomings (pp. 48-62).

The study done by Duwe and Kerschner (2007) of the Minnesota CIP program had some complexity in the definition of recidivism that was used in the study. The study was retrospective and "quasiexperimental" in design where there were two groups: CIP ("boot camp") participants a "control group" of offenders. The control group was selected randomly from other prisoners released in the same time frame. The key is that the control group was selected through a "multistage sampling design" to ensure that the CIP group and the control group were very similar according to a series of explanatory variables and were release during the same period of time (pp. 621-626). The survival analysis of the data used Cox proportional-hazards model (pp. 625-626).

It can be conjectured that Freedman would have been critical of the use of the proportional-hazards model, which will be discussed in greater detail. Yet, the mixed results depending on how recidivism was defined is of great interest, as is the data.

The study serves as an example of the nuances of recidivism. Duwe and Kerschner (2007) conclude that when recidivism is defined as being a return to prison for any reason,

the CIP and control groups were similar. Yet, because the control groups were more likely to return to prison for the commission of a new offense, the CIP group had new stays in prison that were about 40 days less than the control group (p. 614). Hence, the recidivism was not as serious. Can this be considered a small success? This conclusion illustrates very nicely some of the complexities of evaluating recidivism:

> But as this study illustrates, determining whether a program works should not be limited to a simple question of "Did they recidivate or not?" Rather, in assessing whether a program is effective, perhaps the focus should be not only on whether they recidivated, but also on why they returned and for how long. Concentrating merely on whether offenders are rearrested, reconvicted, or reincarcerated following release is often the benchmark used in correctional program evaluations because it is, generally speaking, an easier or more feasible issue to address analytically. But results can vary significantly depending on how one measures recidivism. Moreover, even if multiple measures of recidivism are used, the issue of whether offenders recidivate does not tell the full story about whether a correctional program works. Instead, it is also critical to know why and how long offenders returned to prison because the answers to these two questions will provide a more complete picture as to whether a program is effective. (pp. 639-640)

To return to Schmidt and Witte (1988), recall that the goal of their work was find a statistical model of recidivism, which could be used for prediction, based on data from two cohorts of release prisoners from North Carolina state prisons (pp. v-vii). Their book is a search an accurate model that moves from simple to increasingly complex models, as they deal with various problems that arise from this effort.

Their work can be summarized this way: it begins with simple non-parametric and parametric models without explanatory variables, progresses to parametric "split population models" (which will be discussed in greater depth), then the proportional hazards model is explored (the first one Schmidt and Witte attempt with explanatory variables), then

parametric models with explanatory variables. They then use split parametric models with explanatory variables, and then branch off into two separate issues: the probability of recidivism and the timing of recidivism for those who do return. The model they have at the end is a "logit lognormal model" with explanatory variables where a logit model describes the probability of eventual recidivism while the lognormal model explains the timing of recidivism (p. 19).

One of the items the Schmidt and Witte (1988) felt was a major contribution to the analysis recidivism was the use of "split population models" (p. vii). They did not develop the idea. It is credited to Maltz and colleagues, most notably in a paper by Maltz and McCleary published in 1977. It is discussed in Maltz's often cited study on recidivism (Maltz, 1984, pp. 79-80). To explain what is meant by this term, recall that survival analysis had its origins as evaluating a "lifetime" until an event occurs. In its purest and earliest form, the application was with electric components and how long the component lasted until failure. This idea was extended, most famously, to the realm of medical treatments where people who were afflicted with a medical problem had their remaining life (since diagnosis or starting a treatment regime) as the "lifetime" of survival analysis with death as the final event (Dobson & Barnett, 2008, p. 187).

The underlying assumption is that the final event is inevitable, as in the ultimate failure of an electric component or death in the case of a living thing. This would mean that the cumulative distribution function would ultimately reach a value of one (Schmidt & Witte, 1988, p. 66). In the case of recidivism, this cannot be true, since not everyone will recidivate.

There is always going to be a sizable portion of prisoners released from incarceration who are not going to recidivate (p. 68-69). Hence, the cumulative distribution will not approach 1, but some proportion of 1 that represents the proportion of all who will recidivate. The rest will not. Because the cumulative distribution will not approach 1, Maltz quoted another author who said the function was "'degenerate at infinity'" (p. 80).

What proportion of people will recidivate? That will vary according the ages and kinds of offenders. Schmidt and Witte did provide examples of recidivism studies that rates that had significant range of values. The values went from 36% from a study that began with 10 year old boys (who were apparently at-risk) where recidivism was defined as conviction, to 63% from a study where the subjects were released from a federal prison and recidivism was defined as return to prison for "parole violation, a felony, or a felony-like offense" (p. 68). Schmidt and Witte found raw recidivism rates of .381 and .369 for the two cohorts (pp. 30-31).

It is also noted that recidivism studies find that the hazard rate declines very quickly after reaching a peak. As discussed earlier, Schmidt and Witte found peak hazard rates for their two cohorts at 15 and 6 months and declined rapidly, coming to "essentially zero" by end of the follow-up period in the case of one cohort (Schmidt & Witte, 1988, pp. 29-30).

How is this dealt with? The use of a "splitting parameter" which is intended to be the proportion of those who will ultimately recidivate. In the case of Schmidt and Witte (1988), they referred to this as δ and obtained maximum likelihood estimates of the parameter. For example, for split lognormal model yielded a maximum likelihood estimation of $\delta = .4477$

and .4929 for the two cohorts they examined (p. 72). When the splitting parameter was taken into account, only then did the hazard rate in the model fall as quickly the actual data (p. 69).

This same idea can be found in the idea of "cure models" in medical treatment applications of survival analysis. Recall that survival analysis, in its original form, often assumed that there was one inevitable end that will befall every subject. In the case of electrical components it was failure and in the case of humans it was death. In the case of people who are at risk for contracting a disease or who are undergoing treatment for a condition, in some situations there are going to be survivors—some are "cured" and are not going to die of the disease or the treatment. This idea is developed in Maller and Zhou's book, *Survival Analysis with Long-Term Survivors* (1996). Maller and Zhou (1996) refer to such people as "immunes," saying that "an 'immune' individual means one who is not subject to the event under study…" (p. xi).

In fact, Maller and Zhou (1996) state that a good practice when applying survival analysis is to check for the presence of "immunes." One very basic way is to look at a Kaplan-Meier Estimator graph of the subjects. If the graph "levels off" at the right, seemingly stuck at a proportion less than one, as opposed to having significant increments right up to end of the graph, this can be sign that not everyone will undergo an inevitable fate (pp. 16-17). This does seem reminiscent of the hazard curves of recidivism that drop quickly and stay at a low point after hitting the peak of the hazard curve.

Schmidt and Witte conclude their book by noting that they spent a great deal of effort on "…finding the 'best' distributional assumptions to make, by trying many different distributions and checking the adequacy of the resulting predictions over time. We devoted much less effort to investigating the proper ways to enter individual characteristics into our models…" The also noted explanatory characteristics were entered linearly into their final logit lognormal model (pp. 159-160). Given the fact that the various definitions of recidivism seem to make a great deal of difference, it may in fact matter a great deal how individual characteristics are incorporated into a view of recidivism.

There is larger view of the role of explanatory variables in the phenomenon of recidivism when using survival analysis is this. There are no citations to present for this observation or even if there is a name for what is being describe, but start by thinking of survival analysis used in its most basic applications in the World War II era to test military equipment (Singh & Mukhopadhyay, 2011, p. 145). Now think of survival analysis in its most well-known medical application, where treatment regimens for patients are compared (Singh & Mukhopadhyay, 2011, p. 145). By their nature, such studies occur in much "controlled" situations with fewer variables, although in going from machine units to human patients, the complexity should naturally increase.

Now think of a criminal offender being released from prison. When he walks out of prison, the number of forces, factors, and covariates acting on him must be numerous indeed. Everything from his age, prior criminal history, family situation, to the nature of his friends and associates will have a relevance. It is difficult to see how this list must dwarf the

other situations in which survival analysis was used. This is a human being interacting with a fluid, dynamic situation over a potentially long period of time. In these three applications, which in a sense, represent escalation in complexity, we can see survival analysis being assigned more and more responsibility and being loaded with more and more factors and covariates.

Discussion of the Proportional Hazards Model

Both Schmidt and Witte (1988) and Duwe and Kerschner (2007) used the proportional hazards model. Schmidt and Witte (1988) used the proportional hazards model as one part of the total analysis in their model building and evaluation, and the primary purpose was to investigate the explanatory variables (Schmidt & Witte, 1988, pp. 83-90). Duwe and Kerschner (2007), on the other hand, were evaluating a program to determine an effect on recidivism and used the proportional hazards method to evaluate the program (Duwe & Kerschner, 2007, p. 626).

Statistician David Freedman (2010), in an article that originally appeared in *The American Statistician*, in 2008, expressed skepticism regarding the proportional hazards model. His primary focus in discussing the proportional hazards model was in regard to epidemiological studies, both in randomized controlled studies and observational studies (Freedman, 2010, pp. 169-192). One complaint he had about the use of this model is the difficulty in fulfilling important underlying assumptions of the model (p. 170). In contrast, there is a perception among researchers that the model is actually flexible and undemanding, most likely because a specific distribution need not be specified. Schmidt and

Witte (1988) wrote that the proportional hazards model "…allows one to estimate the effects of individual characteristics on survival times without having to assume a particular form for the distribution function (or the density or hazard)" (p. 83).

Freedman (2010) points out two assumptions of the proportional hazards model as "stationarity of mortality and independence of competing risks" (p. 169). Another source describes an assumption this way: "[s]ince the Cox proportional hazards model relies on the hazards to be proportional, i.e. that the effect of a given covariate does not change over time" (Persson, 2002, p. 17). Freedman states it this way, using his own italics: "…it is also about *hazards that are proportional to the baseline hazard*" (p. 187).

There are various statistical tests of assumptions for the proportional hazards model, but it is extremely difficult to see how recidivism data could fulfill these assumptions needed with the proportional hazards model. Recall the origins of survival analysis and the most famous applications of it. Survival analysis had its beginning with the testing of electrical, electronic, or mechanical components to determine the durability or "lifetime" of such components. The most famous extension of this idea was to people in contexts of disease progression or disease treatment. For example, for a group of unfortunate people with a particular kind of cancer, what is the survival of such individuals with different kinds of treatment? In going from electrical components to people under different treatment regimens to people making individual decisions in daily life, it is important to note the vast difference of circumstance surrounding the subjects of study.

It is common for recidivism data to show a hazard rate that changes greatly over time, as mentioned previously. To be sure, there are aspects to a particular individual that will not change over time, such as race, ethnicity, gender, and childhood situations. Schmidt & Witte (1988) admit the limitations of their recidivism model-building and point out a lack of variables that can better capture the individual circumstance better (pp. 156-157).

Recall the assumption of the proportional hazards model that the effect of a covariate does not change with time. It is difficult to see how this could be true with covariates of recidivism. Given the change in hazard rate that is commonly seen, given that there appears to be the presence of "immunes" (those who will not recidivate), and given much greater likelihood of young people to recidivate, it is very difficult to see how this assumption can be fulfilled when it comes to recidivism—it should seem readily apparent that some covariates are changing with time. Schmidt and Witte (1988) found their models consistently, and by a good margin, under-predicted recidivism in younger offenders in the validation sample, even though age was an explanatory variable (p. 138). This has to be understood as an indication that covariates are changing with time.

Freedman (2010) makes objections to the use of Kaplan-Meier estimators and the Cox proportional-hazard in some epidemiology and social science applications. His concern was that underlying assumptions required for the use of these techniques are difficult to fulfill "such as stationarity of mortality and independence of competing risks" (Freedman, 2010, p. 169). It would appear that recidivism data would have issues with these assumptions also, especially with the idea of the independence of competing risks as far as

recidivism is concerned. A Kaplan-Meier graph can be useful for descriptive purposes rather than inferential and are to be used for this purpose in this thesis.

Chapter 3: Overview of Survival Analysis

An understanding of survival analysis can be thought of as the study of a "lifetime" of an organism, component, or process until some concluding event occurs that ends that lifetime. The term "lifetime" can, but does not necessarily, mean the literal life of an organism, whether plant, animal, or human. Very generally, it refers to the length of time some condition or situation exists before being ended by the concluding event, which can be termed, "failure" (Dobson & Barnett, 2008, p. 187).

This is a very general explanation and is meant to apply to many different situations. A simple example is the running of electric motors until they fail. Setting the motors running was the beginning of the "lifetime" while the failure of each motor marked the end of that motor's lifetime (Dobson & Barnett, 2008, p. 187).

These motors could be the same kind of motor, in which case the differences in the lifetime might indicate hitherto unknown differences between them, such as in the materials or methods of its construction. These might be different kinds of motors, in which case differences may be thought to exist, but until testing such differences would be unverified and unquantified. Or the motors might be operated under different conditions, such as temperature, to find out the performance under differing circumstances.

The comparison of the motors' lifetimes could indicate which motor was better in terms of its durability. The duration of these lifetimes, in addition to any covariates or factors that may have affected these lifetimes, are analyzed statistically. The point of this is

to discover items or relationships of interest that affect the lifetimes of the units in question, as well as offering opportunities for prediction and risk assessment (Lee & Wang, 2003, p. 1).

From another realm of application, a group of cancer patients may be given a treatment of one sort or another, while another group of similar patients is not provided with the treatment. The lifetimes, which could be the individuals literal life or perhaps the period of being tumor-free, can be compared to see if the treatment was effective (Singh & Mukhopadhyay, 2011, p. 145). Also, people can be followed from the development or diagnosis of a disease until death to learn more about the particular disease process. As in the case of engineering applications, covariates and factors can be evaluated and quantified. Risk assessment and prediction can be attempted based on such analysis (Lee & Wang, 2003, p. 1).

Although life tables, or some variant of life tables, have been around for a long time, engineering applications can be considered the impetus of the development of survival analysis as a statistical discipline, beginning in the World War II era as a means of testing and comparing military equipment (Singh & Mukhopadhyay, 2011, p. 145).

The examples given previously, from the realm of engineering and from the realm of medicine, are also applications of survival analysis that are perhaps the most well-known. Survival analysis can be extended to a variety of applications of widely varying situations and circumstances. Examples are time until earthquakes or time until stock market crashes (Singh & Mukhopadhyay, 2011, p. 145). The lengths of marriages can also be studied using survival analysis (Lee & Wang, 2003, p. 1).

Generally speaking, when there is a process or phenomenon which has a defined start and defined ending, it is conceivable to think of this as lifetime, the duration of which, in addition to any covariates or factors, can be subject to survival analysis of one sort or another (Dobson & Barnett, 2008, p. 187).

In using survival analysis to study recidivism, the idea is that an individual has committed a criminal offense in the past, and is put in a position where he has control over whether or not he will commit another offense of some sort. As an example, a prisoner is released from incarceration, and during this time after release, he is viewed as having the power to choose whether or not he will commit another offense. The lifetime, in this example, would begin with the release from incarceration, and the failure would be represented by the commission of a crime.

There is an underlying issue with the use of survival analysis that has become apparent, and for which there are no citations. In the attempt to analyze lifetimes of one sort or another in many different contexts, there is an increasing complexity in the variables, or for lack of a better word, an increasing complexity in the "forces" that are acting on the "unit" of study. For example, in going from analyzing mechanical components in engineering contexts to studying people in medical contexts to studying recidivism, an increasing complexity in the situation is apparent. There is an increasing burden on survival analysis when it is applied to situations of increasing complexity, of which recidivism is an excellent example. One study of recidivism had records which had 149 variables (US Department of Justice, Bureau of Justice Statistics, 1994).

Chapter 4: Application of Survival Analysis to Recidivism

In this case, the interest is in recidivism, which can be defined, very broadly, as whether a person with a criminal history will return to criminal activities. Maltz (1984) gave a more specific definition by stating that "[r]ecidivism, in a criminal justice context, can be defined as the reversion of an individual to criminal behavior after he or she has been convicted of a prior offense, sentenced, and (presumably) corrected" (p. 1).

Survival analysis as applied to the study of recidivism would address questions related to the time it would take for such a thing to occur, as well as the probability of recidivism occurring at all. Survival analysis can help understand the magnitude and dimensions of the problem of recidivism. Survival analysis can also attempt to find factors and covariates related to the process of recidivism, as well as attempt to determine items of interest regarding subgroups of those who are at risk of recidivism (Schmidt & Witte, 1988, pp. 85-88, 95-101). Finally, survival analysis can be used to evaluate the effectiveness of various programs or interventions intended to address the problem of recidivism (Maltz, 1984, pp. 1-5).

Schmidt and Witte (1988) report the earliest use of survival analysis to study recidivism, of which they were aware, was a study done in 1972 (p. 16). Maller and Zhou (1996) have an example of applying survival analysis to the study of recidivism from 1969 (p. 23).

This simple and general definition of recidivism given above can be elaborated upon, as stated in the previous section, by supposing there is an individual who has committed a

criminal offense in the past, and that this individual is put in a position where he has control over whether or not he will commit another offense. The interest, then, is looking at the large numbers of people in this situation.

For example, in reference to the phrase used earlier, “the time it takes for a person with a criminal history to return to criminal activities,” one can ask, “Starting when? And ending when?” There is a need for a starting point to mark the beginning of the time interval (or lifetime) in question. There is a need to know when someone, who has a criminal history, is available for recidivism. There is also a need for an ending point that will determine when recidivism has occurred.

Especially if one is to apply survival analysis, the definition of recidivism will require a clear starting point (Dobson & Barnett, 2008, p. 187). Such a starting point commonly has been the individual’s release from prison. The end of incarceration, the end of parole, or the end of probation would all suffice as a starting point (Maltz, 1984, p. 22). Any of these events would work well as a starting point because each is a well-defined event that is sure to be recorded in data that can be accessed later. Moreover, in a sense, each of these events represents the return of control over the individual from the corrections authority to the individual himself.

The degree of control over the individual is obviously much greater in the case of incarceration as compared to probation or parole, which would most likely mean that the release from incarceration would be more effective as a starting point. The individual in control of his own destiny is at the heart of the issue of recidivism. To most, the conduct of

such an individual when left to his own devices would indicate whether or not the person had been rehabilitated. Again, it should be remembered that such a person may not return to crime at all.

As described in the literature review section, an organization called The Sentencing Project has an on-line database of 99 studies of recidivism made between 1995 and 2009, covering all 50 states and the District of Columbia. The organization does not claim the list is comprehensive for the United States. In fact, the study by Duwe and Kerschner (2007) referred to later, is not listed, so there are undoubtedly more studies that can be found for that time period that are on this list (The Sentencing Project, 2010).

It is not claimed that all of these studies used survival analysis. However, it is of interest to note that the listing of the studies included a definition of recidivism that was used for that particular study (The Sentencing Project, 2010).

Applications of survival analysis to recidivism typically begin with the release of an incarcerated criminal from prison and the lifetime consists of how long that person "stays out of trouble." As stated previously, the release from prison is a very clear event to start with and will have an associated date that will be recorded somewhere.

Survival analysis also needs a well-defined event at the other end of the lifetime, the event that defines a "failure," an end to the lifetime in question. In the application of survival analysis to recidivism, the event to define a failure is more complicated. In the application of survival analysis to recidivism, what event would define the recidivism? As indicated in the literature review, the application of survival analysis to recidivism is plagued

by problems of definition and data, especially when compared to the more well-known applications of survival analysis given previously, in the engineering and medical fields. The failure of a mechanical component or the death of a person are clear events that define an end to a "lifetime." The recidivism of an individual may not be so clear to define.

To begin a discussion of the problems how to define recidivism, it is important to start with the colloquial phrase that captures the spirit of the popular view of recidivism: a person who does not "stay out of trouble" will be a recidivist, meaning a person who does not refraining from criminal activities. However, there is a great deal of ambiguity with that phrase. What does it mean to "stay out of trouble"? More specifically, what is needed is to define recidivism in a clear way that captures the idea of an individual "staying out of trouble" while being a definition that can be supported by actual data (Maltz, 1984, p. 54).

If a lifetime can be said to begin with an offender being released from prison, one can define that person's recidivism, the end of that lifetime, in a surprising number of ways. Commonly used definitions of recidivism include: being arrested, being charged for a new crime, being convicted for a new crime, being re-incarcerated for a new crime, being re-incarcerated for committing a new felony, or being re-incarcerated for any reason (The Sentencing Project, 2010).

Maltz (1984) discusses many of the problems related to defining recidivism, and devotes an entire chapter of his book to the intricacies of defining recidivism (pp. 54-67). Along with the larger problems of definition, such as distinctions between arrest and re-incarceration, there are many smaller issues that may pose problems. As examples of this,

he discusses juvenile offenders who have become adults—should a first offense committed as an adult be considered recidivism? If a person on parole absconds and all contact is lost—should that be considered a recidivism? (pp. 60-62). Maltz (1984) lists eight possible definitions of recidivism that can be used (pp. 64-66).

It is important to note that many crimes go unreported (Mosher et al., 2002, pp. 54-55). When examining the topic of crime and how it is dealt with by society, a very common analogy one encounters is that of a funnel. If one were able to look at all offenses committed and all the offenders who committed those offenses, for example, in one year, the number of people eventually going to prison for the commission of some of those crimes would be much smaller than the real number of offenders. Even when one looks at reported crimes, "from police statistics to imprisonment data" the number of offenders dwindles as the legal process advances through its stages (pp. 59-60).

This issue is intertwined with the statistical definition of recidivism in that the major definitions that have been mentioned thus far—arrest, charge, conviction, and incarceration—are outcomes that descend to the bottom of this funnel. Incarceration, the ultimate outcome of "getting into trouble," will be the result that is encountered the least, given that a crime occurred. Therefore, there would be much "trouble" to society that is missed by using incarceration as the defining event of a recidivism.

As a further example of the complexities of defining recidivism, recall three of the definitions of recidivism that were given previously that could appear to be similar. These three definitions were: being re-incarcerated for a new crime, being re-incarcerated for a

new felony, and being re-incarcerated for any reason. The first definition is a very straight-forward, and would be intuitively understandable. But it is possible that an analyst may have an interest is in "major" crimes, or felonies, and therefore have a greater interest in re-incarceration for felonies as opposed to re-incarceration for misdemeanors or gross misdemeanors.

The third definition is of great importance. Someone who is released from prison on parole can be put back into prison if that person violates a term of parole—this could easily mean going back to prison without committing a new criminal offense. The important distinction has to be made between recidivists returning to prison because new crimes were committed as compared to recidivists returning to prison without new crimes being committed. It is common for people to assume recidivism means the individuals in question are committing new crimes, but that is not always the case. If one wants to equate recidivism with the commission of new crimes, awareness of this distinction is essential.

Earlier it had been stated that an analysis of recidivism is plagued by problems of definition and data. The data issue arises primarily because recidivism data, by its nature, will generally be retrospective observational data. With some exceptions, as in the case of the study by Duwe and Kerschner (2007), which will be discussed later, it is not resulting from an environment where there is some control over confounding variables and covariates.

The important issue in regards to the use of survival analysis of recidivism is that it will necessitate data that records a start of a lifetime. Equally important, it will necessitate

data which will record the defining event of a recidivism, the end of that lifetime. The differing definitions of recidivism elaborated by Maltz (1984) in his chapter of recidivism definitions (pp. 54-67), can be summarized by saying this end of the lifetime, the event that defines the recidivism, will generally mean a legal event that is recorded in data that is accessible later. This would mean legal events such as an arrest, charge, conviction, or incarceration (Maltz, 1984, pp. 63-66).

There is a further subtlety here. In most survival analysis, the defining event, the "failure," is a clear and unambiguous event, such as a mechanical component that no longer functions or a human being who dies. With recidivism, we might think the failure represents a return to crime, but actually, we most likely are unable to tell if an individual has returned to crime. The act of committing crimes is generally an unobservable event and criminal self-reports would not be reliable (Schmidt & Witte, 1988, p. 9).

The proxy for the event of returning to crime is therefore the legal events mentioned previously—such as arrest, charging, conviction, or incarceration. Unless one of these events occurs and is recorded in data that can be studied later, analysts are helpless as to tell whether or not someone has returned to crime. Unfortunately, if someone has returned to crime and is committing criminal acts, unless he is caught and one of the legal events occurs, they would be considered a non-recidivist as far as the data is concerned.

Thus, not only is the failure event that defines a recidivism subject to definition differences, but whatever failure event is decided upon will inevitably be a proxy for the actual event of interest to most investigators—whether or not the individual has returned to

crime (Maltz, 1984, p.55). The event of failure is often very clear to define and observe in the many applications of survival analysis, especially the most well-known applications. This is not so with recidivism.

There has to be an effort made to define the occurrence of recidivism, such as arrest, charge, conviction, re-incarceration, or some other event. Even then, there will often be a need to define recidivism further, based on that event occurring (Maltz, 1984, p. 65; The Sentencing Project, 2010). An example of this might be that an arrest will be counted as recidivism only if it leads to a new charge or conviction. Another example could be that re-incarceration will count as recidivism only if it is for a new felony offense.

To summarize the situation, survival analysis applied to recidivism will generally be interested in the return of an individual to the act of committing new crimes. The problem will be, as stated previously, this is a failure event that is unobservable. Therefore, proxies must be used in place of the failure event, namely, the legal events that can be recorded in data for examination later. As such, the only way the analysis can be thought of as addressing a return to criminality is under the assumption that legal process is reasonably good at apprehending those who have returned to the practice of committing crimes.

The other problem is that even if data is accurate and complete in the sense of lacking blank data values, it may be limited. The data of Schmidt and Witte (1988) covered the prisons of the entire state of North Carolina. In this application of survival analysis, the start of the lifetime was the release of an individual from North Carolina state prison. The failure event, the definition of recidivism that was used, was that person being re-

incarcerated into a North Carolina prison. It is very conceivable that someone who was released from a North Carolina prison may have "gotten into trouble" and been put in a local jail rather than a North Carolina prison, or may have be incarcerated in a neighboring state rather than North Carolina prison.

In fact, it is also very conceivable that a person may have been in prison in another state, come to North Carolina, and ended up in prison there. From the perspective of the North Carolina data, this was that individual's first time in prison, when in reality, that was in itself a case of recidivism.

This will be an inevitable problem when examining data obtained from one jurisdiction. Multiple arrests in separate jurisdictions may not be seen as recidivism if the jurisdiction data is viewed separately.

When examining the Stearns County booking data, there is something very different about this data as opposed to the usual recidivism data seen with other analyses of recidivism. For example, with the work of Schmidt and Witte (1988) along with Duwe and Kerschner (2007), they regarded recidivism as a one-event process. For example, with Schmidt and Witte (1988), release from incarceration started the lifetime, and being re-incarcerated was the failure event, the end of that lifetime. That one failure event marks the end of that individual in the data.

Also, the re-incarceration does not take into account how long the person was in custody, let alone when the offense was committed or when an arrest took place. The individual could have been released from incarceration (which started the lifetime to be

examined), committed a crime the next day, and been arrested a week after that. Then, suppose after the legal proceedings conclude a year later, resulting in re-incarceration in a North Carolina prison. The survival time in the data, which would represent how long before the person recidivated, would be one year (Schmidt & Witte, 1988). The authors were certainly aware of this potential problem with the data. Schmidt and Witte (1988) wrote that they "analyzed the timing of return to prison in North Carolina because this was the only definition of recidivism that our data would support" (Schmidt & Witte, 1988, p. 9).

The booking data used here is very different. Many of the individuals show in the data are booked many times. Almost all bookings have an accompanying time of "release," but release does not mean the person was released to freedom. It could also mean the person was transferred to a prison and turned over to another authority.

The booking data from the Stearns County Jail does not have explanatory variables other than gender. Other examples of the application of survival analysis did have such explanatory variables, yet it could be disappointing in how little of the recidivism could be explained with the variables at hand. Schmidt and Witte (1988) discuss the problem directly both in regard to their particular study, based on data from the North Carolina prison system, and the more general problem of studying recidivism. They describe how their analysis of data obtained R^2 values of .10 to .12, meaning 10-12% of variation in the recidivism was described by their models. In terms of the prediction of recidivism, their models had a false positive rate of 47% with a false negative rate 28%. They actually

considered these results to be quite good compared to other such work in recidivism data (p. 15).

In the previous chapter, observations were made on how recidivism can represent a phenomenon that can be loaded with variables in an attempt to adequately describe an extremely complicated process. This data represents minimal information for each record. The data will have six variables, two of which are not very useful. The only variable that can function as an explanatory variable is that of gender.

Chapter 5: Background of the Stearns County Booking Data

It is best to start with what it means to be "booked." It happens primarily to people when they are arrested. Anywhere in Stearns County, when a person is arrested they are brought to the Stearns County Jail for booking, which means that mugshots are taken, fingerprints and demographic information are collected, the arresting charges are recorded, and the individual's identity is established. The arrest and booking is an entry point into the criminal justice system and into establishing a criminal justice history (Joe Kustritz, personal communication, 05/29/2015).

Booking does not require an arrest. It can also occur when a person is summoned to court, as can happen as a result of an official complaint. Such a court appearance may result in the person being taken to the jail for booking. In fact, Monday afternoons have some of the heaviest traffic at the jail in terms of booking because it is common for courts to schedule such business on Mondays (Joe Kustritz, personal communication, 03/2015).

It is also important to be clear about the function of the Stearns County Jail. It is not a prison in the sense that it is a destination for criminals who have been convicted of major crimes and are sentenced to serve a lengthy period of incarceration. Specific to Minnesota, a prison can be defined as a place of incarceration for persons convicted and sentenced to the custody of the Minnesota Commissioner of Corrections for a term length minimum greater than one year (Joe Kustritz, personal communication, 5/29/2015).

The Stearns County Jail has at least three main purposes: 1) A temporary holding facility for those who are arrested in Stearns County for the purpose of booking, and to

"provide a safe system of release with a court date to be set sometime in the future"; 2) As a pretrial holding facility for those arrested in Stearns County who have been held for court procedures, had bail set, and are unable to post the bail; 3) As a place of incarceration upon conviction for those sentenced by the court to a term of one year or less as a sanction for their crime. Other functions are for the jail to serve as a temporary holding facility for those who have violated a term of parole, or for those who are to be held as a result of a warrant from another jurisdiction (Joe Kustritz, personal communication, 5/29/2015).

The Stearns County Jail is the only jail facility in Stearns County, and when an arrest is made within Stearns County by any law enforcement representative, that person will eventually be booked into this jail (Joe Kustritz, personal communication, 02/2015). The same is true if a local court appearance results in a booking (Joe Kustritz, personal communication, 03/2015). The court building is in close proximity to the jail. Therefore, the booking data reflects an arrest somewhere in Stearns County, or a court appearance that results in a decision to book the individual into the county jail.

An additional detail of importance is that the jail may have individuals who are termed "weekenders," meaning the person is serving a sentence, but is allowed to serve the sentence over weekends, or some parts of weekends (Joe Kustritz, personal communication, 5/29/2015). This means they will check into the jail on Friday afternoons, but this is not a booking event that is included in the data here. The process of checking a weekender into jail on Friday afternoons does not generate a new booking number. The initial booking into

jail, resulting from an arrest or a court appearance, is the only appearance of that individual in the data used here (Personal communication, 2015).

In addition, the major city in the area is St. Cloud, the city limits of which are primarily in Stearns County, but which extend into neighboring counties, Benton and Sherburne. Not only is it the major city in Minnesota, but St. Cloud has a large student population from St. Cloud State University, with additional students who attend St. Cloud Technical and Community College. Because of the proximity of other counties in the city of St. Cloud, as well as the presence of a young and mobile student population, there is undoubtedly many cases of multiple arrests in the different jurisdictions, although there is no data available to quantify this statement.

If someone is arrested in one of the other counties, the arrested person would be brought to that county's jail for booking and it would not be reflected in the Stearns County Jail booking data. It is very conceivable that someone could have multiple arrests in all three counties, but if looking at each county's booking data in isolation, there may be no recidivism to be seen in the data, if one uses booking as a basis for the definition of recidivism.

As discussed previously, this is an inevitable problem that will occur when examining corrections data from a particular jurisdiction. It can be quite difficult to look at even jurisdiction's correction data, obtaining the corrections data of multiple jurisdictions would be even more difficult. Even if one were to obtain the data of multiple jurisdictions, it would be difficult to merge the data and see that one individual has been arrested in the different

jurisdictions. One would need a way to identify a unique individual, thus allowing a way to link records from different data sources. Such personal identifying information may not be present, and the name alone would not be sufficient to uniquely identify an individual.

