SURVIVAL ANALYSIS OF TIME TO FIRST TERTIARY INSTITUTION ADMISSION AMONGST NIGERIAN YOUTH

BY

OLOYEDE, ABULGANIYU OPEYEMI

17/56EG105

BEING A PROJECT REPORT SUBMITTED TO THE DEPARTMENT STATISTICS IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF BACHELOR OF SCIENCE (HONOURS) DEGREE IN STATISTICS OF THE UNIVERSITY OF ILORIN, ILORIN, NIGERIA.

NOVEMBER, 2022

ATTESTATION

This is to certify that the project work entitled **SURVIVAL ANALYSIS OF TIME TO FIRST TERTIARY INSTITUTION ADMISSION AMONGST NIGERIAN YOUTH** is an original work carried out by **Oloyede, Abdulganiyu Opeyemi**, with the Matriculation number **17/56EG105** under my supervision.

_____	_____
Oloyede, Abdulganiyu Opeyemi	**Dr. A. Abiodun**
17/56EG105	**Supervisor**

CERTIFICATION

This work has been read and approved as meeting the partial requirement for the award of Bachelor of Science Degree (B. Sc.), in Statistics, University of Ilorin, Ilorin, Nigeria.

_____ _____

DR. A. A. ABIODUN Date

SUPERVISOR

_____ _____

PROF. G. M. OYEYEMI Date

HEAD OF DEPARTMENT

_____ _____

Prof. F. B. ADEBOLA Date

EXTERNAL EXAMINER

DEDICATION

This project is dedicated to the one who is always there for me in every sense of being there and in every condition since day zero. The one who doesn't deny me my needs even when I don't deserve them. The one who has been holding it down for me even before my existence.

– my amazing mom

AKNOWLEDGEMENTS

First and foremost, all praise belongs to almighty Allah for his rahmah, idayah, and favours upon me, giving me the wisdom, knowledge and strength to complete this project work. Alhamdulillah.

All profound gratitude goes to my supervisor, Dr. Alfred Abiodun whose fatherly love and guidance helped me through out the research, analysis and compilation of this paper. I appreciate his great sense of direction, time squeezed out of busy schedule to be spent with me and wealth of experience which I enjoyed during the period of this project compilation and beyond. Thank you daddy.

It will all be nought if I don't mention my level adviser, Dr. Adeniyi whose motherly love and care I've been enjoying since my first day in this citadel of learning. I appreciate you ma and by extension all the lecturers in the great department of statistics who I've been fortunate enough to pass through their tutelage.

I appreciate my mum and dad, Mr. and Mrs. Oloyede for all their support, be it financial, moral, and spiritual in making this a success and by extension, I appreciate my siblings: Morenikeji, Moh.Shittu and Amirat.

Aknowledgement wouldn't be complete if I don't mention my roomies, Abdhamid Mubarak and Muftaudeen Muheez.

Omotosho Toyeeb deserves a whole page of acknowledgement, right from day one. My sincere gratitude goes to everyone at Baba-Oyo, Ile-Akolu and your roomies, you guys are great for real.

DSN – who showed me I can be a better data professional, without whose impact I wouldn't have added the data imbalance, train- test split, SMOTE and some other contingent parts of this piece of work, the one society that showed me there's more to data analysis tools than R and SPSS and datafest community, which showed me that the labor market will of me more than proving 7 pages of gaussian Probability Distribution Function, the Central Limit Theorem or the mind wrecking confounding and experimental design.

I say a big thank you to everyone that I benefitted from in one way or the other during the course of this research.

ABSTRACT

In this study, gaining first tertiary institution admission is the event of the survival analysis.

Mean cannot be used to obtain average time in survival analysis. The median survival time is 18 years

To test whether the covariates have any significance in the model, the p-value was compared to the significant value of 0.05, and it was observed that the chance of gaining first tertiary institution admission based on gender is significant. We tested whether the covariates have any significance in the model, the p-value was compared to the significant value of 0.05, and it was observed that the chance of gaining first tertiary institution admission based on living in school covariate, that is the accommodation type while at secondary school is significant.

Albeit, it was observed that type of ownership of secondary school contribute significantly to the model. The hazard ratio for the levels of the opinion covariate "education is the best legacy" in comparison to the baseline hazard (strongly disagree) are different for the cohorts - strongly agree, agree, neutral, and disagree respectively. The hazard ratio for the levels of the opinion covariate "school is scam" in comparison to the baseline hazard that`s strongly disagree are not equal for the cohorts strongly agree, agree, neutral, and disagree respectively.

Table of Contents

List of Tables

List of Figures

CHAPTER ONE

INTRODUCTION

1.1 Background of The Study

Survival analysis is the branch of statistics that deals with the collection of statistical procedures and data analytics, for which the dependent variable is the time until an event occurs (Kleinbaum and Klein, 2005).

The time variable is otherwise known as the survival time because it gives the time that an individual has survived over some follow-up period. The event, also known as failure in some cases may range from death, recovery, relapse, loan repayment, load default, churn a product, patent approval, staff promotion, business failure, readmission of patients, the second child, and even admission to a tertiary institution and much more.

Survival analysis, also known as time-to-event analysis is a branch of statistics for analyzing the expected duration of time until an event occurs. In engineering, the topic is known as reliability engineering while the economist knows it as duration modeling while it is known as event-history-analysis in sociology.

Being called different terms in different fields is proof that time-to-event analysis cuts across many fields like Biology, Medicine, Public health, Engineering, Economics, Sociology, Epidemiology, and intersects lots of topics of which biostatistics is one of them.

A key integral part of survival analysis that can't be overlooked is censored data. Censored data arises when an individual under study did not experience the event, otherwise saying we have some information about the individual survival time but we do not know the exact

survival time. This fortunate or unfortunate occurrence may be caused due to the following reasons:

A unit is lost to follow up during the study period.

A unit does not experience the event before the end of the study

A unit withdraws from the study, maybe because of death or other evitable or inevitable reasons

Censoring could be informative or non-informative, a unit could even be right-censored, left-censored, or interval-censored. (Explanation in a bit)

Informative censoring: simply put, this is a term for when units under study are lost to follow-up due to reasons related to the study. for example, in a study comparing time to lung cancer between smokers and non-smokers, frequent death in either of the group may be telling the analyst something.

Non-informative censoring: sometimes, used as **random censoring**. This occurs when the distribution of the time-to-event random variable provides no information about the distribution of the time-to-censorship. In other words, the time of the censored unit is statistically independent of their failure time.

Right censoring: in here, the true survival time (time-to-event) **T** is greater than the observed survival time **t** (censoring time)**. T > t**

Left censoring: the survival time is below a certain value but it is unknown by how much (Wikipedia). **T < t**

Interval censoring: a data point is somewhere in between two values (Wikipedia). The true survival time is within a known time interval. **t1<T<t2**

1.1.1 Basic goals of survival analysis

To estimate and interpret the survival function and the hazard function.

The hazard function: denoted as **h (t)** is the instantaneous portential of a unit to experience event at a particular time t, given that it did not fail before time t (i.e survival up to time t)

The survival function: sometimes called **the survivor function or reliability function in engineering,** denoted as **s(t)**: it gives the probability that a unit survival time **T** exceeds some specified time **t.**

s (t) = P(T > t) . The s(t) ranges from 0 to 1

To compare survival and hazard functions. The relating equation between survival function and the hazard function goes thus:

S(t) = exp[-∫h(u)du]

h(t) = -[$\frac{d\ S(t)/dt}{S(t)}$]

where: h(t) = ≥ 0

s(t) = 1- h(t)

To assess the relationship of covariates (explanatory variables) to survival time

1.2 Statement of Problem

"Only 25% of the candidates seeking admission into Nigerian tertiary institutions via JAMB secure it nowadays. Worst still, many of these seemingly lucky ones are offered courses other than their choices due to **systemic factors**" – Geoffrey A. Ayua | researchgate.net/publication Some belief the admission system in most tertiary institution in Nigeria is biased based on some characteristics of the applicants. Characteristics like:

- Gender
- Ethnicity
- State of Origin
- Types of ownership of secondary school attended
- And many more factors

While we try to analyze claims such as this, we also intend to use survival analysis to really understand what factors affect admissibility of candidates to Nigerian university.

1.2.1 Age at admission into tertiary institutions in Nigeria

It is common knowledge that the minimum requirements for gaining admission into Nigerian tertiary institutions via the Unified Tertiary Matriculation Board, **UTME** is that candidates must possess at least five (5) credits in O level or its equivalent in not more than two sittings and a minimum JAMB score of around 180-200 for universities depending on the university, 160 for polytechnics and 120 for colleges on an overall aggregated estimated average as at the moment, 2021.

"The Nigerian Senate says it is considering the amendment of the law establishing the Joint Admissions and Matriculation Board to make 16 years the minimum age for candidates writing the Unified Tertiary Matriculation Examination (UTME).

Vice-chairman of the Senate Committee on Basic Education, Akon Eyakenyi said this during the committee's oversight visit to JAMB on Monday.

Eyakenyi noted that candidates below 18 years of age should not be admitted into the university explaining that age has a lot to do with learning ability.

Responding, JAMB registrar, Ishaq Oloyede, told the committee that the board does not have powers to disqualify any candidate on the basis of age, stating that individual institutions can decide on who to admit as in the case with the University of Ibadan which does not admit candidates below 16 years"

_businessday.ng, July 13, 2021.

