Relationship between causes of death and education level of South African youth in 2014: A bivariate analysis study

B V Mogale

Dissertation submitted in fulfillment of the requirements for the Degree of Master of Commerce in Statistics at the Mafikeng Campus of the North-West University

Supervisor: Prof Elias Munapo

Co-supervisor: Dr Tshepiso Tsoku

Student number: 17002141

DECLARATION

I, Boipelo Vinolia Mogale, declare that this dissertation is my own original work, and that all sources have been accurately referenced and acknowledged. It is being submitted for the degree of Masters in Statistics. To the best of my knowledge, it has not been submitted before in part or in full for any degree or examination at this or any other University.

Bmogale ..... 30 May 2018

BV Mogale **Date**

ACKNOWLEDGEMENTS

First and foremost, I thank God for giving me the wisdom, strength and excellent health and resilience to complete this dissertation. I thank God for blessing me by surrounding me with such supportive people all round and some I cannot not mention by name here though am truly grateful and humbled by all their support. I give him all the glory, honour and adoration.

I would also like to thank the North West University Bursaries and Statistics South Africa bursary for the financial support without which this dream could not be realised.

To my mom, Kearoma Mogale, the pillar of my strength, I thank you for your steadfast support and unconditional love. Thank you for encouraging and believing in me. To the rest of my family, your support was the wind beneath my wings all the time, I love you all.

To my supervisors, Prof Elias Munapo and Dr Tshepiso Tsoku, I thank you for the guidance, advice and unwavering support you provided me and enabled me to complete this project. Prof Munapo, thank you for your constant encouragement and the time you devoted to me throughout this journey. Dr Tsoku, your statistical guidance really saw me through this paper. Without you all, the masters' degree would have remained but a dream for me.

To all my friends who stood by me and supported me, empathised and appreciated my absence from all our regular fun pursuits, I thank you. A special thank you is due to my best friend, Neo Modibane that listened and kept reinvigorating me till the end.

I am also truly grateful to my manager, Ms Dineo Mokhuwa, for affording me learning space and time demanded by a project such as this. Ms Nthabiseng Makhatha, I also thank you for your unwavering support and guidance.

ABSTRACT

Youth mortality is a challenge in South Africa, where on a daily basis a number of death are reported and are related to youth. The distressing part in it all is that ethically no parent should bury their own child. The school dropout rate is also a correlated concern. The report from the Education Policy and Data Center (2014) states that although the youth in the age category 15 – 24 years may still be in school and striving towards their educational goals, it is notable that approximately 1% of the youth have no formal education and 7% of the youth have attained at most incomplete primary education, connoting that in total 7% of 15 – 24 year olds have not completed primary education in South Africa.

The objective of the study is to determine whether there is a relationship between education and causes of death, in order to utilise education as a factor that can assist reduce youth mortality significantly and speedily. The study therefore investigated this topical issue of association between these factors utilising bivariate techniques such as Chi-Square test, Wilcoxon-Mann-Whitney test, Wilcoxon signed-rank test, ANOVA, Kruskal-Wallis test, Friedman test, Logistics regression and Spearman's rank correlation coefficient. The test adopted here revealed that there is a significant relationship between education and causes of death; and that the odds of the youth mortality could decrease if more youth had higher levels of education. Through the Friedman test it was noted that there is a statistically significant difference in the causes of death depending on education level of the youth. However, the strength of the relationship between education and causes of death was insubstantial.

Keywords: Causes of death, youth, Chi-Square test, Wilcoxon-Mann-Whitney test, Wilcoxon signed-rank test, ANOVA, Kruskal-Wallis test, Friedman test, Logistics regression and Spearman's rank correlation coefficient

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

AIDS	Acquired Immune Deficiency Syndrome
ANOVA	Analysis of Variance
DHA	Department of Home Affairs
HIV	Human Immunodeficiency Virus
HSRC	Human Sciences Research Council
ICD	International Classification of Diseases
MLE	Maximum Likelihood Estimates
NDP	National Development Plan
NYP	National Youth Policy
OECD	Organisation for Economic Co-operation and Development
QLFS	Quarterly Labour force Survey
SAMRC	South African Medical Research Council
SDG	Sustainable Development Goal
Stats SA	Statistics South Africa
TB	Tuberculosis
UNESCO	United Nations Educational, Scientific and Cultural Organization
WHO	World Health Organisation

CHAPTER 1

ORIENTATION OF THE STUDY

1.1. Introduction

From time immemorial, it is a known fact that people are born and they ought to die at some point; that no one is to live forever. A number of deaths are reported daily through communication media such as television, newspapers- to mention a few- about the deaths that occur in and around the country especially those of the youth. What causes these deaths amongst the youth is a concern that needs to be addressed as a matter of urgency.

