2020

MEASURING CHANGE: PREDICTION OF EARLY ONSET SEPSIS

Aric Schadler
University of Kentucky, schadler@uky.edu
Author ORCID Identifier:
https://orcid.org/0000-0001-8814-5834

Digital Object Identifier: https://doi.org/10.13023/etd.2020.393

Aric Schadler, Student

Dr. Arnold Stromberg, Major Professor

Dr. Katherine Thompson, Director of Graduate Studies

MEASURING CHANGE:
PREDICTION OF EARLY ONSET SEPSIS

DISSERTATION

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy in the
College of Arts and Sciences
at the University of Kentucky

By

Aric Daniel Schadler

Lexington, Kentucky

Director: Dr. Arnold Stromberg, Professor of Statistics

Lexington, Kentucky

2020

https://orcid.org/0000-0001-8814-5834

ABSTRACT OF DISSERTATION

MEASURING CHANGE:
PREDICTION OF EARLY ONSET SEPSIS

Sepsis occurs in a patient when an infection enters into the blood stream and spreads throughout the body causing a cascading response from the immune system. Sepsis is one of the leading causes of morbidity and mortality in today's hospitals. This is despite published and accepted guidelines for timely and appropriate interventions for septic patients. The largest barrier to applying these interventions is the early identification of septic patients. Early identification and treatment leads to better outcomes, shorter lengths of stay, and financial savings for healthcare institutions. In order to increase the lead time in recognizing patients trending towards septicemia a multivariate discrimination model was developed to create an early identification sepsis score to identify patients who are starting to show signs of sepsis. The model utilizes the patient's heart rate, respiratory rate, systolic blood pressure, temperature, and oxygen saturation and the change from each of their respective baselines. Patient specific baselines are based on each patient's previous vital sign measures leading up to the current set of measures.

Theoretical assumptions are applied to this sepsis score to investigate distributional properties of the measure for applicable inferences. Finally, a new approximation to the degrees of freedom of a t-distribution, ν_s, is proposed. This new approximation is investigated and compared to the Satterthwaite approximation.

KEYWORDS: Sepsis, Measuring Change, t-Distribution, Degrees of Freedom

Aric Daniel Schadler

(Name of Student)

07/22/2020

Date

MEASURING CHANGE:
PREDICTION OF EARLY ONSET SEPSIS

By
Aric Daniel Schadler

Dr. Arnold Stromberg
Director of Dissertation

Dr. Katherine Thompson
Director of Graduate Studies

07/22/2020
Date

DEDICATION

To my incredible wife who without her support this would not be possible.

ACKNOWLEDGMENTS

I would like to express my sincerest thanks to everyone who has played a role in this pursuit and to each of you on my committee. I especially want to thank Dr. Stromberg, Dr. Bauer, and Dr. Wood who have played vital roles of support and encouragement throughout this process. I am truly thankful for your patience and your advice in working with me throughout this endeavor.

TABLE OF CONTENTS

ACKNOWLEDGMENTS iii

LIST OF TABLES viii

LIST OF FIGURES ix

CHAPTER 1. UNDERSTANDING SEPSIS 1

1.1 Background 1

CHAPTER 2. Contrasting Vital Signs of Standard and Sepsis Patients 8

2.1 Introduction 8

2.2 Body Mass Index (BMI) 8

2.3 Admission Age 11

2.4 Systolic Blood Pressure 14

2.5 Respiratory Rate 17

2.6 Heart Rate 20

2.7 Peripheral Capillary Oxygen Saturation 23

2.8 Body Temperature 26

2.9 Vital Signs Comparison Summary 30

CHAPTER 3. Simulating the vital signs 32

3.1 Introduction 32

3.2 Systolic Blood Pressure for Non-Sepsis Patients........ 32

3.3 Systolic Blood Pressure for Sepsis Patients........ 34

3.4 Respiratory Rate for Non-Sepsis Patients........ 36

3.5 Respiratory Rate for Sepsis Patients........ 38

3.6 Heart Rate for Non-Sepsis Patients........ 40

3.7 Heart Rate for Sepsis Patients........ 42

3.8 Peripheral Capillary Oxygen Saturation........ 44

3.9 Peripheral Capillary Oxygen Saturation for Sepsis Patients........ 46

3.10 Body Temperature for Non-Sepsis Patients........ 48

3.11 Body Temperature for Sepsis Patients........ 50

CHAPTER 4. Measuring Change........ 53

4.1 Introduction........ 53

4.2 Heart Rate........ 54

4.3 Systolic Blood Pressure........ 55

4.4 Respiratory Rate........ 56

4.5 Peripheral Capillary Oxygen Saturation........ 57

4.6 Body Temperature........ 58

4.7 Summary........ 59

CHAPTER 5. Discussion on Variation........ 60

5.1 Introduction 60

5.2 Heart Rate Standard Deviation 61

5.3 Respiratory Rate Standard Deviation 61

5.4 Systolic Blood Pressure Standard Deviation 62

5.5 Peripheral Capillary Oxygen Saturation Standard Deviation 63

5.6 Body Temperature Standard Deviation 64

5.7 Summary 65

CHAPTER 6. The Sepsis Score 67

6.1 Introduction 67

6.2 The Sepsis Score 67

6.3 Utilizing the Sepsis Score 69

6.4 Example Patients 71

6.4.1 Patient One 73

6.4.2 Patient Two 74

6.4.3 Patient Three 75

6.5 Results from Trial 75

6.6 Conclusions 77

CHAPTER 7. Statistical Theory 78

7.1 Theory 78

7.2 Simulations 89

7.2.1 Simulating An from Different Distributions 91

7.2.2 Approximating the Sum of t-distributions 94

7.2.3 Simulation to Test if $Q5$ has a t-Distribution 108

7.2.4 Simulation to Test if $Q15$ has a t-Distribution 112

7.2.5 Normality of Qk when *k* Ranges from Five to Fifty 115

7.2.6 $Qk \sim t$ for Finite Samples when *k* Ranges from Five to Fifty 117

7.2.7 Introduction and Simulation using $vs = 2s2s2 - 1$ 122

7.2.8 The Satterthwaite Approximation 126

CHAPTER 8. Conclusions 133

8.1 Future Research 134

BIBLIOGRAPHY 135

VITA 137

LIST OF TABLES

Table 5.1 Summary of Variation Among Standard Deviations of Vital Signs 65

Table 6.1 Summary of 200 Sepsis Patients 76

Table 7.1 Comparison of Exact vs. Satterthwaite Degrees of Freedom 87

Table 7.2 Comparison of Exact vs. Satterthwaite Confidence Intervals 88

Table 7.3 Proportion of Confidence Intervals that Contained μ 89

Table 7.4 Normality of An when $Xn \sim N(0,1)$ 93

Table 7.5 Testing the Normality of the Sum of Five t-Distributed Random Variables .. 101

Table 7.6 Testing Normality of the Mean of Five t-Distributed Random Variables 104

Table 7.7 Comparing Mean and Standard Deviation for $S5, R5,$ and $Q5$ Computed with t-Distributions with Seven Degrees of Freedom 106

Table 7.8 P-values of KS Test for $H0: Q5 \sim tv$ $(n = 1000)$. The columns represent the seed degrees of freedom in $Q5$ and the rows are the degrees of freedom in tv 111

Table 7.9 P-values of KS Test for $H0: Q15 \sim tv$ $(n = 10{,}000)$. The columns represent the seed degrees of freedom in $Q15$ and the rows are the degrees of freedom in tv 114

Table 7.10 P-values of KS Test for $H0: Qk \sim N\mu, \sigma$ $(n = 10{,}000)$. Values of *k* are going across the columns and the rows represent the seed degrees of freedom in Qk... 117

Table 7.11 P-values for KS Test for Qk Comprised of $t5 \sim tv$ $(n = 1000)$. Values of *k* are going across the columns and the rows represent the degrees of freedom in tv. 119

Table 7.12 P-values for KS Test for Qk Comprised of $t10 \sim tv$ $(n = 1000)$. Values of *k* are going across the columns and the rows represent the degrees of freedom in tv. 120

Table 7.13 P-values for KS Test for Qk Comprised of $t15 \sim tv$ $(n = 1000)$. Values of *k* are going across the columns and the rows represent the degrees of freedom in tv. 121

Table 7.14 Calculated vs for Different k and Underlying tv in Qk. Values of *k* are going across the columns and the rows represent the seed degrees of freedom in Qk. 124

Table 7.15 P-values for $H0: Qk \sim tvs$. Values of *k* are going across the columns and the rows represent the seed degrees of freedom in Qk. 125

Table 7.16 Values of $vs, vsat$, and p-values from KS Tests $(n = 6)$ 129

Table 7.17 Values of $vs,\ vsat$, and p-values from KS Tests $(n = 10)$ 131

LIST OF FIGURES

Figure 1.1 Track and trigger scoring system used at the University of Kentucky Albert B. Chandler Hospital 3
Figure 1.2 Illustration of variation between two different patient's heart rates........ 6
Figure 2.1 BMI of Non-Sepsis Patients Histogram 9
Figure 2.3 BMI of Sepsis Patients Histogram 10
Figure 2.4 BMI Composite Non-Sepsis vs. Sepsis........ 11
Figure 2.5 Age of Non-Sepsis Patients Histogram 12
Figure 2.6 Age of Sepsis Patients Histogram 13
Figure 2.7 Admission Age Composite Non-Sepsis vs. Sepsis 14
Figure 2.8 Systolic Blood Pressure of Non-Sepsis Patients Histogram........ 15
Figure 2.9 Systolic Blood Pressure of Sepsis Patients Histogram........ 16
Figure 2.10 Systolic Blood Pressure Composite Non-Sepsis vs. Sepsis 17
Figure 2.11 Respiratory Rates of Non-Sepsis Patients Histogram 18
Figure 2.12 Respiratory Rates of Sepsis Patients Histogram 19
Figure 2.13 Respiratory Rate Composite Non-Sepsis vs. Sepsis........ 20
Figure 2.14 Heart Rate of Non-Sepsis Patients Histogram........ 21
Figure 2.15 Heart Rate of Sepsis Patients Histogram........ 22
Figure 2.16 Heart Rate Composite Non-Sepsis vs. Sepsis 23
Figure 2.17 Peripheral Capillary Oxygen Saturation of Non-Sepsis Patients Histogram 24
Figure 2.18 Peripheral Capillary Oxygen Saturation of Sepsis Patients Histogram........ 25
Figure 2.19 Peripheral Capillary Oxygen Saturation Composite Non-Sepsis vs. Sepsis . 26
Figure 2.20 Temperature in Degrees Celsius for Non-Sepsis Patients Histogram........ 27
Figure 2.21 Temperature in Degrees Celsius for Sepsis Patients Histogram 28
Figure 2.22 Temperature in Celsius Composite Non-Sepsis vs. Sepsis 30
Figure 3.1 Simulated Systolic Blood Pressure for Non-Sepsis Patients Histogram........ 33
Figure 3.2 Systolic Blood Pressure for Non-Sepsis Patients Composite Simulated vs. Patient Data........ 34
Figure 3.3 Simulated Systolic Blood Pressure for Sepsis Patients Histogram 35
Figure 3.4 Systolic Blood Pressure for Sepsis Patients Composite Simulated vs. Patient Data 36
Figure 3.5 Simulated Respiratory Rate for Non-Sepsis Patients Histogram 37
Figure 3.6 Respiratory Rate for Non-Sepsis Patients Composite Simulated vs. Patient Data 38
Figure 3.7 Simulated Respiratory Rate for Sepsis Patients Histogram 39
Figure 3.8 Respiratory Rate for Sepsis Patients Composite Simulated vs. Patient Data.. 40
Figure 3.9 Simulated Heart Rate for Non-Sepsis Patients Histogram........ 41
Figure 3.10 Heart Rate for Non-Sepsis Patients Composite Simulated vs. Patient Data . 42

Figure 3.11 Simulated Heart Rate for Sepsis Patients Histogram 43
Figure 3.12 Heart Rate for Sepsis Patients Composite Simulated vs. Patient Data........ 44
Figure 3.13 Simulated $SpO2$ for Non-Sepsis Patients Histogram 45
Figure 3.14 $SpO2$ for Non-Sepsis Patients Composite Simulated vs. Patient Data 46
Figure 3.15 Simulated $SpO2$ for Sepsis Patients Histogram 47
Figure 3.16 $SpO2$ for Sepsis Patients Composite Simulated vs. Patient Data 48
Figure 3.17 Simulated Temperature (Celsius) for Non-Sepsis Patients Histogram 49
Figure 3.18 Temperature (Celsius) for Non-Sepsis Patients Composite Simulated vs. Patient Data 50
Figure 3.19 Simulated Temperature (Celsius) for Sepsis Patients Histogram........ 51
Figure 3.20 Temperature (Celsius) for Sepsis Patients Composite Simulated vs. Patient Data 52
Figure 4.1 Non-Sepsis vs. Sepsis – Heart Rate........ 54
Figure 4.2 Non-Sepsis vs. Sepsis – Systolic Blood Pressure........ 55
Figure 4.3 Non-Sepsis vs. Sepsis – Respiratory Rate........ 56
Figure 4.4 Non-Sepsis vs. Sepsis – Peripheral Capillary Oxygen Saturation 57
Figure 4.5 Non-Sepsis vs. Sepsis: Body Temperature (Celsius) 58
Figure 5.1 Histogram of Heart Rate Standard Deviations for Non-Sepsis Patients 61
Figure 5.2 Histogram of Respiratory Rate Standard Deviations for Non-Sepsis Patients 62
Figure 5.3 Histogram of Systolic Blood Pressure Standard Deviations for Non-Sepsis Patients........ 63
Figure 5.4 Histogram of $SpO2$ Standard Deviations for Non-Sepsis Patients 64
Figure 5.5 Histogram of Temperature (Celsius) Standard Deviations for Non-Sepsis Patients........ 65
Figure 6.1 Projection of Data into a Two-Dimensional Space using the Discriminant Model. Sepsis observations are blue circles and non-sepsis observation are denoted with a red x........ 69
Figure 6.2 Visualizing the Sepsis Score. The figure identifies the line of discrimination in black, the sepsis score in blue, and a two standard deviation control chart lines in red. 70
Figure 6.3 Flow Chart for Sepsis Alerts 71
Figure 6.4 Legend for Vital Sign Measures........ 72
Figure 6.5 Sepsis Score and Vital Signs for Patient One........ 73
Figure 6.6 Sepsis Score and Vital Signs for Patient Two 74
Figure 6.7 Sepsis Score and Vital Signs for Patient Three 75
Figure 6.8 Lead Time between Alert and Antibiotics Histogram........ 77
Figure 7.1 Histogram of An when $Xn \sim t5$ 94
Figure 7.2 Histogram of $i = 15ti, 1$ 95
Figure 7.3 Histogram of $i = 15ti, 2$ 96
Figure 7.4 Histogram of $i = 15ti, \nu$ for $\nu = 3, 4, 5,$ and 6 97
Figure 7.5 Histogram of $i = 15ti, \nu$ for $\nu = 7, 8, 9,$ and 10 98

Figure 7.6 Histogram of $i = 15ti, \nu$ for $\nu = 20, 30,$ and 40 99
Figure 7.7 Histogram of $i = 15ti, 10$ with 10,000 Observations 102
Figure 7.8 Comparing the Sum vs. Mean of Five t-Distributed Random Variables with Seven Degrees of Freedom 103
Figure 7.9 Comparing $S5, R5,$ and $Q5$ Comprised of t-Distributed Random Variables with Seven Degrees of Freedom 107
Figure 7.10 Comparing $S5, R5,$ and $Q5$ Comprised of t-Distributed Random Variables with 20 Degrees of Freedom 108
Figure 7.11$Q5$ Comprised of $X\nu, i \sim t10$ Compared to a t-Distribution with Ten Degrees of Freedom 110
Figure 7.12 $Q15$ Comprised of $X\nu, i \sim t10$ Compared to a t-distribution with Ten Degrees of Freedom 113
Figure 7.13 Histogram of $Q10$ Comprised of t-distributions with Six Degrees of Freedom 116
Figure 7.14 Histograms of $\nu k, sat$ for $k = 2, 3, 4,$ and 5 $(n = 6)$ 128
Figure 7.15 Histograms of Wk for $k = 2, 3, 4,$ and 5 $(n = 6)$ 129
Figure 7.16 Histograms of $\nu k, sat$ for $k = 2, 3, 4,$ and 5 $(n = 10)$ 130
Figure 7.17 Histograms of Wk for $k = 2, 3, 4,$ and 5 $(n = 10)$ 131

CHAPTER 1. UNDERSTANDING SEPSIS

1.1 Background

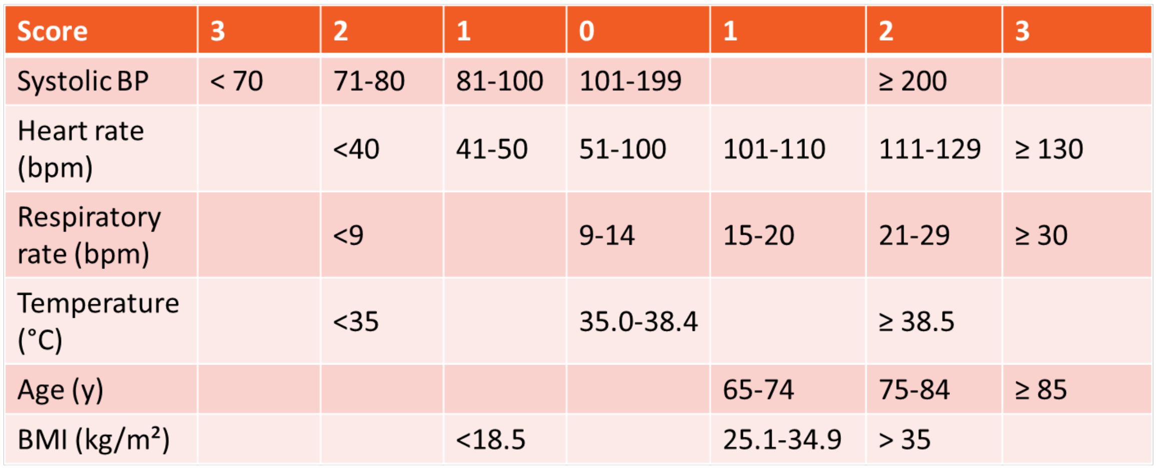
Sepsis occurs in a patient when an infection that started in one part of their body enters into the blood stream and spreads throughout the body causing a systemic response from the immune system. The initial infection can come from a variety of sources including pneumonia, a skin infection, central line infection, a surgical site infection, or a urinary tract infection. The response from the immune system has a sweeping effect across a person's vital signs, leading to a rising temperature, rapid breathing and heart rate, and a change in blood pressure. According to the Centers for Disease Control and Prevention (CDC), more than 1.7 million people develop sepsis each year in the United States and around 270,000 of these expire each year from sepsis. This is a significant increase from what the CDC published as recently as 2015 stating that 1.5 million develop sepsis each year in the United States and with around 250,000 sepsis related mortalities (CDC 2015). A staggering 33% of people who pass away in a hospital have sepsis as a contributing factor. In fact, a CDC study discovered that seven out of 10 patients with sepsis either recently had healthcare services or had a chronic disease requiring frequent medical care (CDC 2020). A study published in The Journal of the American Medical Association (JAMA) which included 173,690 patients with a sepsis diagnosis found that patients in their study with hospital-acquired sepsis had a 25.5% mortality rate compared to only 13.4% of patients who were admitted with sepsis (Rhee, et al 2017).

This connection between sepsis and healthcare is not a coincidence. Healthcare facilities are breeding grounds for bacteria. They are the destination both for a person who has a severe infection and a person who has a serious wound. Additionally, many of the devices used to help patients in hospitals also create easy paths for infections to enter into the body. Central lines, ventilators, catheters, and peripheral intravenous lines give bacteria and other germs access into the body. According to the CDC, one in twenty-five patients contracts a hospital-acquired infection. Infections that originate in hospitals are a serious risk for patients in today's modern healthcare facilities and a looming problem for healthcare systems. The U.S. Department of Health and Human Services declared the prevention and reduction of healthcare-associated infections a top priority (CDC 2015). Healthcare systems are required to absorb the costs of treating the effects of a hospital-acquired condition. The cost of treating a single central line associated infection can be nearly $50,000 (Daley, 2015). According to Michael Eber in a paper published in the Archives of Internal Medicine in 2010, there were $8.1 billion in-hospital costs attributable to healthcare-acquired sepsis and pneumonia in 2006 (Eber 2010). This is a number that will continue to climb higher over time, as infections become more difficult to treat due to antibiotic resistant strains and as the costs of healthcare continue to rise.

Healthcare systems continue to pour money into preventing infections in hospitals. Personal protective equipment, hand hygiene protocols, and countless staff education programs are aimed at reducing and preventing the spread of infections between patients. Hospital-acquired infections that go unnoticed or untreated can lead to sepsis, and sepsis that is not treated in a timely and efficient manner leads to longer hospital stays, additional hospital costs, and too often, death.

Sepsis is one of the leading causes of morbidity and mortality in today's hospitals. This is despite published and accepted guidelines for timely and appropriate interventions for septic patients. The largest barrier to applying these interventions is the timely detection of septic patients. Early identification and treatment leads to better outcomes, shorter lengths of stay, and financial savings for healthcare institutions. Historically, early identification systems have utilized threshold based scoring systems. These systems usually define values on a set of vital signs and demographic variables. When a variable exceeds its threshold, then a point value is assigned and when the sum of these point values reaches a cutoff value, an alert is generated. Different methods will use different sets of variables, different point values, or different threshold values, but they generally follow the same pattern.

UK Healthcare in Lexington, Kentucky utilizes a track and trigger system that follows the scoring criteria depicted in Figure 1, modified from the 2007 work of Dr. Abel Kho, et al.



Score	3	2	1	0	1	2	3
Systolic BP	< 70	71-80	81-100	101-199		≥ 200	
Heart rate (bpm)		<40	41-50	51-100	101-110	111-129	≥ 130
Respiratory rate (bpm)		<9		9-14	15-20	21-29	≥ 30
Temperature (°C)		<35		35.0-38.4		≥ 38.5	
Age (y)					65-74	75-84	≥ 85
BMI (kg/m^2)			<18.5		25.1-34.9	> 35	

Figure 1.1 Track and trigger scoring system used at the University of Kentucky Albert B. Chandler Hospital

Following this criteria, anyone who has a heart rate between 51 and 100 beats per minute (bpm) receives a point value of zero for the heart rate variable. Once a person's heart rate passes the 100 bpm value, they receive a point. When it passes 110 bpm they receive two points, and so on. The point values for each variable are summed and an alert is generated upon this total reaching another threshold. In this example, if the sum reaches a value of six, an alert fires to the rapid response team to evaluate the patient.

There are a few inherent challenges to this method. Many patients will achieve the alerting criteria from the moment they enter the hospital until the moment they leave. This is especially true with more obese and elderly patients. Response teams have tried to counteract this alert fatigue by adjusting alert criteria, or for patients who continually alert, they will try contacting the patient's nurse to first have a conversation before racing to the room.

The contrasting issue is patients who never meet the alert criteria, even when they are in a state of decline. This will often happen for younger or more fit patients. A runner who is less than forty years of age may have a resting heart rate between 40 and 50 beats per minute. When this person is in the hospital, their heart rate may never reach the 100 bpm threshold even when in distress. This type of alert criteria does not account for the differences in individual's baseline metrics where one person's resting heart rate could be 95 beats per minute, another person's resting heart rate could be 55 beats per minute. Thus, a person with a resting heart rate of 55 beats per minute, could have a significant rise in their heart rate and jump to 100 beats per minute, an 82% increase, and only score one additional point on this scale. Conversely, the person who has a resting heart rate of 95 beats per minute could rise to 100 beats per minute, a 5.2% increase would receive the same one point on the scale.

To address these challenges a new methodology is proposed that evaluates not only a patient's current vital signs, but also monitors differentiation of the individual's immediate values from their baselines. This is accomplished by tracking the patient throughout their entire hospital visit in order to continuously update patient specific baselines and to incorporate the change from the current set of vital signs. A patient with a two day, or two week, length of stay has a wealth of data collected which can be used to establish a baseline and a measure of variability for each of their vital signs. Thus, the baseline heart rate of a patient over the course of their stay is leveraged as a comparison to their current heartrate. By comparing the current set of vital sign measures to the patient's baseline measures valuable information is added to the model which aides in the discrimination between patients who are in a state of decline versus patients who are not. At any given point of time, each patient in the hospital has established a baseline heart rate, respiratory rate, blood pressure, etc. for their current visit. When a new set of vital signs are collected and entered into the electronic medical record, the vitals are compared to the patient's baseline values for their visit.

Not every baseline measure is the same, however. Look at the Figure 1.2, below, representing the heart rates of two patients, Patient A (blue) and Patient B (red). The two graphs have approximately the same mean leading up to the final observation of a 98. Notice that the red line has much less variation as compared to the blue line and as such, the observation of 98 is a more striking difference for the patient represented by the red line. The change in the current set of vitals from the baseline is essential, but it is also important to keep the change in perspective with the natural variation exhibited by the patient.

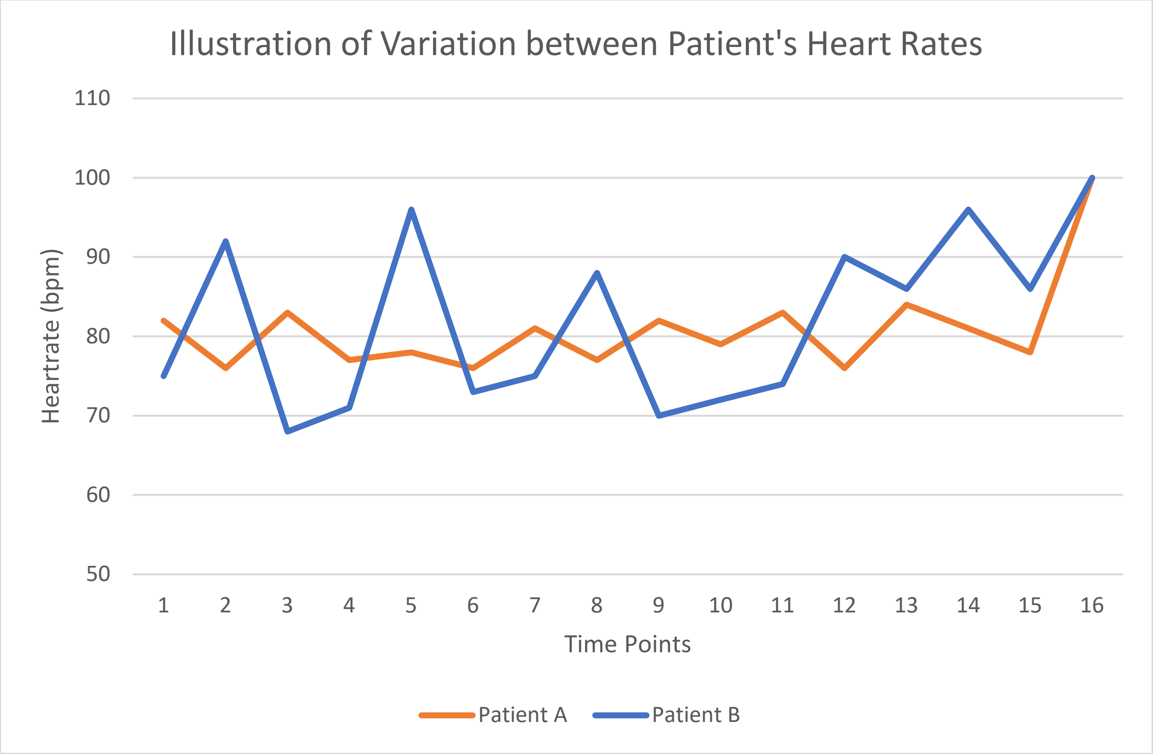


Figure 1.2 Illustration of variation between two different patient's heart rates

Under this new paradigm, if a fit younger patient, who has resting heart rate in the fifties, has a jump up to 98 bpm, the model will recognize this change as a significant jump from the patient's established baseline. In the traditional model as shown in Figure 1.1, this change would have gone unnoticed since this change did not cross the 100 bpm threshold.

The vital sign measures that were included in the model were heart rate, respiratory rate, systolic blood pressure, temperature, and oxygen saturation (SpO_2). Along with the raw scores for these vitals sign measures, the change from baseline variable was also included for each variable along with the age and body mass index of the patient. Lab values were excluded from the model since they are measured much less consistently and would have a dramatic impact to the lead time gained from the model. Where vital sign measurements are usually taken between three and six times daily, lab values, if taken at all, are usually only drawn once a day.

At this point, it is also important to note the scope of this research. This model is focused on identifying patients who are in an acute or progressive level of care who acquire sepsis during their hospital stay. This model is not intended to identify patients with sepsis present on admission, as they would not have baseline information available and they are fully evaluated upon admission. It is also not designed for patients in an intensive care model who are consistently monitored by physicians and nurses. Instead, this model is aimed at finding patients outside of intensive care units who develop sepsis during the course of their stay in a hospital. These patients often develop sepsis from a laceration or device used in the course of their care, which has allowed an infection access into the blood stream. The most common sources are surgical site infections, central line infections, catheter associated infections or ventilator associated infections. The goal of this research is to increase the lead time in recognizing these patients trending towards sepsis who are often left on the floor for hours or days before the deterioration is identified. This delay can have significant impact to patient outcomes and the cost of the care incurred by the hospital.

CHAPTER 2. CONTRASTING VITAL SIGNS OF STANDARD AND SEPSIS PATIENTS

2.1 Introduction

In order to better understand the changes in vital signs between a non-sepsis patient and a sepsis patient, the distribution of each vital sign is examined from a sample of in-patients between 2012 and 2014. This data was pulled from UKHealthCare's electronic medical record located in Lexington, Kentucky. The sample of vital sign data was collected along with diagnosis data in order to distinguish between the patient groups. For those patients with a sepsis diagnosis, data was filtered to only observations directly before and after the sepsis diagnosis in order to obtain information about a patient who was entering into or actively in a septic state. By determining the distributions of the different vital signs and demographics, we can hope to gain a better understanding of the effect of sepsis on patients. Additionally, for each vital sign measure, a simulation of the data is conducted for both groups. This will allow for a better understanding of the information and for inferences to be made on the distributions.

2.2 Body Mass Index (BMI)

Beginning with the demographic data, the body mass index (BMI) of non-sepsis and sepsis patients are visualized with separate histograms. BMI is a measure of body fat in adult men and women and is a calculation based on a person's height and weight. The graph below illustrates the distribution of BMI of non-sepsis patients. The histogram is overlaid with a line representing a gamma distribution with a shape parameter equal to 14.2, a scale parameter of 2.2, and a threshold equal to -2.9.

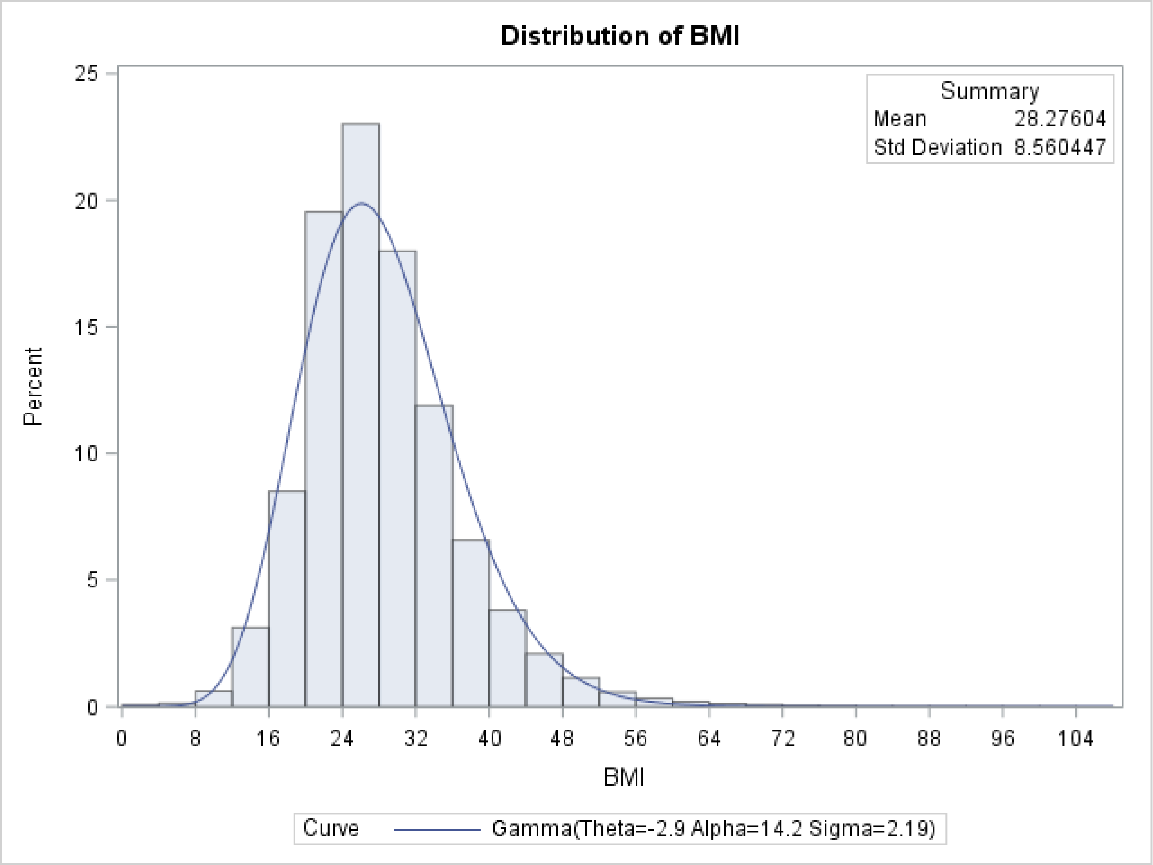


Figure 2.1 BMI of Non-Sepsis Patients Histogram

The gamma distribution seems to fit the data well with the addition of the slightly higher peak on the left side of the distribution.

Looking at sepsis patients, the graph below shows a histogram of BMI measurements. Again, the overlay line shows a gamma distribution with a shape parameter of 9.7, a scale parameter of 3.1 and a threshold of -1.09.

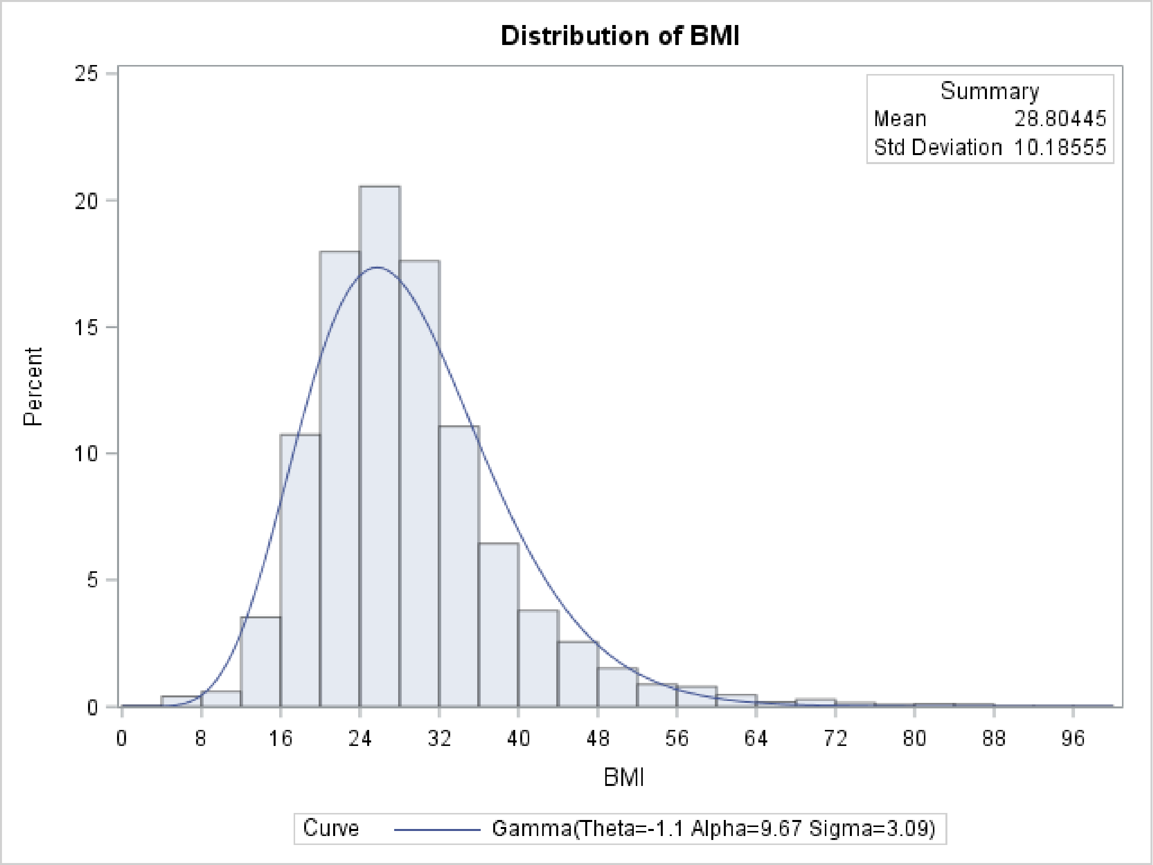


Figure 2.2 BMI of Sepsis Patients Histogram

When comparing the distributions of BMI between non-sepsis and sepsis, the means, 28.3 versus 28.8, and the standard deviations 8.6 versus 10.2, respectively, are quite similar. Looking at the composite of the two distributions, Figure 2.3 below, BMI remains consistent across both groups. As such, it does not appear to be a good predictor of sepsis.

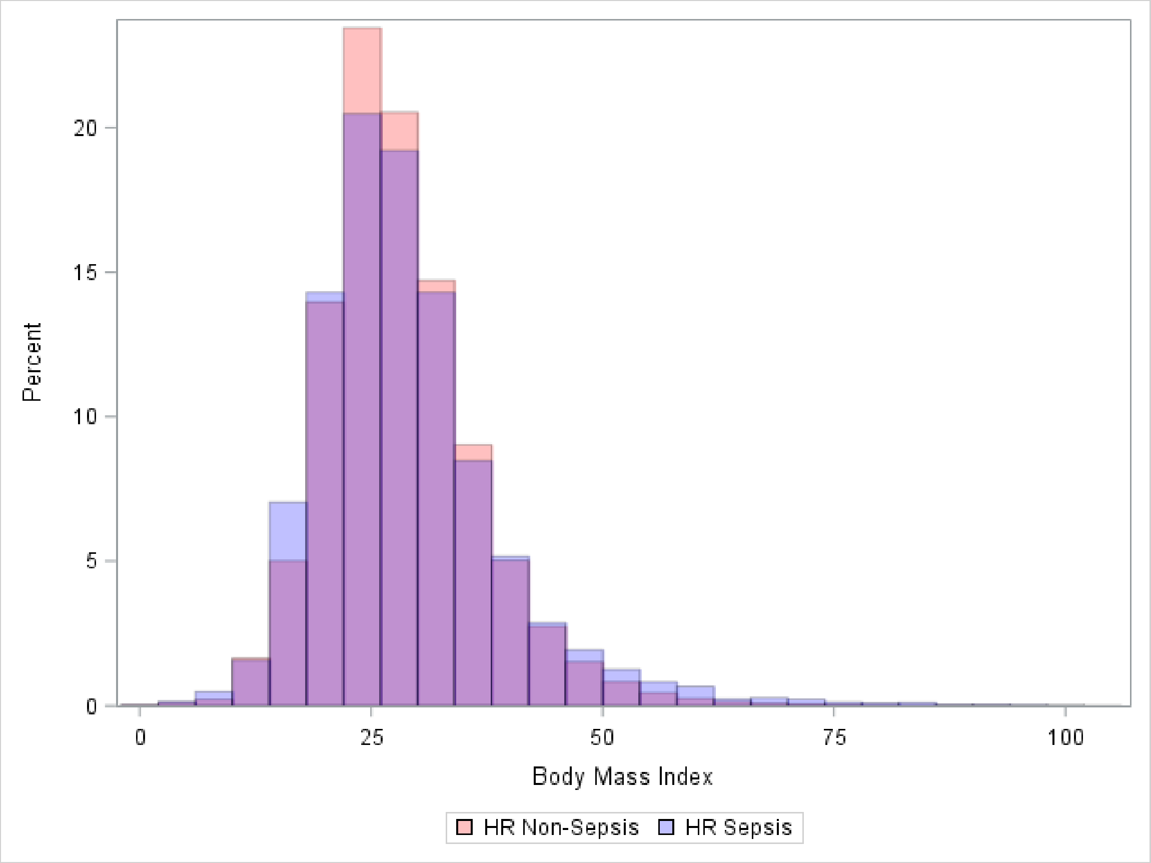


Figure 2.3 BMI Composite Non-Sepsis vs. Sepsis

2.3 Admission Age

The age, measured in years, of the patient at admission is included in the traditional track and trigger methodology. Looking at the distribution of the age of patients with and without a sepsis diagnosis, they will be investigated to observe differences between the groups. The graph below shows the distribution of age of non-sepsis patients. The mean of the distribution is 53.2 years, with a standard deviation of 17.9 years and a median of 54.

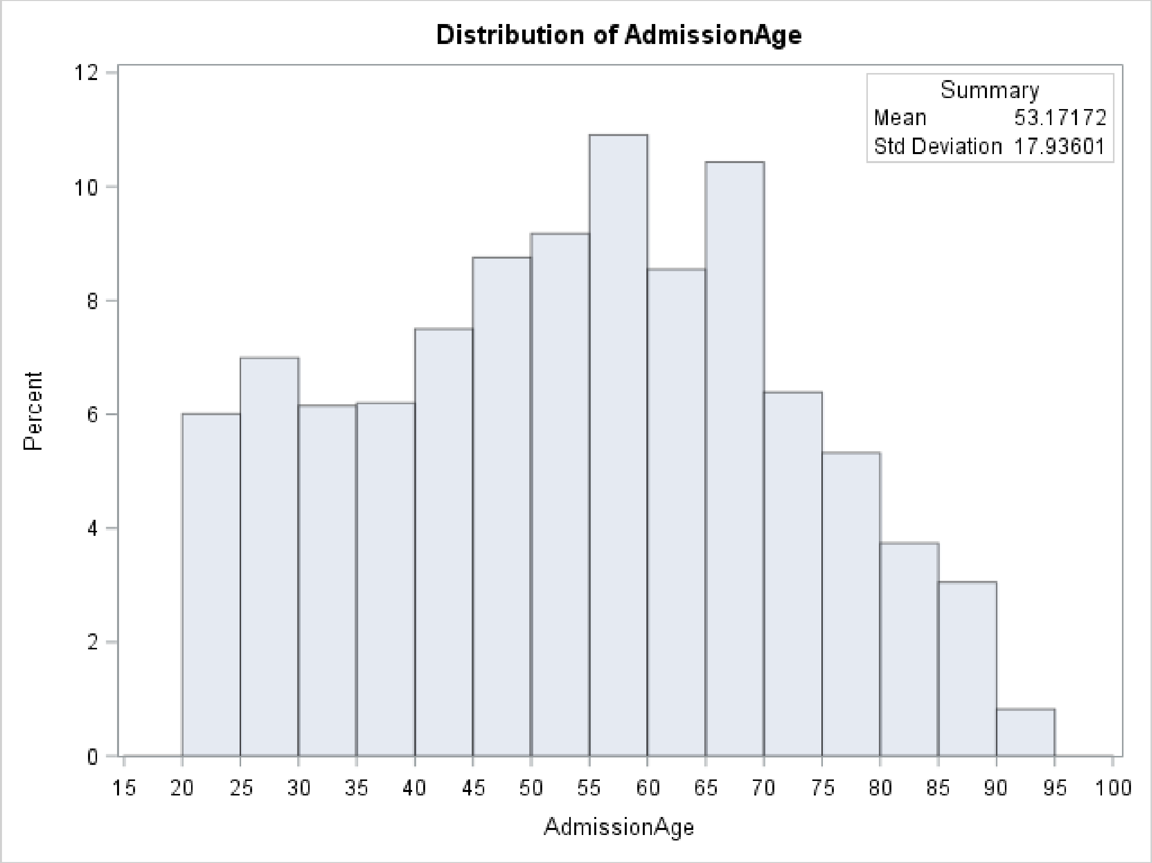


Figure 2.4 Age of Non-Sepsis Patients Histogram

This distribution of age for sepsis patients is similarly shaped with a mean of 57.0 years, a standard deviation of 16.4 years and a median of 57 years. Clearly a patient's age is not affected by the onset of sepsis, however it is worth investigating if different age ranges are more or less susceptible to sepsis.

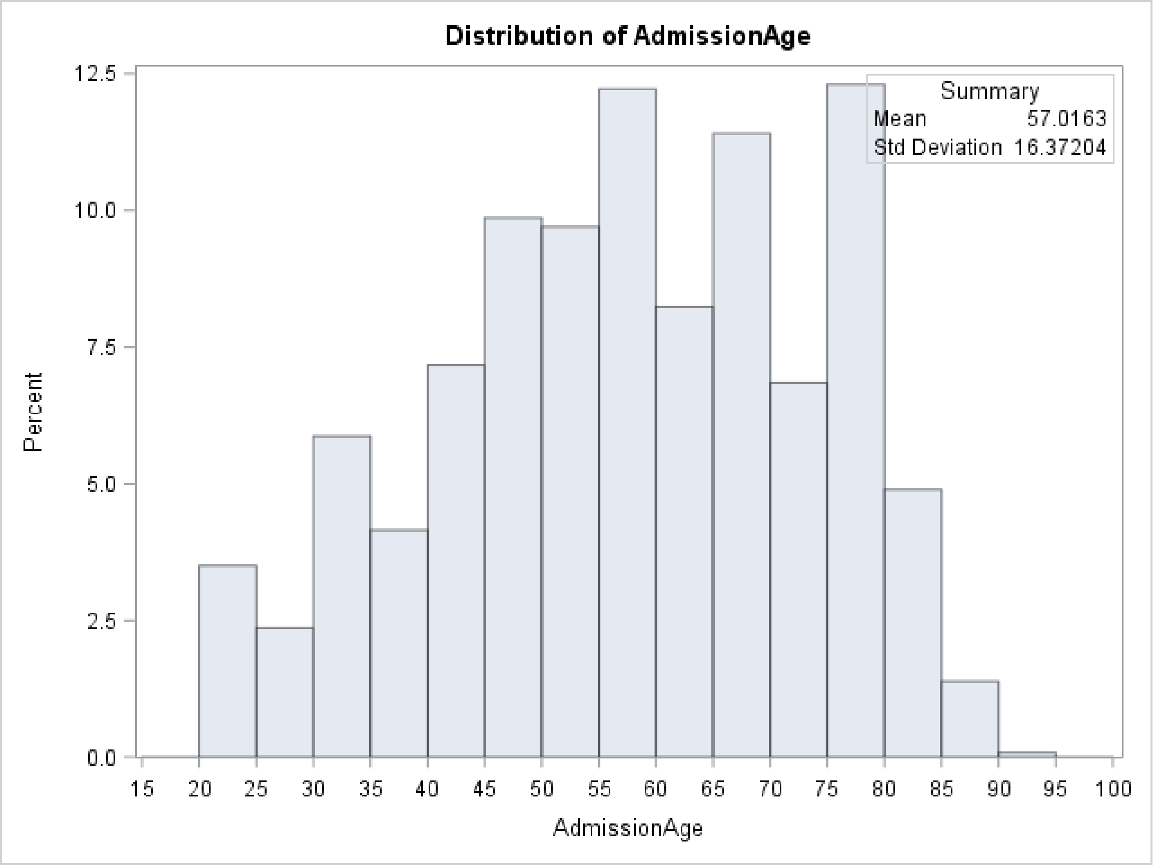


Figure 2.5 Age of Sepsis Patients Histogram

Looking at the measures of centrality and the composite graph below, you will notice that the sepsis population is slightly older, on average 57 years, as compared to the non-sepsis population, 53 years.

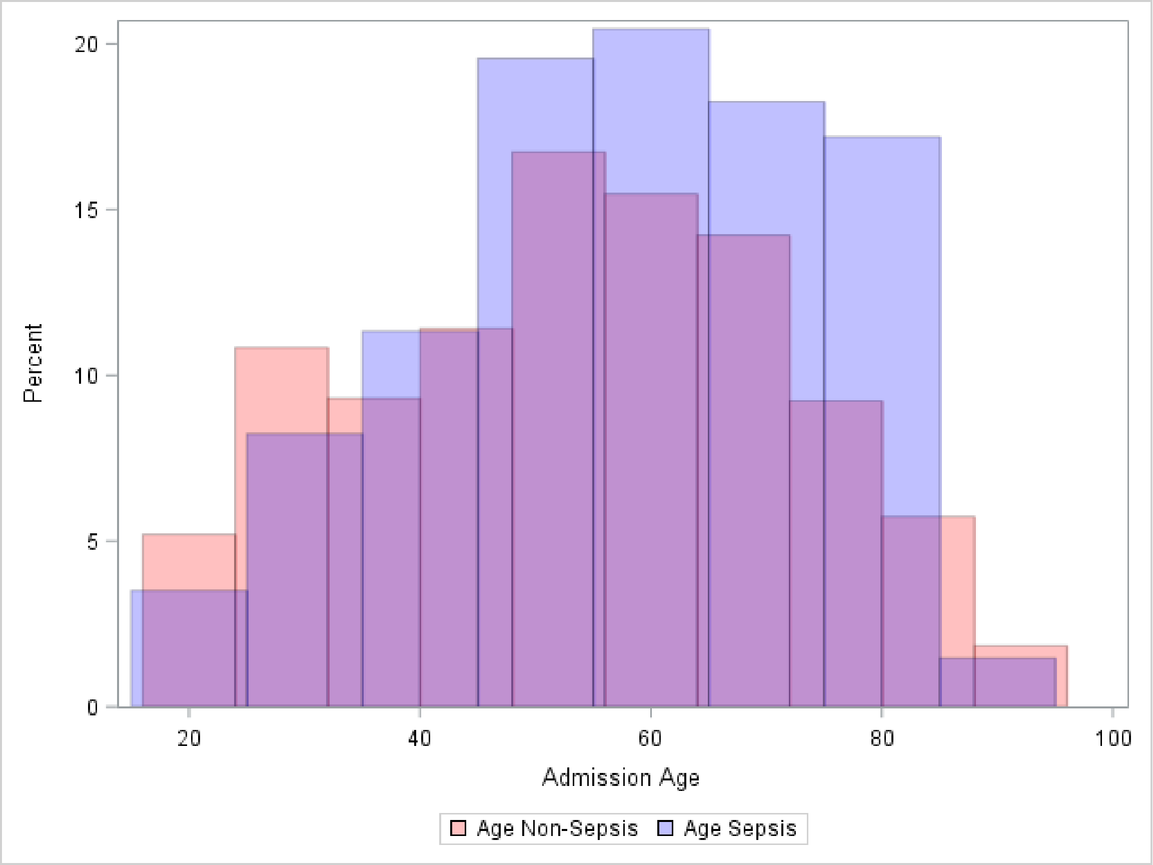


Figure 2.6 Admission Age Composite Non-Sepsis vs. Sepsis

2.4 Systolic Blood Pressure

Beginning with the distributions of the vital signs, Figure 2.7 below shows the distribution of systolic blood pressure measurements of non-sepsis patients. The blue line overlaid onto the graph illustrates a normal distribution with a mean of 126.7 and a standard deviation of 22.1.