From the discussion above, it is obvious there is no clear, rigid construct on the issues surrounding age at admission into Nigerian tertiary institutions. Individual institutions are at the liberty of determining such criteria for their entry students.

With the country's transition from the 6-5-4 education system to the 6-3-3-4 education system in 1982 under the leadership of the then federal commissioner for education, commissioner - Wenike Briggs which means for the majority of the pupils who begins at age 5 or 6, it is expected to spend 6 years in primary school, 3 years of junior secondary education, 3 years of senior education or learning a trade in a technical school and 4 years of tertiary institution and yet another transition from the 6-3-3-4 to the 9-3-4 system by the ministry of education pioneered by Dr. Obi Ezekwesili, similar to the 6-3-3-4 system, pupils are expected to spend 9 years in the basic education which is basically a combination of 6 years of primary school and 3 years of junior school education.

The message all these is passing is that ceteris paribus, a candidate seeking admission in any tertiary institution should not be lesser than 17 years of age (5+6+3+3).

By comparison, this age indirectly structured by the education system conforms with ages in most of the other countries in the world.

1.3 Aim and Objective of the Study

This study is aimed at quite some objectives, the following are but a few of them:

1. To determine if the outcome (age at admission) is dependent on every one of the covariates or some of them.
2. To know which of the covariates affects age at which students gain admission
3. To perform general and non-parametric survival exploration the data
4. To compare the semi-parametric survival analysis models to the parametric survival analysis models.

1.4 Scope of Study - Why Survival Analysis?

Some would argue that other statistical methodologies such as chi-square, logistic regression, and multiple linear regressions could have been used or even ensemble for this dataset, research and project, forgetting the fact that they all have different limitations when compared to survival analysis.

Table 1.1: Alternative models, outcome and measure of effect

Model	**Outcome**	**Measure of effect**
Linear regression	Continuous variable	Regression coefficient **β**
Logistics regression	Dichotomous	Odds ratio
Survival analysis	Time-to-event with censoring	Hazard ratio

Survival analysis is preferred to logistics regression as the survival analysis model uses survival time and censoring while logistics regression uses only categorical outcomes, ignoring survival time.

Likewise, we can't possibly use linear regression for this survival analysis dataset because of the following:

- non-normally distributed outcome
- Censoring- linear regression model will not be able to consider censoring
- may or may not be multivariate analysis
- Finally, chi-square test statistics cannot be used here but only as an approximate analysis of the log-rank test statistics.

$\mathbf{X^2} = \sum_i^n \frac{(O_i - E_i)^2}{E_i}$

log-rank test statistics= $\frac{(O_2 - E_2)^2}{var(O_2 - E_2)}$

j = 1,2

n = number at risk

m = number of failures

i = groups; i = 1,2

log-rank test statistics ~ $\chi 2$

1.4 Source of Data

This project is based on primary data collected from 207 respondents all over various parts of the country in no particular order.

Data was collected via a questionnaire carefully designed, administered and distributed online using **google form**

Data collected include the following variables or responses:

- Gender
- Age
- Name of tertiary institution
- Type of tertiary institution
- Number of times you attempted the O-level examination
- Types of ownership of secondary school attended
- You were a … in secondary school
- Type of family structure
- Department at secondary school
- During secondary school, you were based in a …
- Have you ever been admitted into a tertiary institution?
- If yes, at what age were you when you got your first tertiary institution admission?
- Who sponsored your secondary education?
- Which best describes your educational attainment?
- Still in secondary school
- Completed secondary school

- Already in higher institution
- Agreement with the following assertion?

1. “Education is the best legacy”
2. “school is a scam”

CHAPTER TWO

METHODOLOGY

2.0 Introduction

This chapter gives brief explanation of the statistical tests and statistical methodologies used in the paper with each section outlining the methodologies used accordingly.

2.1 Review of survival pattern

This paper introduces survival analysis as a methodology for analyzing the age of entry into higher institutions using data collected from youths in various part of the country via questionnaires.

2.2 Literature Review

The term "survival analysis" refers to a set of statistical approaches for analyzing data that represents the time leading up to a specific event. Death, recuperation, relapse, debt repayment, loan default, churn a product, patent approval, staff promotion, company failure, readmission of patients, second child, and even admission to a tertiary institution are just a few of the events that can occur.

Study units who have not gained admission at the end of the study's observation period are adequately accounted for in the survival analysis, they are called the **censored data**. The study introduces terminologies, methods, and constraints of survival analysis in this paper. Life tables, the Kaplan-Meier product-limit estimate, the log-rank test, and the Cox proportional hazards model are among the approaches presented.

Survival analysis of time from birth till first tertiary institution admission amongst Nigerian youths

- Aim:

The aim of this paper is to look at some of the fundamental concepts and models in survival analysis. Some of which are:

- To perform non-parametric survival analysis of the data
- To estimate and interpret survival function of the data
- To estimate the hazard function of the data
- To compare survivor functions and rates of the covariates
- To compute the median survival times of the covariates
- semi-parametric models, which are commonly used in survival analysis, will also be covered.

The methodology used in this paper is explained in detail below.

2.3 Analysis of survival data

Before proceeding to the survival analysis of our data, we have to understand some things about our data like its structure, proportion, distribution, the missing values in our data, datatype (categorical or continuous) of the variables (covariates), the contribution of every independent variable to the dependent variable, and visualization charts and graphs were employed to better perform the exploratory data analysis.

2.4 Multicollinearity test

The first step in any analysis is to identify multicollinearity. If it is found in the data, we can fix the problem in a number of ways. The first step is to remove the covariate with the multicollinearity specification bias. If the variable has a lot of multicollinearity, it can be removed by converting it. Different variables can be modified by using the first or second. It can be erased by adding new data. Multicollinearity can be reduced in analysis by obtaining the multicollinearity variable's common score. When the Variance Inflation Factor (VIF > 10) is larger than ten, we know there is a multicollinearity problem.

2.5 Descriptive statistics

Descriptive statistics were used to analyze the properties of the dataset such as frequency, percentage, mean, median, variance, and standard deviation of the study covariates.

Survival curve was used to visualize the survival pattern of levels of the categorical variables. In analysis, the survival curve should look somewhat like the one below using step function and measured in probability (0 to 1).

2.6 Kaplan-Meier Survival Function and Hazard Function

Kaplan-Meier survival estimation was used to estimate survival probability and statistical significance for the categorical covariates. The completed questionnaire was wrangled, edited, and coded at data entry right before the survival analysis. This approach of estimating the survival function is **non-parametric**. Non-parametric approaches are straightforward methods that do not rely on any distributional assumptions, like in the case of the survival time distribution found in a study. Non-parametric approaches are great for summarizing

survival data and making simple comparisons, but they can't handle more complicated circumstances as well.

The survival function S(t) focuses on the survival experience of a cohort while hazard function h(t) focuses on the failure rate of the cohort.

There is a clear connection between the two functions, regardless of whether one selects S(t) or h(t). In reality, if one is familiar with the formula, one may obtain the corresponding h(t) from S(t), and vice-versa. It may then be demonstrated that the duo are corresponding via the formulae below.

$$\mathbf{S(t) = exp[-\int_0^t h(u)du]}$$

$$\mathbf{h(t)} = -[\frac{dS(t)/dt}{S(t)}]$$

where h(t) = λ when $\mathbf{S(t) = e^{-\lambda t}}$

The Kaplan-Meier estimator of the survival function is the most used of the duo, whereas the Nelson-Aalen estimator is an alternative, both of them were used in this paper. The features of these survival function estimators are discussed in this study. The **Nelson-Aalen estimator** has a modest advantage in survival function estimation, according to the data. However, when it comes to percentile estimation, the Kaplan-Meier estimator outperforms the Nelson-Aalen estimator for lowering failure rates whereas the Nelson-Aalen estimator outperforms the Kaplan-Meier estimator for increasing failure rates.

2.7 Life Table Method

Of the various data layout (general data layout, life table method, alternative data layout, and others) that exists for survival analysis, the life table data layout was used for the survival function table in this study.

The fraction of subjects who survive, the cumulative hazard function, and the hazard rates of a large group of subjects followed through time are presented in a life table. Subjects who fail as well as those who are censored are both taken into consideration in the study. In terms of survival analysis, the life-table method competes with the Kaplan-Meier product-limit method. Although the life-table approach was initially devised, the Kaplan-Meier method has proven to be superior and is currently the method of choice in this paper.

The survival function table uses the approach of product-limit formula and it consists of seven (7) columns, which are:

- Time
- Number of event
- Number of units at risk
- Cumulative proportion surviving
- Standard error
- Lower confidence interval
- and upper confidence interval

2.8 Hypothesis Testing

The goal is to assess the effect of one or more treatments or exposures (e.g gender) on a primary outcome (here, first tertiary institution admission), after controlling for other covariates. The survival time is the outcome of interest in survival analysis, and one can compare survival times between groups or analyze the connection between exposure and variables and survival time. Standard data analysis methods (e.g., t-tests, linear or logistic regressions) cannot be used to analyze survival data because censoring is not taken into consideration. The results will be biased if censored observations are not included in the analysis.

We would like to know if the estimated survivor functions are statistically equivalent.

First thing first, we set up a hypothesis, null(H_0) and alternate hypothesis (H_1) where H_0 is our claim as the researcher and it says " there is no statistical difference in survival functions (curve) between the levels of covariate "while the H_1 is trying to debunk our claim saying "there is a difference in at least one pair of survival functions“

The significant level used for this study is 0.05

2.9 Log-Rank Test

The log-rank test is used to test the null hypothesis that there is no statistical difference in the likelihood of an event (in this case, admission into the tertiary institution) between the samples at any time point. The analysis is based on the dates of the event (admission). For each time period, we calculate the number of events observed (Oi) in each cohort and the number expected (Ei) if there were no difference between the groups in reality.