In South Africa, Department of Home Affairs (DHA) is responsible for recording and registering of all the births and deaths happening within the country. The DHA is mandated through the Births and Deaths Registration Act (Act No. 51 of 1992) to ensure that births and deaths are recorded properly and by the relevant people. The deaths in particular, are recorded using the death notification forms by medical practitioners. The International Classification of Diseases (ICD) code list is then used to classify the causes of death for each reported death case. According to Statistics South Africa (Stats SA), Statistical release P0309.3 (2015) there are basically two types of causes of death, namely natural causes or non-natural as certified by a medical practitioner. Natural causes are often attributed to an illness or underlying malfunctioning of the body and non-natural causes refers to deaths that are not natural such as accidents, suicide, and so forth.

South Africa's National Youth Policy (NYP) (2009-2014) defines youth as persons between the age category of 15 and 34 years. According to the *mid-year population estimates 2016* produced by Stats SA there is an overall of 42% of young people in South Africa. In the Stats SA mid-year population estimates (2016) it is reported that in South Africa there has been an improvement in the living conditions of the youth. The youth are now having access to water, housing and sanitation, and are technologically advanced; nevertheless, the high youth unemployment rate, high HIV infection rate as

well as the growing number of households that are headed by young people are still some of the significant challenges that they have to grapple with. The NYP is geared towards prioritising the needs of young people with respect to education, health and well-being, economic participation and social cohesion. The education of the youth is also of paramount importance, the National Development Plan (NDP) 2030, is aiming at ensuring that all people have access to education. The World Health Organisation assets that health is one of the key focus areas of the 2030 Agenda. Within the Sustainable Development Goals (SDGs) and with special reference to SDG 3, health and well-being for all people of a nation receive special emphasis as a global action item. On the other hand, the African Youth Charter (2006) also emphasises that every young person should have the right to education of good quality and the right to enjoy the best attainable state of health physically, mentally and spiritual. Education level of the nation remains the most vital factor in the absence of cure for some of the diseases.

The study focuses on the causes of death of the youth of South Africa and their education level. The study is aimed at determining whether a relationship exist between the two variables.

1.2. Background literature

According to Statistics South Africa, Statistical release P0211.4.2 (2015:1), there was an improvement in the employment levels for the youth as there were about "44.5% of employed youth and 50.3% of employed adults had education below matric", which was an improvement in the education level for the youth and adults during this period as compared to previous years. It was noted that one out of every two young people who were unemployed and seeking employment only had education below matriculation. This therefore made the youth more vulnerable, without education it inevitably fated unemployment and a subsequent risk of contracting diseases whereas unable to access quality medical care.

Borode (2011) suggests that the higher education curriculum in Sub-Saharan African countries should be amended so as to deal with the effects of the current education and training curriculum which have led to the observed high levels of unemployment. He asserts that the drive should focus on training in relevant scarce skills to ensure job

market readiness. Higher education should be directed at producing graduates that are relevant to the economy, who would apply the acquired knowledge and skills to shape the economy and create jobs, rather than seeking for job opportunities as employees versus being entrepreneurs and employers. According to Harvey et al. (2009) there should be intervention measures such as acquiring education, development of skills and change in behaviour programmes for both children and parents, but that strategy should not be used as the only measure since there are other strategies in place. The strategy should therefore be part of the multi-faceted child injury strategy but evidence has shown that this strategy alone does not reduce injuries amongst children especially when other measures are not taken into consideration.

Prior to 1994, the public health care system consisted mostly of hospitals, which were mainly accessible to the elite due to high costs. The public health care services therefore excluded the poor and black people. Post 1994, through the African National Congress (ANC) government's health plan, measures have been put in place to make primary health care accessible and affordable. This resulted in a considerable decrease in the number of patients that use hospitals – many prefer to go to primary health care facilities in their respective communities. Regardless of the large numbers of primary health care facilities that have been built since 1994, the demand far outweighs the supply of services and resources, particularly in the rural areas (Bradshaw and Steyn, 2001).

