USING SURVIVAL ANALYSIS TO INVESTIGATE THE PERSISTENCE OF STUDENTS IN AN INTRODUCTORY INFORMATION TECHNOLOGY COURSE AT METU

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

OĞUZ OZAN KARTAL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF INFORMATION SYSTEMS

SEPTEMBER 2015

USING SURVIVAL ANALYSIS TO INVESTIGATE
THE PERSISTENCE OF STUDENTS IN AN INTRODUCTORY INFORMATION TECHNOLOGY COURSE AT METU

Submitted by Oğuz Ozan KARTAL in partial fulfillment of the requirements for the degree of Master of Science in Information Systems, Middle East Technical University by,

Prof. Dr. Nazife Baykal _____
Director, Informatics Institute

Prof. Dr. Yasemin Yardımcı Çetin _____
Head of Department, Information Systems

Prof. Dr. Yasemin Yardımcı Çetin _____
Supervisor, Information Systems, METU

Assist. Prof. Dr. Tuğba Taşkaya Temizel _____
Co-Supervisor Information Systems, METU

Examining Committee Members

Prof. Dr. Soner Yıldırım _____
CEIT, METU

Prof. Dr. Yasemin Yardımcı Çetin _____
Information Systems, METU

Assoc. Prof. Dr. Aysu Betin Can _____
Information Systems, METU

Assoc. Prof. Dr. Altan Koçyiğit _____
Information Systems, METU

Assoc. Prof. Dr. Tolga Esat Özkurt _____
Medical Informatics, METU

Date: 31.08.2015

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name: Oğuz Ozan Kartal

Signature: _____

ABSTRACT

USING SURVIVAL ANALYSIS TO INVESTIGATE THE PERSISTENCE OF STUDENTS IN AN INTRODUCTORY INFORMATION TECHNOLOGY COURSE AT METU

Kartal, Oğuz Ozan

M.S., Department of Information Systems

Supervisor: Prof. Dr. Yasemin Yardımcı Çetin

Co-Supervisor: Assist. Prof. Dr. Tuğba Taşkaya Temizel

September 2015, 51 pages

The purpose of the study is investigating students' persistence in the IS100 Introduction to Information Technologies and Applications course at the Middle East Technical University (METU). While this course is mandatory for all undergraduate students, its noncredit status allows the students to delay passing the course to upper classes. Hence, they cannot utilize the competencies the course would provide during their studies at METU. In the study, the attendance records of IS100 students are extracted for the fall and spring semesters between the years 2011 to 2013. Survival analysis is used to investigate students' persistence in the course based on different factors including their year, department, and semester information. Their weekly attendance records are used to evaluate their persistence. As a result of the study, it has been found that the survival rates of students do not depend on their year information, except for fourth year students who have a higher survival rate. In addition to this, the survival rate of students does not differ according to students' department information. Except for Spring 2013, the students' survival is similar to each other based on their semester information. In Spring 2013 semester, the students have the lowest survival rates among the terms from Fall 2011 to Spring 2013. We also see that the midterm exams have a negative effect on students' survival. These findings are discussed and suggestions are made.

Keywords: Survival Analysis, Cohort Survival Analysis, Educational Survival, Dropout Rate, Students' Persistence

ÖZ

ODTU'DE BİLGİLENDİRİCİ BİLGİ TEKNOLOJİSİ DERSİNDEKİ ÖĞRENCİLERİN DEVAMLILIGINI İNCELEMEK İÇİN HAYATTA KALMA ANALİZİ KULLANIMI

Kartal, Oğuz Ozan

Yüksek Lisans, Bilişim Sistemleri Bölümü

Tez Yöneticisi: Prof. Dr. Yasemin Yardımcı Çetin

Tez Yardımcı Yöneticisi: Yrd. Prof. Dr. Tuğba Taşkaya Temizel

Eylül 2015, 51 sayfa

Çalışmanın amacı Orta Doğu Teknik Üniversitesi IS100 Bilgi Teknolojileri ve Uygulamalarına Giriş dersindeki öğrenci devamlılığını incelemektir. Bu ders tüm lisans öğrencileri için zorunluyken, kredisiz olması öğrencilere dersi üst sınıflarda geçmeyi ertelemeye izin verir. Bundan dolayı, öğrenciler ODTU'de çalışmaları boyunca, dersin sağladığı kabiliyetlerden faydalanamamaktadır. Çalışmada, IS100 öğrencilerinin 2011-2013 güz ve bahar dönemleri arasındaki katılım kayıtları çıkarılmıştır. Hayatta kalma analizi öğrencilerin devamlılığını, onların yıl, bölüm ve dönem bilgisini içeren farklı faktörlere göre incelemek için kullanılır. Onların haftalık katılım kayıtları, onların devamlılığını değerlendirmek için kullanılır. Çalışmanın sonucunda, öğrencilerin hayatta kalma oranlarının, dördüncü yıl öğrencileri hariç, onların yıl bilgisine bağlı olmadığı bulunmuştur. Buna ek olarak, hayatta kalma oranları, öğrencilerin bölüm bilgisine göre farklılık göstermemektedir. 2013 bahar dönemi hariç, öğrencilerin hayatta kalmaları, dönem bilgisine göre benzerdir. 2013 bahar döneminde, öğrenciler 2011 güz dönemi ve 2013 bahar dönemi arasında en düşük hayatta kalma oranına sahiptirler. Ayrıca, vizelerin öğrencilerin hayatta kalmalarına olumsuz bir etkisi olduğunu gördük. Bulgular tartışıldı ve önerilerde bulunuldu.

Anahtar Kelimeler: Hayatta Kalma Analizi, Toplu Hayatta Kalma Analizi, Eğitimsel Hayatta Kalma, Bırakma Oranı, Öğrencilerin Devamlılığı

This thesis is dedicated:

To the memories of my grandfather Musa Kartal

and

to my family

ACKNOWLEDGMENTS

I would like to thank to my supervisor Prof. Dr. Yasemin Yardımcı Çetin and my co-supervisor Assist. Prof. Dr. Tuğba Taşkaya Temizel for their guidance and support throughout the study.

I would like to thank to my family members Turan Kartal, Günnaz Kartal, Ezgi Kartal, and Özge Kartal for their endless love, support, faith and encouragement. They never stopped believing that I would overcome every difficulty.

I would like to thank to Leyla Koç, Akın Özer and İbrahim Çalışkan for being supportive and making me patient every time.

I would like to thank to my best friends Oğuz Yaralı, Ali Can Onat, Hıdır Yılmaz, Orhan Eroğlu, Musa Kartal and İsmail Vedat Işıldak for their endless support and making me smile among all the difficulties of the study.

I would like to thank to METU assistants Nurcan Alkış and Şeyma Küçüközer for their valuable support and guidance what I confused in the study.

I would like to thank to Gökhan Karatepe, Erdem Işıldar, Fadime Şeyda Yücel, and EG Yazılım company members for their support and making me smile whenever I was tired.

I would also thank to my examining committee members Prof. Dr. Soner Yıldırım, Prof. Dr. Yasemin Yardımcı Çetin, Assoc. Prof. Dr. Aysu Betin Can, Assoc. Prof. Dr. Altan Koçyiğit, and Assoc. Prof. Dr. Tolga Esat Özkurt for their feedbacks and valuable suggestions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

IS100: Introduction to Information Technologies and Applications course
HR: Hazard rate
FAS: Faculty of Arts and Sciences
FEA: Faculty of Economics and Administrative Sciences
FED: Faculty of Education
FEN: Faculty of Engineering

CHAPTER 1

INTRODUCTION

1.1. Background of the Study

"Statistics is the science of learning from data, and of measuring, controlling, and communicating uncertainty; and it thereby provides the navigation essential for controlling the course of scientific and societal advances."(Davidian and Louis, 2012) The learning from data, as specified above, is achieved via analyses by using domain specific methods. The survival analysis, an important branch of statistics, works on the lifetime of a "thing". Its focus may be a machine breakdown, an end of an uptrend/a start of a downtrend of a marketing strategy and a prognosis of a human disease.

"The term survival analysis comes from medical researchers because the methods originate from biomedical interest in studying mortality, or patients' survival times between the time of diagnosis of a certain disease or death. Indeed, the first survival analysis was conducted approximately 350 years ago, when John Graunt derived the very first life table and published his famous paper 'Natural and Political Observations Made Upon The Bills of Mortality' in 1662."(Shenyang Guo, 2009) The term "survival" is from medical research area whereas other analogues are "reliability" in engineering, "duration" in economics, and "event history" in sociology.

The survival analysis has been used in diverse fields from sociology to engineering. Annemette Sorensen and Aage Sorensen (1983) made their study in sociology field. Their objective was the analysis of the timing of the transition to marriage; they aimed to compare the transition among three Norwegian men cohorts from the years between 1939 and 1971. In economics domain; Etzioni, Feuer, Sullivan, Lin, Hu, and Ramsey (1998) study on survival analysis techniques for investigating that these techniques are appropriate for measuring the treatment costs in medical interventions or not. Another field that survival analysis is applied is the medical field. Simsek uses it to examine the impact of the three major breast cancer treatment types (Breast Conserving Surgery (BCS), Mastechomy, Breast Conservation Treatment (BCT), and other), age (groups of 49 and under, 50-64, and 65 and over), and stage at diagnosis on the survival rates of breast cancer patients (Simsek, 2000). Besides comparing the groups, another type of usage of survival analysis is prediction. For example, an application to predict customer churn in the telecommunications industry is made to understand customer churn risk (Junxiang). In another study Electrical Submersible Pump Survival Analysis, Pflueger uses survival analysis in petroleum engineering field to characterize the average "life span" of wells and artificial lift types, to compare production conditions (which ones give better run life), as well as a measure of the performance of a given surveillance and monitoring program (improved surveillance and monitoring should extend run life) (Pflueger, 2011).

In this study, survival analysis is used in the education field for analyzing the students' attendance behavior in the introductory information technologies course at METU. The general focus of

survival analysis in education is to analyze the tendency of students to drop out of the school. For example; Woldehanna, Jones, and Tefera (2006) analyze the determinants of students' persistence and retention in primary education in Ethiopia. This social analysis has important implications for Ethiopian Poverty Reduction Strategy Paper because the results include students' survival according to rural/urban and regional differences children live in, child characteristics, family characteristics, social capital, and community and school factors. In another survival analysis study, Plank, DeLuca, and Estacion (2008) perform their analysis in education field from economical perspective. They state that consequences of a student's dropping out the school include a higher likelihood of unemployment, a greater chance of living below the poverty line and relying on public assistance, more frequent health problems, and increased criminal activity. Additionally, it is important for instructors to examine students' persistence and dropping out because retention is increasingly important both for the student and the university today for economic reasons (Allarcon and Edwards, 2013).

In education field, survival analysis studies focus on;

- comparing of graders' survival rates within the school and the nation; investigating the probabilities of graduation within a specific time (Mortenson, 1999)
- identifying factors that impact a student's ability to persist and graduate (Radcliffe, Huesman, and Kellog, 2006)
- analyzing students' school completion, dropping out rates (Woldehanna, Jones, and Tefera, 2006)
- finding factors that affect students' passing their first-year examination, such as their age group and gender (Bruinsma and Jansen, 2009)
- investigating teachers' retention according to their assignments and workload (Donaldson and Johnson, 2010)

In this study, I aim to analyze the tendency of students' dropping out of IS100- Introduction to Information Technologies and Applications course. The following part includes the purpose of the study and research questions.

1.2. Purpose of the Study and Research Questions

The course investigated in this study is IS100 (IS100 Website) whose aim is to equip students with basic computer skills to enable them to use these skills effectively in their subsequent courses. Therefore, IS100 course program includes lectures about operating systems, hardware, information resources, ethics, document processing, electronic presentations, data analysis and spreadsheets. It is a mandatory non-credit service course. Students are advised to take this course in their first year. However, as it is non-credit, some students tend to drop out of the course.

The main purpose of this study is to use survival analysis to analyze students' dropping out the course by using the students' weekly attendance data to investigate their survivals according to factors of student's year, faculty (aka school), and semester.

We formulated the following hypotheses:
H1: The survival rate of students in IS100 course is independent of their year information.

H2: The survival rate of fourth year students in IS100 course is higher than survival rate of other years' students
H3: The survival rate of students in IS100 course is independent of their school information.
H4: The survival rate of students in IS100 course does not differ among semesters.
H5: The midterm of IS100 course affects students' attendance behavior.

1.3. Significance of the Study

Having information about the persistence of students will enable the instructors to gain insight about students' needs and behavior. By examining the findings, policymakers and instructors may update their strategy to improve the outcomes of this course. Accordingly, the experience here may be applied to university base to identify factors that impact the student's ability to persist and graduate.

1.4. Definition of Terms

IS100: Introduction to Information Technologies and Applications course.