The basic idea behind the test is that there will be n_{1i} units in group 1 and n_{2i} units in group 2 at each event time t. Under the null hypothesis, the likelihood that there is no significant effect under the null hypothesis is zero.

The survival time of interest is seen to be right-censored in the data. For this paper, a non-parametric test is proposed, which is a log-rank test for right-censored data.

We wish to study the survival experience of at least two cohorts of individuals, cohort could be based on gender, family structure, and so on.

The log-rank test, which takes into account the entire follow-up period, is the most common way of comparing the survival experience of different cohorts. It is non-parametric, that is it has the significant advantage of requiring us to have no prior knowledge of the shape of the survival curve or the distribution of survival times.

The log-rank test compares the observed and expected number of events for each group using the same test statistic as the chi-squared test. Estimation of the Test Statistic for comparing two groups:

$$\chi 2 = \frac{(\boldsymbol{Oi-Ei})\mathbf{2}}{\boldsymbol{var}(\boldsymbol{Oi-Ei})} \sim \chi^2$$

i=0,1,2,3,…

where,

Oi is the number of observed units

Var(Oi) is the variance of the observed units

the expected number of events (Ei) is calculated as:

2.10 Median Survival Time

The median survival time is a useful summary of a survival curve, it helps us check if approximately 50% of the group under study is predicted to stay longer/ shorter than the median before gaining their first tertiary institution admission.

Because predicting **mean** from substantially right-censored data without strong parametric assumptions is challenging, the **median** survival is employed as a summary statistic far more commonly than the mean.

By first computing the Kaplan–Meier survival curve, the median is computed non-parametrically using right-censored survival curve.

2.11 Nelson-Aalen Cumulative Hazard Rate

Nelson-Aalen is to hazard function what Kaplan-Meier is to survival function.

The Nelson-Aalen cumulative hazard estimate is non-decreasing.

2.12 Cox Proportional Hazard Regression Model

Cox proportional hazard regression model was applied to estimate the effect of factors. A Cox proportional hazards model has the model

$h(t,X) = h_0(t) \times \exp(\sum_{i=1}^{p} \beta_i x_i)$

where,

$h_0(t)$ is the baseline hazard, it involves t but

not the X's

X is a covariate (factor) that contains the predictor variables

and exp($\sum_{i=1}^{p} \beta_i x_i$) is the exponential, it involves X's but not t

in the Cox proportional hazard model, the covariates are assumed to be time-independent. In this, case the baseline hazard is the hazard when the covariate is equal to 0. The proportional hazards assumption of the cox model, which states that the ratio of the hazard function to the baseline hazard, remains constant throughout time, is the most important assumption. Because the exponential function is used, the hazard is always positive.

A Cox proportional hazards model's estimated amount is understood as a relative risk rather than absolute risk. It is assumed that the covariates have an additive influence on the outcome. similar to the logistics model, exp(β1) is the odds (hazard) ratio comparing the increase/decrease in odds for those with a unit change in the covariate.

The Cox proportional hazards model is a popular way to predict an individual's survival based on their baseline data.

This is the most popular method to evaluate the relationship between covariates and survival with the use of a mathematical model. This is called a semi-parametric model because it does not assume any distribution for the baseline hazard. The model is defined as equation.

where λ0(t) is the baseline hazard at time t and x_1,x_2, …, x_k are k independent covariates. No assumptions are made regarding the baseline hazard function.

We can examine the relationship between each of the independent factors and survival time after controlling for other covariates. Understanding the meaning of the parameters in a proportional hazards model is critical. Assuming that the covariate has two values: 0 and 1. Then, exp(λ1) is the hazard ratio for individuals, that is

Hazard Ratio (HR) = exp(βi)

That is the instantaneous probability of an event in one group divided by the probability of the same event in the other.

The hazard ratio is set to be constant throughout time and for all other factors, according to the model. If the covariate is continuous, the hazard ratio for a unit change for the variable is exp(1). Continuous covariates are frequently divided by their standard deviation to make the units for each covariate comparable. Hazard ratios HR = exp(βi)

Cox proportional hazard model is a semi-parametric model with its baseline hazard h0(t) unspecified and this is the property that makes it semi-parametric.

$h(t,X) = h_0(t) \times \exp(\sum_{i=1}^{p} \beta_i x_i)$

where:

$h_0(t)$ is the baseline hazard

X is the time independent explanatory variable

t= time variable

CHAPTER THREE

ANALYSIS

3.1 Data Analysis

This chapter focuses on the collection process, summary statistics, data analysis of the data collected using the Kaplan-Meier survival function, Cox Proportional Hazard Model, Log-rank test and other methodologies explained in the previous chapter.

3.2 Data Collection

This work is based on primary data collected from 207 respondents all over various parts of the country in no particular order.

Data was collected via a questionnaire carefully designed and distributed using **google form**

Data collected include the following variables or responses:

- Gender
- Age
- Name of tertiary institution
- Type of tertiary institution
- Number of times you attempted the O-level examination
- Types of ownership of secondary school attended
- Living in secondary school (day or border)

- Type of family structure
- Department at secondary school
- Base (Rural or urban)
- Have you ever been admitted into a tertiary institution?
- Age at first tertiary institution admission
- Sponsor
- Education attainment

--Agreement with the following assertion using the Likert scale

- “Education is the best legacy”
- “school is scam”

Table 3.1: first five samples of the data collected

survival time	State	Gender	Type of Sec Sch	Living In Sch	Family Structure	Dept at sec sch	Base	Sponsor	Olevel Attempts	Zone	Fathers edu	Mothers edu	Education is the best legacy
20	1	1	3	1	1	1	1	3	1	6	4	4	5
18	1	1	2	2	2	1	1	2	3	6	1	3	2
17	1	2	3	1	1	2	1	4	2	6	2	3	5
17	1	2	3	2	1	2	1	3	1	6	3	3	3
16	1	2	3	1	1	1	2	3	1	6	3	2	3

The **covariates** were reduced from 13 to 7 based on the relative importance

The **covariates** selected are:

1. Gender
2. zone
3. Type of ownership of secondary school
4. Education is the best legacy
5. School is scam
6. Living in school type

7. O-level attempts

Table 3.2: Summary of the categorical variables

Summary of the categorical variables			
S/N	Variable name	Categorical variables	Categories proportion (%)
1.	status	Event - 1: 191 Censored - 0: 16	1: 92 0: 08
2.	gender	Male - 1: 124, Female - 2: 83	1: 60 2: 40
3.	TOSecSch	Private-3: 103 State owned-2: 63 Federal govt owned-1: 41	3: 50 2: 30 1: 20
4.	livingINsch	Day - 1: 151 Boarder- 2: 56	1: 73 2: 27
5.	Family structure	Monogamous- 1: 157 Polygamous - 2: 50	1: 76 2: 24
6.	Deptatsecsch	Sciences- 1: 147 Arts -2: 42 Commercials - 3: 18	1: 71 2: 20 3: 09
7.	Base	Urban -1: 156 Rural -2: 51	1: 75 2: 25

8.	Sponsor	Father - 1: 28 Mother - 2: 24 Both -3: 149 You- 4: 2 Others -5: 4	1: 13 2: 12 3: 72 4: 0.9 5: 02
9.	zone	NC- 1: 82 NE - 2: 3 NW- 3: 7 SE- 4: 11 SS- 5: 12 SW -6: 92	1: 40 2: 02 3: 03 4: 05 5: 06 6: 44
10.	Fathersedu	Tertiary -4: 126 Secondary -3: 48 Primary -2: 26 No education-1: 7	4: 06 3: 23 2: 13 1: 03
11.	Mothersedu	Tertiary - 4: 118 Secondary -3: 65 Primary -2: 18 Illiterate- 1: 6	4: 57 3: 31 2: 08 1: 02
12.	Educationisthebestlegacy	Strongly agree -1: 61 Agree -2: 92	1: 30 2: 44

		Neutral -3: 41 Disagree- 4: 4 Strongly disagree - 5: 9	3: 20 4: 02 5: 04
13.	Schoolisscam	Strongly disagree -5: 67 Disagree - 4: 85 Neutral -3: 40 Agree -2: 13 Strongly agree - 1: 2	5: 32 4: 41 3: 19 2: 06 1: 0.09
14.	Olevel	0ne or lesser attempt - 0: 154 More than one attempts - 1: 53	0: 74 1: 26

The table above shows the categorical variables, the value count of each categories of the categorical variables (the status of event and the covariates) , and its percentage.

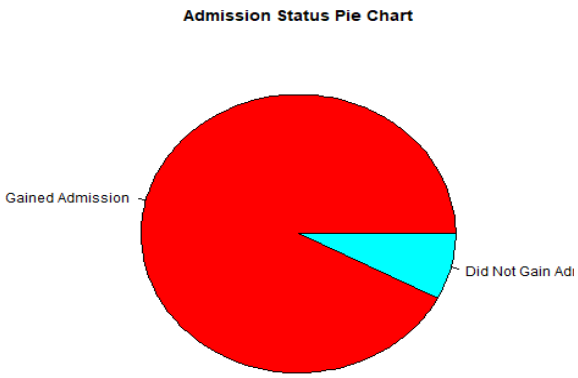

Figure 3.1: Unbalanced Admission Status Pie chart

The figure above shows that the percentage of respondents that gained admission into the tertiary institution (92%) greatly out-weight that of those that did not gain admission into any tertiary institution (8%).