Krueger et al. (2015) makes reference to a study conducted at three universities, namely, Colorado, New York, and North Carolina at Chapel Hill (2015) which illustrated that certain causes of death are associated with particular levels of education, and that the variations across education level have widened greatly over the study period – 2010. Natural experiments have shown a strong association – which is substantially causal - between education level and mortality.

1.3. Statement of the problem

Generally the youth are the most vulnerable group in the population as they are victims of many socio-economic factors such as unemployment, crime, poverty, and many more. Education together with other factors contributes to better health. Income is one of the primary factor that act together with education to have a significant influence on health (Feinstein et al., 2006).

Although the NDP is aimed at ensuring that all people receive education and access to medical facilities, this does not appear to have any impact on youth mortality in South Africa, since cases of young people dying from stress related illness, TB, HIV, and so on are still reported on a daily basis. According to Health 24 (2016), South Africa is however making significant progress in its medical efforts to reduce HIV infections, but a lack of education and social changes has left adolescents vulnerable to infections, with teenage girls terribly particularly at risk. The same principle applies to other infections and diseases as well.

Human Sciences Research Council (HSRC) (2017) overall findings from this study are that despite the country's comprehensive legal and policy framework, which is mostly focused on improving the health of the youth, there are some persistent structural and systematic factors that hinder effective provision and programming of adolescent and youth friendly services.

Stats SA in 2015 reported that there is a decrease in mortality rates of the whole country, despite that the youth mortality is still on the rise. The percentage of the youth dying from the whole population is high. It is therefore important to understand the relationship between education levels and causes of death, where education can be utilised to curb the increasing percentage of youth mortality. The question probed by the study is to determine whether or not there is a relationship between causes of death and education level of the youth. The results of the study will respond to the said question and also assist in making inference about the state of education in South Africa.

1.4. Objectives of the study

The objectives of this study are as follows:

- To determine the relationship between causes of death and education level of the youth in South Africa.
- To determine the relationship between causes of death and other socio-demographic variable (age, gender and province of death) of the youth in South Africa.
- To examine the extent of the relationship between causes of death and education level, age, gender and province of death

1.5. Research questions

Death is one predicament the human race cannot escape. The youth of the country are either affected or infected by natural ailments or non-natural causes. The questions posed by this study are:

- Is there a relationship between causes of death and education level of South African youth?
- Is there a relationship between socio-demographic variables (age, gender and province of death) and causes of death?
- To what extent is the relationship between the education level and causes of death?

1.6. Significance of the study

This study will investigate whether a relationship exists between the causes of death and education level, based on data analysis for South African youth in 2014. The two Sustainable Development Goals (SDG) related to this study are; "(1) to ensure healthy lives and promote well-being for all" (United Nations Economic Commission for Africa, 2015:9); and (2) "ensure inclusive and equitable quality education and promote lifelong learning opportunities for all" (United Nations Economic Commission for Africa, 2015:9). The study will contribute by determining if there is a relationship between causes of death and education level of the youth. Perhaps if they had acquired more education they could have lived longer and not died.

1.7. Data source

This study examines the relationship between causes of death and education level of the youth in South Africa in the year 2014. Data analysed for the purpose of this study was obtained from Stats SA. This data was originally collected by Department of Home Affairs (DHA) for the purpose of recording and registering the deaths, and was then processed and published by Stats SA in 2015.

1.8. Limitations of the study

From the Statistics South Africa Statistical release P0309.3 (2015), it was reported that other variables had high percentage of incomplete information and education was the highest at 48, 5%. That became evident during the cleaning of the data, whereby close to 40% of the data had to be deleted from the dataset, which was either unknown, unspecified or not applicable. In light of the above, that posed as a big limitation to the study at hand, as it also made the data to seem unreliable when tested for reliability. On the other hand, data from Stats SA is used because of the difficulties of accessing the files/records directly from the Department of Home Affairs, which also makes editing or imputing the data very precarious, as it will draw an incorrect picture of the relationship between the causes of death and education level of the youth. Varied literature pointed out that this lack of reporting might be due to the fact that some deaths were certified by traditional leaders in rural areas, and the fact that there is nowhere in the death notification forms to capture that information, to date (Bradshaw et al., 2010). Another limitation was that the variable for race was not included in the data because Stats SA deem it as a very sensitive variable therefore does not permit researchers to use it in their studies.