This data represents bookings at the Stearns County Jail covering the period January 1, 2003, to January 31, 2015. At the end of the previous chapter, it was stated that this booking data was different than the data that is usually analyzed for the purpose of studying recidivism because of the lack of explanatory variables. Only six variables were provided, and two of those are not useful. Only the variable for gender is useable as an explanatory variable.

This data is also different in two other respects. To understand the first of these differences, start by noting that the most common sort of recidivism data will be of the kind used by both Schmidt and Witte (1988) or Duwe and Kerschner (2007). A lifetime begins with a release of an individual from incarceration and can end with an arrest, re-incarceration, or whatever definition is used for recidivism. If such a failure event occurs, that is the end of that person for the purposes of data analysis. There is only one failure.

This is an analog of the more usual kinds of survival analysis where machine units fail and people die. It is a final event, and represents not only the end of the lifetime in question, but the end of any contribution to the data of that particular study. In reality, however, a person can recidivate many times. All they have to do is be incarcerated and then be released from incarceration. As long as such a person has incarcerations that are limited, upon release, the opportunity to recidivate is there. This booking data represents

that situation. This data shows one individual who was booked 72 times and another who was booked 63 times—these are the two individuals with the highest number of bookings that can be determined from the data. One person can be responsible for multiple records in this data. It can be seen that some individuals are booked repeatedly.

The other difference with the usual recidivism data that is seen is regarding the start of the "lifetimes" in the data. Most recidivism data has clearly defined starts to a lifetime, such as the release from an incarceration. With this data, the beginning of a lifetime is extremely ambiguous. There is information about a "release" date and time for the booking. However, the term "release" with this data does not necessarily mean the person was released from custody and went free. It can mean exactly that, but it could also easily mean the person was transferred to another authority or transferred to a prison. This is known because is another Excel spreadsheet from Stearns County that listed bookings had information regarding the reason for a release. This other data will be discussed in more detail later.

The problem is that when the data indicates that an individual was released and provides a date and time for release, that will not necessarily mean the person was released to freedom. In other words, there is no way to know if the person went from being in custody to a state of control over his life within which he could make choices that could result in an event that would define a recidivism. The release could simply mean that the jail transferred custody of the individual to another authority, or to a prison.

The most obvious thing to do with this data would be to base a “lifetime” on a booking incident using the release time. Suppose someone was booked twice. A lifetime could be defined as beginning when the person was released from the first booking, and the lifetime could be defined as ending with the booking time of the second booking. The problem with this approach is that there is no way to know for sure if the “release” represented the person being returned to his real life or being transferred to another authority for custody.

These issues will be discussed again at a later time. But to summarize, there are three differences between this booking data from Stearns County and the usual kind of recidivism data that is seen in most recidivism studies. First, there is only one variable that can be used as an explanatory variable, and that is gender. Second, one individual can account for multiple records, and therefore, can recidivate more than once. Third, for the purposes of survival analysis, the data is unable to supply a clear and understandable start to a lifetime during which a recidivism may or may not occur.

Chapter 6: The Stearns County Jail Booking Data and the Use of Excel

The data used here begins with two Excel spreadsheets provided by Stearns County that lists bookings at the Stearns County Jail. The first of these spreadsheets lists bookings at the Stearns County Jail covering the period, January, 1, 2003, to October 30, 2014. This spreadsheet had 114,215 records. The second and smaller of the two Excel spreadsheets lists bookings at the Stearns County Jail over the period October 31, 2014, to January 31, 2015. This second spreadsheet had 2,139 records. The variables, or Excel columns, were the same and were presented in the same order. The two Excel spreadsheets were easily combined into one Excel spreadsheet of 116,354 records.

When referring to data from the Excel spreadsheets, the term "variable" will be used when referring to an Excel column. There were six variables, and are presented here in the same order presented in the data: Date_Confined, Booking_#, Date_Released, Sex, Offense, and Statute_Description. The Booking_# variable name was changed to Booking_Number since SAS would not accept the # figure in a variable name. Booking_Number was a ten-figure alpha-numeric "number" uniquely assigned to each booking. The construction of this number will be described in greater depth later.

The Date_Confined variable had both the date and time of the booking, therefore containing two pieces of information. Likewise, the Date_Released variable also contained two pieces of information, in this case, the date and time the person was discharged from the jail. The Sex variable assigned M or F to each record, recording the gender of each person booked. The Offense variable described the Minnesota Offense Code of the charge

under which the person was booked. The Minnesota Offense Code (MOC) is a five-figure alpha-numeric "number" that refers to specific offenses in state code. The variable called "Statute_Description" is a text variable that is intended to express the MOC in words. The Statute_Description varied considerably in length and construction. The first seven records are shown here as an example:

Date_Confined	Booking_#	Sex	Date_Released	Offense	Statute_Description
01/01/2003 01:40:00	03J0003-001	M	01/01/2003 10:04:00	X6700	5TH DEGREE DOMESTIC ASSAULT
01/01/2003 01:45:00	03J0002-001	M	01/04/2003 14:22:00	J2R01	TEST REFUSAL
01/01/2003 01:45:00	03J0002-001	M	01/04/2003 14:22:00	J3900	OPEN BOTTLE
01/01/2003 01:45:00	03J0002-001	M	01/04/2003 14:22:00	M6501	DRUG PARAPHERNALIA-POSSESSION
01/01/2003 01:45:00	03J0002-001	M	01/04/2003 14:22:00	J2901	GM DAC - CANCEL IPS
01/01/2003 01:45:00	03J0002-001	M	01/04/2003 14:22:00	J2501	2ND DEGREE DWI
01/01/2003 02:00:00	03J0007-001	M	01/01/2003 14:47:00	J3900	DAS

Figure 1: The First Seven Records in the Excel Spreadsheet

Of these 116,354 records, there were no blank records for Date_Confined, Booking_Number, and Sex. For the other three variables, there were 469 blank records for Date_Released, 378 blank records for the Offense variable, and three blank records for the variable Statute_Description. For Date_Released, the blank records do not appear to solely be the result of error or oversight. No blank records appear until 2012, and then there are only two for that year. All the rest of the blank records occur for the years 2013, 2014, and for January of 2015. This is a sign that the release date record may be blank, at least in some cases, because that individual has not been released yet, especially for bookings that

occurred in January, 2015. For the Offense variable, the blank data values are not as explainable. The blanks appear for each of the years on the record. The situation of the data in regard to blank records will be discussed again later.

The phrase, "booking number" will be used to refer generic concept of a booking number in this data, while "Booking_Number" will refer specifically to the variable name. It is the most fundamental variable in the data. It may be considered with or without the hyphen, but it appears the hyphen is stored as part of the booking number. Some occurrences of the booking number in other Excel spreadsheets have the booking number without the hyphen, however. A booking number is created for each person the first time he or she is booked into the Stearns County jail, and is unique to Stearns County.

The first seven figures of the booking number are assigned to the individual and are unique to that individual. It is assigned to a person the very first time he or she is booked. The last three figures tells us the number of bookings the person has been through. The very first time a certain individual was booked, for example, the booking number assigned was 03J0003-001, which is an actual booking number from the Excel spreadsheet excerpt shown previously. If this person is ever booked again, that booking number will be 03J0003-002. The first seven figures are the same, identifying the individual, the last three figures identify the number of times the person has been booked and will increment each time a person is brought to the Stearns County Jail to be booked. Because of this, an individual can be tracked by the booking number, and the number of times he is booked.

Of those first seven figures in the booking number, the first two digits refer to the year in which the first Stearns County booking was done. There are many booking numbers in the Excel spreadsheets that begin with "99" which should mean that the first booking occurred during the year 1999. There are no booking numbers that begin with "98" or "97" or numbers that would indicate bookings that began in years prior to 1999. This is a good indication that the booking number design and assignment process began in 1999.

The third figure in the booking number is always "J". Ostensibly, it stands for "jail", meaning the booking was done at the jail. In Stearns County, all bookings are done at the jail, and therefore all the booking numbers have a "J" as the third figure. The numbers are assigned by software, and it is likely that the software is meant to be used in a variety of jurisdictions. In other jurisdictions, the third figure would probably have had a variety of options to indicate exactly where someone was booked if that jurisdiction's facilities for bookings are more complicated than Stearns County.

The next four figures in the booking number indicate the order of booking for that year. The very first booking of the year at the jail would have "0001" for this sequence. This is not a perfect relationship between time and this sequence. For example, the very first booking of 2004 was 04J0001-001 but was assigned at 11:30 PM of December 31, 2003.

As an example, consider the booking number, 04J3018-004. This number alone has a great deal of information. The first booking for this person occurred in 2004, which the first two figures indicate. The "J" is not particularly revealing in the case of Stearns County, since all booking numbers have "J" in this position. The sequence "3018" would convey the

information that this was the 3018th booking for that year, which in this case is 2004. That portion of the booking number up to the hyphen, 04J3018, is unique to that individual in Stearns County. The extension "-004" indicates that it was the fourth booking of this person into the Stearns County Jail. This particular booking number was assigned on December 15, 2007, at 11:45 PM.

It needs to be emphasized that the first two figures in the booking number reveal the year in which the individual was first booked. That will follow that person through the ensuing years of interactions with the Stearns County Jail. In this case, the complete booking number 04J3018-004 was generated in 2007, during the person's fourth booking. If this individual is booked a fifth time, whenever that occurs, that booking number will be 04J3018-005. That number for a fifth booking for this person does not appear in this data, however.

It should be repeated again that when the word, "release," is used, it does not mean that the individual was released from the jail's custody and was therefore "free." It simply means that an individual is no longer in the jail's custody. The person may have been transferred to prison, or another county's law enforcement authorities. As an example, another Excel spreadsheet that listed booking information included a text variable for reason for release. Among the reasons were "SENTENCE SERVED", "PERSONAL/PARENTAL RECOGNIZANCE", "BAIL SUPPLIED", "RELEASED TO BENTON CO", "BOOK AND RELEASE", "RELEASED TO ANOTHER FACILITY", "RELEASED TO ANOTHER AUTHORITY", and "RETURNED

TO MCF-RUSH CITY". Unfortunately, there was no simple or reliable way to incorporate the data from this particular spreadsheet, and is therefore incomplete in this important detail.

The list of reasons for release that was given previously are some examples, but many more reasons for release are provided in the particular spreadsheet which included that data. Note that for the first three reasons given as examples, the releases indicate the person was "let go" and was free to leave custody. The remaining examples are cases where the person is "released" from the Stearns County Jail, but is simply transferred to another authority, and the person was not simply "let go."

As discussed previously this is an important detail and goes to the heart of how this data can be evaluated. Most studies of recidivism, whether using survival analysis or not, have a starting event that involves the individual being freed from custody. In most cases, this involves the release of the individual from incarceration, or some form of release from the authority of the criminal justice system, as in the case of probation or parole ending. The point of this is to see what happens to the individual when left to his or her own devices. After all, most would not consider a criminal rehabilitated until that individual uses their own "free will" to "stay out of trouble." The release date provided in this data does not consistently mean the beginning of a period during which the individual can choose whether or not to pursue a criminal lifestyle. This does pose a problem for the application of survival analysis and will be discussed later.

Both the Date_Confined and Date_Released variables were stored in Excel with two parts, the date and the time. An example of this would be: 01/01/2003 10:04:00. The time

was presented in military time format, so no AM or PM designation was needed. Because of Excel formatting options, it is possible to suppress one or the other aspects of the date/time information. In fact, two columns were created that would show the date and time data separately. The underlying data is the same for each column, but the formatting options allowed one aspect of that data to be shown for each column.

As stated previously, the data covers the bookings over the period 1/1/2003 to 1/31/2015 and contains 116,354 booking records. This is an extended period of time and has many bookings from this period. However, this data is not perfect, nor is it complete.

First of all, even though there were 116,354 booking records, it does not mean there were 116,354 separate bookings. As it turned out, if a person was booked for more than one "charge," that would create a separate booking record for each charge. In the spreadsheet excerpt given previously, the booking number, 03J0002-002, has five separate records, corresponding to five separate charges. These five separate charges are described by the Offense and Statute_Description variables. The first four variables would be the same for the five charges. Only the last two variables would contain the differing charge information.

Also, there was some repetition of records, where one record seem identical to another record, even for the Offense and Statute_Description variables. In addition, separate charges seemed very similar, and could make one wonder if there was repetition to some degree. For example, booking number 03J0015-001 had three separate records. Two of these records had the following values for Offense and Statute_Description: U302D,

ISSUE DISHONORED CHECK and U3020, ISSUE WORTHLESS CHECK. Judging from the differing Minnesota Offense Codes, it would appear that separate but similar charges arose from one incident. Is it possible that "U302D" is an error, and the booking authority meant to put "U3020"? However, is it possible that the person charged for two separate, but similar, incidents at the same time? Or, were there different aspects to the same crime that merited two separate charges? Or is it possible that a charge was entered, but the booking officials intended to amend that charge by entering another? There were many such issues with the Offense and Statute_Description variable values.

On some occasions, different records for the same booking incident would have identical Offense variable values but have different Statute_Description text variables. For example, booking number 03J0004-001 has Offense and Statute_Description values of J3500 and TRAF-ACCID-MS-DRIVE UNDER INFLUENCE OF LIQUOR while booking number 03J0005-001 has Offense and Statute_Description values of J3500 and 4TH DEGREE DWI. For other records, the opposite problem exists, where there Offense MOC values are different while the Statute_Description values are identical. An example of this can be shown by a booking record for 03J0938-002 has values of X6700 and 5TH DEGREE ASSAULT - DISORDERLY CONDUCT while booking number 02J1530-004 has variable values of A5000 and 5TH DEGREE ASSAULT - DISORDERLY CONDUCT.

Consider also this combination of Offense and Statute_Description variable values for booking number 02J2336-002: J3901, FAIL PROVIDE PROOF INS - DAR - LEAVE SCENE OF

ACCIDENT. This would appear to be three separate items listed under Statute_Description, yet one Offense code value.

In addition, a question can occur to a person when looking over this data: does this consistently reflect all the charges that someone was booked for at the time of booking? Or is it inconsistent in the sense that one person may have been had five charges recorded in the data, but really be facing seven, while another person shows two charges for a particular booking and that is the extent of the charges against them? It may also occur to someone who is examining the data that it seems likely that someone may have charges added later as Stearns County prosecutors examine the case. Would the additional charges show up in this data? Also, there are 480 records that simply say "DISMISSED" for the Statute_Description. Would this always happen when charges were dismissed, or would this be inconsistently done? Also, how should dismissed charges be treated in any statistical analysis?

There were no answers to these questions about the data. Moreover, a comparison of this data with other Excel spreadsheets that had bookings has shown there were some bookings that were missing from the main Excel spreadsheet of bookings. This discovery of missing bookings did not happen often, but when missing bookings were discovered, the bookings were added to the main Excel spreadsheet of bookings. There was no apparent reason why these bookings were missing from the main booking data.

Lastly, there were many bookings where the Date_Released data was blank, but there was release information for that booking found on other Excel spreadsheets from

Stearns County. It is conceivable that the first Excel spreadsheet, which had booking information covering the period 1/1/2003 to 10/30/2014, was created before the other spreadsheets examined. Some individuals, whose bookings were shown on that spreadsheet, had not been released at the time of its creation, meaning there was no release date information to be shown. The other spreadsheet were created after a period of time and were able to show a number of release dates that were not available when that main spreadsheet was created.

A concerted effort was made to complete data that was missing from the main Excel spreadsheet of data. This meant adding bookings that were missing from the main spreadsheet, and also filling in missing release date data.

It was clear that the use of explanatory variables, other than gender, were not possible with this data. Simply put, there were no explanatory variables other than gender. Other Excel spreadsheets with booking information were no help in this regard, and further inquiries to Stearns County yielded no additional information. The Offense and Statute_Description variables, due to the inconsistencies described previously, appeared to be inconsistent and unreliable.

A decision was made at this time to focus on the timing data. The Booking_Number, Date_Confined, and Sex variables seemed very consistent and reliable. Much effort was spent in cross-checking all the variables, and these three appeared to be reliable. The Date_Released information was also reasonably reliable aside from the issue of when the spreadsheet was created, as described above. One spreadsheet may have been produced

before someone was released, another spreadsheet may have been produced after the individual was release. The first data will not show a release date while the second spreadsheet will show a release date.

The information provided by the Offense and Statute_Description variables was not important in regards to timing. As stated earlier, one booking may generate several records in the spreadsheet because different charges would be described by the different MOC and text descriptions. Each data record, which corresponds to a line in the Excel spreadsheet, would have the same booking number, confined date and time, release date and time, and gender. The following excerpt is an example:

01/02/2003 21:35:00	02J1767-003	M	01/03/2003 18:26:00	P311C	CRIMINAL DAMAGE TO PROPERTY
01/02/2003 21:35:00	02J1767-003	M	01/03/2003 18:26:00	Z2683	PROSTITUTION
01/02/2003 21:35:00	02J1767-003	M	01/03/2003 18:26:00	J3901	DAS

Figure 2: Example of Booking Records in Excel

Here we see three separate records corresponding to one booking incident. Presumably, these records would imply that this individual was arrested and booked with three charges. Note how the booking number, time of booking, time of release, and the gender are the same for every record. It is the MOC and statute description that differs. If the focus is on the booking number, the date and time of booking, the date and time of release, and gender, it is better to have one record and not three.

After adding missing bookings found in other Excel data and filling in missing release information, the combined Excel spreadsheet was in a position to be reduced to get rid of

duplicates. The goal of this was to have each booking number be unique and only appear once. There were two kinds of duplicates present in the original Excel spreadsheet data, both kinds of which have already been mentioned, but will be summarized here. One kind of duplicate record was where two or more records would be seemingly identical. Here is an example of this kind of duplicate:

01/03/2003 14:20:00	02J0870-005	M	01/03/2003 18:12:00	J3900	DAS
01/03/2003 14:20:00	02J0870-005	M	01/03/2003 18:12:00	J3900	DAS

Figure 3: Example of Duplicate Excel Records

The second kind of duplicate are where differing charges are presented for one booking. Here is an example of this sort of duplicate:

01/07/2003 04:15:00	02J0900-006	M	05/06/2003 07:54:00	P1114	CRIM DAM/POSS BURGLARY TOOLSQ
01/07/2003 04:15:00	02J0900-006	M	05/06/2003 07:54:00	P1110	AIDING & ABETING-CRIM DAMAGE TO PROP-1
01/07/2003 04:15:00	02J0900-006	M	05/06/2003 07:54:00	J3900	DAR
01/07/2003 04:15:00	02J0900-006	M	05/06/2003 07:54:00	J3900	DAR - NO INS
01/07/2003 04:15:00	02J0900-006	M	05/06/2003 07:54:00	B3794	3RD DEG BURG - 2ND DEG BURG - POSS STOLEN PROP
01/07/2003 04:15:00	02J0900-006	M	05/06/2003 07:54:00	U2280	FELONY THEFT

Figure 4: Example of One Booking Event with Different Charges

In this case, six separate records have come from one booking. Note that there are two records corresponding to the Minnesota Offense Code (MOC) J3900. It is an open question if there is a legitimate distinction in charge between these two records—is there really a difference in terms of charge between DAR and DAR-NO INS? There is no answer at

this time for that question. If one examines the other records, note that if one is interested in the first four variables, which will be the same for all six of the records, then having six separate records is duplication of the same record. This is why the decision was made to focus on the booking and the timing data, together with the gender, and hence make use of survival analysis.

An analysis that took into account the Offense and Statute_Description variables would have to be a separate analysis. It would also require better data and a better understanding of the Offense and Statute_Description variables. The presence of these variables created duplicate records, the data had problems with the consistency and reliability of these variables, which we have discussed. One further issue of note is that when missing bookings were found in other Excel spreadsheets, these records were always missing the information contained by the Offense variable. The Offense variable was not present in any other data from Stearns County.

Because of these problems associated with the Offense and Statute_Description variables, and also because of the interest in the use of survival analysis, the data was therefore cleaned with the goal of having one occurrence of a booking number in the data.

There were some problems associated with a few of the booking numbers. As an example of the problems associated with the booking numbers, the individual associated with 02J3502 proves to be informative. This number would indicate a first booking in 2002, yet 02J3502-001, 02J3502-002, 02J3502-003 bookings are missing from the main Excel spreadsheet, as well as from the other Excel and JMP data that have been examined. An

obvious explanation is that those bookings all occurred in 2002, when the Excel spreadsheet was constructed from bookings that occurred after January 1, 2003. There is no way to be certain about that, other than the first booking, however. The booking associated with 02J3502-001 would have to have occurred in 2002—the date of the first booking will coincide with the year in which the booking number was generated. Here are bookings listed in the data associated with this individual:

01/28/2006 00:11:25	02J3502-004	M	01/28/2006 16:57:00	P1110	PROP DAMAGE-FE-PRIVATE-UNK INTENT
04/17/2006 15:25:00	02J3502-005	M	04/17/2006 15:55:00	C1200	FORGERY
03/12/2008 21:55:00	02J3502-005	M	08/14/2008 16:04:00	W1643	TERRORISTIC THREATS
08/25/2006 17:30:00	02J3502-006	M	08/28/2006 12:23:00	J3900	NO MN DL
09/29/2010 13:15:00	02J3502-006	M	09/29/2010 15:49:00	A9512	TERRORISTIC THREATS
02/13/2007 12:25:00	02J3502-007	M	03/01/2007 11:13:00	P1110	AID AND ABET CRIMINAL DAMAGE TO PROPERTY
04/02/2012 05:00:00	02J3502-007	M	04/02/2012 16:58:00	A9500	TERRORISTIC THREATS
09/24/2013 08:50:00	02J3502-008	M	09/24/2013 11:41:00	A9500	TERRORISTIC THREATS

Figure 5: Errors in Bookings and Time of Booking Records

In addition, the bookings associated -005, -006, and -007 each had two listings with two separate booking times before a final -008 booking with one booking time. This was indeed puzzling. Duplicate booking records were common enough, but this was a rare case in the data where one booking number would occur more than once with different booking times.

In resolving such conflicts, looking through the other the other Excel spreadsheets was often helpful. After searching through the other data for these booking numbers, the

distinct impression began to form that it was if, after the booking -004, there were two sequences of bookings for the chain -005, -006, and -007. The booking dates associated with the booking numbers ending in -005 were 04/17/2006 and 03/12/2008. The booking dates associated with booking numbers ending in -006 were 08/25/2006 and 09/29/2010. Finally, the booking dates associated with booking numbers ending in -007 were 02/13/2007 and 04/02/2012.

Since the times are assigned to a booking number by computer, one has to start with the assumption that an earlier date cannot be assigned to a later booking number. That is, for example, a booking number ending in -007 cannot get an earlier booking date than the booking number ending in -006. Yet, note how the earlier dates for each of the three bookings forms one chronological sequence, and also the later dates for each of the three bookings forms another chronological sequence. Otherwise, we would have to consider the possibility that booking number ending in -005 had a booking date of 03/12/2008 while the booking number ending in -006 had a booking date of 08/25/2006, which was nearly two years earlier.

Fortunately, there was only one booking time associated with the booking number ending in -008, which was 09/24/2013. This provided a chronological boundary of sorts. The other Excel spreadsheets did not help with this situation. The other data simply had the same data. Lacking any other information in the matter, three deletions were chosen so the sequence of the bookings would fall in chronological order. The corrections resulted in the bookings with associated confinement times as follows:

01/28/2006 00:11:25	02J3502-004	M	01/28/2006 16:57:00	P1110	PROP DAMAGE-FE-PRIVATE-UNK INTENT
04/17/2006 15:25:00	02J3502-005	M	04/17/2006 15:55:00	C1200	FORGERY
08/25/2006 17:30:00	02J3502-006	M	08/28/2006 12:23:00	J3900	NO MN DL
02/13/2007 12:25:00	02J3502-007	M	03/01/2007 11:13:00	P1110	AID AND ABET CRIMINAL DAMAGE TO PROPERTY
09/24/2013 08:50:00	02J3502-008	M	09/24/2013 11:41:00	A9500	TERRORISTIC THREATS

Figure 6: Corrected Sequence of Bookings

As to the problem of why the -001, -002, and -003 bookings are missing, no answer is apparent, other than the bookings occurred prior to January 1, 2003. This is quite common in the data. In fact, proximate to the booking number just considered, the individual identified by 02J3501 has booking numbers ending in -002, -003, and -004 present in the data, but -001 is missing. We can see from the booking number that it was assigned in 2002, and when that number was created and assigned in 2002, it would have had the ending -001.

Also, the individual identified by the partial booking number 02J3504 has booking numbers ending in -003, -004, and -005 present in the data, but the booking numbers ending in -001 and -002 are both missing. Again, the first and most obvious explanation for those bookings being missing would have to be that both bookings likely occurred in 2002, before this spreadsheet was created.

Also, the individual who would be identified by the partial booking number 02J3503 is missing altogether. There is most likely nothing mysterious about this. It is easy to conjecture that there was booking number 02J3503-001 created and assigned in the year

2002, but that individual was never booked into the Stearns County Jail again, and therefore does not appear in any data that begins on January 1, 2003.

Note that the booking number 02J3503-001 has the numerical sequence of 3503. This would mean that it would have been the 3503th booking for the year 2002. This number would have been assigned rather late in the year, so is it possible the bookings never got that high during 2002? The last sequential booking number in the data that was assigned in 2002 ends in 4627, so it most likely that the booking number 02J3503-001 was assigned. In any event, it does not appear in this data.

What is intriguing is the presence of 02J9999-002, 02J9999-003, and 02J9999-004 in the data. If the construction of the booking number holds true, this would mean that this represented the 9999th booking for the year 2002. It would have been extremely unlikely for there to be this many bookings for that year. The highest sequential numbers assigned for any given year tended to be the area of 3500-4500. Even if that were possible in Stearns County for that year, the booking numbers between 02J4627 and 02J9999 should have had a presence in the data, but such booking numbers are missing.

It raises the question as to how the booking number that has the sequence 9999 got generated. It can be conjectured that the computers down at some point and the personnel on duty had to assign a booking number manually. In such a situation, the personnel would have known that it would have been very bad to have duplicate booking numbers, so they would have wanted to issue a number that would be very unlikely to be generated by the computer software, so 02J9999 would be a very logical one to assign manually. This

extended discussion of how this one booking number was generated is intended to provide an idea of the complexities of the booking numbers.

One other matter that needs mention here. There are many booking numbers that begin with 99, 00, 01, and 02, which would refer to the years in which the booking numbers were assigned. There are no booking numbers that begin with numbers smaller than 99, which is a good indication that this system of generating booking numbers began in 1999. It is unclear what was done for booking numbers created before 1999. There are no booking numbers that have a format different than what has been described. This would indicate that people were booked into the current booking number format and any numbers assigned previously were not carried forward, but there is no way to be sure about that.

There are many bookings that appear to be missing. This may not be all due to errors or oversight. Recall that in the composite Excel spreadsheet dates composed of 116,354 booking records that there were 480 cases where the Statute_Description variable had "DISMISSED" in the field. It can be speculated, but unfortunately, cannot be confirmed or refuted with the data available, that perhaps these occurrences of dismissed charges in the data were not intended to be there. If the decision was made, after the booking was performed, to dismiss some or all charges, would that have meant removing the booking records? There is no answer to that question.

Therefore, it can be speculated that perhaps the usual practice was to remove charges from the booking data when charges were dismissed, and these 480 instances represent errors—records that were to have been removed but were not. This would at

least partly explain missing bookings. When bookings were found in other Excel spreadsheets that were missing in our main Excel data, there was nothing about these records to indicate a reason as to why the bookings were missing in the main Excel spreadsheet. Perhaps the charges had been dismissed, perhaps it was oversight. There is no way to explain why some bookings are missing.

In any event, a concerted and time-consuming effort was made to accomplish three things: 1) Find missing bookings; 2) Locate missing release date and time data; and 3) Cull and clean the data so each booking number is unique and is associated with just one date and time of confinement. All of this was done with the use of Excel and JMP, which is how the data was stored.

Also, the decision was made to keep booking records even if the charges were dismissed. There were some corrections that were made to booking numbers based on the other Excel spreadsheet data. There was a case where the booking number had the "J" missing, and this was corrected. There was a case where the three digit extension was incorrect, and was updated. Some deletions of bookings were made to address incidents, as described above, where there were duplicate booking numbers had different booking times.

Before these changes were made, Excel spreadsheet was sorted and ordered by booking number, date/time of booking, date/time of release, and also by the Offense and Statute_Description variable. After the changes and corrections were made, the Excel procedure was performed to remove duplicates.

With Excel, the duplicate removal procedure depends on the order of the data, since the duplicate removal procedure keeps the first occurrence of a data record and eliminates all duplicates that occur after that record. This means that the sorting of the data is critically important before the Excel procedure to remove duplicates is performed.

When sorting in Excel is done, it was desirable to have records occur first that had actual data for a variable and not blank data for that variable value. To that end, it was important to include date/time of release, Offense, and Statute_Description variables in the sorting when it might appear that only the booking number and booking date/time would be the only variable values needed to sort the data by. The reason is that a blank record will come after an actual data value when that variable is included in the sorting criteria. To see why this is so, it is important to imagine one booking number which has resulted in two records in our Excel spreadsheet resulting from two charges made during the booking process. Suppose one of these records was blank for the release date/time while the other record has a value for the release date/time. If all the other variable values are equal, the record that is blank for release date/time variable will be sorted into the last position, and therefore eliminated when Excel removes duplicates.

The same is true for the Offense and Statute_Description variables. Recall that the booking number, the booking date/time, and gender had no blank data values. But for the other three variables, this procedure could be used to selectively keep records that actually had data values for the other three variables.

It is important to now summarize the situation with the Excel spreadsheet that had 116,354 booking records. The records were all sorted by Booking_Number, Date_Time_Confined, Date_Time_Released, Offense, and Statute_Description. The Sex variable was not part of this sorting because there were no blank records for the variable, and its value was not important for ordering the data. Extensive time was spent with this and other booking data stored in various Excel and JMP data sources to find missing booking numbers, find missing variable values for Date_Time_Released, resolving various errors in data (such as duplicate booking numbers). The data was re-sorted to reflect the corrections, additions, and updates. Then the Excel procedure to remove duplicates was performed to result in a situation where every booking number was unique and associated with one booking date/time. This was done in such a way, as described above, so that as many blank data variables as possible were eliminated while keeping records that had data values for those variables.

Chapter 7: The Second Stage of Processing Excel Data

The result of these efforts was an Excel spreadsheet of booking data with 88,768 records. Each record is an individual booking that occurred over the period 01/01/2003 to 01/31/2015. The 88,768 booking records in this spreadsheet are unique with no duplicate booking numbers. For the follow data procedures that were performed in Excel, please see the appendix for the Excel formulae that were used in the creation of new variables, or Excel columns.

The second stage of processing the booking data in Excel had many aspects. The first things that were done were to re-order and re-name some of the variables. Then, two variables were added. In terms of variables, or Excel columns, the data now consisted of Booking_Number, Date_Confined, Time_Confined, Date_Released, Time_Released, Gender, MOC, and Statute_Description. Note that from the original Excel spreadsheets, the following variables, or columns, were re-named: Booking_#, Sex, and Offense were re-named Booking_Number, Gender, and MOC. MOC is the commonly used acronym for Minnesota Offense Code.

Recall how the Date_Confined and Date_Released variables consisted of two portions: the date and the time. In creating the new variables, Time_Confined and Time_Released, no new information was generated. New variables (or Excel columns) were created, the data values were copied from the Date_Confined and Date_Released columns, and the Excel formats for the four columns were chosen to reflect the information that was desired for that particular column. For the Date_Confined and Date_Released variables, the

mm/dd/yyyy format was chosen. For the Time_Confined and Time_Released variables, the military format (hh:mm) was chosen to represent the data. Therefore, for the Date_Confined and Time_Confined variables, the data was the same, but was simply displayed selectively. The same is true for the Date_Released and Time_Released variables.

However, a problem arose when SAS appeared to be unable to understand the time portion of the Excel data. SAS read the date correctly and was able to present the Date_Confined and Date_Released variables accurately, but had trouble with Time_Confined and Time_Released variables. It became clear SAS was having trouble because Excel was storing both the date and the time data in the columns. It was reading and presented the date information, but was having trouble with the time data. Perhaps it was because the time data was stored second after the date information.

Both the Time_Confined and Time_Release variables, as columns in Excel, had to be recreated to contain only the time value. This was different than the original method where the date and time information were both contained in the column, but showing only the time information. Once this was done, SAS had no problem presenting the time variables accurately.