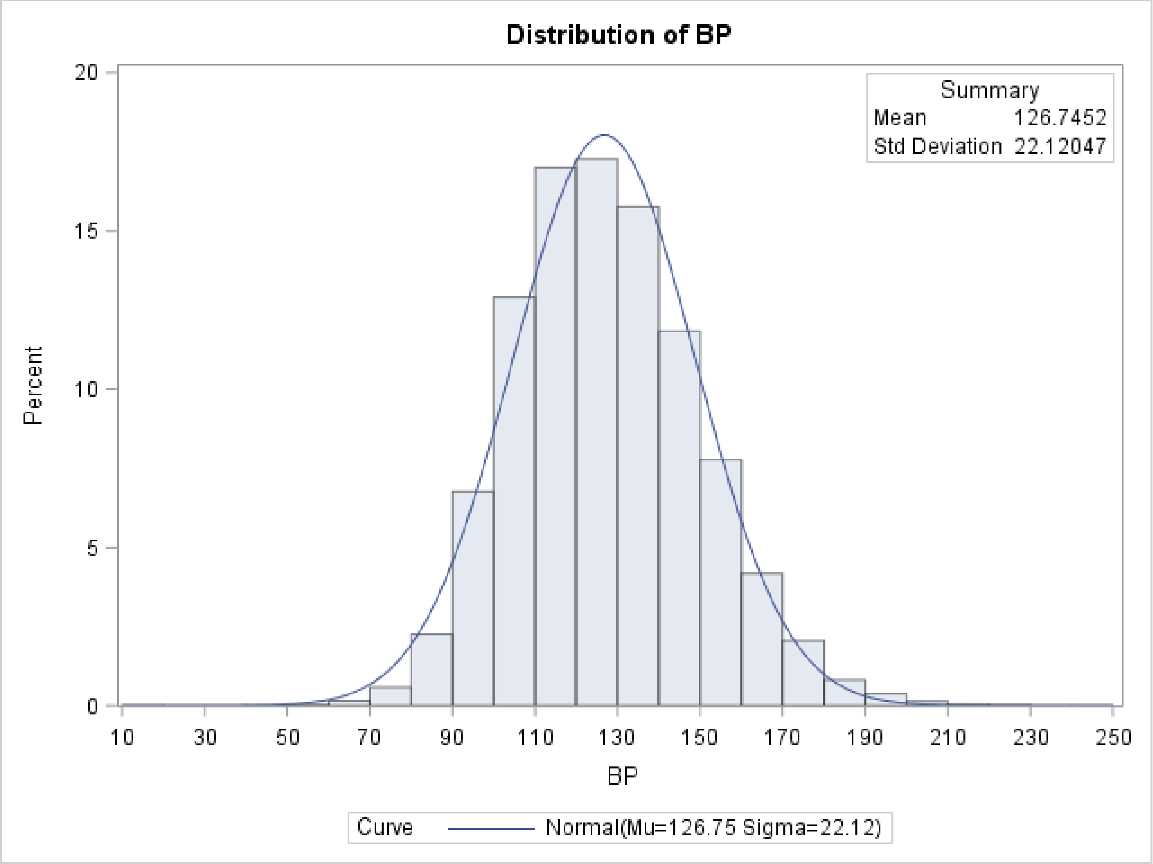


Figure 2.7 Systolic Blood Pressure of Non-Sepsis Patients Histogram

Now looking at a histogram of systolic blood pressure for sepsis patients, the graph below shows this data with an overlaid line in an effort to identify the distribution. The overlaid line illustrates a log normal distribution with a zeta of 4.73 and a sigma of 0.24. When comparing the shape of systolic blood pressure for non-sepsis patients to sepsis patients, not only is there a systematic shift downward, but also the overall shape of the distribution changes from a symmetric normal shape to a non-symmetric distribution skewed to the right.

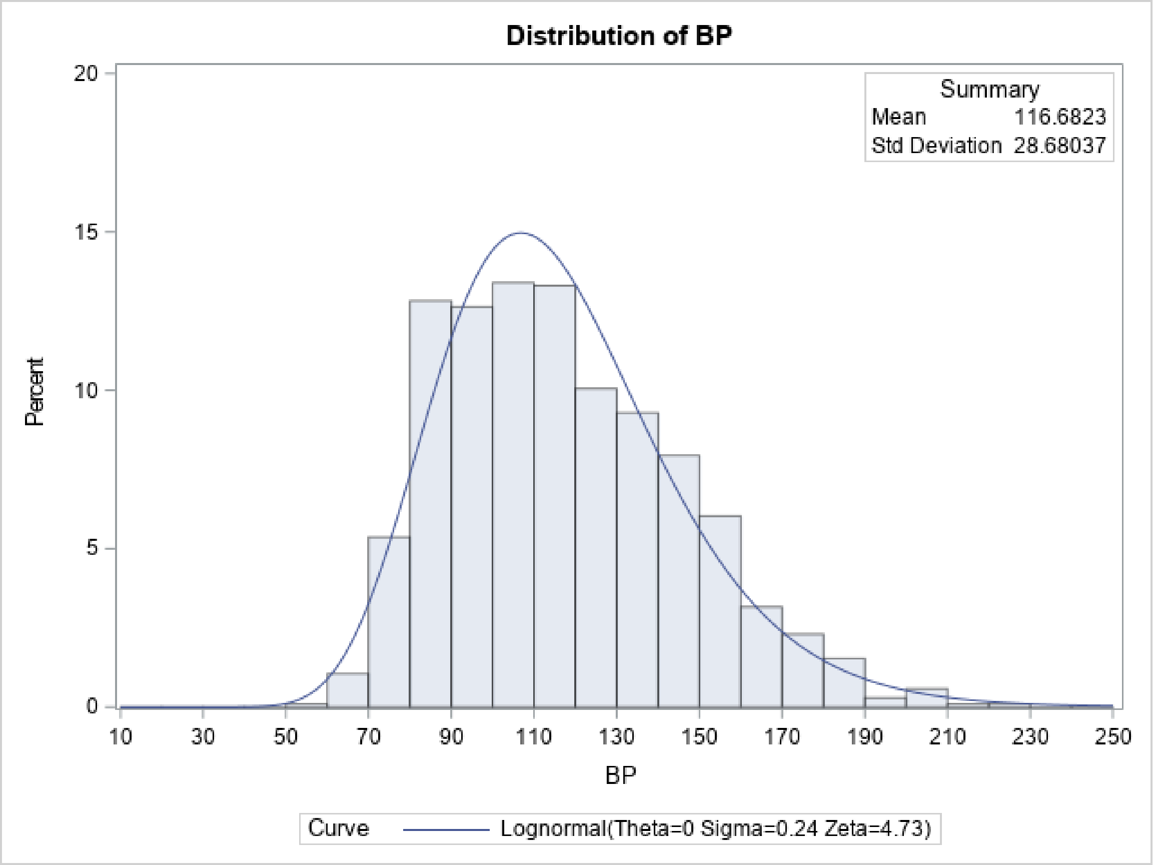


Figure 2.8 Systolic Blood Pressure of Sepsis Patients Histogram

Figure 2.9 below illustrates these two distributions graphed on top of each other. The mean of the distribution drops from 126.7 for non-sepsis patients down to 116.7 for sepsis patients. The standard deviation also changes from a 22.1 for non-sepsis patients to 28.7 for sepsis patients. Considering that a systolic blood pressure of 120 is considered normal (Cleveland Clinic 2014), non-sepsis patients have 39.7 percent of observations that fall below 120 compared to 58.7 percent of observations for sepsis patients.

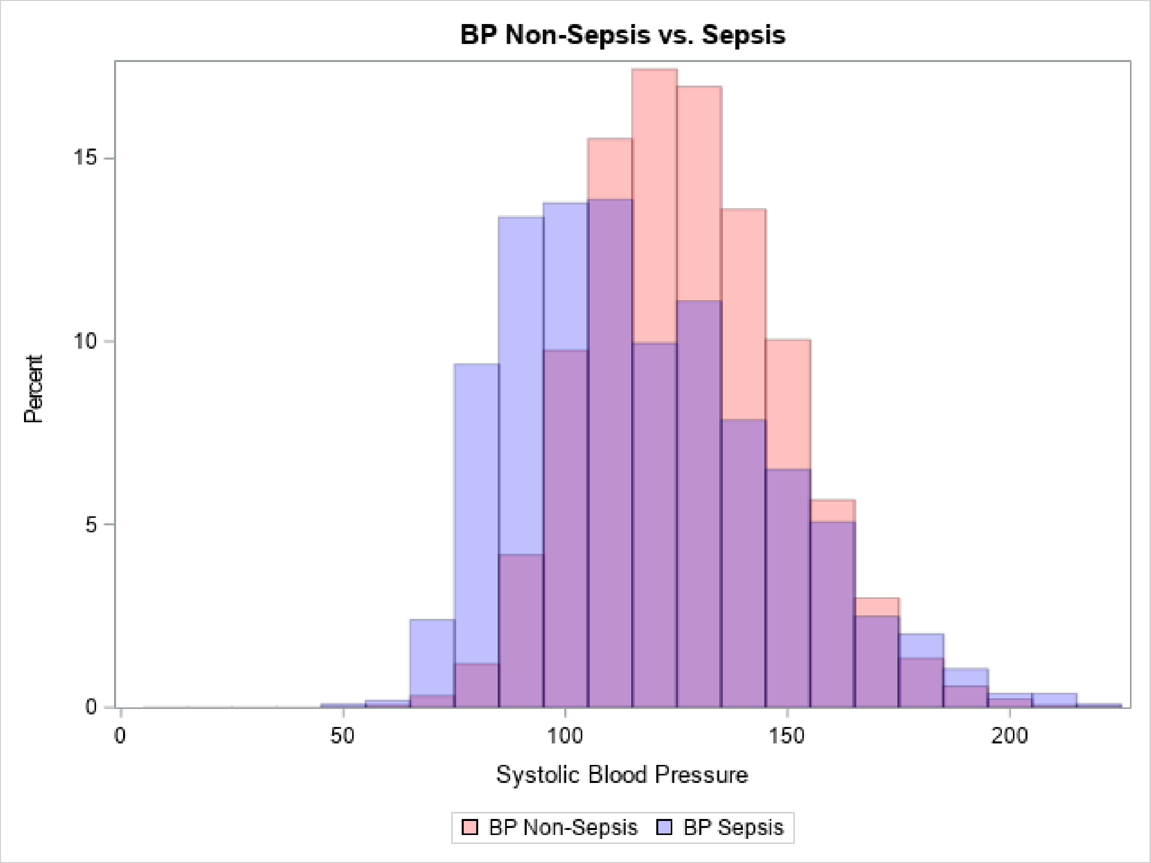


Figure 2.9 Systolic Blood Pressure Composite Non-Sepsis vs. Sepsis

2.5 Respiratory Rate

A patient's respiratory rate measures the number of breaths taken by the patient in a minute. The graph below shows the histogram of respiratory rate readings for non-sepsis patients, where the overlay line illustrates a gamma distribution with a shape parameter equal to 68.4, a scale parameter of 0.450, and a threshold parameter of -13.1. Additionally, the mean of the distribution is 17.7 with a standard deviation of 3.8. As you can see, this distribution fits the data marginally well, with a higher peak at the center and a few outliers on the high side.

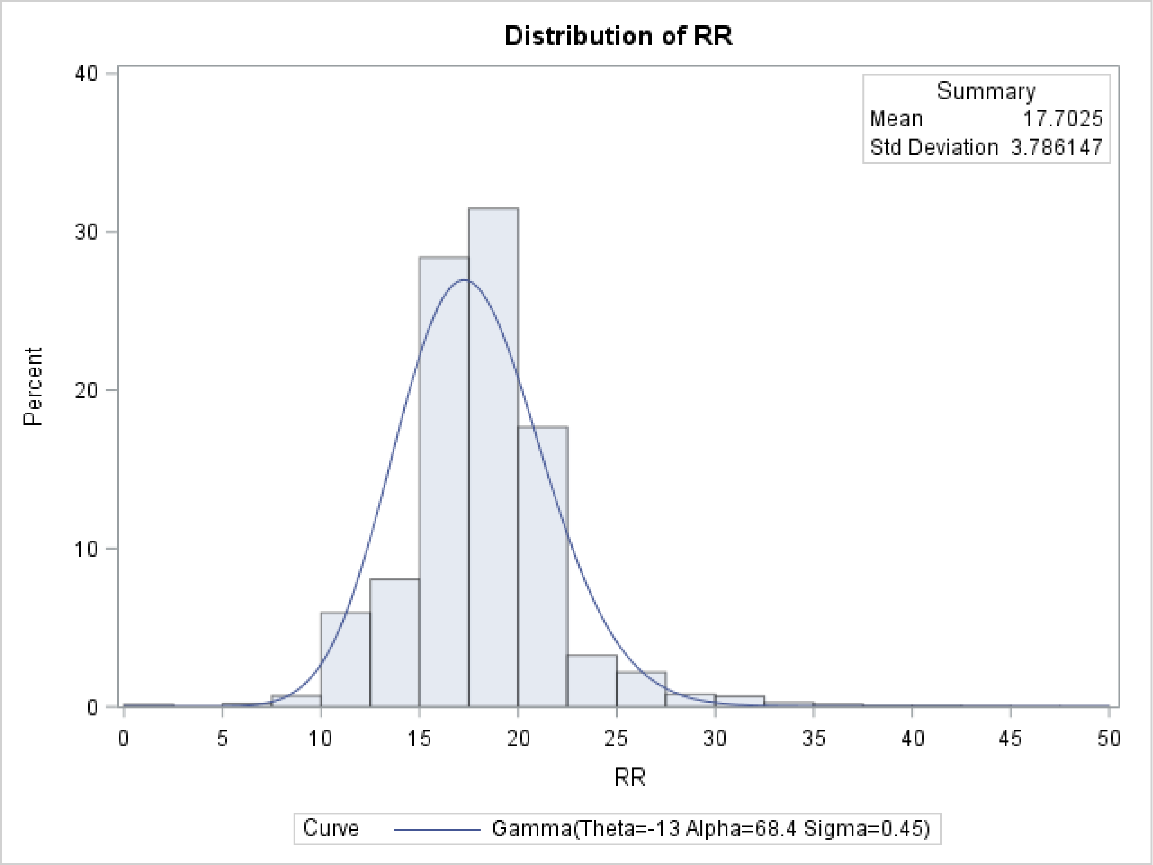


Figure 2.10 Respiratory Rates of Non-Sepsis Patients Histogram

Figure 2.11 below shows the distribution of respiratory rates of patients who had a diagnosis of sepsis. The overlay line illustrates a gamma distribution superimposed onto the distribution with a shape parameter of 16.7, a scale parameter of 1.6 and a threshold equal to -7.0. Notice that the distribution fits the gamma well with the exception of the high peak centered between 15 and 17.5.

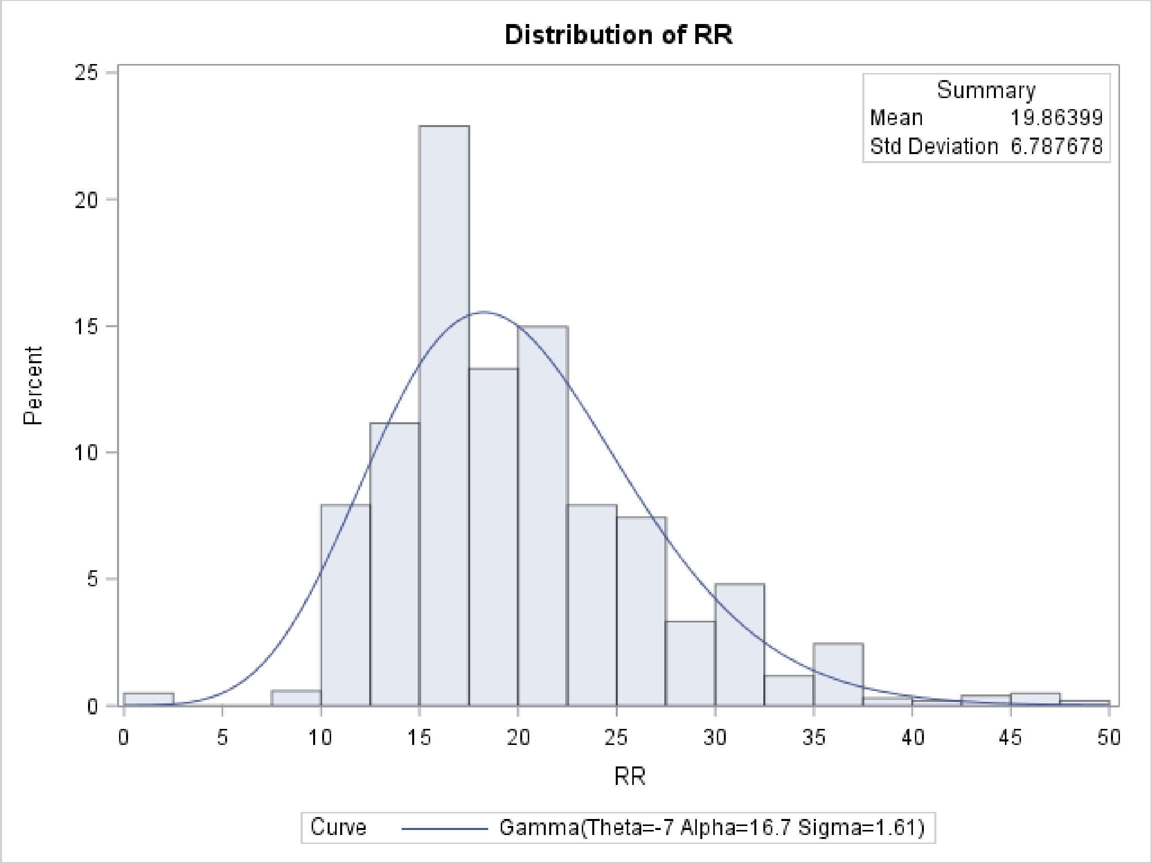


Figure 2.11 Respiratory Rates of Sepsis Patients Histogram

There is an interesting shift in the distribution of a patient's respiratory rate when they have a sepsis diagnosis compared to a non-sepsis patient. Looking at Figure 2.10 of non-sepsis patients, it is observed that the distribution is narrow with the majority, 93.1%, of the observations falling between 12 and 25 breaths per minute, the normal range for an adult's respiratory rate (Cleveland Clinic 2014). With the onset of sepsis into the human system, the overall distribution of respiratory rates has a shift upwards and a larger standard deviation, 6.8 breaths per minute, a 78.9 % increase over the non-sepsis standard deviation of only 3.8. Additionally, more than three times the observations fell outside of 12 to 25 breaths per minute for sepsis patients, 21.4%, comparted to 6.9% for non-sepsis patients. Figure 2.12 below shows the distributions of respiratory rates for non-sepsis patients and sepsis patients together to better illustrate the distributional shift between the patients. Notice the large increase of observations in the tails of the distribution of Sepsis patients (blue) over non-sepsis patients (red).

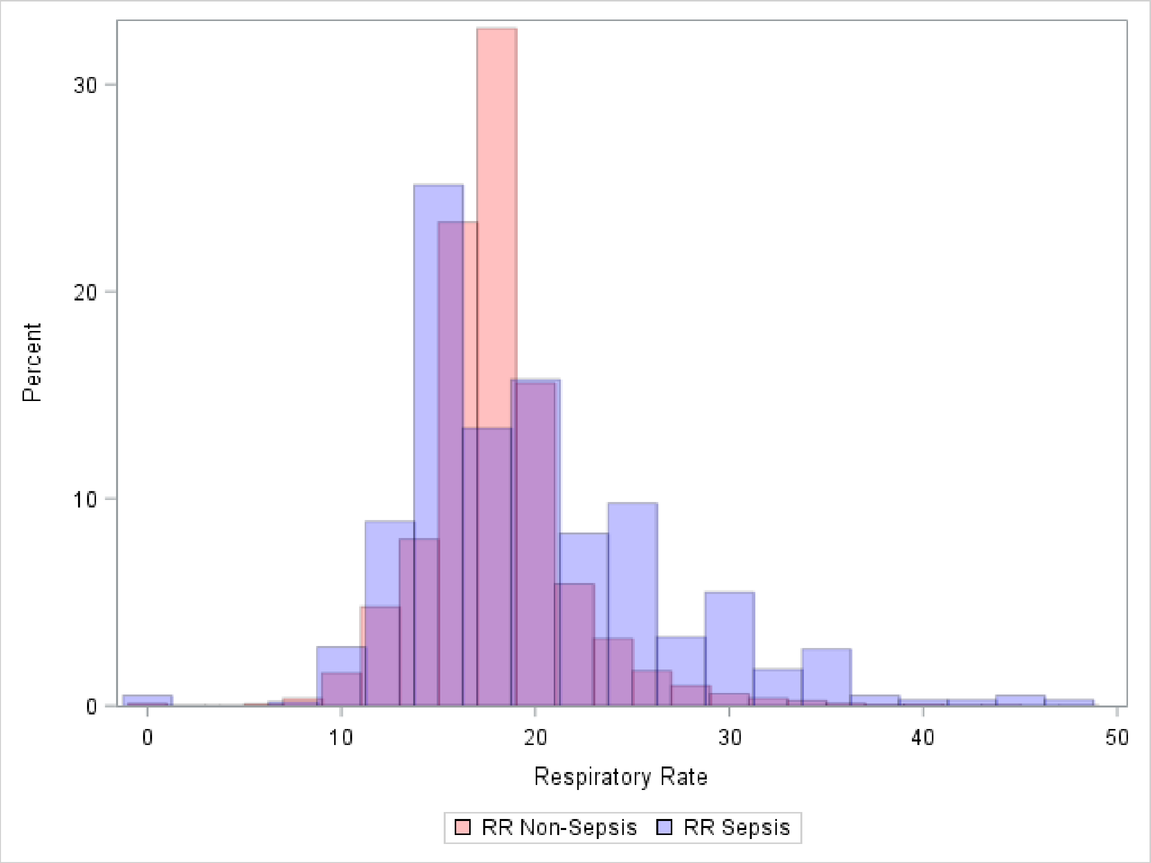


Figure 2.12 Respiratory Rate Composite Non-Sepsis vs. Sepsis

2.6 Heart Rate

The heart rate measures the number of heartbeats a patient has during a span of one minute. The following graph shows a histogram of heart rates for non-sepsis patients. The overlay line illustrates a Normal distribution with a mean of 82.2 and a standard deviation of 16.4. The median of the distribution is 81 heartbeats per minute.

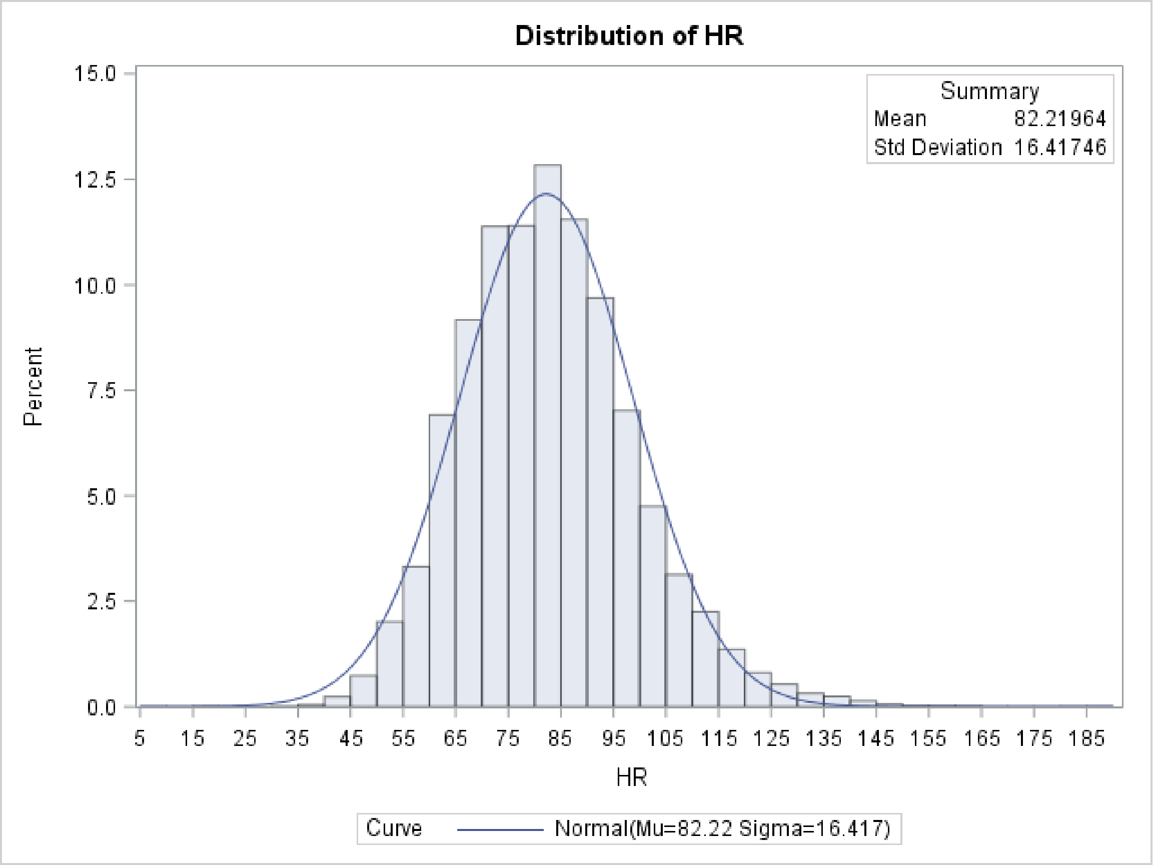


Figure 2.13 Heart Rate of Non-Sepsis Patients Histogram

For sepsis patients, the distribution of heartrates shifts upward with a mean of 99.7, a standard deviation of 20.6 and a median of 99.0. The graph below shows the histogram of heartrates for sepsis patients with an overlay of a normal distribution.

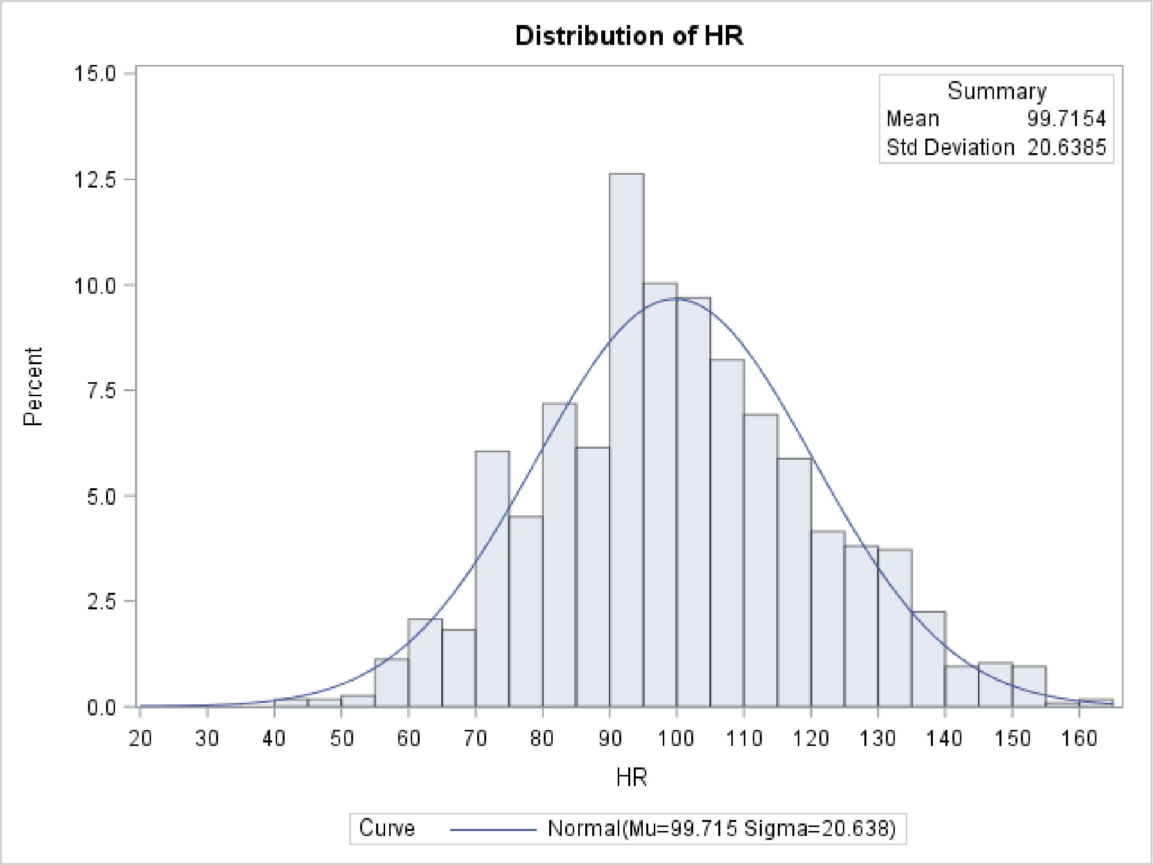


Figure 2.14 Heart Rate of Sepsis Patients Histogram

Comparing the heart rate histograms of the two groups, the mean jumps from 82.2 for non-sepsis patients to 99.7 for sepsis patients. Additionally, the median increases from 81 beats per minute to 99 beats per minute. As can be seen from Figure 2.15 below, the distribution of heartrates of sepsis patients shifts upwards nearly 20 beats per minute and does not follow the normal distribution as non-sepsis patients. The normal heart rate for an adult is usually in the range of 60 to 80 beats per minute (Cleveland Clinic 2014). For non-sepsis patients, 58.4% of the patient sample fell outside of this range, compared to 84.3% for the sepsis patients, a 44% increase of patients outside of the normal range.

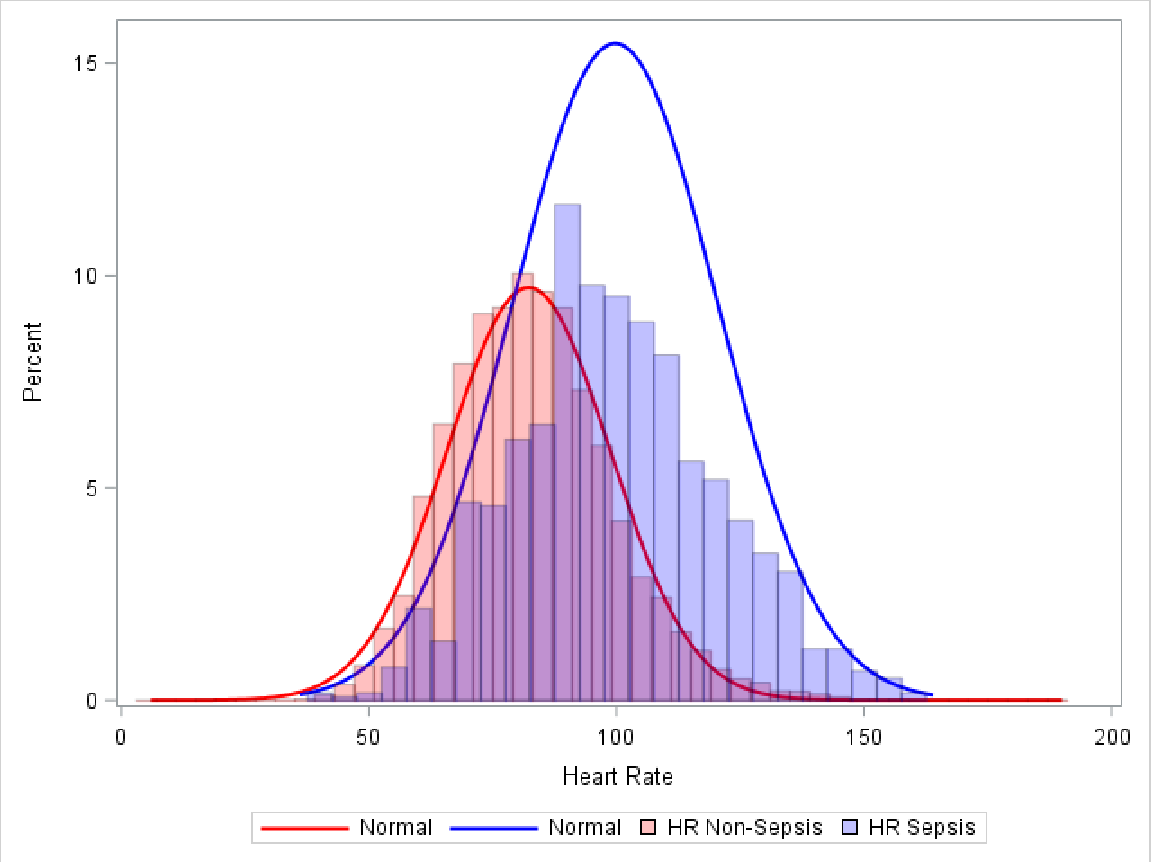


Figure 2.15 Heart Rate Composite Non-Sepsis vs. Sepsis

2.7 Peripheral Capillary Oxygen Saturation

Peripheral Capillary Oxygen Saturation, S_pO_2, measures the saturation of oxygen in the blood by measuring the percentage of oxygenated hemoglobin divided by the total amount of hemoglobin in the blood. Since it is a percentage, it is naturally right truncated at 100%. The normal range for peripheral capillary oxygen saturation should be between 95% and 100%. A S_pO_2 below 92% is an indicator of hypoxia, or low blood oxygenation. Looking at the graph below of the histogram of S_pO_2 measures for non-sepsis patients, the distribution is skewed to the left with a mean of 96.7%, a median 97% and a standard deviation of 2.8.

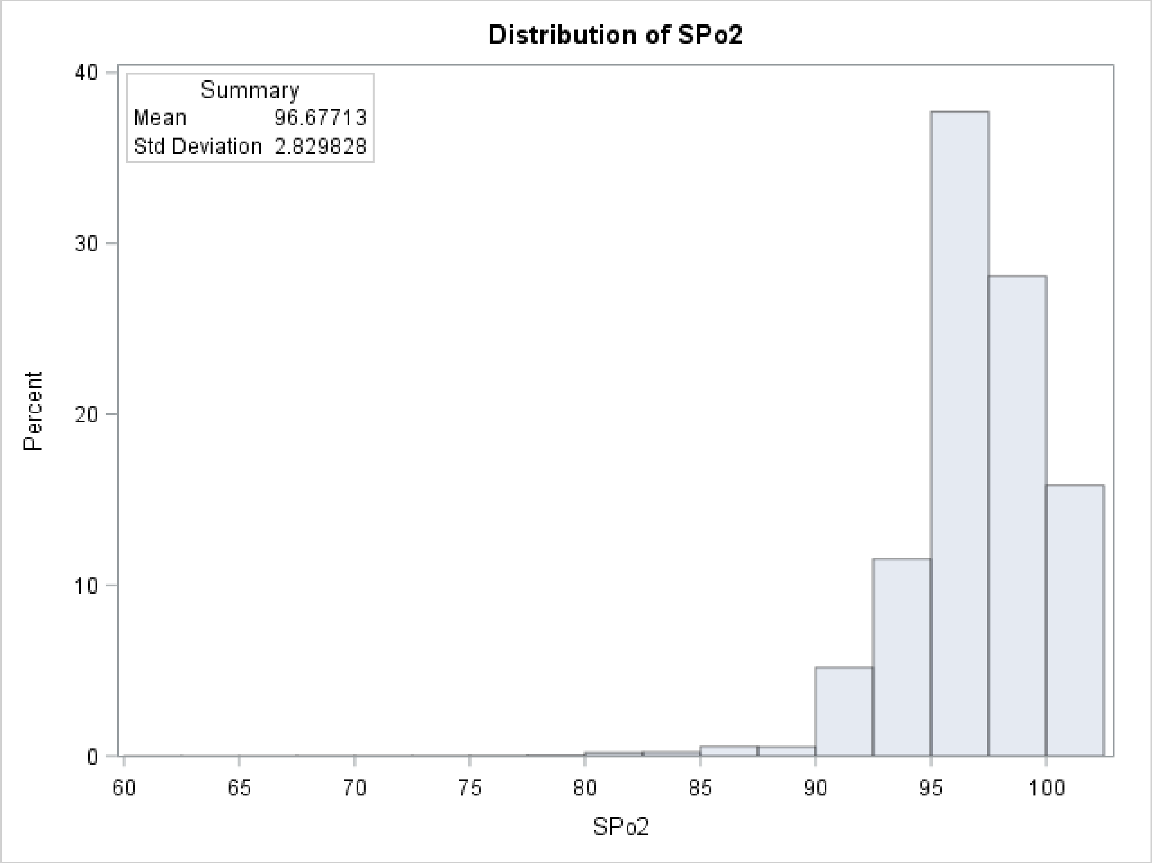


Figure 2.16 Peripheral Capillary Oxygen Saturation of Non-Sepsis Patients Histogram

Comparatively for sepsis patients, the mean of the distribution for peripheral capillary oxygen saturation is 96.9%, the median is 98% and the standard deviation is 4.14. As you can see in the graph below, the shape and skewness of the distribution is the same as that for non-sepsis patients.

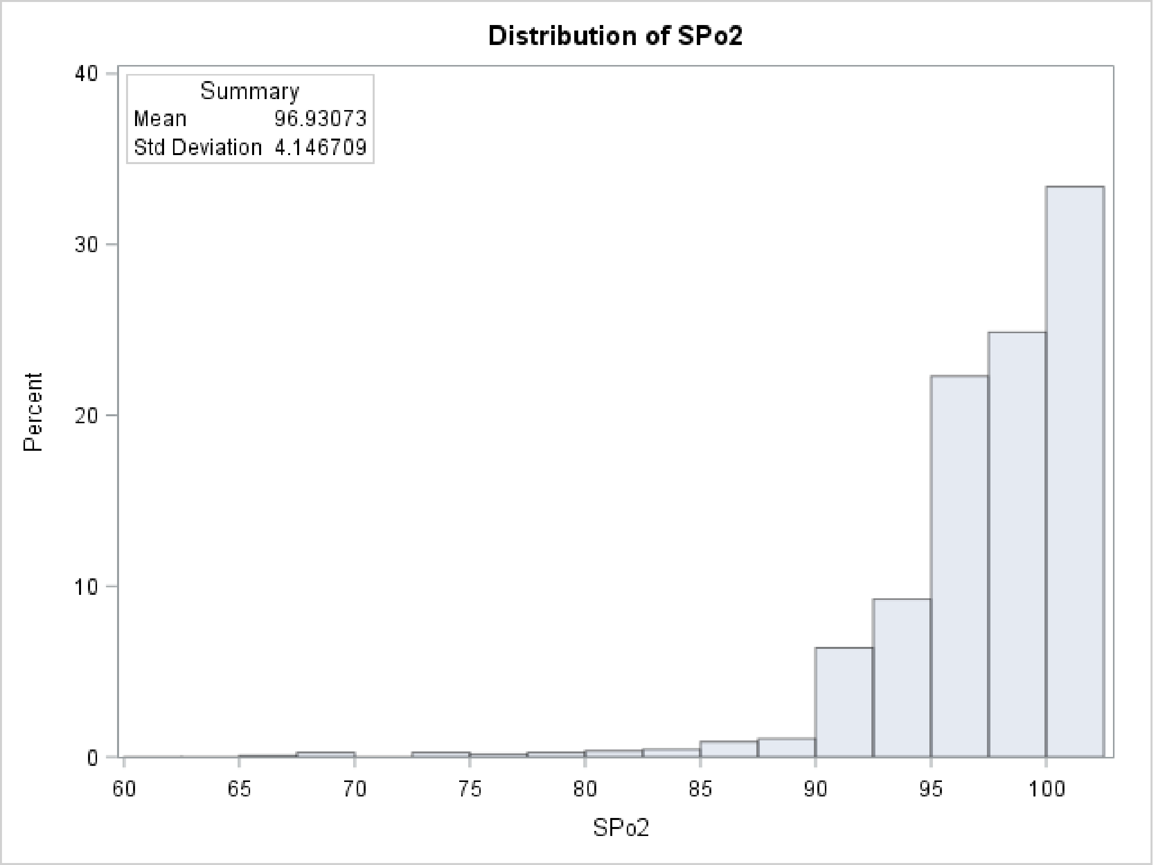


Figure 2.17 Peripheral Capillary Oxygen Saturation of Sepsis Patients Histogram

Comparing the distribution of peripheral capillary oxygen saturation between non-sepsis patients and sepsis patients, the means are similar, 96.68% versus 96.93%, respectively. The median of the distributions are also similar, 97% versus 98%, respectively. However, the spread of the distribution is smaller for non-sepsis patients, 2.83, compared to 4.15 for sepsis patients. Also, looking at the percent of patients whose S_pO_2 value is below 92%, referred to as hypoxia, only 3.8% of non-sepsis patients are classified with hypoxia, which jumps up to 7.7% of sepsis patients. This coincides with the graph below that shows that the distribution for sepsis patients has the same shape, but has a fatter tail.

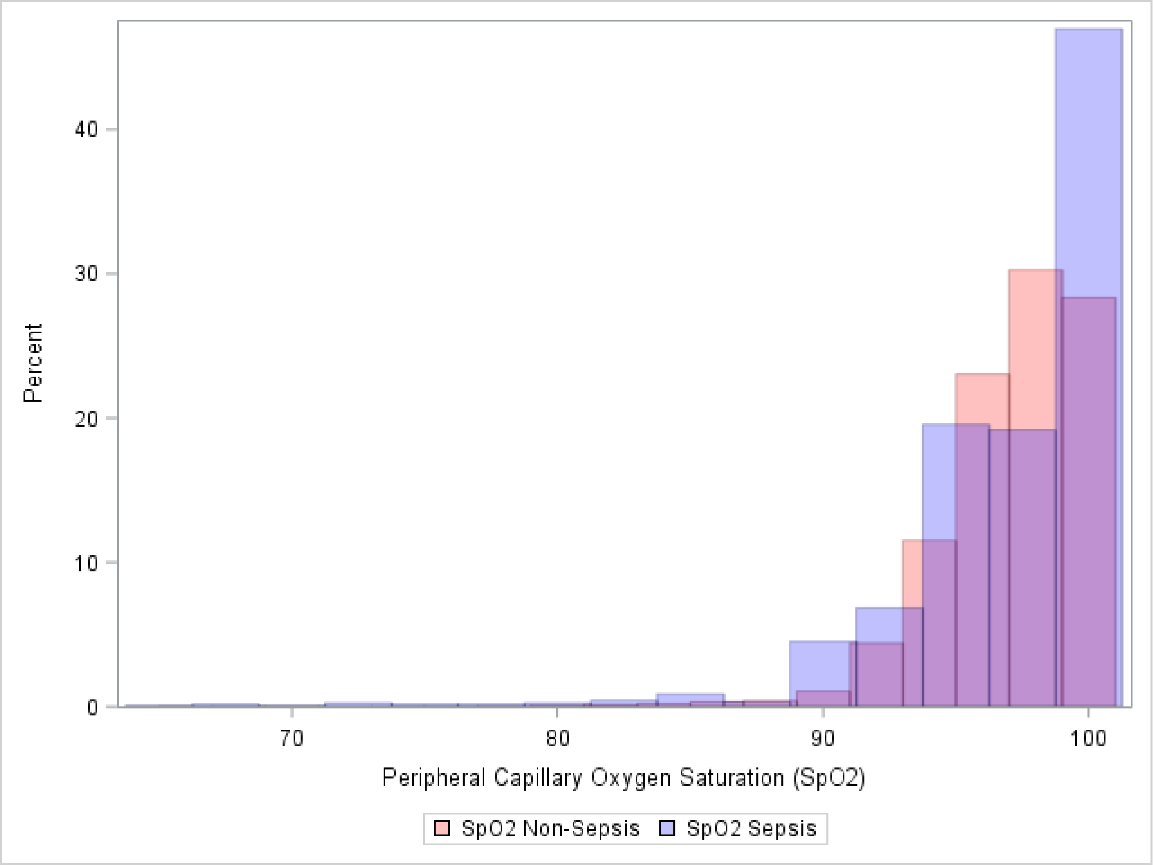


Figure 2.18 Peripheral Capillary Oxygen Saturation Composite Non-Sepsis vs. Sepsis

2.8 Body Temperature

The temperature vital sign is measured in degrees Celsius. The graph below, Figure 2.19, shows the distribution of temperatures for non-sepsis patients. The mean of the distribution is 36.7 degrees, the median is 36.7 degrees and the standard deviation is 0.47 degrees. The overlaid line represents a normal distribution. The line follows the distribution well, with the exception of the higher peak located at the mean.

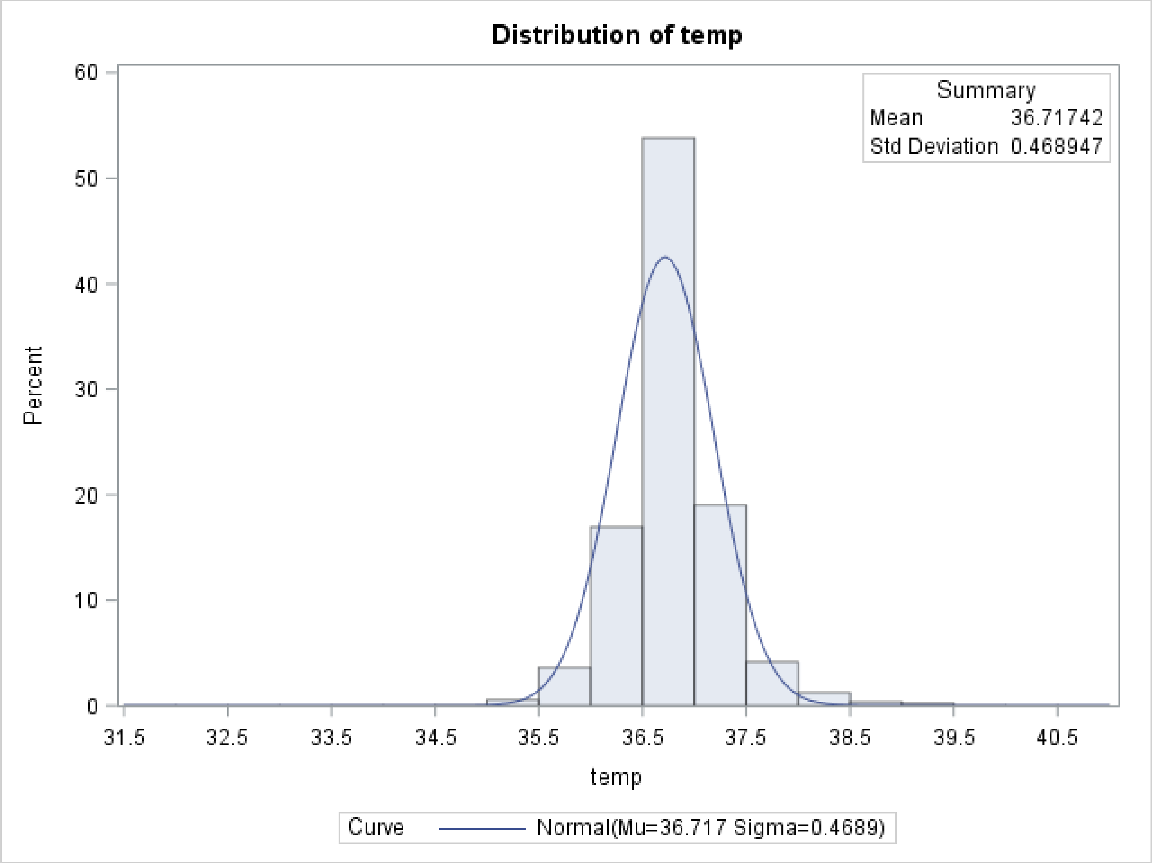


Figure 2.19 Temperature in Degrees Celsius for Non-Sepsis Patients Histogram

Figure 2.20 below shows the histogram of temperatures measured in degrees Celsius for sepsis patients. The mean of the distribution is 37.7 degrees, the median is 37.8 degrees and the standard deviation is 1.3 degrees. It is worth noting that the distribution is bi-modal with what appears to be a mixture of normal distributions. One distribution is centered around 37 degrees Celsius and the other is centered around 39 degrees Celsius, the second of which appears to have a smaller variance.