Survival: The mean of survival arises when someone or something continues to live or exist. It is used here for describing the students' persistence/progression throughout the semester. If a student completes the course which means that the absenteeism record is less than 5, he/she is a survived student. Otherwise, that student is dropped out of the course.

Cohort survival analysis: Survival analysis on groups of people.

Censored data: Censored data is the data which relevant object's event time is not measured. Censoring occurs when a subject (i.e. a patient in medical research, a thing in engineering, a student in educational research) leaves the study before experiences the event (i.e. death in medical research, breakdown in engineering, dropping out of the school in educational research), the research time ends before the event occurs, or the event is experienced before the research time begins. In this study, censoring may occur as the first alternative above. If a student does not drop out of the school until the semester ends, the event "failing the course" is not occurred, so such students' data in the dataset is a censored data.

Hazard ratio: It states the effect of parameter on objects' survival. Small ratio means that the change in the parameter effects the survival, relatively less. The parameter in this study is students' year, school, and semester information. Hazard ratio is defined in Section 2.1.2.

Significance level: It refers to criterion of judgment. While making a decision about validity of a hypothesis, significance level is used as a rejection point (0.05)

Statistical significance:

- $p < 0.05$ means statistically significant result. Denoted as *,
- $p < 0.01$ means statistically highly significant result. Denoted **,

as in Section 4 in this thesis.

Drop-out rate: The rate is calculated for each week in the relevant semester. It is the number of students dropped out of the course for a week denominated by the students who do not drop out of the course until that week.

CHAPTER 2

LITERATURE REVIEW

In this chapter, the literature review is presented. Firstly, the detail about survival analysis is given. Then, the usage of survival analysis in education field is introduced. These are explained in the following sections.

2.1. Survival Analysis

Survival analysis includes the set of methods that analyze the data on the duration until the occurrence of an event. It enables us to compare the effects of different treatments on the lifetime of objects. Figure 1 shows an example of survival analysis. It includes cumulative survivals of objects throughout the timeline.

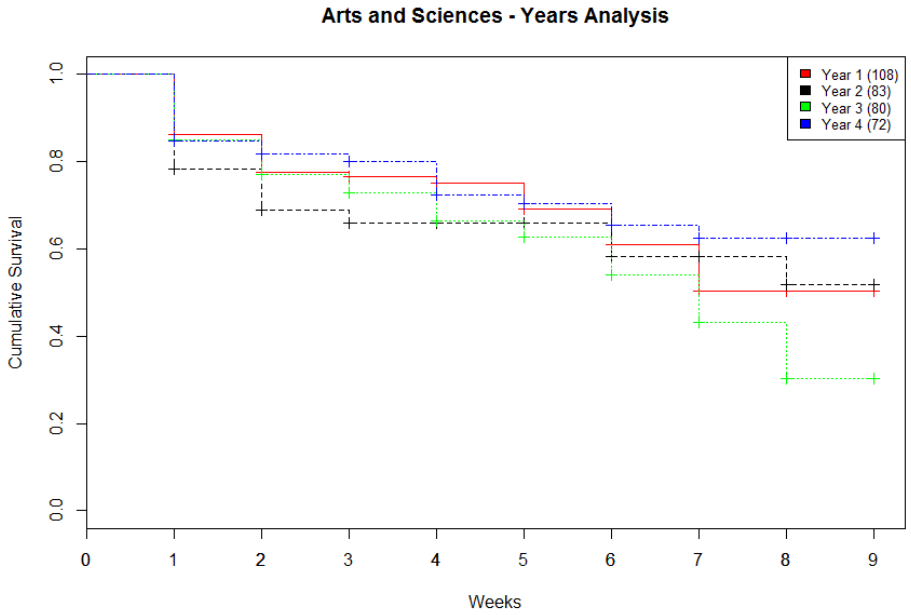

Figure 1: Survival Analysis Figure Including Comparisons of School of Arts and Sciences Students' Survival Rates According to Their Year Information

In Figure 1, the survival analysis of the school of arts and sciences students are displayed. The analysis is done according to their year information. As seen in the figure, fourth year students' survival is highest.

By using survival analysis techniques, researchers can not only extract the cumulative survival statistics of groups and compare the objects but also analyze the relationships between the variables in the data.

These methods are Kaplan-Meier and Cox Regression. The following sections, Section 2.1.1 and 2.1.2, give their details.

2.1.1. Kaplan Meier Method

It is developed by Edward Kaplan and Paul Meier and published in a paper in the Journal of the American Statistical Association in June 1958. Kaplan Meier method is used for extracting cumulative survival rates and so, comparing the groups' survival over time.

$$S(t_i) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right). \quad (1)$$

The formula in Eq. (1) includes:

- i: time
- t_i: time period
- n_i: number of subjects remaining in the cohort in t_i
- d_i: number of subjects who die in t_i

In each time interval, survival probability is the number of subjects living at the start of the interval divided by number of patients at risk (censored patients are not included in denominator because they are accepted as lost). The cumulative probability is to multiply all probabilities for a specific time with its preceding time's probabilities. For example, in a clinic which has 50 patients for the same disease, there are 2 deaths in first month and 3 deaths in second month.
Probability of survivors at the end of the first month is 1-(2/48) = 0.9583 and it is (1-(3/45)) x 0.9583 = 0.8943 for the second month.

2.1.2. Cox Regression

It is also called Proportional Hazards Regression. The Cox model is developed by David Cox and published in a paper in the Royal Statistical Society (Cox, 1972). In 2009, Walters states that a Cox model is a statistical technique for exploring the relationship between the survival of a patient and several explanatory variables. In survival analysis, Cox model is used in estimating the effect of parameters.

$$h_i(t) = h_0(t)e^{\beta_1 x_{i1} + \cdots + \beta_k x_{ik}}. \quad (2)$$

The formula in Eq. (2) includes:

- k: number of coefficients
- β_i: Estimated coefficient for variable i
- x_i: Value of variable i
- $h_i(t)$: hazard function
- $h_0(t)$: null hypothesis

Explanatory variables in our dataset are year, school, and semester variables.

$$HR = \frac{h_1(t)}{h_2(t)} = \frac{h_0(t)e^{\beta x_1}}{h_0(t)e^{\beta x_2}} = e^{\beta(x_1 - x_2)}. \quad (3)$$

By using hazard ratio formula in Eq. (3), the hazard rate describing the difference between the effects of two variables is found. For example, the result of the HR formula including x_2= 2, x_1 = 1, and β value for year = 0.089 means that the effect of year variable (for first and second year) on students' survival is $e^{0.089(2-1)}$ = 1.093, which is the estimated hazard or risk of drop-out increases by 9.3% if a student becomes a second year student.

2.2. Related Studies

Survival analysis has been applied in educational field from student, school and national perspectives. The results have been important for both policymakers and governments of the countries where the studies were conducted. The remaining part of this chapter is devoted to the studies so far in the educational field.

In 1995, a survival analysis is applied on specially admitted students by Han and Ganges. The aim is to investigate the persistence of students, year which students are most likely to leave, and factors that are effective on dropping out risk. The subjects used in this study are 1639 students from Northern Illinois University in 1986, 1987, 1988 and 1989. The dataset include each student's gender, ethnicity, ACT sub-test and composite scores, high school percentile ranks, cumulative credit hours, GPA, and graduation date. They found that student's gender, ethnicity, ACT groups (two groups including grade between 0-10 and 11-20) and high school rank groups (two groups including grade between 0-49 and 50-100) are most effective decisive attributes in the dataset. The semesters that are riskier are second, third and fourth semesters; and black students are more likely to leave the school in these periods. The student group which has ACT scores between 0-10 has a significant risk to leave school compared to the other group. Similarly, the group with HSR scores between 0-49 has a higher risk to leave school than the other group (Han and Ganges, 1995).

Another survival analysis study (Mortenson, 1999) compared the persistence of the students of Kentucky public high schools to that of the USA. The study aimed to examine the rate at which fall 9th grade students in Kentucky's public high schools persisted in their enrollment to fall 10^{th}, 11^{th} and 12^{th} grade enrollments. The data consists of enrollment and regular high school graduate data. It includes all students of Kentucky public high school in period from 1977-78 through 1997-98. Mortenson states that Kentucky ranked behind the nation on the probability of a ninth grader would eventually graduate from high school until the late 1980s. Its rank is nearly identical in years between 1987 and 1991, and it exceeds the nation's rate in years after 1991. "Compared to national data, Kentucky's greatest gains in cohort survival rates have occurred between fall ninth and tenth grade enrollments, and again between fall eleventh and fall twelfth grade enrollments." Tenth to eleventh grade survival is nearly identical to that of the general USA's survival.

The studies specified above compare semesters' data in terms of students' persistence. However, Hoverstad, Sylvester, and Voss applied survival analysis to search student's lifetime for estimating the revenue impact of a student to his university. Their aim is to develop a model to find out the expected monetary value of a student (Hoverstad, Sylvester, and Voss, 2001). Their dataset includes the number of students in each semester; 508 students in first semester, 492 students in second semester, and so on and finally 1 student in 13^{th} semester. The subjects used in this study are the students admitted to the business school at a medium sized private university in the West of the United States from the fall semester of 1993 through the fall semester of 1999. They assume that the tuition per semester is $9000 for a private university. After applying survival analysis, this revenue is multiplied by the prior semester's survival rate.
By this calculation, the expected value of an individual student for the university is $54,370.36 over 13 semesters.

A survival analysis was applied by using family attributes of students besides student characteristics. The study aims to investigate the impact of family attributes on student's persistence/attrition during the primary and secondary school years in Vietnam (Belanger and Liu, 2008). The data includes individual characteristics (school status, duration of schooling, gender, working in past days, birth order), household context (household income level, size of household, number of adults in household, number of school years of household head, gender of household head, household type), and community context (human development level, available of any traditional occupations or handicrafts) to examine the effects of individual and household characteristics on children's dropping out of school. The data is from Vietnam Living Standard Survey (VLSS) of 1992-93 and 1997-98 for 3301 children aged between 6 and 13. Their results include;

- The probability of dropping out of school increases rapidly with age.
- In all age groups, girls have a higher probability of dropping out of the school than boys.
- Children involved in paid work are approximately four times more likely to drop out.
- Birth order is not significant.
- "Children in households where father is the household head are 26% more likely to leave school, compared with those living in households headed by females."
- "Children living in communes with a high human development level have a much higher survival rate than those living in communes with medium or low human development levels." *(Belanger and Liu, 2008)*

Another study aimed to investigate the time students passed their first year examination at a Dutch University (Bruinsma and Jansen, 2009). The period selected for investigation is 24 months from the beginning of the school. In this work, the dataset includes the students' school, gender, age, prior achievement (measured in terms of grade point average in secondary education) and procrastination which is a measured by a set of questionnaire about the student's motivation. The dataset includes 565 first-year students from the University of Groningen (Netherlands) in 1999. They found that 61% of the students passed their first year examination within 24 months; female students passed this earlier than male students; and students which have higher prior achievement passed their examination sooner. According to their investigation, school and gender are not significant variables. Students which have a lower tendency to procrastinate have higher chance of passing their first year examination within 12 months.

Student attrition was also investigated in Vilnius Pedagogical University (Leonavicius and Grazvydas, 2009). Survival analysis was applied to the data including bachelor's degree students in VPU. The aim is to reveal the factors that are effective on student's persistence and attrition. The data includes gender and entrance grades of each student. According to their results, the gender is not a significant factor for students' attrition. In addition to this, survival probabilities are not different among the groups of entrance grades. Important predictors in this work are the

final assessment of Mathematical Analysis in Semesters 1 and 2, and the Theory of Probabilities and Statistics in Semester 4.

Differing from the studies focusing on persistence of students, Leonaviciene and Terese analyzed the migration rate among universities in Lithuania and in Vilnius Pedagogical University. They stated that migration rate is increasing and it is critical to know the reasons (Leonaviciene and Terese, 2009). The aim is to find out the significant factors resulting the migration, and risky semesters. In this study, 578 Mathematics and Informatics students who were entered the VPU in between 2002-2004 are selected. The data consists of gender and entrance grade of each student. Leonaviciene and Terese found that the tendency of girls to persist is twice as that of boys'. The survival rate is the lowest among the students whose entrance marks are low. The persistence of boys and girls who have the lowest grade in their gender is not different.

Another survival analysis concentrating on students' dropping out is applied by Bowers (2010). The aim is to investigate the dropping out rates in between Grade 1 to 12 in United States. In the graduating classes of 2006, 193 students from West Oak and South Pine were selected. The enrollment histories and mean noncumulative GPAs for each grade level of these two cohorts were analyzed. The researcher found that; Grade 7, Grade 8 and 11 are the riskiest grades in students' schooling for dropping out. The students in the lowest GPA group are more likely to drop out of the school.