1.9. Layout of the study

Chapter one provides the orientation of the study. It outline in depth, the introduction, background literature, research objectives, and statement of the problem of the study. It also outlines questions the study intends to answer, the significance and limitations of the study. The chapter also discuss the sources where data was obtained. Chapter two gives an overview of the literature. It discusses what other authors have discoursed in relation

to the issue at hand both internationally and in South Africa. It further provides the theoretical background of the study and the methods utilised for ascertaining the causes of death. Chapter three outlines the research methodology to be used in detail and the justifications thereof.

The study employed Bivariate techniques such as chi-square test, Wilcoxon-Mann-Whitney test, Wilcoxon signed-rank test, ANOVA, Kruskal-Wallis test, Friedman test Spearman's rank correlation and logistics regression to test whether a relationship does exist between the dependent variable (causes of death) and independent variables (education level, age, gender and death province). In Chapter 4, the research finding and their interpretation thereof are outlined. Chapter five discusses research findings, and provides conclusions, recommendation and areas of further study

1.10. Conclusion

This chapter summary is provided in the introduction of the study. It also included the statement of the problem, the objectives and questions intended to be answered by the study. It further included the significance and limitations of the study. In addition to all, it emphasizes the sources were data was obtained. The chapter that follows focuses on the literature review.

CHAPTER 2

LITERATURE REVIEW

2.1. Introduction

This chapter provides a review of what other authors, websites and researchers have deliberated and set parameters on about causes of death and education globally and in South Africa. The chapter 'gives insight into what most people meet their deaths from and also the differences that are brought about by the attainment of higher education.

This chapter is organised as follows: Section 2.2 presents the leading cause of deaths, Section 2.3 presents the education levels of the youth, and section 2.4 focuses on the relationship between causes of death and education level, and also the link with other socio-demographic. Section 2.5 presents the theoretical framework of the study, section 2.6 present methods used to ascertain causes of death and lastly, section 2.7 present literature on bivariate data.

2.2. Leading cause of deaths

This section presents the leading causes of death globally, with the aim of revealing the main causes of death in the continent and around the world, specifically those of the youth.

To start with, Miniño (2010) states that in the United States, accidents/unintentional injuries, homicide, suicide, cancer and heart diseases were the leading causes of death in the period 1999 – 2006. Miniño (2010) further states that deaths causes by motor vehicles claimed 73% of all the deaths. The average annual death rate of teenagers was 49.5 deaths per 100.000 population.

India, on the other hand, which is also a densely populated country and had the largest population of youth globally estimated 356 million (10 – 24) deaths in 2014. Leading causes thereof among the youth in India are self-harm (suicide) which saw 59.366 deaths recorded, followed by road injuries that claimed 37.137 youth and lastly, Tuberculosis

claimed 28.676 lives of youth in 2013 to become the third leading causes of death among youth generation (Sawe, 2017). Similar to India, Canada had the highest death by motor vehicle collisions followed by suicides among the youth, were an average of 294 youth were reported to have committed suicide each year (Center for Suicide Prevention, 2011).

According to Mwaniki (2017) HIV/AIDS, malaria and respiratory infections were the main causes of death in Africa. Annually over a million deaths caused by these diseases are reported. In Africa men have the average life expectancy of 58 years while women has 61 years. When compared to other continent it is notable that Africa has the lowest life expectancy. Approximately 1.1 million people were estimated to have died from HIV/AIDS in 2012. Children under the age of five in Africa suffered mostly from lower respiratory tract infections such as pneumonia, influenza and bronchitis. These infectious diseases contributed to at least a million deaths in 2012. Diarrhoea accounted for 6.7% (603.000 people) of total deaths in 2012. Lastly, malaria accounted for an estimated 554.000 deaths in Africa in 2012 (Mwaniki, 2017).

De Wet and Odimegwu (2012) in their study showed that females were more likely to die from natural causes compared to male counterparts. In addition to this, deaths due to natural causes decreased for males during the period 2001 and 2007 as 26 natural cause of deaths to adolescent males (aged 10 – 19) per 10.000 adolescent male population occurred in 2001. This figure decreased to 25 natural causes to males per 10.000 male population in 2007. On the other hand, the cause specific mortality rate showed that 18 adolescent males per 10.000 adolescent males died from unnatural causes in 2001; and in 2007 there was an increase in the number of male deaths, where it was recorded that there were 26 male deaths per 10.000 adolescent males who died.