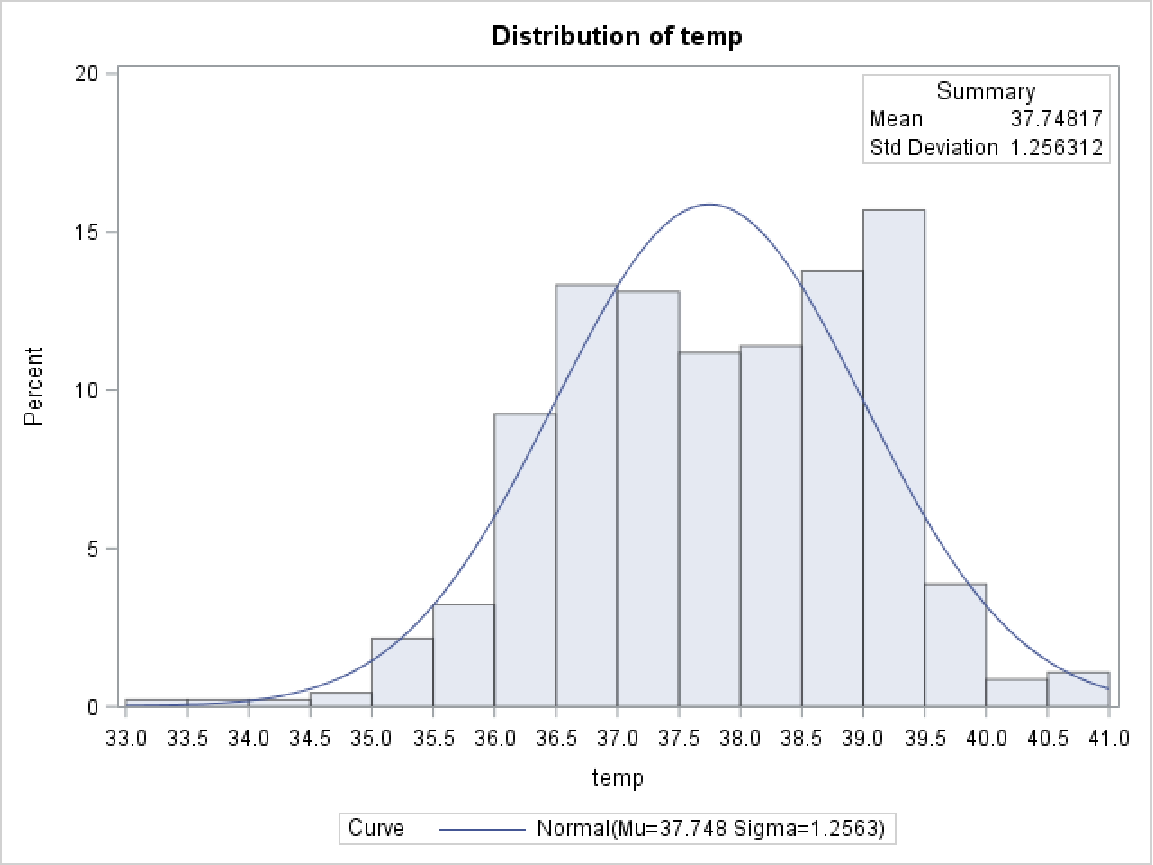


Figure 2.20 Temperature in Degrees Celsius for Sepsis Patients Histogram

Comparing the distributions of temperatures between sepsis and non-sepsis patients, the mean and the median both shift up a degree, while the standard deviation increases from 0.47 up to 1.26, over a 168% increase. This large increase in the standard deviation is partially due to the change in the shape of the distribution. For non-sepsis patients, the distribution is bell-shaped and symmetrical, a typical normally shaped distribution. However, for the sepsis patients, the distribution is bi-modal and appears to be a mixture of two very separate distributions. This is likely due to clinical staff artificially reducing a patient's temperature when they spike a fever since a prolonged exposure to high fever can cause additional complications and long-term effects to the patient. Because the sample of sepsis vital signs are taken both just before and after the patient is diagnosed with sepsis, this bi-modal distribution is due to the recording of some temperature readings after the patient’s temperature has already been artificially reduced. The temperature of a normal patient should be around 37 degrees Celsius, and a patient is considered to have a fever when the temperature rises above 38 degrees Celsius. For the non-sepsis patients, only 1.39% of the patients had a temperature above 38 degrees Celsius. When looking at the distribution of temperatures for sepsis patients, 43.87% of the observations fall above the 38 degree Celsius limit even with the discrepancy noted above. One of the human body’s natural responses to infection is to raise the temperature of the body to aide in the fight against the infection, thus this is not a surprising find. However the bimodal shape of the sepsis patients’ temperatures is somewhat surprising. This mixture of distributions is explained by how healthcare professionals respond to patients with high fevers. Figure 2.21 below shows the composite graph of distributions of temperatures for non-sepsis patients and sepsis patients.

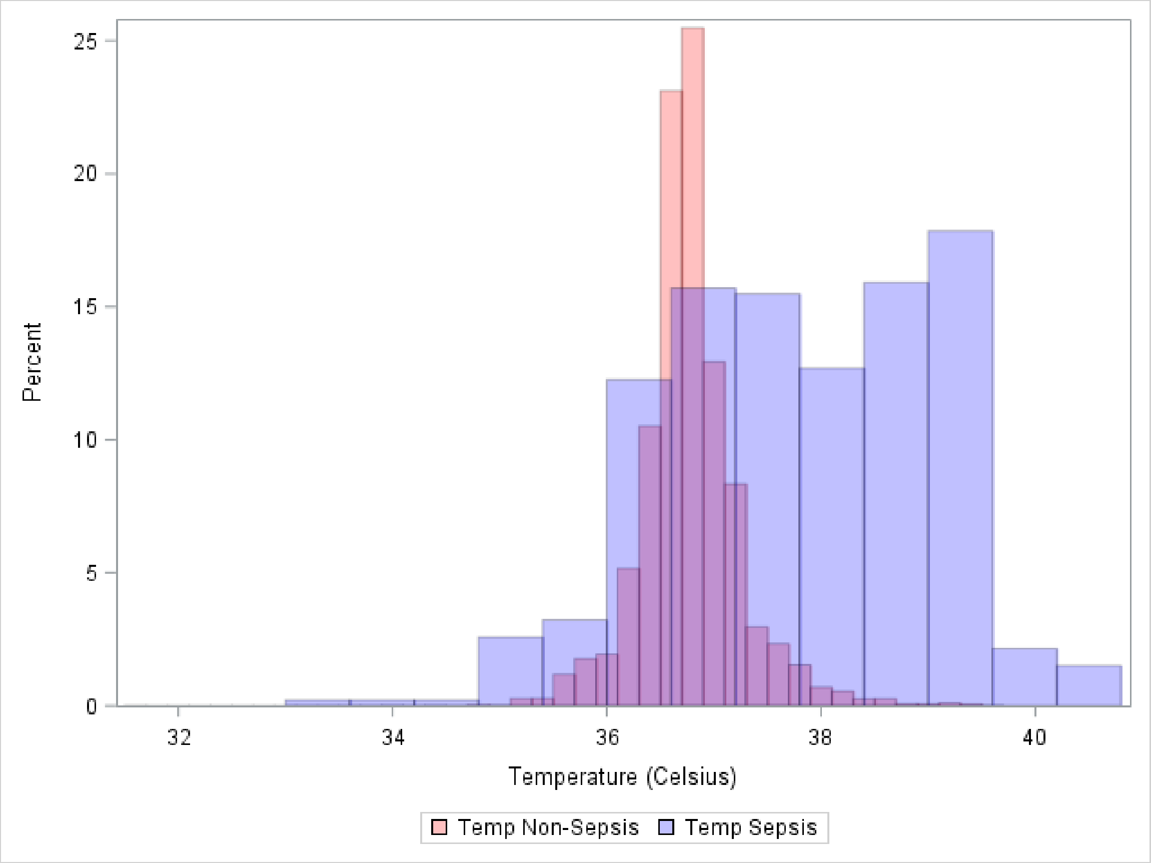


Figure 2.21 Temperature in Celsius Composite Non-Sepsis vs. Sepsis

2.9 Vital Signs Comparison Summary

When looking across the comparisons of distributions of the different vital signs for non-sepsis patients versus sepsis patients there are stark differences between the distributions. However, there is still a significant amount of overlap between the two patient populations. Some comparisons show a complete change in the shape of the distributions. For example, the systolic blood pressure (Figure 2.9) changes from a normal distribution for non-sepsis patients into a log-normal distribution for sepsis patients. This change is accompanied by a systematic shift downward in blood pressure, but still the majority of the two distributions overlap. This highlights why a simple threshold classification is not sufficient in distinguishing between these two populations. The addition of a set of variables looking at the change of each of the vital signs will add to the available information. Looking at the distributions of heart rates (Figure 2.15), the shapes are the same, but there is a dramatic shift upward of the distribution. Thus, a patient with a baseline heart rate on the lower side of the distribution would likely see a jump in their heart rate, but very likely not outside of the range of heart rates for standard patients. A methodology to capture the systematic change across the vital signs of a patient when they are entering into a septic state will be developed to aide in the identification of patients from each population.

CHAPTER 3. SIMULATING THE VITAL SIGNS

3.1 Introduction

In order to more fully understand the distributions and variation changes between the vital signs of sepsis and non-sepsis patients, each of the vital signs will be simulated for both groups. This will also give a better understanding of the distributions that do not follow a standard distribution as noted above for both temperature and peripheral capillary oxygen saturation.

3.2 Systolic Blood Pressure for Non-Sepsis Patients

Begin by looking at the distribution of a non-sepsis patient's systolic blood pressure. Refer back to Figure 2.7, the histogram of systolic blood pressure for non-sepsis patients. The histogram is symmetric and bell-shaped leading to an initial suggestion of a normal distribution. To check this, an overlay of the normal distribution was added to the graph with a mean of 126.75 and a standard deviation of 22.12.

Since the overlay line appears to fit the distribution well, the data was simulated using a normal distribution with a mean of 126.7 and a standard deviation of 22.1. Figure 3.1 shows the histogram of the simulated data with an overlay of the normal distribution.

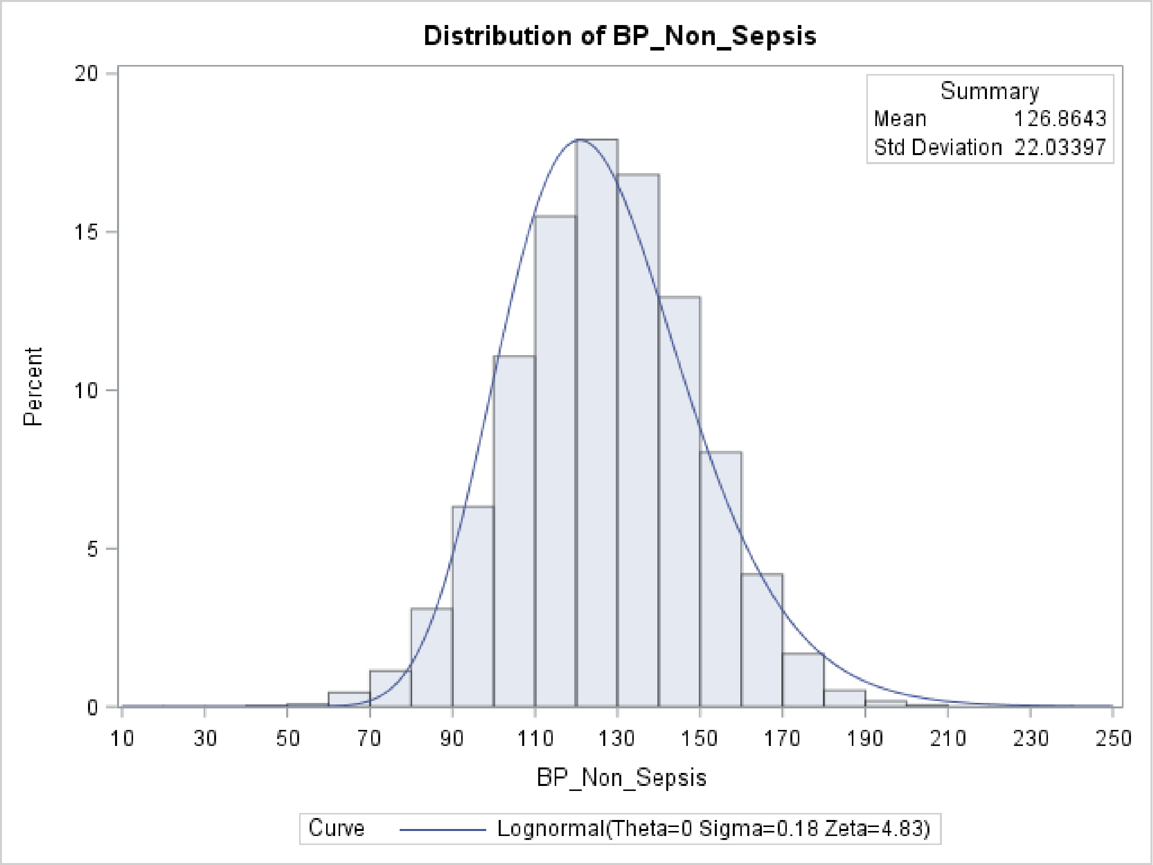


Figure 3.1 Simulated Systolic Blood Pressure for Non-Sepsis Patients Histogram

The graph below shows the two histograms superimposed onto each other. As you can see, the simulated data matches the patient data very well. The original patient data is plotted in red and the simulated data is plotted in blue.

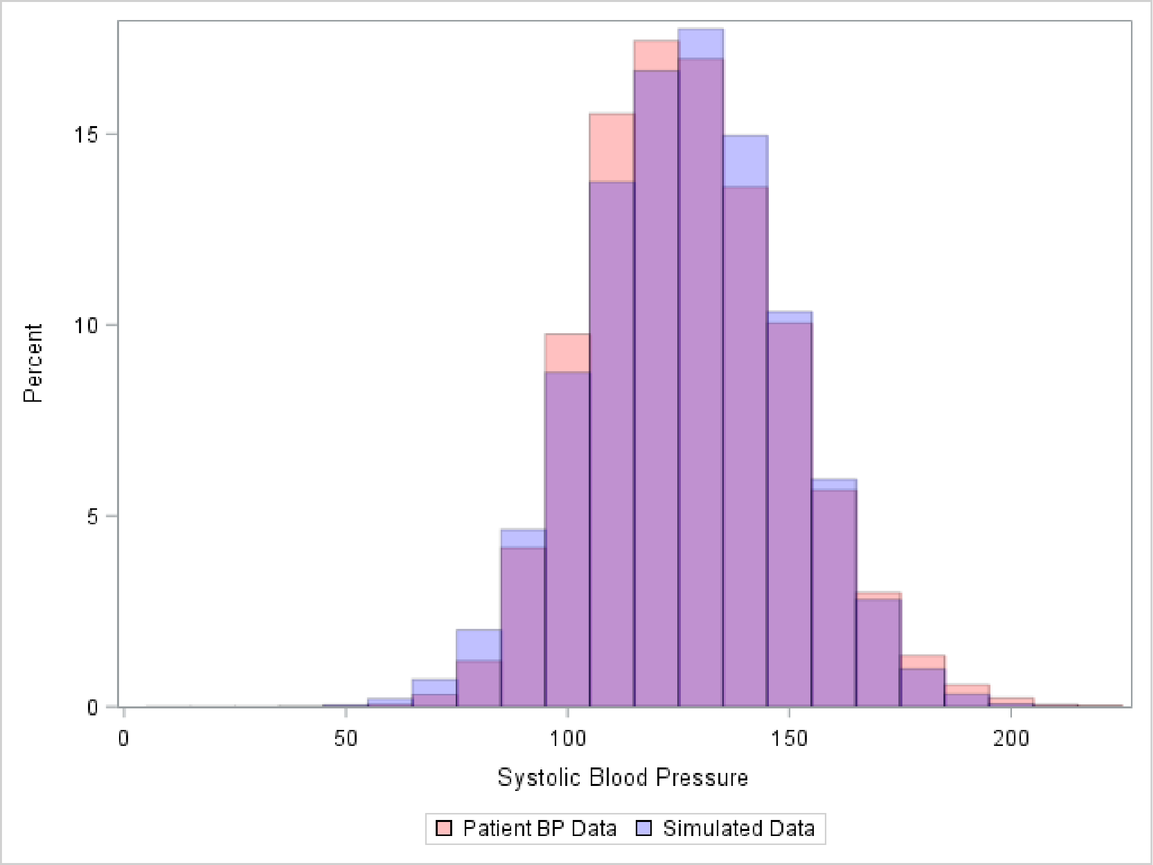


Figure 3.2 Systolic Blood Pressure for Non-Sepsis Patients Composite Simulated vs. Patient Data

3.3 Systolic Blood Pressure for Sepsis Patients

Now looking at simulating the systolic blood pressure for sepsis patients, recall Figure 2.8 showing a histogram of measures for active sepsis patients. The shape of this histogram is no longer symmetric and instead is skewed to the right and resembles a log normal distribution. To check this assertion, a log normal curve was overlaid onto the histogram showing it follows the shape of the distribution.

To simulate the systolic blood pressure of sepsis patients, a log normal distribution is utilized with a mean, or zeta, of 4.73 and a standard deviation, or sigma, of 0.24. Figure 3.3, illustrates the simulated data with an overlay curve of the log normal distribution. The simulated data seems to follow the curve well and also resembles the patient data found in Figure 2.8.

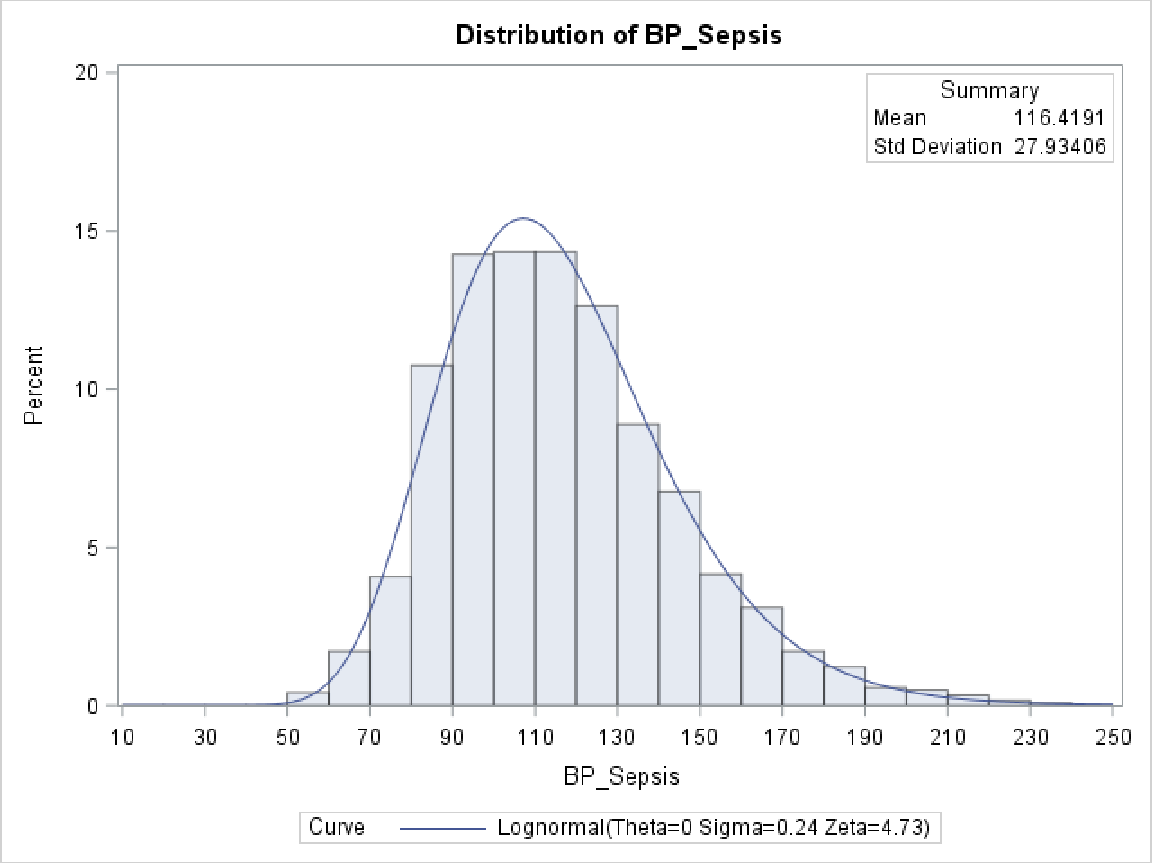


Figure 3.3 Simulated Systolic Blood Pressure for Sepsis Patients Histogram

To better visualize the comparison between the patient data and the simulated data, the graph below shows both histograms superimposed on each other. The patient data is graphed in red and the simulated data is in blue. The shapes of the histograms and the centrality points appear to match well.

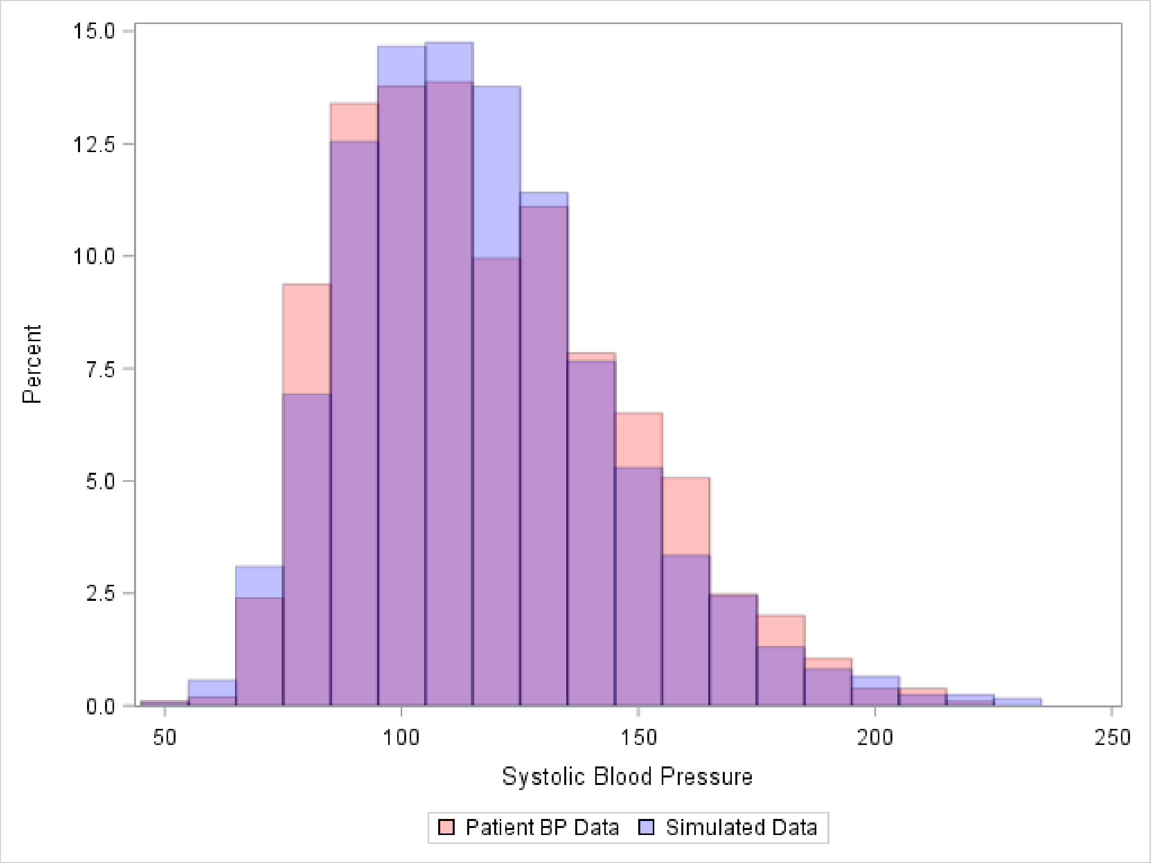


Figure 3.4 Systolic Blood Pressure for Sepsis Patients Composite Simulated vs. Patient Data

3.4 Respiratory Rate for Non-Sepsis Patients

Next, looking at the respiratory rate, recall Figure 2.10 of the histogram of respiratory rates for non-sepsis patients. The graph has an overlay curve of the gamma distribution, however, it is not a perfect fit. Using this gamma distribution with a shape parameter 68.4, a scale parameter of 0.45 and a threshold of -13.1 gave a starting point for the distribution.

To simulate this distribution, a base of a gamma distribution was used with shape parameter, alpha, of 68.4; a scale parameter, sigma, equal to 0.45; and a threshold parameter, theta, of -13.1. Then to capture the higher peak, a narrow normal distribution with a mean of 17.5 and a standard deviation of one was added. The two distributions were combined using a binomial random variable with a probability of 0.7 leading to a 70% mix of the gamma distribution and a 30% mix of the normal distribution. The graph below shows a histogram of the simulated distribution.

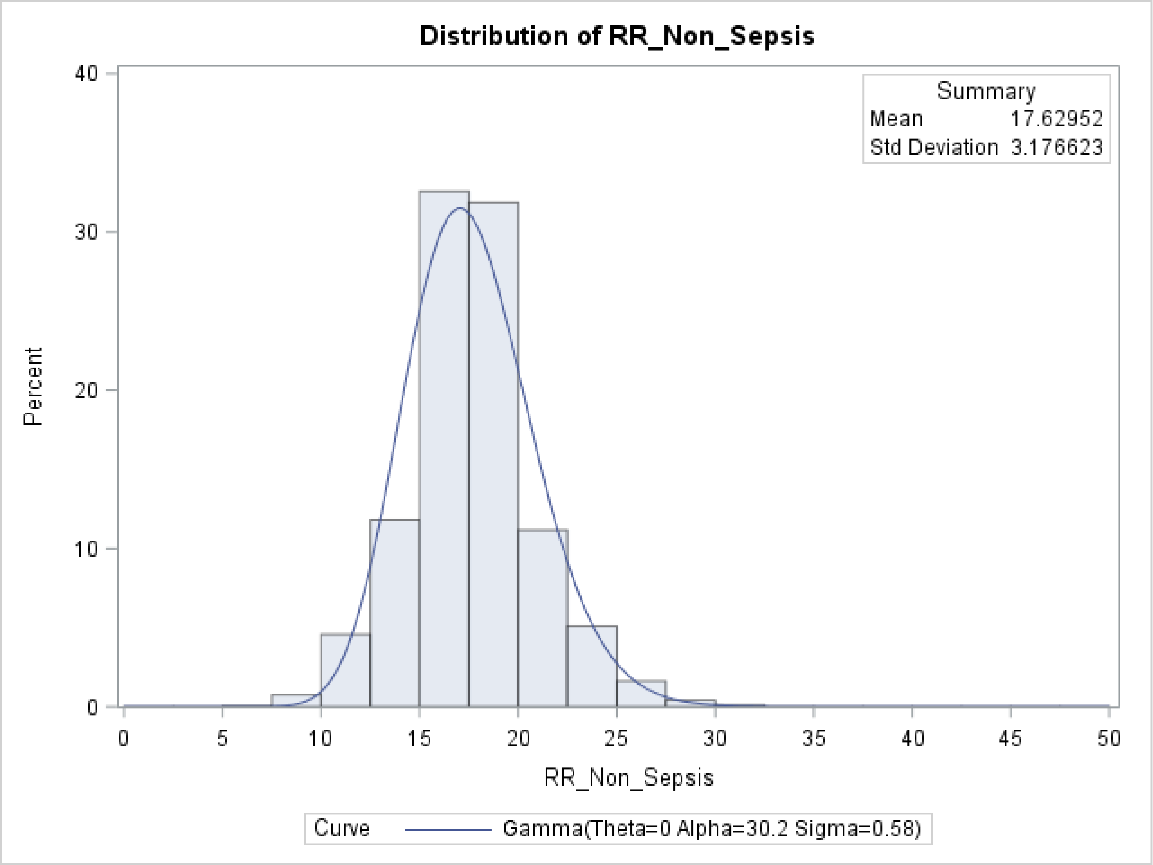


Figure 3.5 Simulated Respiratory Rate for Non-Sepsis Patients Histogram

To visualize the comparison between the patient data and the simulated data, the graph below shows both histograms superimposed on each other. The patient respiratory rate data is graphed in red and the simulated respiratory rate data is in blue. The shapes of the histograms and the centrality points appear to match well with the exception that the peaks are flipped at around 15 breaths per minute and 17 breaths per minute.

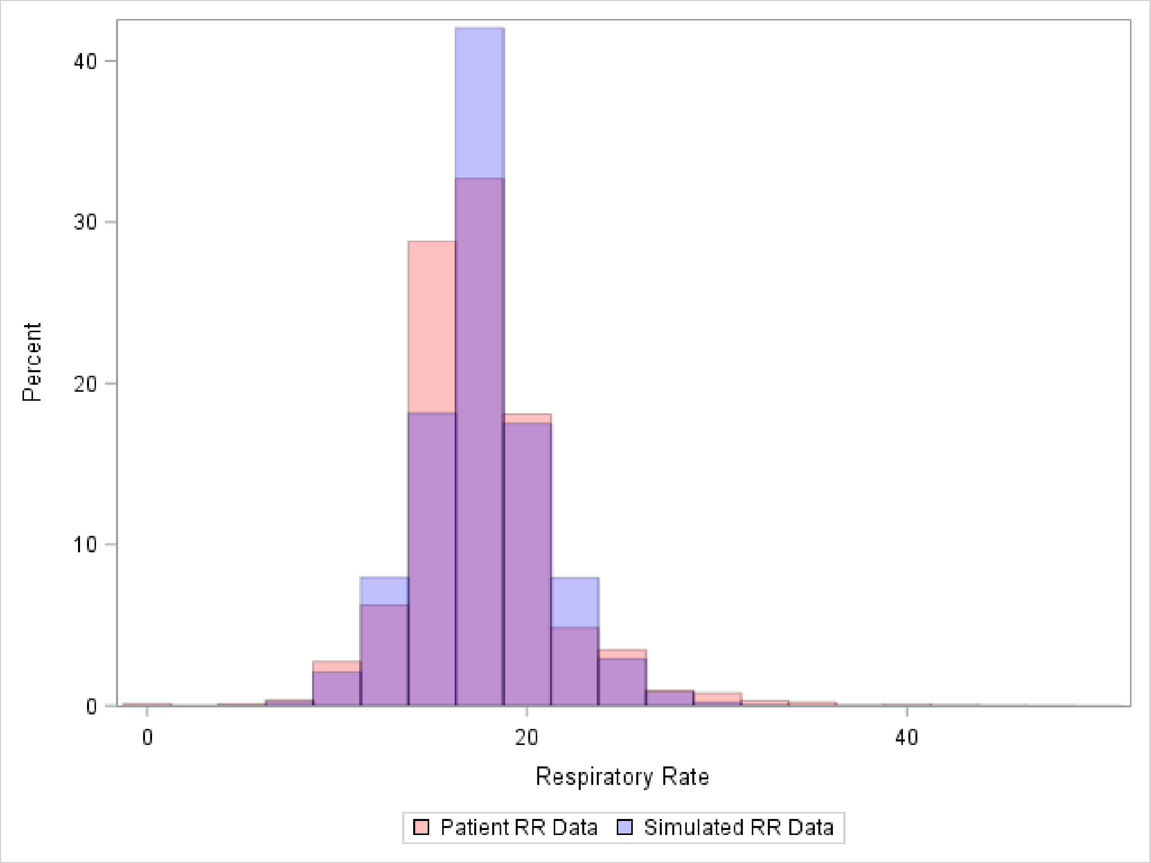


Figure 3.6 Respiratory Rate for Non-Sepsis Patients Composite Simulated vs. Patient Data

3.5 Respiratory Rate for Sepsis Patients

Focusing on the respiratory rate for sepsis patients, recall Figure 2.11 showing a histogram of patient respiratory rates who had an active sepsis diagnosis. The graph shows an overlay of the gamma distribution with a shape parameter of 16.7, a scale parameter of 1.61 and a threshold parameter of -7. The curve fits the data marginally well, with a higher peak around 17.5 breaths per minute.

To simulate this distribution, a gamma distribution with a shape parameter of 16.7, a scale parameter of 1.61 and a threshold parameter of -7 was used as the base distribution. Then, a small portion of a normal distribution with a mean of 16 and a standard deviation of 0.5 is mixed with the gamma distribution to account for the uncharacteristic peak. The final distribution has an 80% mix of the gamma distribution and a 20% mix of the normal distribution.

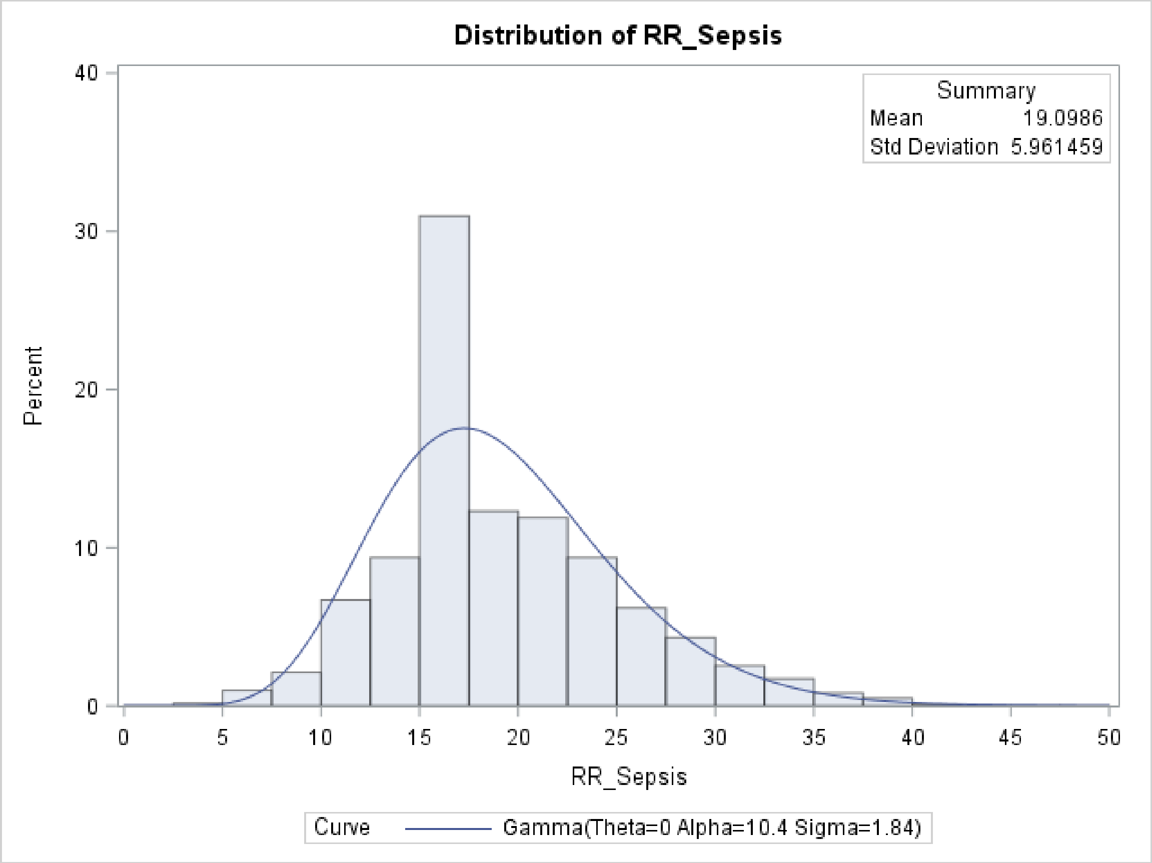


Figure 3.7 Simulated Respiratory Rate for Sepsis Patients Histogram

To compare the histograms between the patient data and the simulated data, the graph below shows both histograms graphed together utilizing a transparency feature. The patient respiratory rate data is graphed in red and the simulated respiratory rate data is in blue. The shapes of the histograms and the centrality points appear to match well with the some exceptions in the tails.

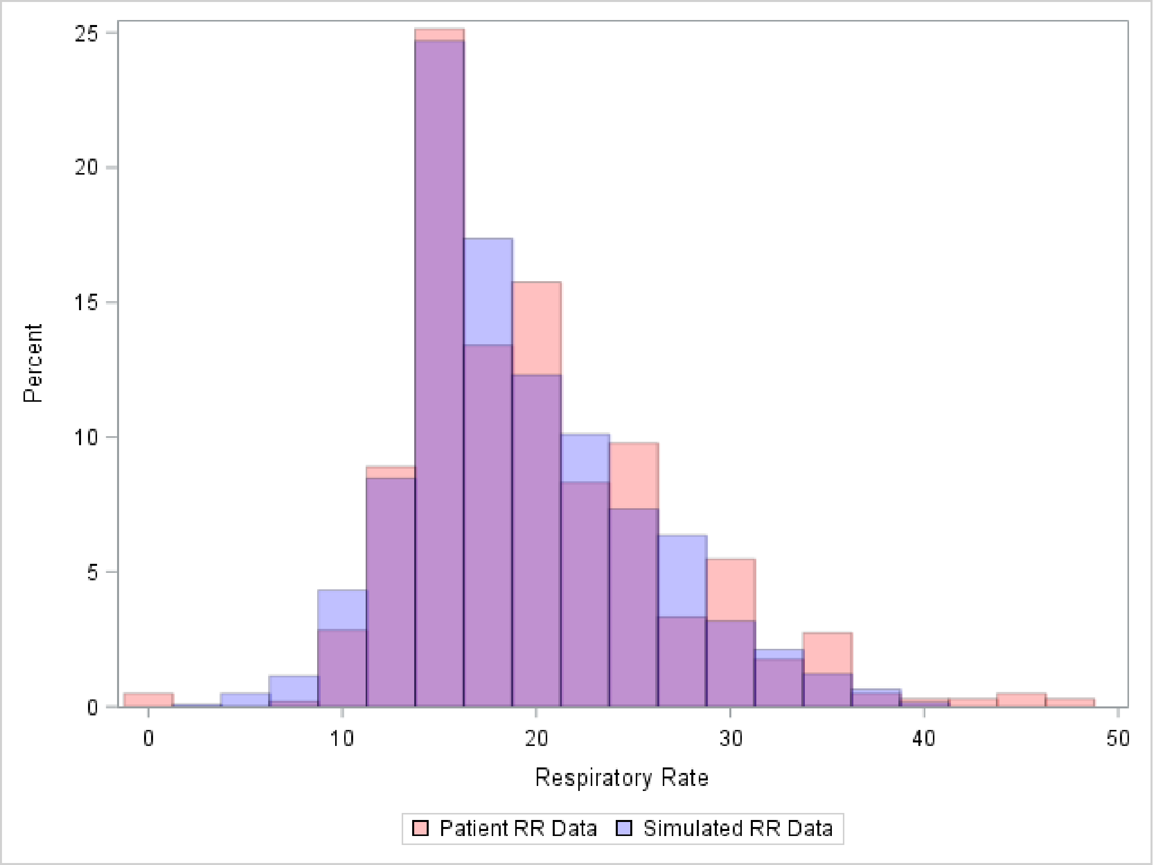


Figure 3.8 Respiratory Rate for Sepsis Patients Composite Simulated vs. Patient Data

3.6 Heart Rate for Non-Sepsis Patients

Consider Figure 2.13 showing the distribution of heart rates for non-sepsis patients. The histogram is shown with an overlay of a normal distribution which has a mean of 82.2 and a standard deviation of 16.4. The normal curve is a good approximation to the heart rate histogram for non-sepsis patients. To simulate the heart rate for this subset of patients, a normal distribution is used with a mean of 82.2 and a standard deviation of 16.4. The graph below, Figure 3.9, illustrates the simulated data with an overlay curve of the normal distribution.

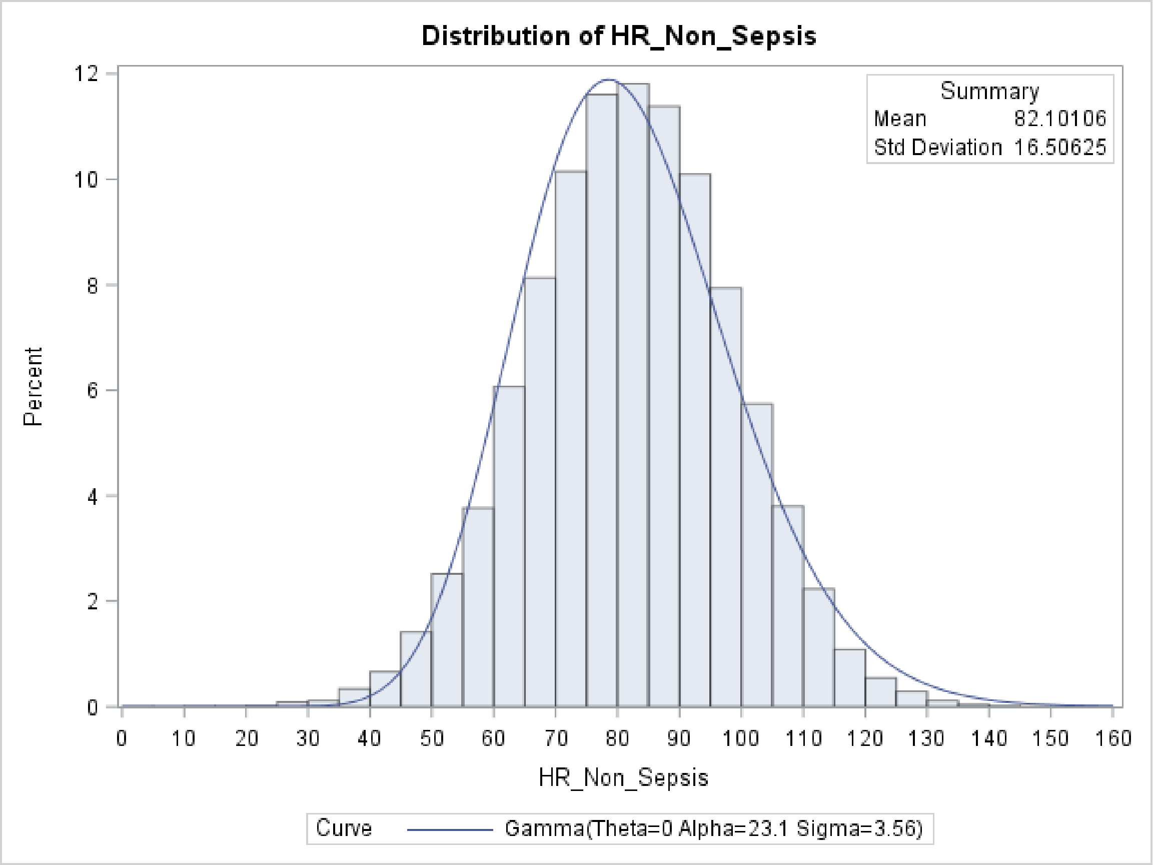


Figure 3.9 Simulated Heart Rate for Non-Sepsis Patients Histogram

The figure below shows the two histograms superimposed onto each other. The simulated heart rate data matches the non-sepsis patient data very well. The original patient data is plotted in red and the simulated data is plotted in blue.

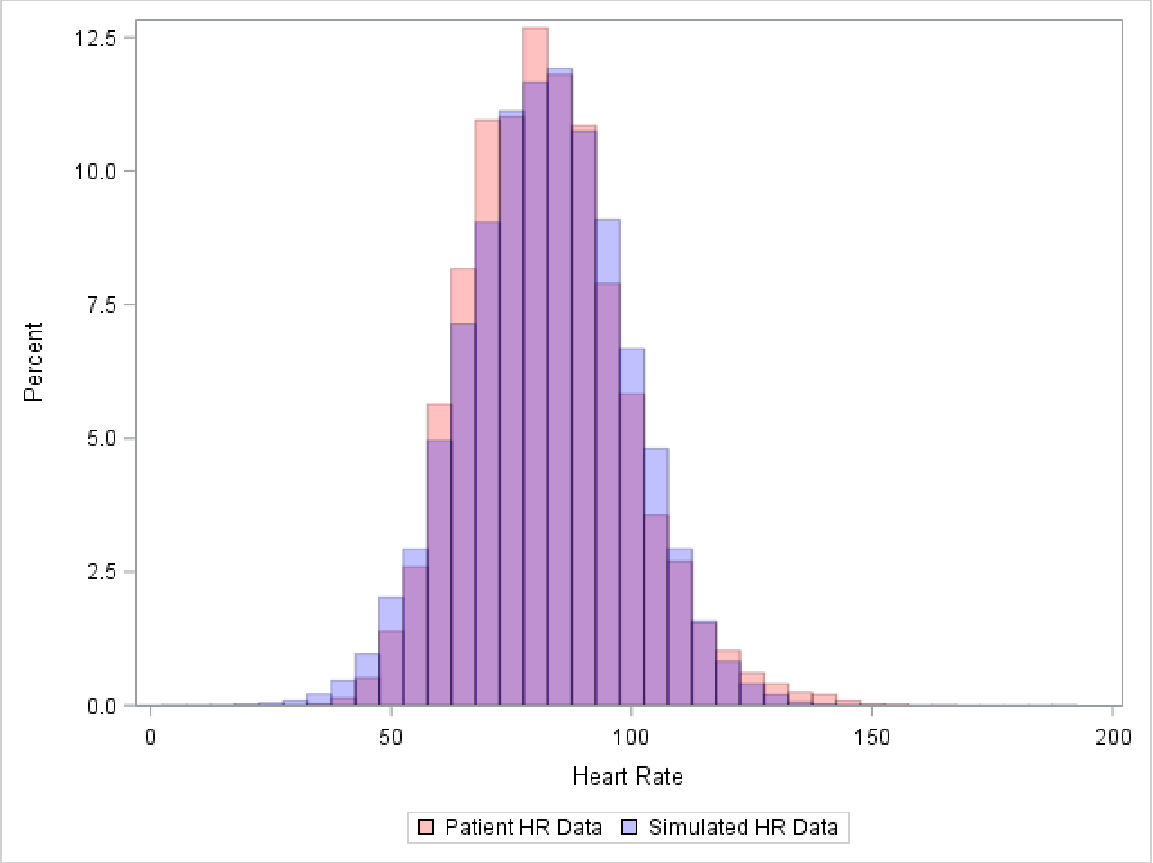


Figure 3.10 Heart Rate for Non-Sepsis Patients Composite Simulated vs. Patient Data

3.7 Heart Rate for Sepsis Patients

Looking at the distribution of heart rates for patients with an active sepsis diagnosis, recall Figure 2.14 showing the distribution. The graph has an overlay of the normal distribution, which fits the data marginally well, with the exception of two peaks around 75 and 95.

To examine if this distribution fits is normal distribution or if it needs to be mixed with an additional distribution, an Anderson-Darling test of normality is run using the Univariate procedure in SAS. The test gives a p-value greater than 0.25 indicating that the distribution of heart rates for sepsis patients is normally distributed.

To simulate the data a normal distribution with a mean of 99.7 and a standard deviation of 20.6 will be utilized. The graph below shows a histogram of the simulated heart rate data for the sepsis population with an overlay of the normal distribution.

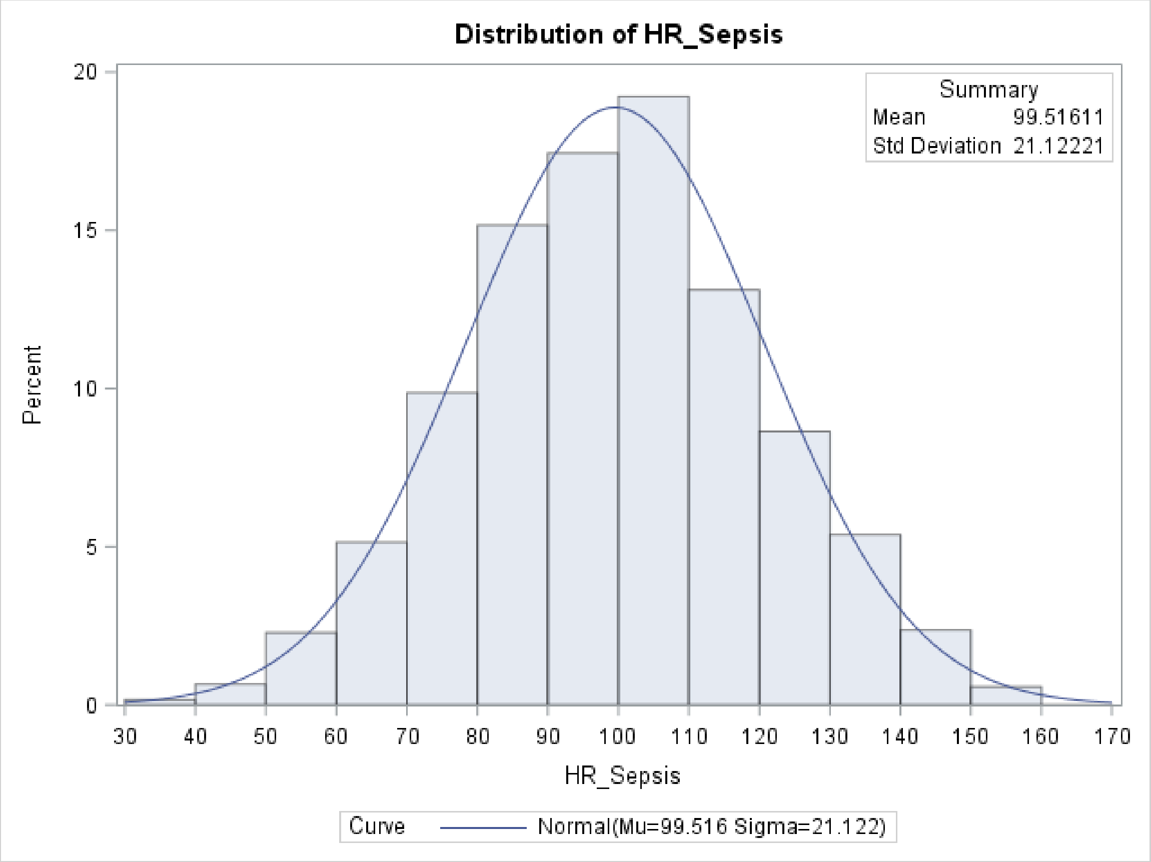


Figure 3.11 Simulated Heart Rate for Sepsis Patients Histogram

In order to better visualize the fit between the patient data and the simulated data, the graph below shows the two histograms graphed together with the patient data in red and the simulated data in blue. The simulated data is a good fit with the patient data.