While most of the studies focus on gender, scores, demographic and parental information of each student, another study investigated the dropping out tendency of students according to their mental disorders (Borges, Icaza, Benjet, Lee, Lane and Breslau, 2011). The study works on the dataset including students who have one of 16 DSM-IV (Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition) mental disorders and their degree on four educational milestones which are primary school completion, high school graduation, college entry, and college termination. The dataset assessed from Composite international Diagnostic Instrument (CIDI) during fieldwork in 2001-2002 in Mexico includes 2362 students. They found that students with impulse control disorders and substance use disorders have higher risk for dropping out of school, secondary school dropping out and college entrance. In addition to this, students with anxiety disorders are associated with lower risk for dropping out, especially at secondary school and primary school.

In 2003, an education reform called Quality Reform was realized in Norway. Survival analysis methods were applied to investigate dropping out and transfer rate in university undergraduate education (Hovdhaugen, 2011). The dataset includes each student's parent's education level, gender, and geographical background. It includes two cohorts of first-time entrants to higher education among all students who entered undergraduate program of humanities, social science and science in 1999 and 2003 autumn terms. As a result, Hovdhaugen stated that there is a significant reduction in transfer rate after the reform, but the dropping out rate is not significantly affected by the reform, which is that there were 18730 students who dropped out of the school and 11485 students who were transferred to another school before the reform while there were 16554 students who dropped out of the school and 10681 students who transferred to another

school after the reform. Additionally, parental education level affects student persistence positively.

Alarcon and Edwards investigated students' survival from ability and motivation perspectives. They aimed to find out decisive factors of college students' retention in their first-year (Alarcon and Edwards, 2013). The data consists of each student's ACT test score (ability); conscientiousness, affectivity or retention (motivation); parent's education, and gender. It includes 584 freshman students enrolled to a Midwestern university in Unites States. As a consequence; gender, ACT scores, and conscientiousness are significant predictors of students' university retention, but parent's education is not a significant predictor.

The studies which apply survival analysis in education field generally use dataset including student's gender, parental education, mean cumulative GPAs, ACT scores and so on. However, Zuilkowski and Jukes investigated schooling by relating it with early childhood malaria in Gambia (Zuilkowski and Jukes, 2014). According to their references, malaria is linked to absenteeism in school age because it effects cognitive development negatively. Their aim is to find whether Intermittent Preventive Treatment (IPT) reduces the drop-out rate and changes survival in later grades while cognitive demands of schooling increase. To investigate the impacts of IPT treatment on student's persistence, they worked on the data consisting of 562 youth born between 1981 and 1986. It includes the variables; student's gender, student's grade, whether student's cohort was eligible for government mass treatment program or not, and the student attended government school or madrassa. They state that government school drop-out rate is lower than madrassa's rate; attrition is lower in government school in each grades up to the completion of ninth grade; students' drop-out is highest in third, fifth and sixth grades for madrassa students.

Another study is "Cause analysis of students' dropout rate in higher education study program" (Paura and Arhipova, 2014). The aim is to find out decisive factors of the first year students' dropping out of the school. The dataset consists of each student's gender, secondary school scores, the priority of the program to study, and its financial source (government-financed or self-finance). Priority means that "According to the Latvian enrolment rules all potential students may choose several programs during the application process. Students must indicate the priority for each program separately (first, second, third etc.)" (Paura and Arhipova, 2014). The dataset includes 677 students enrolled in 2011-2012 at the Latvia University of Agriculture. The results indicated that, a male student's dropping out rate is 1.5 times higher than female student's rate; area of study is a decisive factor on survival; the students in the Faculty of Information Technology and Food Technology schools are more risky to drop out of the school; the finance source and priority are not statistically significant variables; and students who have scores lower than 25 are in the most risky score group to drop out.

CHAPTER 3

RESEARCH METHODOLOGY

In this chapter, the research methodology is applied in this thesis is given. The design of the study is explained in detail.

3.1. Design of the Study

The aim of this study is to investigate persistence of students for IS100 course. In order to do that, the dataset includes the absenteeism information of all students on a weekly basis. If a student's total absenteeism reaches 5 weeks, he or she is considered as a student who dropped out of the course. The details about the content of the dataset are given in Section 3.2 Design of the Data.
The survival analysis is applied according to year, school, and semester information of the students. In order to apply survival analysis, R Software is used (see Appendix C). It is an open source software environment for statistical computing and graphics.
In this study, Middle East Technical University students who were enrolled in the course in the fall and spring semesters between 2011 and 2013 are selected.

3.2. Design of the Data

The dataset of this study is provided by Informatics Institute of Middle East Technical University. It includes id, name, surname, class, section, school; final, midterm, and assignment grades; and absenteeism records of each student. Absenteeism records include participation information of the student for 13 weeks in corresponding semester.

This data is then processed for being used in the software. The new version of the dataset consists of the following variables:

- status: the binary variable that shows whether the student dropped out of the course or not. 0 stands for a censored data.
- time: the variable that shows the week when dropping out is realized. All semesters have 13 weeks except for Fall 2013 term. It includes 12 weeks.
- school: the variable shows the school of the student. Students in the dataset are grouped under four schools: FAS, FEA, FED, and FEN.
- year: the variable shows the year of the student. The year in the dataset is from 1 to 4.

The design of the data is decided according to input the software requires. The data are analyzed by using R software and survival analysis is applied to observe the survival difference of student groups according to their year, school (dept), and semester information. Table 1 depicts a sample R software input.

Table 1: The Example of R Software Input

time	status	dept	class
1	1	4	3
1	1	4	3
1	1	4	2
6	0	2	4
2	0	4	3
1	1	2	3
9	1	2	1
6	0	4	2
3	1	1	1

The dept column stands for the school of the student, and

- 1 stands for FAS
- 2 stands for FEA
- 3 stands for FED
- 4 stands for FEN.

Seventh record in Table 1, for instance, corresponds to a first year student who is from FEA and drops out of the course in ninth week. Additionally, fifth record represents a third year student who is from FEN does not drop out of the course and this student results in censored data.

CHAPTER 4

DATA ANALYSIS

All hypotheses of this study are stated in Section 1.2 Purpose of the Study and Research Questions and the following parts display relationships of those and the results of data analysis.

4.1. Survival Rate Analysis of Students According to Their Year Information

In order to investigate the survival rate according to students' year groups - first year, second year, third year, and fourth year – we use their year information in the dataset. In Section 2.2 Related Studies, it was stated that Bowers (2010) has searched for the riskiest grades in students' schooling for dropping out while Han and Ganges have studied on risky semesters (1995). Actually, we assume that fourth year students' survival rate is more than the rates of other years' students, but it is also important for us to search the difference of the persistence of first, second and third year students. Therefore, in this section, we aim to figure out the following hypotheses: the survival rate of students in IS100 course is independent of their year information (H1) and the survival rate of fourth year students in IS100 course is higher than the survival rate of other years' students (H2).
Figure 2 shows their cumulative survival rate throughout weeks.

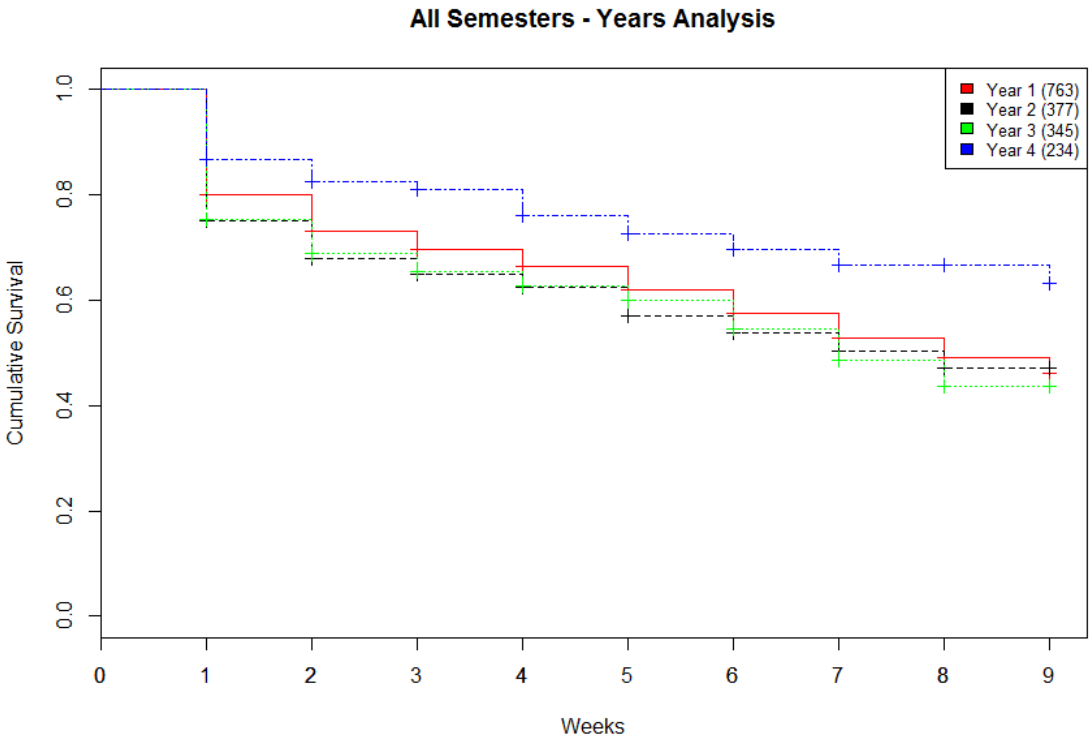

Figure 2: Survival Analysis According to Students' Year

In Figure 2, it is clearly seen that fourth year students' survival is highest among the groups and first, second, and third year have close survival rates throughout the semester.

Table 2: Proportional Hazard Rates, p-values, and Sample Sizes of Year Group Comparisons

	Year 1-2	Year 1-3	Year 1-4	Year 2-3	Year 2-4	Year 3-4
HR	1.116	1.061	0.86	1.01	0.76	0.56
HR (95% C.I)	0.921-1.352	0.961-1.169	0.786-0.940	0.808-1.261	0.658-0.879	0.425-0.761
p-value	0.26	0.24	**0.00094	0.93	**0.0002	**0.000015
N	763-377	763-345	763-234	377-345	377-234	345-234

In Table 3, the statistical information about survival analyses is included. The hazard ratios (effect of the change in the year parameter) between survival of first and second students, first and third year students, and second and third year students are close to one, which means that these students' survival in IS100 course are similar to each other. However, as in the Figure 2 and Table 3**,** we can observe that discrepancy between fourth year students' survival and others is higher and HR values of years 1-4, 2-4, and 3-4 are farther than others according to 1.

The p-values in Table 3 show statistical significance of the results. p-values of survival analysis between first and second year students(1.116, 95% CI, 0.921-1.352, p = 0.26), first and third year students (1.061, 95% CI, 0.961-1.169, p = 0.24), and second and third year students (1.01, 95% CI, 0.808-1.261, p = 0.93) are high, so the difference between first, second and third year students' survival comparisons are not statistically significant. Therefore, they retain the hypothesis that is the survival rates of students in IS100 course are independent of their year information (H1). However, this is not the case for the fourth year students; in survival analysis, fourth years' students' survival analysis compared to others rejects the null hypothesis (H1) because p-values of survival analysis of first and fourth years students (0.86, 0.786-0.940, **p = 0.00094), second and fourth years students (0.76, 95% CI, 0.658-0.879, **p = 0.0002), and third and fourth years students (0.56, 95% CI, 0.425-0.761, **p = 0.000015) are less than our significance criteria (0.05). Therefore, fourth year students retain the hypothesis that the survival rate of fourth year students in IS100 course is higher than survival rate of other years' students (H2).

4.2. Survival Rate Analysis of Students According to Their School Information

Another important parameter of the study is students' school information, which consists of FAS, FEA, FED, and FEN. In Section 2.2 Related Studies, there is a study which finds the most risky school about students' drop-out (Paura and Arhipova, 2014) and it is important also for us to see whether the students of any school are riskier than others between Fall 2011 to Spring 2013 semesters. Therefore, in this section, we aim to investigate the following hypothesis: the survival rate of students in IS100 course is independent of their school information (H3). Figure 3 shows their cumulative survival rate throughout weeks.