In its refined form of 88,768 booking records, there was no missing data for the variables Booking_Number, Date_Confined, Time_Confined, and Gender. For both the Date_Released and Time_Released variables, there are 134 missing data values, for each variable. For the Offense variable, there were 288 missing data values, and for the Statute_Description variable there were only two blank records. It is important to point out

that when missing bookings were added from other Excel or from JMP data, these records were always missing the Offense variable, so such additions resulted in a blank data for that variable.

From this point, much time and effort was spent on creating new variables with an eye towards diagnosing data problems, and eventually attempting survival analysis, with the booking data. The booking number is the most fundamental information in the data. In fact, the first action taken was to move the Booking_Number variable to the left, to be first in the order of Excel columns.

From Booking_Number, two more variables were created: Booking_7 and Booking_3. Booking_7 represented the first seven figures of the full booking number while Booking_3 represented the last three figures from the booking number. Recall that the first seven figures from the booking number are unique to the individual and represent one person. The last three figures from the booking number represent the number of bookings that individual has undergone. The two highest number of bookings seen in this data, as revealed by the Booking_3 variable, are the 63rd and 72nd bookings of two individuals.

The data does not have all 135 bookings for these two individuals. For the person who has been through 63 bookings, the first booking is missing. For the person with 72 bookings, the first four are missing. Both booking numbers begin with 02, meaning that the very first booking was in 2002 for each person, so those particular bookings are missing most likely because the bookings occurred prior to the time period covered by the data. Here is

an example of how this segment of the data now appears, along with the Date_Confined and Time_Confined variables:

Booking_Number	Booking_7	Booking_3	Date_Confined	Time_Confined
00J0068-005	00J0068	005	4/25/2005	14:00
00J0068-006	00J0068	006	12/7/2006	18:32
00J0068-007	00J0068	007	4/24/2009	1:35
00J0102-004	00J0102	004	6/17/2003	11:05
00J0102-005	00J0102	005	7/25/2003	19:55
00J0102-006	00J0102	006	12/4/2003	20:45
00J0102-007	00J0102	007	2/8/2004	0:15
00J0102-008	00J0102	008	6/10/2004	23:10
00J0102-009	00J0102	009	11/2/2004	9:50

Figure 7: Example of Corrected Excel Records

The booking number was broken up into its component parts by using the Replace() formula in Excel. The green color for one record indicated it was added from another Excel spreadsheet and was therefore missing from the two main Excel spreadsheets obtained from Stearns County. There is no obvious reason why it was missing from the main Excel spreadsheets obtained from Stearns County. The booking occurred in 2009, so it fell into the time period where it should have been captured by the main Excel spreadsheets. Since it was obtained from another Excel data source, the Offense variable has a blank data value for this booking. None of the other Excel spreadsheets had this variable other than the two Excel spreadsheets that formed the fundamental source of the data used here.

During these two stages of cleaning, correcting, and culling the data, then re-naming, re-ordering, and refining the variables, as well as creating some new variables, some columns were copied and kept to the right in the Excel spreadsheet. Another column of

booking numbers, called Booking_Number_2 was kept. This consisted of the original booking numbers from the main Excel spreadsheets. Occasionally, booking numbers were recognized as having an error. The Booking_Number variable received this correction, but Booking_Number_2 did not. The intention was to keep its values in original form. This was done in case it was ever useful to see a booking number in the original form. This was also done to make sure the records were kept in sequence and guard against unintentional alterations.

This was particularly important with the Date and Time released information, given the attention that the variable received, as well as data additions. To that end, Date_Release_2 was created and had the date and time released information copied. Therefore, Booking_Number_2 and Date_Released_2 were created and shifted off the right, when viewing the Excel columns, and were away from the main focus of work on the variables that was taking place.

At this point, the Excel data consisted of twelve variables in the following order: Booking_Number, Booking_2, Booking_7, Booking_3, Date_Confined, Time_Confined, Date_Released, Time_Released, Booking_Number_2, Gender, Date_Released_2, MOC, and Statute_Description.

Chapter 8: Data Issues and Creating New Variables in Excel

The variables created and described previously were derived from the booking numbers, confinement and release dates and times. Beyond this, 21 new variables were created in Excel to both discover problems with the data and also to prepare the data for survival analysis using SAS. The first of these variables created was a dummy variable named, Skip_In_Booking. This was a dummy variable whose purpose was to determine if there was a missing booking number, and assign 0 if the record did not represent a skip in booking number and assign 1 if the record did represent a skip.

Recall that the three digit extension of a booking number is intended to indicate number of bookings the individual has undergone. The first booking, for example, will always have -001 as an extension. By a skip in booking number, what is meant is that if there were multiple bookings for an individual, in evaluating the three digit extensions for those multiple bookings, there was a gap in the sequence of resulting booking number extensions. For example, suppose an individual showed booking numbers with the -003 and -005 extensions, but not the -004.

After making the concerted effort (mentioned previously) to find missing data, almost all such gaps were filled in. Rarely, some of these gaps were due to the data having an incorrect booking number extension, such as a booking number having -004 ending that should have been -003. This was determined by examining the other Excel spreadsheet data, along with the booking times. Most of the time, the gap was simply because a booking record was simply missing from the main Excel spreadsheet.

This specifically excluded situations where the first booking numbers were missing. Because of when the data was created, recall that the composite Excel spreadsheet covered the period, 1/1/2003 to 1/31/2015, bookings that occurred in the time prior to 1/1/2003 would obviously not be included. Therefore, the first booking number for a particular individual in the spreadsheet could have an extension, for example, of -004, meaning that bookings -001, -002, and -003 probably occurred before 1/1/2003. Therefore, when constructing the Skip_In_Booking variable, it only considered the situations where booking number extensions began with some number, whether it be -001, -002, or higher, yet had a gap in whatever sequence of booking number extensions that had been started.

After the efforts to correct errors, fill in missing release dates and times, and adding missing bookings, there was only one booking gap left in the data. The good news is that this is the only skip in booking seen in data that amounts to 88,768 records. The following partial records show this gap, and note that only the first five variables are shown for each record, including the booking date and time:

02J2406-002	02J2406	002	3/28/2003	15:25
02J2406-003	02J2406	003	6/6/2003	17:40
02J2406-005	02J2406	005	2/11/2005	22:41
02J2406-006	02J2406	006	4/8/2005	15:15
02J2406-007	02J2406	007	2/22/2007	19:45

Figure 8: Excel Record with Missing Booking

This particular individual had 16 separate bookings in the main Excel spreadsheet data, but the only the first five bookings are shown here. This sequence is missing the bookings that have the extensions -001 and -004. Note that the entire booking number

begins with 02, meaning it was generated in 2002. That means the absence from the Excel spreadsheets of 02J2406-001 is very understandable—that particular booking occurred before the data begins chronologically. As stated previously, the dummy variable Skip_In_Booking ignored missing bookings at the beginnings of sequences for that very reason. The skips of interest were ones that occurred in the midst of established sequences in the data. In any event, the bookings associated with -001 and -004 were simply not found in other Excel or JMP spreadsheets.

In the original composite Excel spreadsheet that is used as the basis for the booking information, the bookings associated with -001, -003 and -004 are all missing. As mentioned previously, it is understandable why the -001 booking number would be missing, but no reason is apparent as to why -003 and -004 are missing. These booking numbers are missing from all other Excel data from Stearns County.

However, in the original Excel data, there were two occurrences of the booking number 02J2406-002 with different booking dates and times, 3/28/2003 and 6/6/2003. This is an oddity because booking numbers are intended to have one associated date and time. The other Excel and JMP data was of no help to resolve this discrepancy, as well as resolving the problems with missing booking numbers. The missing booking number with the extension -001 was understandable, as has been discussed. So, a decision was made to assume that the booking number with the extension -002 with the later date was in fact the booking number with the extension -003, and to alter the data by making this correction.

That is why the record associated with the booking number 02J2406-003 appears in yellow. That color means a record that was corrected.

Note that it could have been assumed that the -002 booking with the earlier date should be -001, but that would create a problem. The fact that the booking number begins with 02 means the booking number with the extension -001 was generated in 2002, not in 2003. This attempted correction would have meant the booking number 02J2406-001 had a booking date of 3/28/2003, which should be considered unlikely. It is also possible to assume that the later booking number with the -002 extension was -004, leaving -003 as the missing extension. However, given the gap between booking times for the two bookings ending in the -002 was less than three months, it seemed safer to assume the missing booking number had the -003 extension and not -004. There was no resolution to the problem of why the booking number ending in -004 was missing. Still, discovering this problem, and determining the degree to which it was a problem was made possible because of the development of this variable.

The second Excel variable that was created was a dummy variable called Previous_Booking and was designed to address the problem mentioned previously. Because this has bookings beginning 1/1/2003, many times in the data, the first booking for a person that is shown in the data was not the first booking for that individual. That person may have been booked one or more times prior to 1/1/2003, so the first time a booking appears for them in the Excel data, it may have a -002 extension or perhaps higher.

This variable will check to see if a booking was the first for that individual to appear in the data, and if so, will check to see if it has the -001 extension. If the booking is not the first for that person to appear in the data, or if it is and has the -001 extension, the variable takes on a value of 0. If the booking record is the first for that individual to appear in the data and does not have the -001 extension, the variable takes on a value of 1. It turns out that there are 2493 cases of this in the data.

The next Excel variable created is a dummy variable called First_Booking. The idea for this variable is based on the previous variable, Previous_Booking. Because of the booking number extension, one knows that all bookings that end in -001 would be first bookings. However, there may be an interest in first bookings as they appear in this data. As revealed through the previous variable, there are many bookings in this data that are the first bookings for an individual that appear in the data, but are not the first bookings for that individual at the Stearns County Jail.

Such bookings are revealed by the Previous_Booking variable, but is of interest to combine those bookings with all those that end in -001. The First_Booking variable will count all those revealed by the Previous_Booking variable plus all those booking numbers that end in -001. This variable attempts to identify all bookings in the data that are first bookings for an individual that appear in the data, no matter what the booking number extension may be. It assigns 0 to a booking number that is not the first for a person to appear in the data and assigns a 1 to a booking number that is the first booking number for an individual to appear in the data. It turns out there are 36,609 such bookings. By

coincidence, this is only one more than the number of censored bookings, as revealed by the Censor_Flag variable discussed later.

The fourth Excel variable to be created is Booking_001. This is a dummy variable that keeps track of booking numbers that end in -001. This will assign 0 to booking numbers that do not end in -001 and assign 1 to booking numbers that do end in -001. The count of these variable values shows that there are 34,116 booking numbers that end in -001. When combined with the count of 2493 provided by the Previous_Booking variable, the total comes to 36,609. The same procedure was done to create dummy variables for the bookings with extensions -002, -003, -004, and -005. These variables were called Booking_002, Booking_003, Booking_004, and Booking_005.

The next Excel variable created was called, No_Release_Date. This a dummy variable that was to indicate whether or not the booking number had a date (and time) for release of the individual after booking. There would be two explanations for this: the person had not been released and was still in custody at the Stearns County Jail at the time the Excel data had been compiled, or there had been an error and the information had not been recorded in the Excel spreadsheet.

As mentioned previously, many cases had been found where the main Excel spreadsheet had no information for release date and time. Recall that for the original unedited Excel spreadsheet, which had 116,354 records, the Date_Released variable had 469 blank records. With the corrected and edited Excel spreadsheet that was used for the following analysis, there were 88,768 records and this variable had 134 blank data values for

this variable. There were some cases where the main Excel spreadsheet that was used as a basis for the data simply did not have the date released information. This could have been simply an error when creating the spreadsheet. Many of those cases actually had the release date information on other Excel or JMP spreadsheets. Other cases of a booking number missing release date and time values seemed to have occurred because the spreadsheet was created before the individual was released.

One could easily see how many data values for this variable were missing by simply using Excel to count blank records in this column. However, that would have been useless in SAS if there were a need to take this into account for some analysis. It was necessary to create a dummy variable, No_Release_Date, which assigned 0 if the was a release date for that booking or 1 if the release date was missing. Fortunately, the count of 1's totaled 134, the same number as given by the Excel procedure to count blank fields in the release date column.

The next Excel variable created was Only_Booking_No_Release_Date. This is a dummy variable used to determine if there was only one booking for an individual, and if so, was the release date and time for that booking was missing. If there was one booking with a release date and time, the value is 0. If there was one booking for an individual, but with the release date and time was missing, the variable value is 1. There is a total of 27 such cases in the data, and this would have been a subset of the 134 records that lacked a release date and time.

The next Excel variable that was created is a dummy variable used to determine if that record was the first booking for an individual in the data, given that he had more than one booking, and if that booking had no release date and time. This variable was called, First_Booking_No_Release_Date. The previous dummy variable described, Only_Booking_No_Release_Date, would look to see if an individual had only one booking and then check to see if the release date and time was missing.

First_Booking_No_Release_Date, on the other hand, will check to see if an individual had more than one booking, and if so, did the first booking miss a release date and time. This variable gives a value of 0 when the first booking has a release date and time, and a value of 1 if the first booking does not have a release date and time. There were no such cases in this data.

The next variable that was created was the Skip_In_Release_Date variable. This was also a dummy variable the purpose of which was the check for a sequence of multiple bookings for an individual where there was a missing release date and time. Because of the way recidivism was defined with this data—if someone had been booked previously, being booked again was considered recidivism. The survival time was the length of time from release date of the earlier booking to the confinement time of the later booking—it became important to see if there had been a skip in release time.

For example, if an individual had three bookings ending in extensions -001, -002, and -003, but suppose the booking number ending in -002 had no release date, this would present a problem in calculating the recidivism time. This dummy variable took a value of 0

if a booking did not represent such a skip in release date and took a value of 1 if a booking did represent a skip in release date.

The dummy variable describe earlier, No_Release_Date, was not sufficient for the analysis of recidivism. For example, the records where there was no release date may be the only booking for that person, in which case, there is no recidivism. Or the missing release date information may be for the last booking of a sequence of bookings. If that is the case, there is no recidivism, and no recidivism calculation is necessary. Recall that there were 134 bookings that, for one reason or another, lacked a release date and time.

However, having a missing release time poses a problem because if a booking number shows no release time, yet that person was booked again later, obviously there is a recidivism event for which a calculation of recidivism time is needed, there is a problem created if the release date is missing. Therefore, there is a need to determine the extent to which this is a problem.

The next Excel variable created was a dummy variable to determine if the last booking in the data for an individual lacked a release date, provided there was more than one booking for an individual. In keeping with the spirit in which the previous variables were name, this variable was called, Last_Booking_No_Release_Date.

The underlying idea for this variable was that if an individual is accused of an offense bad enough or if he considered a severe risk to public, and therefore has a high bail amount, this would be revealed as a final booking (if there had been more than one booking), where there is no release. It is important to add that this variable did not evaluate cases where

there was only one booking for an individual. This variable was triggered only if there was more than one booking, but the last of which did not have a release date. This also has to be understood with the caveat that "released" does not mean "freed," and even the most dangerous people will eventually be released, although it may well mean transferring custody to a state prison or some other authority. As it turned out, there were 102 occurrences of this sort.

These last five variables can be described together as checking the different situations in which the release date and time for a booking may be missing. The reason for doing so is concern over correctly calculating the recidivism time, as well as the following time. If the release date and time is needed to perform such important functions in the survival analysis, it is necessary to find out the extent to which the problem of missing data values and to compensate for the missing data values.

Fortunately, it does not appear to be a bad problem with this data. As was discussed in regard to the No_Release_Date variable, there were 134 records for which there was no release date. In perform data explorations to determine the extent of the problem, it became clear that there were four separate situations to be considered. The dummy variable No_Release_Date provided the details on the extent of the problem. The remaining four dummy variables that were created were designed to detect these occurrences, and the number of such occurrences in the data. What follows is a summary of these four situations, the dummy variables designed to detect the specific issues, and the number of occurrences in the data.

1) Where there was only one booking for a person in the data and that booking was missing a release date and time, Only_Booking_No_Release_Date, 27 occurrences.
2) Where there were multiple bookings, and the first booking is missing a release date and time, First_Booking_No_Release_Date, 0 occurrences.
3) Where there are multiple bookings, and one of the middle bookings is missing a release date and time, Skip_In_Release_Date, 5 occurrences.
4) Where there are multiple bookings, and the last booking is missing a release date and time, Last_Booking_No_Release_Date, 102 occurrences.

Each of these four situations had to be evaluated in order to determine how recidivism would be detected and how recidivism would be calculated given that it occurred. The sum of the occurrences for the four situations comes to 134, which is the total number of blank records for the release date and time. This situation will be returned to later on.

The next Excel variable created was Hours_Spent_In_Custody. As the name indicates, its intent is to determine how many hours someone spent in custody, and was derived from subtracting the confinement date and time from the release date and time. The result was converted in a decimal number of hours. In cases where the release information was missing, the subtraction was not done and the field was blank. There were 134 records left blank because of the 134 records that lacked a release date and time.

The next Excel variable created was Days_Spent_In_Custody. This was similar to the previous variable, but expressed the value in days instead of hours. This became important

when it became clear that some individuals were spending long periods of time in the Stearns County Jail. Because the Days() function was used, it did not provide decimal number extensions. There were 134 records left blank because of the 134 records that lacked a release date and time.

The next Excel variable to be create was Recidivism_Flag. This is a dummy variable that will take a value of 0 if the booking did not have a recidivism event and a value of 1 if the booking did have a recidivism event. A recidivism event was defined for a booking if there was a subsequent booking for that person. For example, suppose a person had been booked for the first time and received a booking number that ended in -001. If that was the only booking, the Recidivism_Flag would be 0. If there is a second booking, then the recidivism flag for the -001 booking record would be 1, while the Recidivism_Flag for the -002 booking would be 0.

This is an important detail, so it bears repeating. The -002 booking creates a recidivism event for the -001 booking record, not for the -002 booking record. There is no recidivism for the -002 records unless there is a -003 booking. The result is ultimately that out of the total data of 88,768 bookings in the data, 52,160 bookings have recidivism events.

The next Excel variable created was Censor_Flag and is a dummy variable that checks to see if the booking had a recidivism event. If the booking had a recidivism event, Censor_Flag would return a value of 0. If the booking did not have a recidivism event, Censor_Flag would return a value of 1. This variable would simply check to see if the individual had another booking after the current record. If this was true, there was a

recidivism event. If this was false, the booking event did not have a recidivism event, and should therefore be censored.

As they were designed, the Recidivism_Flag and Censor_Flag should be complementary. If one had a value of 0, the other should have a value of 1, and vice versa. The next variable, Sum_Of_Flags, was designed to make sure this was true. It simply added Recidivism_Flag and Censor_Flag together. If designed correctly, the sum should always be 1.

As a result, these three variables are linked together. Recidivism_Flag showed that 52,160 of the bookings had a recidivism event. Censor_Flag showed that 36,608 of the booking records were to be censored. The Sum_Of_Flags had a column sum of 88,768. These results were as they should and showed the variables were working adequately with the data.

The next three variables that were created are also linked. The first of these is called Recidivism_Time_Days and is designed to give the time of a recidivism event for a booking. For example, if there were two bookings for a person, with the -001 and -002 extensions, -001 would have a recidivism event while -002 would not, because there was no -003 booking. The recidivism time for booking -001 would be calculated by subtracting the release date for the -001 booking from the confinement date that accompanied the -002 booking.

The next variable was Follow_Time_Days. If the booking did not have a recidivism event, it was to be censored, and therefore needed the follow time. If a booking record was

to be censored. The follow time needs to have an end to the observation window. At first, it would seem the date that should define the end of the observation window would seem to be 1/31/2015. This date was gotten from the second of the Excel spreadsheets received from Stearns County that listed bookings until 1/31/2015.

The last bookings listed were ones that occurred on 1/31/2015. However, the release dates for many of these bookings were in February, 2015. The latest such release date was 2/23/2015. This is a good indication that even though the bookings ended on 1/31/2015, the Excel list was obviously done on or after 2/23/2015, and listed release dates that were available to be included.

After numerous attempts, it was decided to consider the end of the observation window to be 2/23/2015. This was the date incorporated into the Excel formula to determine follow time for censored observations.

The next and last Excel variable created was simply a combination of the previous two columns. The variable was called, Recidivism_Follow_Time. If there was a recidivism time, it equaled the value of the variable Recidivism_Time_Days. If there was a follow time, it equaled the value of the variable Follow_Time_Days. This variable was necessary because SAS and JMP needed the recidivism and follow time figures to be held in the same variable. Whether this variable value was a recidivism or follow time depended on the value of the Censor_Flag variable, although the Recidivism_Flag variable could have served this purpose. If the value of Censor_Flag was 0, the value of Recidivism_Follow_Time was a failure time for

a recidivism event. If the value of Censor_Flag was 1, the value of Recidivism_Follow_Time was a follow time for a censored variable.

It needs to be noted that not all of these variables created in Excel were used or were useful. Many, such as the two related to time in custody, were created early in the process of data exploration in case the information was needed at some point. Many of these variables turned out to be superfluous, but they were created because of a potential value to some unforeseen analysis.

There are many details regarding the calculation of the recidivism and follow time variables that are left to the next chapter. That chapter deals specifically with the recidivism and censor variables, as well as the recidivism and follow time values. An important feature of that discussion will be regarding how to deal with cases in which the released date and time are missing.

Chapter 9: Censoring, Calculation of Recidivism and Follow Time

This collection of 88,768 bookings is not a random sample of bookings. It is intended to be a comprehensive list of the bookings over the period 1/1/2003 to 1/31/2015, although admittedly, some bookings were found that were not on the Excel spreadsheet. Also, the construction of the booking number allows one to see that there are some bookings that are not on the list. Why some bookings are not present on this list of bookings is not certain, although speculations have been made previously. In addition, there are concerns over the consistency and reliability of the MOC and Statute_Description variables.

A more positive view, however, is that although there are some bookings that are undoubtedly missing, this list of bookings is very complete. Exhaustive cross-checking with other data has left the overall impression that, for a collection of data describing real-world activity for covering a long period of time, and in regard to the variables related to booking number, confinement time, release time, and gender, it is very solid. It is comprehensive, and it is possible to view this as a census of bookings over this time period. As data approaches being a census, it can be considered a representative of future bookings that will occur, especially on a large scale.

Recidivism is defined as an event that occurs for a booking record if there is a subsequent booking. For example, if someone is booked into the jail and given a booking number ending in -001, which would mean it was the first booking for that person, it would be said to have a recidivism event if there was a subsequent booking. This is a definition that applies to any booking, whether it is the first booking or any booking. Suppose there is

a ninth booking for an individual, with a booking number that ends with -009. It would have a recidivism event attached to that record if there was another booking.

A booking which is the last booking for a person cannot be said to have a recidivism event, the reason being no booking follows it. It will be censored. The same is true if only one booking for a person appears in the data. Since no booking follows it, it cannot be said to have a recidivism event. Such data seems to be purely a matter of right censoring. A first booking event creates a presence among the data, while the lack of subsequent bookings will mean a booking record will need to be censored at the end of the observation window.

Recidivism is defined this way because it is the only way the data allows. As stated many times previously, release from the Stearns County Jail did not necessarily mean release to freedom, it just means an exit from the custody of the Stearns County Jail. That could mean a release to freedom or a transfer to a state prison. This differs substantially from the ideal way to apply survival analysis to recidivism—where there is a clear start to a lifetime that can be easily derived from the data, such as a release from incarceration, to begin a state of "freedom" and the ability to choose whether or not to engage in criminal activities. This sort of situation would be ideal—a clearly seen event defined in the data that would be a start to a lifetime where the individual has the opportunity to choose whether or not he can pursue criminal activities. This situation is not fulfilled by the release date and time in the data from the Stearns County Jail.

In lieu of any other way to define the beginning of a lifetime, which a recidivism event would end, the release date and time will have to suffice. The release date, such as it

is, often does mean the person was free to go, but there was no way to quantify how often this was the case with the Excel data. The release date may mean they were free to go or it could mean the individual was transferred to another criminal justice authority or to another facility. One is reminded of Schmidt and Witte (1988) discussing how they defined recidivism despite many obvious objections and concluding, "[w]e analyzed the timing of return to prison in North Carolina because this was the only definition of recidivism that our data would support" (p. 9).

It should be noted that despite this less than ideal situation with the release date and time, from the perspective of the Stearns County Jail, the analysis of this data would still be of interest. The reason is that the data can describe the large-scale behavior of those who are booked into the jail. It will provide them with an idea of the numbers who will return as well as the statistical distribution that best describes the behavior of those who are booked into their facility.

It is beneficial to review how recidivism and censoring are to be decided, along with the details of how the calculation of recidivism and follow time is to proceed. The first detail is regarding the date used as the end of the observation window, 2/23/2015. This date is used in calculation of follow time for booking records where censoring is to occur. This date comes about because the second of the two Excel spreadsheets from Stearns County listed bookings from 10/31/2014 to 1/31/2015, inclusive of those dates. However, even though the bookings shown on the spreadsheets end on the date 1/31/2015, release dates up to 2/23/2015 are listed for these bookings. The latest of these release dates seen in the data is

2/23/2015, so it can be considered the end of the observation window. The spreadsheet must have been created on or after 2/23/2015.

To show how these details will work with the data from Stearns County, below is an excerpt from the Excel data. There are seven bookings for one individual. These are all the bookings for this particular person that were in the Excel data, but note the -001 and -002 bookings are missing, most likely because they occurred before 1/1/2003.

The first three columns present the individual booking number and its two main components. The fourth and fifth columns represent the date and time of the booking, or as it is often referred to, the confinement date and time. The sixth and seventh columns present the date and time of release from the jail, and the last column presents the recidivism or follow time, in days.

99J5436-003	99J5436	003	5/8/2003	10:34	9/4/2003	7:25	58
99J5436-004	99J5436	004	11/1/2003	3:45	11/26/2003	15:26	41
99J5436-005	99J5436	005	1/6/2004	1:40	4/2/2004	14:05	26
99J5436-006	99J5436	006	4/28/2004	0:42	9/3/2004	22:00	424
99J5436-007	99J5436	007	11/1/2005	14:00	11/2/2005	8:25	903
99J5436-008	99J5436	008	4/23/2008	14:25	5/5/2008	11:50	123
99J5436-009	99J5436	009	9/5/2008	9:56	9/22/2008	13:26	2345

Figure 9: Example of Records after Re-arranging and Creating Variables

With this data, it worked best to define a recidivism event as occurring for a particular booking if the person was booked again. Therefore, in regard to the booking for

-003, that booking is considered as having a recidivism event because of the -004 booking. The recidivism time is calculated by subtracting the release date for the -003 from the confinement date for the -004 booking, which in this case comes to 58 days.

The -004 booking has a recidivism event because of the -005 booking, and the recidivism time is gotten by subtracting the release date for the -004 booking from the confinement date for the -005 booking, which comes to 41 days. This process continues through the bookings. All the bookings through, and including the -008 booking are considered as having recidivism events because of a subsequent booking.

The exception to this is the -009 booking. Because there is no booking that follows it, this booking is not considered to have a recidivism event. It is to be censored. The follow time is arrived by subtracting the release time for the -009 booking from the date 2/23/2015. This comes to a total of 2345 days. For both recidivism and follow time, the Days() function in Excel is used. In the case of recidivism, this function is used to find the number of days between the release date of a booking and the confinement date of the subsequent booking. In the case of follow time, the Days() function is used to find the number of days between the release date for the censored booking and the date 2/23/2015, the date used as the end of the observation period.

This discussion describes how the procedure is to work with the great majority of the 88,768 booking records in the Excel data. A difficulty is posed by the 134 bookings for which there was no release date. As discussed previously, there are four categories in which these 134 occurrences of no release data. These four categories are listed again, along with the

dummy variable used to detect its occurrence, and the number of occurrences of this situation:

1) Where there was only one booking for a person in the data and that booking was missing a release date and time, Only_Booking_No_Release_Date, 27 occurrences.
2) Where there were multiple bookings, and the first booking is missing a release date and time, First_Booking_No_Release_Date, 0 occurrences.
3) Where there are multiple bookings, and one of the middle bookings is missing a release date and time, Skip_In_Release_Date, 5 occurrences.
4) Where there are multiple bookings, and the last booking is missing a release date and time, Last_Booking_No_Release_Date, 102 occurrences.

In the case of item 1, this is not recidivism because there is only one booking and therefore will have no effect on recidivism calculations. However, a follow time is needed. Ordinarily, follow time would be calculated by subtracting the release date from 2/23/2015. For all booking records where there is a release date, a mean time in custody was found to be 8.14 days. In these cases, where there is no release date, it was decided to use the Days() Excel function to determine the number of days between the confinement date and 2/23/2015, then subtract 8.14 days from that result. In effect, this is adding the mean confinement time of 8.14 days to confinement date to make up for the lack of a release date and time.

This mean value of 8.14 days for confinement time was derived in the following manner. With 134 total missing release dates, there were 88,768 – 134 = 88,634 records

with release dates. These 88,634 totals for confinement time in days were added and divided by 88,634 to give an average length of confinement in days. This sum of confinement time in days was 721,107. 721,107/88,634 = 8.14

In the case of item 2, there were no occurrences of this in the data. Therefore, it had no impact on the calculation of recidivism or follow time.

Item 3, however, had a profound effect on recidivism calculation at first. It counts as an event of recidivism since there is a subsequent booking. In the first calculation of survival time based on a release date, since the record was blank, the recidivism length of time yielded an absurd result that profoundly affected the results, even though this was just five records. For each record, the result was over 40,000 days for the survival time.

The Excel formula for calculating survival time had to be adjusted accordingly. Therefore, for these five cases, a substitute for the release date was to use the midpoint of the confinement date and the subsequent booking's confinement date. Ordinarily, when there is a recidivism event listed in the data, the release date for the earlier booking is subtracted from the confinement date of the later booking. For these five cases, the recidivism time was calculated by subtracting the confinement time of the earlier booking from the confinement time of the later booking, and then dividing this result by two. For example, if there is a gap of 84 days between the confinement dates of the two bookings, the recidivism time will be 42 days. Because this was a recidivism event, there was no need for a follow time calculation.

The problem described by number 4 on the list above, is that there is no release date for 102 of the records where it was the last booking record for a particular individual. This was not a recidivism event, but a censored one. Following time needed to be calculated without the use of a release date.

The method to get around this problem is similar to the method done for case 1. When there was a lack of a release date, the solution applied to these records was to use the average length of confinement in days, the same figure of 8.14 that was used previously. In effect, for these 102 booking records that lacked a confinement date and time, 8.14 days was added to the confinement date to get a usable release date and time. As with all censored observations, the date of 2/23/2015 was used as the end of the observation window.

As discussed earlier, the data appears to be quite comprehensive despite some bookings that were missing for unknown reasons. Extensive efforts were made to find missing bookings, but there are clearly some that are still missing. Still, the data appears to be comprehensive enough for the period, 1/1/2003 to 1/31/2015, that the data can be considered on the level of census.

There are two assumptions about the data that also appear to reasonably solid. First, the survival times of the different individuals appear to be independent, random, and identically distributed. There is no reason to suppose that the bookings of any individual or individuals are connected with others in a consistent or systematic way.

The second assumption is that the survival time is independent of censoring time. To this end, consider the following two graphs of the recidivism and censoring times obtained by using SAS. The first shows the recidivism time and the second shows the censoring time. These appear to be very different distributions. In particular, the censoring times appear to come reasonably close to a uniform distribution. One must bear in mind that censoring times would seem to have a reason to be weighted towards the smaller, or left side of the graph. The reason is that bookings that occurred at the end of 2014 and in January, 2015, the people involved would not have had much time to recidivate, and such records would be censored with small time values for the censoring.