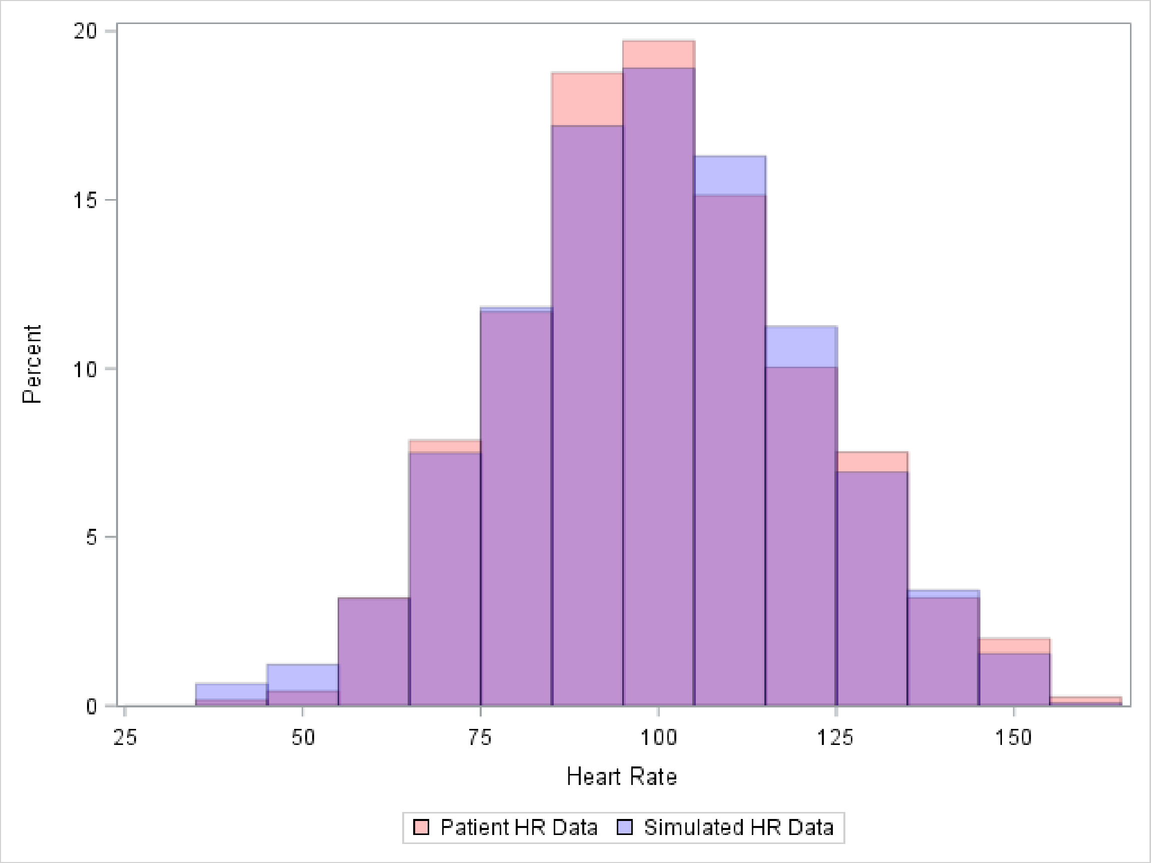


Figure 3.12 Heart Rate for Sepsis Patients Composite Simulated vs. Patient Data

3.8 Peripheral Capillary Oxygen Saturation

Next focusing on the peripheral capillary oxygen saturation (S_pO_2) for non-sepsis patients, recall Figure 2.16 which visualizes the distribution of patient data. The data is truncated on the right at 100% and the vast majority of patient measures stay above 90% saturation with a shape skewed to the left. To simulate S_pO_2 for non-sepsis patients, an exponential distribution is used as a base and mixed with a Poisson distribution to achieve the desired shape and tail of the distribution. To simulate the right truncation, this mixture is subtracted from 100. The final mixture uses 70% of a Poisson distribution with a mean parameter of four and 30% of a standard exponential distribution. Figure 3.13 shows the histogram of the simulated distribution.

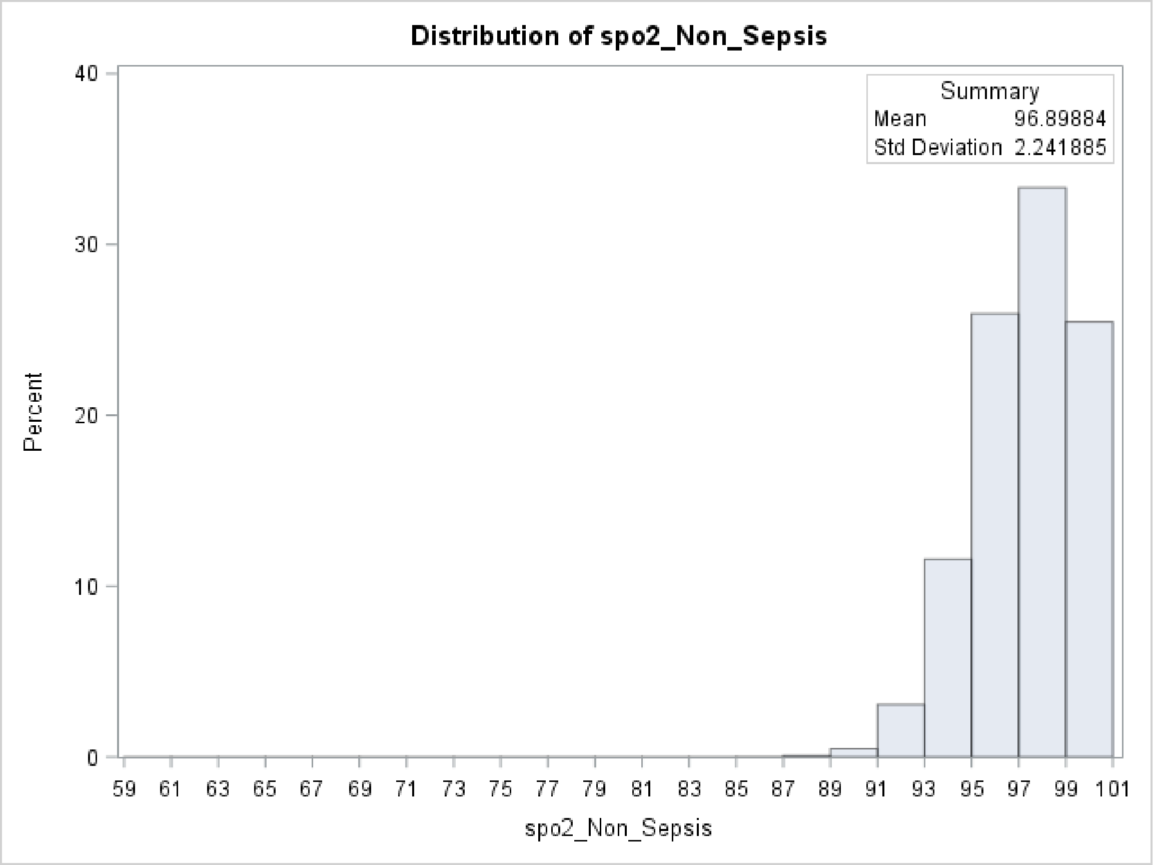


Figure 3.13 Simulated S_pO_2 for Non-Sepsis Patients Histogram

For comparison, of the distributions on the same size and scale, the two histograms are graphed together below. The actual patient data is graphed in red and the simulated data is graphed blue. The simulated data follows the patient data for peripheral capillary oxygen saturation well.

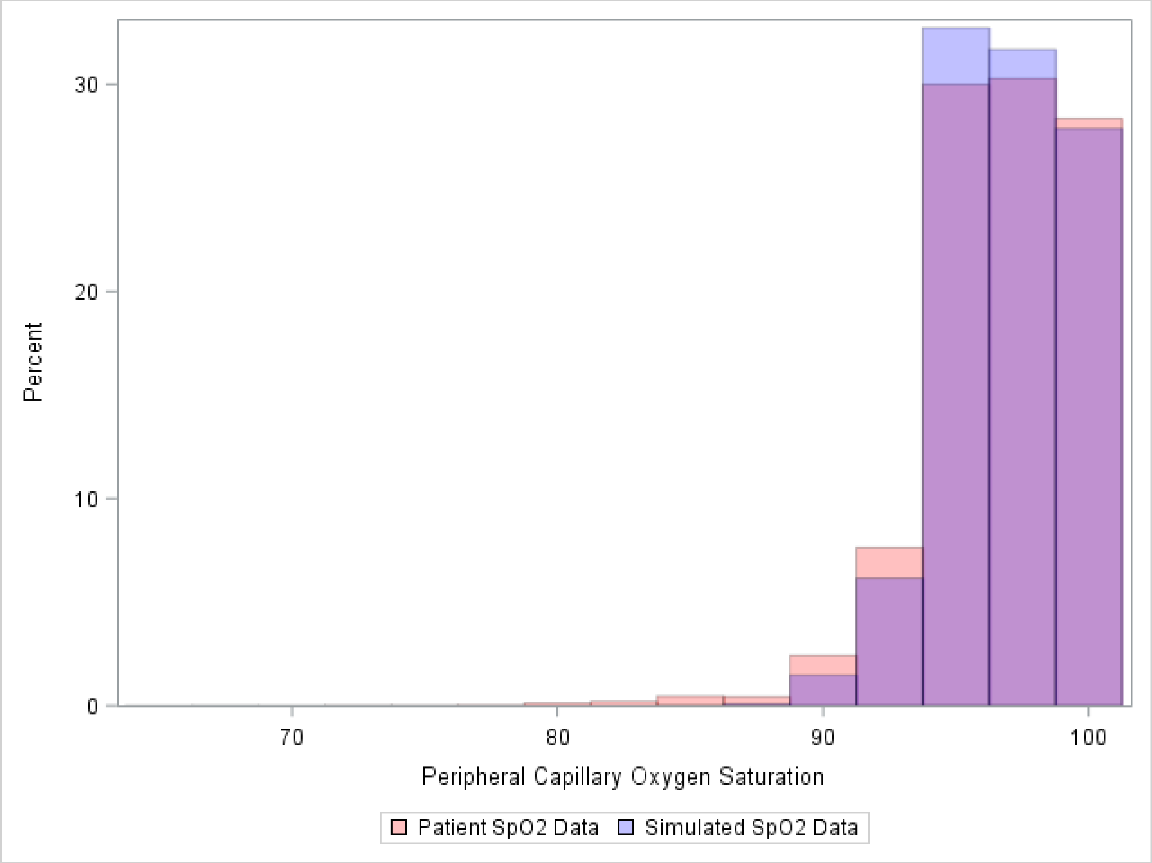


Figure 3.14 S_pO_2 for Non-Sepsis Patients Composite Simulated vs. Patient Data

3.9 Peripheral Capillary Oxygen Saturation for Sepsis Patients

Turning our attention to patients with an active sepsis diagnosis, recall Figure 2.17 showing the peripheral capillary oxygen saturation. The main difference between non-sepsis patients and sepsis patients is the higher spread of the data in the sepsis population. This spread is manifested in the longer and fatter tail. To simulate this distribution, the same base exponential distribution is utilized, but the spread of the Poisson distribution is increased and the mixture percentages are adjusted to better simulate this patient group. For the active sepsis population, the mixture is 70% of a standard exponential distribution with 30% of a Poisson distribution with a mean parameter of 7 to increase the spread in the tail.

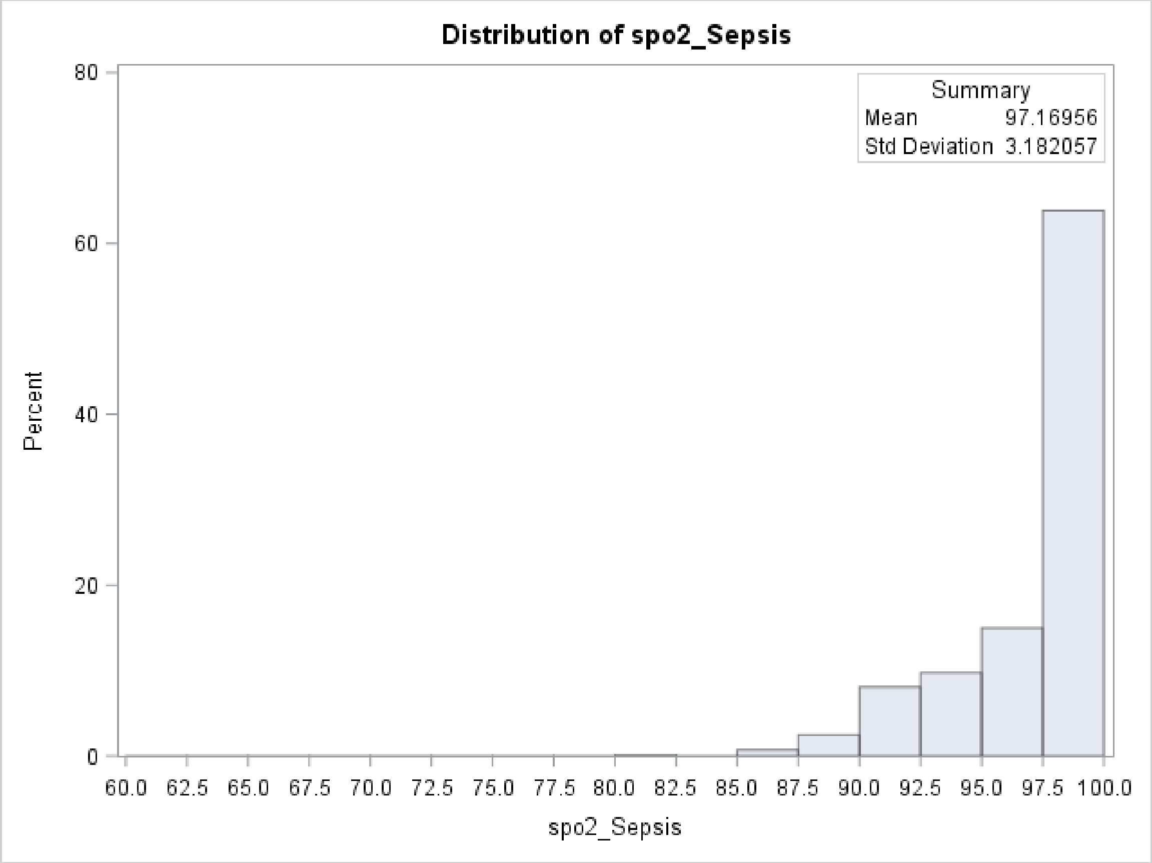


Figure 3.15 Simulated S_pO_2 for Sepsis Patients Histogram

In order to compare the simulated data with the patient data, the histogram of the simulated data is superimposed on the sepsis patient data for peripheral capillary oxygen saturation in the graph below. The patient data is shown in red and the simulated data is displayed in blue. The simulated data fits the patient data well with a slightly lower peak in the mid 90% bin.

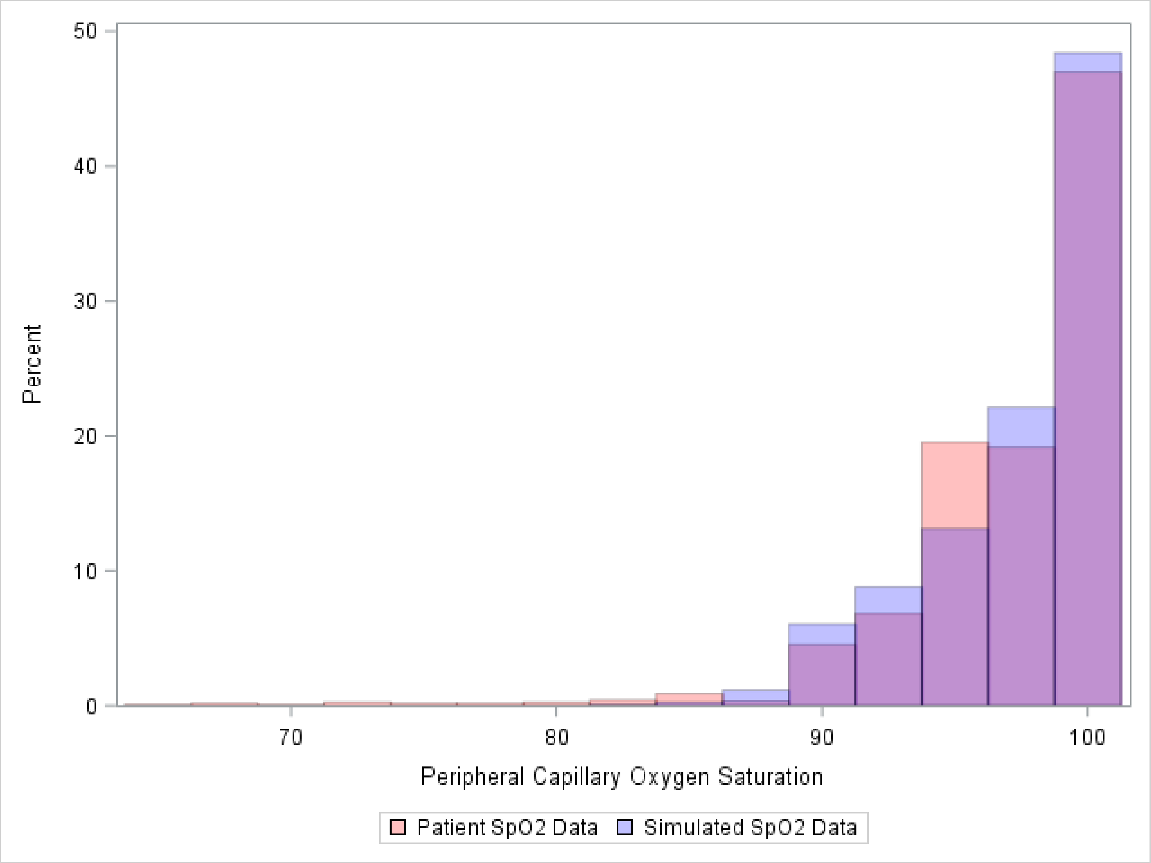


Figure 3.16 S_pO_2 for Sepsis Patients Composite Simulated vs. Patient Data

3.10 Body Temperature for Non-Sepsis Patients

Refer to the distribution of temperature across non-sepsis patients in Figure 2.19. The histogram has an overlay of a curve with a normal distribution with a mean of 36.7 degrees and a standard deviation of 0.47 degrees. The curve fits the distribution well except for the higher peak at the mode of the distribution. To simulate the temperature in degrees Celsius for non-sepsis patients, a foundation of a normal distribution was utilized with a mean of 36.7 degrees and a standard deviation of 0.47 degrees. This distribution is mixed with a second normal distribution which has a mean of 36.7 degrees and a standard deviation of 0.1 degrees to accommodate the higher peak. The mixture of normals is combined at a ratio of four to one.

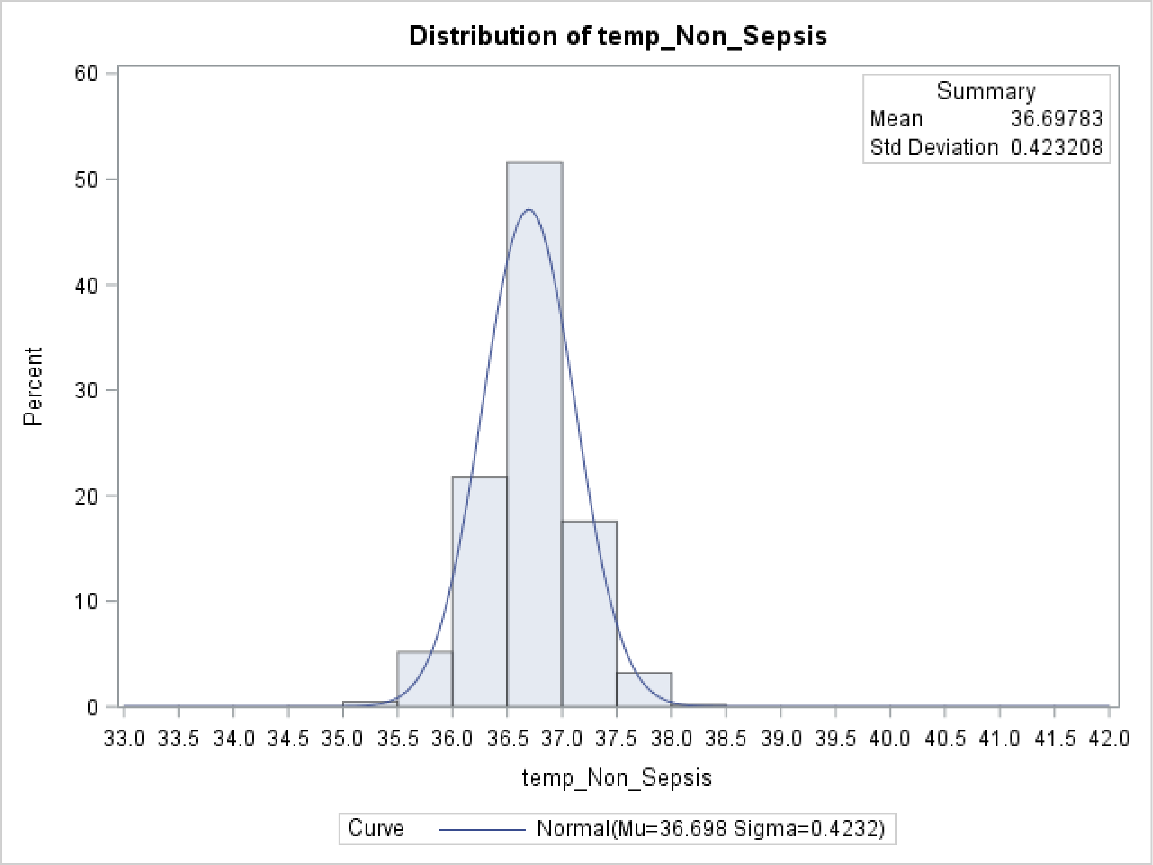


Figure 3.17 Simulated Temperature (Celsius) for Non-Sepsis Patients Histogram

To visualize the comparison between the patient temperature data and the simulated data for non-sepsis patients, the graph below shows the two histograms superimposed onto each other. The sample of non-sepsis patient data is plotted in red and the simulated temperature data is plotted in blue. The simulated data appears to fit the patient data well.

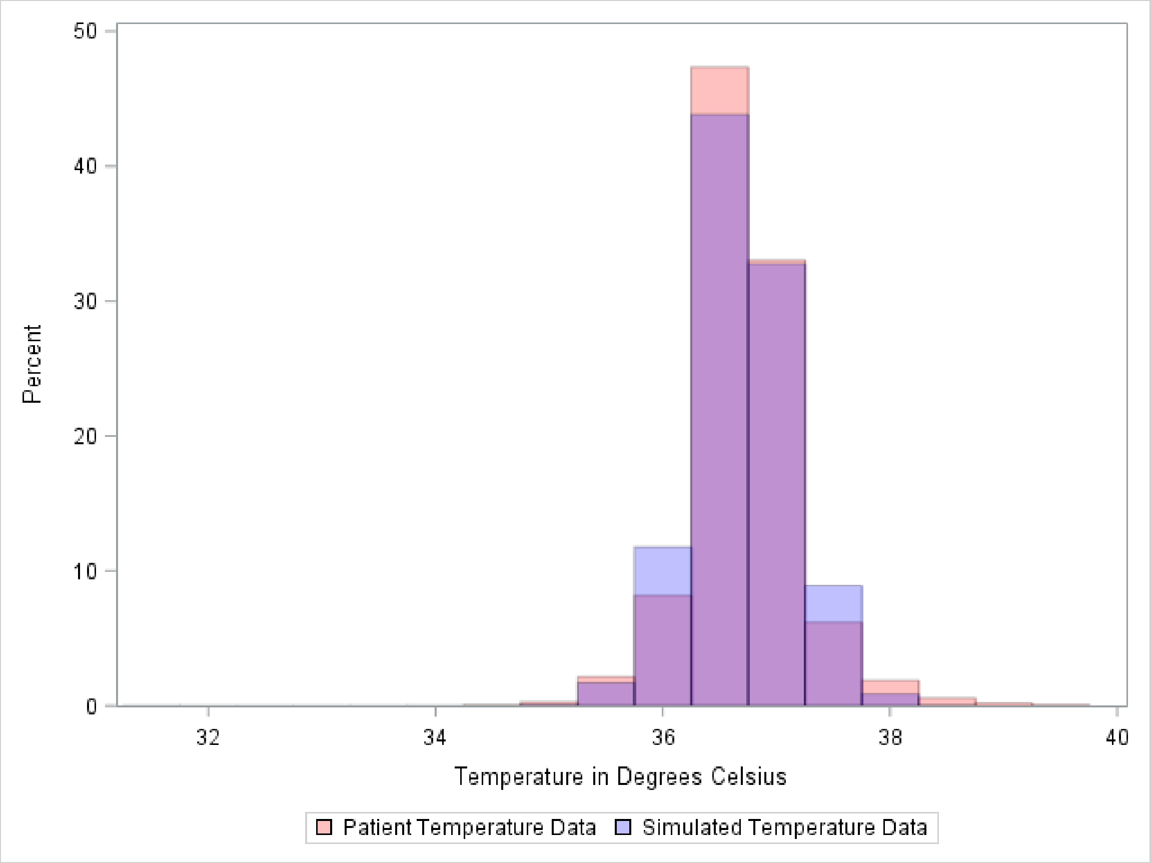


Figure 3.18 Temperature (Celsius) for Non-Sepsis Patients Composite Simulated vs. Patient Data

3.11 Body Temperature for Sepsis Patients

Next, looking at the temperatures for sepsis patients, recall Figure 2.20 showing a histogram of their temperatures in degrees Celsius. Notice again that the graph is bi-modal showing what appears to be a mixture of two normal distributions. One is centered around 36.7 degrees, similar to the non-sepsis patients, and the other has an elevated mean showing patients with an active fever.

To simulate the temperature data for patients with sepsis, a normal distribution with a mean of 36.9 degrees and a standard deviation of 0.8 degrees is mixed with a second normal having a mean of 39 degrees and a standard deviation of 0.5. A mixture of 57% of the first normal distribution and 43% of the second normal distribution approximated the temperature data with the highest level of accuracy. The histogram of this simulated data is shown below.

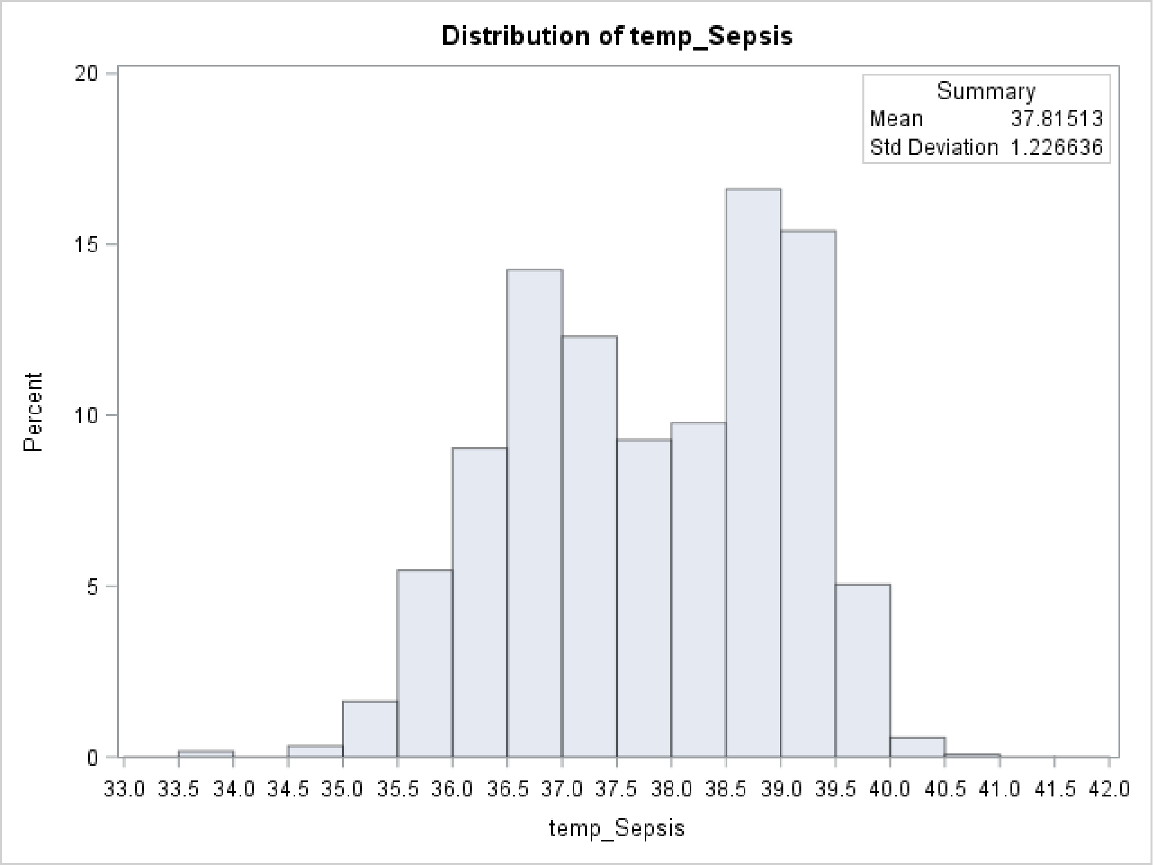


Figure 3.19 Simulated Temperature (Celsius) for Sepsis Patients Histogram

To better visualize the comparison between the patient data and the simulation, the graph below shows the two histograms plotted together. The patient data is graphed in red and the simulated data is plotted in blue. The simulated data for the temperature of patients with an active sepsis diagnosis fits the sample data well.

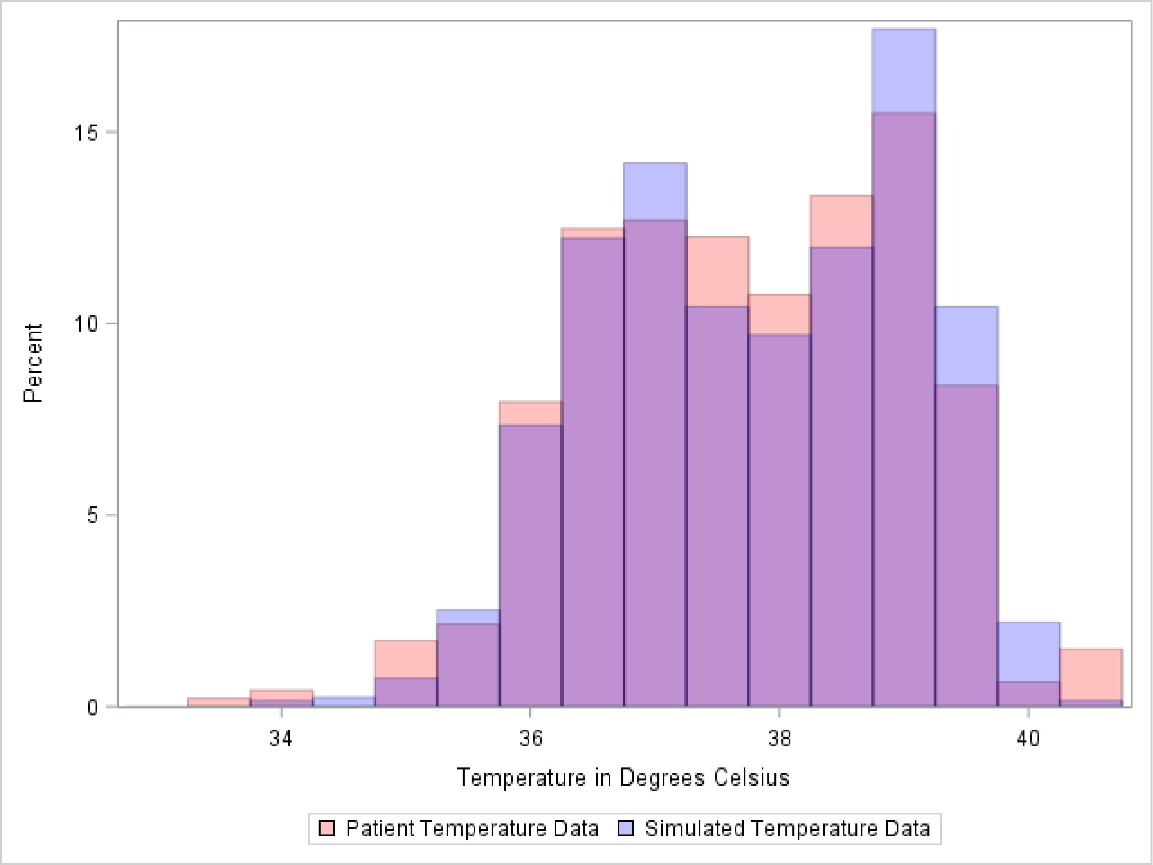


Figure 3.20 Temperature (Celsius) for Sepsis Patients Composite Simulated vs. Patient Data

CHAPTER 4. MEASURING CHANGE

4.1 Introduction

The onset of sepsis spreading throughout a person's body causes systematic changes that can be detected across a person's vital signs. However, not every person reacts to sepsis in the exact same way. As mentioned above, when a sample of patients begin to develop sepsis there are marked changes in the distributions of non-sepsis patients versus sepsis patients, though those distributions are not disjoint. Recall the graphs below, which are repeated for the reader's convenience, that illustrate the distributions of each of the five vital signs contrasting sepsis patients from non-sepsis patients. In each of the vital sign measures, notice the change between non-sepsis patients and sepsis patients. The distributions are significantly different and the measures of centrality are different, however much of the distributions still overlap.

4.2 Heart Rate

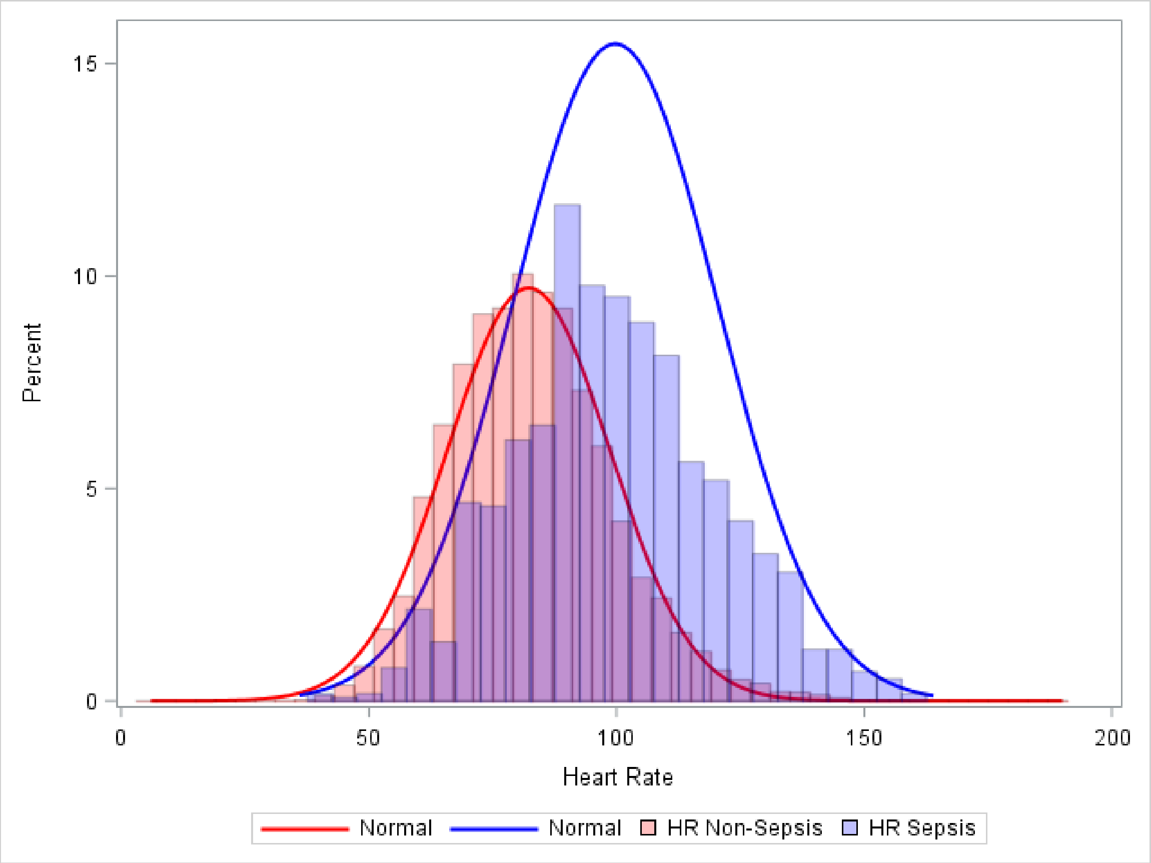


Figure 4.1 Non-Sepsis vs. Sepsis – Heart Rate

As can be seen in Figure 4.1, when comparing the distribution of heart rates of non-sepsis patients to sepsis patients, the mean of the distribution jumps over 17 beats per minute from 82.2 beats per minute up to 99.7 beats per minute, a 21.3% increase in the mean. This rise in the mean coincides with an increase of 25% more of the distribution falling outside of what is generally accepted as the normal range of an adult, 60 to 80 beats per minute (Cleveland Clinic 2014). Even with this dramatic change in the distribution, the majority of the distributions for the two groups overlap. A subset of the non-sepsis population exists who have a higher heart rate as their baseline as compared to other patients. Given a single observation of 96 beats per minute, it would be difficult to state conclusively from which distribution this observation originated. However, notice how the information about the new observation would change if we knew that the baseline heart rate for this patient was 70 beats per minute as opposed to 92 beats per minute.

4.3 Systolic Blood Pressure

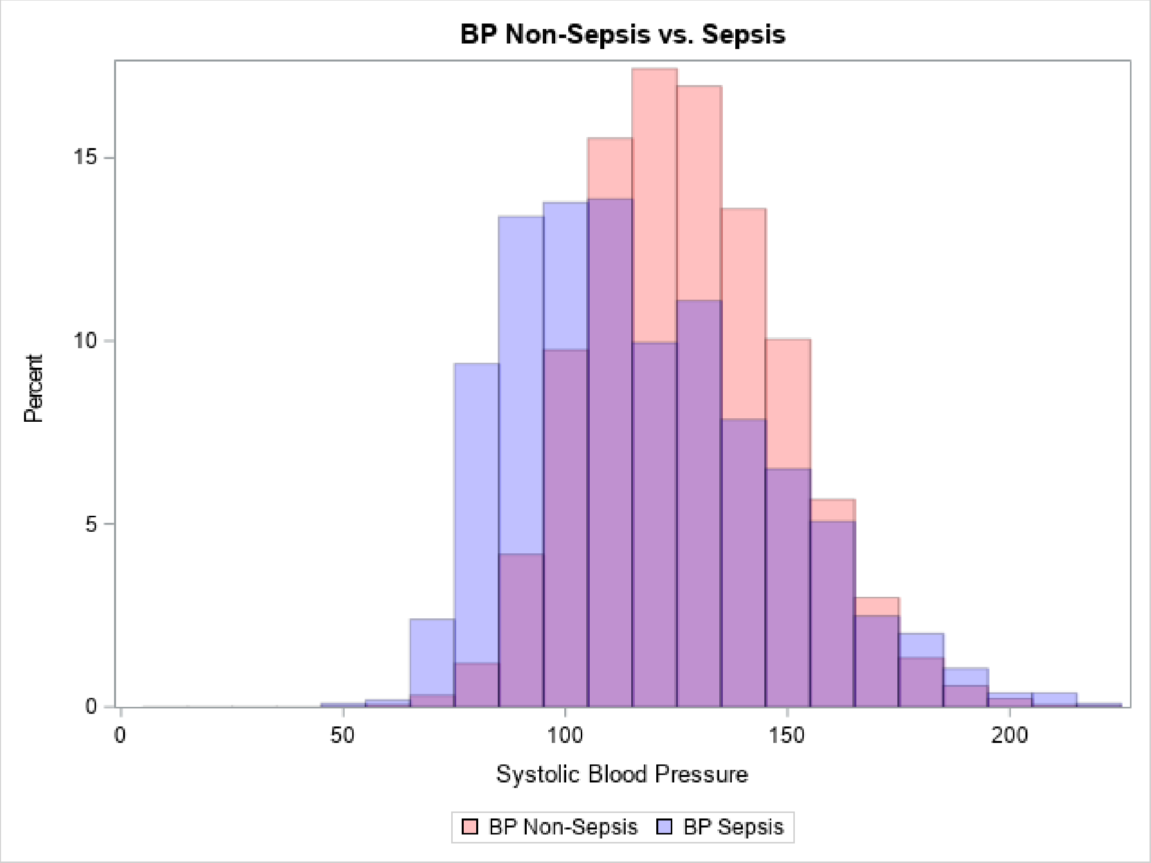


Figure 4.2 Non-Sepsis vs. Sepsis – Systolic Blood Pressure

The mean systolic blood pressure for a sepsis patient drops 10 mmHG compared to the mean of a non-sepsis patient. This change in the central tendency coincides with nearly a 30% increase in the standard deviation of the distribution. As can be seen in Figure 4.2, the two distributions almost completely overlap. An observation of a systolic blood pressure measure of 105 mmHG reasonably fits into both distributions. However, the observation points to a change due to sepsis if the baseline observation for the patient was observed to be 140 mmHG as compared to 110 mmHG.

4.4 Respiratory Rate

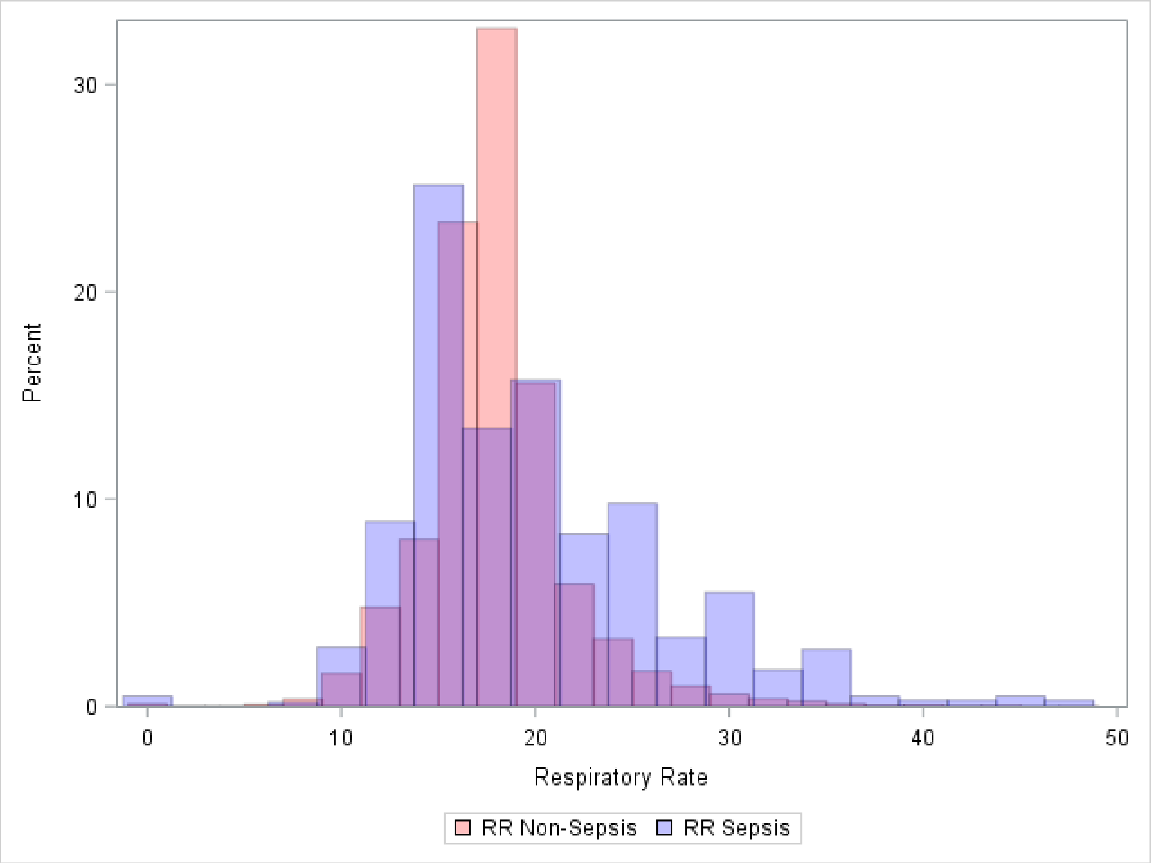


Figure 4.3 Non-Sepsis vs. Sepsis – Respiratory Rate

The mean respiratory rate of a sepsis patient is only 2.2 breaths per minute higher than the respiratory rate for a non-sepsis patient. However, in Figure 4.3, you can see that the distributions are quite different. The standard deviation of the distribution jumps from 3.79 breaths per minute up to 6.79 breaths per minute, nearly an 80% increase. The small change in mean along with the large increase in spread leads to the non-sepsis distribution being completely encompassed by the sepsis distribution. An observation of 23 breaths per minute easily fits into both distributions, however the story changes if the baseline for the patient was 17 breaths per minute compared to 21.

4.5 Peripheral Capillary Oxygen Saturation

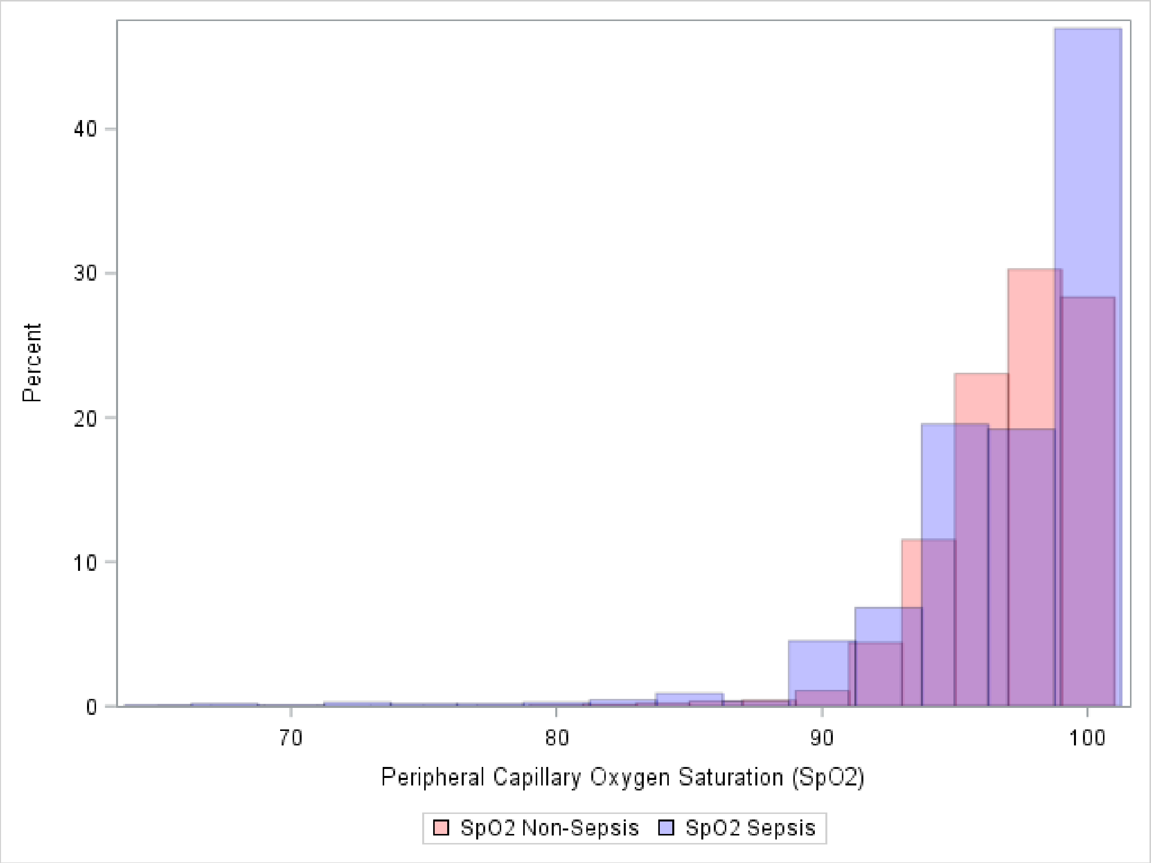


Figure 4.4 Non-Sepsis vs. Sepsis – Peripheral Capillary Oxygen Saturation

The mean peripheral capillary oxygen saturation (S_pO_2) of both groups are nearly identical at 96.68% and 96.93%. However, the standard deviations do show a larger change. As can be seen in the larger tail of the sepsis patients, the standard deviation jumps from 2.83% up to 4.15%, a 47% increase in sepsis patients compared to non-sepsis patients. Looking at Figure 4.4, it is worth noting that not all sepsis patients see a drop in their S_pO_2, but when they do, they generally see a larger drop as compared to non-sepsis patients.