Just as it is expected in almost every Survival analysis datasets, the dependent variable – that is the event class tends to be imbalance, for example - a dataset that contains information about Leukemia patients/time in remission might have 98% of the patients positive and 2% of the patients negative, A dataset for studying Elderly (60+) population/time until death (years) may contain 90% dead and 10% alive by the end of the study for the cohort, time till full loan repayment, time from graduation till employment amongst fresh graduates - in the examples above, we will all agree that the probability of having class imbalance in the datasets to be used for the analysis will be high.

Such an imbalance is what we now experience in our data. Imbalances such as this will have a negative effect on the output of our analysis where the predictive model will perform *too good* on the

majority class -the 92% that gained admission and thereby predict more of gained admission class but very poorly on the minority class – the not gained admission class. To handle this imbalance in the dependent covariate (status) scientifically, we could:

- Synthesize samples of new minority of class instances
- Over-sample the minority class instance
- Under-sample the majority class instance

but we did a combination of the first two approaches, synthesize samples and over-sampling of the minority class instance using a statistical methodology called SMOTE – **Synthetic Minority Oversampling Technique** which we will use to create new synthetic cases based on existing cases of the minority class.

And the result thus:

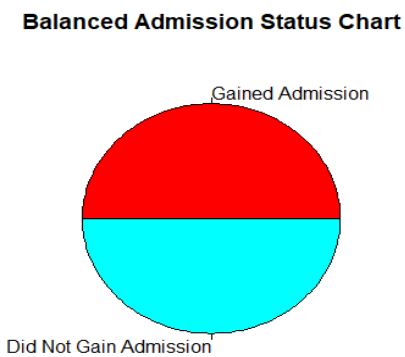

Figure 3.2: Balanced Admission Status Pie Chart

3.3 General Summary Of The Data

Table 3.3: count of data points and variables collected

	Count
Data points	207
Variables	15

Interpretation: the table above shows the number of data points and features of the dataset.

Table 3.4: Count of categorical and discrete variables

Count summary	
Categorical	13
Discrete	2

There are 13 categorical variables and 2 discrete variables in the dataset including the response variables of which there is 1 categorical independent variable and 1 discrete variable of the 15 variables in the data.

3.4 Summary Of The Survival Time

Table 3.5: Summary of the survival time

Minimum (0% quantile)	Median (50% quantile)	Mean	Maximum (100% quantile)
15	18	18.5	28

This is the summary of how long it took the units in the sample to gain admission or censor out.

summary of the survival time, that is time till admission of which 50% is the median survival time.

The minimum time is 15, the maximum time is 28, the median time is 18 while the mean time is 18.5. this is evident that the average is either 18 or 18.5 depending on the average metric being used.

3.5 Contingency Tables

The contingency tables below are of the following format for each cell content:

Table 3.6: Labelling of the contingency tables

Cell Content
N- number of units
table proportion

Table 3.7: The gender – status in a school contingency table

STATUS	GENDER		
	1	2	Status Total
0	9 0.043	7 0.034	16 0.077
1	115 0.556	76 0.367	191 0.923
Gender Total	124 0.599	83 0.401	207 1

Table 3.9: The base – living in a school contingency table

BASE	LIVINGINSCH		
	1	2	Base Total
1	110 0.531	46 0.222	156 0.754
2	41 0.198	10 0.048	51 0.246
Livinginsch total	151 0.729	56 0.271	207 1.00

Table 3.10: The father's education X mother's education contingency table

FATHERSEDUCATION	MOTHERSEDUCATION				
	1	2	3	4	fathersedu total
1	1 0.005	2 0.010	4 0.019	0 0.00	7 0.034
2	2 0.010	7 0.034	10 0.048	7 0.034	26 0.126
3	2 0.010	8 0.039	24 0.116	14 0.068	48 0.232
4	1 0.005	1 0.005	27 0.130	97 0.469	126 0.609
mothers total	6 0.029	8 0.087	65 0.314	118 0.570	207 1.000

Table 3.11: The contingency table between the two opinions ‘education is the best legacy’ and ‘school is scam’

		School Is Scam					
Education Is The Best Legacy		1	2	3	4	5	“Education Is The Best Legacy ” Total
	1	1 0.005	3 0.014	5 0.024	15 0.072	37 0.179	61 0.295
	2	0 0.000	3 0.014	16 0.077	54 0.261	19 0.092	92 0.444
	3	0 0.000	6 0.029	15 0.072	13 0.063	7 0.034	41 0.198
	4	0 0.000	1 0.005	1 0.005	0 0.000	2 0.010	4
	5	1 0.005	0 0.000	3 0.014	3 0.014	2 0.010	9 0.043
	School Is Scam Total	2 0.010	13 0.063	40 0.193	85 0.411	67 0.324	207 1.000

3.6 Survival Table And Graphs

The table below is the life table showing the survival experience of all the units in the sample. It is the overall survival function table with the lower and upper confidence interval.

Table 3.12: Survival Table

Time	**number at risk**	**number of Event**	**Survival rate**	**Standard error**	**lower 95% CI**	**upper 95% CI**
15	207	7	0.9662	0.0126	0.94187	0.9911
16	200	28	0.8309	0.0261	0.78139	0.8836
17	170	47	0.6012	0.0342	0.53781	0.6720
18	122	46	0.3745	0.0339	0.31364	0.4472
19	72	32	0.2081	0.0289	0.15847	0.2732
20	38	15	0.1259	0.0240	0.08662	0.1831
21	20	6	0.0882	0.0212	0.05501	0.1413
22	13	5	0.0542	0.0177	0.02866	0.1027
23	7	1	0.0465	0.0168	0.02295	0.0942
24	5	2	0.0279	0.0143	0.01021	0.0762
25	3	1	0.0186	0.0122	0.00515	0.0672
28	1	1	0.0000	NA	NA	NA

The time column in table 3.12 gives ordered failure times. The failure times are denoted by *t* when writing it in a formula. To get ordered failure times from survival times, we must first remove from the list of unordered survival times all those times that are censored (still haven't gained their first tertiary institution admission) we are thus working only with those

times at which people failed (gained admission). We then order the remaining failure times from smallest to largest, and count ties only once.

The number at risk column in the table is self-explanatory; the number at risk is not a numerical value or count, but rather a collection of individuals who have survived at least to the time; that is, each person included in the 'number at risk' has a survival time that is the corresponding figure in the 'time' column or longer, regardless of whether the person failed (gained admission) or is censored.

The survival rate column is obtained from the survival function, which is the likelihood that a person will live longer than some defined period, i.e., survival function S(t) offers the probability that the random variable T will live longer than the stated time t. The survivor function is essential in a survival study because calculating survival probabilities for various values of t offers critical summary information from survival data.

To illustrate, $S(15) = P(T > 15)$

3.6.1 Kaplan-Meier Survival Curves And Log-Rank Test

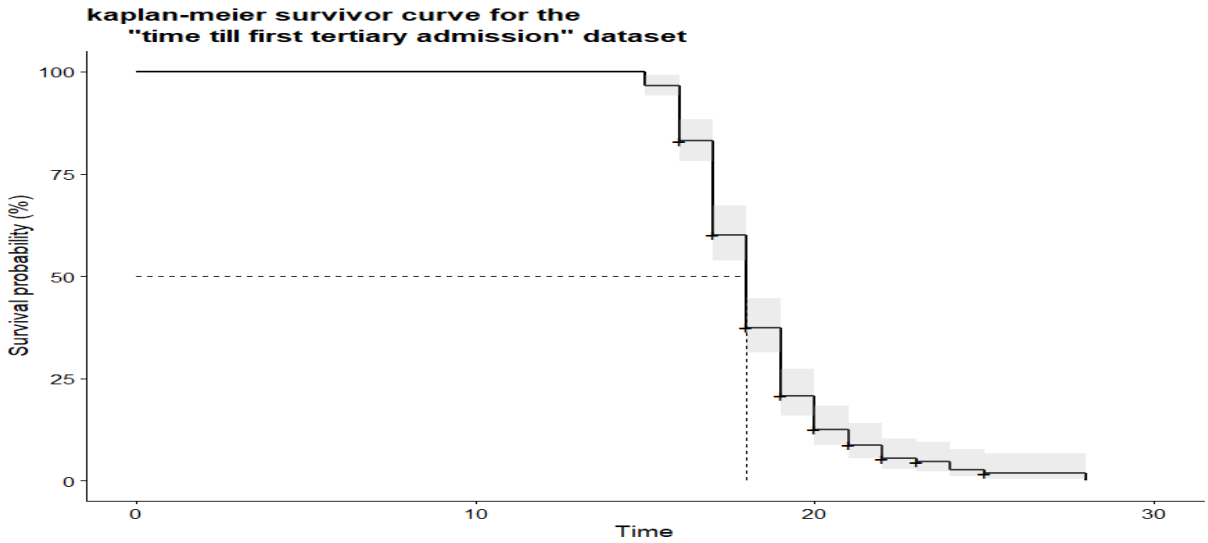

Figure 3.3 : survival curve for the time till first tertiary institution admission

From the Kaplan Meier curve, we can observe the survival step function which is basically indicating to us that the survival rate is reducing with respect to time. We can also observe the median survival time which is **18.** This is saying, on average, 18 is the age people gain their first tertiary institution admission.