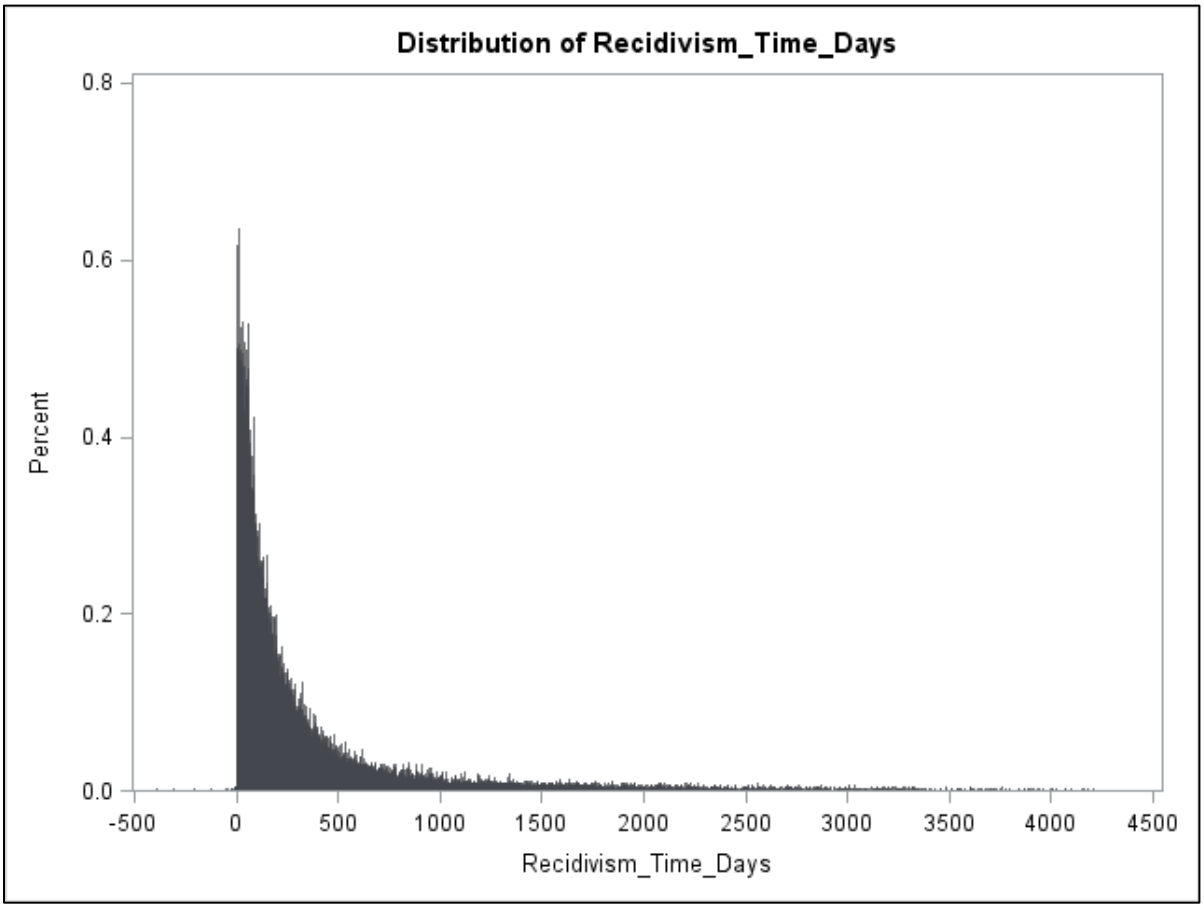


Figure 10: Distribution of Recidivism Time

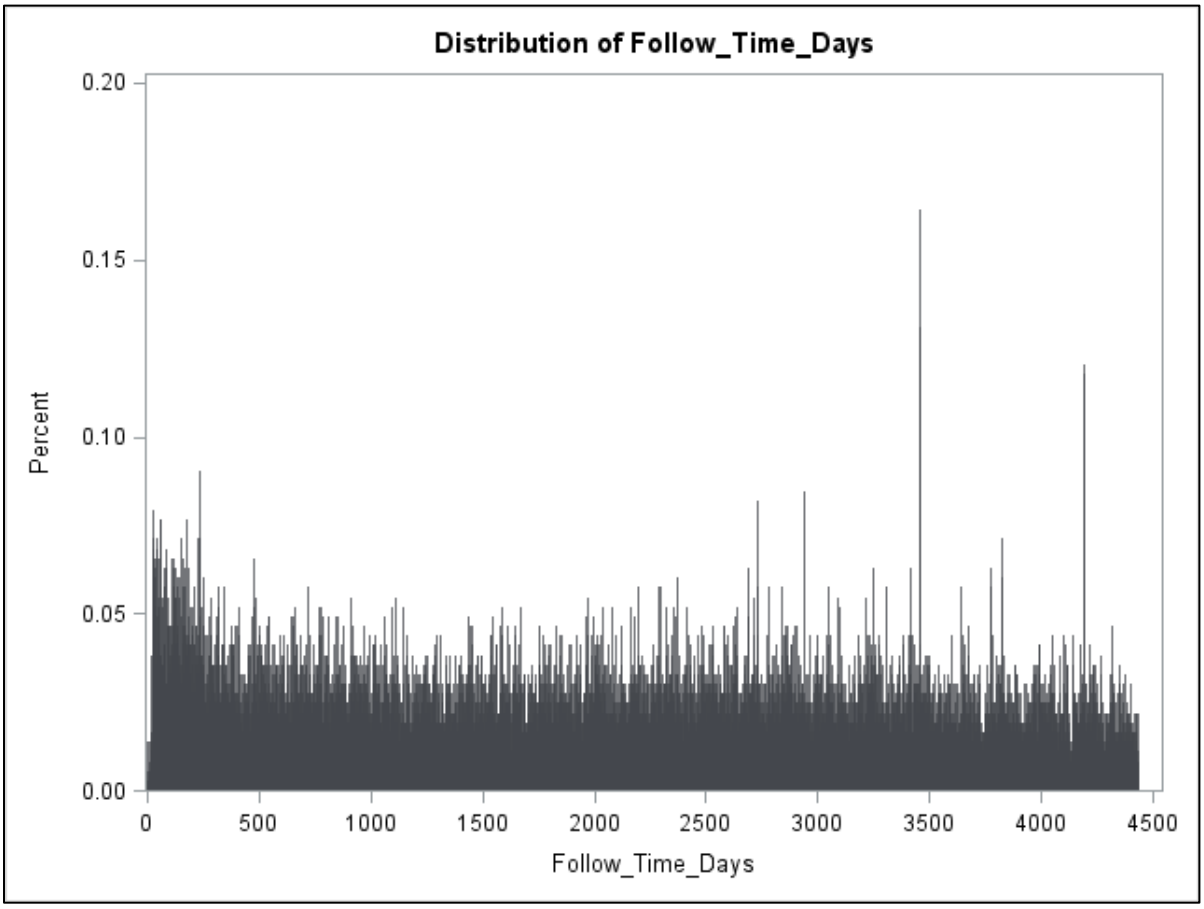


Figure 11: Distribution of Censor Time

Chapter 10: Descriptive Statistics of the Booking Data

What follows are some statistical measures intended to describe the booking data. Most basically, there are 88,768 bookings cover the period 1/1/2003 to 1/31/2015. These bookings are separate booking events that have been applied to a total of 36,608 different people. There are also 36,608 bookings that are censored. This is not a coincidence.

As it turns out, everyone will eventually be censored, because every person will have a last booking, even if it is the only booking that person might have recorded in the data. That last record will be inevitably censored, because recall that a recidivism event for a booking record is defined as having a subsequent booking. Therefore, there will always be a last booking, which will invariably be censored, because a last booking cannot have a subsequent booking.

This Excel data is intended to be a comprehensive list of bookings, but in fact, there are certainly bookings that are missing, for unknown reasons. There may have been some effort on the part of Stearns County to exclude bookings where the charges were later dropped or perhaps expunged. Some bookings were found in other Excel spreadsheets provided at other times by Stearns County, and it certainly appears that these should have been in the main Excel data.

Nonetheless, the data from the two Excel spreadsheets is a very complete list of bookings for the time period 1/1/2003 to 1/31/2015. No other list of data provided by Stearns County is remotely comparable in completeness and overall reliability. Much time and effort was spent in looking for errors and missing information, and in correcting such

deficiencies. A few errors in booking numbers were found, but no errors in confinement date and time were ever found. Missing release dates and times were an occasional problem, as discussed previously. The MOC and Statute_Description variables were found to be relatively unreliable, mostly because of inconsistent values, which has also been discussed previously. Overall, the booking records that came from the two Excel spreadsheets are strikingly reliable for the first four variables.

There are two aspects of the booking number that can be described quantitatively. The first aspect is the booking extension, which reveals the number of bookings that an individual has been through. The five highest extensions attained by separate individuals in the data are -048, -049, -058, -063 and -072. These booking totals accumulated by these five individuals are exceptional. If one examines the frequency of -001 bookings, -002 bookings, -003 bookings, etc., the number of bookings dwindles for each booking number level. This is especially true for the decrease in the number of bookings for the -001 booking as compared to the -002 booking. The decrease is by well over one-half.

What follows is a listing of the first 25 booking number extensions, which would be attached to the 1st through the 25th booking. This data will present numbers of such bookings present among the data of 88,768 bookings, as well as the proportion, and the relative rate of decrease as the bookings go up each, for example in going from -001 to -002. This table is derived by using the Excel pivot table application.

Table 1

Booking Records with Extensions 1-20, Frequency, and Proportion

Extension	Number of Bookings	Proportion	Rate of Decrease
001	34116	0.3843	-
002	15129	0.1704	-0.5565
003	9195	0.1036	-0.3922
004	6404	0.0721	-0.3035
005	4724	0.0532	-0.2623
006	3657	0.0412	-0.2259
007	2851	0.0321	-0.2204
008	2230	0.0251	-0.2178
009	1834	0.0207	-0.1776
010	1459	0.0164	-0.2045
011	1189	0.0134	-0.1851
012	981	0.0111	-0.1749
013	811	0.0091	-0.1733
014	676	0.0076	-0.1665
015	579	0.0065	-0.1435
016	472	0.0053	-0.1848
017	392	0.0044	-0.1695
018	335	0.0038	-0.1454
019	275	0.0031	-0.1791
020	233	0.0026	-0.1527
021	201	0.0023	-0.1373
022	168	0.0019	-0.1642
023	133	0.0015	-0.2083
024	111	0.0013	-0.1654
025	88	0.0010	-0.2072
Sum	**88243**	**0.9941**	**-**

As can be seen, bookings that end in -001 to -025 account for the vast majority of the booking quantity, 99.41% of the total out of 88,768 bookings. The bookings that end in -001 by themselves account for 38.43% of the total number of bookings. For brevity, this was stopped at -025. It is important to remember that the levels shown on the table do not show

the highest booking level attained, it shows how often a booking number with that particular extension is present among the 88,768 bookings in the data. For example, someone at the -009 level will also likely have the -008 booking, the -007 booking, and earlier bookings on this table, provided those bookings took place on or after 1/1/2003.

Next to consider are the years in which the booking numbers are generated, which should be the same as the years in which the first booking for a person took place. It is important to remember that the software that generates the booking numbers most likely began in 1999, and that the data analyzed here begins keeping track of bookings beginning in 1/1/2003 and ending in 1/31/2015.

These two facts mean that no booking numbers are shown that begin with 98, 97, or numbers that would indicate the initial booking happened in the years before 1999. It also means that booking numbers that begin with 99, 00, 01, and 02 are present in the data because the individual was booked for new charges after 1/1/2003. Someone who was booked before that date, and never again, will not be present in the data. Individuals with booking numbers that begin with 99, 00, 01, or 02 make their first appearance in this data with bookings that end with the extension -002 or higher.

If we make use of the Excel pivot table application once again, a table is presented which shows the years for which a booking number was generated and the quantity for that year, as seen in the data. Unlike the table for the booking extensions, the following table can summarize all 88,768 bookings, since all bookings begin with one of these two pairs of numbers.

Table 2

Bookings by Year, with Frequency and Proportion

Year	Number	Proportion	Rate of Decrease
1999	173	0.0019	-
2000	248	0.0028	-
2001	121	0.0014	-
2002	12040	0.1356	-
2003	10406	0.1172	-
2004	9215	0.1038	-0.1145
2005	8750	0.0986	-0.0505
2006	7732	0.0871	-0.1163
2007	7148	0.0805	-0.0755
2008	6143	0.0692	-0.1406
2009	5726	0.0645	-0.0679
2010	5136	0.0579	-0.1030
2011	4661	0.0525	-0.0925
2012	4630	0.0522	-0.0067
2013	3680	0.0415	-0.2052
2014	2779	0.0313	-0.2448
2015	180	0.0020	-
Sum	**88768**	**1.0000**	**-**

The relative rate of decrease column does not show any figures for the years 1999-2003 and for 2015. The reason is that the bookings for the years 1999-2002 are not comprehensive, and only appear in this data if the individual was booked again after 1/1/2003. For the year 2003, since it was the first year, there is no year with which to compare it. For the year 2015, only January was included. For those years, in this table, showing a yearly change does not make sense.

The most surprising aspect of this table is the very large number of bookings for the booking numbers that were generated in 2002, which comes to 12,040. This is something which is extremely surprising since these bookings were all created in 2002, before this data

began its accumulation of booking numbers. Since the bookings beginning with 02 were generated in 2002, all those first bookings are not part of the data. For all those 2002 bookings shown, there are no -001 bookings. The year 2002 must have had an inordinately large number of bookings at the Stearns County Jail, and by a large number. Why would this be so? No answer is available, but it can be conjectured that there may have been an administrative change in policy where certain law enforcement practices may no longer have resulted in bookings. In fact, even if one includes the year 2002 in this table, there has not been one year where there has not been a decrease in the number of bookings.

In view of the surprisingly large number of bookings for 2002, the few bookings for 1999, 2000, and 2001 is also surprising. When these three years are combined, it accounts for only 542 bookings, not even 5% of the number of bookings for 2002. This seems surprisingly small in view of how large the number of bookings for numbers that were generated in 2002. Was there an administrative change in policy, or was there a data decision, not applied perfectly, to exclude some old booking numbers? Unfortunately, there is no way to know.

Chapter 11: Non-Parametric Explorations of the Booking Data

Most statistical texts will advise a student to initially approach data from the non-parametrical perspective (Lee & Wang, 2003, p. 64). As stated previously, recidivism is a phenomenon that is extremely dependent on definition and data—how one defines recidivism and the nature of the data one has available. Because of this, it would seem to be unwise to assume a particular distribution for the data based on the results obtained elsewhere with different data.

As an example, Schmidt and Witte (1988) found a lognormal distribution worked well for their data on the time to recidivism, although this is a very simplified summation of their result (p. 17). When the same data was put into JMP, it was reported that a Frechet distribution was a slightly better fit. The data used by Schmidt and Witte (1988) defined a lifetime as starting with a release from incarceration from a North Carolina state prison, and a recidivism was defined as being re-incarcerated in a North Carolina state prison. This booking data from Stearns County is substantially different, and the definition of recidivism will, of necessity, also be very different.

In this case, as discussed at length previously, the Stearns County Jail bookings are separate events, and do not define separate people. There are 88,768 bookings in the data, but 36,608 people that account for these bookings. In most studies of recidivism, such as that done by Schmidt and Witte (1988), the evaluation period begins with an event such as a release from incarceration, and the individual who was release was followed to see if he or she recidivated, depending on the definition of recidivism that was used. This will generally

mean that separate people are being evaluated. If that person recidivates, that is the end of that person as far being a source of data.

There is another item of interest to mention. There were 19 cases of recidivism that could not be included in the various survival analyses, such as the Kaplan-Meier graphs. The SAS warnings are displayed along with the tables that accompany the Kaplan-Meier graphs. The reason that these recidivism time lengths could not be included is because the lengths of the recidivism time was negative. How was this possible? It appears that these were people who were in custody in the jail and were booked again, while still in custody. This was a new booking event, and therefore creating a new booking number. An example is given:

02J1389-003	02J1389	003	2/7/2003	4:10	2/10/2003	16:12
02J1389-004	02J1389	004	7/3/2003	10:40	7/12/2003	11:13
02J1389-005	02J1389	005	7/7/2003	14:00	7/7/2003	16:22
02J1389-006	02J1389	006	8/19/2003	16:15	8/19/2003	17:01

Figure 12: Example of Record with Negative Recidivism Length

This individual had a total of 18 bookings, the first two of which are not in the Excel data, most likely because the bookings occurred in 2002. However, only the relevant bookings are shown here. Note how the booking ending in -004 show a confinement from 7/3/2003 to 7/12/2003, but the booking which ended in -005 occurred on the 7/7/2003, and showed a release on 7/7/2003, while the person was in custody for the booking ending in -004. If a recidivism is defined as another booking, and the lifetime is defined as the length of time from the release date for the earlier booking to the confinement date of the next

booking, this particular record will yield a negative length of time for recidivism. In this case, having the date subtraction 7/7/2003 (the booking date for the -005 booking) minus 7/12/2003 (the release date for the -004 booking) will give a recidivism time of -5 days. The other 19 bookings that showed a negative length of time for recidivism appear to have similar explanations.

As to how these booking records came about, the first conjecture that can be made is that the individual must have committed an offense while in custody. From the information provided by the Offense and Statute_Description variables that accompany the bookings that occur while the individual is in custody, that is certainly possible for some of these occurrences, such as when the offense is assault or obstruction of legal process. Other occurrences, however, list crimes that are difficult to imagine as something that occurred while in custody. For example, with the booking provided previously, the Statute_Description described the charge as felony theft of a motor vehicle, which is difficult to imagine if that individual were in custody. Then again, the Statute_Description variable has a questionable level of reliability.

In these cases, it appears unlikely that the person in custody would have the opportunity to commit the offense for which they were booked. It can be conjectured that authorities became aware of another offense that had been committed by the individual before they were initially booked into jail. The person was then re-booked for that new offense with a new booking number while in custody. Was this the correct procedure, or should the charge have been added to list of charges that followed the original booking?

There is no available answer to that question. Since there were only 19 occurrences, this situation would seem to be an exception. Since SAS would not include these data values when conducting survival analysis procedures, it was decided not to attempt some kind of correction for the problem of negative survival times in the Excel or SAS stage of data processing.

There are two very important aspects of the data that must be kept in mind when evaluating the initial results. The first item of importance is that the following graphs do not distinguish between individuals, as indicated by the previous discussion. Another way of saying this is that, at first, the 88,768 bookings are treated as separate entities as to whether or not recidivism did occur. It is as if 88,768 individuals were released from custody and then evaluated for recidivism.

The second item of importance to consider is that the lifetime being measured is the "release date" when, as has been discussed at length previously, the release from the jail did not necessarily mean the individual was freed. It can be stated that this is from the perspective of the jail—they may be only interested if and when someone does return once they leave the jail.

However, if someone is interested in the details of recidivism, and the accuracy of those details, then it matters a great deal if someone leaves the jail to freedom, or if that person leaves to serve a multi-year sentence at a state correctional institution. The former person is in a position to recidivate immediately, while the latter individual will be unable to recidivate until their release from incarceration. The lifetime of that person will be given a

"boost" by being in this position of being incarcerated. Of course, it is possible for someone to commit crimes while incarcerated, but this is also another challenge to how one defines recidivism. It raises the question of whether or not this would count as recidivism if the person is already in custody. If so, it raises the further question as to how to define the lifetime of this sort of recidivism if a usual event to start a lifetime, such as release from incarceration, has not even occurred yet.

Suppose there are two individuals arrested and booked in the jail on the same day in separate incidents. Suppose further that the two individuals are there for the same length of time and are "released" on the same day. However, because the offenses and charges are different, for one person, the release means they are freed and no longer in custody. For the other person, suppose the release means they were transferred to another authority as part of this person's legal problems. Suppose further, that person remains in custody until the legal process decides that he is to be incarcerated for two years, two-thirds of which he serves in a state prison until being released.

Then suppose both are again arrested and booked into the Stearns County Jail three years after they were initially booked. As far as Stearns County is concerned, the survival time is the same: they were released on the same day and were then booked again on the same day some three years later. However, the first person, in reality, had a far greater survival time because he was released to freedom while the second person was released to another authority and was, by comparison, unavailable to recidivate for a substantial period of time after that initial release from the custody of the Stearns County Jail.

These are two large precautions to bear in mind when viewing the following graphs in this chapter. In most descriptions of survival analysis, it is stated that the most fundamental functions, and the associated plots, are the survival function and curve, the probability density function and curve, and the hazard function and curve (Lee & Wang, 2003, pp. 8-9). All three are reproduced here through SAS using life table method, within the lifetest procedure, which is needed for the hazard and pdf plots.

What is notable about all three curves is the danger of recidivism is early. Relatively soon after a booking, the individual is at the greatest risk for another booking, thus creating a recidivism. In each graph, the highest point on the curve is the first point on the left. According to Schmidt and Witte (1988), the peak hazard rate for each cohort examined was 15 months and 6 months, there was some recidivism earlier on, recidivism increased very quickly, and then fell off drastically (p. 29). In that examination, recidivism was defined as being re-incarcerated in a North Carolina state prison (p. 9). Thus the definition of recidivism and the nature of the data was quite different, yet this is a basic similarity in the data.

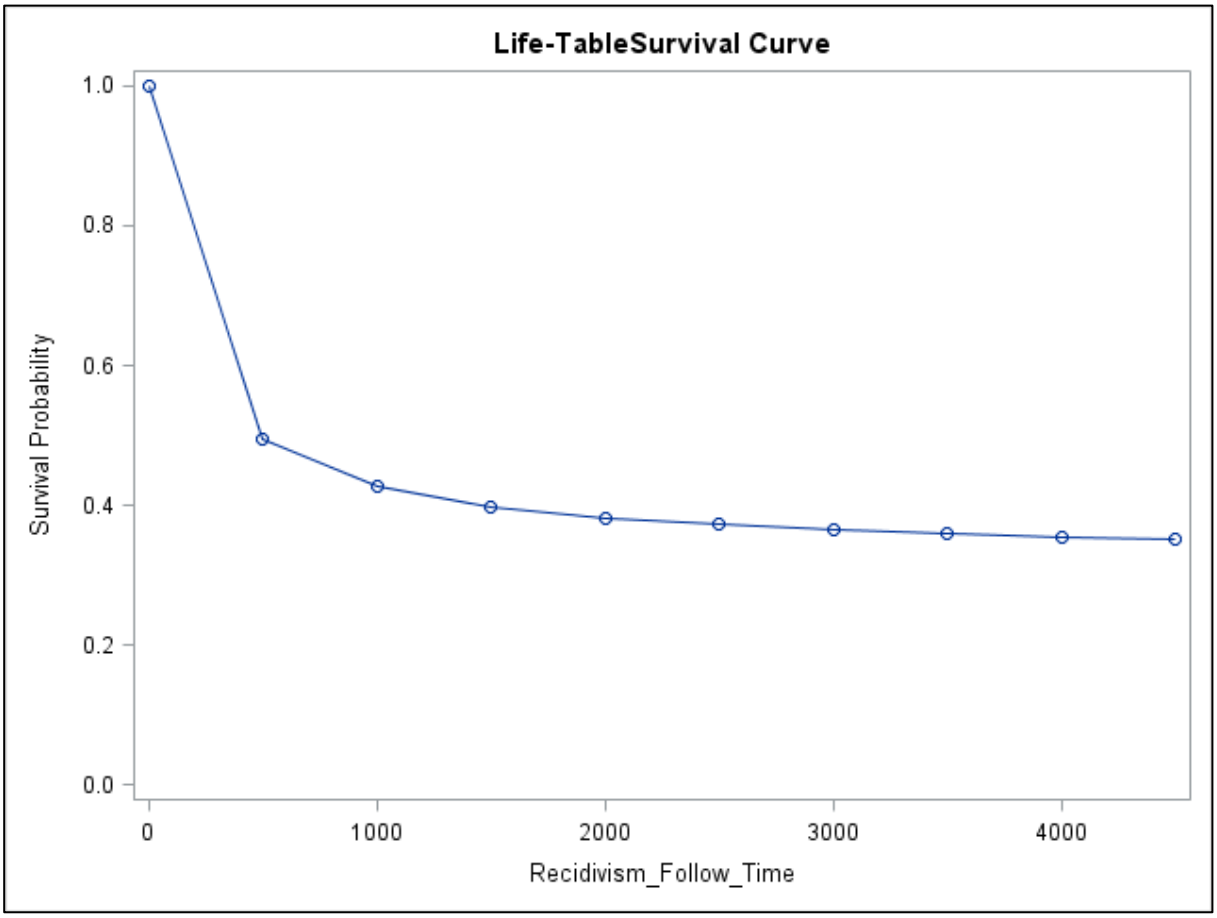


Figure 13: Survival Curve Calculated by Life Table Method in SAS

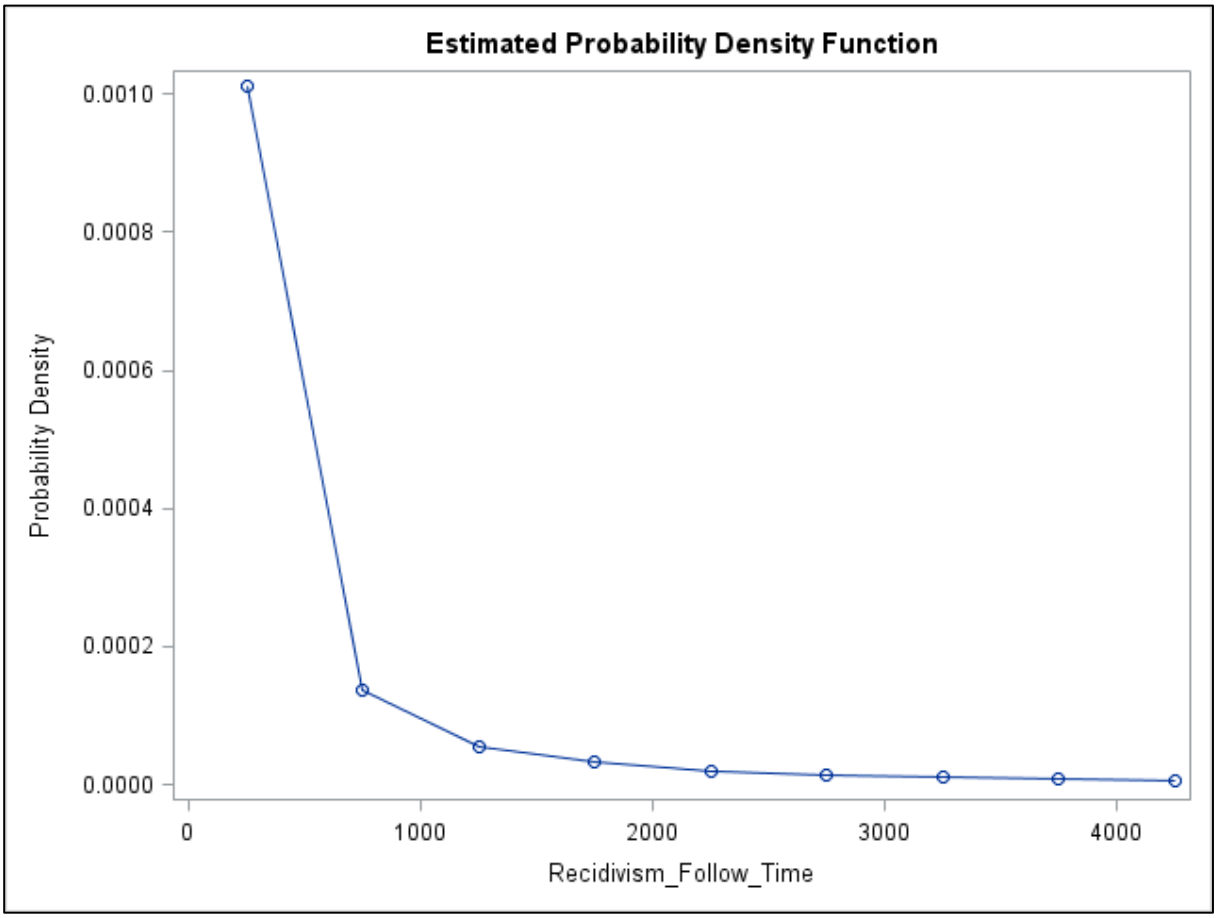
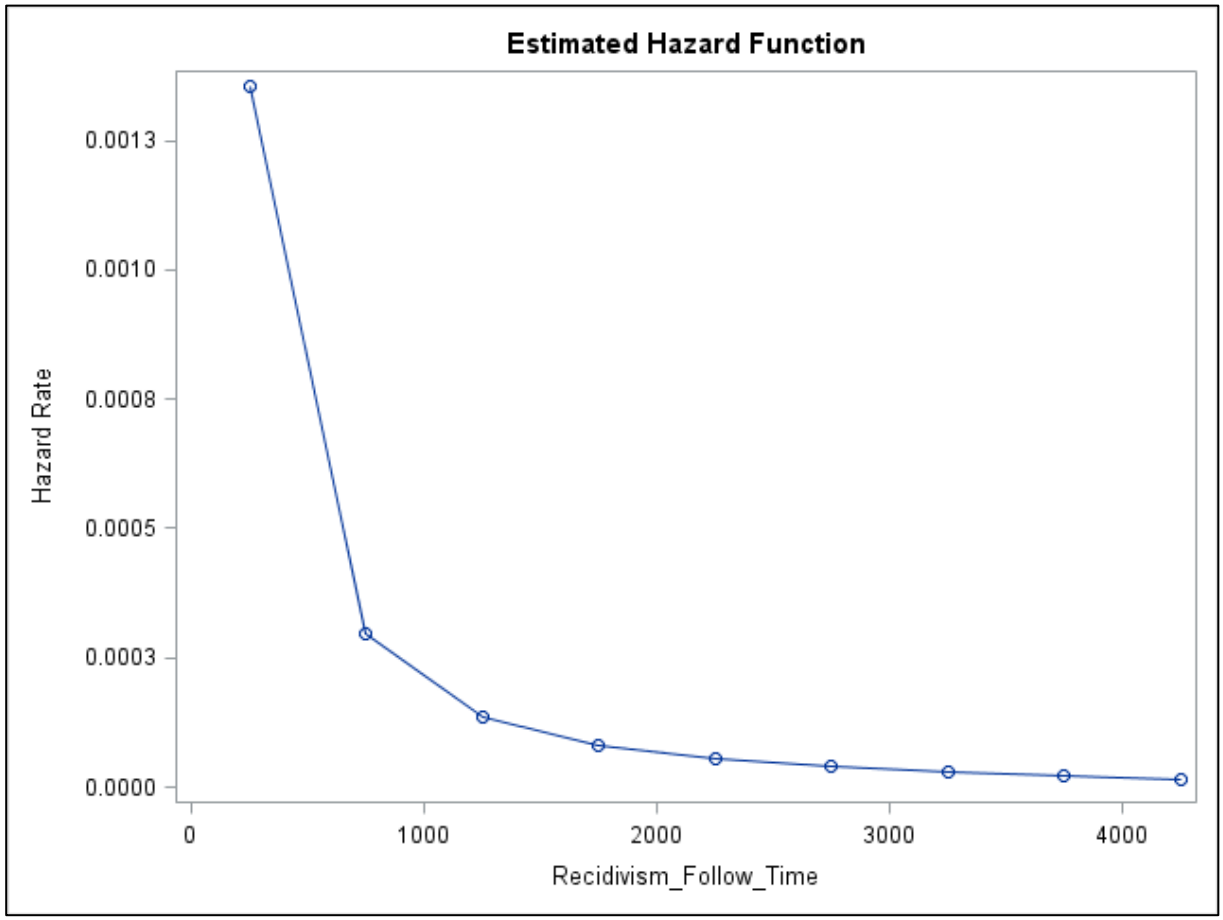


Figure 14: Probability Density Function Calculated by Life Table Method in SAS



Summary of the Number of Censored and Uncensored Values

Total	Failed	Censored	Percent Censored
88749	52141	36608	41.25

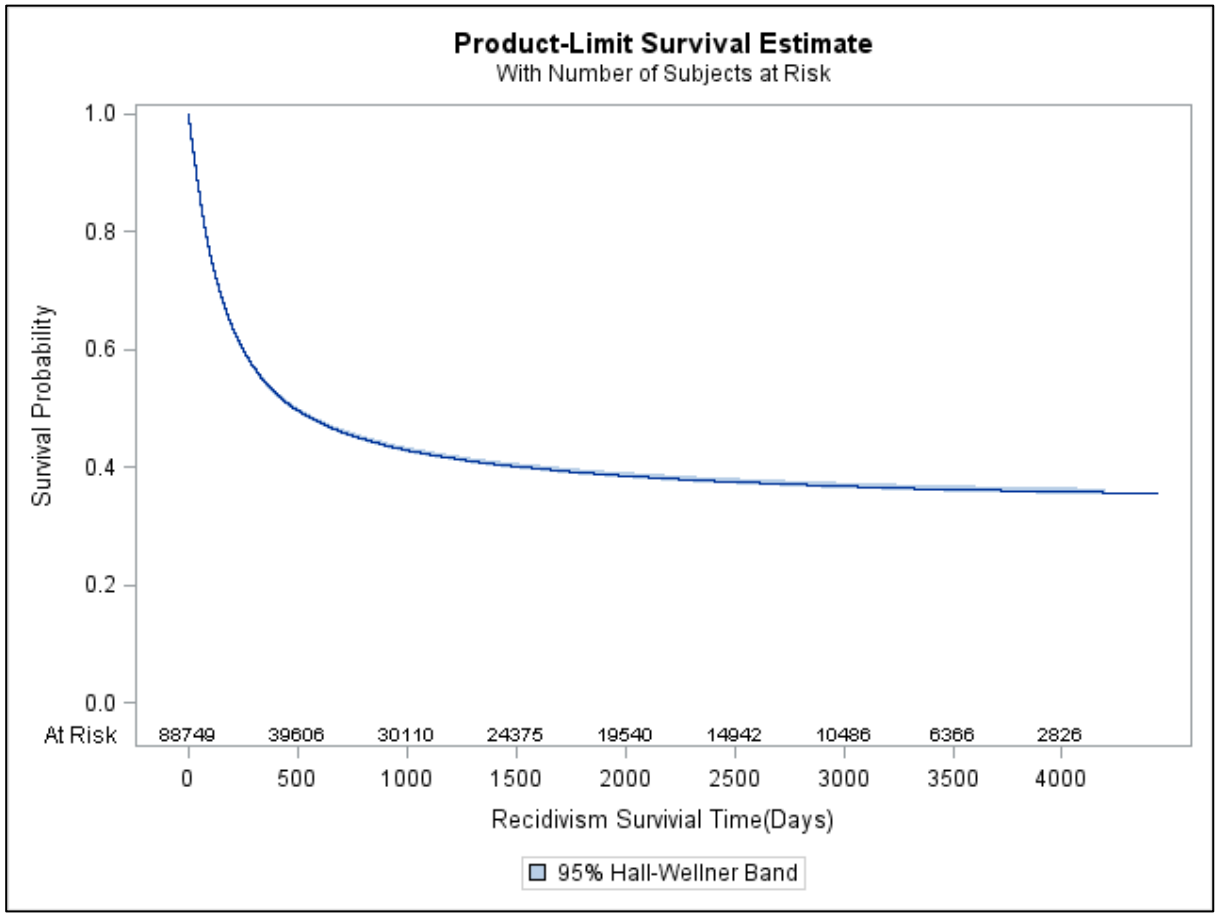
Note: 19 observations with invalid time or censoring values were deleted

Figure 15: Hazard Function Calculated by Life Table Method in SAS

What follows is a Kaplan-Meier (or product-limit) survival graph. It bears repeating that the variable Recidivism_Follow_Time contains both recidivism and follow time, and the dichotomous censor variable Censor_Flag determines which kind of time it is for a given record. A value of 0 for Censor_Flag means Recidivism_Follow_Time is a recidivism time while a value of 1 means it is a follow time. The time unit is days.