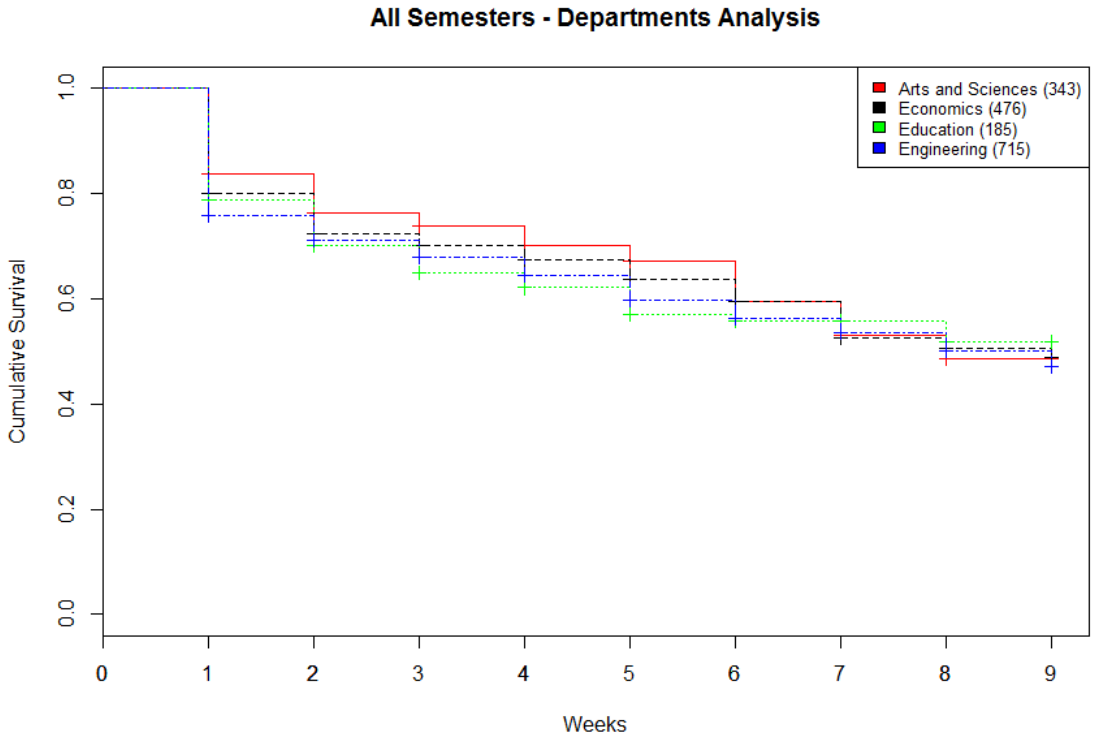

Figure 3: Survival Analysis According to Students' School

It is seen that the persistence of students in IS100 course is similar to each other based on their school information.

Table 3: Proportional Hazard Rates, p-values, and Sample Sizes of School Parameter Combinations

	Arts&Sciences Economics	Arts&Sciences Education	Arts&Sciences Engineering	Economics Education	Economics Engineering	Education Engineering
HR	1.06	1.066	1.043	1.068	1.035	1.004
HR (95% C.I)	0.846-1.328	0.924-1.23	0.973-1.118	0.816-1.397	0.943-1.135	0.778-1.294
p-value	0.612	0.378	0.228	0.629	0.464	0.976
N	343-476	343-185	343-715	476-185	476-715	185-715

It can be seen from Table 4 that the proportional hazard ratios are close to one. Additionally, all p-values in the Table 4 are higher than the 5% significance level (0.05). Therefore, we retain the hypothesis that the survival rate of students in IS100 course is independent of their school information (H3).

4.3. Survival Rate Analysis of Students According to Their Semester Information

The dataset also includes semester information of each student. Survival analysis according to students' semester provides us a general insight about the change of students' attendance to the course throughout years. Therefore, in this section, we aim to investigate the following hypothesis: the survival rate of students in IS100 course does not differ among semesters (H4). Figure 4 shows survival rates of students from Fall 2011 to Spring 2013 semester.

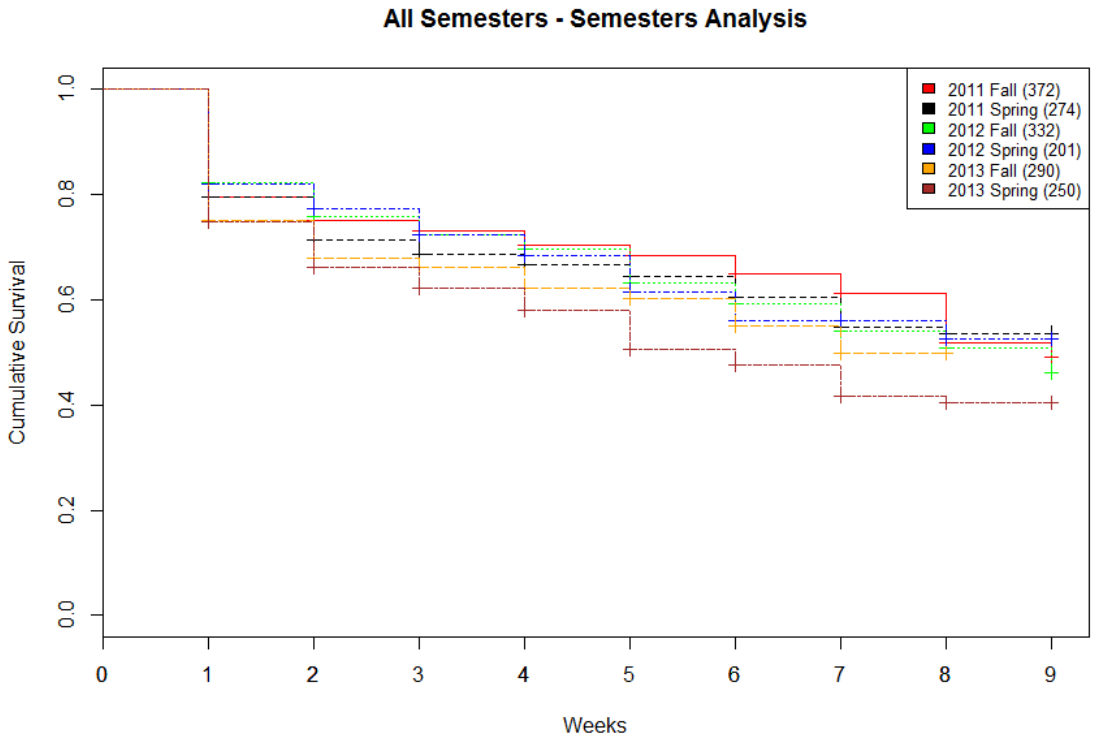

Figure 4: Survival Analysis According to Students' Semester Information

It is seen that students' survival is similar to each other, but the discrepancy between Spring 2013 semester students and others is more pronounced. In the figure, survivals of students of each semester are generally close to each other in each week. Therefore, it is beneficial to separate them into two graphs to make the visualization of the analysis clearer. In addition to Figure 4, Figure 5 and 6 show the survival rates of students of fall and spring semester.

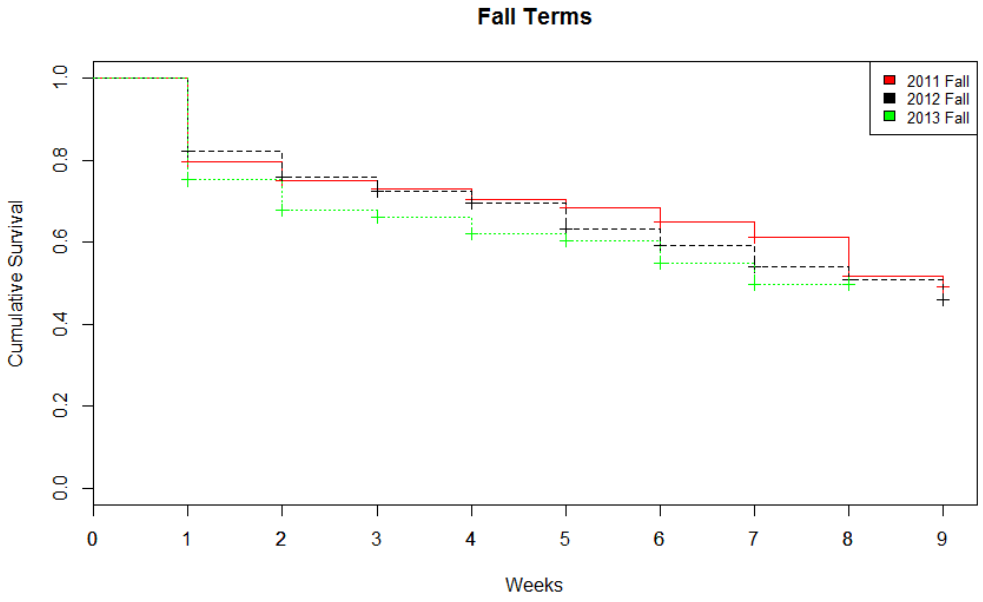

Figure 5: Students' Survival in Fall Terms

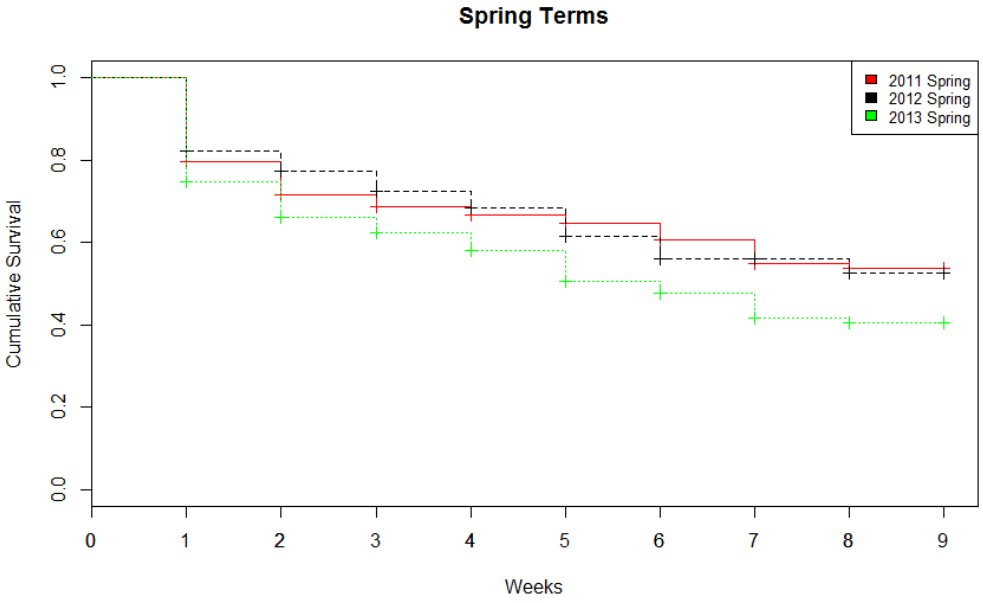

Figure 6: Students' Survival in Spring Terms

Table 4: Proportional Hazard Rates, p-values, and Sample Sizes of Semester Parameter Combinations

	Fall 2011 Spring 2011	Fall 2011 Fall 2012	Fall 2011 Spring 2012	Fall 2011 Fall 2013	Fall 2011 Spring 2013	Spring 2011 Fall 2012	Spring 2011 Spring 2012	Spring 2011 Fall 2013
HR	1.071	1.04	1.02	1.056	1.078	1.014	0.992	1.045
HR (95% C.I)	0.827-1.384	0.920-1.175	0.928-1.121	0.992-1.124	1.026-1.133	0.784-1.312	0.8572-1.15	0.9571-1.14
p-value	0.602	0.529	0.673	0.086	*0.0029	0.914	0.921	0.328
N	372-274	372-332	372-201	372-290	372-250	274-332	274-201	274-290

Table 5: (Cont.) Proportional Hazard Rates, p-values, and Sample Sizes of Semester Parameter Combinations

	Spring 2011 Spring 2013	Fall 2012 Spring 2012	Fall 2012 Fall 2013	Fall 2012 Spring 2013	Spring 2012 Fall 2013	Spring 2012 Spring 2013	Fall 2013 Spring 2013
HR	1.08	0.977	1.08	1.108	1.161	1.179	1.181
HR (95% C.I)	1.011-1.152	0.736-1.297	0.952-1.225	1.02-1.204	0.8706-1.547	1.022-1.36	0.9166-1.522
p-value	*0.021	0.873	0.23	*0.015	0.31	*0.024	0.198
N	274-250	332-201	332-290	332-250	201-290	201-250	290-250

By looking at the proportional hazard ratios, the difference between Spring 2013 and others is a little more than other hazard rates.

p-values in Tables 5 and 6 show that it is retained that the survival rate of students in IS100 course does not differ among semesters (H4) except for Spring 2013 semester. p-values of Fall 2011 and Spring 2013 (1.078, 95% CI, 1.026- 1.133, *p=0.0029), Spring 2011 and Spring 2013 (1.08, 95% CI, 1.011- 1.152, *p=0.021), Fall 2012 and Spring 2013 (1.108, 95% CI, *p=0.015), and Spring 2012 and Spring 2013 (1.179, 95% CI, 1.022- 1.36, *p=0.024) which means students' survival in Spring 2013 is different and it has the least survival rate according to students' semester information.

4.4. Attendance Behavior Investigation Based on Midterm Date

In this study, we also investigate the absenteeism records in all terms to see whether the midterm results have an effect on students' attendance behavior. In this section, we aim to investigate the following hypothesis: the midterm of IS100 course affects students' attendance behavior (H5). It is important to see drop-out rates per week to see whether the midterms have an effect on students' attendance behavior.