Analysis of youth death rates as outlined in the Statistics South Africa report 03 – 09 – 12 (2015), show that for all the deaths that occurred in 2013, the youth accounted for 16.4% of the overall mortality rate. In 1997 youth mortality accounted for 18% of the overall deaths in the country, which was deemed high at the time. In 2004, this figure further increased to 24%, gradually decreasing over the years to the lowest rates observed in 2013 (16.4%). In South Africa the most common cause of death is the natural causes which accounted for 71.9% of the total deaths that occurred in 2013 for the youth, while

non-natural causes only contributed approximately 29% of the total death. The report confirms the findings of De Wet and Odimegwu (2012) that showed that young males were at most at the risk of dying from the non-natural causes when compared to females. Females on the other hand were most likely to die from natural causes.

Recently, it was reported that the main factors that influence life expectancy are chronic diseases and accidental injuries. According to the estimated projections life expectancy will increase by 0.86 year in the next five years, however the goal of increasing life expectancy by one year was not reached meaning that there is still a lot that needs to be done to achieve this goal (Liu et al., 2014).

2.3. Education levels of the youth

According to the UNESCO factsheet, education attainment can be used to reduce poverty and hunger, and also to promote sustainable developments in the world. UNESCO further reports that globally about 10.6% of the youth are uneducated and lack basic numerical and reading skills. As a result there is high unemployment rate among the youth, since many do not have sufficient knowledge and skill to apply for decent jobs. Lack of access to and unavailability of gender sensitive educational infrastructure characterises the gender inequality in education. Furthermore high teenage dropout in secondary school increases the gender gaps in education. The resultant inequality deters development amongst the youth.

Over the years in all OECD and partner countries, there has being an increase in the levels of educational attainment. It was reported that 80% of young adults had education up to secondary level in 2000 in about 20 out of 35 OECD members. In 2016 the OECD countries on average had 84% of the people aged between 25 – 54 years who had at least attained upper secondary education. This number had increased from the 75% reported in 2000. There was also a steady increase in the number of adults (aged 25 – 34) with secondary, post-secondary and tertiary education during the period 1970 to 2016.

According to the Child Trends Data Bank (2016), there is a probability that young adults who have completed higher levels of education were more likely to attain economic

success than those with lower levels of education. The Ministry of education and research (2016), Mailis Reps, further put an emphasis on the point that many young people limit themselves by completing only basic education or failing to acquire professional qualification and therefore, support services and early interventions should be prioritised.

It was established in the Australian census of 2006 that approximately 63%, which is 2.2 million person in Australia were attending some form of education, either secondary, tertiary or Vocational Education and Training (VET). Forty-three percent of the youth were in secondary schools, while 1 in 5 had registered at a tertiary institution and only 6% were attending at a VET facility. The study suggests that working youth or who are more educated suffered from psychological distress compared to youth without employment or who had low levels of education. (Australian Government, 2009).

Excellent health was associated with young adults in possession of a bachelor's or higher degree, (National Center for Education Statistics, 2011). The young adults with lower levels of education were linked to poor health. According to the National Center for Education (2011) 65% of the young adults did not complete high school, while those who did complete were reported to be at 72% and 82% were reported to have some college education and/or an associate's degree. Hanushek and Wößmann (2007) further argued that the notion that skills and human capital can only be acquired through formal schooling is short sighted as other factors like family, peers and other individuals are more influential to an individual's knowledge level and ability to think rationally.

2.4. Relationship between education and causes of death

Hummer and Hernandez (2013) state that "higher educated adults in the United States have lower yearly mortality rate than less-educated people in every age, gender and racial/ethnic subgroup of the population". (Hummer and Hernandez, 2013:3)

Education has a potential to create opportunities for better health. The relationship between the education and health of the people in general and the youth in particular has existed for generations, whether tests were conducted or not, in a sense that it was noted that people with no education were most likely to have poor health which ultimately resulted in untimely death. An assumption is that people who are exposed to poverty and

high unemployment are unable to get better health cure for many of their ailments. It is therefore believed that eventually there will be an increase in a number of these diseases amongst the least educated. People who have acquired more education tend to have better health, perhaps due to exposure to information that enables them to make informed decisions about their health and life in general and also having the means to look after their health should the need arise (Feinstein et al., 2006).