For the following graphs, the horizontal axis has tick marks every 500 days and shows the number of people at risk at those 500 day marks. The censored data is not shown with indicators on the graph because the number of records is so numerous that the censoring indicators all run together and are indistinguishable from each other, so the graphs were produced without showing censor markers. Also, the steps that one customarily sees in a Kaplan-Meier graph are not visible because of the volume of records being shown. Instead, the table below the graph will provide the information regarding the number of censored data records.

With the realization that the data reflects booking events and not individual people, the Kaplan-Meier plot shows a survival proportion of around 40%, meaning a recidivism rate of approximately 60%.



Summary of the Number of Censored and Uncensored Values

Total	Failed	Censored	Percent Censored
88749	52141	36608	41.25

Note: 19 observations with invalid time or censoring values were deleted.

Figure 16: Kaplan-Meier Survival Plot of Recidivism Time in Days

What follows are summary statistics for the survival time, obtained by using the non-parametric lifetest procedure on SAS. Calculation of the 75th percentile was troublesome and was left blank by the system, which is perhaps understandable considering that 75% are extremely unlikely to recidivate given the data and how recidivism is defined. It bears

repeating the variable Recidivism_Follow_Time expresses the time to event or follow time, in this case recidivism, as defined by being booked again, and is in days. The censoring variable, which tells SAS whether to regard the time variable as failure or follow time is Censor_Flag.

Table 3

Summary Statistics of Survival Time

Quartile Estimates

Percent	Point Estimate	95% Confidence Interval		
		Transform	[Lower	Upper)
75	.	LOGLOG	.	.
50	487.00	LOGLOG	474.00	501.00
25	106.00	LOGLOG	105.00	109.00

Mean	Standard Error
1767.46	6.70

Note: The mean survival time and its standard error were underestimated because the largest observation was censored and the estimation was restricted to the largest event time.

Chapter 12: Gender and the Booking Data

Gender is the only usable explanatory variable available in the data. The variable itself can be examined by the booking incidents or by the people who were booked in these incidents. As stated previously, the 88,768 booking incidents are by a total of 36,608 people. For the 88,768 booking incidents, 69,313 were by males and 19,455 were by females. This would give a 78.08% and 21.92% gender proportion by booking incident.

If one looks at the people involved, of the 36,608 people involved in these bookings, 27,280 were male and 9,328 were female. This would yield gender proportions of 74.52% and 25.48% by people involved in the bookings. The fact that males are higher proportion of booking incidents then males are of the people involved, 78.08% compared to 74.52%, will by itself illustrate recidivism is a worse problem for the males in the data than it is for the females.

Whether by bookings or by people, the data is decidedly male. However, much criminal and recidivism data is male-centric. Schmidt and Witte (1988, pp. 30-31) and Duwe and Kerschner (2007, p. 626) show males in the respective analyses as being represented in proportions of over 90%. The following output is a frequency table produced using SAS regarding gender, followed by a frequency table that takes censoring into account. Note that with the second table, where censoring is taken into account, 19 records are left out because of the negative values for survival time. This only applied to recidivism events.

Table 4

Gender Frequency Table

Gender				
Gender	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
F	19455	21.92	19455	21.92
M	69313	78.08	88768	100.00

Table 5

SAS Output for Frequency Table When Censoring is Considered

Summary of the Number of Censored and Uncensored Values					
Stratum	**Gender**	**Total**	**Failed**	**Censored**	**Percent Censored**
1	**F**	19450	10122	9328	47.96
2	**M**	69299	42019	27280	39.37
Total		88749	52141	36608	41.25

Note: 19 observations with invalid time, censoring, or strata values were deleted.

What is of interest with this aspect of the data is that the male-domination is "only" by a 78.08% to 21.92% proportion for bookings and by a 74.52% to 25.48% for the individuals. If one is used to the 90+ percentages that usually seem to describe male proportions in criminal justice matters, 78.08% can seem mildly surprising. However, this data represents bookings at a county jail as opposed to incarceration, which is the standard used in much recidivism analysis. It can be speculated, although there is nothing in the data that would reliably bear this out (recall that the MOC data is relatively inconsistent and

unreliable), that bookings is near the top of the aforementioned "funnel" while incarceration is at the bottom of this "funnel." It could be conjectured that perhaps women are less likely to be ultimately incarcerated following an arrest and booking, but that conjecture cannot be tested with this data. One way to evaluate this would be examine the seriousness of the MOC offenses for which the individuals are booked.

For example, below is a SAS output table that shows gender breakdown by booking number extension. The overall gender proportion in the booking numbers is 78.08% to 21.92% for the entire set of data. By looking at by booking number extension, one can see how the gender proportion changes according to booking number extension. For example, for the first booking, that is, for the extension -001, the male to female percentage is 73.87% to 26.13%. The first ten bookings are shown for each gender, and with each increase in the level of booking, the proportion of males to females grows slightly.

Now, one could conjecture that if males are more likely to be incarcerated following an arrest, then it should follow for a given level of booking, more males than females are being incarcerated and are no longer available for re-arrest. One can contend that if that is true, the proportion of males to females, if anything, should decrease as the booking level increases. One could counter this by contending that if males are simply more likely to recidivate (as is being defined here), this may more than overcome a greater likelihood to be incarcerated. The following table helps demonstrate the second view is more likely. Males are more likely to experience the higher levels of bookings at the Stearns County Jail.

Table 6

Table Showing Gender by Booking Extension

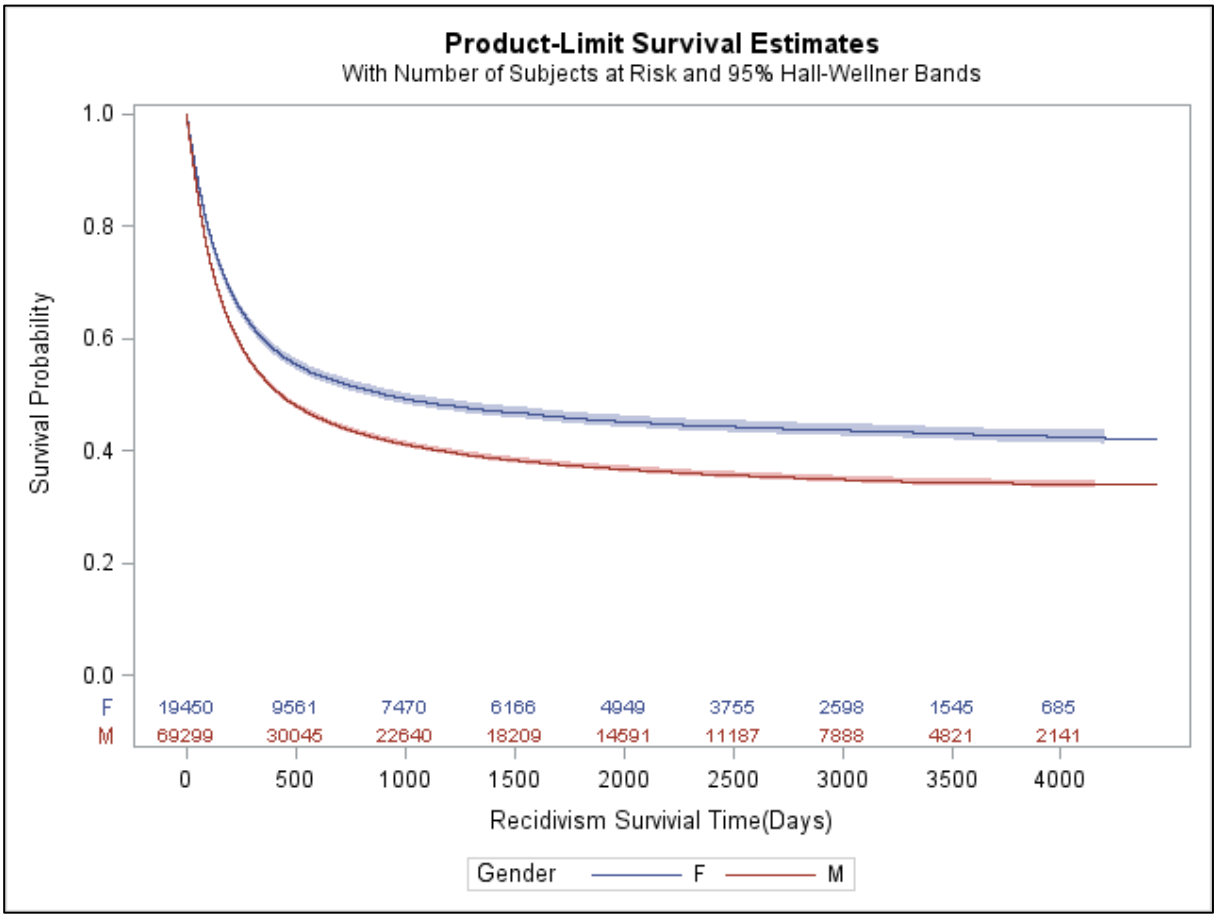
Table of Gender by Booking_3					
Gender	**Booking_3**	**Frequency**	**Percent**	**Row Percent**	**Column Percent**
F	**001**	8913	10.04	45.81	26.13
	002	3366	3.79	17.30	22.25
	003	1888	2.13	9.70	20.53
	004	1243	1.40	6.39	19.41
	005	904	1.02	4.65	19.14
	006	681	0.77	3.50	18.62
	007	509	0.57	2.62	17.85
	008	373	0.42	1.92	16.73
	009	299	0.34	1.54	16.30
	010	231	0.26	1.19	15.83
M	**001**	25202	28.39	36.36	73.87
	002	11763	13.25	16.97	77.75
	003	7308	8.23	10.54	79.47
	004	5161	5.81	7.45	80.59
	005	3820	4.30	5.51	80.86
	006	2976	3.35	4.29	81.38
	007	2342	2.64	3.38	82.15
	008	1857	2.09	2.68	83.27
	009	1535	1.73	2.21	83.70
	010	1228	1.38	1.77	84.17

If one thinks of a ratio of male percentage of bookings over female percentage of bookings, for example, the first booking yields a gender ratio of 28.39/10.04 = 2.83. For the second booking, this gender ratio becomes, 13.25/3.79 = 3.50. For the tenth booking, this ratio grows to 1.38/0.26 = 5.31.

Previously, a Kaplan-Meier graph was presented which presented the survival of the data as a whole. The following Kaplan-Meier graph separates the two genders. In addition,

some non-parametric tests comparing the differences in survival time of the two genders are also presented. As has been stated many times previously, there is a lack of explanatory variables that accompanies this booking data. The only variable that can fulfill this role is that of gender. Therefore, with the same cautions as given previously--as to the records being undifferentiated by individuals, meaning the bookings are treated as separate entities, along with the problems with the release variable—the same Kaplan-Meier graphs are done, but now separate curves for each gender.

This Kaplan-Meier graph offers support to the idea that recidivism, in general, is a worse phenomenon for males as opposed to females. Criminal activity and recidivism have consistently been a worse problem for males (National Research Council, 1986, p. 24). What is of interest are how similar the shapes are for each gender. In fact, it looks almost as if the survival curve for females was shifted downward. Critical opinions were expressed in the literature review about the use of the proportional hazards model applied to explanatory variables with respect to recidivism, especially regarding age. This graph could, if viewed in isolation, support a view of gender, or more specifically, being male, as a proportional hazard. The reason is that one could look at the plot and see the gender curves are roughly parallel. On the other hand, one might also say the plot shows the gap between the genders as widening one as time goes by. This plot does not offer strong support to the view that gender is a proportional hazard as far as this data is concerned. That would require data that would show the problem of recidivism and gender over time. Perhaps an accelerated failure model is more applicable.



Summary of the Number of Censored and Uncensored Values					
Stratum	**Gender**	**Total**	**Failed**	**Censored**	**Percent Censored**
1	**F**	19450	10122	9328	47.96
2	**M**	69299	42019	27280	39.37
Total		88749	52141	36608	41.25

Note: 19 observations with invalid time, censoring, or strata values were deleted.

Figure 17: Kaplan-Meier Survival Plot of Recidivism Time in Days, Comparing Genders

Testing Homogeneity of Survival Curves for Recidivism_Follow_Time over Strata

Rank Statistics		
Gender	**Log-Rank**	**Wilcoxon**
F	-1994.1	-1.156E8
M	1994.1	1.1556E8

Covariance Matrix for the Log-Rank Statistics		
Gender	**F**	**M**
F	9280.58	-9280.58
M	-9280.58	9280.58

Covariance Matrix for the Wilcoxon Statistics		
Gender	**F**	**M**
F	3.547E13	-3.55E13
M	-3.55E13	3.547E13

Test of Equality over Strata			
Test	**Chi-Square**	**DF**	**Pr > Chi-Square**
Log-Rank	428.4708	1	<.0001
Wilcoxon	376.5341	1	<.0001
-2Log(LR)	750.5198	1	<.0001

Figure 18: Non-parametric Statistical Tests Comparing Genders

As can be seen above, after the Kaplan-Meier graph representing the genders, there is SAS output that provides the results of the log-rank test, which is an applicable non-parametric test of survival functions in the presence of observations that are right-skewed and have censoring. Also, the survival functions for each gender do not cross, which would make it difficult for the log-rank test (Lee & Wang, 2003, pp. 111-112, 119-120). Although a hypothesis test of the recidivism difference between the two genders has not been formally described here, as can be seen by the results of the log-rank test, gender is significant factor.

The Kaplan-Meier graphs that show the different genders are not differentiated into people. These curves represent booking events separated by gender. Recall that there are 27,280 males and 9,328 females. One way to get at the genders in the form of people rather than booking events is to focus on booking levels. If the focus is on booking numbers that end in the extension -001, the booking numbers will indicate individuals. What follows is a graph that makes use of a dummy variable Booking_001. This variable assigns a value of 1 to a booking number that ends in the extension -001 and a value of 0 to all other booking numbers. At the same time, each booking level is divided by gender.

The resulting graph is extraordinary. The top two curves are individuals who have first bookings, separated by gender. The bottom two curves are booking events that end in extensions other than -001, separated by gender. Thus, the top curves are individuals in their first booking, the bottom two curves are booking events.

The first thing that is noticed is the similarity in the shape of the curves. It appears as if one curve has been shifted into the respective positions. The second aspect that is noticed

how the booking levels seem to form two separate prongs where the genders further differentiate each prong of the booking level curves. In each case, the females have the curve with the higher survival rate, the males have the curve with the lower survival rate.

The major differentiation seems to be based on booking level. The highest survival rate is for women who have their first booking. The final survival rate achieved at the right of the graph appears to over 60%. Men who have had their first booking seem to have a final survival rate of approximately 55%. Females in booking events beyond the -001 level have a survival rate of approximately 30%, while males in booking events beyond the -001 level appear to have a final survival rate of approximately 25%.

It would appear from the graph that there are two factors—booking level and gender--that operate within the group of people who are being booked. One is a personal characteristic, as if being booked makes it increasingly more likely to get booked again. This is an aspect that will be explored in the next chapter. Also, gender appears to make it more likely to someone will recidivate. Of the four groups seen. The least likely to recidivate are females who have been booked once, the most likely to recidivate are males who have been booked more than once. The issue of gender will continue to arise in subsequent chapters.

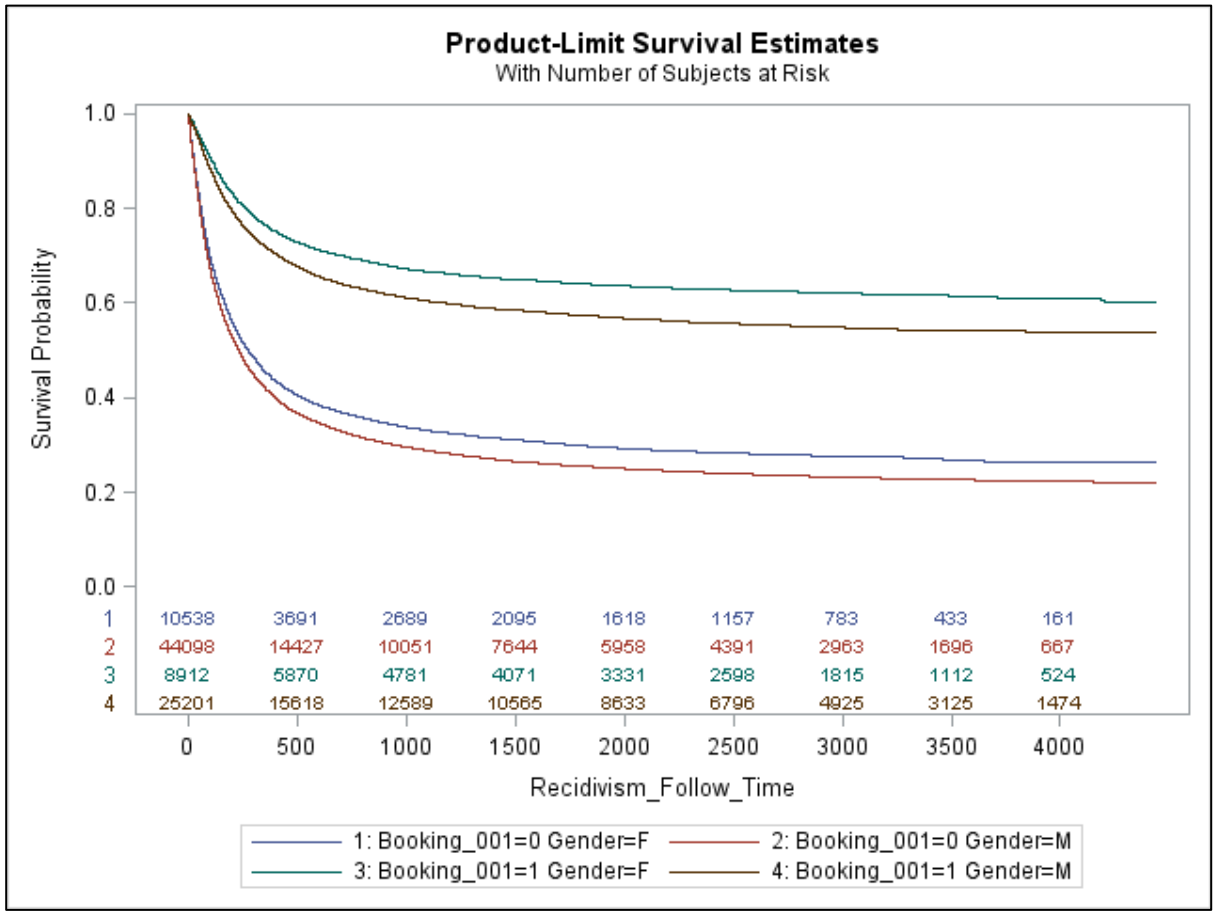


Figure 19: Graph Separating the Data into Booking -001 Level and Gender

Chapter 13: Further Explorations of Gender, Comparisons with BJS Study

This chapter begins with another view of the direct comparison of the recidivism rate of the respective genders. The differing recidivism experiences of the genders has been discussed previously, but the data will be organized into time intervals. Tables 7-15 presents SAS output that will be summarized by Tables 16 and 17. Because of the limitations of SAS, this raw output was not easily readable, therefore, it was put directly into Excel tables (Tables 7-15) so it could be formatted and organized into a more readable form. However, even with the formatting changes, the time intervals are not in an easily understandable form. Therefore, Tables 16 and 17 are purely Excel tables that presents the SAS output of Tables 7-15 in a succinct form with time intervals that are easier to read and understand, along with some results from a Bureau of Justice Statistics study on recidivism that was completed in 2014.

The time period and time intervals of interest was six month intervals over a five year period after "release" by the Stearns County Jail. The final summaries offered by Tables 16 and 17 present the recidivism rate at certain points in time after release: six months, one year, two years, three years, four years, and five years. Both genders, different booking levels, as well as the overall figures are presented.

The first view is that of the booking data as a whole divided into the genders, which is from the perspective of booking events and not individuals. This is presented in two sets of SAS output, presented in Tables 7 and 8.

The second view divides the data into four parts: female and first booking, male and first booking, female and beyond first booking, and male and beyond first booking. This is presented in Tables 9-12. It bears repeating that "First Booking" incidents means that the booking number had the extension -001. Thus, these are separate individuals. What is meant by the phrase "Beyond First Booking," is that the booking number has an extension of something other than -001, and therefore does not represent separate individuals.

Note that the SAS output presented in Table 7-15 uses six month intervals for the five year period, but is given in day quantities. The starting point is that a year is 365.25 days. This means that a six month interval is taken to be 182.625 days, while the year-and-a-half translates into 547.875 days. Two years is 730.50 days, three years is 1095.75 days, four years is 1461.00 days, and five years is 1826.25 days. The final two table summaries provide the important time divisions in the more easily understandable forms: six months, one year, two years, three years, four years, and five years.

What is striking, once again, is the apparently consistent separation in the proportions that recidivate, based on gender and booking level. As stated earlier, it is as if gender and booking level are two factors at work at the individual level. The question is again raised if gender can be considered a proportional hazard. If one examines the two summarizing tables at the end of this chapter, the difference in the male and female proportions tends to get larger, not quite as consistent of a separation as it might first appear. An accelerated failure model may be more applicable.

Booking level, it would appear at least initially, cannot be viewed as an inherent factor that remains unchanged over time, as would be demanded by the proportional hazards view. As can be seen in the preceding graphs, it seems booking level would appear to be a dynamic factor where each successive booking seems to affect the likelihood of being booked again. The alternative would be to view individuals as inherently having one, two, or some number of bookings "in them" that is revealed by the booking history as time goes on. Either alternative would seem to be too dynamic to view as simply a "proportional hazard" at work.

This data only listed bookings into the Stearns County Jail. Because of necessity, the data could only support a definition of recidivism that involved being booked again into the Stearns County Jail. The majority of recidivism studies appear to use a release from incarceration as a starting point, and either re-arrest, re-conviction, or re-incarceration as defining recidivism.

The most recent authoritative study of recidivism in the US was mentioned previously. It was the report issued by the Bureau of Justice Statistics in 2014 by Durose, Cooper, and Snyder. This examination used all three of the definitions of recidivism mentioned previously: re-arrest, re-conviction, and re-incarceration. The closest comparison to this data would be using re-arrest. Recall that booking into the Stearns County Jail usually meant an arrest had taken place, but it did not have to mean an arrest had occurred. A booking could also have resulted from a court appearance where the

person was summoned to appear in court, the result of which meant the person was taken to the jail from the court and booked.

The results of the survey done by Durose et al. (2014) are presented in Table 16 for direct comparison with the results from this data. For both females and males, the recidivism rate was higher for Stearns County for the first two years and then got progressively worse for the BJS study beyond the two year mark.

For several reasons, these comparisons must be approached with caution. Any study of recidivism is highly dependent on data and definition. For the Durose et al. (2014) BJS study, the starting point (for the lifetime that could be ended by a recidivism event) was much clearer than is possible with this data. For that study it was release from incarceration (p. 1), whereas for the Stearns County booking data, the starting event was the ambiguous "release date" which has been discussed previously. With the BJS study, they were able to follow individuals to where they had been arrested in many different jurisdictions (p. 3). With the Stearns County booking data, the bookings are only in regard to the Stearns County Jail. An individual could easily have been booked into another jail elsewhere, but it would not be noted in this data.

Also, note that since the study by Durose et al. (2014) uses release from incarceration as the starting point (p. 1), it would imply that the initial offenses of these individuals were bad enough to warrant incarceration. With the data used here, the presence of an individual in the data is simply being booked into the county jail. Arrest or booking does not necessarily imply the person was later incarcerated. By perusing the Statute_Description

variable values, one can see offenses such as traffic offenses, underage drinking, or loud parties. It can be conjectured that offenses such as these are unlikely to have resulted in incarceration. Thus, the pool of individuals in each group is likely quite different, with the individuals in the study by Durose et al. (2014) being closer to the "hardened criminal" that most would think comprise a prison population.

With that in mind, why would the recidivism rates, as shown on Table 16, be worse for the first 2 years for the Stearns County Jail individuals? That would bring up the most profound difference of all when comparing the two groups of people. The data used here, unless the focus is on booking number extension, is based on booking events. The data used by Durose et al. (2014) is based on individuals: it begins with an individual being released from an incarceration, and ends with the re-arrest, reconviction, or re-incarceration of that person, or the end of the observation window. In the case of the Durose et al. (2014) study, that amounted to five years (p. 1).

To summarize, comparisons of these results with other studies of recidivism can certainly be made, but only with great awareness of the differences in data and definition. Ironically, the study by Durose et al. (2014) issued a cautionary note about comparison in their work with a seemingly similar study of recidivism done in 1994 by the Bureau of Justice Statistics. They stated clearly that "direct comparisons between the published recidivism statistics should not be made," with the major reasons being the difference in states covered by the studies in addition to some demographic differences between the people comprising the two different studies (p. 2). The difference in data and definition between the Stearns

County booking data used here and the 2014 BJS study are much more profound than the differences between the 1994 and 2014 BJS studies.

Despite the note of caution in the Durose et al. (2014) study, the case can be made that comparisons are worth making between different examinations of recidivism. The reason is that fruitful discussions of the differences in data and definition can take place, but there always needs to be an awareness of these differences.

Tables 7-15 represent SAS output that is included for completeness. The SAS output was copied into Excel tables so it could be formatted and made more presentable. The important information on Tables 7-15 is summarized in a more succinct and readable form on Tables 16 and 17. Results from the BJS study by Durose et al. (2014) are included in Table 16 for comparison. Table 17 contains much information for its size. Overall recidivism proportions are presented along with that of booking levels -001, -002, and beyond booking -002.

Table 7

SAS Output Put Into Excel Table for Overall Female Recidivism in Six Month Intervals for Five Years--Summarized in Table 16

LowerTime	UpperTime	Failed	Failure
0.000	182.625	5727.000	0.000
182.625	365.250	1986.000	0.299
365.250	547.875	839.000	0.408
547.875	730.500	431.000	0.456
730.500	913.125	306.000	0.482
913.125	1095.750	195.000	0.501
1095.750	1278.375	141.000	0.514
1278.375	1461.000	103.000	0.524
1461.000	1643.625	89.000	0.531
1643.625	1826.250	72.000	0.538
1826.250	.	233.000	0.544

Table 8

SAS Output Put Into Excel Table for Overall Male Recidivism in Six Month Intervals for Five Years--Summarized in Table 16

LowerTime	UpperTime	Failed	Failure
0.000	182.625	24452.000	0.000
182.625	365.250	7839.000	0.358
365.250	547.875	3376.000	0.477
547.875	730.500	1889.000	0.530
730.500	913.125	1162.000	0.561
913.125	1095.750	787.000	0.582
1095.750	1278.375	592.000	0.596
1278.375	1461.000	422.000	0.607
1461.000	1643.625	312.000	0.616
1643.625	1826.250	259.000	0.622
1826.250	.	929.000	0.628

Table 9

SAS Output Put Into Excel Table for Booking -001 Female Recidivism in Six Month Intervals for Five Years--Summarized in Table 17

LowerTime	UpperTime	Failed	Failure
0.000	182.625	1383.000	0.000
182.625	365.250	695.000	0.158
365.250	547.875	327.000	0.240
547.875	730.500	181.000	0.280
730.500	913.125	134.000	0.303
913.125	1095.750	89.000	0.321
1095.750	1278.375	58.000	0.334
1278.375	1461.000	47.000	0.342
1461.000	1643.625	35.000	0.349
1643.625	1826.250	30.000	0.355
1826.250	.	111.000	0.360

Table 10

SAS Output Put Into Excel Table for Booking -001 Male Recidivism in Six Month Intervals for Five Years--Summarized in Table 17

LowerTime	UpperTime	Failed	Failure
0.000	182.625	4769.000	0.000
182.625	365.250	2254.000	0.192
365.250	547.875	1119.000	0.285
547.875	730.500	666.000	0.333
730.500	913.125	403.000	0.363
913.125	1095.750	291.000	0.382
1095.750	1278.375	210.000	0.396
1278.375	1461.000	137.000	0.407
1461.000	1643.625	113.000	0.414
1643.625	1826.250	111.000	0.420
1826.250	.	395.000	0.427

Table 11:

SAS Output Put Into Excel Table for Female Recidivism for Booking Levels Greater Than -001 in Six Month Intervals for Five Years--Summarized in Table 17

LowerTime	UpperTime	Failed	Failure
0.000	182.625	4344.000	0.000
182.625	365.250	1291.000	0.420
365.250	547.875	512.000	0.552
547.875	730.500	250.000	0.607
730.500	913.125	172.000	0.635
913.125	1095.750	106.000	0.656
1095.750	1278.375	83.000	0.669
1278.375	1461.000	56.000	0.680
1461.000	1643.625	54.000	0.688
1643.625	1826.250	42.000	0.696
1826.250	.	122.000	0.703

Table 12

SAS Output Put Into Excel Table for Male Recidivism for Booking Levels Greater Than -001 in Six Month Intervals for Five Years--Summarized in Table 17

LowerTime	UpperTime	Failed	Failure
0.000	182.625	19683.000	0.000
182.625	365.250	5585.000	0.452
365.250	547.875	2257.000	0.587
547.875	730.500	1223.000	0.644
730.500	913.125	759.000	0.676
913.125	1095.750	496.000	0.697
1095.750	1278.375	382.000	0.712
1278.375	1461.000	285.000	0.724
1461.000	1643.625	199.000	0.733
1643.625	1826.250	148.000	0.740
1826.250	.	534.000	0.746

Table 13

SAS Output Put Into Excel Table for Overall Recidivism for Only Booking Level -001 in Six Month Intervals for Five Years--Summarized in Table 17

LowerTime	UpperTime	Failed	Failure
0.000	182.625	6152.000	0.000
182.625	365.250	2949.000	0.183
365.250	547.875	1446.000	0.273
547.875	730.500	847.000	0.320
730.500	913.125	537.000	0.348
913.125	1095.750	380.000	0.366
1095.750	1278.375	268.000	0.380
1278.375	1461.000	184.000	0.390
1461.000	1643.625	148.000	0.397
1643.625	1826.250	141.000	0.403
1826.250	.	506.000	0.410

Table 14

SAS Output Put Into Excel Table for Overall Recidivism for Only Booking Level -002 in Six Month Intervals for Five Years--Summarized in Table 17

LowerTime	UpperTime	Failed	Failure
0.000	182.625	4419.000	0.000
182.625	365.250	1831.000	0.296
365.250	547.875	825.000	0.423
547.875	730.500	417.000	0.482
730.500	913.125	287.000	0.513
913.125	1095.750	181.000	0.536
1095.750	1278.375	150.000	0.550
1278.375	1461.000	119.000	0.563
1461.000	1643.625	88.000	0.574
1643.625	1826.250	52.000	0.583
1826.250	.	247.000	0.588

Table 15

SAS Output Put Into Excel Table for Overall Recidivism for Booking Levels Beyond -002 in Six Month Intervals for Five Years--Summarized in Table 17

LowerTime	UpperTime	Failed	Failure
0.000	182.625	19608.000	0.000
182.625	365.250	5045.000	0.504
365.250	547.875	1944.000	0.641
547.875	730.500	1056.000	0.696
730.500	913.125	644.000	0.728
913.125	1095.750	421.000	0.749
1095.750	1278.375	315.000	0.763
1278.375	1461.000	222.000	0.775
1461.000	1643.625	165.000	0.783
1643.625	1826.250	138.000	0.790
1826.250	.	409.000	0.796

Table 16

Summarizing Table of Gender Differences in Booking Events at Different Times After "Release" and Results from BJS Recidivism Study from 2014

Cumulative Recidivism Proportion	6 Months	1 Year	2 Years	3 Years	4 Years	5 Years
Stearns County Female	0.299	0.408	0.482	0.514	0.531	0.544
Stearns County Male	0.358	0.477	0.561	0.596	0.616	0.628
Difference (Male - Female)	0.058	0.069	0.080	0.082	0.085	0.084
Stearns County All	0.345	0.462	0.544	0.578	0.597	0.610
BJS 2014 Study for Females	0.221	0.344	0.498	0.585	0.639	0.681
BJS 2014 Study for Males	0.289	0.445	0.607	0.690	0.741	0.776
BJS Difference (Male - Female)	0.068	0.101	0.109	0.105	0.102	0.095
BJS All	0.282	0.434	0.595	0.678	0.730	0.766

Table 17

Summarizing Table of Gender Differences for Booking Level for the Different Time Intervals, Different Booking Levels for the Different Time Intervals

Cumulative Recidivism Proportion	6 Months	1 Year	2 Years	3 Years	4 Years	5 Years
Female and First Booking	0.158	0.240	0.303	0.334	0.349	0.360
Male and First Booking	0.192	0.285	0.363	0.396	0.414	0.427
Difference (Male - Female)	0.034	0.046	0.060	0.062	0.065	0.067
Female and Beyond First Booking	0.420	0.552	0.635	0.669	0.688	0.703
Male and Beyond First Booking	0.452	0.587	0.676	0.712	0.733	0.746
Difference (Male - Female)	0.033	0.035	0.041	0.043	0.045	0.043
Booking -001 Overall	0.183	0.273	0.348	0.380	0.397	0.410
Booking -002 Overall	0.296	0.423	0.513	0.550	0.574	0.588
Beyond Booking -001 and -002	0.504	0.641	0.728	0.763	0.783	0.796

Chapter 14: Booking Levels and Recidivism

If the data is examined only for those with a first booking, with the booking number extension of -001, then it is possible to view the bookings as separate individuals. More specifically, in regard to the data as it is configured, all bookings that end with the -001 extension can be accessed using the dummy variable Booking_001, which assigned a value of 0 to booking numbers that did not have this extension and a value of 1 to booking numbers that did have -001 as an extension. Similar dummy variables were created for the second through the fifth bookings: Booking_002, Booking_003, Booking_004, and Booking_005.