4.6 Body Temperature

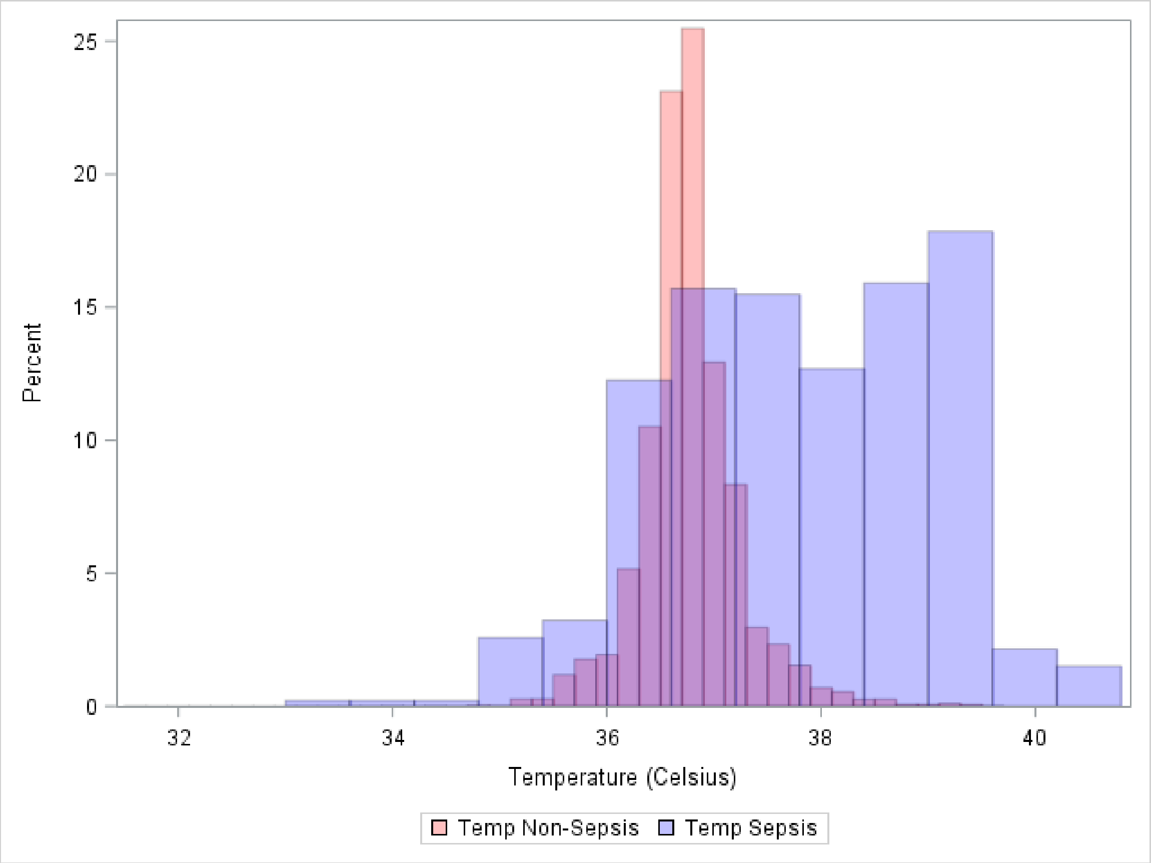


Figure 4.5 Non-Sepsis vs. Sepsis: Body Temperature (Celsius)

Temperature is perhaps the measure that shows the most dramatic change during the onset of sepsis. However, as can be seen in Figure 4.5, it is one of the most easily masked of the vital signs. The distribution of a non-sepsis patient is quite consistent with a mean of 36.7 degrees Celsius in our dataset and a small standard deviation of 0.47 degrees. The mean of the sepsis distribution is only one degree higher. However, as noted previously, the distribution for sepsis patients shows a clear bimodal distribution which has been shown to be a mixture of two different distributions. The mean of the higher distribution is 39 degrees Celsius. The onset of sepsis, with the spreading of an infection through the blood stream of a patient, leads to a fever. This alone is not uncommon in sick patients and especially not uncommon in a hospital setting. However, the treatment could mask the symptoms of sepsis and again illustrates why it is important to incorporate the change in the vital sign along with the measure.

4.7 Summary

There is a consistent pattern across most of the vital signs, a dramatic shift in the central measure or shape of the distribution, but the majority of the distributions between sepsis and non-sepsis patients still overlap. This leads to an interesting problem where a patient's single observation from either distribution is difficult to discriminate. However, if that singular observation is compared to the patient's baseline in each of the vital signs, the shift across each of the vital signs could be captured. This has the potential to greatly improve the discrimination of a singular set of observations. Thus, a patient who has a current heart rate of 100 beats per minutes and a baseline heart rate of 75 beats per minute is distinguished from a patient with a current heart rate of 100 beats per minute and a baseline of 103 beats per minute. To establish a patient specific baseline, the mean measure for each observation prior to the current measure is used during that specific hospital visit.

CHAPTER 5. DISCUSSION ON VARIATION

5.1 Introduction

As a patient establishes a set of baseline measures for their vital signs, they not only establish a baseline measure of centrality, but also a baseline variation. To investigate the variability of variation of patients across the different vital sign measures, the following graphs show histograms of standard deviations of non-sepsis patients for each measure. This allows us to determine if it is important to consider patient specific standard deviations, for example, when evaluating the change in a person's heartrate. If one patient shows a very consistent heart rate with a standard deviation of five, and has a change of 20 beats per minute, this is distinguishable between a patient who has an erratic heart rate with a standard deviation of 18, and has a change of 20 beats per minute. Since the focus of this research is to determine when patients residing in non-ICU settings enter into a septic state, only the standard deviations of non-sepsis patients are examined. The histograms are shown in repetition with discussion held to the end.

5.2 Heart Rate Standard Deviation

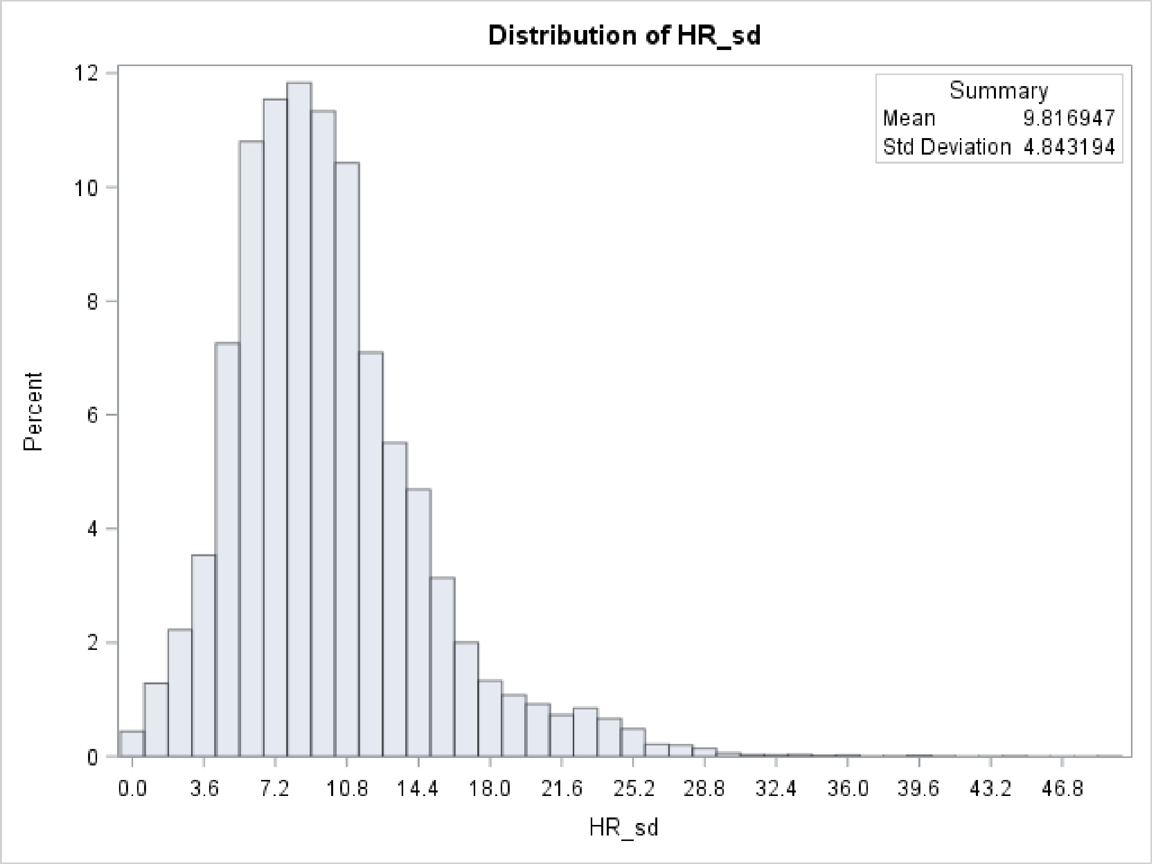


Figure 5.1 Histogram of Heart Rate Standard Deviations for Non-Sepsis Patients

5.3 Respiratory Rate Standard Deviation

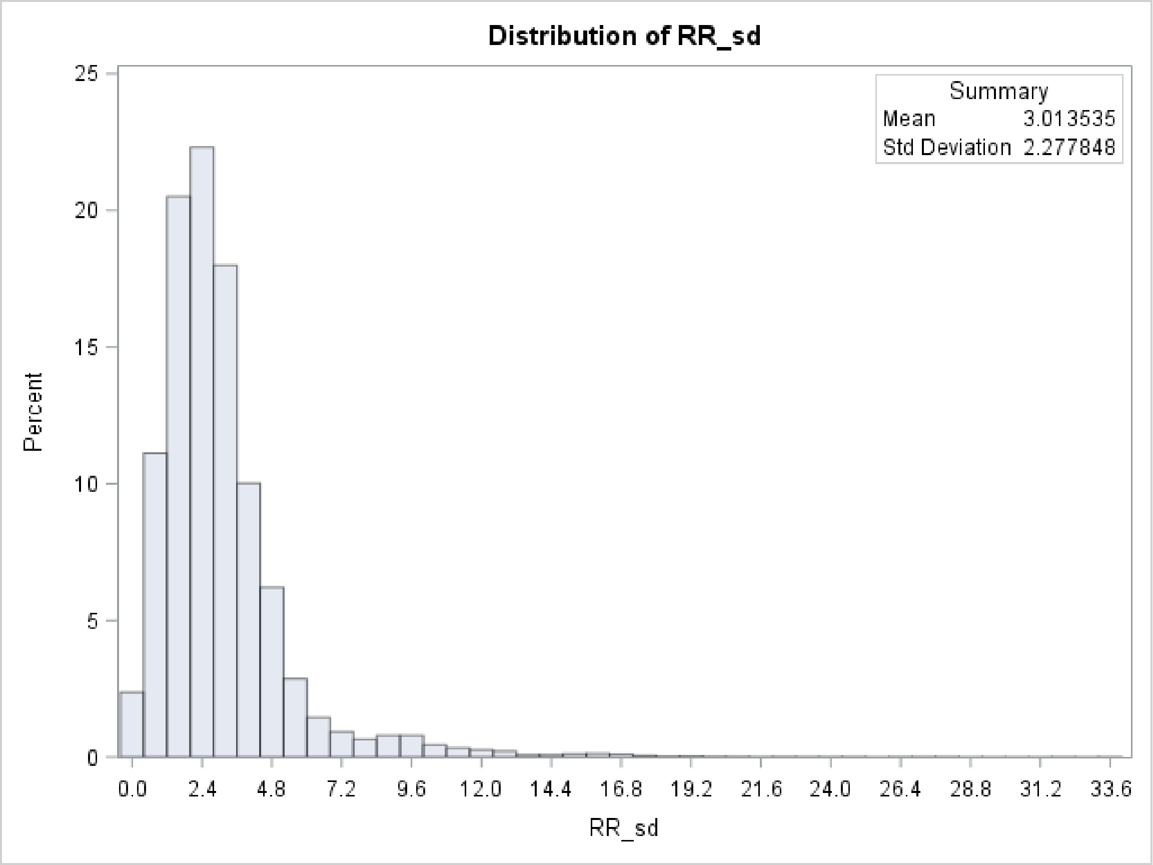


Figure 5.2 Histogram of Respiratory Rate Standard Deviations for Non-Sepsis Patients

5.4 Systolic Blood Pressure Standard Deviation

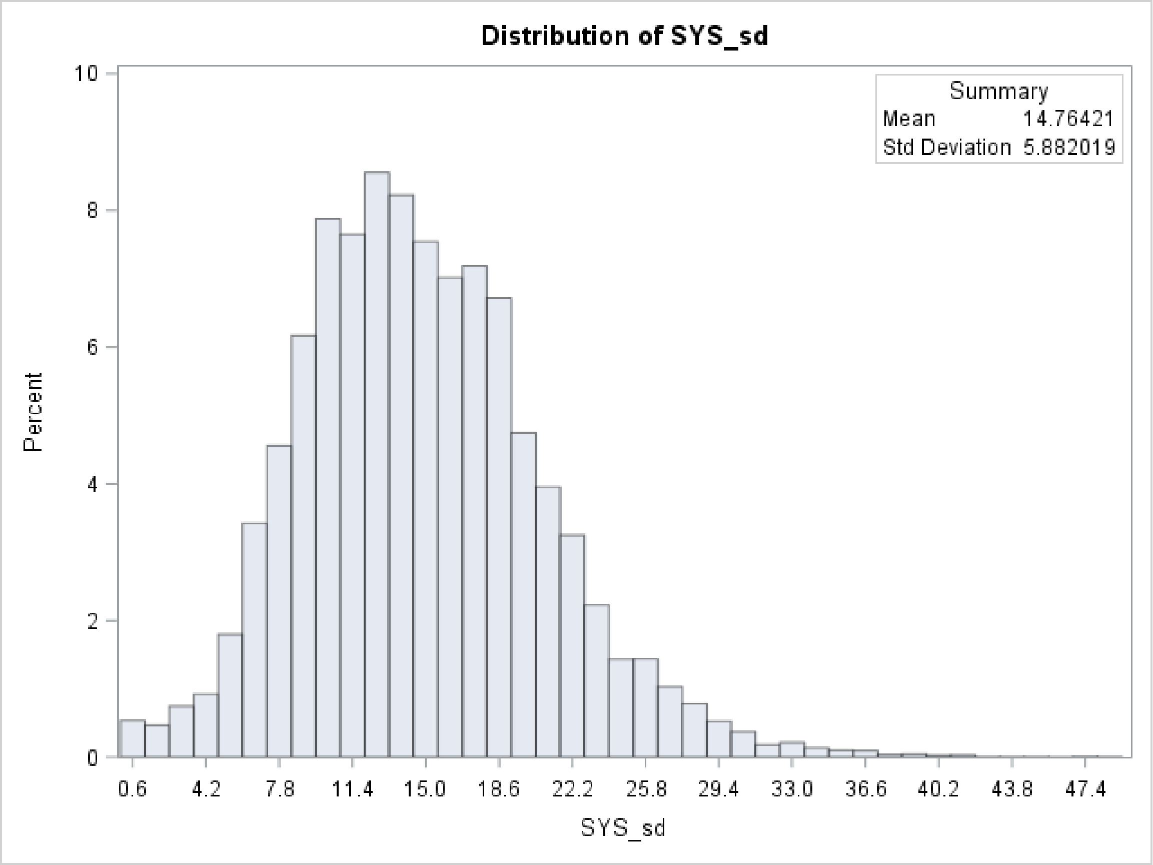


Figure 5.3 Histogram of Systolic Blood Pressure Standard Deviations for Non-Sepsis Patients

5.5 Peripheral Capillary Oxygen Saturation Standard Deviation

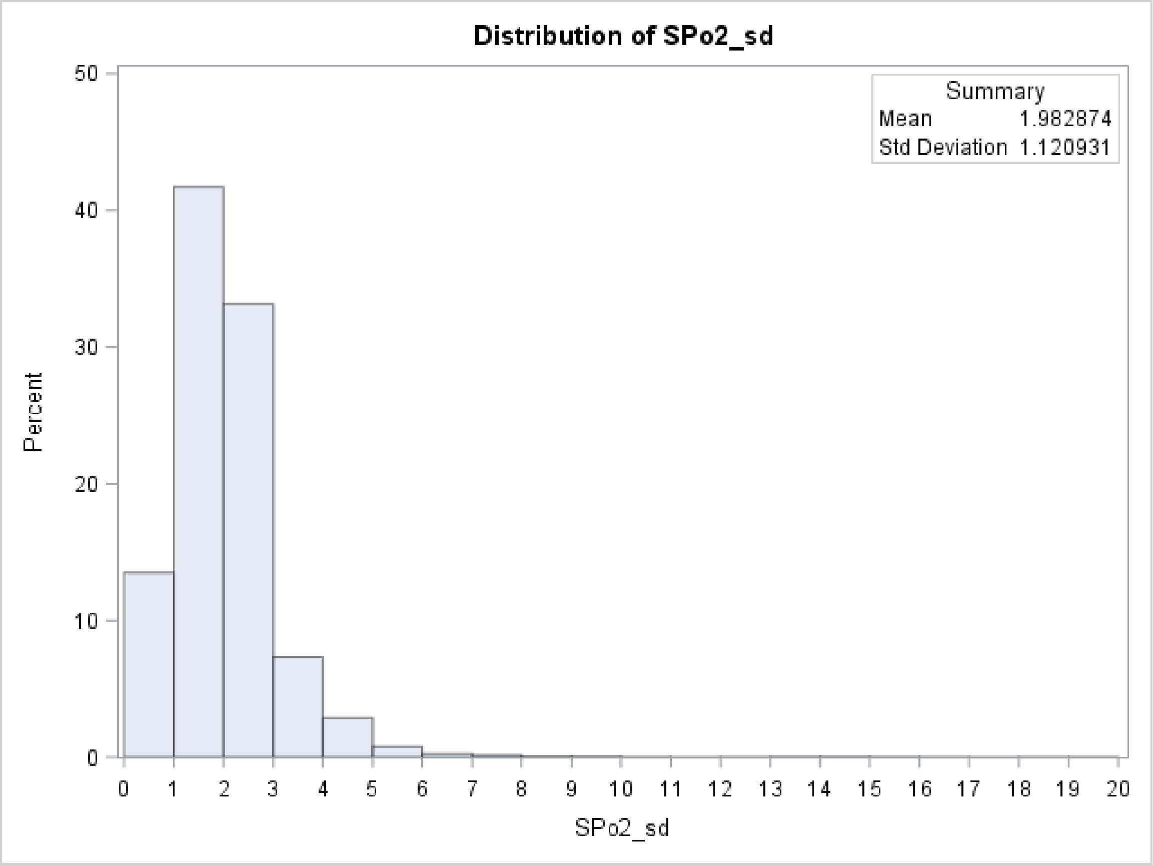


Figure 5.4 Histogram of S_pO_2 Standard Deviations for Non-Sepsis Patients

5.6 Body Temperature Standard Deviation

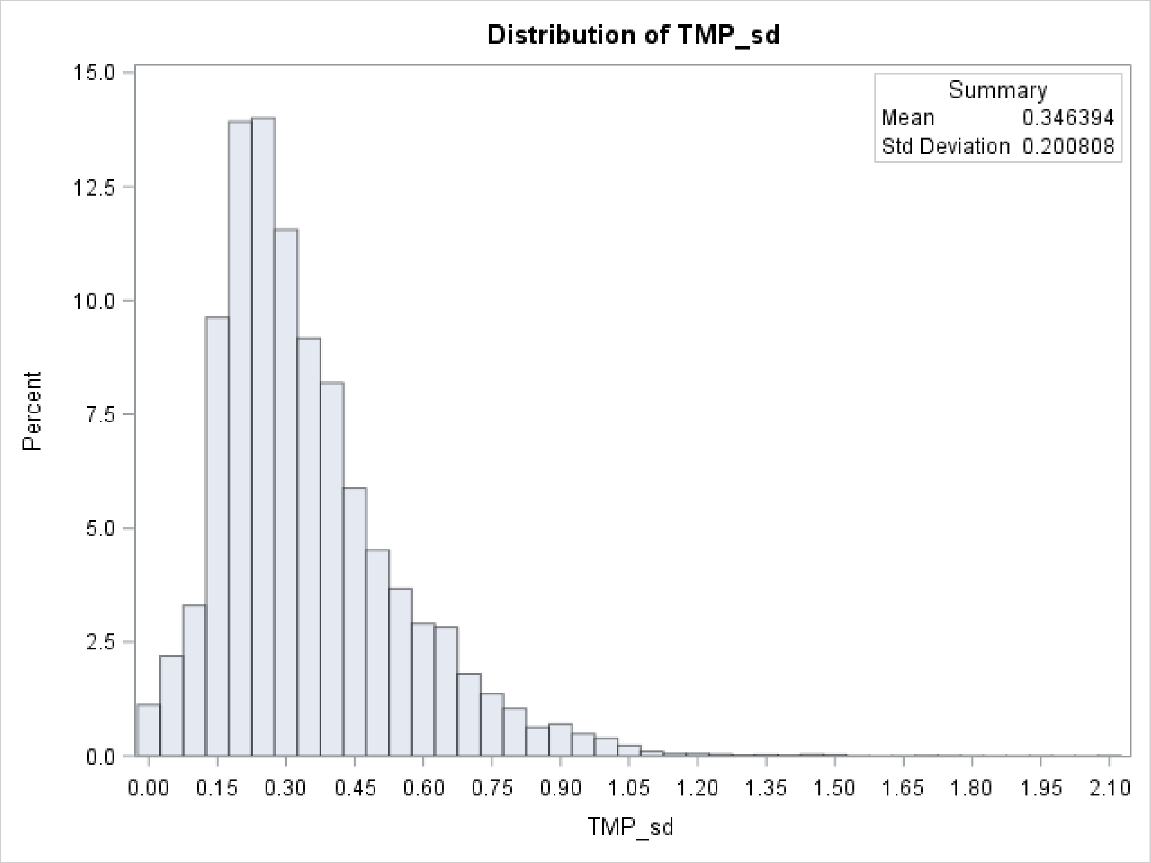
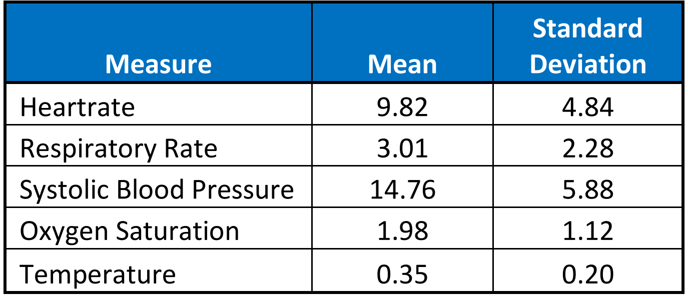


Figure 5.5 Histogram of Temperature (Celsius) Standard Deviations for Non-Sepsis Patients

5.7 Summary

Table 5.1 summarizes the five histograms for the standard deviations of the vital signs. As can be seen in each vital sign, the between patient variation of standard deviations is significant and needs to be considered when evaluating the change.

Table 5.1 Summary of Variation Among Standard Deviations of Vital Signs



Measure	Mean	Standard Deviation
Heartrate	9.82	4.84
Respiratory Rate	3.01	2.28
Systolic Blood Pressure	14.76	5.88
Oxygen Saturation	1.98	1.12
Temperature	0.35	0.20

The amount of variation that is present across the standard deviations of the vital sign measures offers that it is also important to consider the patient's baseline variation in each measure. It is also informative that the variation in heart rate and systolic blood pressure is likely more patient specific given the amount of variation found between patients. Recall Figure 1.2 that illustrates the heart rates of two different patients who have an average heart rate of approximately 80 beats per minute. Patient A has a standard deviation of three beats per minute, and Patient B has a standard deviation of ten. The final observation of 100 for each patient is likely telling a different story between patients even though they have very similar means. For Patient A, the observation of 100 is over seven standard deviations from the mean and represents a significant change for the patient. Conversely, the observation for Patient B is under two standard deviations from the mean, a much less significant change.

CHAPTER 6. THE SEPSIS SCORE

6.1 Introduction

In order to increase the lead time in recognizing patients trending towards septicemia, a partial least squares discriminate model was developed to identify patients who are starting to show signs of septicemia in their vital signs. The multivariate model was utilized in order to keep the full set of variables without reducing the model. This allows us to include variables where change may be detected even though they are not, on average, the best indicators. This decision also allows for the detection of large single system changes, that although may not be indicative of sepsis, could be important to healthcare providers. These changes present differently in the model than the onset of sepsis. Large single system changes cause the sepsis score to change significantly, but not always cross into the sepsis zone. These are often significant changes to the patient that need the attention of a doctor or nurse, but can also be results of less emergent circumstances. Examples include a nurse who accidently mistyped 205 instead of 105 when entering a heart rate, or a patient who had recently been taken off of heart medicine causing their heart rate to jump. In both of these examples, a notification to the nurse is all that was needed to address the issue.

6.2 The Sepsis Score

A random sample of 1275 observations from patients with and without sepsis was collected from the electronic medical records at the University of Kentucky Albert B. Chandler Hospital. The sample included 872 non-sepsis observations and 403 sepsis observations. The non-septic patients were selected only if they did not have a diagnosis related to sepsis and they did not expire during their visit. Vital signs taken of the sepsis cohort were during a two hour window before and after the time they were identified with a sepsis diagnosis and interventions were started. This sample was employed to ensure the measures were indicative of vital sign changes during and just prior to the onset of sepsis. A partial least squares discriminate model was developed utilizing the patient's age, current vitals, and the change variables defined below for each vital sign.

$$A_n = \frac{X_n - \bar{X}_{n-1}}{s_{n-1}}$$

Where $\bar{X}_{n-1} = \frac{1}{n-1}\sum_{j=1}^{n-1} X_j$ and $s_{n-1}^2 = \frac{\sum_{k=1}^{n-1}(X_k - \bar{X}_{k-1})^2}{n-2}$, the mean and standard deviation of the patient's previous observations, respectively. The overall model was significant with a p-value < 0.0001. A resubstitution strategy was used for grading the linear discriminant function. The specificity, identifying a non-sepsis patient as non-sepsis, was 89.7%, giving a false positive rate of 10.3%. Conversely, the sensitivity, identifying a sepsis patient as having sepsis, was 77.9%, giving us a false negative rate of 22.1%. The figure below shows the separation between groups given by the discriminant model. Sepsis observations are denoted by a blue circle and non-sepsis observations are shown with a red x.

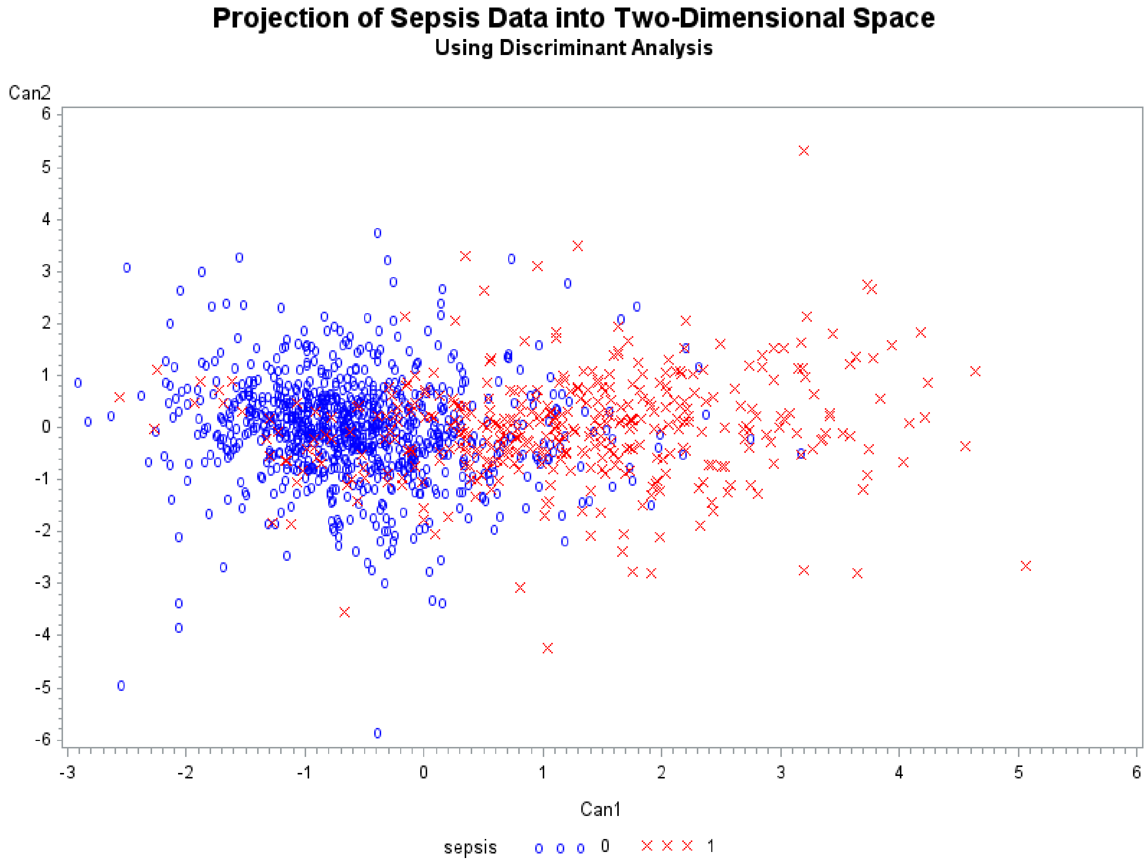


Figure 6.1 Projection of Data into a Two-Dimensional Space using the Discriminant Model. Sepsis observations are blue circles and non-sepsis observation are denoted with a red x.

6.3 Utilizing the Sepsis Score

The sepsis score can be viewed over time for each patient as seen in Figure 6.2 below. The model designates zero as the line of discrimination, where a score above zero signifies a patient who is trending towards sepsis.

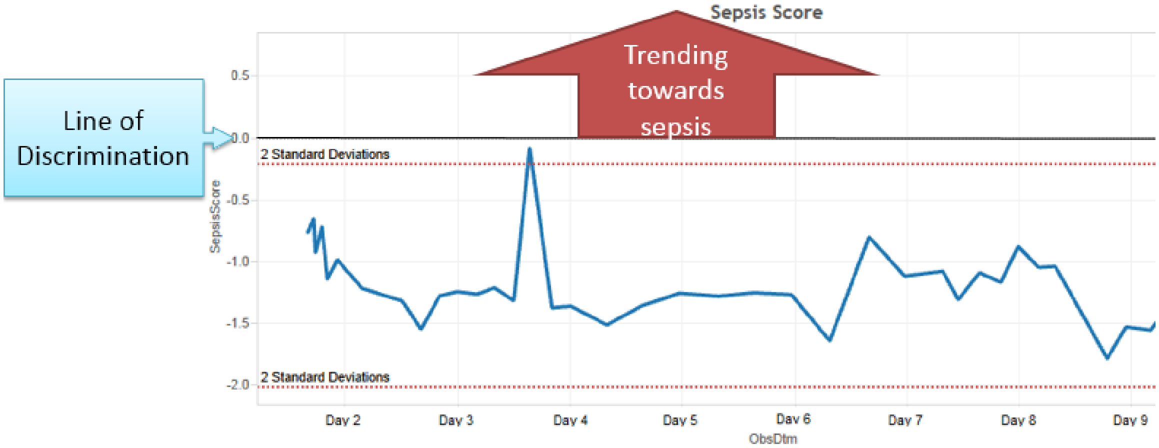


Figure 6.2 Visualizing the Sepsis Score. The figure identifies the line of discrimination in black, the sepsis score in blue, and a two standard deviation control chart lines in red.

There are three different types of alerts which are triggered by this calculated sepsis score. A yellow alert is generated when a patient's sepsis score has gone outside of their patient specific control chart. A control chart is generated by calculating plus and minus two standard deviations from their mean sepsis score, and is designated above with the red dotted lines. A visual example of a yellow alert can be seen between days three and four in Figure 6.2. Generally, these alerts are caused by large single system changes. This is a nursing centric alert that only notifies the nurse, who evaluates the situation and determines if any escalation is needed. An orange alert is generated when a patient's sepsis score touches the line of discrimination but does not jump across it. This alert signifies that a patient is trending in the wrong direction. This is also a nursing centric alert that notifies the nurse to perform a sepsis screening on the patient and look for a possible source of infection. Based upon this screening the nurse can alert the rapid response team and the physician if necessary. The final alert is a red alert and is generated when a patient's sepsis score crosses the model's line of discrimination for sepsis. This alert generates a page to the rapid response team to evaluate the patient. In the UKHealthCare environment, the rapid response team also carries a tool to take a point of care lactate value to give baseline lab information to the physician. Figure 6.3, below, outlines the process flow for the three alerts.

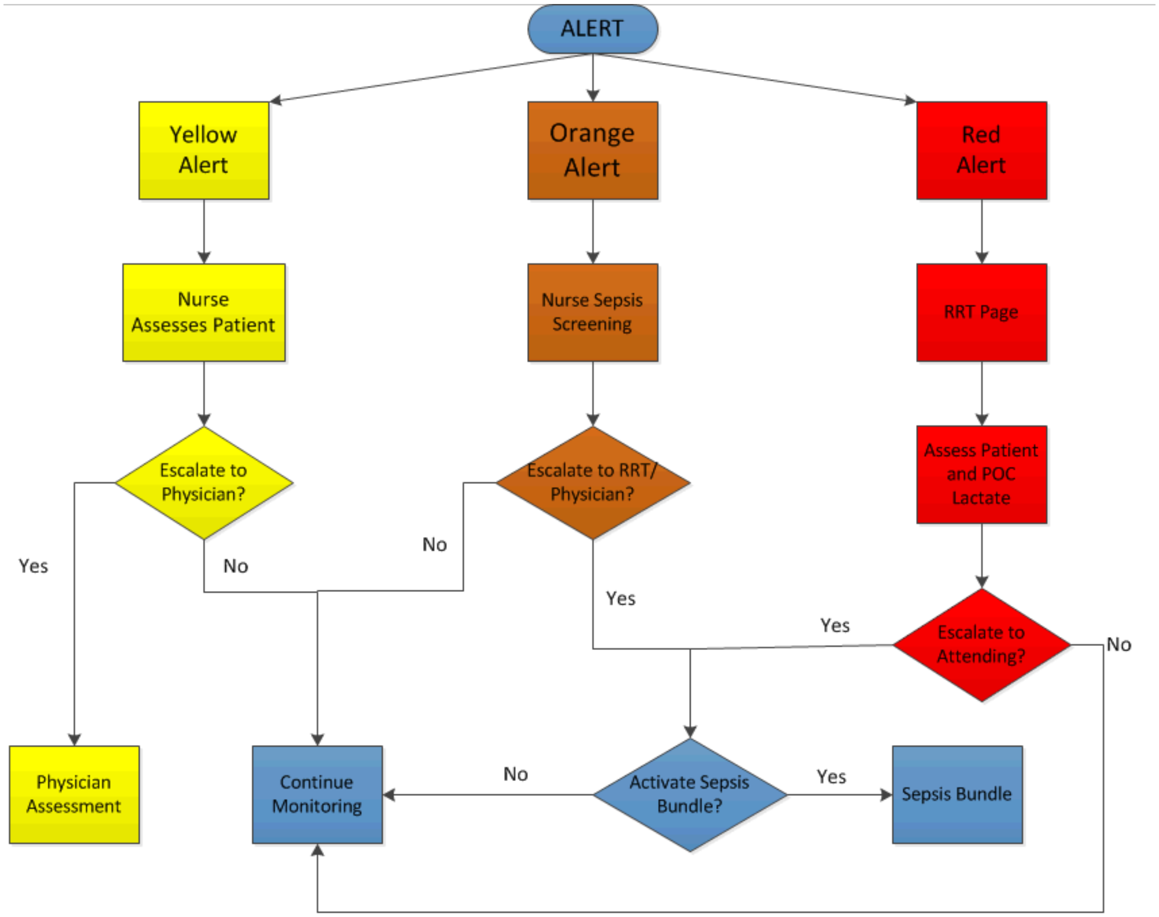


Figure 6.3 Flow Chart for Sepsis Alerts

6.4 Example Patients

Three examples are provided to illustrate the sepsis score on patients while also showing the change in their vital signs. Figure 6.5 is an example of the sepsis score actively being calculated during a patient's visit. The last two are retrospective patients who were coded with a sepsis diagnosis and had the sepsis score calculated on their vital signs post hoc. A graph of the patient's vital signs below the sepsis score graph is provided in order to see what changes in the vital signs caused the changes in the sepsis score. Note that the axis on the left is for the heart rate, respiratory rate, peripheral capillary oxygen saturation, and systolic blood pressure. The axis on the right is utilized for the patient's body temperature. Figure 6.4, below, is a legend for the vital signs graphs of all three patients.

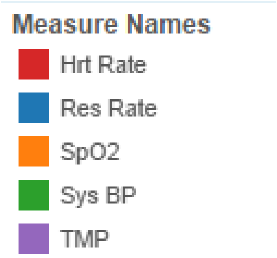


Figure 6.4 Legend for Vital Sign Measures

6.4.1 Patient One

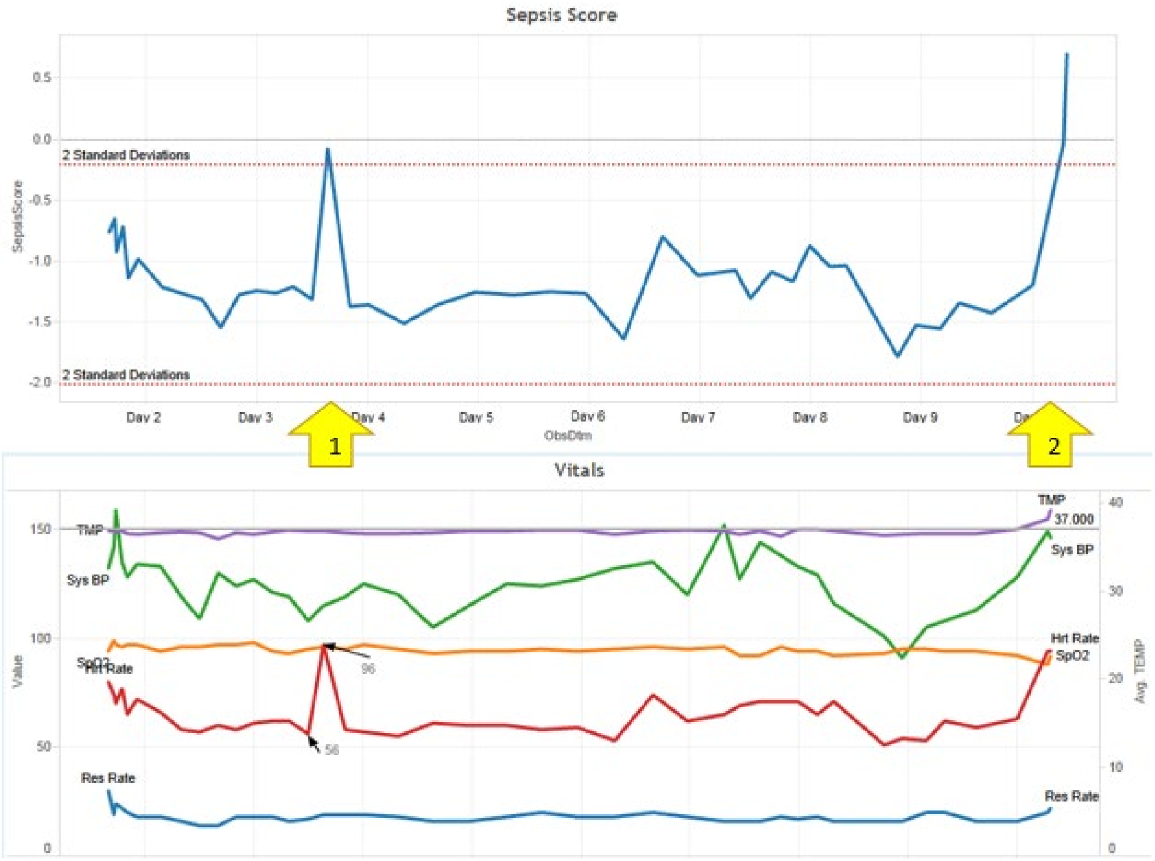


Figure 6.5 Sepsis Score and Vital Signs for Patient One

At time point one, signified by the yellow arrow, the patient's heart rate jumped from a baseline around 60 bpm to 96 bpm. However, at this time there was not a significant change in the patient's other vital signs. This single system change resulted in the sepsis score jumping outside of the control chart, a yellow alert, but not crossing the line of discrimination. Contrast this with time point 2, when the patient had a similar jump in heart rate, but this time it coincided with a jump in temperature, a slight rise in the respiratory rate, and a jump in systolic blood pressure. These combined changes resulted in the sepsis score jumping over the line of discrimination and a red alert.

6.4.2 Patient Two

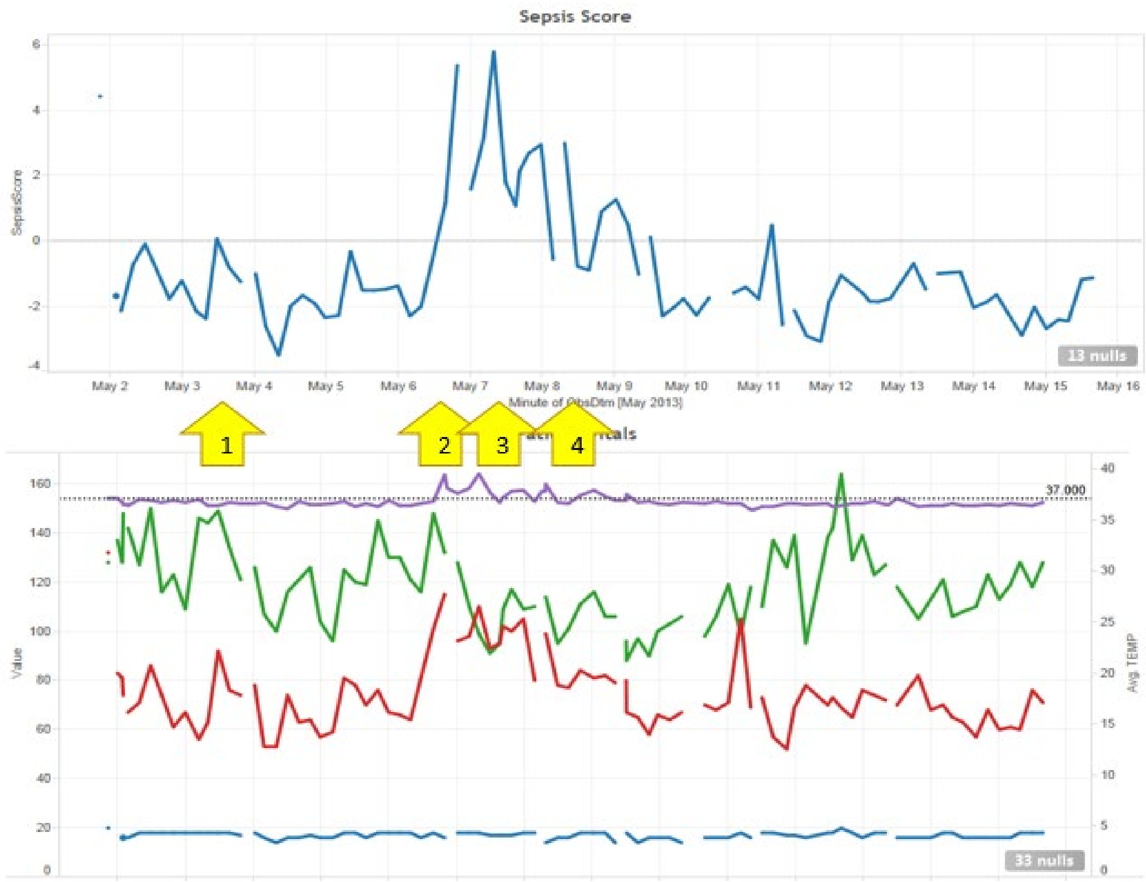


Figure 6.6 Sepsis Score and Vital Signs for Patient Two

Patient two had a length of stay of fifteen days and coded out with bacteremia, an early form of sepsis. At time point 1, the patient had a rise in their heart rate and systolic blood pressure. This resulted in the sepsis score touching the line of discrimination, but not jumping over it. This would have resulted in an orange alert and a sepsis screening from a nurse. Time point two occurred at 3:38 PM on day six when the patient had jumps in their heart rate, systolic blood pressure and temperature resulting in a red alert. Time point three on day seven at 10:33 AM, over eighteen hours after the red alert, is the first time that antibiotics were ordered. At time point four, the antibiotics were changed and the patient began showing improvement.

6.4.3 Patient Three

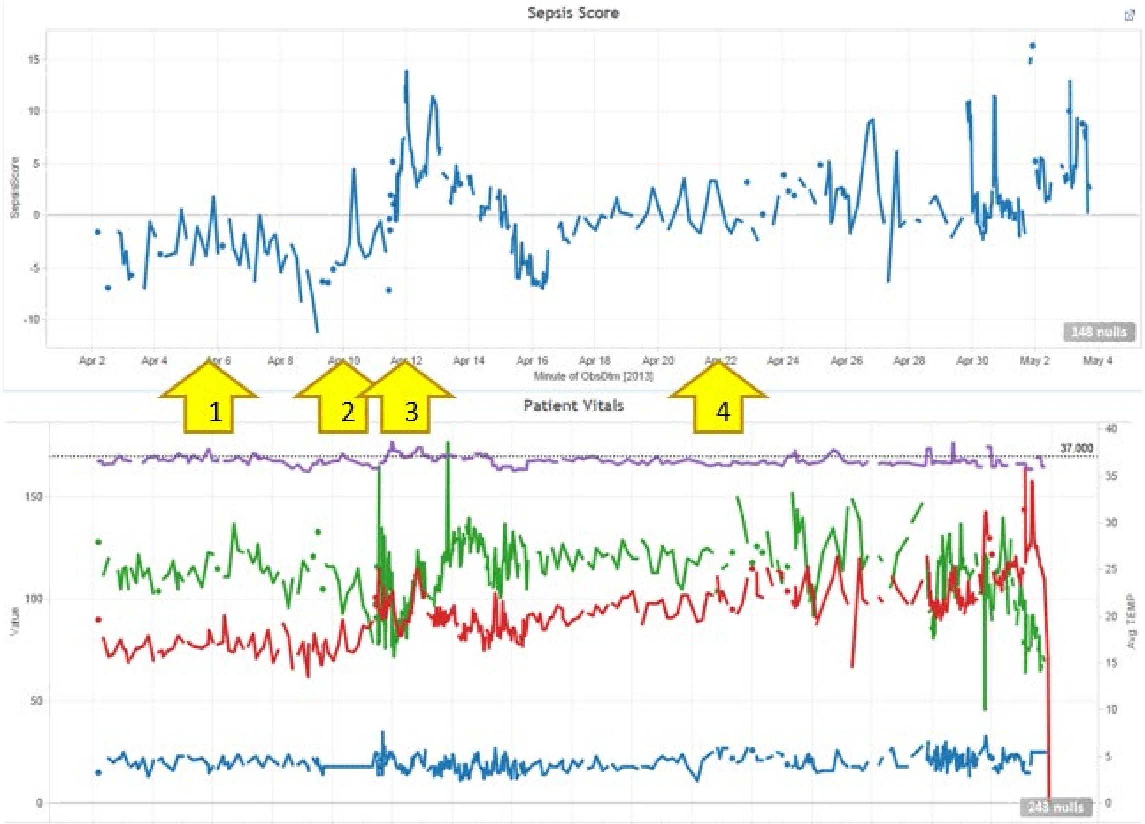


Figure 6.7 Sepsis Score and Vital Signs for Patient Three

Patient three had a length of stay of 31 days, was diagnosed with septic shock, and expired during their hospital stay. At time point 1 on day three at 8:00 am, the patient would have had a red alert and a baseline lactate taken. At time point 2 on day eight at noon, the patient would have had a second red alert. At time point 3 on day nine at 10:14 am is the first time that antibiotics were ordered, 22 hours later. Time period 4 on day 20 is the first time that the traditional track and trigger system fired on the patient.

6.5 Results from Trial

SAS was utilized to generate a partial least squares discriminant model. The overall model was significant, determined by an F test, to evaluate if the canonical correlations are significantly different from zero (p < 0.0001). The sepsis score was run in the background for four months at the University of Kentucky HealthCare. During this time 200 patients were coded as having a hospital acquired diagnosis of sepsis. For each of these patients, the first time each patient had any of the following antibiotics ordered was documented: cefepime, fluconazole, piperacillin, tobramycin, or vancomycin. This time was compared with the time of a sepsis alert that would have fired preceding the start of the antibiotics to examine the possible lead time that could have been achieved through the alert. Of the 200 patients, 117 of the patients had an alert before their first dose of antibiotics, 21 did not have an occurrence of the antibiotics in their charts, and 62 received antibiotics before any alerts. However, of the 62 patients, 49 of them received antibiotics within the first 24 hours of their first vital signs which leaves little data to build baseline information on the patient. Table 6.1 shows this breakdown.

Table 6.1 Summary of 200 Sepsis Patients

Categories	**Average Lead Time before Antibiotics**	**Number of Patients**
Alerted	25:37:09	117
No Antibiotics (Abx)	0:00:00	21
No Preceeding Alert		
Abx after 24 hrs	0:00:00	13
Abx within first 24 hrs	0:00:00	49
Grand Total	**14:59:14**	**200**

As seen above, the average lead time for the 117 patients who would have alerted before they received antibiotics was approximately twenty-five and a half hours. Removing outliers beyond three days results in a mean of 14:46:49 (hh:mm:ss) and a median of 8:31:29 (hh:mm:ss) of lead time. Figure 6.8 shows the histogram of lead times excluding outliers over three days.