From theory it behoves one to state that people with lower education levels are the most likely affected by diseases, resulting to their ultimate death; vice versa it seems like the most educated peoples are most likely to die from accidents, murder and the like. In a study conducted by Mwamwenda et al. (2014), results indicated that university students were more knowledgeable on HIV/AIDS as compared to secondary students, According to the SAMRC Burden of disease research unit the causes of death statistics are an essential data source for understanding, monitoring and making policy decisions that are aimed at improving the health of the nation.

In relation to education, the study conducted by Garriga et al. (2015) showed that there was proportionately a higher intravenous drug usage amongst a group of patients with lower levels of education, 26%, compared to those with higher levels of education, 1,5%. Additionally, the study further suggests that patients with lower levels of education had a lower CD4 count compared to those with higher levels of education. This further, emphasises that the probability of people with lower levels of education being exposed to some of these diseases were higher.

The Steinhardt School of Culture, Education and Human Development (2015) found that in the United States, more than 10% of the adults did not have high school degree, more than a quarter, had some college and no bachelor's degree. The study submits that generally, higher education is associated with longevity because of factors such as higher income and social status, healthier behaviours and improved social and psychological well-being of individuals. In support of this finding, Feinstein et al. (2006) suggests that education is directly correlated with income levels which in turn are directly correlated with health status in that better education is likely to lead to improvement in the income levels of individuals which in turn have a positive influence on the health status of that individual.

Generally, life expectancy is increasing for the human population, and people with higher levels of education are realising most of the benefits. Narrowed to a daily basis, being educated implies getting better jobs and access to better health due to the benefits provided by the employer whereas these basics are non-existent for the uneducated who land even more hazardous jobs with no benefits and get exposed to sickness as they are not able to take care of their medical conditions/needs. All need to take care of health equally because poor health can conversely put even hard earned educational attainment at risk. One of the proposed goals by UNESCO in 2014 was to "ensure healthy lives and promote well-being for all at all ages". It further states that educated people are better informed about scientific diseases, so they can take measures to prevent them or act on the early signs. Education plays a major role in ensuring the diseases are contained and not exacerbated.

The Centre on Society and Health (2015) reported that in the United States 27% of adults without higher education were unable to consult a medical practitioner as they could not afford the cost of the consultation and medication. The figure for high school graduates was less than 18% while amongst college graduates this figure stood at only 8%. Equally, people need to take care of their health because poor health can put education attainment at risk in more ways than one.

Similarly a study that was conducted to examine whether gender disparity existed in AIDS mortality rates among youth with secondary education reported that AIDS mortality is higher among females than males in South Africa (De Wet, 2016).

Recent studies showed that causes of death differed by educational attainment. Hummer and Hernandez (2013) argue that mortality rates linked to causes that have less human control have less variability by educational attainment while those that are linked to social and behavioural factors are spread wide amongst the different levels of education.

While it has been proven that lower education levels are associated with poor health and therefore higher mortality rates, a report by the New York University (2015) submits that there are health conditions that are positively correlated with higher education levels. Educated people were more at risk of suffering from cardiovascular diseases.

2.5. Theoretical framework

WHO (2010) suggests that the knowledge and skills attained through education may affect a person's reasoning, equipping them with better decision making capability that enables them to receive and respond appropriately to health education messages, or enable them to communicate with and access appropriate health services.

According to LaVeist (2005), the following two general theories attempts to explain the link between socioeconomic status and health: "the theories of social causation and social selection (also referred to as social drift). These two theories offer different perspectives on whether low socioeconomic status causes poor health – or poor health causes low socioeconomic status" (LaVeist, 2005:171).

Researchers that have investigated the association between education and mortality in industrialized countries had consistently shown that higher levels of education are directly correlated with decreased mortality risk. However, the relationship between education and mortality, and the variations based on demographic group have received less attention (Everett et al., 2014). Everett et al. (2014) further suggest that models accurately predict the relationship and variations by cohort both between and within race/ethnic and gender population groupings.