Table 6: Drop-out Rates Per Week for All Terms

Term	week 5	week 6	week 7	week 8	week 9	week 10	week 11	week 12	week 13
Fall 2011	20.43	5.06	2.13	2.54	1.49	2.27	1.93	3.55	0.4
Spring 2011	20.43	9.63	3.55	2.1	2.15	3.29	3.97	0.59	0
Fall 2012	17.77	6.95	3.54	2.85	5.88	3.12	3.22	1.42	1.44
Spring 2012	17.91	5.45	5.76	4.08	6.38	4.54	0	1.58	0
Fall 2013	24.82	8.71	2.01	4.1	1.6	3.8	3.38	0	0
Spring 2013	25.2	10.69	4.79	5.03	7.28	2.85	4.41	0.76	0

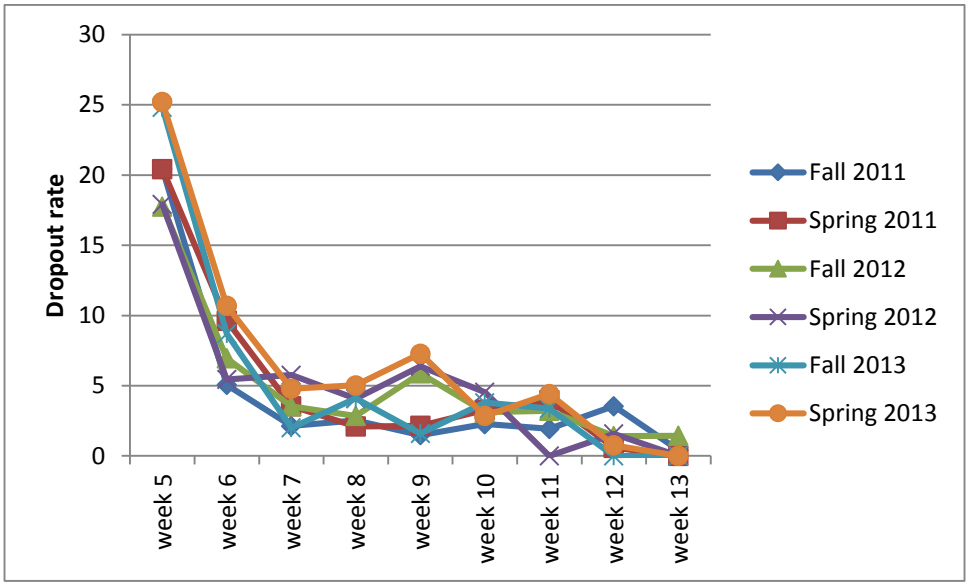

Figure 7: Drop-out Rates per Week In All Terms

The midterm is at the end of 8th week in almost all terms. In Figures 7-13, we see that the drop-out rates until 8th week are decreasing continually, but after that week, they are generally increase or they are steady, so it is seen that midterm results affect students' attendance behavior in a negative way in IS100 course.

Students' drop-out rates in Spring 2011, Fall 2012, Spring 2012, and Spring 2013 are decreasing until 8th weeks. In Fall 2013 semester, the midterm is held at the end of 9th week and the same attendance behavior is also valid for that term. The Figures 8-13 and Tables 8-13 include the drop-out rate in all semesters and display the midterm effects on students' attendance behavior.

Table 7: Drop-out Rates in Fall 2011

Term	week 5	week 6	week 7	week 8	week 9	week 10	week 11	week 12	week 13
Fall 2011	20.43	5.06	2.13	2.54	1.49	2.27	1.93	3.55	0.4

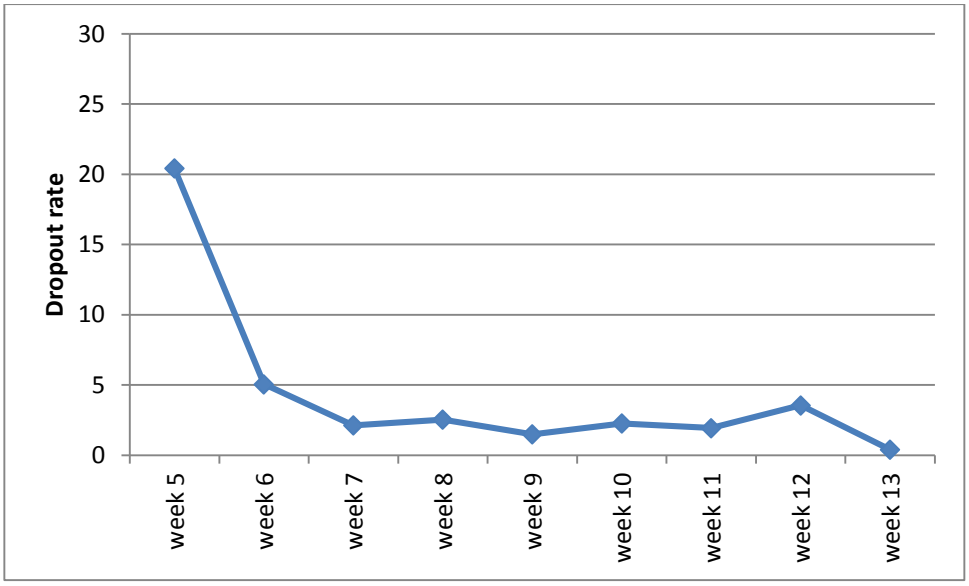

Figure 8: Drop-out Rates in Fall 2011

Table 8: Drop-out Rates in Spring 2011

Term	week 5	week 6	week 7	week 8	week 9	week 10	week 11	week 12	week 13
Spring 2011	20.43	9.63	3.55	2.1	2.15	3.29	3.97	0.59	0

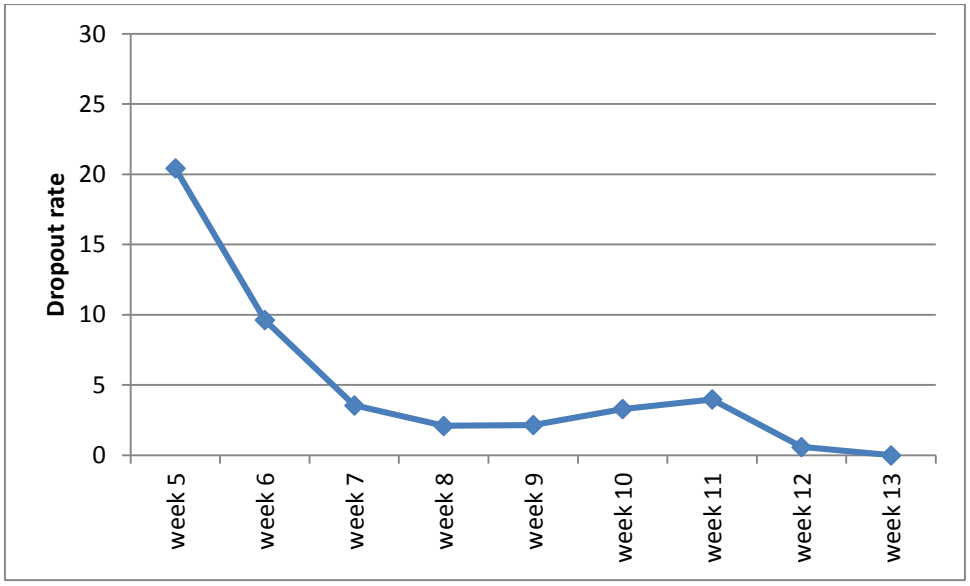

Figure 9: Drop-out Rates in Spring 2011

Table 9: Drop-out Rates in Fall 2012

Term	week 5	week 6	week 7	week 8	week 9	week 10	week 11	week 12	week 13
Fall 2012	17.77	6.95	3.54	2.85	5.88	3.12	3.22	1.42	1.44

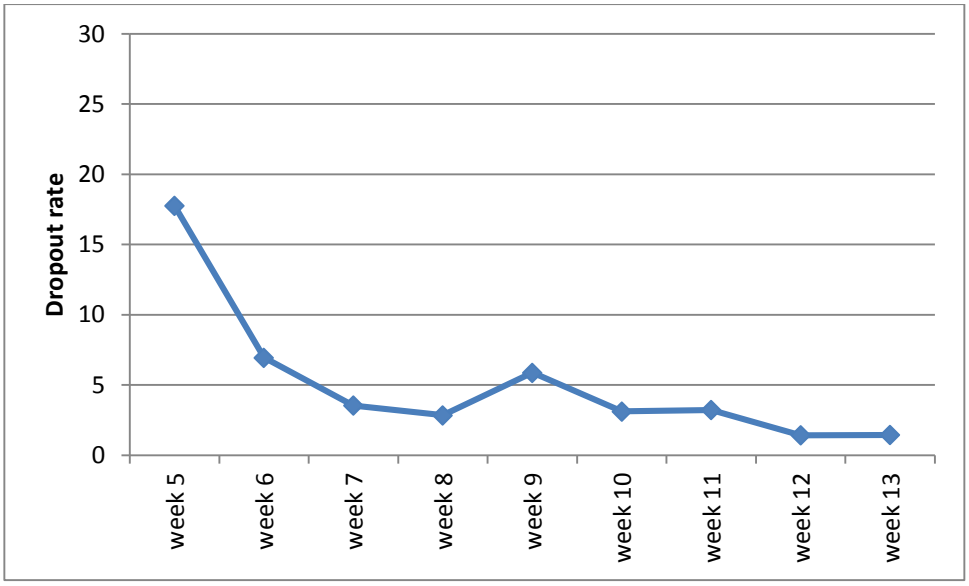

Figure 10: Drop-out Rates in Fall 2012

Table 10: Drop-out Rates in Spring 2012

Term	week 5	week 6	week 7	week 8	week 9	week 10	week 11	week 12	week 13
Spring 2012	17.91	5.45	5.76	4.08	6.38	4.54	0	1.58	0

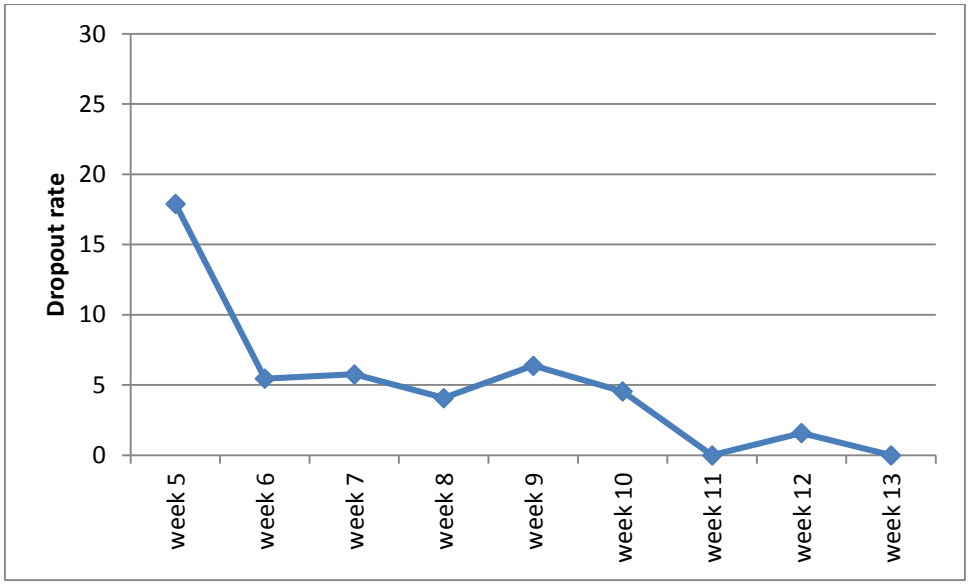

Figure 11: Drop-out Rates in Spring 2012

Table 11: Drop-out Rates in Fall 2013

Term	week 5	week 6	week 7	week 8	week 9	week 10	week 11	week 12	week 13
Fall 2013	24.82	8.71	2.01	4.1	1.6	3.8	3.38	0	0