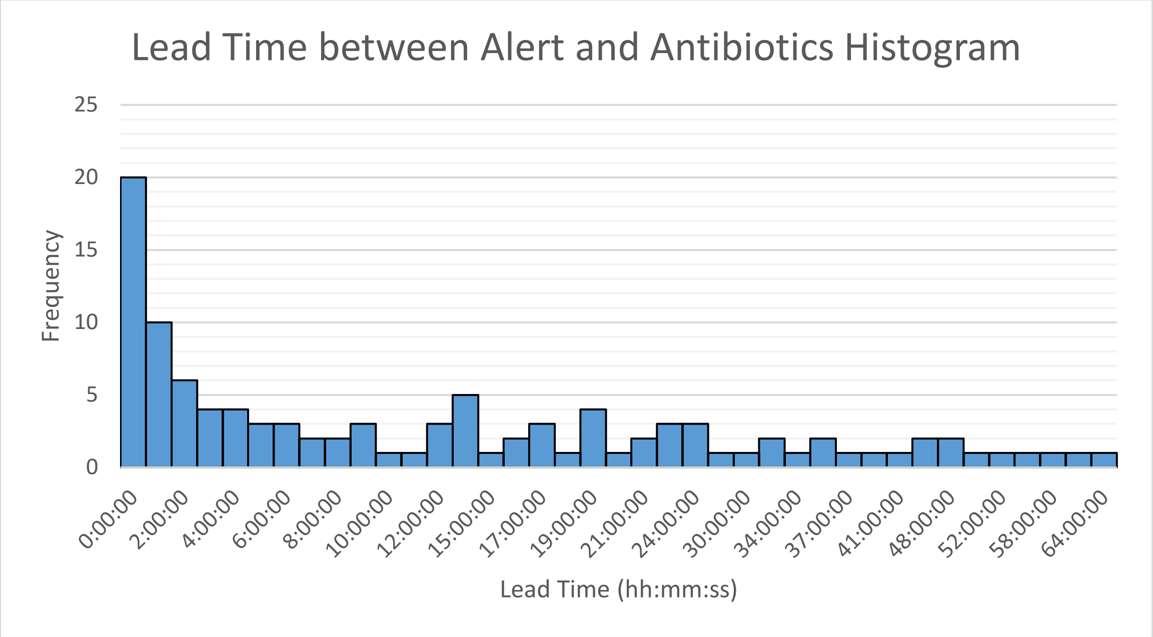


Figure 6.8 Lead Time between Alert and Antibiotics Histogram

6.6 Conclusions

The model requires patient baseline data is kept up-to-date for each patient's vital signs throughout their stay. When a new set of vital signs is entered in the electronic medical record, the raw scores and the change from baseline data is utilized to calculate a sepsis score.

The use of an automated system that utilize both raw data and the change of the data from patient specific baselines may prove useful for earlier detection of sepsis in progressive and acute care patients. The results show that a median of eight and a half hours of lead time to start of interventions could be achieved using similar methods. However, given the large amount of variability within the human body, the onset of sepsis cannot be predicted with complete accuracy. The goal is to identify patients at risk and bring them to the attention of qualified healthcare workers as early as possible.

CHAPTER 7. STATISTICAL THEORY

7.1 Theory

We will use the subscript a, b to denote a sum going from $a+1$ to b.

Lemma 1: Let $\{X_i\}_{i=1}^{n}$ be i.i.d. random variables from a normal distribution with mean μ and variance σ^2. Define $\bar{X}_{0,n-1} = \frac{1}{n-1}\sum_{j=1}^{n-1} X_j$ and $s_{0,n-1}^2 = \frac{\sum_{k=1}^{n-1}\left(X_k-\bar{X}_{0,n-1}\right)^2}{n-2}$, and $A_n = \frac{X_n-\bar{X}_{0,n-1}}{s_{0,n-1}}$. Then, $\frac{1}{\sqrt{1+\frac{1}{n-1}}} A_n \sim t_{n-2}$.

Proof:

(1) $$A_n = \frac{X_n-\bar{X}_{0,n-1}}{s_{0,n-1}} = \frac{X_n-\bar{X}_{0,n-1}}{\sqrt{\sigma^2+\sigma^2/_{n-1}}\cdot\frac{s_{0,n-1}}{\sqrt{\sigma^2+\sigma^2/_{n-1}}}} = Z\cdot\frac{1}{\frac{s_{0,n-1}}{\sqrt{\sigma^2+\sigma^2/_{n-1}}}},$$

where $Z = \frac{X_n-\bar{X}_{0,n-1}}{\sqrt{\sigma^2+\sigma^2/_{n-1}}} \sim N(0,1)$ since $X_n \sim N(\mu,\sigma^2)$ and $\bar{X}_{0,n-1} \sim N\left(\mu,\frac{\sigma^2}{n-1}\right)$.

Thus,

(2) $$A_n = Z\cdot\frac{1}{\frac{s_{0,n-1}}{\sqrt{\sigma^2+\frac{\sigma^2}{n-1}}}\cdot\sqrt{1+\frac{1}{n-1}}} = \frac{Z}{\sqrt{\frac{\sum_{k=1}^{n-1}\left(X_k-\bar{X}_{0,n-1}\right)^2}{(n-2)\sigma^2}}}\cdot\sqrt{1+\frac{1}{n-1}}$$

(3) $$A_n = \frac{Z}{\sqrt{\frac{\chi_{n-2}^2}{(n-2)}}}\cdot\sqrt{1+\frac{1}{n-1}}$$

where χ_{n-2}^2 is a Chi-Square distribution with *n-2* degrees of freedom.

Therefore, $\frac{1}{\sqrt{1+\frac{1}{n-1}}}A_n \sim t_{n-2}$.

Lemma 2: Let $\{X_i\}_{i=1}^{n+m}$ be i.i.d. random variables from a normal distribution with mean μ and variance σ^2. Define $\bar{X}_{n,m} = \frac{1}{m}\sum_{j=n+1}^{m} X_j$, $\bar{X}_{0,n} = \frac{1}{n}\sum_{j=1}^{n} X_j$, $s_{n,m}^2 = \frac{\sum_{k=n+1}^{m}\left(X_k - \bar{X}_{n,m}\right)^2}{m-1}$, and $B_{n,m} = \frac{\bar{X}_{n,m} - \bar{X}_{0,n}}{s_{n,m}/\sqrt{m}}$.

Then, $\frac{1}{\sqrt{\frac{m}{n}+1}} \cdot B_{n,m} \sim t_{m-1}$.

Proof:

$$(4) \qquad B_{n,m} = \frac{\bar{X}_{n,m} - \bar{X}_{0,n}}{s_{n,m}/\sqrt{m}} = \frac{\bar{X}_{n,m} - \bar{X}_{0,n}}{\sqrt{\frac{m\sigma^2}{n} + \frac{m\sigma^2}{m}} \cdot \frac{s_{n,m}}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}}} = Z \cdot \frac{1}{\frac{s_{n,m}}{\sqrt{\frac{m\sigma^2}{n} + \frac{m\sigma^2}{m}}}} = Z \cdot \frac{1}{\frac{s_{n,m}}{\sigma}} \cdot \sqrt{\frac{m}{n} + \frac{m}{m}}$$

where $Z = \frac{\bar{X}_{n,m} - \bar{X}_{0,n}}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} \sim N(0,1)$ since $\bar{X}_{0,n} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ and $\bar{X}_{n,m} \sim N\left(\mu, \frac{\sigma^2}{m}\right)$.

$$(5) \qquad B_{n,m} = \frac{Z}{\sqrt{\frac{\sum_{k=n+1}^{m}\left(X_k - \bar{X}_{n,m}\right)^2}{(m-1)\sigma^2}}} \cdot \sqrt{\frac{m}{n} + \frac{m}{m}} = \frac{Z}{\sqrt{\frac{\chi_{m-1}^2}{(m-1)}}} \cdot \sqrt{\frac{m}{n} + \frac{m}{m}}$$

Therefore, $\frac{1}{\sqrt{\frac{m}{n}+1}} \cdot B_{n,m} \sim t_{m-1}$.

Lemma 3: Let $\{X_i\}_{i=1}^{n+m}$ be i.i.d. random variables from a normal distribution with mean μ and variance σ^2. Define $\bar{X}_{n,m} = \frac{1}{m}\sum_{j=n+1}^{m} X_j$, $\bar{X}_{0,n} = \frac{1}{n}\sum_{j=1}^{n} X_j$ $s_{0,n}^2 = \frac{\sum_{k=1}^{n}\left(X_k - \bar{X}_{0,n}\right)^2}{n-1}$, and

$B_{0,n} = \frac{\bar{X}_{n,m} - \bar{X}_{0,n}}{s_{0,n}/\sqrt{n}}$. Then $\frac{1}{\sqrt{1+\frac{n}{m}}} \cdot B_{0,n} \sim t_{n-1}$.

Proof:

$$(6) \qquad B_{0,n} = \frac{\bar{X}_{n,m} - \bar{X}_{0,n}}{s_{0,n}/\sqrt{n}} = = \frac{\bar{X}_{n,m} - \bar{X}_{0,n}}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}} \cdot \frac{s_{0,n}}{\sqrt{\frac{n\sigma^2}{n} + \frac{n\sigma^2}{m}}}} = Z \cdot \frac{1}{\frac{s_{0,n}}{\sqrt{\frac{n\sigma^2}{n} + \frac{n\sigma^2}{m}}}} = Z \cdot \frac{1}{\frac{s_{0,n}}{\sigma}} \cdot \sqrt{\frac{n}{n} + \frac{n}{m}}$$

where $Z = \frac{\bar{X}_{n,m} - \bar{X}_{0,n}}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} \sim N(0,1)$ since $\bar{X}_{0,n} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ and $\bar{X}_{n,m} \sim N\left(\mu, \frac{\sigma^2}{m}\right)$.

$$(7) \qquad B_{0,n} = \frac{Z}{\sqrt{\frac{\sum_{k=n+1}^{m}\left(X_k - \bar{X}_{0,n}\right)^2}{(n-1)\sigma^2}}} \cdot \sqrt{\frac{n}{n} + \frac{n}{m}} = \frac{Z}{\sqrt{\frac{\chi_{n-1}^2}{(n-1)}}} \cdot \sqrt{\frac{n}{n} + \frac{n}{m}}$$

Therefore, $\frac{1}{\sqrt{1+\frac{n}{m}}} \cdot B_{0,n} \sim t_{n-1}$.

Lemma 4: Let $\{X_i\}_{i=1}^{n}$ and $\{Y_j\}_{j=1}^{m}$ be independent and identically distributed random variables with distribution *F*, with mean μ and variance σ^2. Define $\bar{Y}_{0,m} = \frac{1}{m}\sum_{j=n+1}^{m} Y_j$, and $\bar{X}_{0,n} = \frac{1}{n}\sum_{i=1}^{n} X_i$, and $B_{n,m} = \frac{\bar{Y}_{0,m}-\bar{X}_{0,n}}{\sigma/\sqrt{2m}}$.

Then as $n \to \infty$ and $m \to \infty$ and $\frac{m}{n} \to 1$, $B_{n,m} = \frac{\bar{Y}_{0,m}-\bar{X}_{0,n}}{\sigma/\sqrt{2m}} \xrightarrow{d} N(0,1)$.

Proof:

(8) $B_{n,m} = \frac{\bar{Y}_{0,m}-\bar{X}_{0,n}}{\sigma/\sqrt{2m}} = \frac{\bar{Y}_{0,m}-\mu}{\sigma/\sqrt{2m}} - \frac{\bar{X}_{0,n}-\mu}{\sigma/\sqrt{2m}}$

(9) By the CLT, as $m \to \infty$, $\frac{\bar{Y}_{0,m}-\mu}{\sigma/\sqrt{m}} \xrightarrow{d} N(0,1)$

(10) Also by the CLT, as $n \to \infty$, $\frac{\bar{X}_{0,n}-\mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$.

(11) Since $\frac{m}{n} \to 1$, as $n \to \infty$, then $m \to \infty$.

(12) Thus by Slutsky's Theorem, $\left(\frac{\sqrt{m}}{\sqrt{n}}\right)\left(\frac{\bar{X}_{0,n}-\mu}{\sigma/\sqrt{n}}\right) = \frac{\bar{X}_{0,n}-\mu}{\sigma/\sqrt{m}} \xrightarrow{d} N(0,1)$.

(13) Using (9) and (12) and the i.i.d. assumption, $\frac{\bar{Y}_{0,m}-\mu}{\sigma/\sqrt{m}} - \frac{\bar{X}_{0,n}-\mu}{\sigma/\sqrt{m}} \xrightarrow{d} N(0,2)$.

(14) Thus as $n \to \infty$ and $m \to \infty$ and $\frac{m}{n} \to 1$, $B_{n,m} = \frac{\bar{Y}_{0,m}-\mu}{\sigma/\sqrt{2m}} - \frac{\bar{X}_{0,n}-\mu}{\sigma/\sqrt{2m}} \xrightarrow{d} N(0,1)$.

(Casella 2002) Lemma 5: If $X_1, \ldots, X_n$ are independent Normal random variables with means $\mu_1, \ldots, \mu_n$ and variances $\sigma_1, \ldots, \sigma_n$ respectively, then

$$Y = \sum_{i=1}^{n} c_i X_i \sim N\left(\sum_{i=1}^{n} c_i \mu_i, \sum_{i=1}^{n} {c_i}^2 {\sigma_i}^2\right).$$

Let $X_{v_1}, \ldots, X_{v_n}$ be independent random variables with standard Student's t distributions with degrees of freedom $v_i > 0$ for all i. Using numerical methods, the sum, $X_{v_1} + X_{v_2} + \cdots + X_{v_n}$can be described as follows.

As discussed in Ahsanullah(2014), a closed form proof of the general sum of t-distributions has yet to be proven with the exception of special cases. Fisher(1935) explored the special case of a weighted sum of two random variables with a Student's t-distribution called the Behrens-Fisher statistic. Walker and Saw(1978) expressed that the sum of Student's t-distributions will converge to a normal distribution as the degrees of freedom approach infinity and that it does not have a closed form for degrees of freedom, v_i, where $0 < v_i < \infty$. Additionally, Walker and Saw (1978) showed that if the $v_i = 1$ for $i = 1,2, \ldots, n$, the sum will have a Cauchy distribution.

By Walker and Saw above, let X_i be i.i.d. from a t-distribution with ν degrees of freedom where $\nu > 0$ and $Y_k = \sum_{i=1}^{k} X_{i,}$. Then as $\nu \to \infty$, then $Y_k \overset{d}{\to} N(0, k)$.

Lemma 6: Consider $Q_k = \frac{\sum_{i=1}^{k} X_i}{\sqrt{k}}$, where $X_i \sim t_v$. As $v \to \infty$, then, $Q_k \overset{d}{\to} N(0,1)$.

The proof is trivial given the above, we are only multiplying a Normal distributed random variable with a mean of 0 and a variance = k by a constant $\frac{1}{\sqrt{k}}$.

Consider $Q_k = \frac{\sum_{i=1}^{k} X_{v_n}}{\sqrt{k}}$, where $X_i \sim t_v$. For finite v, we will investigate if Q_k has approximately a t-distribution and if so, can we approximate the degrees of freedom.

Different methods were considered and investigated in order to approximate the degrees of freedom for a t-distribution for Q_k. Two of these methods were further investigated using simulations.

The second moment of the t-distribution, $\delta^2 = \frac{v}{v-2}$, is defined by the degrees of freedom of the distribution. Using a simulation to approximate the distribution of Q_k, the sample statistic for the second moment can then be calculated from the distribution. Solving the equation above for v we get $v = \frac{2\delta^2}{\delta^2 - 1}$. Then substituting the sample statistic for the parameter we get $v = \frac{2s^2}{s^2 - 1}$.

A second approach to investigate if the Satterthwaite approximation for degrees of freedom for a t-test could be extended to this problem was also explored. Following Satterthwaite, 1946: If MS_1 and MS_2 are independent mean squares with r_1 and r_2 degrees of freedom, the approximate degrees of freedom for $a_1(MS_1) + a_2(MS_2)$ are given in equation (6) as:

$$(15) \quad r_s = \frac{[a_1E(MS_1)+a_2E(MS_2)]^2}{\frac{[a_1E(MS_1)]^2}{r_1}+\frac{[a_2E(MS_2)]^2}{r_2}}$$

Letting $a_1 = a_2 = 1$ gives,

$$(16) \quad r_s = \frac{[E(MS_1)+E(MS_2)]^2}{\frac{[E(MS_1)]^2}{r_1}+\frac{[E(MS_2)]^2}{r_2}}$$

For an unbiased estimate, $E(MS)$ is the second moment or the variance of the random variable. Thus, considering $X_{11}, X_{12}, \ldots$ and $X_{21}, X_{22}, \ldots$ coming from a standard normal distributions and substituting σ^2/n for $E(MS)$ yields:

$$(17) \quad r_s = \frac{\left[\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right]^2}{\frac{\left[\frac{\sigma_1^2}{n_1}\right]^2}{r_1}+\frac{\left[\frac{\sigma_2^2}{n_2}\right]^2}{r_2}}$$

Considering the two sample t-test, r_1 and r_2 are the degrees of freedom from the two independent samples and thus, $r_1 = n_1 - 1$ and $r_2 = n_2 - 1$. Giving,

$$(18) \quad r_s = \frac{\left[\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right]^2}{\frac{\left[\frac{\sigma_1^2}{n_1}\right]^2}{n_1-1}+\frac{\left[\frac{\sigma_2^2}{n_2}\right]^2}{n_2-1}}$$

In practice, the parameters σ_1^2 and σ_2^2 are unknown and are substituted with their observed sample statistics giving:

$$(19)\quad \hat{r}_s = \frac{\left[\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right]^2}{\frac{\left[\frac{s_1^2}{n_1}\right]^2}{n_1-1}+\frac{\left[\frac{s_2^2}{n_2}\right]^2}{n_2-1}}$$

Which is the Satterthwaite approximation to the degrees of freedom for a two-sampled t-test.

Next, again starting from equation (6) in (Satterthwaite, 1946), we will extend this for more than two samples. Given k samples $X_{11}, X_{12}, \dots, X_{1n_1}, X_{21}, X_{22}, \dots, X_{2n_2}, X_{k1}, X_{k2}, \dots, X_{kn_k}$ coming from normal distributions with means $\mu_1, \mu_2, \dots, \mu_k$ and standard deviations $\sigma_1, \sigma_2, \dots, \sigma_k$:

$$(20)\quad r_s = \frac{[a_1E(MS_1)+a_2E(MS_2)+\cdots]^2}{\frac{[a_1E(MS_1)]^2}{r_1}+\frac{[a_2E(MS_2)]^2}{r_2}+\cdots}$$

$$(21)\quad = \frac{\left[\sum_{m=1}^{k} a_mE(MS_m)\right]^2}{\sum_{m=1}^{k}\frac{[a_mE(MS_m)]^2}{r_m}}$$

Setting $a_1, a_2, \dots, a_k = 1$, gives:

$$(22)\quad r_s = \frac{\left[\sum_{m=1}^{k} E(MS_m)\right]^2}{\sum_{m=1}^{k}\frac{[E(MS_m)]^2}{r_m}}$$

For unbiased estimators $E(MS_m) = \sigma_m^2/n_m$, giving:

$$(23) \quad r_s = \frac{\left[\sum_{m=1}^{k} \frac{\sigma_m^2}{n_m}\right]^2}{\sum_{m=1}^{k} \frac{\left[\frac{\sigma_m^2}{n_m}\right]^2}{r_m}}$$

Again, the r_m equals the degree of freedom for each of the k samples, giving:

$$(24) \quad r_s = \frac{\left[\sum_{m=1}^{k} \frac{\sigma_m^2}{n_m}\right]^2}{\sum_{m=1}^{k} \frac{\left[\frac{\sigma_m^2}{n_m}\right]^2}{n_m - 1}}$$

Finally, substituting the observed sample statistics for the unknown parameters gives our Satterthwaite approximation for the degrees of freedom for the linear combination of k variables.

$$(25) \quad \hat{r}_s = \frac{\left[\sum_{m=1}^{k} \frac{s_m^2}{n_m}\right]^2}{\sum_{m=1}^{k} \frac{\left[\frac{s_m^2}{n_m}\right]^2}{n_m - 1}}$$

Thus, the Satterthwaite approximation for the degrees of freedom for a linear combination of five random variables from a normal distribution with means $\mu_1, \mu_2, \mu_3, \mu_4, \mu_5$ and standard deviations $\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2, \sigma_5^2$ is:

$$\hat{r}_s = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} + \frac{s_3^2}{n_3} + \frac{s_4^2}{n_4} + \frac{s_5^2}{n_5}\right]^2}{\frac{\left[\frac{s_1^2}{n_1}\right]^2}{n_1 - 1} + \frac{\left[\frac{s_2^2}{n_2}\right]^2}{n_2 - 1} + \frac{\left[\frac{s_3^2}{n_3}\right]^2}{n_3 - 1} + \frac{\left[\frac{s_4^2}{n_4}\right]^2}{n_4 - 1} + \frac{\left[\frac{s_5^2}{n_5}\right]^2}{n_5 - 1}}$$

The table below shows the results for the exact degrees of freedom given by r_s above and the Satterthwaite approximation given by $\hat{r}_s$ above for a single sample of the sum of two through five random variables with 30 observations each from standard normal distributions.

Table 7.1 Comparison of Exact vs. Satterthwaite Degrees of Freedom

Linear combination	$\hat{r}_{sat}$	r_{sat}
$\bar{X}_1 + \bar{X}_2$	57.3988	58
$\bar{X}_1 + \bar{X}_2 + \bar{X}_3$	84.1006	87
$\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \bar{X}_4$	113.0751	116
$\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \bar{X}_4 + \bar{X}_5$	138.8825	145

The formulas for the 95% confidence intervals utilizing the exact and Satterthwaite approximation for the sum of the five variables are below.

$$\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \bar{X}_4 + \bar{X}_5 \pm t_{r_{sat},95\%} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} + \frac{1}{n_3} + \frac{1}{n_4} + \frac{1}{n_5}\right)}$$

$$\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \bar{X}_4 + \bar{X}_5 \pm t_{\hat{r}_{sat},95\%} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} + \frac{1}{n_3} + \frac{1}{n_4} + \frac{1}{n_5}\right)}$$

In this case, $n_1 = n_2 = n_3 = n_4 = n_5 = 30$ and thus $\frac{1}{n_1} + \frac{1}{n_2} + \frac{1}{n_3} + \frac{1}{n_4} + \frac{1}{n_5} = \frac{1}{6}$ and s_p^2 is the pooled variance estimate. Using the exact and Satterthwaite approximation for the degrees of freedom above, the two confidence intervals are given in the table below.

Table 7.2 Comparison of Exact vs. Satterthwaite Confidence Intervals

Linear combination	95% CI using $\hat{r}_{sat}$	95% CI using r_{sat}
$\bar{X}_1 + \bar{X}_2$	(-0.7525, 0.2638)	(-0.754, 0.2637)
$\bar{X}_1 + \bar{X}_2 + \bar{X}_3$	(-0.7226, 0.5957)	(-0.7222, 0.5953)
$\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \bar{X}_4$	(-0.6663, 0.8501)	(-0.6661, 0.8499)
$\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \bar{X}_4 + \bar{X}_5$	(-0.2277, 1.4070)	(-0.2274, 1.4067)

To further investigate the Satterthwaite approximation and the exact degrees of freedom, the confidence intervals for 5000 samples were generated and the following coverage probabilities were found. The table below shows the proportion of 5000 samples where the confidence intervals for the sum of t distributions contained the true $\mu = \mu_1 + \mu_2 + \mu_3 + \mu_4 + \mu_5 = 0 + 0 + 0 + 0 + 0 = 0$.

Table 7.3 Proportion of Confidence Intervals that Contained μ

Linear combination	$CI_{sat} \supset \mu$	$CI_{exact} \supset \mu$
$\bar{X}_1 + \bar{X}_2$	0.9524	0.952
$\bar{X}_1 + \bar{X}_2 + \bar{X}_3$	0.9514	0.9512
$\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \bar{X}_4$	0.949	0.9488
$\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \bar{X}_4 + \bar{X}_5$	0.9524	0.9522

7.2 Simulations

Numerical methods will be utilized to further explore the distributional properties of the change variable, $A_n = \frac{X_n - \bar{X}_{n-1}}{s_{n-1}}$, the sum of t-distributions and transformations of the sum such as $Q_k = \frac{\sum_{i=1}^{k} X_{i,v}}{\sqrt{k}}$. There are three variables in these simulations that can be scaled and manipulated to understand their impact on the problem. All three have direct connections into the real world sepsis prediction model. The number of different clinical measures, or vital signs, that are measured on a patient, the number of touch points, or sets of observations on each patient, and the total number of patients sampled. All three of these can vary and have an impact on the outcome of the simulation. The number of touch points gives an indication of the information that is known about each patient. If the patient has been in the hospital for less than a day and only a few sets of observations exist, much less information about the patient is known compared to someone who has been in the hospital for over a week. To better understand their effects, the notation must be clearly defined.

The first variable, the number of clinical measures, is the number of t-distributed random variables (X_i). The logistics of the sepsis prediction problem focuses on five random variables, as this is the number of vital signs that are being monitored to use in the prediction: heart rate, respiratory rate, temperature, peripheral capillary oxygen saturation, and systolic blood pressure. However, additional numbers of random variables are considered to better understand how this fluctuation impacts the distribution on Y_k and Q_k as defined above. For the purposes of this paper, the number of clinical measures will be referred as k, thus Q_5 would imply a problem that includes the sum of five clinical measures.

The second variable that will be scaled and manipulated is the number of degrees of freedom (ν) of each of the t-distributed random variables. In the context of the sepsis prediction problem, the ν represents the number of observations of vital sign measurements that have been collected on the patient. This is a highly variable number, however, a couple of assumptions are made in order to maintain the integrity of the problem. First, since vital signs are generally captured as a set and recorded at the same time, it is assumed that the degrees of freedom for each of the t-distributed random variables are equal, effectively saying that a patient will not have 15 heart rate observations and only 10 blood pressure observations. Additionally, degrees of freedom ranging from two through 25 will be the focus, and an additional observation with 40 degrees of freedom will be added to represent an approximation to the normal case. The special case of random variables with one degree of freedom $(\nu = 1)$ has already been expressed as mentioned above. The simulation will focus on $\nu < 25$ and resolve that observations significantly larger than 25 will be approximated with normally distributed random variables.

The third variable that will be influenced in the simulation is the number of observations used to determine the underlying distribution of the random variables Y_k and

Q_k. This is equivalent to the number of patients that have a series of measurements documented. These measurements will be used to approximate the overall distribution of the variables. Since this has a direct impact on the power and sensitivity of the test, the number will vary from smaller samples of 1000 in simulations looking for a t-distribution up to samples of 10,000 when looking for normality.

7.2.1 Simulating A_n from Different Distributions

A simulation was created to investigate the distributional properties of the change variable:

$$A_n = \frac{X_n - \bar{X}_{n-1}}{s_{n-1}}$$

where $\bar{X}_{n-1}$ and s_{n-1} are the mean and standard deviation of the first *n*-1 observations and X_n is the *n*th observation. Simulations were created where the underlying X_n come from a normal distribution, a gamma distribution, a binomial distribution, and a Student's t-distribution. During this simulation the number of X_n, ie. the number of observations per patient, was iterated from *n*=3 to *n*=20. In regards to the clinical problem, this is equivalent to the number of observations of a single vital sign that are obtained on a patient. Thus, if n was 20, A_n was calculated by finding the mean and standard deviation of the first 19 observations, $\bar{X}_{n-1}$ and s_{n-1} respectively, and subtracting $\bar{X}_{n-1}$ from the 20th observation, X_{20}, and then dividing the difference by s_{n-1}. For each underlying distribution, this was repeated 1,000 times, representing having clinical measures from 1,000 patients. Then, histograms were constructed and the distribution was tested for normality utilizing the

Anderson-Darling and Kolmogorov-Smirnov test for normality inside of the Univariate procedure in SAS.

As can be seen in the table below, when X_n originates from a normal distribution, the distribution of A_n begins to fail to reject the hypothesis, $H_0 : A_n$ is normally distributed, when the sample size, $n \geq 14$.

Table 7.4 Normality of A_n when $X_n \sim N(0,1)$

Distribution	n	Anderson-Darling	Kolmogorov-Smirnov
Normal	3	<0.0050	<0.0100
Normal	4	<0.0050	<0.0100
Normal	5	<0.0050	<0.0100
Normal	6	<0.0050	<0.0100
Normal	7	<0.0050	<0.0100
Normal	8	<0.0050	<0.0100
Normal	9	<0.0050	<0.0100
Normal	10	<0.0050	0.0237
Normal	11	<0.0050	0.032
Normal	12	<0.0050	0.0352
Normal	13	<0.0050	0.0144
Normal	14	>0.2500	>0.1500
Normal	15	0.0234	0.0949
Normal	16	0.1305	>0.1500
Normal	17	0.2173	>0.1500
Normal	18	0.1384	>0.1500
Normal	19	0.2494	0.0537
Normal	20	>0.2500	>0.1500
Normal	40	>0.2500	>0.1500

When the X_n come from a gamma distribution, the distribution of A_n is skewed to the right and never trends towards a normal distribution. When the underlying observations, X_n, are distributed with a student's t-distribution, the $A_n$exhibit a distribution which is unimodal, symmetric and centered at zero. However, this distribution has some outliers and fatter tails and as such it does not reach normality. Figure 7.1 shows the distribution of A_n when the underlying data originates from t-distributions with five degrees of freedom. Notice the fatter tails and the large outliers in the distribution.

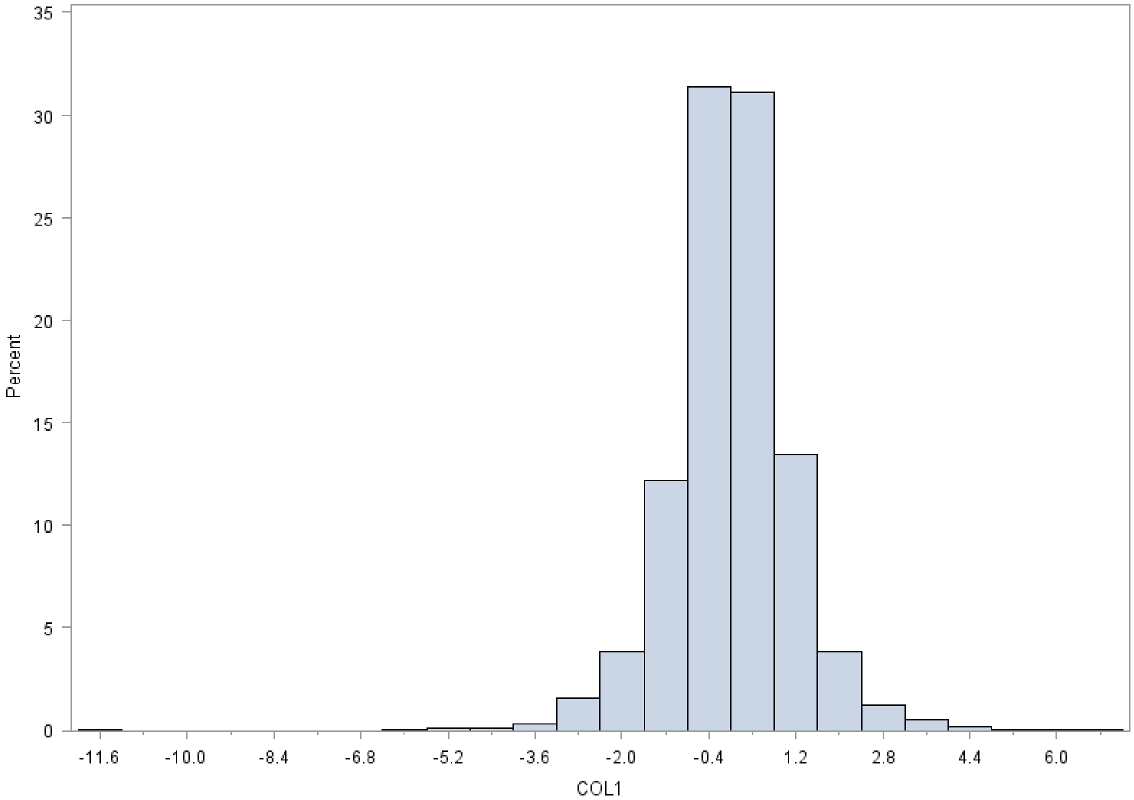


Figure 7.1 Histogram of A_n when $X_n \sim t_5$

Data that originates from a normal distribution produces an A_n that reaches normality with a sufficiently large sample size. However, the distribution of A_n for finite n has previously been shown above to have a t-distribution plus an error term. The distribution of the sum of t-distributions and transformations of that distribution will be further explored.

7.2.2 Approximating the Sum of t-distributions

Let $X_{1,v_1}, \dots, X_{k,v_k}$ be independent random variables with standard Student's t-distributions with degrees of freedom $v_i > 0$. Numerical methods will be used to describe the sum, $X_{1,v_1} + X_{2,v_2} + \cdots + X_{k,v_k}$. For the purposes of our problem, it will be assumed that $v_1 = v_2 = \cdots = v_k = v$. The problem where k = 5 and the sum of $X_{1,v} +$

$X_{2,\nu} + X_{3,\nu} + X_{4,\nu} + X_{5,\nu}$ for different values of ν will be explored first. Five independent and identically distributed variables from a t-distribution were randomly generated 1000 times each with degrees of freedom from one to ten and additionally for twenty, thirty and forty and the variable $S_k = \sum_{i=1}^{5} X_{\nu,i}$ was calculated. The histograms for each S_k were created using the Univariate procedure in SAS. Figure 7.2 shows the histogram for $\nu = 1$. Note the distribution is highly skewed left and ranges from -2750 up to 750. The mean is -4.04 with a standard deviation of 113.40.

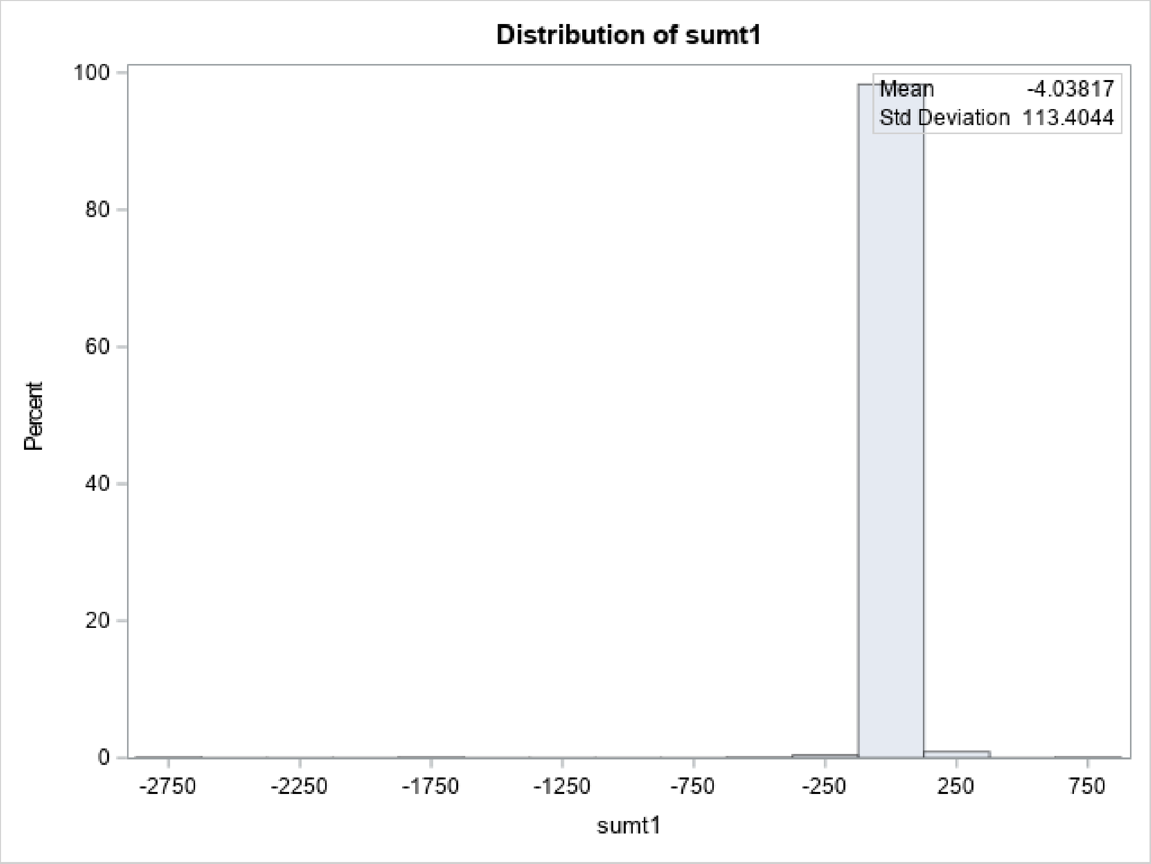


Figure 7.2 Histogram of $\sum_{i=1}^{5} t_{i,1}$

Figure 7.3 shows the histogram of the sum of five t-distributions with two degrees of freedom. Note that the distribution has a much smaller variance than Figure

7.2 but still ranges from -135 up to 75. It has a mean of -0.03 and a standard deviation of 8.03.

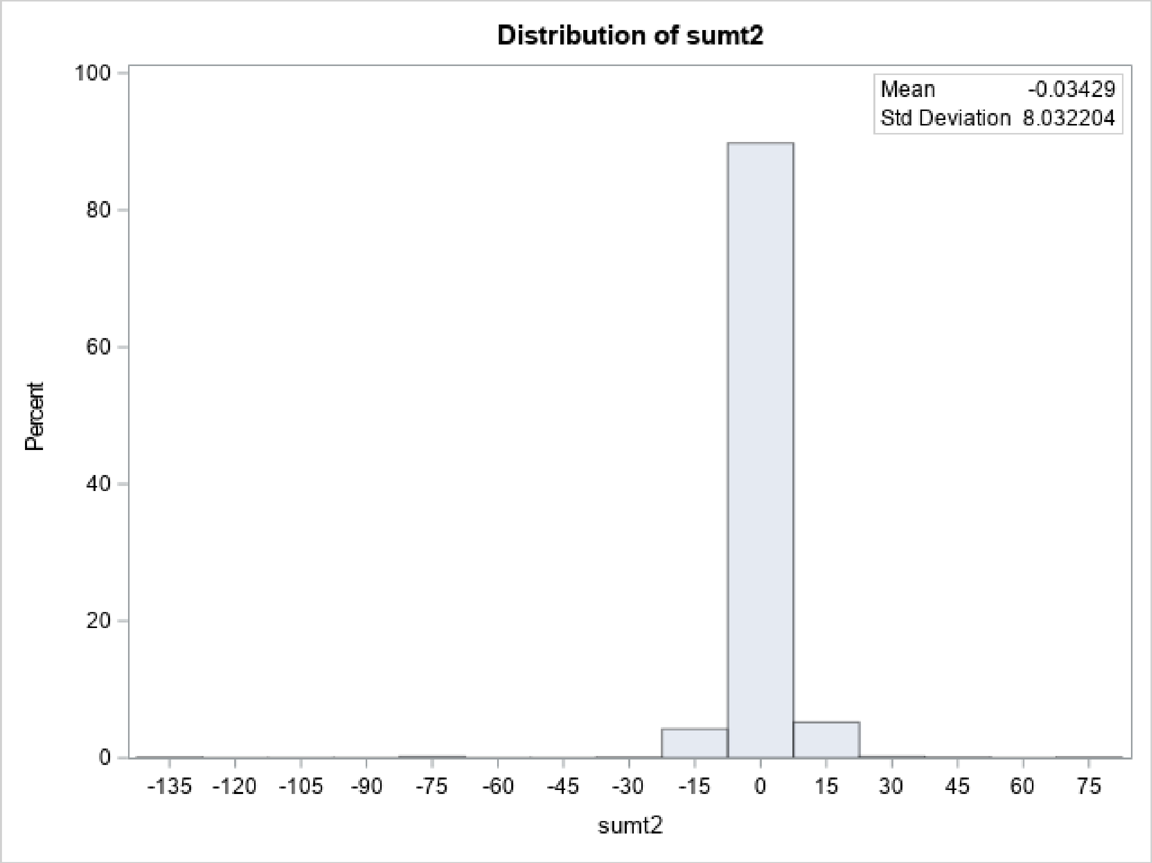


Figure 7.3 Histogram of $\sum_{i=1}^{5} t_{i,2}$

Figure 7.4 shows the distributions of the sum of five t-distributions with degrees of freedom ranging from three to six. All four histograms remain centered around zero, but their standard deviations are decreasing as the degrees of freedom increases.

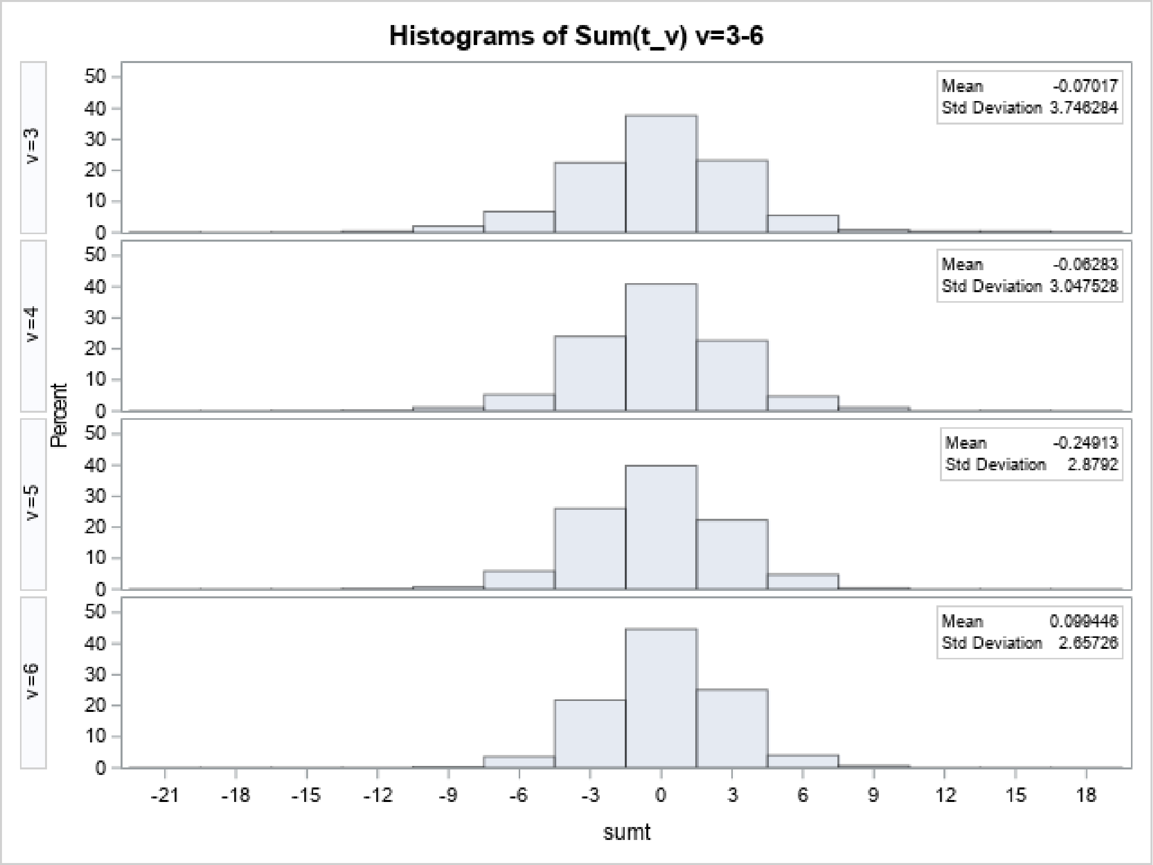


Figure 7.4 Histogram of $\sum_{i=1}^{5} t_{i,\nu}$ for $\nu = 3, 4, 5,$ and 6

Figure 7.5 shows the distributions of the sum of five t-distributions with degrees of freedom ranging from seven to ten. All four histograms remain centered around zero and their ranges continue to decrease.

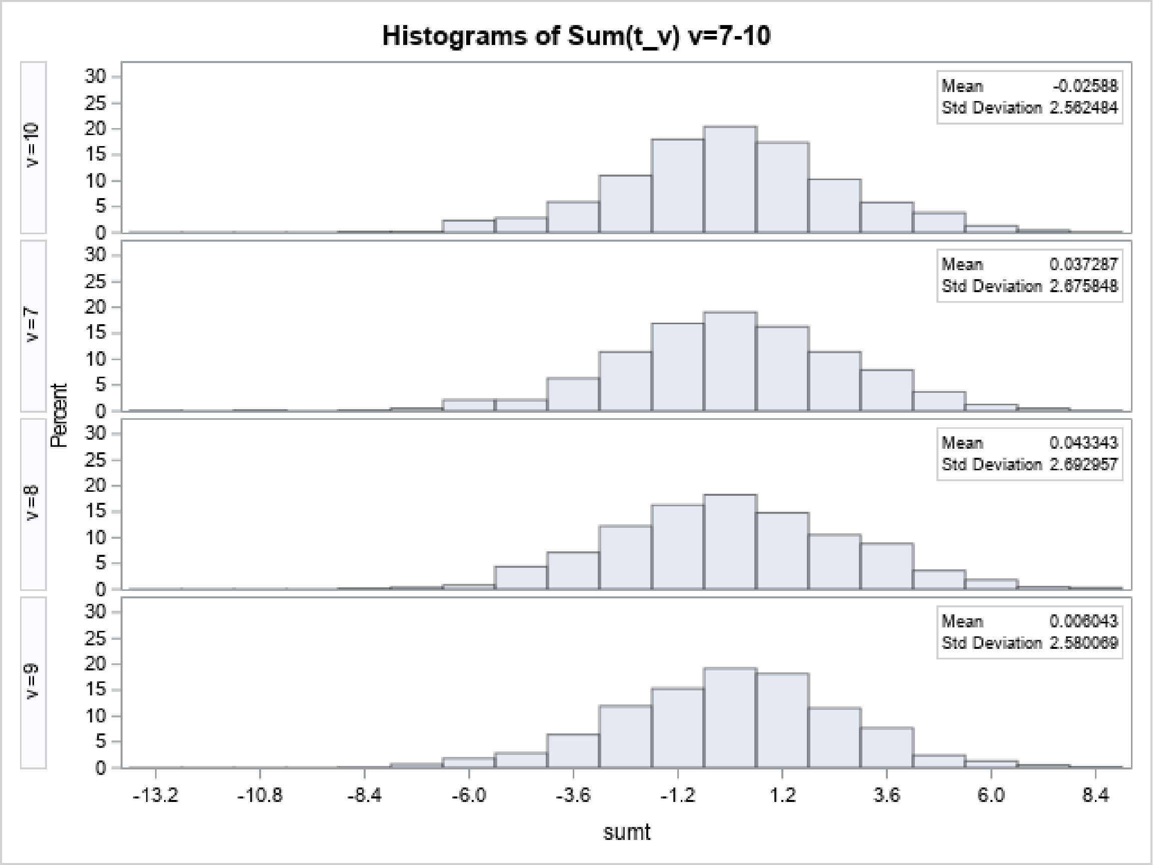


Figure 7.5 Histogram of $\sum_{i=1}^{5} t_{i,\nu}$ for $\nu = 7, 8, 9,$ and 10

For completeness, the graphs of the sum of five t-distributions with degrees of freedom of twenty, thirty and forty are shown in Figure 7.6, below. As the degrees of freedom increases, the histograms follow a normal distribution more closely. For each of the histograms below the p-value for the Anderson-Darling test for normality is greater than 0.25.

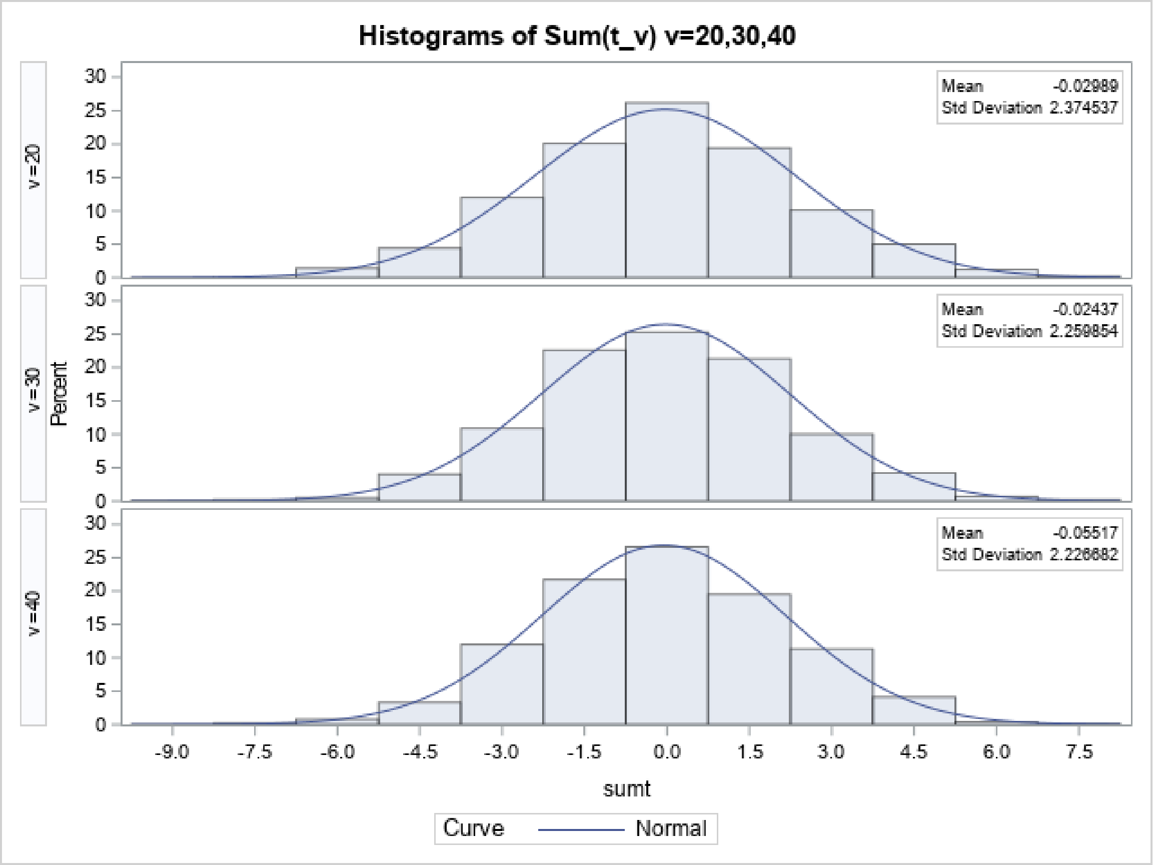


Figure 7.6 Histogram of $\sum_{i=1}^{5} t_{i,\nu}$ for $\nu = 20, 30,$ and 40

The distribution of the sum of the five t-distributions appears to converge to a normal distribution as the degrees of freedom increase. To determine when the sum of t-distributions converges to a normal distribution, the simulation is expanded to capture the Kolmogorov-Smirnov test for normality. Additionally, what are the characteristics of the distribution of a sum of t-distributions for finite samples? Five simulated t-distributions were generated with 10,000 repetitions with degrees of freedom ranging from one to twenty. The distribution of the sum of these random variables was tested for normality using the Kolmogorov-Smirnov (KS) test with the Univariate procedure in SAS. Next, utilizing the Npar1way procedure, a KS test was run to compare the sum to a set of t-distributions with different degrees of freedom. Early in the process it was determined that though the distribution was symmetric and unimodal, the variance of the sum was too large to be a t-distribution. A simple transformation could reduce the variance and possibly change the distribution to a t-distribution. The mean, dividing the sum by the number of random variables, was initially investigated, followed by dividing the sum by the square root of *n*. Finally, a macro was developed to iterate through t-distributions with different degrees of freedom to narrow down the selection which fail to reject the KS.