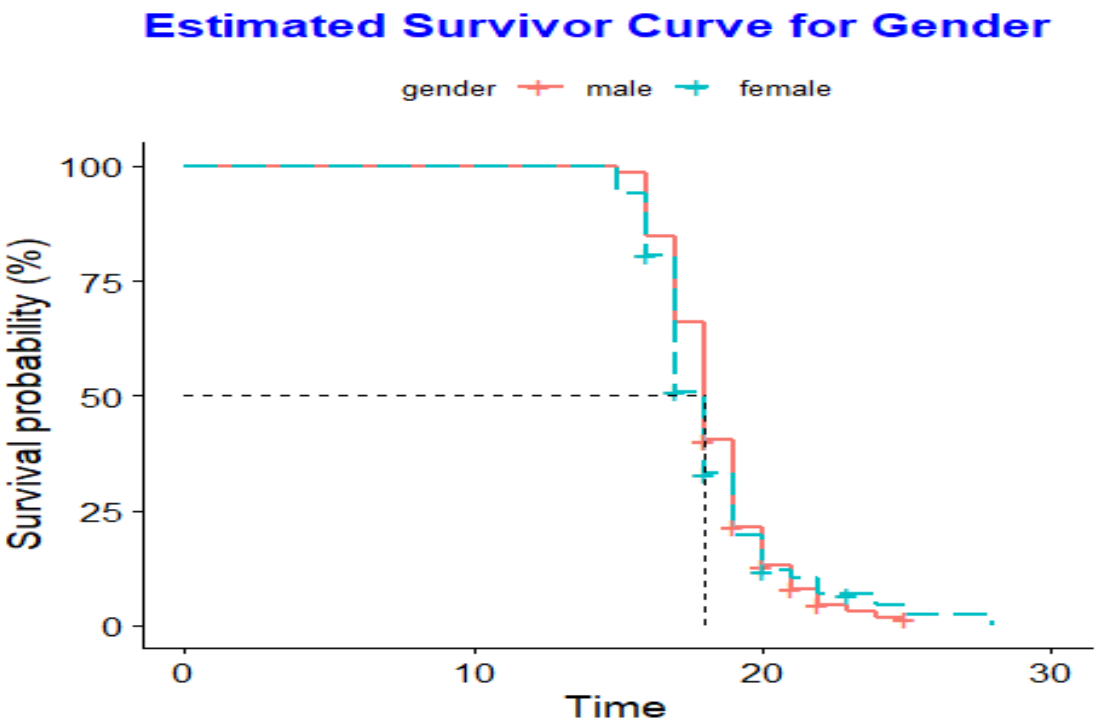

Figure 3.4: Survival curve based on gender.

The survival curve above shows the difference in the survival experience between the male and female in the sample. A test of significance of this difference is provided by the log-rank test.

Hypothesis Testing

H0: there is no difference in survival functions (curve) between the two levelsof gender

H1: difference in at least one pair of survival function.

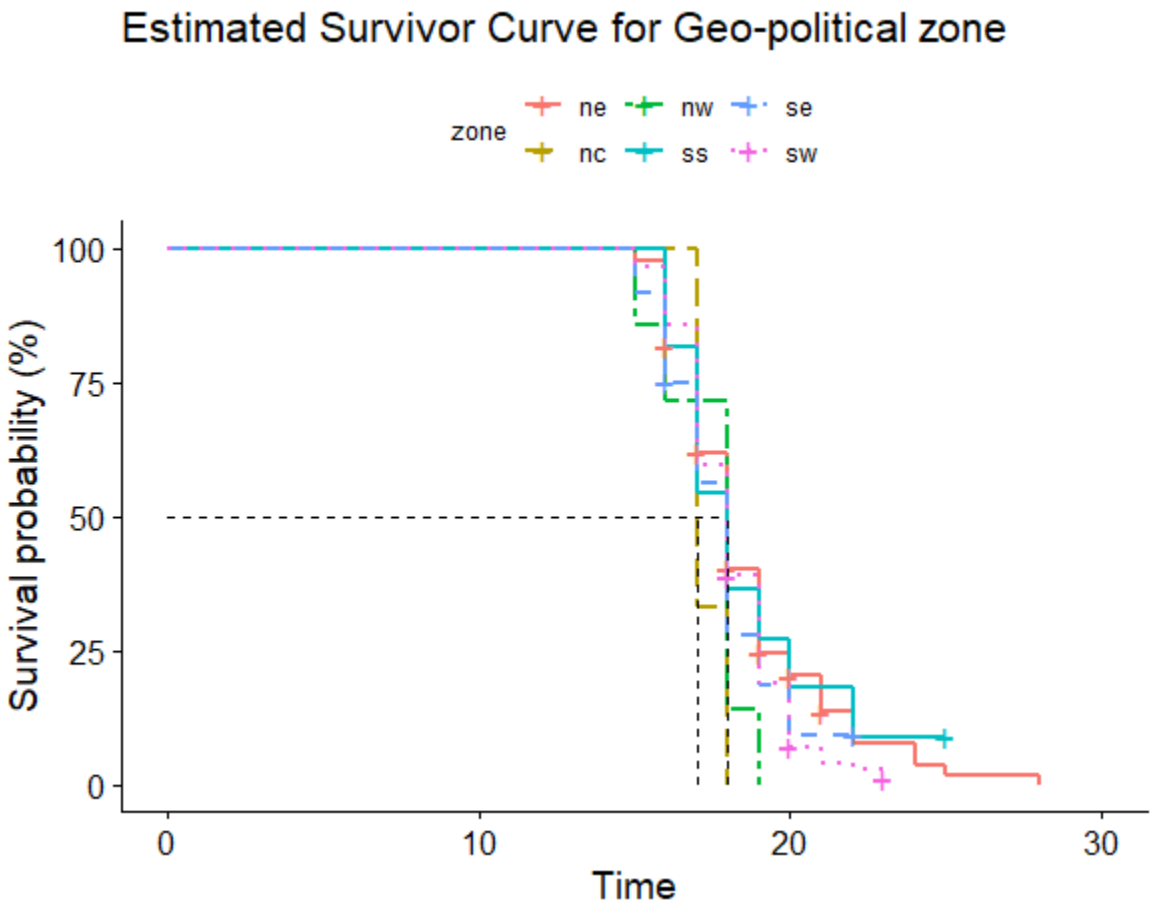

Figure 3.5: Survival curve based on geopolitical zone.

The survival curve above shows the difference in the survival experience between the various geopolitical zones in Nigeria in the sample. A test of significance of this difference is provided by the log-rank test.

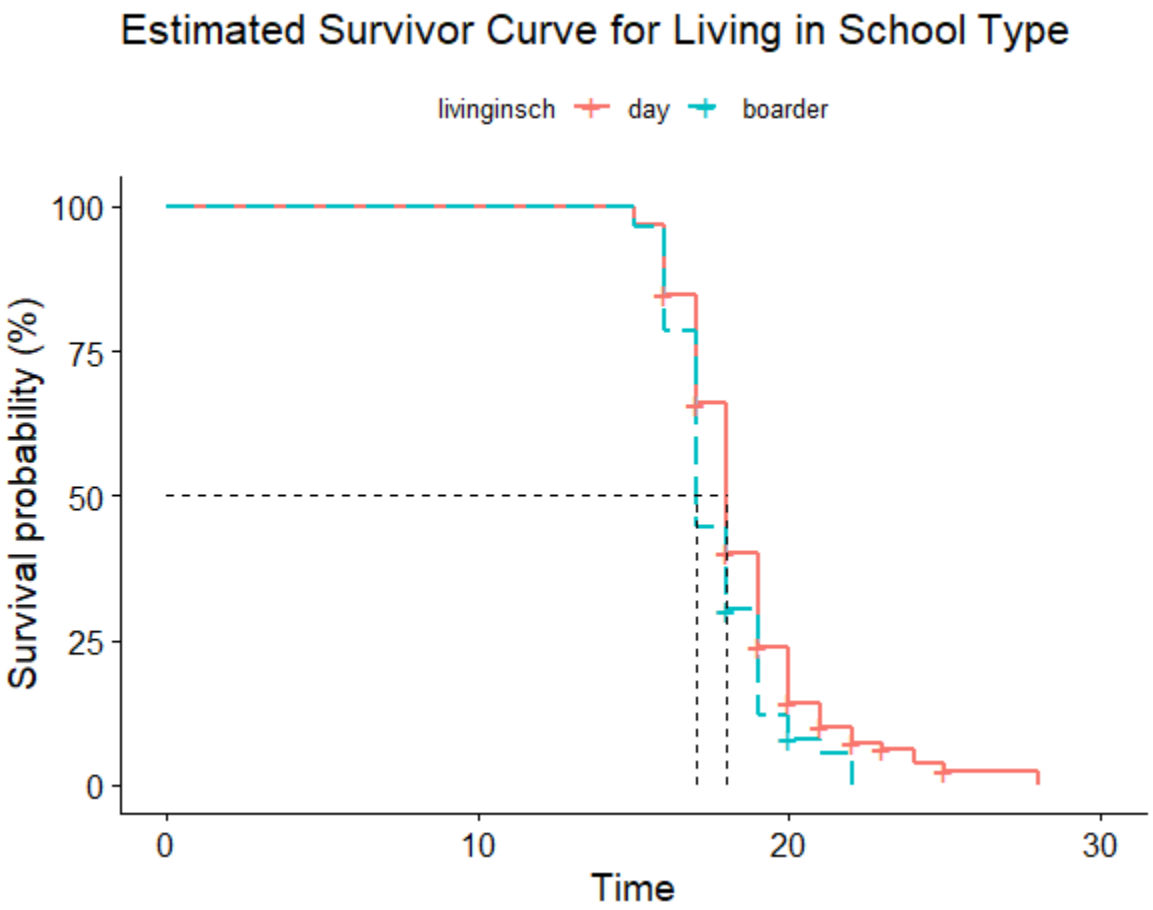

Figure 3.6: Survival curve for living in school type

Survival curve based on living in school while in secondary school, that is either day student or boarder.