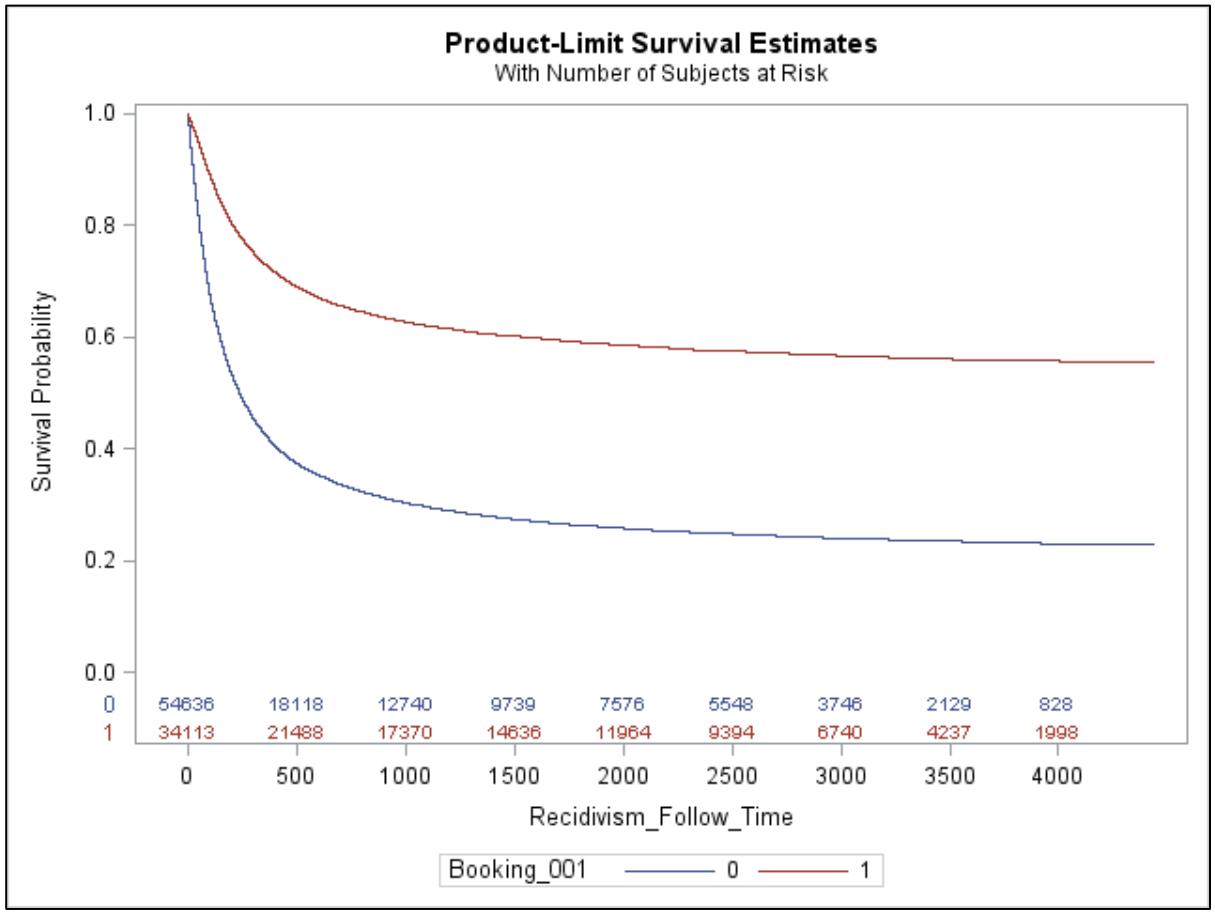
A recidivism event for the first booking would be created by a second booking, or booking number that had an extension of -002. The focus is first on all individuals who had first bookings, which had a booking number ending in -001, and whether or not that person recidivated by having another booking with a booking number extension of -002. This is very different from the previous way the booking data was viewed. In this way, the bookings are in fact separate people. The same Kaplan-Meier graph was therefore constructed using SAS and are reproduced below.

Following this presentation, the similar procedure is done for the second booking, through using the variable Booking_002. That is, booking records that had an extension -002, indicating that it was the second booking. If such a record had a recidivism event, that meant there was a subsequent booking, or one with a booking number extension of -003.

One difference between the bookings ending in -001 compared to -002 is that given the existence of an -002 booking number, the -001 booking may not appear in the data if

that first booking occurred before 1/1/2003. The same is true of all booking levels beyond -001. Someone may have been booked several times prior to 1/1/2003, therefore the first booking that appears in the data will not have the -001 extension. In fact, the dummy variable called, Previous_Booking, was designed to track bookings like this. It examines the first occurrence of the individual booking number, that is, the first seven figures of the full booking number. If that first occurrence does not have an extension ending in -001, the dummy variable was assigned a value of 1. There were 2,493 such bookings out of the total of 88,768.

The data, in this presentation, can be examined for this flattening of the Kaplan-Meier curve on the right. The upper curve, which represents separate people under the guise of a booking number that ends in -001, does indeed appear to be flattening out on the right at a survival proportion of approximately 60%, meaning a recidivism rate of approximately 40%. The surprise here is that the curves are very similar in shape. The bottom curve does not describe individuals in the way the top curve does. It is comprised of bookings that end in something other than -001, in other words, all bookings that do not end in -001. This second curve can be viewed as the previous show survival curve for the entire group of bookings with all bookings ending in -001 removed. The recidivism is much worse when the first bookings are removed. The Kaplan-Meier survival curve for this group seems to approach 20%, meaning about 80% are failing, or recidivating. Non-parametric tests of this separation are also shown.



Summary of the Number of Censored and Uncensored Values					
Stratum	**Booking_001**	**Total**	**Failed**	**Censored**	**Percent Censored**
1	**0**	54636	38583	16053	29.38
2	**1**	34113	13558	20555	60.26
Total		88749	52141	36608	41.25

Note: 19 observations with invalid time, censoring, or strata values were deleted.

Figure 20: Kaplan-Meier Survival Plot of Recidivism Time in Days, Comparing First Bookings with Non-first Bookings

Testing Homogeneity of Survival Curves for Recidivism_Follow_Time over Strata

Rank Statistics		
Booking_001	**Log-Rank**	**Wilcoxon**
0	11265	6.8779E8
1	-11265	-6.878E8

Covariance Matrix for the Log-Rank Statistics		
Booking_001	**0**	**1**
0	12785.0	-12785.0
1	-12785.0	12785.0

Covariance Matrix for the Wilcoxon Statistics		
Booking_001	**0**	**1**
0	4.936E13	-4.94E13
1	-4.94E13	4.936E13

Test of Equality over Strata			
Test	**Chi-Square**	**DF**	**Pr > Chi-Square**
Log-Rank	9926.1897	1	<.0001
Wilcoxon	9584.2302	1	<.0001
-2Log(LR)	20006.7446	1	<.0001

Figure 21: Non-parametric Statistical Tests for Comparing First Bookings to Non-first Bookings

The same statistical tests as used previously are presented here, and are decisive, but not as informative as the Kaplan-Meier graph of the data. Also, the shape of the graph is remarkably similar to the graphs shown previously, even though the data assumptions has changed rather profoundly. This is an interesting result.

If we continue this line of thought, here is a Kaplan-Meier graph that shows first booking, second booking, and all other bookings. This is done by using the dummy variables Booking_001 and Booking _002. The graph shows the curves of the booking records that have a value of 1 for the variable Booking_001, booking records that have a value of 1 for Booking_002, and all other booking records.

The striking feature of this graph is similarity in shape of the curves. This is true even though the definition of recidivism for the bottom curve is very different than for the top two curves. The bottom curve is all bookings that have extension numbers other than -001 or -002, which means it is composed of booking events instead of individuals. The survival proportion for the curve at the right extreme is approximately 60% for the first booking, while the survival rate is approximately 40% for the second booking, and about 20% for bookings beyond the second booking. The graph is followed by non-parametric tests of equality and homogeneity.

Following this graph, subsequent graphs show the Kaplan-Meier graphs show more booking levels. Each graph adds a booking level. The last graph shows the first five booking levels, that is, booking numbers with extensions of -001, -002, -003, -004, and all bookings beyond -004.

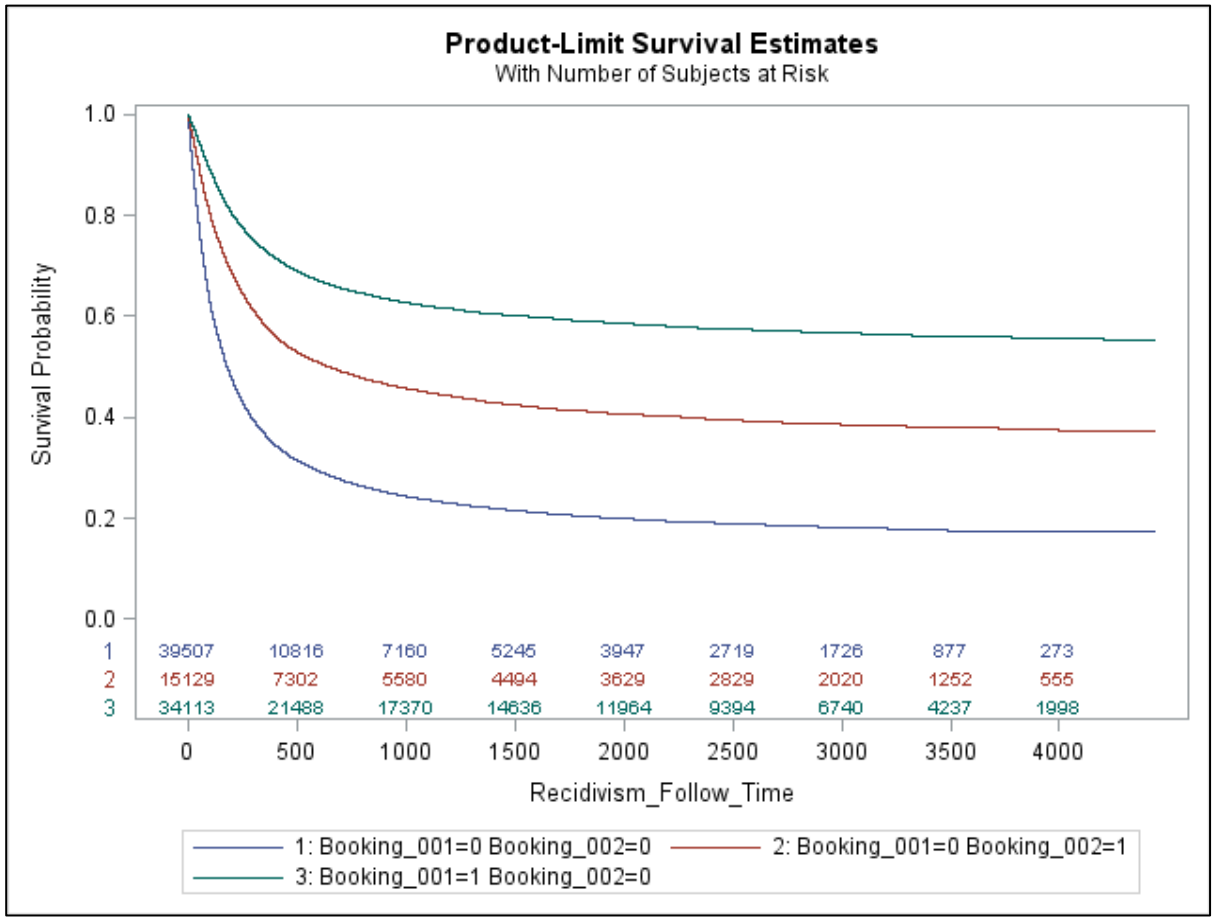


Figure 22: Kaplan-Meier Graph Showing Bookings -001, -002, and Other Levels

Summary of the Number of Censored and Uncensored Values						
Stratum	**Booking_001**	**Booking_002**	**Total**	**Failed**	**Censored**	**Percent Censored**
1	**0**	**0**	39507	29967	9540	24.15
2	**0**	**1**	15129	8616	6513	43.05
3	**1**	**0**	34113	13558	20555	60.26
Total			88749	52141	36608	41.25

Note: 19 observations with invalid time, censoring, or strata values were deleted.

Rank Statistics		
Stratum	**Log-Rank**	**Wilcoxon**
1	12082	7.5601E8
2	-817	-6.822E7
3	-11265	-6.878E8

Covariance Matrix for the Log-Rank Statistics			
Stratum	**1**	**2**	**3**
1	11503.6	-3215.3	-8288.3
2	-3215.3	7712.0	-4496.7
3	-8288.3	-4496.7	12785.0

Test of Equality over Strata			
Test	**Chi-Square**	**DF**	**Pr > Chi-Square**
Log-Rank	13651.8418	2	<.0001
Wilcoxon	13060.5580	2	<.0001
-2Log(LR)	27047.9733	2	<.0001

Figure 23: Non-parametric Tests of Booking Level Strata -001, -002, and Beyond

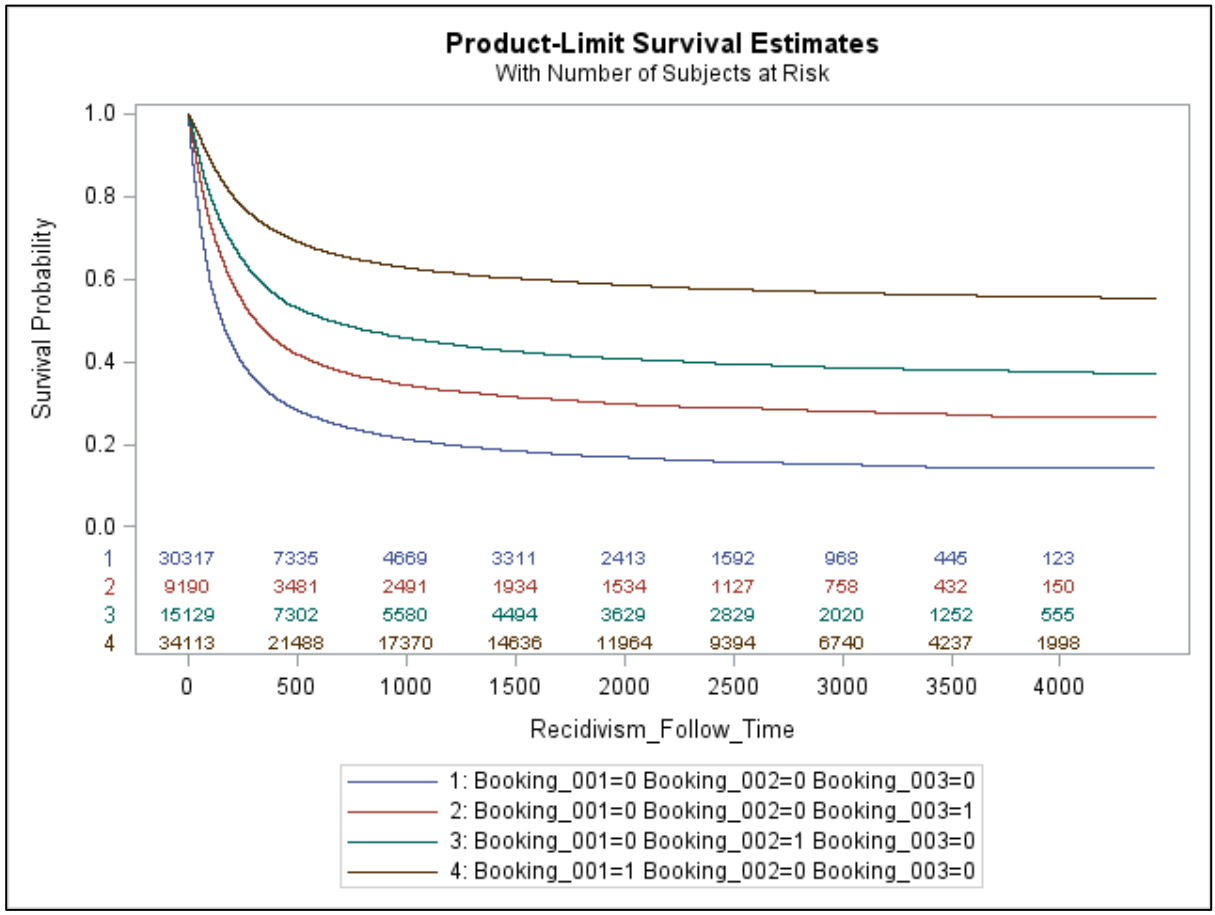


Figure 24: Kaplan-Meier Graph Showing Bookings -001, -002, -003, and Others

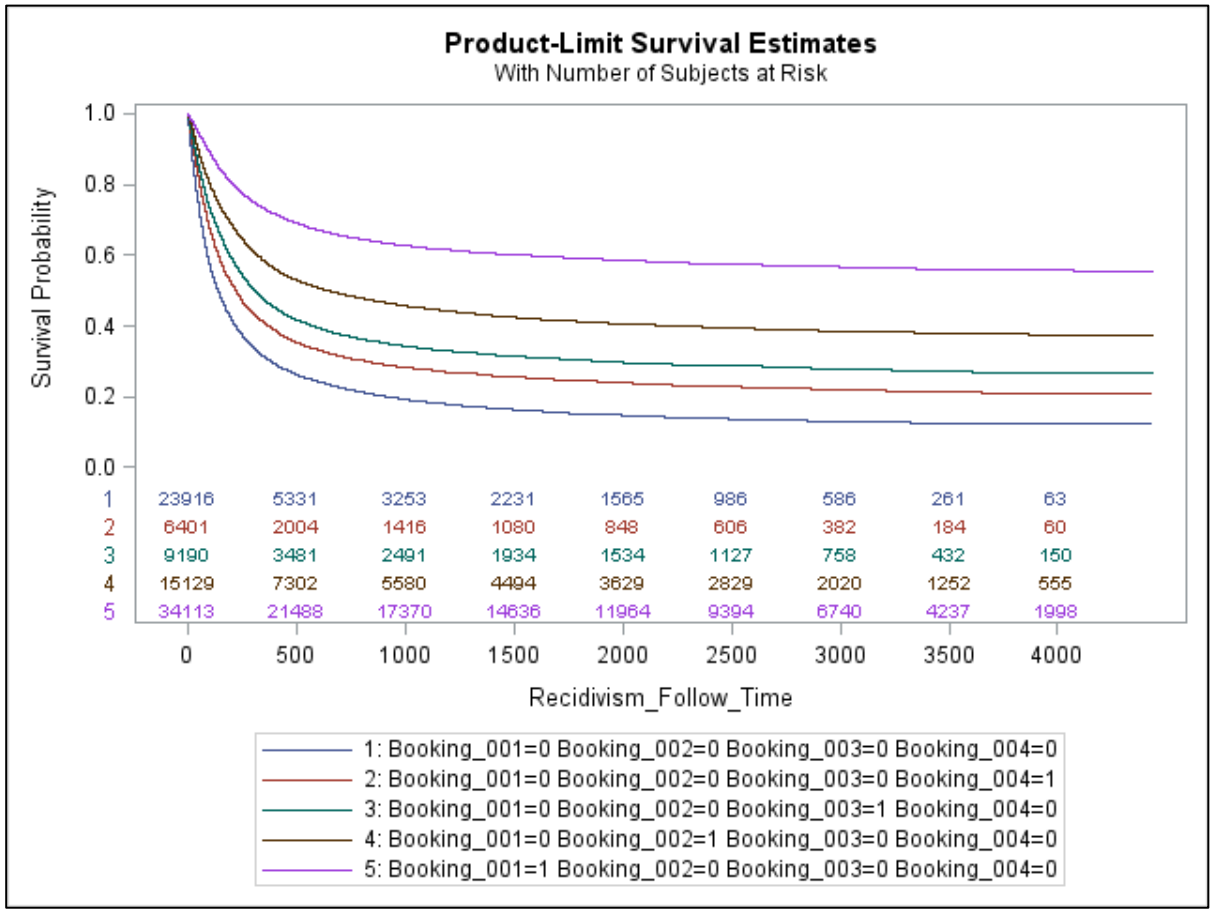


Figure 25: Kaplan-Meier Graph Showing Bookings -001, -002, -003, -004 and Others

The same tests for equality and homogeneity that were shown for the first graph are not shown for the subsequent graphs because these became non-informative, replicating the kind of results gotten from the first analysis show. Likewise, if the graphs escalated booking levels beyond the last level shown, the results became repetitious.

Let it suffice to say that each increase in booking level increases the recidivism. To link the wording to the graphs presented, it can be said that each increase in booking level decreases the survival rate. By looking at the where the survival curves flatten out on the right, the decrease in survival rate is apparent. It is an example of diminishing returns, also.

By going from the booking extension -001 to -002, the survival rate goes from approximately 60% to 42%. In going from booking number extension -002 to -003, the survival rate goes from approximately 42% to 30%. In going from booking number extension -003 to -004, the survival rate goes from approximately 30% to 22%. If one examines the curves on the right, the gaps between successive booking levels gets progressively narrower.

Based on what has been seen in the last two chapters, it would appear that there are two factors—booking level and gender--at work within the group of people who have been booked into the Stearns County Jail. One is a personal characteristic, as if being booked makes it increasingly more likely to get booked again. Also, gender appears to make it more likely that someone will recidivate. In the previous chapter, the last graph showed four groups of people being booked: females on the first booking, males on the first booking, females at booking levels beyond the first booking, and males at booking levels beyond the first booking.

The least likely to recidivate are females who have been booked once, the most likely to recidivate are males who have been booked more than once. The similarity in the shapes of the curves and gaps between the curves are of great interest. To illustrate this point further, another graph follows that shows the genders for first booking (-001), second booking (-002), and beyond the second booking. Note how the gap between genders gets narrower with the increase in booking level.

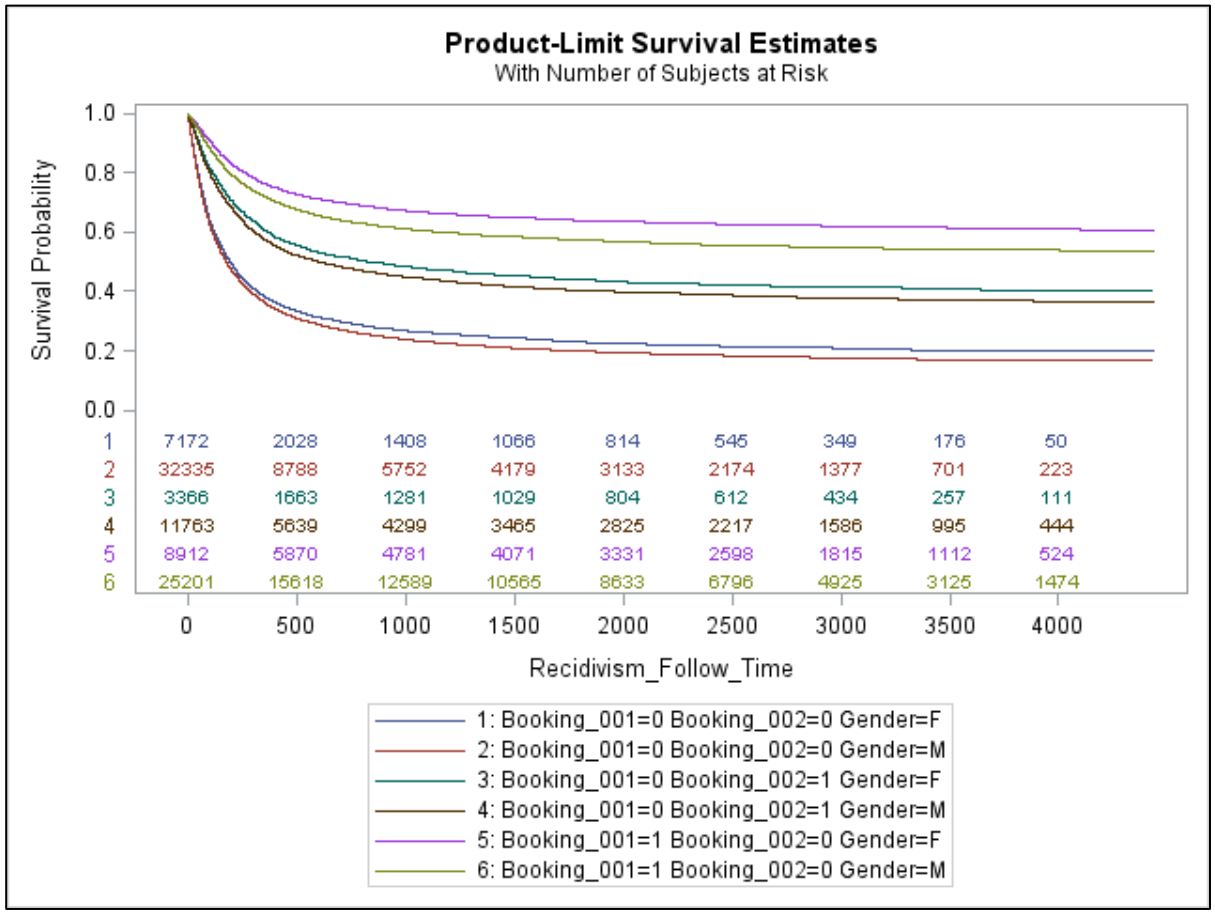


Figure 26: Graph Separating the Data into Booking -001 and -002 Levels and Gender

Chapter 15: Parametric Description

In their analysis of North Carolina prison data, Schmidt and Witte (1988) found that a lognormal distribution fit their data best, however, imperfectly. More specifically, the time until recidivism was based on a lognormal distribution. Through a series of developments based on this lognormal distribution of time until recidivism, the model that functioned best was what they termed a "logit/individual lognormal model." This worked by using logit distribution with explanatory variables to determine a probability to recidivate, while a lognormal distribution, also based on explanatory variables, determined the time to recidivism for those that do recidivate (pp. 106-109). They did not use the term, but this could be thought of as a mixture model. The key insight, which is largely ultimately attributable to the works of Maltz and other co-authors in the late 1970s and 1980s, is simply that not everyone will recidivate.

There was an aspect of the recidivism data that drew Schmidt and Witte (1988) to the use of the lognormal distribution. It was how that, with their data from the North Carolina prisons, the recidivism would rise very soon after release from incarceration and then drop very quickly. This behavior was very difficult to model using a specific distribution, but the lognormal distribution was able to do this better than others they tried (p. 54).

When the data from the work of Schmidt and Witte (1988) was put into JMP, the Frechet distribution was found to be slightly better than a lognormal for explaining time until recidivism, although they were probably unaware of this distribution.

With the Stearns County Jail booking data, early attempts at data analysis made use of JMP to fit a distribution to explain time until recidivism. The result was that lognormal was initially found to work best. When better procedures were found to calculate recidivism and follow time, JMP began to report that the Frechet distribution worked better at describing the data than other distributions. Interestingly, the lognormal was listed as the second best distribution, both when evaluating the data from Schmidt and Witte (1988) and the Stearns County booking data after the data was re-worked. Here is the result of the result of model comparison by JMP:

Parametric Survival Fit - Distribution

Distribution	AICc	BIC	
Weibull	822743.63	822762.42	
LogNormal	810948.58	810967.36	
Exponential	880497.11	880506.51	
Frechet	805146.82	805165.60	Best
Loglogistic	814035.27	814054.05	

Figure 27: JMP Comparison of Distribution Models

For survival data where there are many events of interest relatively early, yet a long right tail, a lognormal distribution tends to work better than many other distributions. A recent article on the lifetime of restaurants used a lognormal distribution, apparently for this very reason (Luo & Stark, 2015, pp. 25-29). It would appear that the Frechet distribution is also suited for this purpose, although an obscure one for students.