The National Institutes of Health (2015) notes that there is a positive correlation between formal education, adult health and longevity. The arguments was in favour of increasing education in order to improve public health, but that depends on the assumption that educational attainment has a causal association to improved adult health. While a lot of studies have described the relationship between education and health status, there is still a need to establish and prove whether there is a cause and effect link between education level and health status. The following are three proposed "mechanisms to explain the association, and each has been empirically supported:

- Higher education causes better adult health by conferring access to resources that contribute to better health, such as fulfilling jobs, more social ties, increased awareness and education about disease and health, healthy lifestyles, economic security and personal control.

- Poor health (in childhood) causes lower levels of educational attainment and is predictive of poor health later in life
- The association is not causal, but points to common factors that influence both schooling and health".

2.6. Methods of ascertaining deaths in South Africa

Stats SA in collaboration with DHA, derived the following methods to verify and certify the causes of death using the death notification forms. It is of paramount importance to understand how the causes of death are ascertained in South Africa. This is carried out in order to ensure that proper procedures are followed when recording and registering a death, and that the underlying cause is recorded properly according to the tenth version of International Classification of Disease (ICD 10). In South Africa medical practioners have the responsibility of ascertaining the causes of death. The DHA prescribes that the BI-1663 and DHA-1663 forms be used for recording information pertaining to registering a death. The forms are used differently; the BI-1663 is used to record all deaths with the exception of deaths that occurred within the first week after birth and stillbirths. The latter two are recorded on the DHA-1663.

Even though the study does not use any information from the registry of DHA, it is still important to understand the purposes of the forms as this is the primary source of the data. The information recorded by the DHA for administrative purposes is later handed over to Stats SA for processing and analyses. This information is used to formulate policy and make informed decisions with regards to the health of the people in the country.

2.7. Literature on bivariate data

2.7.1. Chi-square test of independence

The chi-square test in statistics is used to evaluate whether there is a statistically significant difference between observed and expected frequencies on the basis of a certain set of theoretical assumptions being met (Blalock, 1979). According to Blalock (1979), when testing the independence of a contingency table, the null hypothesis is such that the two attributes or characteristics of the elements of a given population are not related (i.e. they are independent) against the alternative hypothesis that the two

characteristics are related (i.e. they are dependent).It is also important to test how strong is the relationship between variable, and for this study the Cramer' V test is used to perform such test. According to Jackson (2005) the assumptions underlying the χ^2 test of independence are the same as for the χ^2 goodness-of-fit test:

- The sample must be random
- The observations must be independent
- The data are nominal

Ha and Ha (2012) further elaborates on the assumptions for the χ^2:

- Each subject has only one entry, and the categories are mutually exclusive. This ensures that there is independence between observations.
- In χ^2 designs that are larger than 2X2, the expected frequency in each cell must be at least five. There is recent evidence to suggest that this assumption is unnecessary, or at least that χ^2 is not seriously affected by violations of this assumption.

2.7.2. Wilcoxon-Mann-Whitney (U) test

Mann-Whitney test is designed to evaluate the separation between the groups. A larger separation suggests that the populations are unlikely to be the same, while a smaller separation suggests that both groups come from the same underlying null hypothesis population (Ha and Ha, 2012).

Ha and Ha (2012) suggested the following assumptions for Mann-Whitney test:

- The dependent variable must be measured in at least the ordinal scale, but the interval or ratio data can be "collapsed" into ordinal data via the ranking.
- It is only appropriate when you have an independent or between-groups design and two groups/conditions
- Random sampling is required

Blalock (1979) suggests that an alternative form of exactly the same test may be used with the normal approximation. Instead of obtaining *U* directly, *Z* statistics can be used to test for significance. To obtain the *Z*, compute the sum of the ranks for each of the samples. Compute the difference of the sums of the ranks for each sample, and subtract

from this difference a quantity representing the expected difference under the null hypothesis. Then divide the difference of differences, which is analogous to $(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)$, by the standard error to obtain Z. When Z is larger than the p value, the null hypothesis will be rejected (Blalock, 1979).

2.7.3. Wilcoxon signed-rank test

As with most nonparametric tests, the calculations for the Wilcoxon are quite simple. To compute Wilcoxon, rank the absolute values of the difference scores, then separate the ranks into two groups: those associated with positive differences and those associated with negative differences. Sum the ranks for each group. The smallest from these two sums is the test statistics for Wilcoxon test (Gravetter and Wallnau, 2004). It replaces the correlated (or paired) *t* test. This test considers both the magnitude and direction of effect, but it is not as powerful as the *t* test. This test is approximate when you have a nominal independent variable with two groups, but the dependent variable is at least on an ordinal scale (Ha and Ha, 2012).