The survival curve above shows the difference in the survival experience between the day students and the boarders in the sample. A test of significance of this difference is provided by the log-rank test.

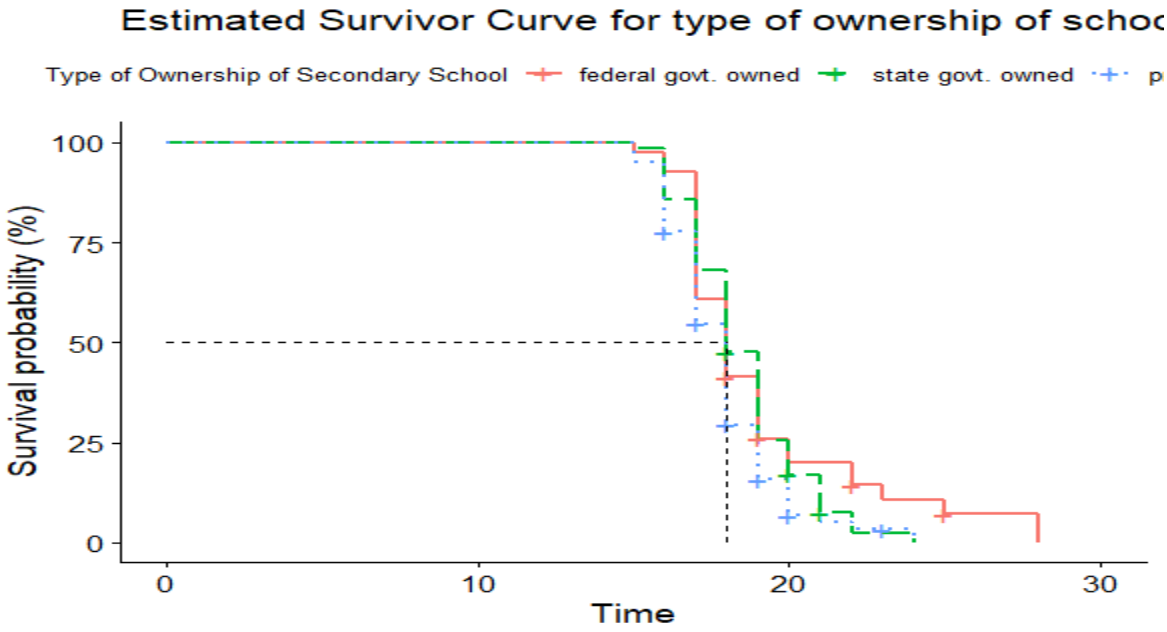

Figure 3.7: Survivor curve for type of ownership of school

Survival curve based on ownership type of secondary school attended.

The survival curve above shows the difference in the survival experience between the federal government owned, state government owned and private owned in the sample. A test of significance of this difference is provided by the log-rank test.

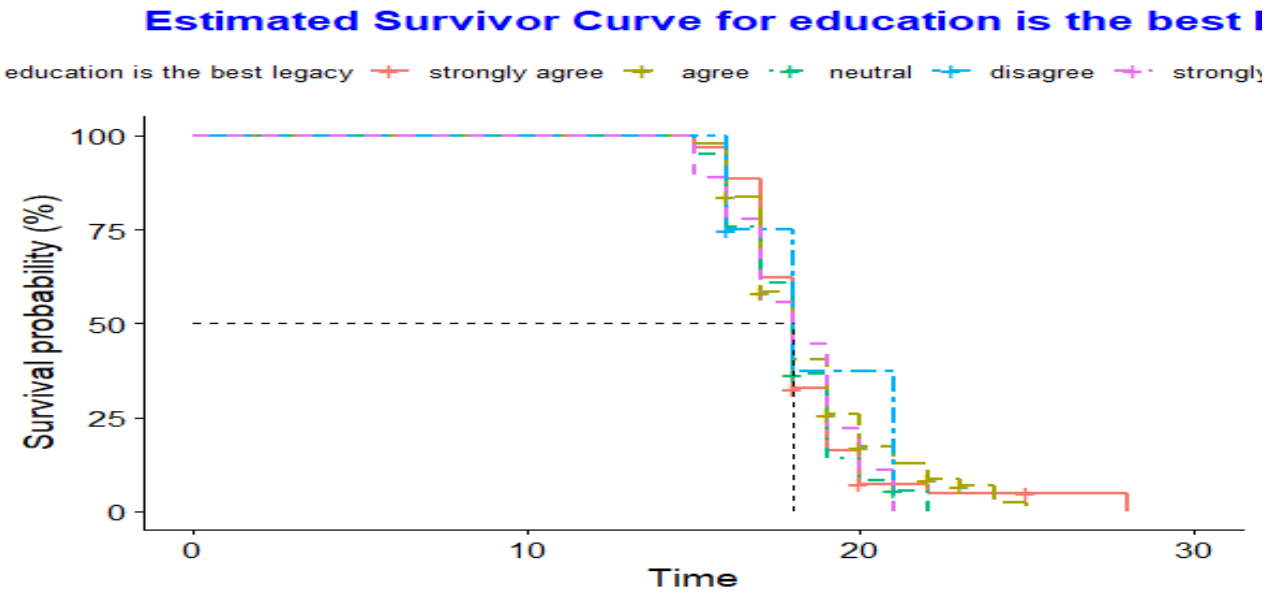

Figure 3.8: Survival curve based on the likert scale of "education is the best legacy."

The survival curve above shows the difference in the survival experience between strongly agree, agree, neutral, disagree and strongly disagree in the sample. A test of significance of this difference is provided by the log-rank test.

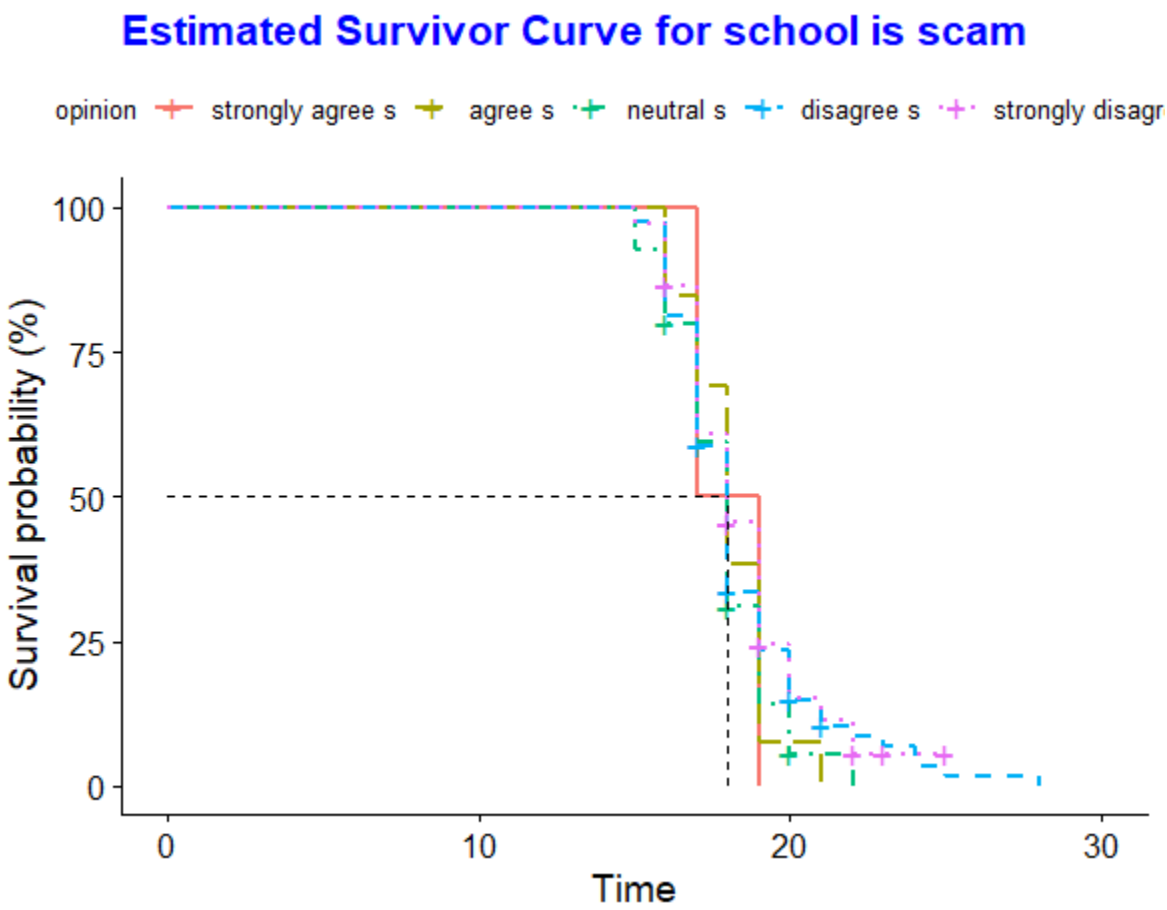

Figure 3.9: Survival curve based on “school is scam”.