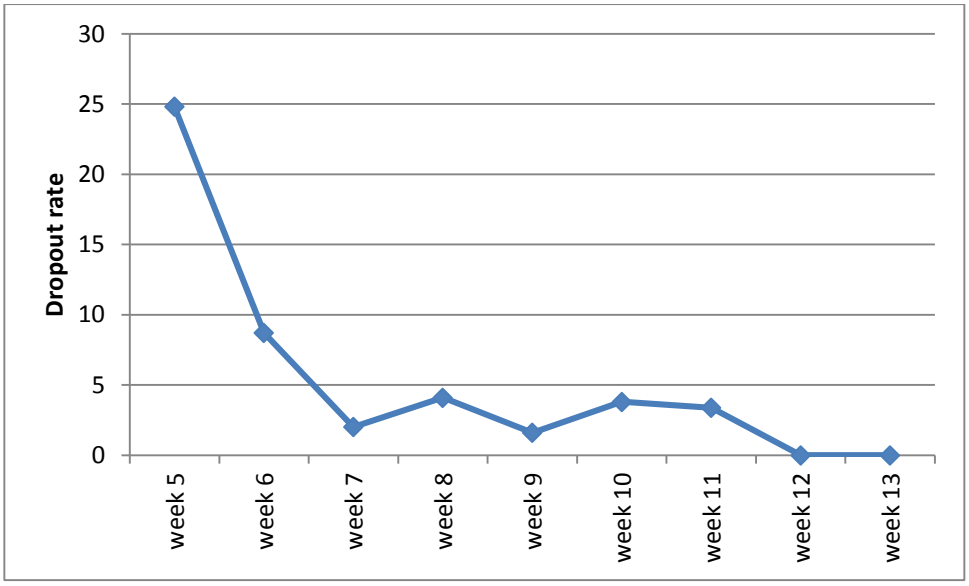

Figure 12: Drop-out Rates in Fall 2013

Table 12: Drop-out Rates in Spring 2013

Term	week 5	week 6	week 7	week 8	week 9	week 10	week 11	week 12	week 13
Spring 2013	25.2	10.69	4.79	5.03	7.28	2.85	4.41	0.76	0

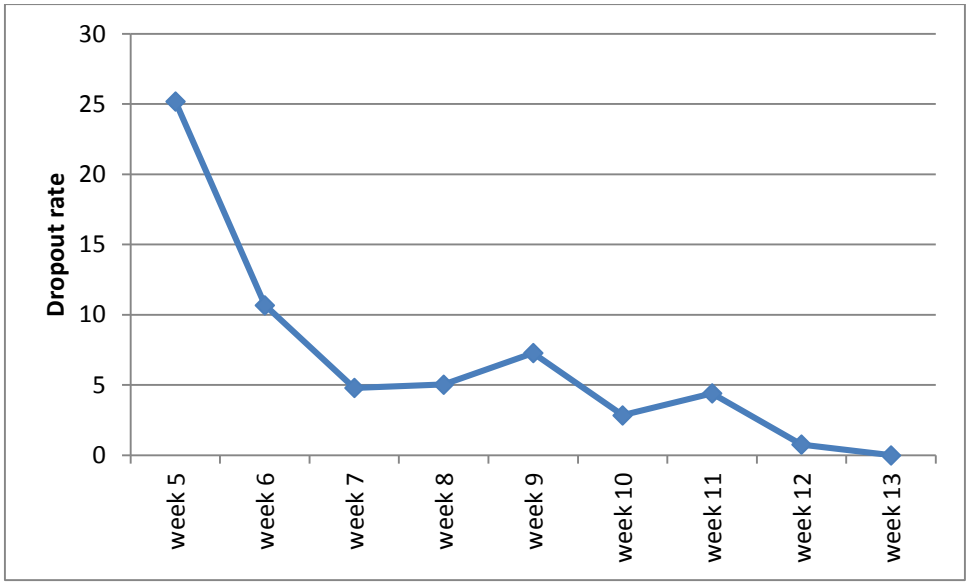

Figure 13: Drop-out Rates in Spring 2013

To see whether the change of attendance behavior just after the midterm week depends on the students' midterm results or not, the students are grouped under two cohorts which are the students who get 0-49 and 50-100. Appendix A includes drop-out rates of two cohorts on a weekly basis in all terms from Fall 2011 to Spring 2013. The results show us that students in midterm groups have the similar attendance behavior with behavior of all students in Figures 7-13. This means that the midterm affects the students' persistence to attend the course and the effect depends on the midterm but not the midterm result.

CHAPTER 5

DISCUSSION AND CONCLUSION

5.1. Discussion and Conclusion

When reviewing the literature, we see that several investigations are made in order to apply survival analysis in education field. These studies generally focus on persistence of students in their education and schooling. They compare the graders' survival rates within the school and nation, identify the factors affecting student's persistence and graduation, analyze students' drop-out rates, and find effects of age and gender of students on survival and so on. In the papers, we see that their results and recommendations are important for the policymakers of schools and governments. These studies and their results all motivate us throughout our study.

In this study, we aim to analyze students' survival for the course IS100. It is a common course for almost all students at METU, so it is worthwhile to conduct an investigation. Before starting the analysis, we acquire our dataset. It includes year, school, semester, and absenteeism records of each student in all terms from Fall 2011 to Spring 2013. In total, our dataset includes 1719 records.

Firstly, we transform our dataset into an acceptable input for R software. We generate the figures that show the cumulative survival rates for each week. Secondly, we find out proportional hazard ratios (i.e. hazard rates between first and second year in years analysis or hazard rates between FAS and FED) between of all factors. Then, in order to test the hypotheses, we calculate the p-values to determine the significance of the results.

There are 9 weeks in the timeline of survival analysis graphs even if a semester includes 13 weeks because there is no drop-out in those times. Therefore, all graphs have the same curve in first 4 weeks and survival rate is on 1.0 cumulative survival point and so, we do not put them in graphs and the timeline is started with week 5. In this study, we evaluate all data including 1 to 13 weeks in data analysis for all semesters.

Students' survival can be affected by a number of factors. First of all, IS100 is a non-credit course and a student can take the course in any year. Secondly, in this course, the students are given basic and fundamental lectures to learn information systems and to use the learned instruments in other courses of the university and in social life. However, the students may think that they have already known the required technology and they see the course as a waste of time. Therefore, it is needed to find out their drop-out rates on weekly basis to see their attendance behavior from the beginning to the end of the week in a term and it is also important to see that being non-credit and having basic supplements for the course have a negative effect on students' motivation to participate to the course, especially for the first and second year students. Additionally, another candidate that students' attendance behavior needs to be investigated is students' school. Students select their

school according to their ability and capability, and also success on university selection exam. It is also crucial to see whether their educational background has an impact on their survival. Finally, it needs to be taken into consideration that how students' attendance behavior changes throughout years. To compare survival rates as years goes on may provide us to see general inclination of students to the course.

When we examine statistics of the analyses in Section 4 Data Analysis and then interpret them, we find the following results.
Survival rates of first, second and third year students are close to each other on weekly basis. Their HR values are close to 1, and the differences of their survival are not statistically significant. On the other hand, fourth year students' survival is different from others, the highest among all year groups. This difference is statistically significant, and it can be clearly seen from the Figure 2. It is seen that a student can postpone passing the course unless he or she is not a fourth year students. A fourth year student has to pass the course in order to graduate from the university, so s/he attends more to the lectures. In addition to this, it is stated above that IS100 course is non-credit and its basic and fundamental structure is not seen critically important by the students (except for fourth year students), this points to the case why first, second, and third year students' survivals are similar to each other.

Different schools need different abilities from the students. However, the analyses show us that students' survival does not depend on their educational background because p values are high on their survival difference investigation. It can be seen from Figure 3 and HR values that their survivals are close to each other, which means that their difference is not statistically significant according to their school information. Students may give more importance to their field specific lectures because they have at least 3-4 credits and these lectures are generally prerequisite course of another lecture, which is that the student has to pass a course to enroll another course in the next term. Therefore, this situation removes the effect of the type of school and its requirements on students. In addition to fundamentals of the course, having high number of credit also affects students' motivation. A high-credit course is more attractive than non-credit course for students. These factors are important to indicate why students from different schools show similar survival for the course.

In addition to work on survival of students, we study on students' attendance behavior on a weekly basis. Therefore, we examine drop-out rates of all students in all terms between 2011 and 2013. A student drops out of the course if he/she does not attend to the lectures as five times. Tables 7-13 and Figures 7-13 show that the fifth week has the highest drop-out rate and the rate decreases up to eight weeks. After that week, the drop-out rate increases or it is steady for a while. This is the week when the midterm is held in IS100 course, and the results are announced in a few days. This means that, midterm results increases the rate of dropping out of the course and not only the students whose midterm result is not good but also the students whose motivation is low drop out the course. In addition to midterm, social proof and social circle are also important factors causing the drop-out of students (Temizel T, Alkış N, 2014). In the paper, it is stated that friends of the students and class environment are leading factors in education context. Students may not persist to attend to the class and this affects others in the same way because of following the lead of

others and this happens mostly at the beginning of the term. After midterm results, the students who continue to attend the lectures have the highest motivation among all students.

Another important case in students' survival investigation is to analyze them according to their semester information. As years goes on, if the survival rate of all students decreases continually, it is critically crucial to be aware of this for the policymakers. Therefore, one of the aims of this study is to see whether this case exists or not. Figure 5-6 show that 2013 terms have the lowest survival rates in both fall and spring terms. Additionally, Spring 2013 has the least survival rate among all semesters and this is statistically significant. This means students' attendance decreases by time; social proof and social circle affects students more, and students give less importance to the lecture as a whole.

The causes what make survival analysis crucial for education field are two major methods: Kaplan-Meier method and Cox regression model. By using cox regression, a researcher can show the relationship between the survival of a student and several explanatory variables. Eventually, the effects of each variable on students' survival are found.

Another important step in survival analysis is Kaplan-Meier method that handles the case of lost students, which results in censored data. Censored data exists when a student leaves the study. If a researcher starts to work on students' real-time data since the beginning of the term and a few students leave the study, s/he removes them from the dataset and s/he makes his/her analyses from the scratch. Campbell at al., (2014) investigated students' behavior week by week. They created the first smartphone sensing system that provided them information including the impact of the students' stress, mood, workload, sociability, sleep and mental well-being on their educational performance. As a result, they presented correlations between sensor data from smartphones and mental well-being and academic performance outcomes. Their results are important for providing awareness of students' behavior during the term and causes of those behaviors. However, by using survival analysis, one could find the effects on depression level on students' survival by finding hazard rate of depression variable (low, moderate, high) on students' behavior by using Cox model, in addition to correlations (i.e. negative correlation between sleep duration and depression). By finding the hazard rates of all factors, the policymakers and other researchers in education field may decide priorities of the results. Moreover, they could have compared the students' attendance behavior for all students' activities, conversation, sleep, and location information gathered during the term by using Kaplan-Meier method. Recall that, this method takes into consideration of censored data so it can be advantageous to use in that study. Because they stated that seven students dropped out of the study and five dropped the class in the term in which they have studied, so they removed their data from their dataset. It is clear that survival analysis can make important contributions to the research above and all researches in education field.

It is seen that first, second and third year students may tend to drop out of the course to pass it in another term(s). What's more, they give more importance to their field specific and credit-course(s). Therefore, firstly, policymakers and instructors are informed with this study. Secondly, students' must be motivated to get it in their first year and they are told in detail that this lecture has basic and important supplements to help them in their other courses and in social life; IS100

course provides them working well in their assignment, projects and challenges in their social life. Additionally, instructors make the students known what are social proof, its effect, and its results. In this way, they reduce the drop-out risk in first four weeks. We see in semester analysis that the survival rate decreases continually in years, so instructors and policymakers may change the structure of the course and/or revise the assignments and topics in the course.

To sum up, we investigate on students' attendance behavior in our study. We compare students' survival according to their information (year, school, and semester) and compare their survivals. Additionally, we examine their schooling week by week by extracting their drop-out rates. This study has an importance to see decisive factors on students' survival, how students' attendance behavior change over time and what reasons have an effect on the results.

5.2. Limitations and Further Research

The study has some limitations. Firstly, our actual sample size is one of the limitations in this study. It is important to have a larger dataset for a more comprehensive research. Actual sample sizes and expected sizes of the study are included in Appendix B. For the further research, more semesters and students can be added to the dataset of the study.

Another one is the demographic parameter diversity. Persistence of students in the course can also be investigated based on students' gender, GPA, entrance grade to the school, parental information, and geographical background. To do so, the data set needs to be enlarged for further studies. Besides the demographic data, smart devices can be used to get real-time data about students' behavior like the study of Campbell at all. (2014).

For the further research, a survey can be conducted to see the reasons of the students for dropping out of the course. They can be asked why they do no attend the course in the first four weeks and why they drop out of the course after the week in which midterm results are announced.

It is also important to observe students' attendance behavior in response to different triggers they receive from their instructors. Methods applied by instructors may affect students' persistence or withdrawal from the course. Different triggers may be mails/messages sent to students at different times and pop-up quizzes. They might affect students' survival throughout the term in different ways. In addition to the survey above, these triggers can be applied in order to observe the students' attendance behavior.

REFERENCES

Alarcon, G.M., Edwards, J.M. (2013). Ability and Motivation: Assessing Individual Factors That Contribute to University Retention, *Journal of Educational Psychology*, 105(1), 129-137.

Belanger, B., Liu, J. (2008). Education and inequalities in rural Vietnam in the 1990s, *Asia Pacific Journal of Education,* 28(1), 51-65.
doi: 10.1080/02188790701845980

Bowers, A. G. (2010). Grades and Graduation: A Longtudinal Risk Perspective to Identify Student Dropouts. *The Journal of Educational Research*, 103, 191-207.