The initial question to address was: does the sum of t-distributions with varying degrees of freedom converge to a normal distribution as *n* goes to infinity? To determine where this distribution begins to converge to a normal distribution, a simulation was conducted using 10,000 iterations calculating the sum of five generated t-distributions with degrees of freedom varying from one to twenty. The table below shows the results of these simulations with the mean, standard deviation, and the p-value of a Kolmogorov-Smirnov test using the Univariate procedure in SAS, testing the hypothesis that the distribution is a normal distribution.

Table 7.5 Testing the Normality of the Sum of Five t-Distributed Random Variables

d.f.	Mean	Standard Deviation	p-value for H_0: Distribution is Normal
1	1.1658	226.9733	< 0.01
2	-0.0010	8.0925	< 0.01
3	-0.0459	4.0433	< 0.01
4	-0.0225	3.0829	< 0.01
5	0.0196	2.9026	< 0.01
6	0.0175	2.7464	< 0.01
7	-0.0320	2.6410	0.03
8	0.0085	2.5934	> 0.15
9	-0.0100	2.5585	0.14
10	0.0166	2.5186	0.09
11	-0.0001	2.4706	> 0.15
12	-0.0077	2.4554	0.14
13	0.0328	2.4326	> 0.15
14	0.0008	2.4020	> 0.15
15	-0.0191	2.3847	0.03
16	-0.0144	2.3724	> 0.15
17	0.0303	2.3603	> 0.15
18	0.0218	2.3691	> 0.15
19	0.0185	2.3724	> 0.15
20	-0.0436	2.3571	> 0.15

As can be seen above, the sum of five t-distributions begins to reach normality when the original t-distributions have degrees of freedom between eight and fourteen. The graph below shows the histogram of 10,000 observations from the sum of t-distributions with ten degrees of freedom and a normal curve superimposed on the graph.

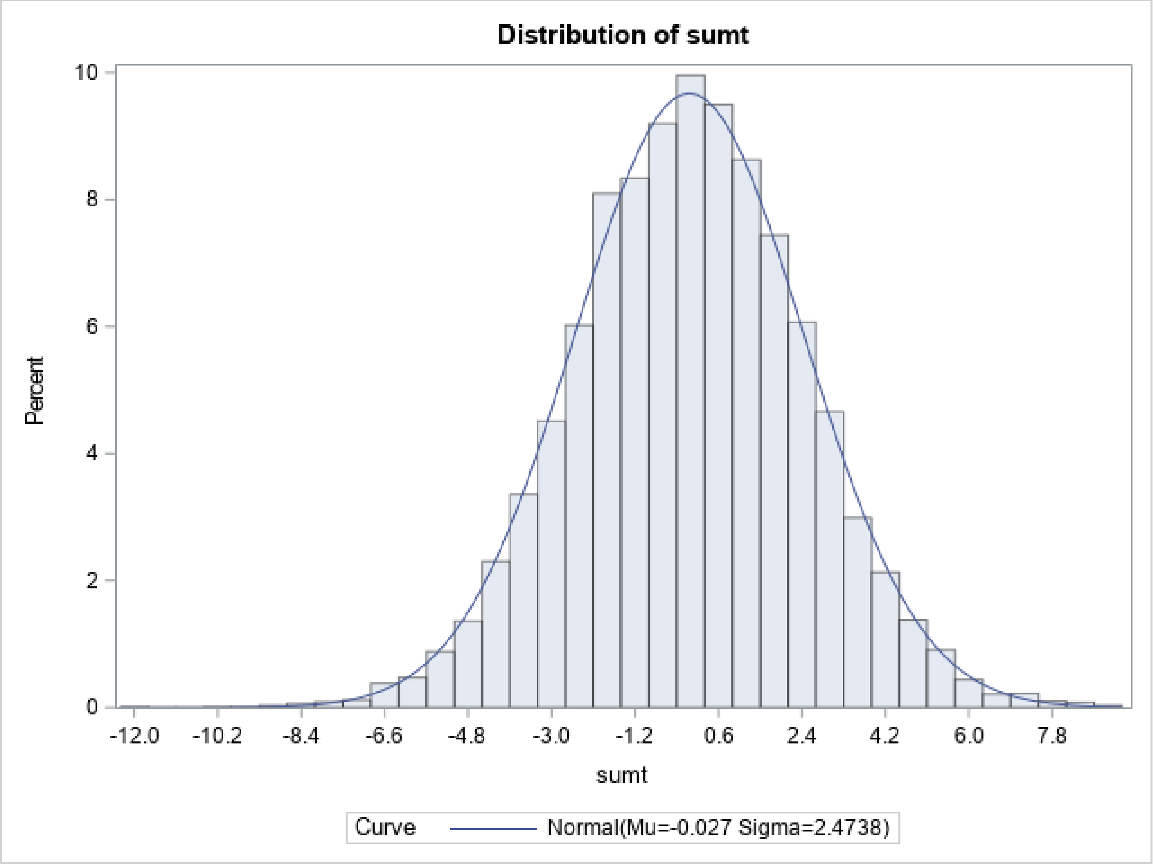


Figure 7.7 Histogram of $\sum_{i=1}^{5} t_{i,10}$ with 10,000 Observations

It can be seen from the table and graph above that although the distribution of the sum of t-distributions attains normality, the variance of the distribution is too large to be a t-distribution at smaller degrees of freedom. Looking at Table 7.5 above, the mean of the distributions are centered on zero and have a standard deviation around 2.4.

It was determined to see if the standard deviation of the distribution could be better described and if it could be transformed into a standard normal. This would allow for a better description of the standard deviation and for further investigation into whether it follows a t-distribution at lower degrees of freedom of the original random variables in the sum.

The distribution of the sum of the t-distributed random variables is already centered at zero, so the focus was reducing the standard deviation to one. The first hypothesis tested was if the average of the five t-distributed random variables converged to a standard normal distribution. Figure 7.8 shows an illustration of the sum of five random variables compared to the average, or $R_k = \frac{\sum_{i=1}^{k} t_i}{k}$, of k = 5 t-distributed random variables with seven degrees of freedom.

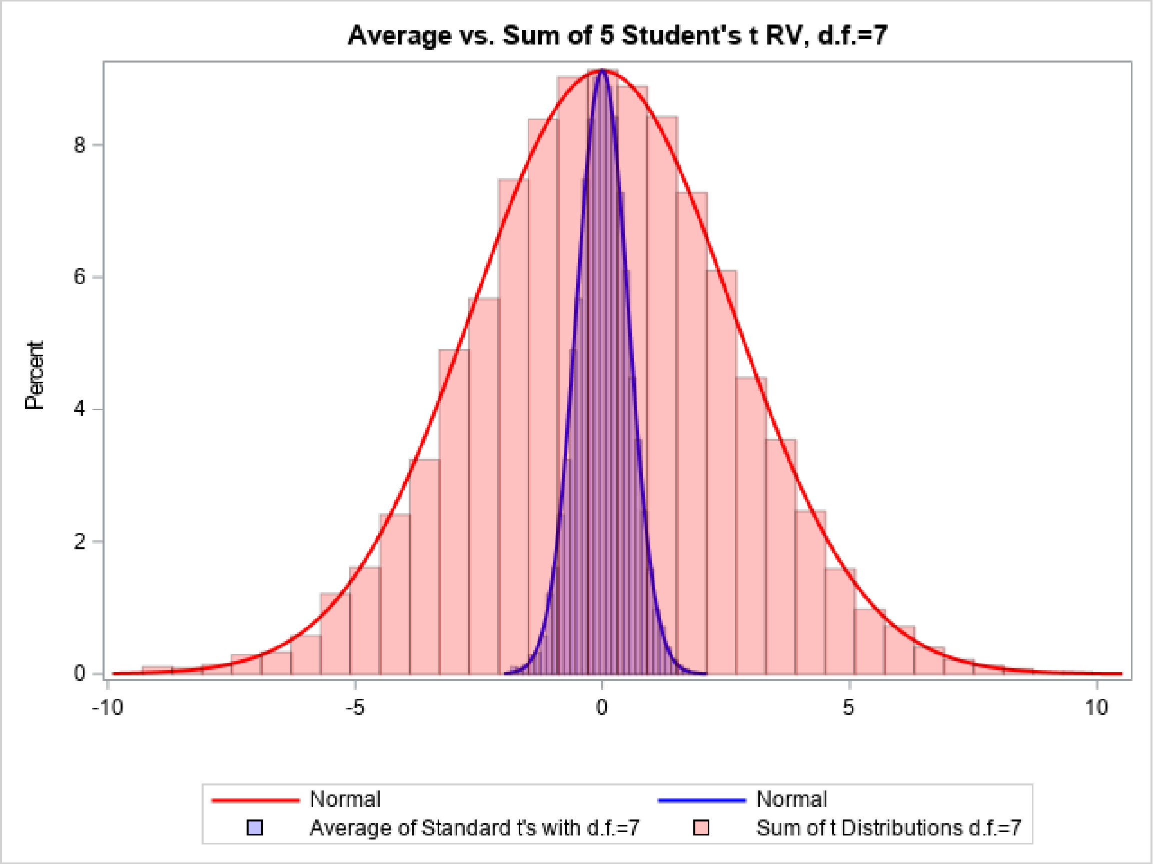


Figure 7.8 Comparing the Sum vs. Mean of Five t-Distributed Random Variables with Seven Degrees of Freedom

Both distributions follow a normal distribution and the standard deviation of the mean of the variables is significantly smaller than the standard deviation of the sum of variables. Using the simulated data from above that generated the five t-distributed random variables with varying degrees of freedom, the mean and standard deviation was calculated across the 10,000 observations. The simulation was repeated using degrees of freedom from one to twenty seeding the initial t-distributed random variables. Table 7.6

shows the results of this simulation along with the p-value from the Kolmogorov-Smirnov test of normality within the Univariate procedure in SAS.

Table 7.6 Testing Normality of the Mean of Five t-Distributed Random Variables

d.f.	Mean	Standard Deviation	p-value for H_0: Distribution is Normal
1	0.2332	45.3947	< 0.01
2	-0.0002	1.6185	< 0.01
3	-0.0092	0.8087	< 0.01
4	-0.0045	0.6166	< 0.01
5	0.0039	0.5805	< 0.01
6	0.0035	0.5493	< 0.01
7	-0.0064	0.5282	0.03
8	0.0017	0.5187	> 0.15
9	-0.0020	0.5117	0.14
10	0.0033	0.5037	0.09
11	0.0000	0.4941	> 0.15
12	-0.0015	0.4911	0.14
13	0.0066	0.4865	> 0.15
14	0.0002	0.4804	> 0.15
15	-0.0038	0.4769	0.03
16	-0.0029	0.4745	> 0.15
17	0.0061	0.4721	> 0.15
18	0.0044	0.4738	> 0.15
19	0.0037	0.4745	> 0.15
20	-0.0087	0.4714	> 0.15

The p-values for the normality of the distributions are identical since the resulting variable is only divided by a constant. However, notice the standard deviation is now approximately 0.48, or 2.4/5. In order to transform this distribution into a standard normal distribution, the sum of the t-distributions needs to be divided by a smaller number to bring the standard deviation up to one.

The sum of the random variables divided by the square root of *k* was next tested to find if the standard deviation of this distribution was approaching one. Using the previously simulated data, this variable was calculated for each of the 10,000 observations across the different degrees of freedom from one to 20. Let $X_{\nu,i}$ be k independent and identically distributed t-distributions with ν degrees of freedom, then define:

$$S_k = \sum_{i=1}^{k} X_{\nu,i}\, , R_k = \sum_{i=1}^{k} \frac{X_{\nu,i}}{k}, \text{ and } Q_k = \sum_{i=1}^{k} \frac{X_{\nu,i}}{\sqrt{k}}$$

The table below shows the mean, standard deviation and the p-value for the Kolmogorov-Smirnov test for normality for S_k, R_k and Q_k from the simulation while iterating the degrees of freedom from one to 20. The p-value is only reported once as the p-values for S_k, R_k and Q_k are equal as the transformation is only multiplying the original sum by a constant.

Table 7.7 Comparing Mean and Standard Deviation for S_5, R_5, and Q_5 Computed with t-Distributions with Seven Degrees of Freedom

d.f.	R_k Mean	S_k Mean	Q_k Mean	R_k St. Dev	S_k St. Dev	Q_k St. Dev	p-value
1	0.233	1.166	0.521	45.395	226.973	101.506	0.010
2	0.000	-0.001	0.000	1.619	8.093	3.619	0.010
3	-0.009	-0.046	-0.021	0.809	4.043	1.808	0.010
4	-0.004	-0.022	-0.010	0.617	3.083	1.379	0.010
5	0.004	0.020	0.009	0.581	2.903	1.298	0.010
6	0.004	0.018	0.008	0.549	2.746	1.228	0.010
7	-0.006	-0.032	-0.014	0.528	2.641	1.181	0.028
8	0.002	0.008	0.004	0.519	2.593	1.160	0.150
9	-0.002	-0.010	-0.004	0.512	2.558	1.144	0.137
10	0.003	0.017	0.007	0.504	2.519	1.126	0.095
11	0.000	0.000	0.000	0.494	2.471	1.105	0.150
12	-0.002	-0.008	-0.003	0.491	2.455	1.098	0.145
13	0.007	0.033	0.015	0.487	2.433	1.088	0.150
14	0.000	0.001	0.000	0.480	2.402	1.074	0.150
15	-0.004	-0.019	-0.009	0.477	2.385	1.066	0.025
16	-0.003	-0.014	-0.006	0.474	2.372	1.061	0.150
17	0.006	0.030	0.014	0.472	2.360	1.056	0.150
18	0.004	0.022	0.010	0.474	2.369	1.059	0.150
19	0.004	0.019	0.008	0.474	2.372	1.061	0.150
20	-0.009	-0.044	-0.019	0.471	2.357	1.054	0.150

As has earlier been discussed, the sum of t-distributions with one degree of freedom is a special case that has a Cauchy distribution, (Walker 1978). Additionally, the standard deviation for Q_k is approaching one which would make it possible that it could have a t-distribution for lower degrees of freedom. Figures 7.9 and 7.10 below show the distributions for $S_k, R_k\ and\ Q_k$ calculated using t-distributions with seven degrees of freedom superimposed on each other and 20 degrees of freedom, respectively.

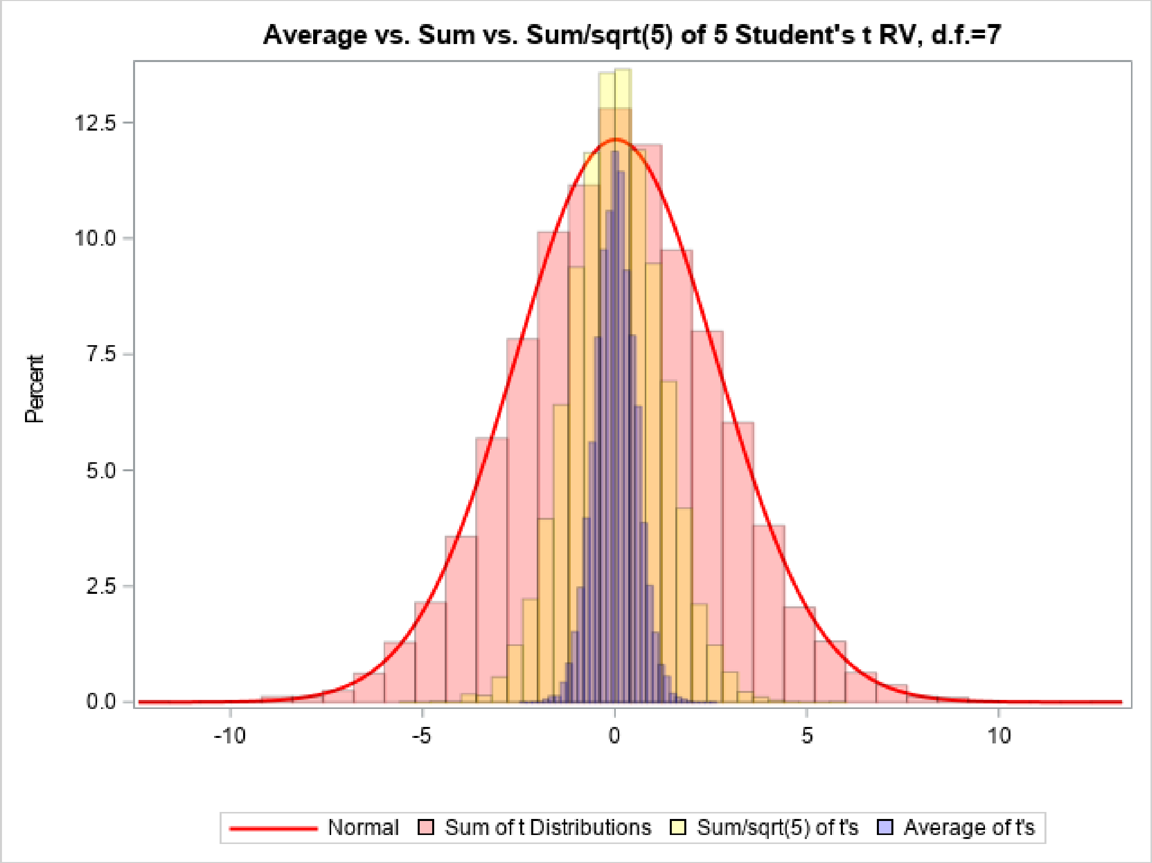


Figure 7.9 Comparing S_5, R_5, and Q_5 Comprised of t-Distributed Random Variables with Seven Degrees of Freedom

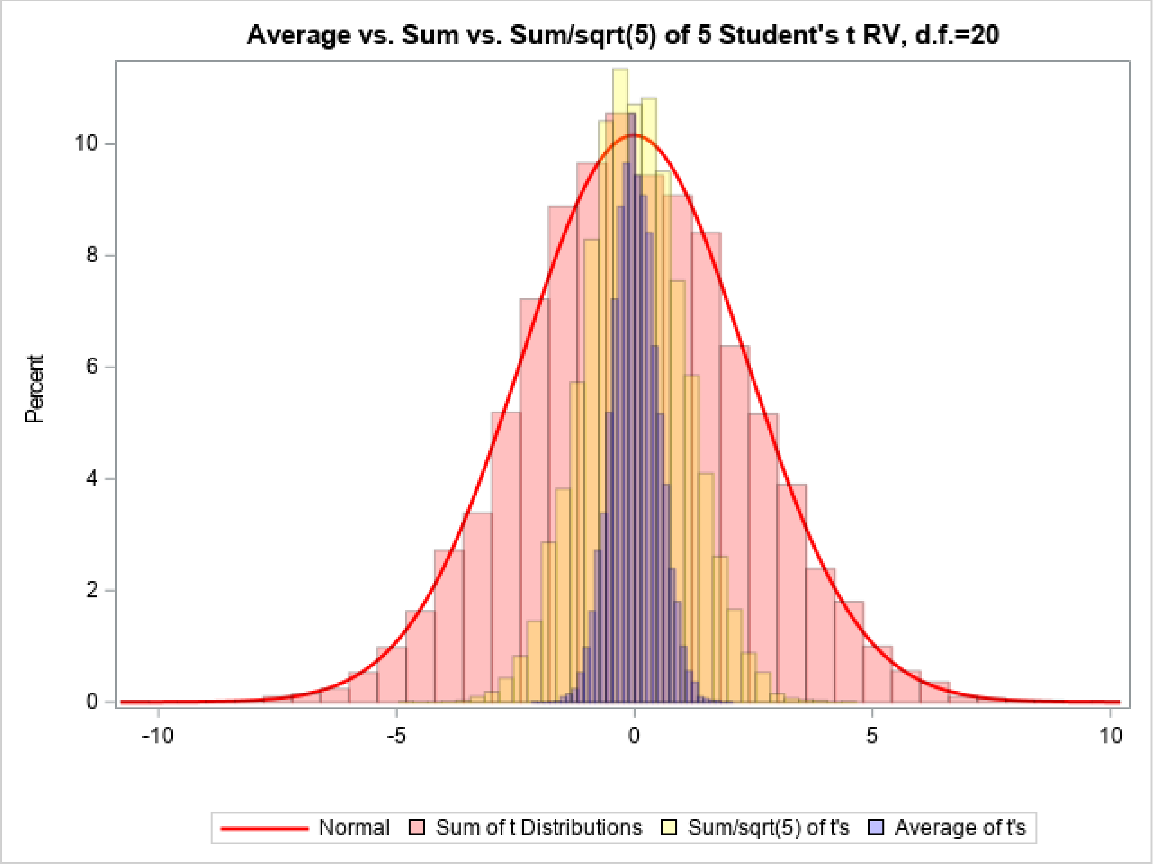


Figure 7.10 Comparing S_5, R_5, and Q_5 Comprised of t-Distributed Random Variables with 20 Degrees of Freedom

Notice that the standard deviation of Q_k is now trending towards one. This convergence gives us a couple opportunities. First, the sum of five t-distributed random variables can more accurately be describe as converging to a normal distribution with a mean of zero and a variance of five, or conversely, the distribution has a standard deviation of $\sigma = \sqrt{5}$. Additionally, we can begin to investigate whether the distribution of the sum divided by the square root of n follows a Student's t-distribution at lower degrees of freedom for the seeding t-distributed random variables.

7.2.3 Simulation to Test if Q_5 has a t-Distribution

In pursuit of this question, a simulation was created following the pattern of the previous simulation. This simulation is designed to study both varying the degrees of freedom of the underlying t-distributions and test the distributions of Q_k to see if they are possibly t-distributions. The simulation is designed such that it has a smaller loop inside of a larger loop. The outer part is a loop that iterates the degrees of freedom of the underlying t-distributions from two to 16. For each iteration, 1000 samples of five t-distributions is generated and the variable Q_k is calculated. Inside of this loop is a second loop that takes each sample of 1000 observations and tests to see if the distribution of Q_k is significantly different from a t-distribution with degrees of freedom ranging from one to 20 plus one additional observation with forty degrees of freedom. This test is accomplished by generating a t-distribution with degrees of freedom *D*, and utilizing a Kolmogorov-Smirnov (KS) *D* statistic for two-sample data based on the empirical distribution function found in the Npar1way procedure in SAS to test the hypothesis: $H_0 \colon Q_k \sim t_D$. In each instance Q_k is calculated using five randomly generated t-distributions.

Figure 7.11 shows the results of Q_k which is comprised of t-distributions with ten degrees of freedom superimposed on the graph of a t-distribution with ten degrees of freedom. Table 7.8 shows the p-value for this KS test is 0.90, meaning the distributions are not significantly different from each other.

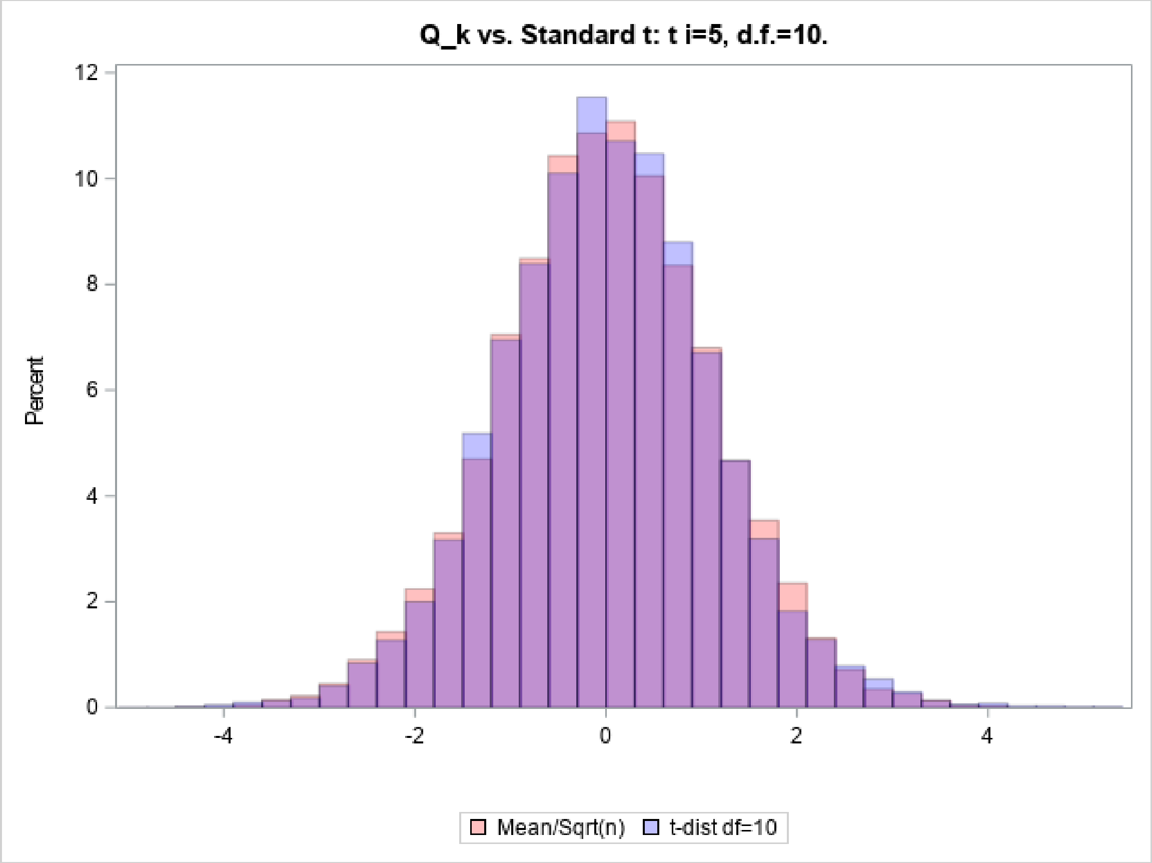


Figure 7.11Q_5 Comprised of $X_{\nu,i} \sim t_{10}$ Compared to a t-Distribution with Ten Degrees of Freedom

The table below shows the results of the simulation. The degrees of freedom for the five t-distributions that comprise Q_k are going across the top of the table and the degrees of freedom for the t-distribution that being tested against Q_k are going down the left of the table. The p-values of the KS test for each combination is displayed in the table. P-values that fail to reject the null hypothesis are highlighted in yellow.

Table 7.8 P-values of KS Test for $H_0: Q_5 \sim t_\nu \ (n = 1000)$. The columns represent the seed degrees of freedom in Q_5 and the rows are the degrees of freedom in t_ν.

d.f. t_v	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.01	0.26	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00
3	0.00	0.02	0.10	0.46	0.50	0.07	0.07	0.18	0.07	0.22	0.01	0.01	0.16
4	0.00	0.01	0.02	0.31	0.16	0.25	0.02	0.32	0.53	0.20	0.15	0.07	0.16
5	0.00	0.00	0.11	0.30	0.23	0.69	0.11	0.76	0.91	0.35	0.33	0.17	0.26
6	0.00	0.00	0.01	0.23	0.53	0.67	0.34	0.30	0.21	0.61	0.60	0.03	0.63
7	0.00	0.00	0.06	0.01	0.09	0.02	0.49	0.15	0.04	0.84	0.99	0.48	0.41
8	0.00	0.00	0.00	0.11	0.27	0.28	0.03	0.25	0.57	0.01	0.33	0.85	0.42
9	0.00	0.00	0.00	0.04	0.09	0.04	0.01	0.08	0.28	1.00	0.06	0.18	0.37
10	0.00	0.00	0.00	0.02	0.11	0.01	0.08	0.08	0.90	0.08	0.17	0.11	0.75
11	0.00	0.00	0.00	0.00	0.24	0.18	0.04	0.05	0.31	0.56	0.30	0.11	0.67
12	0.00	0.00	0.00	0.01	0.34	0.38	0.01	0.03	0.11	0.55	0.32	0.28	0.60
13	0.00	0.00	0.00	0.04	0.04	0.13	0.01	0.63	0.43	0.42	0.59	0.34	0.72
14	0.00	0.00	0.01	0.00	0.16	0.30	0.03	0.34	0.05	0.54	0.35	0.55	0.08
15	0.00	0.00	0.00	0.05	0.11	0.14	0.01	0.09	0.62	0.28	0.07	0.04	0.73
16	0.00	0.00	0.00	0.00	0.02	0.05	0.12	0.29	0.01	0.01	0.34	0.02	0.89
17	0.00	0.00	0.00	0.02	0.10	0.07	0.13	0.02	0.16	0.04	0.11	0.43	0.51
18	0.00	0.00	0.00	0.00	0.00	0.24	0.02	0.01	0.05	0.32	0.31	0.61	0.83
19	0.00	0.00	0.00	0.00	0.01	0.03	0.14	0.12	0.03	0.25	0.47	0.22	0.67
20	0.00	0.00	0.00	0.00	0.30	0.02	0.18	0.07	0.22	0.34	0.03	0.37	0.02
21	0.00	0.00	0.00	0.00	0.02	0.22	0.12	0.02	0.21	0.08	0.09	0.41	0.16
22	0.00	0.00	0.00	0.00	0.16	0.11	0.29	0.25	0.17	0.15	0.13	0.27	0.23
23	0.00	0.00	0.00	0.02	0.07	0.01	0.00	0.12	0.51	0.22	0.61	0.04	0.30
24	0.00	0.00	0.01	0.02	0.02	0.06	0.07	0.07	0.02	0.11	0.07	0.29	0.93
25	0.00	0.00	0.00	0.02	0.03	0.01	0.03	0.23	0.07	0.13	0.53	0.34	0.65
40	0.00	0.00	0.00	0.09	0.01	0.08	0.00	0.16	0.39	0.44	0.14	0.38	0.11

Table 7.8 shows that Q_5 is not significantly different from some t-distributions using a KS test when the degrees of freedom of the underlying variables is greater than three. Also, there are a range of t-distributions that fail to reject the KS test when Q_5 is computed using t-distributions with four or more degrees of freedom. At this point, Q_k has only been investigated with the sum of five variables.

7.2.4 Simulation to Test if Q_{15} has a t-Distribution

Another simulation is constructed to investigate how the number of t-distributions, k, used in the calculation of Q_k affects the distribution. The previous simulation was repeated expanding k from five random variables to include the sum of fifteen t-distributions in the calculation of Q_k.

Again, the simulation is constructed utilizing two nested loops. The outer loop iterates the degrees of freedom for the underlying t-distributions used in the construction of Q_k from two degrees of freedom up to 20. Once the selection is made, 15 t-distributed variables are randomly generated and Q_{15} is calculated. This is repeated 10,000 times in order to create a distribution of Q_{15} at each level of the degrees of freedom of the underlying t-distributions. For each iteration, an inner loop tests the distribution of Q_{15} against t-distributions with degrees of freedom ranging from one to 25 and one additional observation with forty degrees of freedom. The two distributions are tested using a Kolmogorov-Smirnov *D* statistic for two-sample data based on the empirical distribution function found in the Npar1way procedure in SAS. Q_k is constructed using 15 randomly generated t-distributions using the rand call in SAS.

Figure 7.12 shows the distribution of Q_{15} comprised of $X_{v,i} \sim t_{10}$compared to a t-distribution with ten degrees of freedom. Looking at Table 7.9 below, you can see the results from the KS test showing that there is a significant difference between these distributions with a p-value of 0.02.

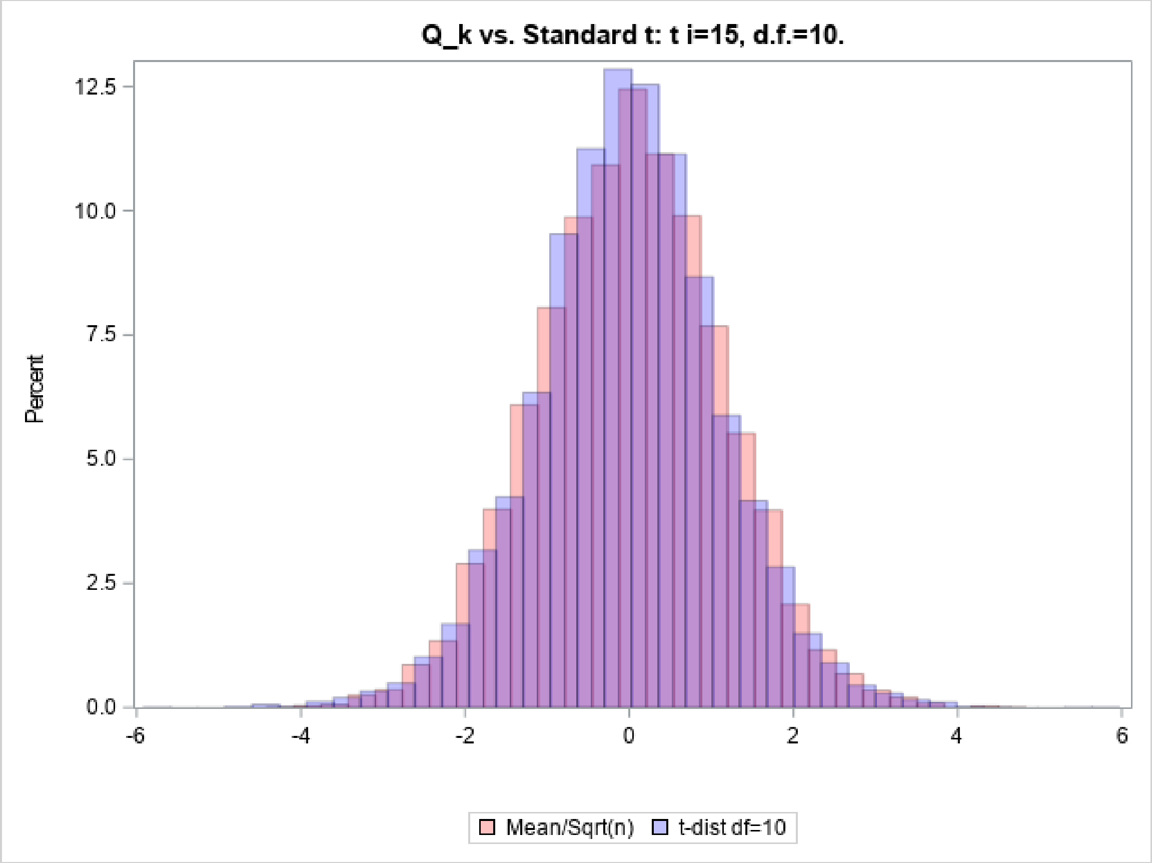


Figure 7.12 Q_{15} Comprised of $X_{\nu,i} \sim t_{10}$ Compared to a t-distribution with Ten Degrees of Freedom

Table 7.9 shows the results of the simulation with 10,000 observations at each level. The columns of the table are the degrees of freedom for the 15 t-distributions that comprise Q_{15} and the degrees of freedom for the t-distribution that is being tested against Q_k are going down the rows of the table. The p-values of the KS test for each combination are displayed in the table. P-values that fail to reject the null hypothesis that $Q_{15}{\sim}t$-distribution are highlighted in yellow.

Table 7.9 P-values of KS Test for $H_0: Q_{15} \sim t_\nu$ $(n = 10{,}000)$. The columns represent the seed degrees of freedom in Q_{15} and the rows are the degrees of freedom in t_ν.

d.f. t_v	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.14	0.04	0.02	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.09	0.01	0.26	0.26	0.27	0.10	0.05	0.01	0.01	0.03	0.01
4	0.00	0.00	0.02	0.01	0.03	0.07	0.56	0.47	0.04	0.48	0.10	0.29	0.32
5	0.00	0.00	0.00	0.06	0.02	0.13	0.59	0.97	0.68	0.19	0.80	0.80	0.50
6	0.00	0.00	0.00	0.10	0.06	0.02	0.12	0.12	0.84	0.23	0.45	0.97	0.47
7	0.00	0.00	0.00	0.01	0.05	0.03	0.11	0.00	0.36	0.30	0.25	0.77	0.49
8	0.00	0.00	0.00	0.00	0.04	0.42	0.03	0.05	0.04	0.02	0.07	0.65	0.29
9	0.00	0.00	0.00	0.01	0.08	0.05	0.84	0.05	0.02	0.24	0.86	0.40	0.28
10	0.00	0.00	0.00	0.00	0.29	0.23	0.03	0.29	0.67	0.27	0.13	0.25	0.14
11	0.00	0.00	0.00	0.00	0.00	0.06	0.09	0.09	0.27	0.10	0.31	0.79	0.20
12	0.00	0.00	0.00	0.00	0.07	0.30	0.18	0.13	0.05	0.32	0.10	0.45	0.27
13	0.00	0.00	0.00	0.00	0.01	0.13	0.14	0.24	0.23	0.12	0.19	0.10	0.66
14	0.00	0.00	0.00	0.01	0.04	0.22	0.01	0.29	0.29	0.28	0.14	0.83	0.72
15	0.00	0.00	0.00	0.00	0.00	0.03	0.07	0.12	0.16	0.13	0.53	0.55	0.39
16	0.00	0.00	0.00	0.00	0.00	0.09	0.07	0.11	0.17	0.30	0.51	0.13	0.35
17	0.00	0.00	0.00	0.03	0.00	0.05	0.01	0.00	0.04	0.25	0.27	0.48	0.04
18	0.00	0.00	0.00	0.00	0.07	0.01	0.18	0.23	0.47	0.45	0.39	0.03	0.37
19	0.00	0.00	0.00	0.00	0.00	0.03	0.21	0.08	0.04	0.05	0.10	0.59	0.09
20	0.00	0.00	0.00	0.00	0.00	0.00	0.18	0.21	0.20	0.05	0.12	0.15	0.13
21	0.00	0.00	0.00	0.00	0.01	0.01	0.04	0.02	0.08	0.24	0.96	0.75	0.03
22	0.00	0.00	0.00	0.01	0.01	0.03	0.01	0.40	0.25	0.10	0.21	0.53	0.46
23	0.00	0.00	0.00	0.00	0.00	0.03	0.01	0.00	0.13	0.24	0.19	0.83	0.19
24	0.00	0.00	0.00	0.00	0.01	0.01	0.07	0.14	0.01	0.13	0.46	0.30	0.03
25	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.05	0.30	0.22	0.78	0.63	0.26
40	0.00	0.00	0.00	0.00	0.00	0.01	0.06	0.01	0.02	0.82	0.18	0.63	0.93

The random variable, Q_k, where $k = 15$ is not significantly different from some t-distributions when the degrees of freedom for the underlying t-distributions is greater than three. However, there is still a large range of t-distributions that fail to reject the hypothesis.

7.2.5 Normality of Q_k when k Ranges from Five to Fifty

It has been shown Q_5 fails to reject the hypothesis that it has a normal distribution when the underlying t-distributions have around seven or more degrees of freedom. Is the normality a function of k? Standard rhetoric states that normality is achieved around a sample size of 30. It is postulated that the normality achieved around seven degrees of freedom may be a function of k as in $5 \times 7 = 35 > 30$. To test this hypothesis, a new simulation was devised to investigate Q_k for $k = 5, 10, 15, 20, 25, 30, 35, 40, 45$ and 50. A macro was written to cycle k from five to 50 in intervals of five. For each instance a loop was written to iterate over the degrees of freedom of the underlying t-distributions from two to 25 and one additional loop with 40 degrees of freedom. For each iteration, Q_k was calculated and tested for normality using a Kolmogorov-Smirnov test for normality in the Univariate procedure in SAS.

The figure below shows the histogram of Q_{10} calculated with randomly generated t-distributions with six degrees of freedom. The graph is overlaid with a normal distribution. Looking at Table 7.10, the p-value of this KS test for normality is 0.04.

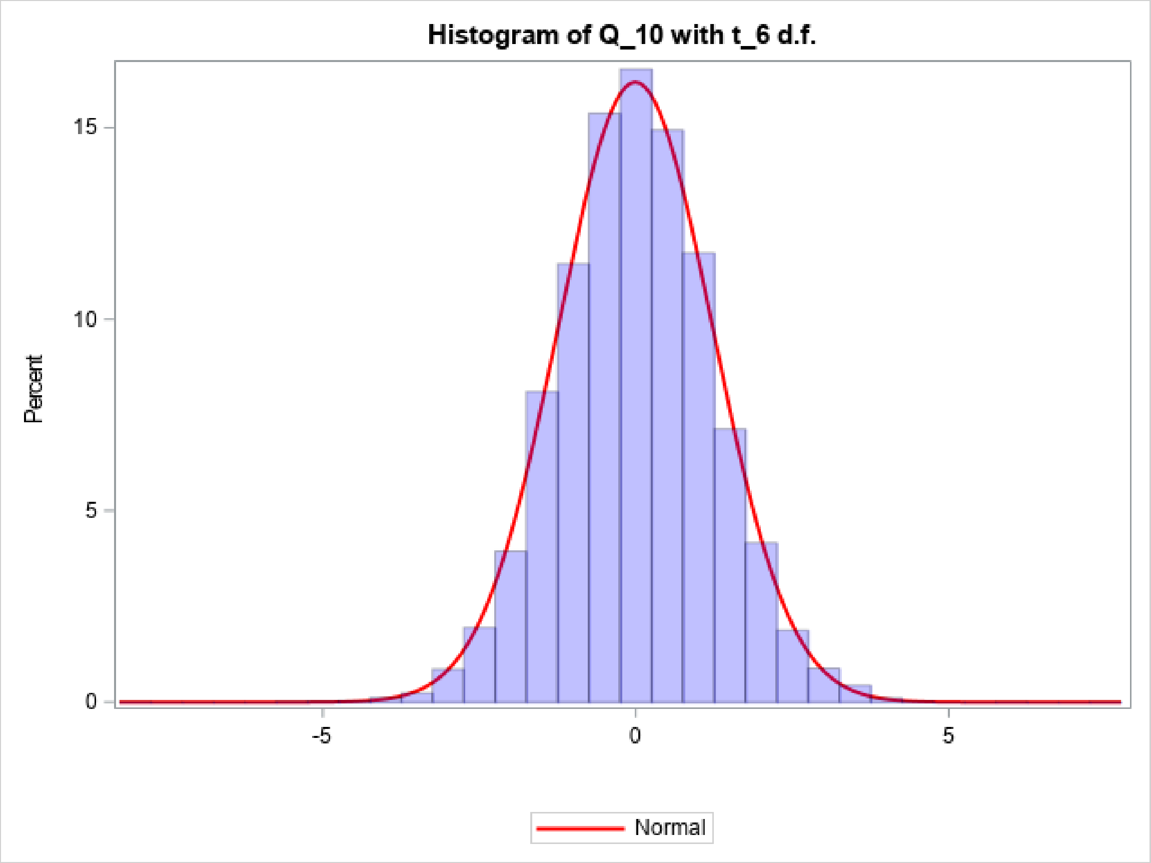


Figure 7.13 Histogram of Q_{10} Comprised of t-distributions with Six Degrees of Freedom

The table below shows the p-value for the Kolmogorov-Smirnov test for normality from the Univariate procedure in SAS. The degrees of freedom used in the underlying t-distributions in the calculation of Q_k are going down the rows on the left of the table and the number of variables, k, is going across the top of the table. Thus the p-value for Q_{35}, when it is calculated using t-distributions with seven degrees of freedom, is 0.01.

Table 7.10 P-values of KS Test for $H_0: Q_k \sim N(\mu, \sigma)$ $(n = 10{,}000)$. Values of k are going across the columns and the rows represent the seed degrees of freedom in Q_k.

d.f. of t	5	10	15	20	25	30	35	40	45	50
2	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
3	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
4	0.01	0.01	0.01	0.03	0.01	> 0.15	0.07	0.04	0.04	> 0.15
5	0.01	0.03	0.02	0.06	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15
6	0.01	0.04	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	0.08	> 0.15
7	0.01	> 0.15	0.15	> 0.15	> 0.15	> 0.15	0.01	> 0.15	> 0.15	> 0.15
8	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	0.04	> 0.15	> 0.15	> 0.15
9	0.05	0.02	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15
10	0.05	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	0.02	> 0.15	> 0.15	> 0.15
11	0.03	> 0.15	> 0.15	> 0.15	> 0.15	0.09	> 0.15	> 0.15	0.14	> 0.15
12	0.01	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	0.13	0.02
13	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15
14	0.09	> 0.15	> 0.15	> 0.15	0.10	> 0.15	0.04	> 0.15	> 0.15	> 0.15
15	> 0.15	> 0.15	0.11	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15
16	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15
17	> 0.15	> 0.15	> 0.15	0.15	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	0.01
18	0.03	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	0.06	> 0.15	> 0.15
19	> 0.15	0.10	> 0.15	> 0.15	> 0.15	0.01	> 0.15	> 0.15	> 0.15	> 0.15
20	> 0.15	> 0.15	0.10	0.03	> 0.15	> 0.15	> 0.15	> 0.15	0.06	0.02
21	0.01	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	0.15
22	> 0.15	> 0.15	> 0.15	> 0.15	0.03	0.09	> 0.15	> 0.15	> 0.15	0.09
23	0.08	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	0.01	> 0.15
24	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	0.10	0.15	> 0.15	0.07	> 0.15
25	0.09	> 0.15	> 0.15	0.04	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	0.14
40	> 0.15	0.01	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15	> 0.15

Normality is achieved by Q_k with lower degrees of freedom as k increases up to k = 30. At $k \geq 30$, normality is achieved around four degrees of freedom.

7.2.6 $Q_k \sim t$ for Finite Samples when k Ranges from Five to Fifty

Given what was learned in the previous simulation, it is questioned how k impacts the distribution of Q_k with smaller finite samples. Specifically, under what circumstances is the distribution of Q_k a t-distribution with different values of *k*. With this in mind, a larger simulation was created that iterated over three variables: the degrees of freedom of the underlying t-distributions, the degrees of freedom of the t-distribution tested against Q_k, and the size of *k*. The variable k was investigated from five to 50 at intervals of five.

A simulation is created that cycles through three nested loops. The first loop takes initial values for k and the degrees of freedom for the underlying t-distributions, randomly samples 1000 iterations of *k* t-distributions, calculates Q_k, tests the distribution of Q_k against t-distributions from two to 25 and one additional test with 40 degrees of freedom. This is repeated for the degrees of freedom for the underlying t-distributions from two to 20. This cycle is repeated at different levels of *k* from five to 50 at intervals of five. Q_k is compared to different t-distributions with a Kolmogorov-Smirnov *D* statistic for two-sample data based on the empirical distribution utilizing the Npar1way procedure in SAS.

Tables 7.11, 7.12, and 7.13 show the p-values from the KS test for each of the levels of k for Q_k calculated with t-distributions with degrees of freedom five, 10, and 15, respectively. Each table is highlighted with higher p-values in blue and lower p-values in red. The goal is to identify trends where the highest p-values are occurring. The different levels of k are going across the columns at the top and the different degrees of freedom of the t-distributions that are compared with Q_k are going down the rows on the left of the table.

Table 7.11 P-values for KS Test for (Q_k Comprised of t_5) ~ t_ν ($n = 1000$). Values of k are going across the columns and the rows represent the degrees of freedom in t_ν.

df of t	5	10	15	20	25	30	35	40	45	50
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.14	0.06	0.12	0.00	0.01	0.03	0.11	0.13	0.03	0.06
3	0.49	0.17	0.10	0.01	0.18	0.01	0.11	0.54	0.03	0.01
4	0.01	0.04	0.06	0.01	0.36	0.13	0.10	0.13	0.02	0.01
5	0.09	0.20	0.02	0.13	0.00	0.04	0.00	0.02	0.09	0.10
6	0.00	0.03	0.01	0.01	0.08	0.01	0.00	0.03	0.00	0.00
7	0.00	0.01	0.05	0.02	0.03	0.03	0.01	0.00	0.03	0.01
8	0.01	0.00	0.00	0.04	0.00	0.02	0.01	0.00	0.00	0.03
9	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00
10	0.00	0.04	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.02
11	0.01	0.00	0.00	0.00	0.01	0.01	0.00	0.01	0.00	0.00
12	0.00	0.00	0.00	0.02	0.02	0.01	0.00	0.01	0.00	0.00
13	0.01	0.00	0.00	0.00	0.02	0.00	0.01	0.00	0.01	0.00
14	0.04	0.01	0.05	0.00	0.00	0.00	0.02	0.00	0.00	0.01
15	0.01	0.03	0.00	0.04	0.01	0.00	0.00	0.00	0.01	0.00
16	0.01	0.00	0.05	0.00	0.00	0.04	0.00	0.00	0.00	0.00
17	0.02	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
18	0.02	0.01	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00
19	0.00	0.00	0.01	0.00	0.00	0.02	0.00	0.00	0.00	0.02
20	0.06	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
21	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
23	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
25	0.00	0.01	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00
40	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00

Focusing on the table above, most of the non-significant p-values for Q_k comprised of t-distributions with five degrees of freedom are found in the comparison with t-distributions with two to four degrees of freedom. As previously shown in Table 7.10, Q_k fails to reject the hypothesis that it is normally distributed for higher levels of k.