For a simple overview of the Frechet distribution, which is not easily found in textbooks, the following displays are reproduced from the Wikipedia page on the Frechet distribution. The cumulative distribution function is shown first, "…where $\alpha > 0$ is a shape parameter. It can be generalised to include a location parameter m(the minimum) and

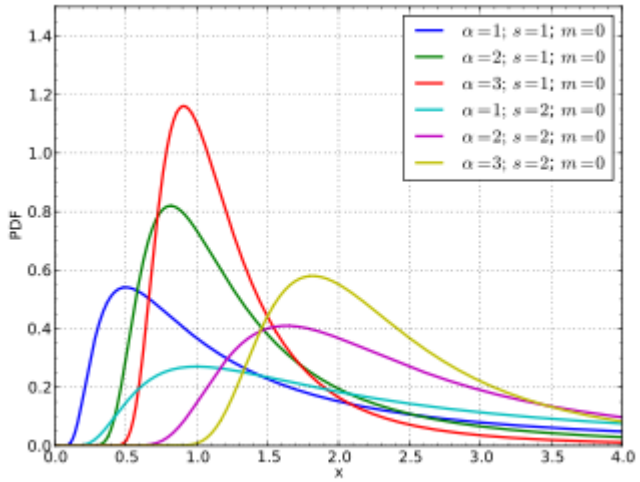
a scale parameter *s* > 0 with the cumulative distribution function" (Wikipedia, retrieved 7/18/2015).

$$\Pr(X \leq x) = e^{-\left(\frac{x-m}{s}\right)^{-\alpha}} \text{ if } x > m.$$

Parameters	$\alpha \in (0, \infty)$ shape. (Optionally, two more parameters) $s \in (0, \infty)$ scale (default: $s = 1$) $m \in (-\infty, \infty)$ location of minimum (default: $m = 0$)
Support	$x > m$
PDF	$\frac{\alpha}{s}\left(\frac{x-m}{s}\right)^{-1-\alpha} e^{-(\frac{x-m}{s})^{-\alpha}}$
CDF	$e^{-(\frac{x-m}{s})^{-\alpha}}$

Figure 28: Frechet Distribution (from Wikipedia)

Probability density function



Cumulative distribution function

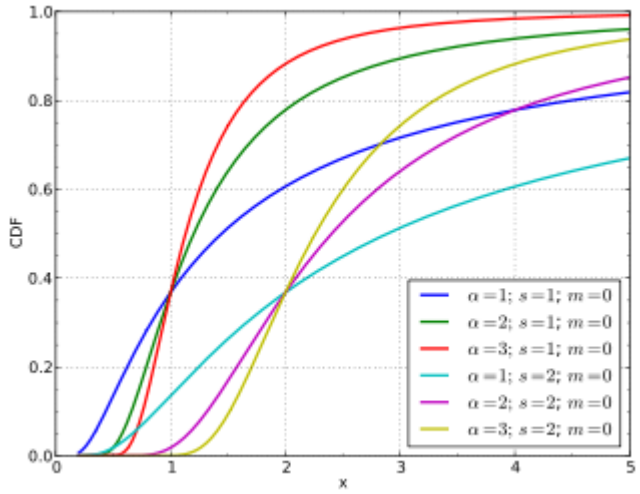


Figure 29: Frechet Distribution PDF and CDF (from Wikipedia)

What follows is output from JMP that describes the Frechet and lognormal applications to the Stearns County booking data, including distribution parameters:

Parametric Survival Fit

Time to event: Recidivism_Follow_Time

Distribution: Frechet

Censored By: Censor_Flag

Time=zero encountered 150 times in table

AICc	805146.8
BIC	805165.6
-2*LogLikelihood	805142.8

Observation Used	88618
Uncensored Values	52010
Right Censored Values	36608

Parameter Estimates

Term	Estimate	Std Error	
Intercept	5.57106533	0.0099067	
σ	2.64233722	0.0087331	

Parametric Survival Fit

Time to event: Recidivism_Follow_Time

Distribution: LogNormal

Censored By: Censor_Flag

Time=zero encountered 150 times in table

AICc	810948.6
BIC	810967.4
-2*LogLikelihood	810944.6

Observation Used	88618
Uncensored Values	52010
Right Censored Values	36608

Parameter Estimates

Term	Estimate	Std Error	
Intercept	6.64755624	0.0099383	
σ	2.59975065	0.00883	

Figure 30: JMP Output Regarding Frechet and Lognormal Distributions

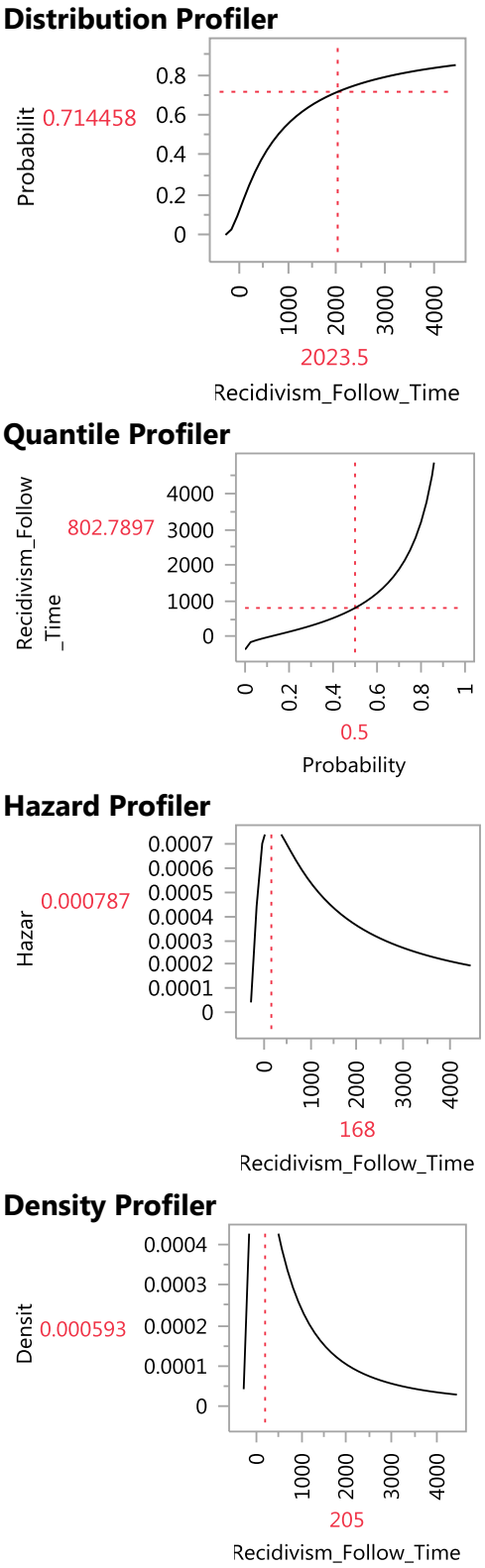


Figure 31: Frechet Distribution Graphs

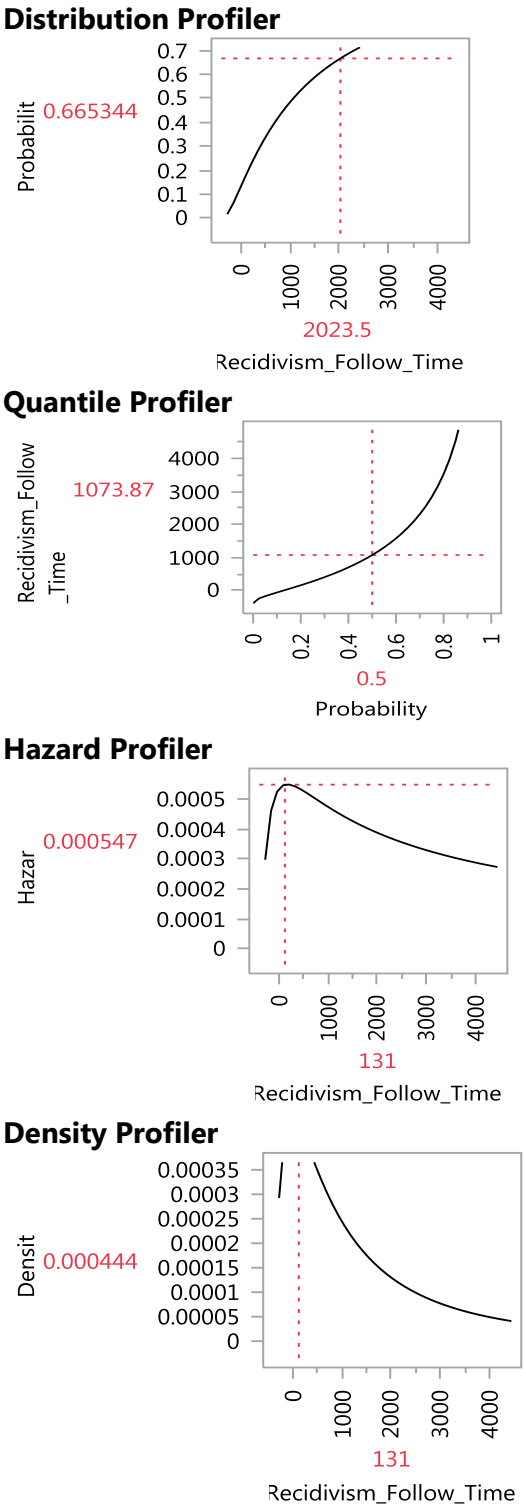


Figure 32: Lognormal Distribution

There is a large problem in that SAS is unable to handle the Frechet distribution, unlike JMP. In fact, using the search feature to search for the term, "Frechet," in the help section yields, "No results." Therefore, given this limitation with use of SAS, an attempt to create a rudimentary model based on a lognormal distribution applied to the survival time, as opposed to the Frechet distribution.

This is why JMP results for both Frechet and lognormal distributions are provided above, courtesy of JMP's "Life Distribution" analysis application. For each distribution, four graphs are shown: the failure (inverse of survival), what is called a "Quantile Profiler," the hazard graph, and the pdf graph. The first graph is labeled as "Distribution Profiler," but it is the failure curve. The second is labeled as "Quantile Profiler," and is useful for its clear identification of the median survival value of days. The third and fourth graphs are the hazard and pdf curves, labeled as "Hazard Profiler" and "Density Profiler," respectively.

The failure curves are of particular interest with each distribution. If both are examined, one can see the behavior of the graph at the right extreme of the x-axis, corresponding to over 4000 days. Both appear to show the failure curve going to the right of the graph at approximately the 80% level and higher, meaning that 80% are recidivating (and 20% not recidivating, or surviving). The failure curve is the inverse of the survival curve, and recall the basic Kaplan-Meier curve previously showed a survival rate of 40% at the right, meaning the failure rate of approximately 60%. This would appear to be a limitation of using a distribution to describe and approximate the actual data—the distributions are over-estimating the results of the data.

This same pattern can be seen when examining the medians provided by the two distributions. Recall the non-parametrically derived median value of 487 days. By examining the second graph for each distribution, the median can easily be seen. The Frechet distribution gives a median value of approximately 803 days, while the lognormal distribution gives a median value of approximately 1074 days. Recall also how the non-parametric analysis was unable to provide an estimate of the 75^{th} percentile, most like because such a level was not being approached when the survival rate was getting to approximately 40%, or a failure rate of 60%. Both distributions can provide a 75^{th} percentile estimate.

The difference in the median values also provide an insight as to why the Frechet describes the data more effectively, since the median is closer to the non-parametrically derived median value. In any event, the pattern is emerging of over-estimation. This will be discussed more at length later, but let it suffice to state for now that a distribution fit to survival data will make the assumption that failure, or in this case, recidivism, is inevitable. All will fail, where the more realistic viewpoint, in regard to recidivism data, is that some individuals will not recidivate.

The hazard curve allows for the identification of the peak hazard time, which coincides with the peak of the pdf, turns out to be at approximately 168 days after release from the Stearns County Jail, according to the Frechet distribution. The lognormal distribution provides a peak of the hazard curve at approximately 131 days. Finally, note that no use of an explanatory variable was made in applying these distributions.

Before continuing onto the SAS results, a brief review of the variables involved in the survival time modeling is in order. The variable Recidivism_Follow_Time, which is in days, is a recidivism time if the Censor_Flag has a value of 0. This would mean the recidivism event happened and a "failure" occurred. However, Recidivism_Follow_Time is a censored value with the time value being a follow time if the Censor_Flag has a value of 1. The only explanatory variable offered by the data is that of gender.

The first SAS output presented is that of the lognormal model without gender as an explanatory variable. The second round of SAS output regarding the lognormal model is where gender is an explanatory variable. The figures for AIC, AICc, and BIC that are provided as part of the results are not particularly meaningful in isolation, without other models for comparison. Therefore, the JMP results for the measures of AICc and BIC, provided previously, which provide the measures for comparison with other parametric distributions, are more instructive.

However, if the AIC, AICc, and BIC results are compared between the models with and without gender, the lognormal model is more effective when gender is considered. Graphs of the cumulative distributions are presented after the results of each analysis. If one compares the fit statistics between the two lognormal models, one can see that the one where gender is taken into account does "work" somewhat better. Note how the graphs of the cumulative distribution function approaches levels, in the 70%-80% range that the actual does not approach. Recall that for the data as a whole, the failure rate approaches 60%.

Model Information		
Data Set	MYLIB.TIMING_DATA_2015_07_15	
Dependent Variable	Log(Recidivism_Follow_Time)	Recidivism_Follow_Time
Censoring Variable	Censor_Flag	Censor_Flag
Censoring Value(s)	1	
Number of Observations	88598	
Noncensored Values	51991	
Right Censored Values	36607	
Left Censored Values	0	
Interval Censored Values	0	
Number of Parameters	2	
Zero or Negative Response	170	
Name of Distribution	Lognormal	
Log Likelihood	-152217.9449	

Number of Observations Read	88768
Number of Observations Used	88598

Fit Statistics	
-2 Log Likelihood	304435.9
AIC (smaller is better)	304439.9
AICC (smaller is better)	304439.9
BIC (smaller is better)	304458.7

Fit Statistics (Unlogged Response)	
-2 Log Likelihood	810944.6
Lognormal AIC (smaller is better)	810948.6
Lognormal AICC (smaller is better)	810948.6
Lognormal BIC (smaller is better)	810967.4

Algorithm converged.

Analysis of Maximum Likelihood Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**95% Confidence Limits**		**Chi-Square**	**Pr > ChiSq**
Intercept	1	6.6476	0.0099	6.6281	6.6670	447408	<.0001
Scale	1	2.5997	0.0088	2.5825	2.6171		

Figure 33: SAS Description of the Lognormal Model without Gender

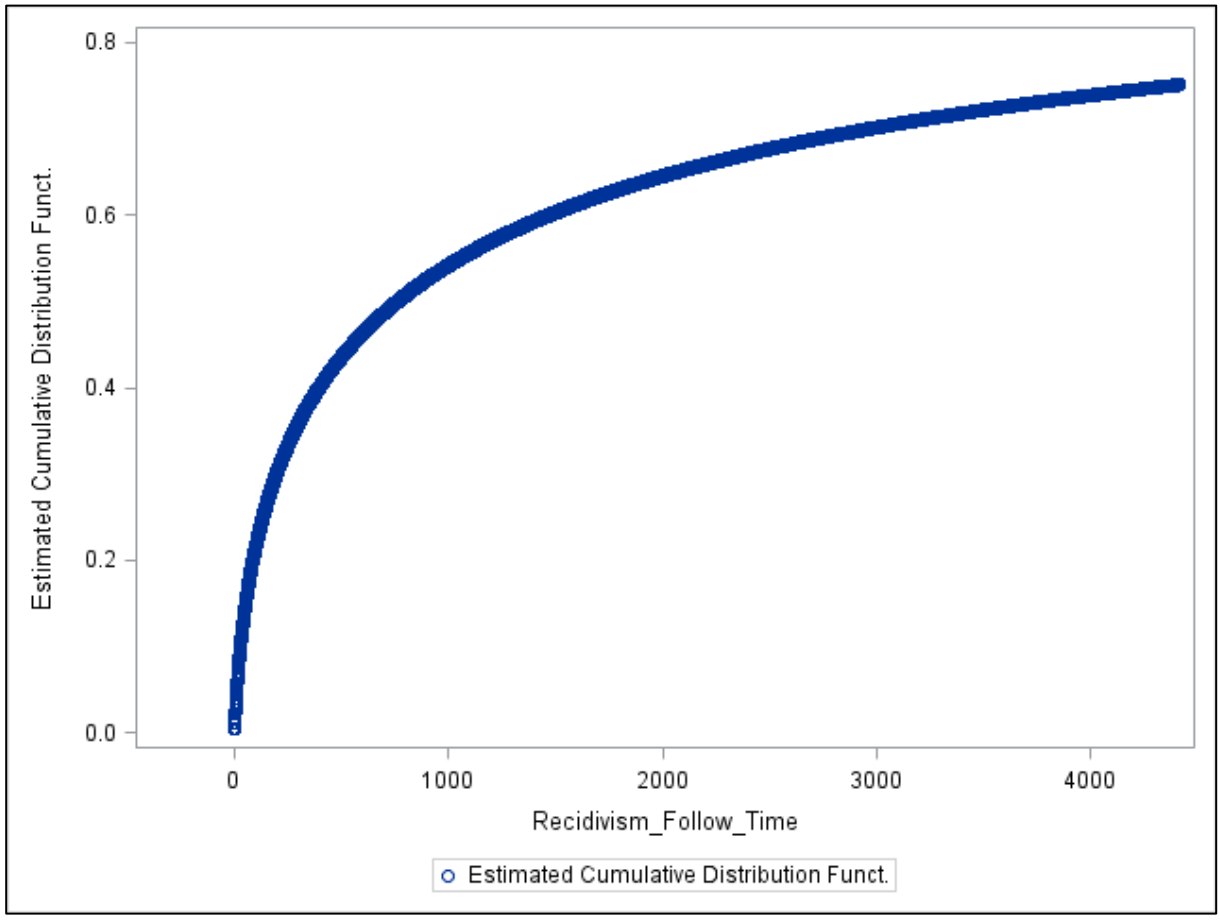


Figure 34: CDF Based on the Lognormal Distribution without Gender

Model Information		
Data Set	MYLIB.TIMING_DATA_2015_07_15	
Dependent Variable	Log(Recidivism_Follow_Time)	Recidivism_Follow_Time
Censoring Variable	Censor_Flag	Censor_Flag
Censoring Value(s)	1	
Number of Observations	88598	
Noncensored Values	51991	
Right Censored Values	36607	
Left Censored Values	0	
Interval Censored Values	0	
Number of Parameters	3	
Zero or Negative Response	170	
Name of Distribution	Lognormal	
Log Likelihood	-152010.6789	

Number of Observations Read	88768
Number of Observations Used	88598

Class Level Information		
Name	**Levels**	**Values**
Gender	2	F M

Fit Statistics	
-2 Log Likelihood	304021.4
AIC (smaller is better)	304027.4
AICC (smaller is better)	304027.4
BIC (smaller is better)	304055.5

Fit Statistics (Unlogged Response)	
-2 Log Likelihood	810530.0
Lognormal AIC (smaller is better)	810536.0
Lognormal AICC (smaller is better)	810536.0
Lognormal BIC (smaller is better)	810564.2

Algorithm converged.

Type III Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Gender	1	412.6482	<.0001

Analysis of Maximum Likelihood Parameter Estimates								
Parameter		**DF**	**Estimate**	**Standard Error**	**95% Confidence Limits**		**Chi-Square**	**Pr > ChiSq**
Intercept		1	6.5463	0.0110	6.5248	6.5678	355317	<.0001
Gender	**F**	1	0.4673	0.0230	0.4222	0.5124	412.65	<.0001
Gender	**M**	0	0.0000	.	.	.	.	.
Scale		1	2.5940	0.0088	2.5767	2.6113		

Figure 35: SAS Description of the Lognormal Model with Gender

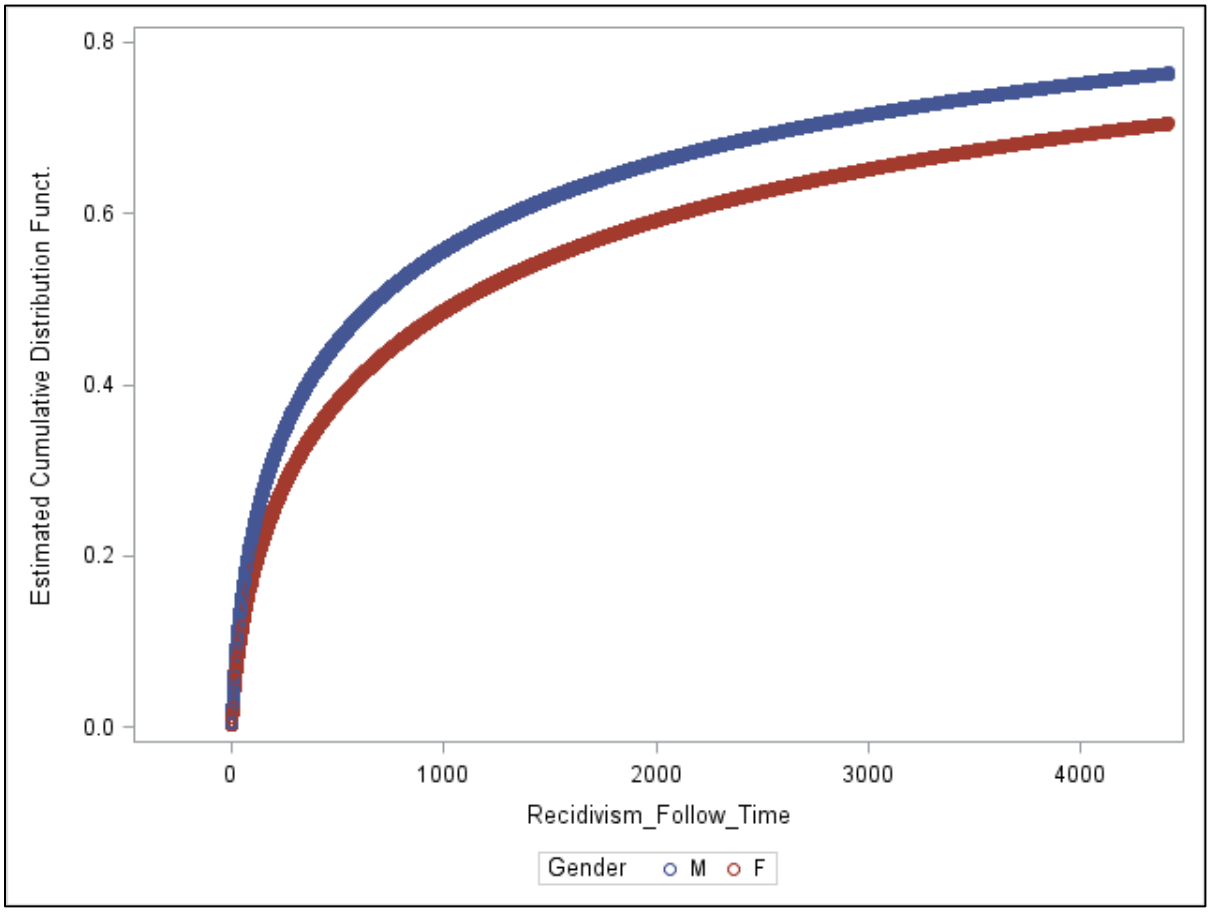


Figure 36: CDF Based on the Lognormal Distribution with Gender

Next, probability plots are examined under the assumption of a lognormal distribution through SAS. Once again, the lognormal model without explanatory variables is dealt with first and then the lognormal model with gender as an explanatory variable. Much of the SAS output is the same as shown previously, so only the probability plots will be shown.

The probability plots show a reasonable fit for the lognormal model, though the graphs are difficult to distinguish from each other. In conjunction with the fit statistics provided earlier, it would appear that a lognormal model that takes gender into account is the better description of the data.

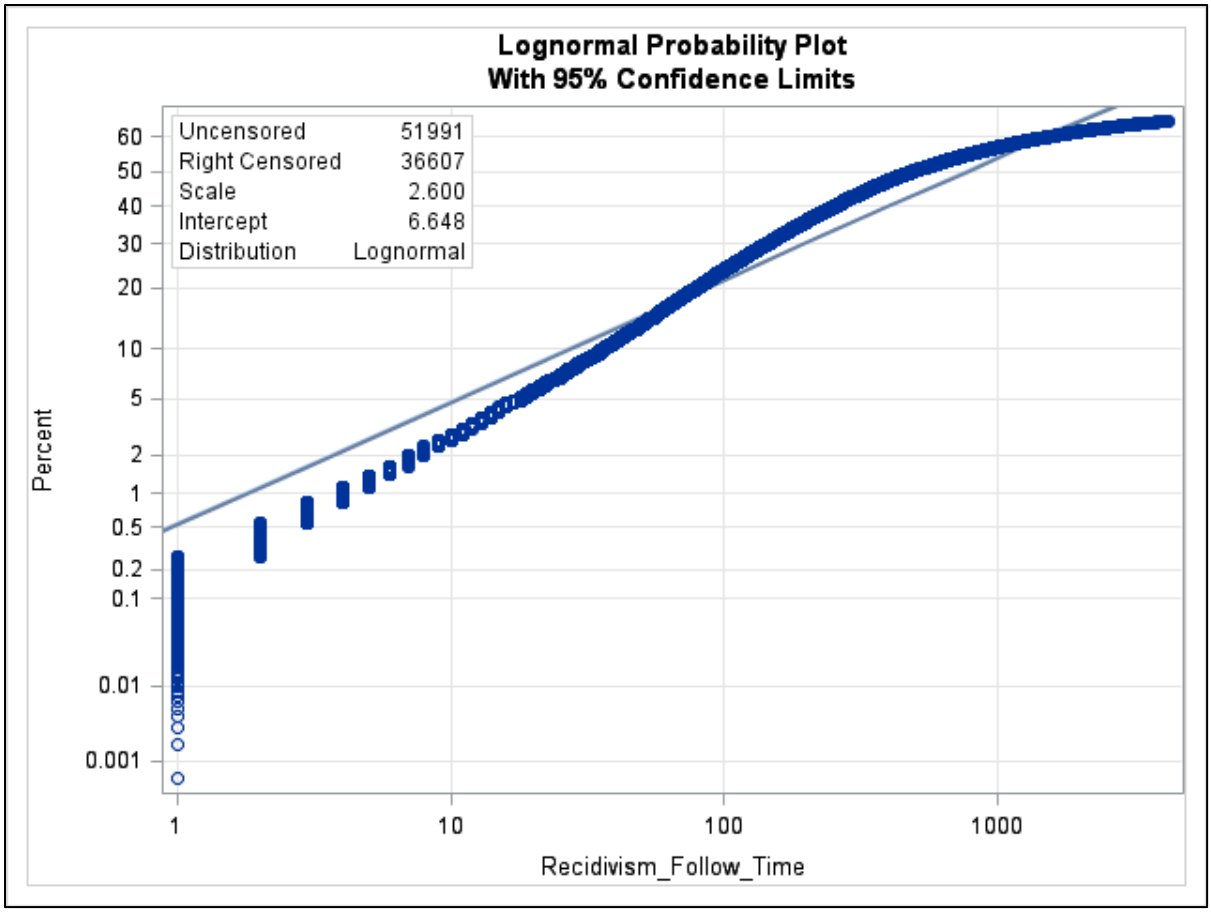


Figure 37: Probability Plot for Lognormal Distribution without Gender

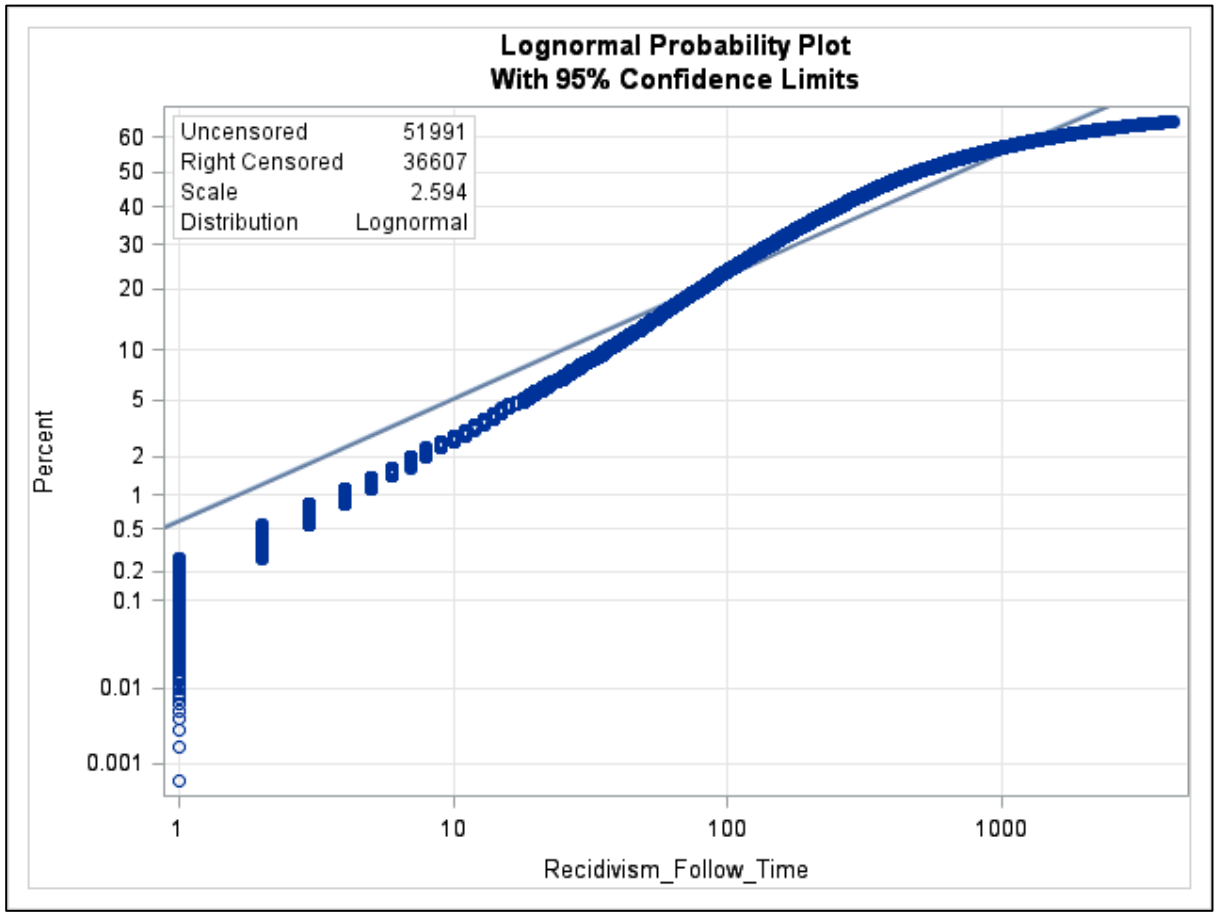


Figure 38: Probability Plot for Lognormal Distribution with Gender

Chapter 16: Testing for the Presence of Immunes

The following survival analysis concepts are from Lee and Wang (2003, pp. 8-9). From basic definitions from general mathematical statistics regarding cumulative distribution functions, there are two properties of cumulative distribution functions that are important in the discussion here. One is that F(t) is a non-decreasing function of T. The second can be stated in the following manner:

$$F(\infty) \equiv \lim_{t\to\infty} F(t) = 1$$

Based on the definition of a cumulative distribution function, note there is an underlying assumption in most survival analysis applications of these concepts that F(t) will approach 1 as t, or time, approaches ∞. Thus, as time accumulates, the proportion of the population that fails will approach 1.

In view of the fact that a survival function is the complement of the cumulative distribution function, that is:

S(t) = P(an observed unit survives longer than time t)

S(t) = P(T > t)

S(t) = 1 – P(an observed unit fails before time t)

S(t) = 1 – F(t)

Based on this, the survival function, S(t), can be viewed as a "non-increasing function of time t" with the following important properties:

S(t) = 1 for t=0

and

$S(t) = 0$ for $t=\infty$

The presence of "immunes" will disrupt this line of thought. Maller and Zhou (1996), in reference to survival analysis, generally speaking, believed that there are various survival phenomena where not everyone will fail. If this is not taken into account, the results could suffer some distortion. If there are immunes, the analysis may need to take that fact into account. If there were no indications of a presence of immunes, then the "usual" survival analysis could proceed (p. xi).

Maller and Zhou (1996) developed some procedures for testing for the presence of immunes. Without some sort of procedure, all one is left with is the idea that if the Kaplan-Meier failure plot of the data levels off at the right end of the graph, and then exits the graph on the right "somewhere" below the value of 1, this might be an indication that the cumulative distribution function:

$$\lim_{t\to\infty} F(t) < 1$$

As has been mentioned previously, Maller and Zhou (1996) have pointed out that with some applications of survival analysis, not all the lifetimes will end in a failure event. Even when applied to non-medical applications, not everyone will succumb to the event in question. They also thought about other applications of survival analysis, including the issue of recidivism (pp. 23-24). Maller and Zhou (1996) referred to such individuals as "immunes" (pp. 1-4), meaning those who will not succumb to the failure event. In certain medical applications, they are the ones who will not undergo tumor recurrence or who will not die of the disease process.