Bruinsma, M., Jansen, E.P.W.A. (2009). When will I succeed in my first-year diploma? Survival analysis in Dutch higher education. *Higher Education Research & Development*, 28(1), 99-114.
doi: 10.1080/07294360802444396.

Cox D.R. (1972). Regression Models and Life-Tables, *Journal of the Royal Statistical Society*, 34(2).
Davidian M., Louis T. (2012). *Why Statistics?*. Washington: Science.
doi: 10.1126/science.1218685
Donaldson, M.L., Johnson, S.M. (2010). The Price of Misassignment: The Role of Teaching Assignments in Teach For America Teacher's Exit From Low-Income Schools and the Teaching Profession, *Educational Evaluation and Policy Analysis*, 32(2), 299-323.
doi: 10.3102/0162373710367680

Etzioni, R. D., Feuer, E.J., Sullivan S.D., Lin, D., Hu, C., Ramsey, S.D. (1999). On the use of survival analysis techniques to estimate medical care cost, *Journal of Health Economics*, 18, 365-380.

Hovdhaugen, E. (2011). Do structured study programmes lead to lower rates of dropout and student transfer from university? *Irish Educational Studies*, 30(2), 237-251.

Borges, G., Medina, M.I.M.E., Benjet, C., Lee, S., Lane, M., Breslau, J. (2011). Influence of mental disorders on school dropout in Mexico. *Rev Panam Salud Publica*, 30(5), 477-83.

Hoverstad, R., Sylvester, R., Voss, K.E. (2001). The Expected Monetary Value of a Student A Model and Example, *Journal of Marketing for Higher Education*, 10(4).

IS100 Website. *Course Information*. Retrieved August, 2014, from http://ii.metu.edu.tr/course-information

Leonaviciene, T. (2009). Analysis of the study process using survival analysis methods. *Pedagogika Studies*, 93, 98-110.

Leonavicius, G. (2009). Research into bachelor's degree studies of informatics at VPU using survival analysis methods. *Pedagogika Studies*, 94, 95-98.

Lu, J. Predicting Customer Churn in the Telecommunications Industry — An Application of Survival Analysis Modeling Using SAS, Sprint Communications Company Overland Park, Kansas, 114(27).
Mortenson, T.G. (1999). Kentucky Public High School Cohort Survival Analysis, *Prichard Committee for Academic Excellence*

Paura, L., Arhipova, I. (2014). *Cause Analysis of student's dropout rate in higher education study program*. Paper presented at 2nd World Conference on Business, Economics and Management.
doi: 10.1016/j.sbspro.2013.12.625

Pflueger, M. (2011). Electrical Submersible Pump Survival Analysis, *Department of Statistics, Texas A&M, College Station.*

Shenyang, G. (2009). *Survival Analysis*. New York: Oxford University Press.

Simsek, F. (2000). Five Year Survival Analysis of Patients with Clinical Stages I and IIA Breast Cancer who Received Initial Treatment at North Carolina Hospitals, *Center for Health Informatics and Statistics*, 123.

Sorensen A., Sorensen A.B. (1983). An event history analysis of the process of entry into first marriage. *Harvard University,* 83(26).
Temizel T., Alkış N. (2014). How to Persuade Students for Active Participation in Course Activities? : A Qualitative Study, Poster session presented at the meeting of the 9th International Conference on Persuasive Technology, Padova, Italy.

Tianqi, H., Ganges, T.W. (1995, April). *A Discrete-Time Survival Analysis of the Education Path of Specially Admitted Students*. Paper presented at the Annual Meeting of the American Educational Research Association, San Frascisco, CA.

Wang R., Chen F., Chen Z., Li T., Harari G., Tignor S., Zhou X., Ben-Zeev D., and Campbell T. (2014) . StudentLife: Asessing Mental Health, Academic Performance and Behavioral Trends of College Students using Smartphones, *In Proc. of Ubicomp.*

Woldehanna, T., Jones, N., Tefera, B. (2006, August). *Children's educational completion rates and dropouts in the context of Ethiopia's national poverty reduction strategy*. Paper presented at the International Association of Agricultural Economists Conference, Gold Coast, Australia.

Zuilkowski, S.S., Jukes, M.C.H. (2014). Early childhood malaria prevention and children's pattern of school leaving in the Gmabia, *British Journal of Educational Psychology*, 84, 483-501.

APPENDICES

Appendix A: Dropout rates of Students in Semesters

Table 13: Drop-out Rates of Students whose Midterm Result is Less Than 50 in All Terms

Term	week 5	week 6	week 7	week 8	week 9	week 10	week 11	week 12	week 13
Fall 2011	47.61	19.48	9.67	8.92	5.88	10.41	4.65	9.75	2.70
Spring 2011	44.06	30.30	15.21	10.25	5.71	15.15	7.14	3.84	0.00
Fall 2012	47.50	30.15	18.18	19.44	27.58	23.80	31.25	9.09	10.00
Spring 2012	41.86	20.00	22.50	16.12	30.76	33.33	0.00	16.66	0.00
Fall 2013	56.19	30.18	10.81	21.21	11.53	21.73	16.66	0.00	0.00
Spring 2013	52.54	35.71	13.04	18.42	22.58	12.50	19.04	0.00	0.00

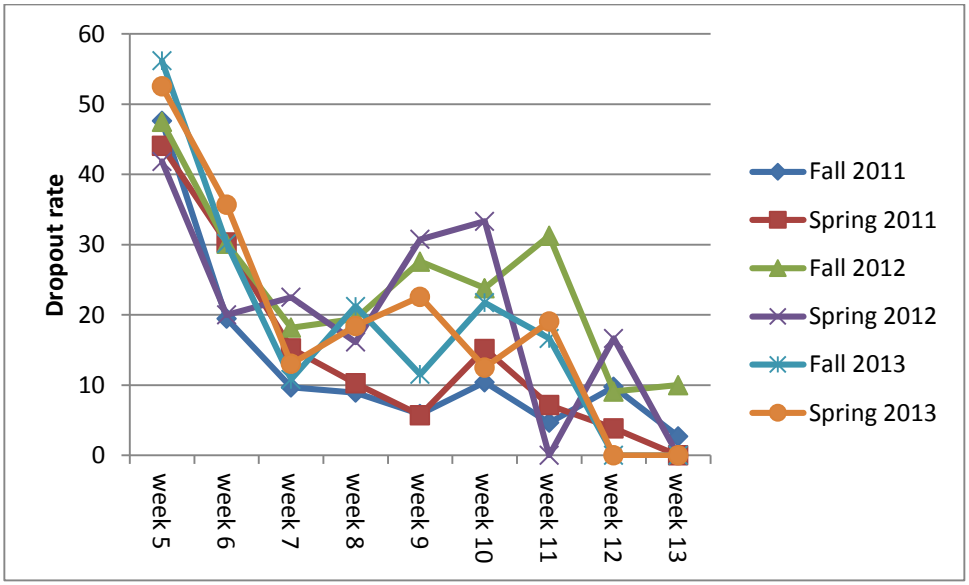

Figure 14: Drop-out Rates of Students whose Midterm Result is Less Than 50 in All Terms

Table 14: Drop-out Rates of Students whose Midterm Results are Higher Than or Equal to 50 in All Terms

Term	week 5	week 6	week 7	week 8	week 9	week 10	week 11	week 12	week 13
Fall 2011	2.66	0	0	0.91	0.46	0.462	1.39	2.35	0
Spring 2011	2.54	0.65	0	0	1.31	0.66	3.35	0	0
Fall 2012	0.93	0	0.47	0	2.85	0.98	0.99	1	1.01
Spring 2012	0.84	0	0	0.84	0.85	0	0	0	0
Fall 2013	2.87	1.77	0	0.602	0.606	1.21	0.61	0.62	0
Spring 2013	0.75	0	1.52	0.77	3.12	0.8	1.62	0.82	0

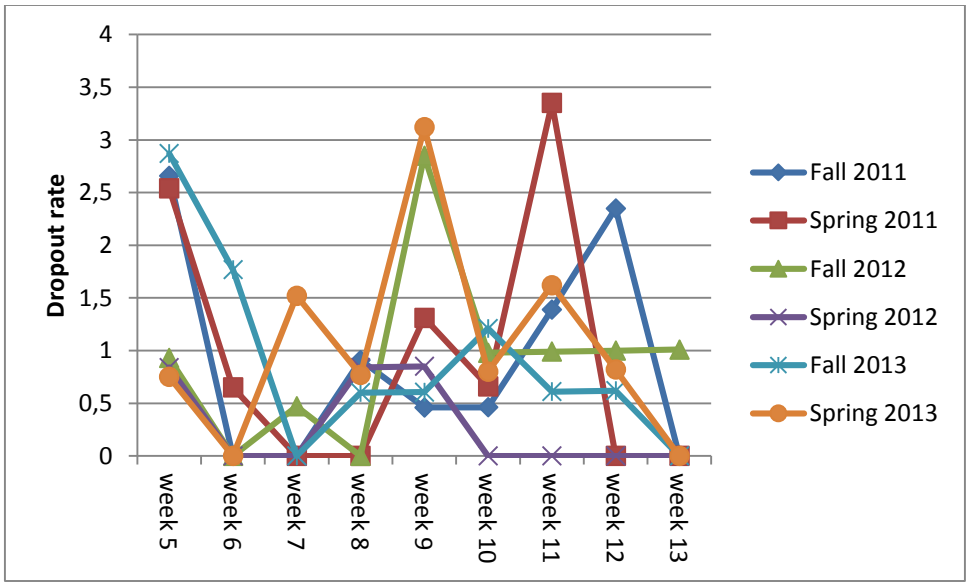

Figure 15: Drop-out Rates of Students whose Midterm Result is Higher Than or Equal to 50 in All Terms

Table 15: Drop-out Rates of Students According to Their Midterm Result in Fall 2011 Semester

Fall 2011	week 5	week 6	week 7	week 8	week 9	week 10	week 11	week 12	week 13
< 50	47.61	19.48	9.67	8.92	5.88	10.41	4.65	9.75	2.70
> 50	2.66	0.00	0.00	0.91	0.46	0.46	1.39	2.35	0.00

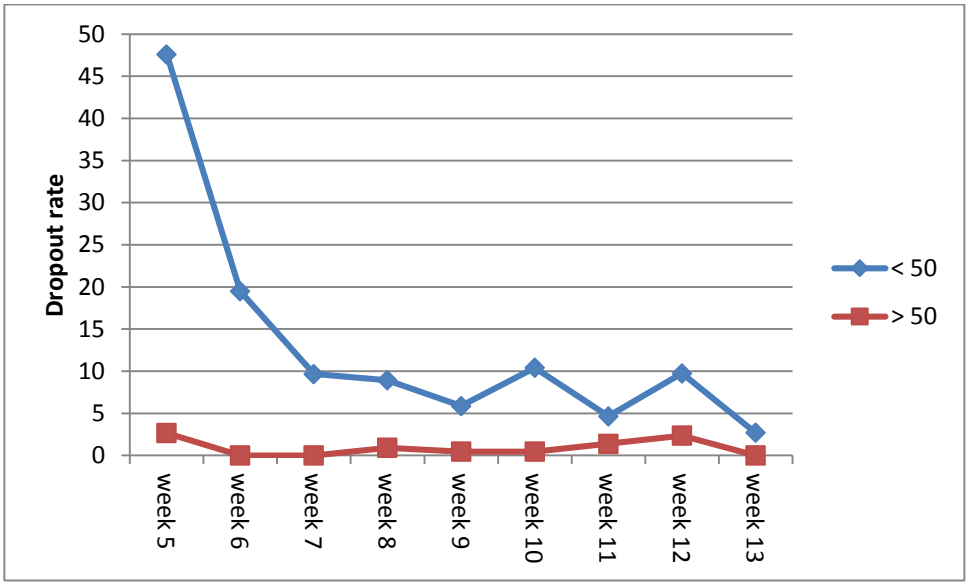

Figure 16: Drop-out Rates of Students According to Their Midterm Result in Fall 2011 Semester

Table 16: Drop-out Rates of Students According to Their Midterm Result in Spring 2011 Semester

Spring 2011	week 5	week 6	week 7	week 8	week 9	week 10	week 11	week 12	week 13
< 50	44.06	30.3	15.21	10.25	5.71	15.15	7.14	3.84	0
> 50	2.54	0.65	0	0	1.31	0.66	3.35	0	0