Table 7.12 P-values for KS Test for $(Q_k$ Comprised of $t_{10}) \sim t_v$ $(n = 1000)$. Values of k are going across the columns and the rows represent the degrees of freedom in t_v.

df of t	5	10	15	20	25	30	35	40	45	50
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.26	0.11	0.04	0.03	0.04	0.06	0.03	0.19	0.11	0.11
4	0.39	0.15	0.30	0.22	0.50	0.06	0.16	0.22	0.34	0.08
5	0.57	0.01	0.16	0.32	0.21	0.75	0.15	0.46	0.38	0.30
6	0.05	0.60	0.29	0.29	0.25	0.13	0.30	0.88	0.43	0.01
7	0.32	0.12	0.60	0.02	0.38	0.40	0.64	0.07	0.09	0.63
8	0.85	0.53	0.46	0.11	0.36	0.90	0.70	0.12	0.36	0.08
9	0.80	0.09	0.83	0.20	0.47	0.13	0.62	0.84	0.06	0.78
10	0.20	0.16	0.08	0.24	0.33	0.22	0.36	0.14	0.01	0.06
11	0.54	0.39	0.05	0.05	0.10	0.26	0.37	0.42	0.03	0.19
12	0.26	0.21	0.70	0.29	0.64	0.07	0.22	0.42	0.08	0.01
13	0.09	0.01	0.06	0.14	0.18	0.69	0.22	0.51	0.02	0.20
14	0.32	0.10	0.05	0.21	0.02	0.15	0.81	0.01	0.37	0.10
15	0.22	0.28	0.12	0.01	0.88	0.18	0.31	0.10	0.40	0.05
16	0.21	0.01	0.03	0.13	0.04	0.14	0.59	0.19	0.25	0.29
17	0.18	0.00	0.28	0.01	0.06	0.16	0.02	0.40	0.37	0.40
18	0.12	0.06	0.00	0.25	0.24	0.62	0.07	0.51	0.24	0.20
19	0.29	0.41	0.11	0.23	0.03	0.02	0.16	0.17	0.11	0.07
20	0.67	0.03	0.05	0.17	0.08	0.35	0.01	0.14	0.12	0.09
21	0.21	0.15	0.01	0.08	0.02	0.12	0.42	0.05	0.51	0.45
22	0.82	0.00	0.14	0.01	0.06	0.25	0.08	0.14	0.17	0.11
23	0.34	0.06	0.08	0.04	0.71	0.03	0.24	0.09	0.01	0.02
24	0.05	0.07	0.46	0.04	0.02	0.01	0.26	0.02	0.01	0.00
25	0.01	0.44	0.07	0.08	0.15	0.19	0.16	0.03	0.00	0.20
40	0.09	0.00	0.11	0.07	0.02	0.02	0.41	0.31	0.52	0.03

Looking at the table above it appears there is a pattern of higher p-values in the range of t-distributions with seven to nine degrees of freedom.

Table 7.13 P-values for KS Test for $(Q_k$ Comprised of $t_{15}) \sim t_\nu$ $(n = 1000)$. Values of k are going across the columns and the rows represent the degrees of freedom in t_ν.

df of t	5	10	15	20	25	30	35	40	45	50
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.01	0.09	0.02	0.10	0.03	0.03	0.04	0.04	0.17	0.01
4	0.16	0.07	0.05	0.36	0.13	0.27	0.18	0.43	0.00	0.41
5	0.28	0.34	0.31	0.32	0.68	0.04	0.12	0.68	0.31	0.69
6	0.33	0.08	0.82	0.30	0.32	0.66	0.54	0.19	0.42	0.28
7	0.85	0.17	0.43	0.34	0.45	0.46	0.92	0.80	0.49	0.56
8	0.93	0.08	0.89	0.60	0.36	0.07	0.48	0.35	0.48	0.35
9	0.19	0.02	0.72	0.71	0.69	0.01	0.19	0.42	0.86	0.35
10	0.49	0.20	0.79	0.50	0.92	0.82	0.28	0.93	0.24	0.42
11	0.86	0.76	0.85	0.28	0.43	0.02	0.18	0.89	0.86	0.51
12	0.98	0.74	0.44	0.00	0.07	0.39	0.47	0.70	0.77	0.41
13	0.74	0.01	0.96	0.09	0.73	0.24	0.84	0.37	0.33	0.39
14	0.10	0.92	0.65	0.66	0.05	0.12	0.00	0.99	0.40	0.13
15	0.60	0.51	0.00	0.16	0.82	0.90	0.46	0.24	0.05	0.34
16	0.12	0.65	0.02	0.32	0.59	0.02	0.22	0.02	0.96	0.64
17	0.48	0.78	0.44	0.06	0.95	0.16	0.46	0.71	0.29	0.13
18	0.40	0.23	0.37	0.52	0.21	0.06	0.83	0.45	0.93	0.33
19	0.97	0.80	0.61	0.31	0.52	0.24	0.56	0.05	0.59	0.13
20	0.09	0.35	0.02	0.16	0.29	0.10	0.06	0.20	0.02	0.51
21	0.68	0.47	0.32	0.15	0.19	0.37	0.10	0.28	0.87	0.11
22	0.00	0.30	0.24	0.40	0.69	0.03	0.96	0.02	0.55	0.07
23	0.38	0.16	0.42	0.57	0.54	0.24	0.20	0.05	0.76	0.13
24	0.93	0.40	0.24	0.29	0.34	0.41	0.62	0.25	0.24	0.04
25	0.11	0.29	0.04	0.35	0.24	0.01	0.29	0.04	0.19	0.07
40	0.04	0.80	0.21	0.91	0.14	0.07	0.04	0.09	0.77	0.13

The table above shows a pattern of higher p-values for t-distributions with eight to ten degrees of freedom. Across the tables there are many instances of Q_k which fail to reject the null hypothesis that they have t-distribution. In each case there is a range of possibly acceptable t-distributions, but not a clear cut solution. Additionally, it is likely that the optimal t-distribution does not have an integer degrees of freedom.

7.2.7 Introduction and Simulation using $\nu_s = \frac{2s^2}{s^2-1}$

This thought led to possible ways to narrow down the optimal degrees of freedom for the t-distribution. It was noted that the second moment for a t-distribution, $\delta^2 = \frac{\nu}{\nu-2}$, is defined by its degrees of freedom, ν. Using a simulation to approximate the distribution of Q_k, the sample statistic for the second moment can then be calculated from the distribution. Solving the equation above for ν gives $\nu = \frac{2\delta^2}{\delta^2-1}$. Then substituting the sample statistic for the parameter gives $\nu_s = \frac{2s^2}{s^2-1}$. A simulation is constructed to test if $Q_k \sim t_{\nu_s}$ where $\nu_s = \frac{2s^2}{s^2-1}$.

A simple example is provided before a larger simulation to illustrate the use of ν_s. Five hundred independent observations from a t-distribution with 4.6 degrees of freedom were randomly generated. The univariate procedure in SAS was utilized to test the distribution for normality. As expected, the hypothesis that the distribution was normally distributed was rejected with a p-value < 0.0001. The sample variance, s^2, of the distribution was equal to 1.950. Using the sample variance, ν_s was calculated and found to be equal to 4.106. A one sample KS test was utilized to test the hypothesis that the distribution had a t-distribution with $\nu_s = 4.106$ degrees of freedom. Using the ks.test.t function in R, the p-value from the one sample KS test was equal to 0.896, thus the hypothesis fails to be rejected. Additionally, the two sample KS test using a randomly generated t-distribution with $\nu_s = 4.106$ degrees of freedom was completed using the Npar1way procedure in SAS finding a p-value of 0.935. Repeated tests showed similar results where the two KS test with a randomly generated dataset aligned with the one sample KS test utilizing R software. Given these results the decision was made to move forward with the two sample test in SAS in order to allow for a larger simulation that can be automated in SAS and remove the need for manual manipulation of the dataset in order to transfer the individual dataset to R for the one sample KS test.

A simulation was created that randomly generated k t-distributions each with ν degrees of freedom. The degrees of freedom were iterated from two to 20 and the number of variables, k, included inside the calculation for Q_k was iterated from five to 50 in intervals of five. For each cycle, Q_k was calculated 1000 times and then the sample variance of the distribution was computed. Using the sample variance ν_s was calculated and the distribution of Q_k was compared with $t_{\nu_s}$with a Kolmogorov-Smirnov *D* statistic for two-sample data based on the empirical distribution with the Npar1way procedure in SAS.

Table 7.14 shows the calculated ν_s for each iteration of *k* across the columns at the top and each iteration of the degrees of freedom used to generate the underlying t-distributions used in the calculation for Q_k go down the rows on the left of the table.

Table 7.14 Calculated ν_s for Different k and Underlying t_ν in Q_k. Values of k are going across the columns and the rows represent the seed degrees of freedom in Q_k.

df of t_v	5	10	15	20	25	30	35	40	45	50
2	2.39	2.16	2.11	2.22	2.22	2.24	2.25	2.26	2.23	2.13
3	3.16	3.00	3.07	3.11	3.17	2.97	3.16	3.05	3.01	3.15
4	4.19	3.83	3.76	4.07	3.82	4.05	3.87	3.85	3.86	3.95
5	5.03	5.29	4.51	5.31	4.65	5.69	5.39	4.93	4.77	5.03
6	5.13	7.55	5.73	5.90	5.65	6.64	6.08	6.68	6.43	6.18
7	6.77	6.93	6.63	6.09	6.66	7.06	9.02	6.68	7.81	5.68
8	8.31	8.77	7.81	7.42	8.65	5.88	6.37	8.17	7.74	7.98
9	11.55	8.91	9.01	13.20	8.31	10.44	8.58	9.31	8.22	6.82
10	12.11	11.76	10.40	13.17	10.26	8.66	11.32	9.84	11.23	9.82
11	11.81	10.32	15.18	11.30	9.57	13.21	9.92	10.26	13.71	17.73
12	12.77	13.80	15.75	13.01	11.49	13.97	10.65	15.96	19.24	11.07
13	11.35	14.64	10.45	12.01	14.24	8.58	18.38	10.35	8.92	13.26
14	21.06	12.05	10.17	17.71	17.34	10.24	11.10	15.69	14.48	14.89
15	21.84	12.68	12.21	27.29	11.84	14.46	14.50	17.98	15.59	12.90
16	29.79	14.40	38.10	11.50	12.49	12.32	25.46	12.09	17.16	13.89
17	12.13	38.20	18.78	14.89	16.88	12.09	21.55	33.96	22.52	20.03
18	17.39	50.52	33.12	12.21	17.94	21.97	22.77	13.56	20.44	15.30
19	14.03	23.45	24.54	27.25	13.89	14.32	51.45	43.93	15.31	16.74
20	23.74	94.11	19.03	38.57	26.41	12.67	19.09	31.80	14.67	15.05

Table 7.15 shows the p-values for the KS test comparing the distribution of Q_k with the simulated distribution t_{ν_s} where ν_s is shown in the table above. Again, the levels of k go across the columns at the top of the table and the degrees of freedom used in the underlying t-distributions used in the calculation of Q_k go down the rows on the left of the table.

Table 7.15 P-values for $H_0: Q_k \sim t_{\nu_s}$. Values of k are going across the columns and the rows represent the seed degrees of freedom in Q_k.

df of t_v	5	10	15	20	25	30	35	40	45	50
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.02	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.04	0.06	0.00	0.02	0.00	0.00	0.00
5	0.37	0.61	0.15	0.12	0.37	0.16	0.31	0.03	0.12	0.03
6	0.26	0.20	0.43	0.31	0.15	0.26	0.47	0.22	0.34	0.47
7	0.12	0.69	0.29	0.10	0.08	0.16	0.99	0.43	0.07	0.18
8	0.22	0.89	0.43	0.43	0.91	0.11	0.01	0.22	0.50	0.29
9	0.24	0.26	0.01	0.31	0.20	0.79	0.08	0.99	0.24	0.50
10	0.16	0.31	0.18	0.06	0.20	0.20	0.72	0.26	0.76	0.00
11	0.03	0.12	0.03	0.24	0.50	0.18	0.24	0.29	0.43	0.29
12	0.83	0.95	0.22	0.72	0.47	0.43	0.43	0.76	0.99	0.95
13	0.91	0.11	0.47	0.12	0.50	0.40	0.12	0.12	0.29	0.54
14	0.57	0.18	0.72	0.99	0.05	0.76	0.10	0.29	0.04	0.24
15	0.95	0.47	0.03	0.40	0.89	0.47	0.31	0.86	0.76	0.69
16	0.69	0.97	0.65	0.13	0.83	0.02	0.97	0.15	0.02	0.15
17	0.37	0.50	0.12	0.57	0.24	0.79	0.94	0.34	0.01	0.61
18	0.20	0.08	0.50	0.40	0.65	1.00	0.15	0.95	0.29	0.54
19	0.94	0.79	0.37	0.94	0.54	0.08	0.26	0.76	0.40	0.72
20	0.57	0.89	0.65	0.04	0.43	0.26	0.29	0.15	0.91	0.91

The table shows t-distributions with ν_s degrees of freedom appear to accurately match the distributions of Q_k. For Q_k comprised of t-distributions with degrees of freedom greater than four, 91.9% of the tests fail to reject the null hypothesis, providing evidence that $Q_k \sim t_{\nu_s}$. There are a couple of instances where the ν_s seemed uncharacteristically large. Delving into these occurrences, the issue was found in the structure of the definition.

$$\nu_s = \frac{2s^2}{s^2 - 1}$$

Notice if the sample variance is very close to one, then the denominator is nearly zero and ν_s can be inflated. It is worth noting that $\nu_s = 94.11$ for k = 10 and the underlying t-distributions used in the calculations of Q_k have 20 degrees of freedom. However, the p-value for this KS test is 0.89 showing that the null hypothesis fails to reject that $Q_{10} \sim t_{94.11}$. Another limitation is that the distribution of the test statistic would need to be known in order to compute ν_s. This limitation can be overcome with simulations or resampling techniques.

7.2.8 The Satterthwaite Approximation

The idea to test the Satterthwaite approximation to the degrees of freedom was brought up as a possibility for a solution. In order to test this hypothesis and compare the Satterthwaite approximation to the degrees of freedom, ν_{sat} with ν_s a simulation was created to calculate both approximations and compare $t_{\nu_{sat}}\ and\ t_{\nu_s}$ with the appropriate test statistic, which will be designated W_k. Let X_{ij} be independent and identically distributed observations from a standard normal distribution where $i = 1, \dots, k\ and\ j = 1, \dots, n$, then define:

$$W_k = \frac{\sum_{i=1}^{k} \bar{X}_i}{S_p\sqrt{\frac{k}{n}}}$$

Where S_p is the pooled variance estimator and $\bar{X}_i = \frac{\sum_{j=1}^{n} x_{ij}}{n}$.

A simulation was created with 5000 iterations. With each iteration six observations of each X_i were randomly generated from a standard normal distribution which resulted in 5000 replicas of W_k. The mean and variance for each X_i array were calculated along with the pooled variance estimators, Satterthwaite approximations for the degrees of freedom ($\nu_{2,sat}, \nu_{3,sat}, \nu_{4,sat}, and\ \nu_{5,sat}$) and W_k for $W_2, W_3, W_4, and\ W_5$. This led to a sample of 5000 observations for the four Satterthwaite approximations and W_k. The sample variance of each W_k, $s_{W_k}^2$, was calculated in order to calculate $\nu_{k,s}$ defined as:

$$\nu_{k,s} = \frac{2s_{W_k}^2}{s_{W_k}^2 - 1}$$

Since each distribution of W_k has only one $\nu_{k,s}$, but has 5000 observations of $\nu_{k,sat}$, the mean of the Satterthwaite approximations was utilized in order to make comparisons with $\nu_{k,s}$.

Next, 5000 observations from random t-distributions were generated with each of the degrees of freedom $\nu_{k,s}$ and $\nu_{k,sat}$. Finally, the distribution of each W_k was compared to the appropriate t-distribution with $\nu_{k,s}$ and $\nu_{k,sat}$ degrees of freedom utilizing a KS test in the Npar1way procedure in SAS.

The simulation was also repeated for n = 10, where n is the number of observations of each of the k standard normally distributed random variables in W_k.

Figure 7.14 shows a comparison of the histograms of $\nu_{k,sat}$ for k = 2, 3, 4 and 5 for the 5000 iterations of the simulation with n = 6.

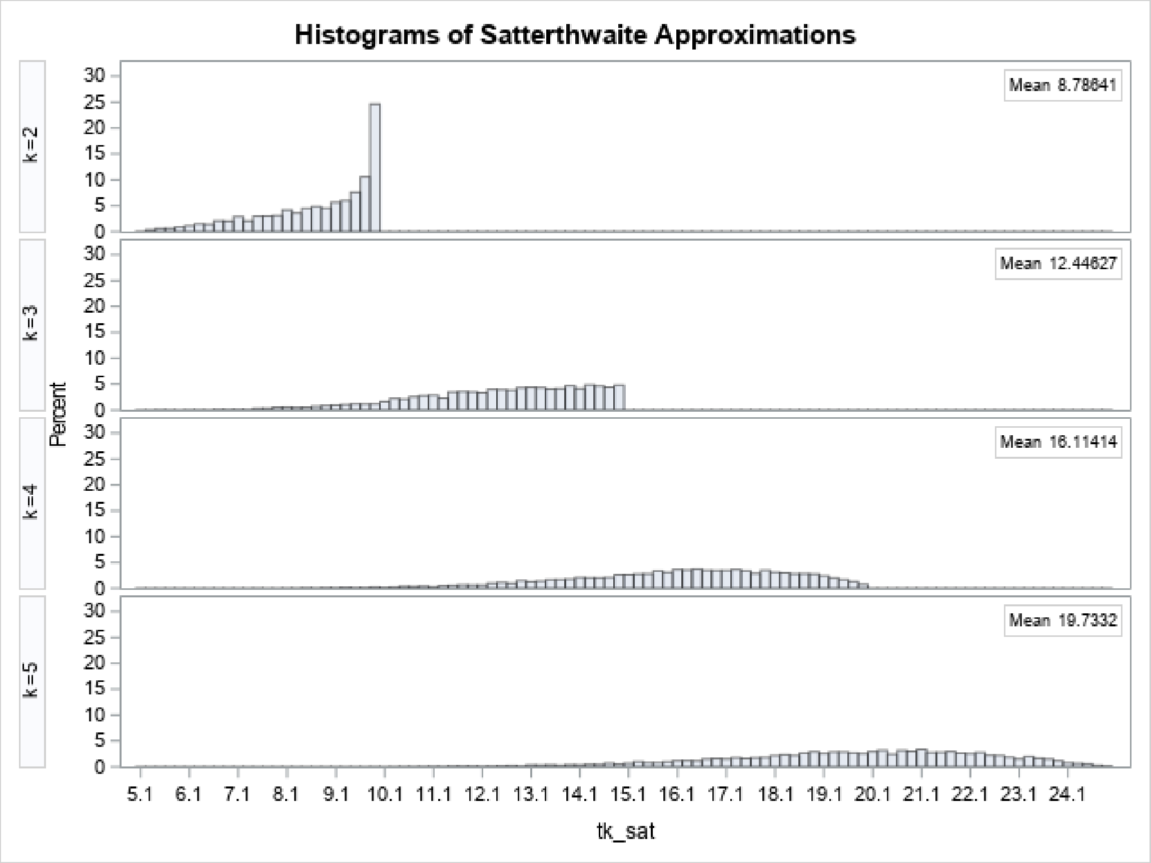


Figure 7.14 Histograms of $\nu_{k,sat}$ for $k = 2, 3, 4,$ and 5 ($n = 6$)

It is worth noting that each histogram is bounded on the right by the exact degrees of freedom. For k = 2, the exact degrees of freedom is equal to $n_1 + n_2 - 2 = 6 + 6 - 2 = 10$, and for k = 5 the exact degrees of freedom is $n_1 + n_2 + n_3 + n_4 + n_5 - 5 = 25$.

Figure 7.15 shows a comparison of the histograms for W_2, W_3, W_4 and W_5 for n = 6. Notice as k increases, the degrees of freedom for the Satterthwaite approximation increases and the standard deviation of the distributions decreases.

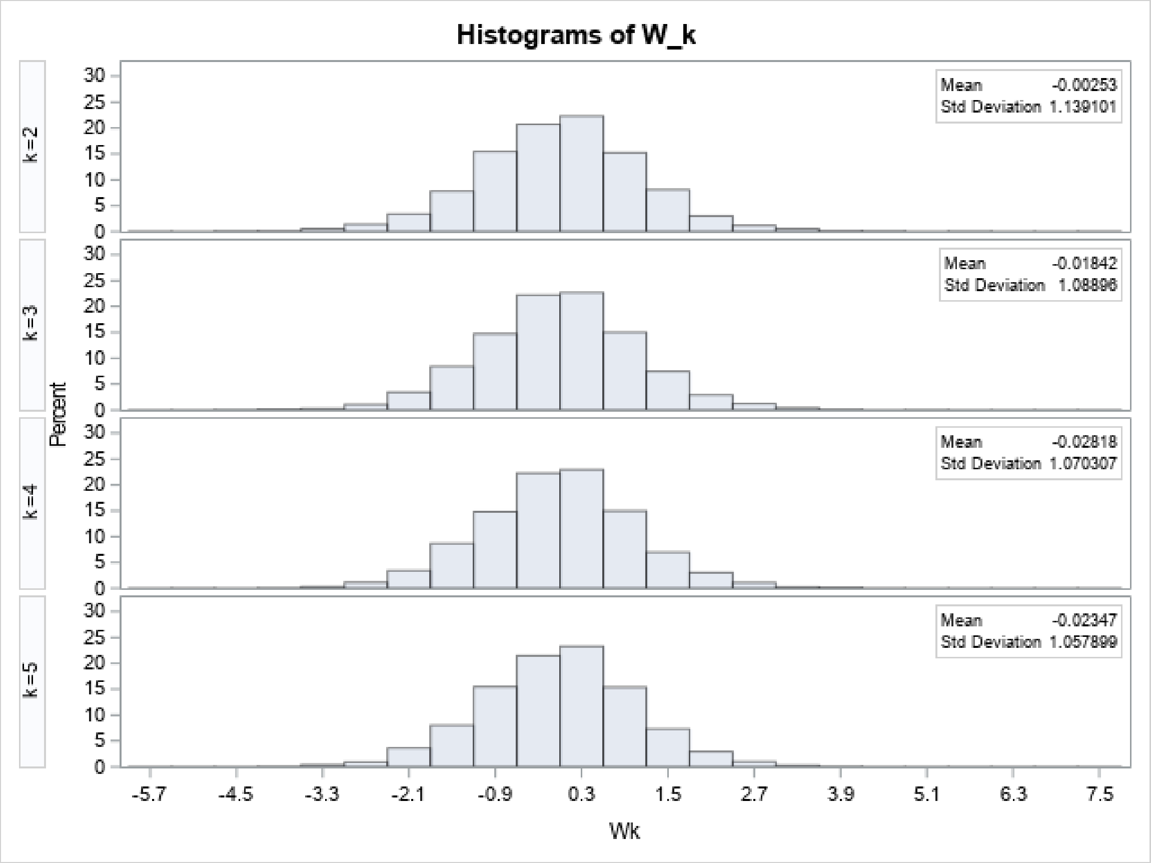
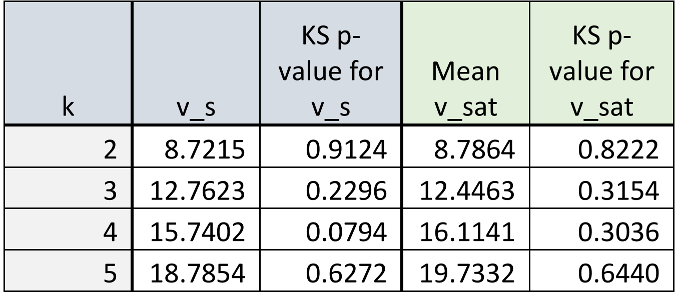


Figure 7.15 Histograms of W_k for $k = 2, 3, 4,$ and 5 $(n = 6)$

Table 7.16 shows the two approximations to the degrees of freedom where $n = 6$ for $k = 2, 3, 4,$ and 5. In addition, the table displays the p-value for the Kolmogorov-Smirnov D statistic for two-sample data based on the empirical distribution for the null hypothesis $H_0: W_k \sim t_{\nu_{k,sat}}$ and $H_0: W_k \sim t_{\nu_{k,s}}$.

Table 7.16 Values of ν_s, ν_{sat}, and p-values from KS Tests $(n = 6)$



k	v_s	KS p-value for v_s	Mean v_sat	KS p-value for v_sat
2	8.7215	0.9124	8.7864	0.8222
3	12.7623	0.2296	12.4463	0.3154
4	15.7402	0.0794	16.1141	0.3036
5	18.7854	0.6272	19.7332	0.6440

The simulation was repeated for n = 10. Figure 7.16 shows the histograms for the Satterthwaite approximations for the degrees of freedom for k = 2, 3, 4 and 5. Again, note that the histograms are bounded on the right by the exact degrees of freedom. For k = 2, the exact degrees of freedom are $n_1 + n_2 - 2 = 18$ where $n = n_1 = n_2 = 10$ and for k = 5 the exact degrees of freedom is equal to 45.

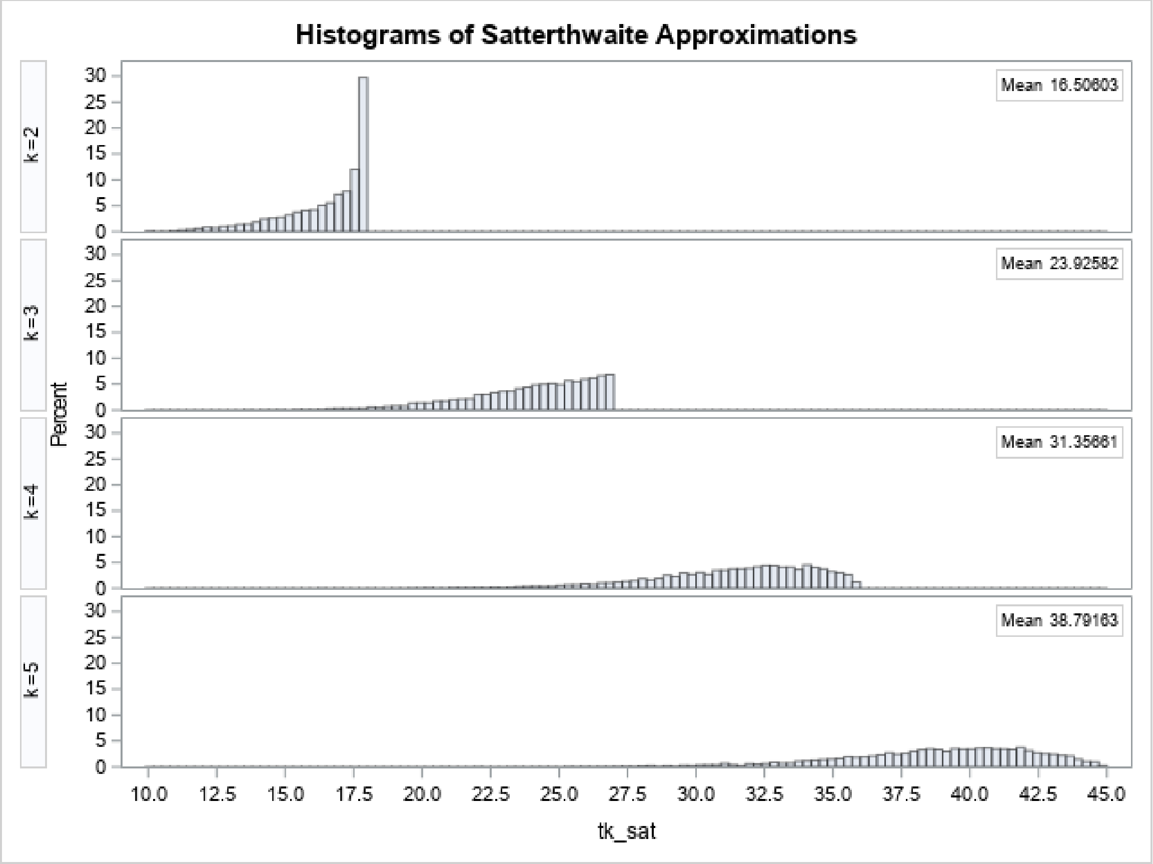


Figure 7.16 Histograms of $\nu_{k,sat}$ for $k = 2, 3, 4,$ and 5 $(n = 10)$

Figure 7.17 shows the histograms for the test statistics, W_k, for k = 2, 3, 4 and 5 where n = 10.

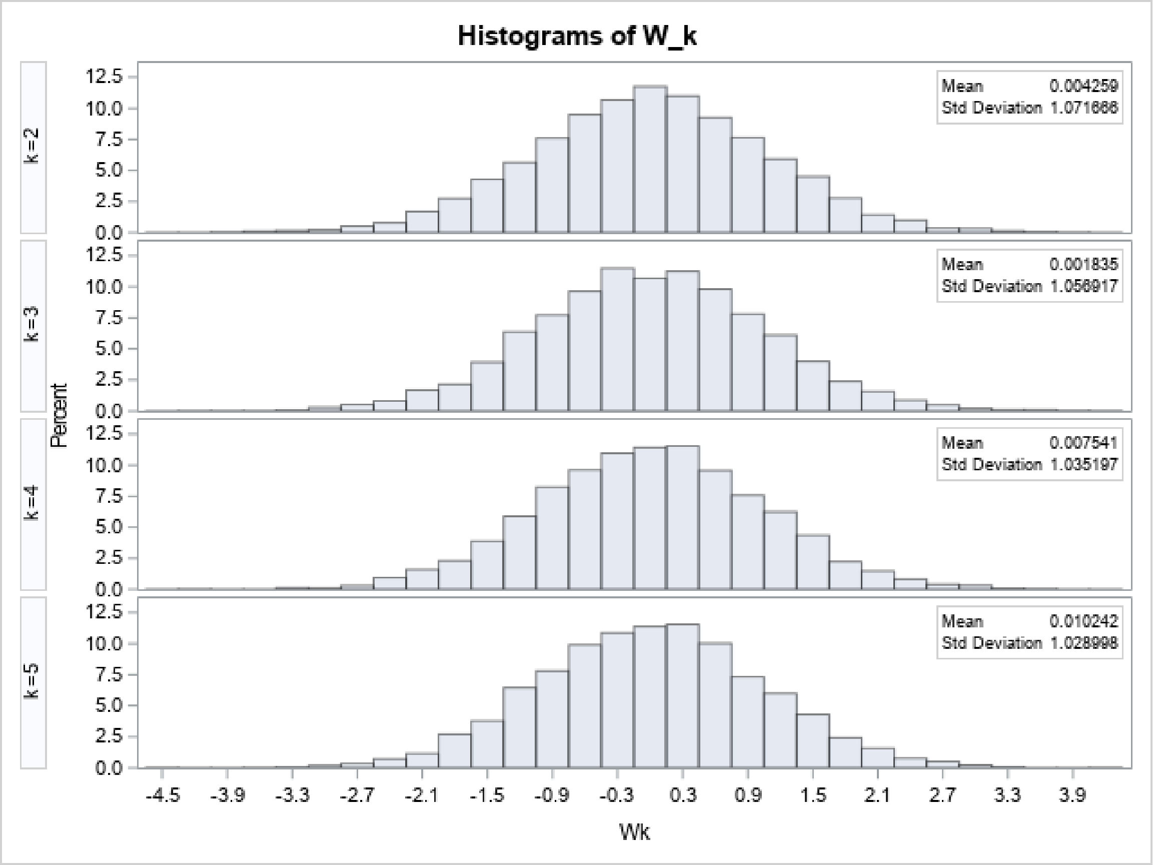


Figure 7.17 Histograms of W_k for $k = 2, 3, 4,$ and 5 $(n = 10)$

Table 7.17 shows the approximations for the degrees of freedom $\nu_{k,s}$ and $\nu_{k,sat}$. Additionally, it shows the p-value for the KS test comparing the distribution of the test statistic W_k with the two hypothesized distributions t_{ν_s} and $t_{\nu_{sat}}$.

Table 7.17 Values of ν_s, ν_{sat}, and p-values from KS Tests $(n = 10)$

k	v_s	KS p-value for v_s	Mean v_sat	KS p-value for v_sat
2	15.4709	0.3791	16.5060	0.1177
3	19.0833	0.6104	23.9258	0.2024
4	29.9204	0.2921	31.3566	0.7112
5	35.9918	0.1777	38.7916	0.2809

The estimate to the degrees of freedom, ν_s, tracks well with the Satterthwaite degrees of freedom for both of the samples. Additionally, ν_s is tracking with the mean of the distribution of ν_{sat} given multiple samples, adding confidence that ν_s is an accurate approximation to the degrees of freedom. Since the Satterthwaite approximation is calculated from the sample, it is easier to calculate with a single sample. However, it has a greater variability, as can be seen in Figures 7.14 and 7.16. The ν_s approximation of the degrees of freedom is easily applied in situations, like simulations and resampling techniques, where the distribution can be estimated and where the test statistic is non-standard. Both of these situations make ν_s an attractive solution for the approximation of the degrees of freedom for Q_k with finite samples.

CHAPTER 8. CONCLUSIONS

Sepsis is an increasing problem in modern medical facilities. Additionally, patients who develop sepsis while in the hospital have worse outcomes and the costs for care are attributed to the healthcare facility. The largest barrier to better sepsis outcomes is the early identification and treatment of septic patients. Early interventions lead to better outcomes, shorter lengths of stay, and financial savings. It has been shown that the distributions of vital signs exhibit large measurable changes during the onset of sepsis. Utilizing variables that measure the change for each vital sign from the patient's specific baselines, a discriminant model was constructed in order to proactively identify patients at risk for developing sepsis while on acute and progressive levels of care in hospitals. Running this sepsis score in the background at the University of Kentucky Chandler Hospital, it has been shown to be successful in the early identification of sepsis patients.

Extensive simulations were completed to investigate the sum of t-distributions with $\nu > 0$ degrees of freedom. Results from simulations showed that $Q_{k,\nu}$, the sum of random variables that have a t-distribution divided by the square root of k, fails to reject the hypothesis that $Q_{k,\nu}$ has an approximate normal distribution. In addition, for finite $\nu > 4$, simulations also showed that $Q_{k,\nu}$ failed to reject the hypothesis that it is well approximated by a t-distribution. However, each $Q_{k,\nu}$ failed to reject the hypothesis that had an approximate t_{ν^*} distribution for a range of ν^*. In order to better approximate the degrees of freedom for the approximate t-distribution an estimate, ν_s, was developed from the second moment of a t-distribution. Simulations show that $Q_{k,\nu}$ failed to reject the hypothesis that it had an approximate t-distribution with ν_s degrees of when the underlying t-distributions in $Q_{k,\nu}$ had degrees of freedom $\nu > 4$. A simulation was constructed to compare the approximate for the degrees of freedom for a t-distribution ν_s with the traditional Satterthwaite approximation for degrees of freedom. The approximate ν_s compared favorably with the Satterthwaite approximation. Additionally, the approximate ν_s can be utilized for a wider range of situations than the Satterthwaite approximation.

8.1 Future Research

Opportunities to apply the change in vital signs prediction technique in a children's hospital setting are planned for future research. Decline in a pediatric setting is most often associated with respiratory distress and decline. This presents challenges to add additional variables to the model including calculated variables that explore trends over time. Early research has included looking at changes in respiratory settings and devices utilized for the patient.

Research directions will also be explored in PICU and NICU settings where more continuous measurements are being taken on each patient. This leads to occasions where multiple measurements may be able to be used for the current state instead of only one.

Additionally, opportunities to add a fully automated simulation that includes a one sample KS test comparing Q_k with a t-distribution with ν_s degrees of freedom will be explored. Given a sample distribution, an automated package could be developed such that ν_s is calculated from the sample variance and a one sample KS test is run with a single function call. This would greatly simplify the process since it would not necessitate the researcher providing a guess for the degrees of freedom.

BIBLIOGRAPHY

Ahsanullah, M., Kibria, B. M. G., & Shakil, M. Normal and student's t distributions and their applications. Amsterdam: Atlantis Press: 2014. 93-101 p.

Aminikhanghahi S, Cook DJ. A Survey of Methods for Time Series Change Point Detection. Knowledge and Information Systems. 2017 May;51(2):339-67.

Casella, G. Berger, R. L. Statistical Inference. Pacific Grove: Cengage Learning & Wadsworth, ©2002. Print.

Centers for Disease Control and Prevention. Heathcare-Associated Infections [Internet]. Atlanta (GA): Centers for Disease Control and Prevention; 2015 Nov [cited 2015 Dec 5]. Available from https://www.cdc.gov/hai/prevent/prevention.html.

Centers for Disease Control and Prevention. Prevention Healthcare-associated Infections [Internet]. Atlanta (GA): Centers for Disease Control and Prevention; 2015 Nov 13 [cited 2015, Dec 3]. Available from: https://www.cdc.gov/hai/prevent/prevention.html.

Centers for Disease Control and Prevention. Sepsis: Data and Reports [Internet]. Atlanta (GA): Centers for Disease Control and Prevention; 2015 Feb 14 [cited 2015 November 20]. Available from: https://www.cdc.gov/sepsis/datareports/index.html.

Centers for Disease Control and Prevention. Sepsis: Data and Reports [Internet]. Atlanta (GA): Centers for Disease Control and Prevention; 2020 Feb 14 [cited 2020 May 1]. Available from: https://www.cdc.gov/sepsis/datareports/index.html.

Cleveland Clinic. Vital Signs, What are Vital Signs? [Internet]. Cleveland (OH): Cleveland Clinic: [cited 2014 Apr 1]. Available from: https://my.clevelandclinic.org/health/articles/10881-vital-signs.

Daley, John. For Colorado Mom, Story of Daughter's Hospital Death is Key to Others' Safety [Internet]. Centennial (CO): Colorado Publich Radio: 2015 Feb 17 [cited 2015 Nov 15]. Available from: https://www.cpr.org/2015/02/17/for-colorado-mom-story-of-daughters-hospital-death-is-key-to-others-safety/.

Eber MR, Laxminarayan R, Perencevich EN, Malani A. Clinical and Economic Outcomes Attributable to Health Care–Associated Sepsis and Pneumonia. Arch Intern Med. 2010;170(4):347–353.

Fisher, RA. The fiducial argument in statistical inference. Annals of Eugenics. 1935 Dec;6(4):391-398.

Fisher RA. The use of multiple measurement in taxonomic problems. Annals of Eugenics. 1936 Sep; 7(2): 179-188.

Kho A, Rotz D, Alrahi K, et al. Utility of commonly captured data from an EHR to identify hospitalized patients at risk for clinical deterioration. *AMIA Annu Symp Proc*. 2007;2007:404-408. Published 2007 Oct 11.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rhee C, Dantes R, Epstein L, et al. Incidence and Trends of Sepsis in US Hospitals Using Clinical vs Claims Data, 2009-2014. *JAMA.* 2017;318(13):1241–1249. doi:10.1001/jama.2017.13836

SAS Institute Inc. 2011. Base SAS® 9.5 Cary, NC: SAS Institute Inc.

Satterthwaite, F. E. An Approximate Distribution of Estimates of Variance Components. Biometrics Bulletin. 1946 Dec;2(6):110-114.

Satterthwaite, F.E. Synthesis of variance. Psychometrika. 1941 Oct;6**:** 309–316.

Swartz C. Recognition of clinical deterioration: a clinical leadership opportunity for nurse executive. J. Nurs Adm. 2013 Jul-Aug;43(7-8): 377-81.

Walker, G. A., & Saw, J.G. The distribution of linear combinations of t-variables. Journal of the American Statistical Association. 1978;73(364):876-878.

Welch, B. L. The generalization of "student's" problem when several different population variances are involved. Biometrika. 1947 Jan;34(1/2): 28-35.

The data analysis for this paper was generated using SAS Software. Copyright, SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary NC, USA.

VITA

Aric D. Schadler

Education

M.S. in Statistics, June 2005, University of Kentucky, Lexington.
B.S. in Mathematics and CIS, May 2001, University of the Cumberlands.

Academic Experience

Director, Biostatistics Support Group, Department of Pediatrics, Sept 2018 – Current.
College of Medicine, University of Kentucky.

Statistician/Data Science Manager, Nov 2015 – Current.
IPOP College of Pharmacy, University of Kentucky.

Business Intelligence Statistician, May 2014 – Nov 2015.
UK HealthCare, University of Kentucky.

Business Intelligence Analyst, Aug 2011 – May 2014.
UK HealthCare, University of Kentucky.

Grant Statistician/Research Faculty, Jul 2010 – Aug 2011.
VA Medical Center, Lexington KY.

Director of SSTARS Center, Jan 2008 – Jul 2010.
SSTARS Center UKIT, University of Kentucky.

Interim Director of SSTARS Center, Jan 2007 – Jan 2008.
SSTARS Center UKIT, University of Kentucky.

Instructor, Aug 2006 – May 2009.
Department of Statistics, University of Kentucky.

Research Statistician, Sep 2005 – Jan 2008.
SSTARS Center UKIT, University of Kentucky.

Instructor, Aug 2003 - May, 2005.
Department of Statistics, University of Kentucky.

Lecturer, Jan 2002 - May 2003.
Mathematics Department, Northern Kentucky University.

Publications

1. Sara A Atyia, PharmD, Frank P Paloucek, PharmD, Allison R Butts, PharmD, Douglas R Oyler, PharmD, Craig A Martin, PharmD, MBA, Aric D Schadler, MS, Aaron M Cook, PharmD, Impact of PhORCAS references on overall application score for postgraduate year 1 pharmacy residency candidates, *American Journal of Health-System Pharmacy*. (2020 Jun) zxaa152, https://doi.org/10.1093/ajhp/zxaa152
2. Kebodeaux, C., Woodyard, J., Kachlic, M., Allen, S., Schadler, A., Vouri, S. Student Pharmacists' Ability to Organize Complex Medication Regimens According to the Universal Medication Schedule. American Journal of Pharmaceutical Education. (2020 Jun). 84(6). https://doi.org/10.5688/ajpe7531
3. Fuller, M., Schadler, A., Cain, J. An Investigation of Prevalence and Predictors of Disengagement and Exhaustion in Pharmacy Students. American Journal of Pharmaceutical Education. (2020 Apr);84(6). DOI: https://doi.org/10.5688/ajpe7945.
4. Wilsey HA, Bailey AM, Schadler A, Davis GA, Nestor M, Pandya K. Comparison of Low- Versus High-Dose Four-Factor Prothrombin Complex Concentrate (4F-PCC) for Factor Xa Inhibitor-Associated Bleeding: A Retrospective Study [published online ahead of print, 2020 Apr 3]. *J Intensive Care Med*. 2020;885066620916706. doi:10.1177/0885066620916706
5. Craker, N. C., Gal, T.J., Wells, L., Schadler, A., Pruden, S., & Aouad, R. K. (2020 Apr). Chronic Opioid Use after Laryngeal Cancer Treatment: A VA Study. *Otolaryngology_Head and Neck Surgery*. https://doi.org/10.1177/0194599820904693
6. Adcock, B. Carpenter, S., Bauer, J., Schadler, A., *et al.* Acute Kidney Injury, Fluid Balance and Risks of Intraventricular Hemorrhage in Premature Infants. *J Perinatol* (2020). https://doi.org/10.1038/s41372-020-0613-5
7. Winstead RJ, Waldman G, Autry EB, Evans RA, Schadler A, Kays L, Baz M, Anstead MI, Shafii A, Goetz ME. Outcomes of Lung Transplantation for Cystic Fibrosis in the Setting of Extensively Drug-Resistant Organisms. *Prog Transplant*. 2019 Sep; 29(3):220-224.
8. Blackmer J, Lindahl E, Strahl A, Schadler A, Freeman PR. Regulating gabapentin as a drug of abuse: A survey study of Kentucky community pharmacists. *J Am Pharm Assoc*. 2019 May;59(3):379-382.
9. Swiney M, Moga D, Hatton-Kolpek J, Schadler A, Kebodeaux C. HELP: Health Literacy and Polypharmacy, Exploring the Relationship Between Medication Burden and Health Literacy. *J Am Pharm Assoc.* 2019;59(4): Article 187(e112).
10. Real, K, Fay L., Isaacs, K., Carll-White, A., & Schadler, A. (2018). Using systems theory to examine patient and nurse structures, processes, and outcomes in centralized and decentralized units. *Health Environments Research and Design (HERD) Journal* https://doi.org/10.1177/1937586718763794
11. Talbert J, Schadler A, Freeman P. Rural/Urban Disparities in Pneumococcal Vaccine Service Delivery Amond the Fee-for-Service Medicare Population. Lexington, KY: Rural and Underserved Health Research Center Publications; 2018.
12. Department for Behavioral Health, Developmental and Intellectual Disabilities. Grant Support: Statistician. Kentucky Cabinet for Health and Family Services. Grant #: 3048113571.

13. Stringfellow V, Young M, Yoder K, Schadler A, Shenoi A. Burnout among Pediatric Intensive Care Advanced Practice Providers in the United States. *Critical Care Medicine* 46(1):626 January 2018. Issn Print 0090-3493.
14. Winstead R, Pandya K, Flynn J, Davis G, Sieg A, Guglin M, Schadler A, Evans R. Factor VIIa Administration in Orthotopic Heart Transplant Recipients and its Impact on Thromboembolic Events and Post-Transplant Outcomes. J of Thrombosis and Thrombolysis, 45(3):452-456, Apr 2018.
15. Lawrence J, Yoder K, Schadler A, Shenoi A. Burnout among Nurses in the Pediatric Intensive Care Unit in the United States. *Critical Care Medicine* 46(1):626 January 2018. Issn Print 0090-3493.
16. Ramey W, Lohr KM, Zeltner M, Herrell Postonl H, Johannemann A, Schadler AD, Lenert A. Biological and Targeted Synthetic Dmands' Prior Authorization Time Is Significantly Reduced with Pharmacy Presence in the Rheumatology Clinic [abstract]. *Arthritis Rheumatol.* 2017; 69 (suppl 10). http://acrabstracts.org/abstract/biological-and-targeted-synthetic-dmards-prior-authorization-time-is-significantly-reduced-with-pharmacy-presence-in-the-rheumatology-clinic/. Accessed October 13, 2017.
17. Gregory E, Wallace K, Burgess DR, Schadler A, Burgess DS. Impact of an Antimicrobial Stewardship Initiative Focused on Staphylococcus Aureus Bacteremia. *Open Forum Infect Dis.* 2017; 4:S494.
18. Thompson, Z., Gardner, B., Carter, T., Schadler, A., Allen, J., & Bailey, A. Decreasing the Time to Oral Antibiotics in a University Hospital Pediatric Emergency Department. *J of Pediatric Pharmacology and Therapeutics*, http://jppt.org. Volume 22, Issue 4 July-August 2017.
19. Fay, L., Carll-White, A., Schadler, A., Isaacs, K., & Real, K. (2016). Shifting Landscapes: The Impact of Centralized and Decentralized Nursing Station Models on the Efficiency of Care. *Health Environments Research & Design*. Doi: 10.1177/1937586717698812.
20. Brewer, A, Divine, H, & Schadler, A. Patient Awareness, Willingness, and Barriers to Point-of-Care Hepatitis C Screening in Community Pharmacy. J of the American Pharmacists Association, www.japha.org. May 2017.
21. Buckler, LT, Teasdale, C, Turner, M, Schadler, A, Schwieterman, TM, Campbell, CL. "The Patient-Centered Discharge An Electronic Discharge Process is Associated with Improvements in Quality and Patient Satisfaction" *J Healthcare Quality*. Nov 2014.
22. Page, Cecilia K. and Schadler, Aric D. "A Nursing Focus on EMR Usability Enhancing Documentation of Patient Outcomes" *Nursing Clinics of North America*. March 2014. Print.

Abstracts: Peer Reviewed

Slone A, Bauer J, Schadler A, Ballard H. Use of CRRT for neonatal non-cardiac ECLS. *Pediatric Academic Societies*. April 2019.

Sanchayan D, Bauer J, Schadler A, Giannone P. Chorioamnionitis and Retinopathy of Prematurity: An institutional review. *Pediatric Academic Societies*. April 2019.

Murphy M, Bauer J, Huang H, Schadler A, Kiessling S, Chishti A. Impact of obesity on diurnal blood pressure assessment and cardiovascular risk markers in pediatric patients. *Pediatric Academic Societies*. April 2019.

Kebodeaux C, Woodyard J, Kachlic M, Allen S, Schadler A, Vouri SM. Analysis of Student Pharmacists' Errors When Simplifying Complex Medication Regimens. *Am J Pharm Educ*. 2018; 82(5):7158(p.492)

Blust W, Grise W, Hudspeth B, Schadler A, Divine H. Implementing a patient engagement strategy to improve comprehensive medication review completion rates in a chain community pharmacy. *Journal of the American Pharmacists Association,* 2017;57:e1-e142. doi: http://dx.doi.org/10.1016/j.japh.2017.04.011

Kebodeaux C, Sewell K, Schadler A, Beaumont K. Assessing Student Performance in the Medication Use Process Using Community Pharmacy Simulation (MyDispense). Am J Pharm Educ. 2017;81(5):S5(p.55)

Kebodeaux C, Jones M, Schadler A, Vouri SM. Measuring Student Pharmacists' Ability to Simplify Complex Medication Regimens. Am J Pharm Educ. 2017;81(5):S5(p.83)

Brewer A, Hanna C, Eckmann L, Schadler A, Divine H. Patient awareness, willingness, and barriers to point-of-care Hepatitis C screening in community pharmacy. *Journal of the American Pharmacists Association,* 2017;57:e1-e142. doi: http://dx.doi.org/10.1016/j.japh.2017.04.011

Villwock K, Roberts P, Petchimuthu M, Schadler A, Divine H. Impact of a pharmacy technician training program on completion of medication therapy management encounters in a chain community pharmacy. *Journal of the American Pharmacists Association,* 2017;57:e1-e142. doi: http://dx.doi.org/10.1016/j.japh.2017.04.011