Plain survival models often operate under the assumption that all the subjects will ultimately succumb to the event in question, whether before or after the period of observation. That is, in making use of the most popular applications of survival analysis, electrical components will all fail eventually and all the subjects will die. In more mathematical terms, the cumulative distribution function will eventually sum to 1 (Maller & Zhou, 1996, p. 18). Maller and Zhou (1996) were referring to larger issues in survival analysis and not simply recidivism. Schmidt & Witte (1988) were thinking only of recidivism in this regard (p. 66-67).

Another term commonly encountered for this idea is that of "cure models," which reflects a medical perspective. In his paper on taking into account the "cure fraction," Lambert stated, "In these situations, interest often lies in estimating the proportion of subjects who do not experience the event" (Lambert, 2007, p. 351).

Without a method or procedure to evaluate survival for the presence of immunes, all one can do is observe the behavior of the Kaplan-Meier graph and see if it tends to flatten out as it goes to the right (Maller & Zhou, 1996, pp. 5-6). Obviously, this is subjective and relies on an interpretation. Examples of which can be contained in such questions such as, what is meant by "flatten out"—if it is rising, but at a shallow angle relative to prior behavior, does that count as sufficiently flattening? This would certainly seem to be particularly important in medical applications.

The work of Maller and Zhou (1996) is widely referenced, but it appears to be referenced mostly in discussions of the importance of evaluating the presence of immunes.

This is opposed to being referenced for the actual methods of evaluating the presence of immunes. The non-parametric and parametric methods they developed appear to be of questionable value. Very simply, it is a rather obtuse procedure, it appears to rely on the calculation of the proportion of censored data in a hypothesis test procedure. These values are compared with table values in the back of their book. These table values were derived from computer simulations based on an exponential distribution. Surprisingly, time does not seem to enter into the evaluation, when, for example, it would certainly appear that a five-year observation window would be more revealing of the presence of immunes than a one-year window (pp. 76-81, 109-115).

Despite this problem, the book by Maller and Zhou (1996) is illuminating, and they do offer one important result that is of great value. Kaplan-Meier survival curves are largely unbiased and good estimators of the immune proportion of the population in question (pp. 41-47).

The great majority of the discussion and interest in the presence of immunes is on the part of the medical community for understandable reasons. The question does arise as to whether or not the idea of “cure” or “immune” could have the same meaning from medical applications as it does in the evaluation of recidivism. The larger idea that if one examines survival data for a disease treatment or for recidivism, there may be a proportion that will not fail. This would appear to be a point of unity.

In medical science, it seems clear that questions surrounding the existence of immunes would be extremely important: questions such as to whether or not immunes are

present, and if so, there would further questions as to the proportion of immunes. It would seem that further questions would then revolve around how to take this into account when evaluate the course of the disease or method of treatment.

With recidivism, it would seem the presence of immunes is already known to exist in nearly any group of people subject to recidivism. There is unlikely to be much recidivism data that shows 100% recidivism. Recidivism data consistently shows substantially less than 100% recidivism rate. Depending on exactly the nature of the data, the definition, and people being examined, rates can vary. When Schmidt and Witte (1988) examined the various rates of recidivism in their survey, the range of rates went from 36% to 63%, while their own data had a rate of 35% during the window observation (pp. 68-69).

The most recent and authoritative examination was the already reference BJS study. For a five year period, 55.1% had a re-incarceration recidivism rate and a 76.6% re-arrest recidivism rate (Durose et al., 2014, p. 1). The rate of recidivism is then extremely dependent on data, definition, and assumptions. In contrast, medical applications have the constant pressure to find results that can be replicated and applied to other groups of people. In other words, the context is different. Testing for the presence of immunes would certainly appear to be of great importance in the medical realm, and that is where the focus of testing for immunes appears to be.

It can be conjectured that being an “immune” would mean different things between medical science and the study of recidivism. In medicine, being immune would mean the medical event is not going to befall certain individuals. With recidivism, being immune is

more nuanced—a person can recidivate but has not done so. This is a different matter from immunity. To summarize, testing for the presence of immunes in recidivism data is not illuminating. It is understood that immunes, or individuals who are not recidivating, are present. What proportion such people exist will depend heavily on data, definition, and assumptions.

Chapter 17: Summary and Conclusions

For retrospective observational data of real events, the data was excellent as far as the variables for booking numbers, the dates and times of booking, and the gender. The only problem in regard to the booking number variable was the inexplicably missing bookings from what was intended to be a comprehensive list over the period of time 1/1/2003 to 1/31/2015. The data on booking date and time, along with gender, appeared remarkably reliable. In the constant cross-checking with other Excel data, no errors were found with these variables. On the whole, these aspects of the data were excellent.

The data was more frustrating in the date and time of release information. More specifically, the times when the data was simply missing and the amount of time and effort to locate missing release information was considerable. When the date and time of release information was present, or when it was found in other sources, the information proved to be very reliable. Despite these problems with the data, as far as the four variables used in this analysis are concerned, this collection of data is comprehensive enough to be considered a census. In regard to these four variables, the data is sufficiently complete to make reliable and useful data analyses for Stearns County officials. It is also reliable enough to make projections and predictions for future use of the jail.

The one caveat that must be mentioned at this point is the single most difficult aspect of the data. The "release" did not necessarily mean the individual was released to a state of freedom. It simply meant the person was no longer in the Stearns County Jail's custody. This issue alone threatened the basis of an analysis that purports to examine the

issue of recidivism, since it means that time that most ideally thought of as “survival time” may in fact have been spent in the custody of some other criminal justice entity. It may mean, and perhaps in most cases, it does mean the person was released from the facility to a state of freedom. One is unable to even give a proportion for the releases to freedom as compared to a transfer to another authority.

This situation with a “release” date and time, though obviously not ideal, can remind one of other situations with recidivism data. Recidivism data is imperfect. When discussing their own data and definition of recidivism with which they had to work, Schmidt and Witte (1988), in a very telling passage, they “analyzed the timing of return to prison in North Carolina because this was the only definition of recidivism that our data would support” (p. 9). Moreover, even with the ambiguity over the release date and time, Stearns County corrections officials would still be interested about the large group behavior of those who have been in their custody.

The data is very large at 88,768 bookings. Excel, SAS, and JMP applications run on a laptop computer often seemed to have trouble handling the volume of data. Slowness and occasional non-response of the computer applications were reminders of the size of the data being used. Given this size, the absence of some bookings or some release dates and times should probably not have been regarded as being so troubling.

This entire analysis was based on four variables: booking number, confinement date and time, release date and time, and gender. Only one of these, gender, can be considered to be an explanatory variable for the purposes of examining recidivism. The other two

variables in the original Excel data, MOC and Statute_Description, were another problem. The variables appeared to be inconsistent, as discussed earlier.

In future analyses, however, some use must be found for the MOC variable, or as labeled in the original version of the data, the "Offense" variable. This variable has the potential to be revealing, especially as an explanatory variable in regard to recidivism. The nature of the offenses, and it will be more than one offense for many of the bookings, should be expected to be revealing for recidivism. Different crimes can be expected to result from different personalities, or different problems, of the people who are accused of committed them.

Care must be taken in this regard. As discussed earlier, one problem with this variable in the Excel spreadsheet was that it appeared to have inconsistent values, but with no reason or understanding behind this inconsistency. Here are some questions that needs to be addressed in order to better use this variable: Is the list of charges shown for a particular booking is consistently comprehensive—does it tend to reflect all the charges that lead to a booking, or can it be partial list? What happens to charges that are added, dismissed, or changed? If changes are made to the charges, will these changes appear on the Excel list? Why do different booking records for different people have different MOC values even though the Statute_Description seems to be identical? On the other hand, how can different booking records have the same MOC value, but different Statute_Description wordings? How is it that a booking can have one MOC value, but the Statute_Description appears to list multiple offenses that in another booking warrant individual MOC values or

separate booking records? What does it mean if one booking event but have identical MOC and Statute_Description on different Excel rows? If an entire booking is dismissed or expunged, is it taken off the list before it is seen? Should dismissed or expunged charges be included in statistical analyses? How should blank data values be treated?

It is a recommendation here that the data should reflect the booking event at the jail. As much as is possible, separate charges should merit a separate booking record, or row in Excel data. The list should strive to be consistently comprehensive, meaning that if there are three rows of Excel data for one booking event, it means there were only three charges for which the booking is occurring. The list should not altered retroactively, unless it is to fix errors. It should not add or subtract charges later in the process since it is the actual booking event that is most useful to study.

In other words, the number of rows in an Excel spreadsheet should reflect the totality of the charges for which the individual was booked at that particular time, with no charges left out, and no duplication. There should be no additions, subtractions, or deletions, other than to correct errors that were made. If this is done, it would be a powerful explanatory variable.

One feature of using the MOC variable is that if there are multiple charges, care must be taken to separate the seriousness of the different charges. In the parlance of criminal justice officials, there needs to be awareness of the "controlling charge." If someone is booked into the jail for a violent assault and driving without insurance, an analyst will need to take care that that such a booking is included in the proper subgroups. The MOC coding is

standardized for the State of Minnesota, so once the questions listed above are addressed, it should be no problem to sort through the codes and order them by seriousness.

In fact, this difference in the seriousness of charges was a difficulty when this data was being culled to get data where there was one unique booking associated with one unique date and time. One unfortunate result was that it was difficult to keep a more serious charge and delete the less serious.

Another item can make the data even more powerful. This would be to add another variable so that the reason for release from the Stearns County Jail can be quantified. This can be done, since this information was seen on another, but limited Excel spreadsheet. This was text information, much like the Statute_Description variable present in the Excel spreadsheet used here. If such information is text information, extra work will have to be done to make the outcome of "release" from the Stearns County Jail something that can be used in analysis.

The fact that the lognormal distribution worked best was not surprising in view of the experience of Schmidt and Witte (1988), who found that their recidivism peaked very quickly and dropped very quickly, a behavior best approximated by a lognormal distribution (p. 54). The definition of recidivism used here was very different, at the other end of the funnel—meaning that booking was used instead of incarceration. Yet, that same behavior was seen here.

That definition of recidivism as denoting an event at the top of the funnel (being booked) is also offered as a partial reason why women were a more substantial portion of

the data. Women were approximately 22% of the data, whereas much criminal justice data has less than 10% of the data is represented by women.

Gender and booking level were powerful indicators of recidivism, as seen in previous chapters. The graphs which sorted the recidivism by booking level and gender were extremely telling and appeared to reveal that these two aspects affected an individual in both ways. Booking level affected the recidivism of each gender, and gender affected the recidivism of each booking level.

When the booking is separated by individuals, which was done by focusing on bookings ending in -001, the recidivism rate is approximately 40%. Recidivism rates can vary quite a bit depending on the group of people being studied, the length of time the individuals are followed, and the definition of recidivism that is being used. Schmidt and Witte (1988) examined the different results from recidivism studies available to them at the time and found the results varied from 36% to 63%, while their own group had a recidivism rate of about 35% (pp. 68-69). Durose et al. (2014) found in their examination of a sample, 55.1% were re-incarcerated for a new crime within five years of release and 76.6% were arrested for a new crime within five years of release (p. 1). With this data, where booking is a more common event to befall an individual, as opposed to incarceration, a 40% recidivism rate for people booked for the first time is not an unusual number.

However, in using a lognormal distribution to model the data, some additional problems emerged. First, the lognormal was really the second best distribution. The Frechet was actually better. However, recall that JMP actually reported that the lognormal

distribution was better in early attempts at analyzing the data. Then, better methods were used to calculate recidivism and follow time. When these changes were made and the data was re-analyzed, the Frechet distribution became a better fit for the data.

There is a lesson here. The data is real data, observational and imperfect. Given this nature, it is very solid, but it is still messy, with many problems as has been described in detail earlier. On top of that, there were many decisions to be made regarding the calculation of recidivism and follow time, the length of the observation window, and how to deal with missing data.

Moreover, the data exists in its current form because of many decisions made by people that could have gone in other directions. The immensity of the bookings made in 2002 is a case in point, as well as the unknown factors in removing certain bookings from the data. It can wondered if being more or less diligent in filling in missing data would have had much impact on the final distribution. Another question is whether it is right or wrong to remove dismissed or expunged charges.

In other words, this is messy, unclear, and far from ideal social science data. To take such data and use a particular distribution to describe that data can seem artificial. It has to occur to the analyst that if certain decisions had gone another direction, another distribution would then work better. This is what happened here when JMP analysis went from recommending a lognormal to recommending a Frechet distribution. Seen this way, it is not a great sacrifice in analytical quality to use the lognormal distribution to describe the data, even though JMP said it was the second best distribution. It would have been preferable to

use the Frechet had SAS been capable of using this distribution, but it should not be seen as devastating blow to analytical quality.

The remaining issue was the over-estimation by using a survival distribution that assumes S(t) will approach 0 as t approaches ∞. This was a problem that can be seen quite simply by looking at the very different estimates of the median value of the data obtained from the non-parametric and lognormal estimates. According to the non-parametric analysis, the median recidivism time was 487 days, while the 75^{th} percentile was unable to be calculated. According to the Frechet distribution, the median was approximately 803 days, while the lognormal distribution offered a median value of approximately 1074 days. The underlying assumption in fitting a lognormal distribution was that the failure event was inevitable. In fact, according to the JMP graphs presented earlier, the survival curve appeared to approach 0, while the "Quantile Profiler" approached 1, meaning the assumption was being made that the failure event would befall everyone.

As mentioned earlier, Schmidt and Witte (1988) found a lognormal model fit best, but was far from perfect initially. The model worked much better when it was taken into account that not everyone would recidivate. They introduced a "splitting parameter" that would yield a probability of an individual recidivating. Details and specifics were not provided as to how they applied this in practice, and later this splitting parameter was made dependent on explanatory variables. It appears they were using a "degenerate" probability function where the cumulative distribution function did not sum to 1.

Such splitting parameters may be determined by explanatory variables, but do not have to be. In one version of their model, Schmidt and Witte (1988) used a flat number proportion as the splitting parameter that applied to every person in the data (pp. 66-67). This did not work as well as using explanatory variables, which were ultimately used to determine a unique splitting parameter for every person (p. 69).

Lastly, the construction of the booking numbers suggest the use of a Poisson model to describe recidivism. The booking number construction was described in detail earlier. It allows individuals to be uniquely identified as well as the number of times each person has been through the booking process, thus allowing for the definition of recidivism as being re-booked. The number of bookings over a period of time suggests a Poisson model can describe recidivism. However, if a person is booked once or twice over ten years, such a distribution choice may not be helpful. It would seem that a certain number of bookings—for example, five bookings—a Poisson distribution to describe the sequence of bookings would seem to be potentially useful.

This distribution might actually be useful to the jail authorities themselves in ways that a lognormal distribution would not. After a certain number of bookings, a Poisson distribution could be used to predict the probability of more bookings over a period of time in the future, and explanatory variables would not be necessary.

References

Asimov, I. (1982). *Foundation's edge*. Garden City, NY: Doubleday.

Auerhahn, K. (1999). Selective incapacitation and the problem of prediction. *Criminology*, *37*(4), 703-734. DOI: 10.1111/j.1745-9125.1999.tb00502.x

Barnett, A. (1987, January-February). Prison populations: A projection model. *Operations Research*, *35*(1), 218-234. DOI: 10.1287/opre.35.1.18

Carson, E. (2014, September). *Prisoners in 2013*, NCJ 247282. Bureau of Justice Statistics. Retrieved from http://www.bjs.gov/index.cfm?ty=pbdetail&iid=5109.

Cunniff, M. (2008, April). *Jail bed utilization analysis for February/March, 2003 & 2008*. Unpublished study, Stearns County, Minnesota. NIC TA#: 08J10.

Dobson, A., & Barnett, A. (2008). *An introduction to generalized linear models* (3rd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Durose, M., Cooper, A., & Snyder, H. (2014, April). *Recidivism of prisoners released in 30 states in 2005: Patterns from 2005 to 2010*, NCJ 244205. Bureau of Justice Statistics. Retrieved from http://www.bjs.gov/index.cfm?ty=pbdetail&iid=4986

Duwe, G., & Kerschner, D. (2007, October). Removing a nail from the boot camp coffin: An outcome evaluation of Minnesota's Challenge Incarceration Program. *Crime and Delinquency*, *54*, 614-643. DOI: 10.1177/0011128707301628

Frechet Distribution. (2015, April 14). In *Wikipedia, the free encyclopedia*. Retrieved July 18, 2015, from https://en.wikipedia.org/wiki/Fr%C3%A9chet_distribution.

Freedman, D. (2010). *Statistical models and causal inference*. New York, NY: Cambridge University Press.

International Centre for Prison Studies. (2014). *Highest to lowest-prison population total.* Website application. Retrieved from http://www.prisonstudies.org/highest-to-lowest/.

Lambert, P. (2007). Modeling the cure fraction in survival studies. *The Stata Journal*, *7*(3), 351-375.

Lee, E., & Wang, J. (2003). *Statistical methods for survival data analysis*. Hoboken, NJ: John Wiley & Sons, Inc.

Luo, T., & Stark, P. (2015). Nine out of ten restaurants fail? Check, please. *Significance*, *12*(2), 25-29. DOI: 10.1111/j.1740-9713.2015.00813.x

Maltz, M. ([1984] 2001). *Recidivism*. Originally published by Academic Press, Inc., Orlando, Florida. Retrieved from http://www.uic.edu/depts/lib/forr/pdf/crimjust/recidivism.pdf.

Maller, R., & Zhou, X. (1996). *Survival analysis with long-term survivors*. West Sussex, UK: John Wiley & Sons, Ltd.

Mosher, C., Miethe, T., & Phillips, D. (2002). *The mismeasure of crime*. Thousand Oaks, CA: Sage Publications.

Nagin, D. (2014). Deterrence and the death penalty: Why the statistics should be ignored. *Significance*, *11*(2), 9-13. DOI: 10.1111/j.1740-9713.2014.00733.x

Nagin, D., & Pepper, J. (Eds.). (2012). Report by the Committee on Deterrence and the Death Penalty, National Research Council. *Deterrence and the death penalty*. Washington, DC: The National Academies Press. Retrieved from https://www.law.upenn.edu/live/files/1529-nagin-full-reportpdf.

National Institute of Justice. (2014, June 17). *Recidivism*. Retrieved from http://www.nij.gov/topics/corrections/recidivism/Pages/welcome.aspx.

National Research Council. (1986). In A. Blumstein, J. Cohen, J. Roth, & C. Visher (Eds.), *Criminal careers and "career criminals," Volume 1.* Washington, DC: The National Academies Press.

Persson, I. (2002). *Essays on the assumption of proportional hazards in Cox regression* (Unpublished doctoral thesis), Uppsala University, Uppsala, Sweden. Retrieved from http://www.diva-portal.org/smash/get/diva2:161225/FULLTEXT01.pdf.

Schmidt, P., & Witte, A. (1988). *Predicting recidivism using survival models*. New York, NY: Springer-Verlag.

Shinnar, S., & Shinnar, R. (1975, Summer). The effects of the criminal justice system on the control of crime: A quantitative approach. *Law and Society Review*, *9*(4), 581-611. Article DOI: 10.2307/3053340

Singh, R., & Mukhopadhyay, K. (2011, October-December). Survival analysis in clinical trials: Basics and must know areas. *Perspectives in Clinical Research*, *2*(4), 145-148. DOI: 10.4103/2229-3485.86872

The Sentencing Project. (June, 2010). State Recidivism Database. Retrieved from http://www.sentencingproject.org/detail/Publication.cfm?Publication_id=311.

TheFreeDictionary.com (2014). *Definition of "Recidivism*." Retrieved from http://legal-dictionary.thefreedictionary.com/recidivism.

Upton, G., & Cook, I. (2006). *Oxford dictionary of statistics*. New York, NY: The Oxford University Press.

U.S. Dept. of Justice, Bureau of Justice Statistics. (1994). *Recidivism of felons on probation, 1986-1989 [United States]*. Conducted by Mark A. Cunniff and the National Association of Criminal Justice Planners. 2nd ICPSR ed. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [producer and distributor], 1994. DOI 10.3886/ICPSR09574.v2

Appendix A

List of Variable Names (Excel Columns) and Excel Formulae Used to Create Them

1. Booking_2

   To separate the first two figures from the full booking number, since this will indicate the year in which the booking number was generated.

   =REPLACE(A2,3,9,)

2. Booking_7

   To separate the first seven figures from the full booking number, since this will indicate the individual who is being booked.

   =REPLACE(A2,8,4,)

3. Booking_3

   To separate the last three figures from the full booking number, since this will indicate the number of bookings an individual has undergone up to that point in time.

   =REPLACE(A2,1,8,)

4. Skip_In_Booking

   Dummy variable to determine if there was a skip in booking number order, as indicated by the booking number extension.

   =IF(AND(C2=C1,D2-D1<>1),1,0)

5. Previous_Booking

   Dummy variable to determine if an individual had previous bookings not shown in the data.

   =IF(AND(C2<>C1,D2<>"001"),1,0)

6. First_Booking

   Dummy variable to determine if a booking was the first booking for an individual to appear in the data.

   =IF(OR(J2=1,D2="001"),1,0)

7. Booking_001

   Dummy variable to determine if a booking was a first booking.

   =IF(D2="001",1,0)

8. Booking_002

   Dummy variable to determine if a booking was a second booking.

   =IF(D2="002",1,0)

9. Booking_003

   Dummy variable to determine if a booking was a third booking.

   =IF(D2="003",1,0)

10. Booking_004

    Dummy variable to determine if a booking was a fourth booking.

    =IF(D2="004",1,0)

11. Booking_005

    Dummy variable to determine if a booking was a fifth booking.

    =IF(D2="005",1,0)

12. No_Release_Date

    Dummy variable to determine if a booking record has a release date and time.

    =IF(ISBLANK(G2),1,0)

13. Only_Booking_No_Release_Date

    Dummy variable to determine if a booking was the only booking for that person and if that booking lacked a release date and time.

    =IF(AND(ISBLANK(G2),C2<>C1,C2<>C3),1,0)

14. First_Booking_No_Release_Date

    Dummy variable to determine if an individual had multiple bookings and if the first booking had no release date and time.

    =IF(AND(ISBLANK(G2),C2<>C1,C2=C3),1,0)

15. Skip_In_Release_Date

    Dummy variable to determine if an individual had multiple bookings and if there was a missing release date between the first and last bookings.

    =IF(AND(ISBLANK(G2),C2=C1,C2=C3),1,0)

16. Last_Booking_No_Release_Date

    Dummy variable to determine if an individual had multiple bookings and if the last booking in the data had no release date and time.

    =IF(AND(ISBLANK(G2),C2=C1,C2<>C3),1,0)

17. Hours_Spent_In_Custody

    Decimal value for the length of time an individual spent in custody for a booking, in hours.

    =IF(AND(ISBLANK(G2)=FALSE,ISBLANK(H2)=FALSE),24*(G2-E2),"")

18. Days_Spent_In_Custody

    Decimal value for the length of time an individual spent in custody for a booking, in days.

    =IF(AND(ISBLANK(G2)=FALSE,ISBLANK(H2)=FALSE),DAYS(G2,E2),"")

19. Recidivism_Flag

    Dummy variable to determine if an individual's booking constituted a recidivism.

    =IF(C2=C3,1,0)

20. Censor_Flag

    Dummy variable to determine if a booking should be censored.

    =IF(C2<>C3,1,0)

21. Sum_Of_Flags

    Integer value to determine the sum of recidivism and censor flags.

    =T2+U2

22. Recidivism_Time_Days

    Integer value to determine how long it took for a recidivism to occur, in days.

    =IF(AND(X2=1,T2=0),DAYS(E3,G2),IF(AND(X2=1,T2=1),DAYS(E3,E2)/2,""))

23. Follow_Time_Days

    Integer to determine how long an individual's follow time should be, in days.

    =IF(AND(Y2=1,ISBLANK(G2)=FALSE),DAYS("2/23/2015",G2),(IF(AND(Y2=1,ISBLANK(G2
)=TRUE),DAYS("2/23/2015",E2)+8.14,"")))

24. Recidivism_Follow_Time_Days

    This column combined the information contained in the Recidivism_Time_Days and Follow_Time columns. This was to make it possible for SAS to process.

    =IF(W2="",X2,W2)

Appendix B

SAS Program Codes

1. SAS code to import the data from Excel.

```
proc import out=work.Timing
datafile= "C:\SAS_test\2015_07_15_Timing_Data.xlsx"
DBMS = xlsx replace;
run;
```

2. SAS code to plot the recidivism and censor times.

```
ods graphics on;
ods select ParameterEstimates GoodnessOfFit FitQuantiles MyHist;
proc lifereg data = mylib.Timing_Data_2015_07_15;
model Recidivism_Follow_Time * Censor_Flag(1) =  /d = lnormal;
output out=New;
proc univariate data = New;
histogram / midpoints=0.2 to 1.8 by 0.2;
run;
ods graphics off;
```

3. SAS code to generate the survival, probability density, and hazard function using the lifetable method.

```
ods graphics on;
proc lifetest method = lt data = mylib.Timing_Data_2015_07_15 plots = (s h p)
notable;
        time Recidivism_Follow_Time * Censor_Flag(1);
run;
ods graphics off;
```

4. SAS code to produce the fundamental Kaplan-Meier survival plot of the booking data from Stearns County.

```
ods graphics on;
proc lifetest data = mylib.Timing_Data_2015_07_15 atrisk plots = survival(nocensor
atrisk (atrisktick) = 0 to 4000 by 500 cb) notable;
        time  Recidivism_Follow_Time * Censor_Flag(1);
label Recidivism_Follow_Time = "Recidivism Survivial Time(Days)";
run;
ods graphics off;
```

5. SAS code to obtain mean and median statistics of booking data.

```
proc lifetest data = mylib.Timing_Data_2015_07_15;
        time  Recidivism_Follow_Time * Censor_Flag(1);
run;
```

6. SAS code to print a frequency table regarding the gender variable.

```
PROC freq DATA = mylib.Timing_Data_2015_07_15;
tables Gender;
RUN;
```

7. SAS code to produce frequency table for the genders (that takes into account censoring) as well as Kaplan-Meier curves for each gender.

```
ods graphics on;
proc lifetest data=mylib.Timing_Data_2015_07_15 atrisk plots=survival(nocensor
atrisk (atrisktick)= 0 to 4000 by 500 cb)notable;
        strata gender;
        time  Recidivism_Follow_Time * Censor_Flag(1);
label Recidivism_Follow_Time = "Recidivism Survivial Time(Days)";
run;
ods graphics off;
```

8. SAS code to produce Kaplan-Meier plots that separate genders and first bookings.

```
proc lifetest data = mylib.Timing_Data_2015_07_15 plots = survival(nocensor atrisk
(atrisktick)= 0 to 4000 by 500) notable;
        time Recidivism_Follow_Time * Censor_Flag(1);
        strata Booking_001 Gender;
run;
ods graphics off;
```

9. SAS code to product gender vs. booking number extension.

```
ODS RTF FILE = 'C:\SAS_test\SAS_Results.RTF';
options nodate nonumber ;
PROC freq DATA = mylib.Timing_Data_2015_07_15;
tables gender * booking_3 /crosslist;
RUN;
ODS RTF CLOSE;
```

10. SAS code to produce the survival proportion in six month intervals over a five year period for each gender.

```
ods output  LifetableEstimates=est2;
proc lifetest data= mylib.Timing_Data_2015_07_15 method=lt intervals=(0 182.625
to 1826.25 by 182.625);
        time Recidivism_Follow_Time * Censor_Flag(1);
        strata Gender;
proc print data = est2 noobs;
var lowertime uppertime failed failure;
run;
```

11. SAS code to produce the survival proportion in six month intervals over a five year period for each gender.

```
ods output  LifetableEstimates=est2;
proc lifetest data= mylib.Timing_Data_2015_07_15 method=lt intervals=(0 182.625
to 1826.25 by 182.625);
        time Recidivism_Follow_Time * Censor_Flag(1);
        strata Gender Booking_001;
proc print data = est2 noobs;
var lowertime uppertime failed failure;
run;
```

12. SAS code to produce the survival proportion in six month intervals over a five year period for booking levels -001 and -002.

```
ods output  LifetableEstimates=est2;
proc lifetest data= mylib.Timing_Data_2015_07_15 method=lt intervals=(0 182.625
to 1826.25 by 182.625);
time Recidivism_Follow_Time * Censor_Flag(1);
        strata Booking_001 Booking_002;
proc print data = est2 noobs;
var lowertime uppertime failed failure;
run;
```

13. SAS code to produce the Kaplan-Meier survival plots for first booking and non-first booking individuals.

```
ods graphics on;
proc lifetest data = mylib.Timing_Data_2015_07_15 plots = survival(nocensor atrisk
(atrisktick)= 0 to 4000 by 500) notable;
        time Recidivism_Follow_Time * Censor_Flag(1);
        strata Booking_001;
run;
ods graphics off;
```

14. SAS code to produce the Kaplan-Meier survival plots for first booking, second booking, and higher booking levels.

```
ods graphics on;
proc lifetest data = mylib.Timing_Data_2015_07_15 plots = survival(nocensor atrisk
(atrisktick)= 0 to 4000 by 500) notable;
time Recidivism_Follow_Time * Censor_Flag(1);
        strata Booking_001 Booking_002;
run;
ods graphics off;
```

15. SAS code to produce the Kaplan-Meier survival plots for the first booking, second booking, third booking, and higher booking levels.

```
ods graphics on;
proc lifetest data = mylib.Timing_Data_2015_07_15 plots = survival(nocensor atrisk
(atrisktick)= 0 to 4000 by 500) notable;
        time Recidivism_Follow_Time * Censor_Flag(1);
        strata Booking_001 Booking_002 Booking_003;
run;
ods graphics off;
```

16. SAS code to produce the Kaplan-Meier survival plots for the first booking, second booking, third booking, fourth booking, and higher booking levels.

```
ods graphics on;
proc lifetest data = mylib.Timing_Data_2015_07_15 plots = survival(nocensor atrisk
(atrisktick)= 0 to 4000 by 500) notable;
        time Recidivism_Follow_Time * Censor_Flag(1);
        strata Booking_001 Booking_002 Booking_003 Booking_004;
run;
ods graphics off;
```

17. SAS code to produce the Kaplan-Meier survival plots for the first booking, second booking, third booking, and higher booking levels.

```
ods graphics on;
proc lifetest data = mylib.Timing_Data_2015_07_15 plots = survival(nocensor atrisk
(atrisktick)= 0 to 4000 by 500) notable;
        time Recidivism_Follow_Time * Censor_Flag(1);
        strata Booking_001 Booking_002 Gender;
run;
ods graphics off;
```

18. SAS code to create the lognormal model of survival time without gender.

```
proc lifereg data = mylib.Timing_Data_2015_07_15;
model Recidivism_Follow_Time * Censor_Flag(1) =  /d = lnormal;
output out=New cdf=Prob;
run;
```

19. SAS code to produce the CDF of the lognormal model without gender.

```
ods graphics on / ANTIALIASMAX=88800;
proc sgplot data=New;
scatter x = Recidivism_Follow_Time y = Prob; discretelegend;
run;
ods graphics off;
```

20. SAS code to create the lognormal model of survival time with gender.

```
proc lifereg data = mylib.Timing_Data_2015_07_15;
class gender;
model Recidivism_Follow_Time * Censor_Flag(1) = gender /d = lnormal;
output out=New cdf=Prob;
run;
```

21. SAS code to produce the CDF of the lognormal model with gender.

```
ods graphics on / ANTIALIASMAX=88800;
proc sgplot data=New;
scatter x = Recidivism_Follow_Time y = Prob / group=Gender;
discretelegend;
run;
ods graphics off;
```

22. SAS code to produce the lognormal probability plot without gender.

```
ods graphics on / ANTIALIASMAX=88800;
proc lifereg data = mylib.Timing_Data_2015_07_15;
        model Recidivism_Follow_Time * Censor_Flag(1) = /d = lnormal;
probplot nocenplot;
inset;
run;
ods graphics off;
```

23. SAS code to produce the lognormal probability plot with gender.

```
ods graphics on / ANTIALIASMAX=88800;
proc lifereg data = mylib.Timing_Data_2015_07_15;
        class Gender;
        model Recidivism_Follow_Time * Censor_Flag(1) = gender /d = lnormal;
probplot nocenplot;
inset;
run;
ods graphics off;
```