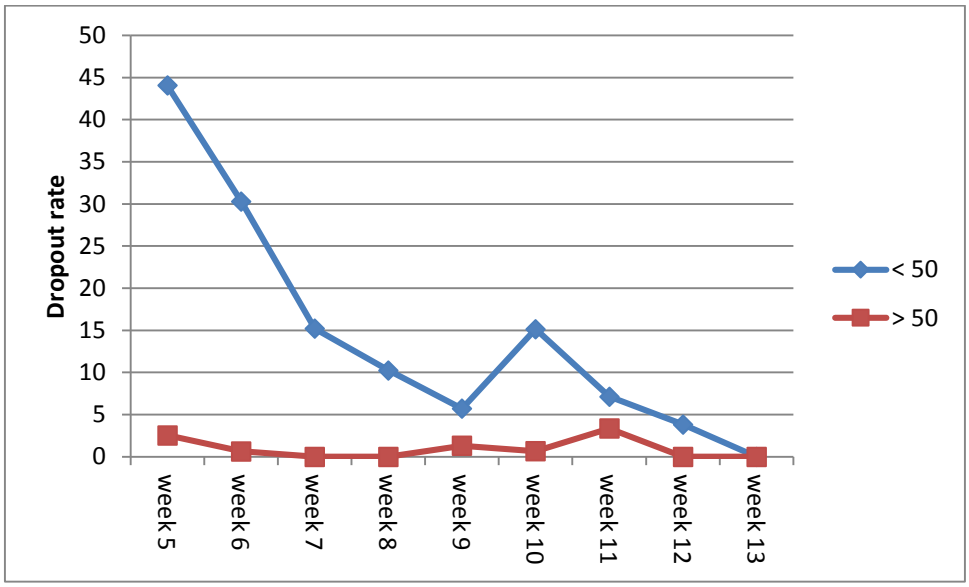

Figure 17: Drop-out Rates of Students According to Their Midterm Result in Spring 2011 Semester

Table 17: Drop-out Rates of Students According to Their Midterm Result in Fall 2012 Semester

Fall 2012	week 5	week 6	week 7	week 8	week 9	week 10	week 11	week 12	week 13
< 50	47.5	30.15	18.18	19.44	27.58	23.8	31.25	9.09	10
> 50	0.93	0	0.47	0	2.85	0.98	0.99	1	1.01

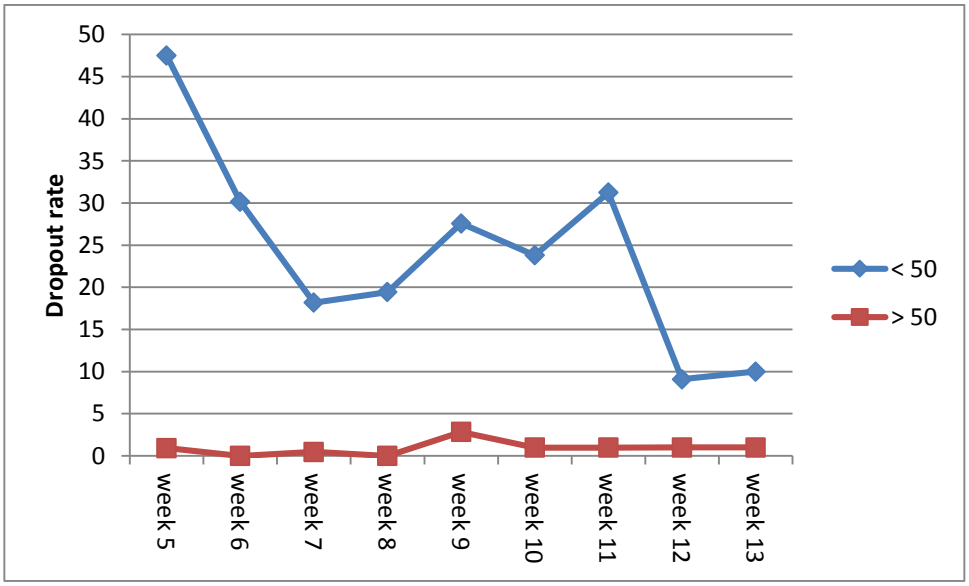

Figure 18: Drop-out Rates of Students According to Their Midterm Result in Fall 2012 Semester

Table 18: Drop-out Rates of Students According to Their Midterm Result in Spring 2012 Semester

Spring 2012	week 5	week 6	week 7	week 8	week 9	week 10	week 11	week 12	week 13
< 50	41.86	20	22.5	16.12	30.76	33.33	0	16.66	0
> 50	0.84	0	0	0.84	0.85	0	0	0	0

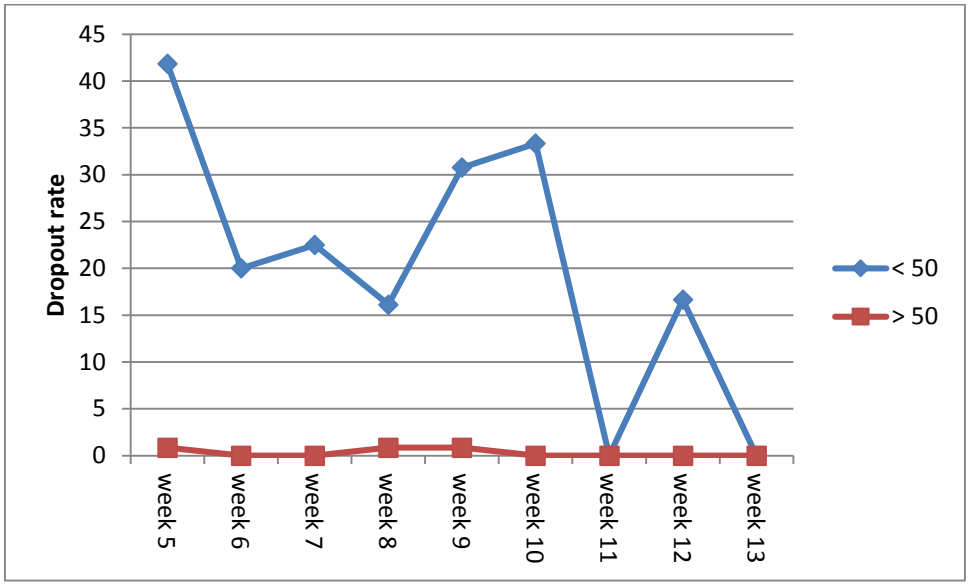

Figure 19: Drop-out Rates of Students According to Their Midterm Result in Spring 2012 Semester

Table 19: Drop-out Rates of Students According to Their Midterm Result in Fall 2013 Semester

Fall 2013	week 5	week 6	week 7	week 8	week 9	week 10	week 11	week 12	week 13
< 50	56.19	30.18	10.81	21.21	11.53	21.73	16.66	0	0
> 50	2.87	1.77	0	0.602	0.606	1.21	0.61	0.62	0

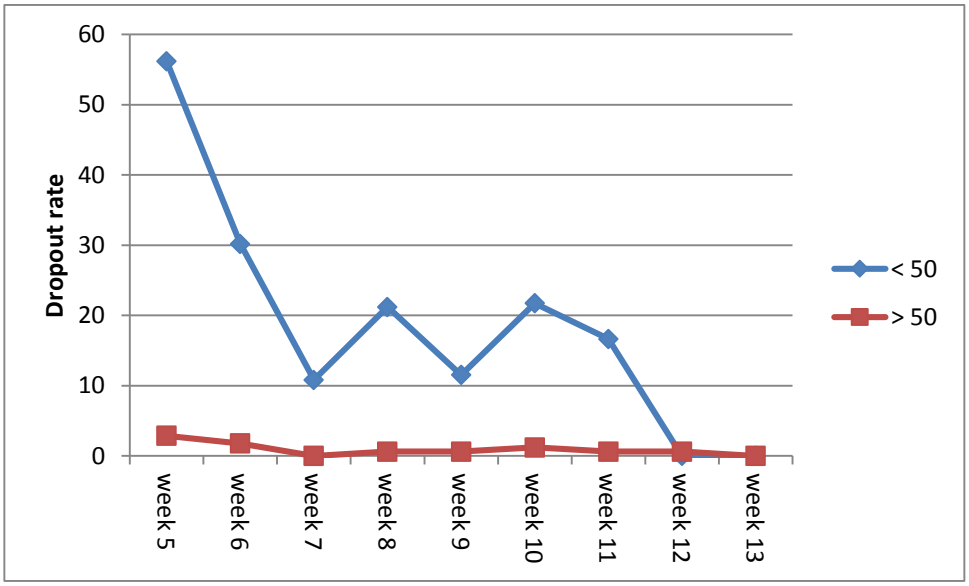

Figure 20: Drop-out Rates of Students According to Their Midterm Result in Fall 2013 Semester

Table 20: Drop-out Rates of Students According to Their Midterm Result in Spring 2013 Semester

Spring 2013	week 5	week 6	week 7	week 8	week 9	week 10	week 11	week 12	week 13
< 50	52.54	35.71	13.04	18.42	22.58	12.5	19.04	0	0
> 50	0.75	0	1.52	0.77	3.12	0.8	1.62	0.82	0

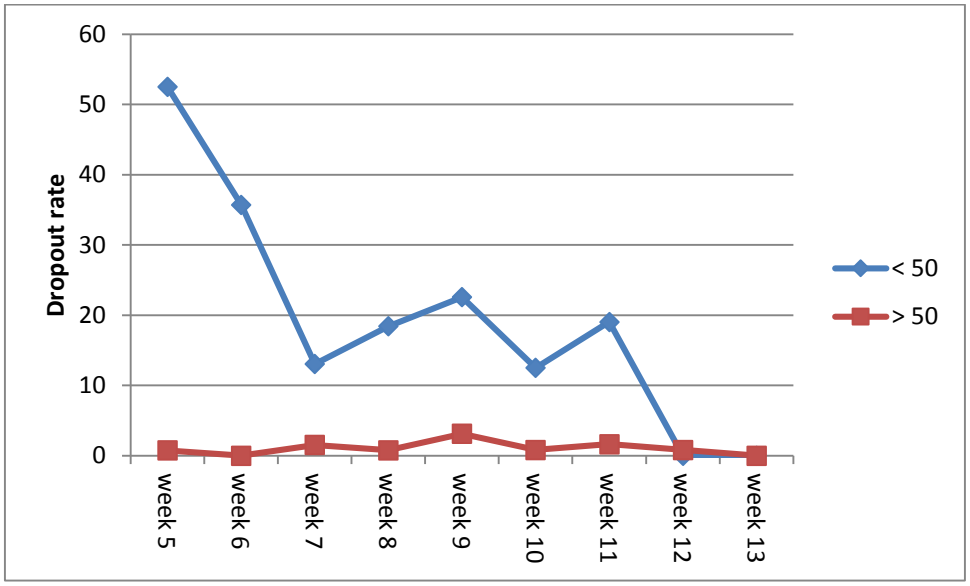

Figure 21: Drop-out Rates of Students According to Their Midterm Result in Spring 2013 Semester

Appendix B: Actual and Expected Sample Sizes that Survival Analysis needs

Year Parameter

Table 21: Data Sample Size of Year Parameter in Dataset

	Year 1	Year 2	Year 3	Year 4
Data Sample Size	763	377	345	234

Table 22: Calculated Sample Size with Given 5% Significance Level and 80% Power

Expected Size	Year 2	Year 3	Year 4
Year 1	3913/1935	14879/6730	4585/1407
Year 2		345140/316062	909/564
Year 3			220/150

School Parameter

Table 23: Data Sample Size of School Parameter in Dataset

	FAS	FEA	FED	FEN
Data Sample Size	343	476	185	715

Table 24: Calculated Sample Size with Given 5% Significance Level and 80% Power

Expected Size	FAS	FED	FEN
FAS	8756/12161	11908/6423	13988/29203
FEA		14139/5498	23770/35744
FED			1353673/5246793

Semester Parameter

Table 25: Data Sample Size of Semester Parameter in Dataset

	Fall 2011	Spring 2011	Fall 2012	Spring 2012	Fall 2013	Spring 2013
Data Sample Size	372	274	332	201	290	250

Table 26: Calculated Sample Size with Given 5% Significance Level and 80% Power

Expected Size	Spring 2011	Fall 2012	Spring 2012	Fall 2013	Spring 2013
Fall 2011	9152/6745	23549/21026	137519/74335	12770/9961	6508/4374
Spring 2011		162552/197032	697823/511976	16661/17649	5182/4728
Fall 2012			93614/56702	5956/5206	3235/2436
Spring 2012				1193/1722	9341161
Fall 2013					1116/962

Appendix C: R Software and Its Methods Used In Survival Analysis

R Software and relevant information can be found in https://www.r-project.org/. It is an open source software including statistical computing methods.

The methods used in this study are Surv, coxph, survfit, and ssizeCT. ssizeCT is in 'powerSurvEpi' package and others are included in 'survival' package.

Surv method is used for creating a survival object.

coxph method is used for fitting proportional hazards regression model. It gives p-values and hazard rate.

survfit method is used for estimating Kaplan-Meier survival curves.

ssizeCT method is used for sample size calculation for the comparison of survival curves between two groups.

'survival' package information can be found at
https://cran.r-project.org/web/packages/survival/survival.pdf

'powerSurvEpi' package can be found at
https://cran.r-project.org/web/packages/powerSurvEpi/powerSurvEpi.pdf