The survival curve above shows the difference in the survival experience between the levels of the Likert of the notion “school is scam” in the sample. A test of significance of this difference is provided by the log-rank test.

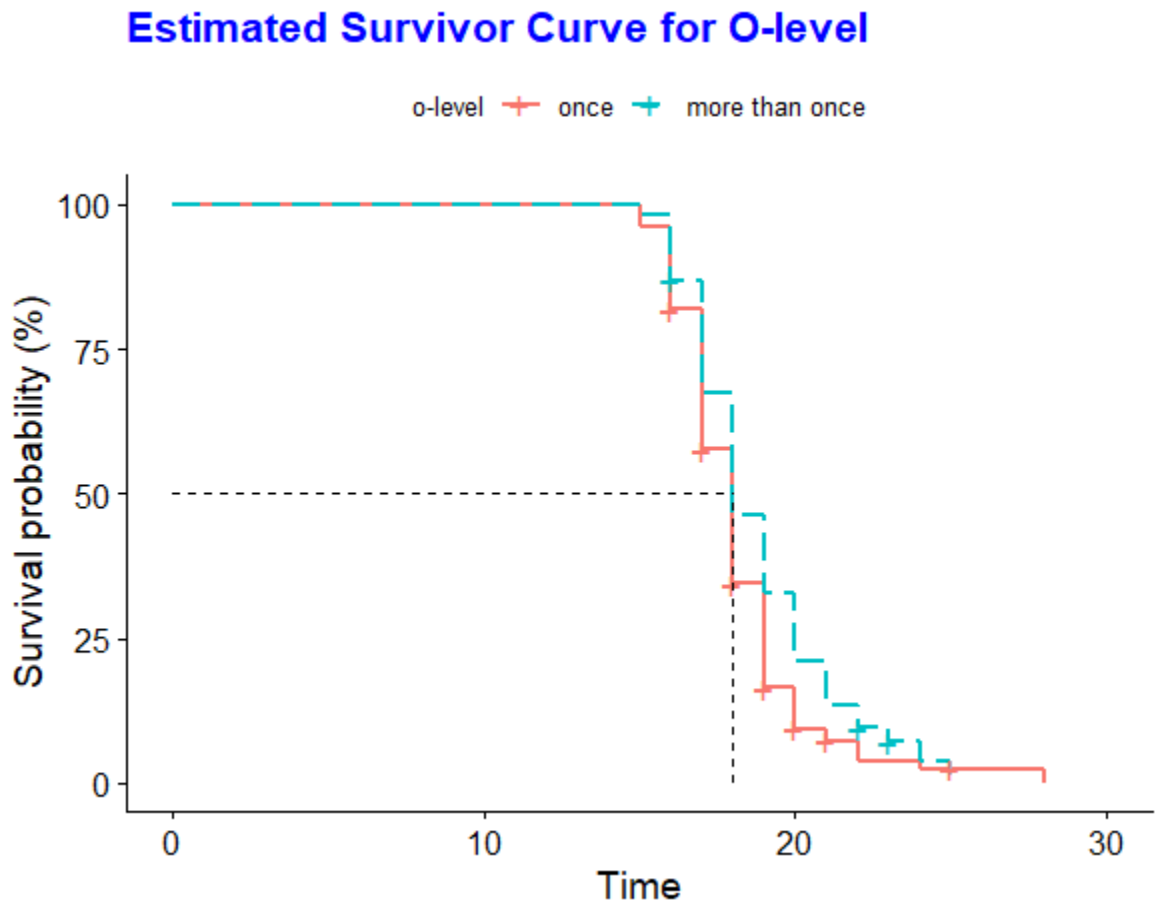

Figure 3.10: Survival curve based on number of attempts the units tried O-level exam.

The survival curve above shows the difference in the survival experience between the units who tried the o-level exam once and those who did the examination more than once in the sample. A test of significance of this difference is provided by the log-rank test.

3.6.2 Nelson Aalen Cummulative Hazard

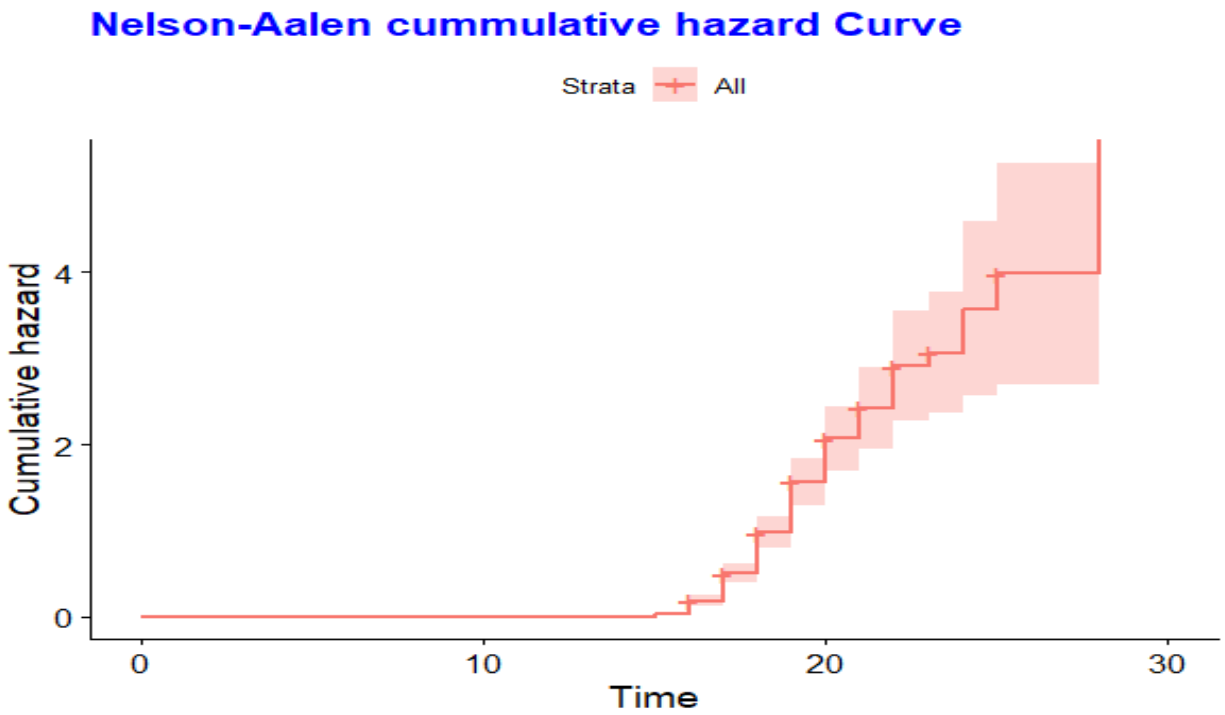

Figure 3.11: Nelson-Aalen cumulative hazard curve including the confidence interval

Nelson-Aalen estimator is a non-parametric estimator that was used to analyze cumulative hazard functions. While Kaplan-Meier estimator was used to analyze cumulative survival functions.

3.7 Cox Proportional Hazard Model

Cox proportional hazard model is a semi-parametric model with its baseline hazard $h_0(t)$ unspecified and this is the property that makes it semi-parametric.

$H(t,x) = h_0(t)[\exp\sum_{i=1}^{p} \beta ixi]$

where $h_0(t)$ is the baseline hazard

X is time-independent explanatory variable; status

t= time variable

number of units = 207

number of events = 191

Likelihood-Ratio test = 16.8

P-value = 0.3308

In the table below:

- **E ---> "Education is the best legacy"**
- **S ---> "school is scam"**

Table 3.13: Table for the cox probability hazard model

Covariate	Coefficient	Hazard Ratio	Std. error	Z	P-Value
South West	0.14	1.15027	0.16536	0.847	0.3972
South South	0.19045	1.20979	0.35109	0.542	0.5875
South East	-0.29048	0.74791	0.34934	-0.831	0.4057
North West	0.28559	1.33055	0.40753	0.701	0.4834
North East	0.9054	2.47291	0.59788	1.514	0.1299
Female	0.02901	1.02943	0.158	0.184	0.8543
State	0.11232	1.11887	0.22436	0.501	0.6166
Private	0.46331	1.58933	0.20824	2.225	0.0261
Day	-0.41602	0.65967	0.17654	-2.356	0.0185
strongly agree-E	-0.20407	0.81541	0.81679	-0.25	0.8027
agree-E	0.45507	1.57629	0.49094	0.927	0.354
neutral-E	-0.01453	0.98558	0.39241	-0.037	0.9705
disagree-E	-0.18292	0.83283	0.4207	-0.435	0.6637
strongly agree-S	0.46331	0.81541	0.3422	2.351	0.3972
agree-S	-0.29048	1.11887	0.4237	-0.435	0.8027
neutral-S	0.02901	0.85423	0.1585	0.1299	0.6755
disagree-S	0.18922	2.47291	0.45194	0.184	0.1299

The Cox proportional hazard ratio was employed to determine the hazard ratio of various covariates of the data.

The estimate regression coefficient and the hazard ratio (i.e expected coefficient) between groups of the covariates were obtained . the hazard ratio which was used to interpret the cox proportion hazard model is compared based in its closeness to 1.

Thus, The geopolitical zone's risk ratio, or the risk ratio for North central in relation to South West, is1.15027. North central is used as the baseline risk in this model.

The risk ratio for the geopolitical zone, which uses North Central as the baseline risk and compares that region to South-South, is 1.20979. Contrary to what is commonly believed, since the hazard ratio is bigger than 1, South-South has a lower risk and has survived longer than North Central before receiving first admission. Since 0.05 <0.5875, it was determined that the chance of being admitted to the first tertiary institution based on zone is not significant enough to be included in the cox model when the p-value (0.5875) was compared to the significant value of 0.05.

The risk ratio for the geopolitical zone, which uses North Central as the baseline risk and compares that region to South East, is 0.74791. Contrary to what is commonly believed, since the hazard ratio is below 1, South East has a larger risk and has survived less time than North Central before being admitted. Since 0.05 0.4057, it was determined that the chance of being admitted to the first tertiary institution based on zone is not significant enough to be included in the cox model when the p-value (0.4057) was compared to the significant value of 0.05.

Contrary to what is commonly believed, because the hazard ratio is greater than 1, North West has lower risk and has survived longer than North Central before receiving first admission. This is true for the geopolitical zone with North Central as the baseline hazard. Its hazard ratio is therefore 1.33055. The p-value (0.4834) was compared to the significant value of 0.05 to determine whether the covariates were significant in the model. Since 0.05 0.4834,

it was determined that the chance of being admitted to the first tertiary institution based on zone is not significant enough to be included in the cox model.

The risk ratio for the geopolitical zone, which uses North Central as its baseline risk, is 2.47291. This means that North Central's risk is more than twice as great as North East's risk. Contrary to what is commonly believed, the North East has a lesser danger than the North Central since the hazard ratio is more than 1, and it also managed to survive longer before being admitted. In order to determine whether the covariates in the model are significant, the p-value (0.1299) was compared to the significant value of 0.05. Since 0.050.1299, it was determined that the chance of being admitted to the first tertiary institution based on zone is not significant enough to be included in the cox model.

The hazard ratio for males compared to females, or the hazard ratio of gender using male as the baseline hazard, is 1.02943. Contrary to what is commonly believed, since the hazard ratio is greater than 1, females have a lower risk and have survived longer before being admitted than males. The p-value (0.8543) was compared to the significant value of 0.05 to determine whether the covariates were significant in the model. Since 0.050.8543, it was determined that the chance of being admitted to the first tertiary institution based on gender is not significant enough to be included in the cox model.

The risk ratio for the kind of school ownership using Federal as the baseline risk, or the risk ratio for Federal in relation to State, is 1.11887. Contrary to what is commonly believed, since the hazard ratio is bigger than 1, State has a lower risk and has survived longer than Federal before receiving first admission. The p-value (0.6166) was compared to the

significant value of 0.05 to determine whether the covariates were significant in the model. Since 0.050.6166, it was determined that the chance of being admitted to the first tertiary institution based on the type of ownership of the school is not significant enough to be included in the cox model.

The risk ratio for the kind of ownership of a school is 1.58933, with federal ownership serving as the baseline risk. Contrary to what many people instinctively believe, since the hazard ratio is higher than 1, Private has a lower risk and has survived longer than Federal before receiving first admission. The likelihood of being admitted to the first tertiary institution depending on the kind of ownership of the school was not significant enough to be included in the cox model, which was determined by comparing the p-value (0.0261) to the significant value of 0.05.

the hazard ratio of Living in school type with boarding as the baseline hazard, that is the hazard ratio for boarding relative to Day is 0.65967. This means that since the hazard ratio is lesser than 1, Day has higher risk and survived shorter than boarding before gaining first admission, contrary to the popular instinctive belief. To test whether the covariates have any significance in the model, the p-value (0.0185) was compared to the significant value of 0.05, and since 0.05>0.0185, it was observed that the chance of gaining first tertiary institution admission based on gender is not significant enough to be included in the cox model.

Strongly disagree is the baseline danger, and the hazard ratio for Strongly disagree relative to Strongly agree-E is 0.81541. This makes education the best legacy. Contrary to what is

generally believed to be the case, as the hazard ratio is less than 1, strongly agree-E has a higher risk and has lived less time before being admitted than strongly disagree. The p-value (0.8027) was compared to the significant value of 0.05 to determine whether the covariates were significant in the model. Since 0.05 0.8027, it was determined that the chance of being admitted to the first tertiary institution based on gender is not significant enough to be included in the cox model.

The hazard ratio for Strongly disagree relative to agree-E is 1.57629, making it the best legacy with Strongly disagree as the baseline hazard. Contrary to what many people instinctively believe, since the hazard ratio is bigger than 1, agree-E has a lower risk and has lasted longer before receiving initial admission than Strongly Disagree. According to the belief that "education is the best legacy," the probability of being admitted to the first tertiary institution was not significant enough to be taken into account in the cox model when the p-value (0.354) was compared to the significant value of 0.05.

Contrary to what many people instinctively believe, neutral-E has a higher risk and has survived for a shorter period of time before receiving their first admission because the hazard ratio is less than 1, which is why education is the best legacy with strongly disagree as the baseline hazard is 0.98558. Since 0.05 0.9705, it was determined that the chance of being admitted to the first tertiary institution based on the belief that "education is the best legacy" is not significant enough to be included in the cox model. This was done to determine whether the covariates have any significance in the model.

The hazard ratio of Education is the best legacy with Strongly disagree as the baseline hazard, that is, the hazard ratio for Strongly disagree relative to disagree-E is 0.83283. This means that, contrary to popular belief, disagree-E has a higher risk and survives shorter than Strongly disagree before gaining first admission. To see if the covariates had any significance in the model, the p-value (0.6637) was compared to the significant value of 0.05, and since 0.050.6637, it was determined that the chance of gaining first-year tertiary institution admission based on the perspective that "education is the best legacy" is not significant enough to be included in the cox model.

The hazard ratio of Schools is a scam with Strongly disagree as the baseline hazard is 0.81541. This means that because the hazard ratio is less than one, strongly agree-S has a higher risk and survives shorter than Strongly disagree before gaining first admission, contrary to popular instinctive belief. To see if the covariates had any significance in the model, the p-value (0.3972) was compared to the significant value of 0.05, and since 0.050.3972, it was determined that the chance of gaining first-year tertiary institution admission based on the perspective that "school is a scam" is not significant enough to be included in the cox model.

School is a scam with Strongly disagree as the baseline hazard, hence the hazard ratio for Strongly disagree compared to neutral-S is 0.85423. This indicates that, contrary to popular assumption, neutral-S has a higher risk and survives longer than Strongly disagree before receiving initial admission since the hazard ratio is smaller than one. To see if the covariates had any significance in the model, the p-value (0.6755) was compared to the significant value

of 0.05, and since 0.050.6755, it was determined that the chance of gaining 1st tertiary institution admission based on the perspective that "school is a scam" is not significant enough to be included in the cox model.

the hazard ratio of School is a scam with Strongly disagree as the baseline hazard, that is the hazard ratio for Strongly disagree relative to disagree-S is 2.47291. This means that since the hazard ratio is greater than 1, disagree-S has a lower risk and survived longer than Strongly disagree before gaining first admission, contrary to the popular instinctive belief. To test whether the covariates have any significance in the model, the p-value (0.1299) was compared to the significant value of 0.05, and since 0.05<0.1299, it was observed that the chance of gaining first tertiary institution admission based on the perspective that “school is a scam” is not significant enough to be included in the cox model.

CHAPTER FOUR

SUMMARY AND CONCLUSION

4.1 Summary

The summary of the data and analysis carried out in this paper are as follows:

The data was collected via a questionnaire conducted with Nigerian youths as the targeted subjects were presented via graphs, tables, and other visualization tools.

The Kaplan-Meier survival function and curves were used to show the relationship and difference between survival curves.

The Nelson Aalen survival curve was used to show the rate of failure, in other words, the rate of admission.

The use of a log-rank test with the aid of a hypothesis was used to compare the survival experience of units in different cohorts of covariates.

The cox probability hazard model was used to model the data.

4.2 Conclusion

From the data collected on the survival analysis of time from birth till first tertiary institution admission, the following conclusion was drawn:

Low or no multicollinearity in the variables used.

All the factors used were significant except ‘school is a scam and ‘education is the best legacy. The study shows that one’s opinion about school and education is not significant to the age of admission or censorship for that matter.

There is a relationship between the Kaplan-Meier survival curve and the Nelson-Aalen cumulative hazard curve.

The median survival time for the overall survival curve is 18 years that is, on average, a Nigerian youth will gain his or her first tertiary institution admission by the age of 18 years.

Both males and females have equal survival times.

REFERENCES

- Ani Katchova (2013). Survival analysis
- Cox, D.R. and Oakes, D. (1984). Analysis of survival data, Chapman and Hall, New York.
- Dr. Nicholas Girerd, MD, Ph.D.Survival Analysis: interaction with time- how to deal with it in SPSS
- Fleming, T.R. and Harrington, D.P. (1981). Counting processes and survival analysis, Wiley, New York.
- Introduction to stata by Ani Katchova
- Kalbfleisch, J.D. and Prentice, R.L. (1980). The statistical analysis of failure time data, Wiley, New York.
- Klein, J.P. and Moeschberger, M.L. (1997). Survival analysis, techniques for censored and truncated data, Springer, New York.
- Kleinbaum, D.G. et Klein, M. (2005). Survival analysis, a self-learning text, Springer, New York.
- Lectures in statistics by Prof. Mike Marin
- Statisticsmentor.com
- Wilson Fandino and Karla Loss. Stata for running survival analysis
- Zedstatistics.com