Rice (2007) suggests that "if some of the differences are equal to zero, the most common technique is to discard those observations. If there are ties, each $|D_i|$ is assigned the average value of ranks for which it is tied. If there are too many ties, the significance level of the test is not greatly affected. If there are a large number of ties, modification must be made".

2.7.4. Analysis of Variance (ANOVA)

ANOVA "is used to test the null hypothesis that the means of three or more populations are the same against the alternative hypothesis that not all population means are the same" (Mann, 2001). An analysis of variance (ANOVA) test is performed using the F distribution. According to Mann (2001) the following are assumptions of one-way ANOVA:

- Samples are drawn from a normally distributed population.
- Samples drawn have the same variance (or standard deviation).
- Samples drawn from different populations are random and independent.

Ross (2005) suggests that ANOVA must be performed on a normal distribution with the same (though unknown) variance σ^2. When testing a null hypothesis with multiple

parameters the analysis of variance approach derives two estimators of the common variance σ^2. The first estimator is a valid estimator of the variance, while the second estimator is only valid when the null hypothesis is true, irrespective of whether the null hypothesis is true of not. In cases were the null hypothesis is not true, the second estimator will over estimates the variance σ^2. Anova will then compare the values of the two estimators and rejects the null hypothesis when the ratio of the second estimator is larger than the first estimator. When null hypothesis is true, the two estimators are close to each other, and the second estimator expected to be larger than the first estimator when null hypothesis is not true.

R-squared is always between 0 and 100%.

- 0% indicates that the model explains none of the variability of the response data around its mean.
- 100% indicates that the model explains all the variability of the response data around its mean."

According to Frost (2013) the following are key limitations of the R-squared:

- R-squared is unable to determine whether the coefficient estimates and predictors are biased.
- R-squared cannot indicate whether a regression model is adequate or not.

2.7.5. Kruskal-Wallis (H) test

This test assumes that the observations are independent, no assumptions of normality is made. Since the data is ranked, outliers have less influence on this nonparametric test compared to the *F* test. The sample variance among the group averages is an estimate of σ^2/n under the null hypothesis. Because the within Mean Squares also estimates σ^2, the two independent estimate of variance where obtained which is necessary for an *F* test from ratio (Dowdy et al., 2004).

Ha and Ha (2012) stated the following as assumptions to be met for the *H* test:

- The dependent variable (data) are measured on an ordinal or better scale
- The scores comes from the same underlying distributions

- have at least five scores per group to use the chi-square table
- Random sampling

2.7.6. Friedman test

Friedman's test like other nonparametric tests does not make assumptions of normality. The observations in different rows are independent, but the columns are not because of some unit of association. In order to avoid making the assumptions required for the classical analysis of variance test that the *n* treatments are the same, Friedman (1937, 1940) suggested that replacing each treatment observation within the i^{th} block by its rank relative to the other observations in the same block (Gibbons & Chakraborti, 2011).

Ott and Longnecker (2010) describes the conditions under which the Friedman test would be valid:

- The experimental design is a randomized block design, with *t* treatments randomly assigned to exactly one experimental unit per block, yielding *N* = tb responses.
- The *N* responses y_{ij} are mutually independent.
- The *N* responses are related by the model: $y_{ij} = \theta + \tau_i + \beta_j + \varepsilon_{ij}$ where θ is the overall median, τ_i is an effect due to the i^{th} treatment, β_j is an effect due to the j^{th} block, and the *N* ε_{ij} 's are a random sample from a continuous distribution within a median equal to 0.

2.7.7. Logistic regression

Regression methods are an important part of data analysis that are concerned with describing the relationship between a response variable and one or more explanatory variables. The goal of logistic regression is to find the best fitting and most parsimonious, yet reasonable model to describe the relationship between the dependent and response variable.

Lee and Peters (2016) states that in essence, logistic regression predicts the probability that some dichotomous outcome will occur for a given case, based on observations of whether or not that outcome did actually occur. So, for any case, the actual value of the outcome y must equal either 0 (did not occur) or 1 (did occur), and the predicted value of

y, denoted as either $P(y)$ or $\hat{y}$, will lie somewhere between 0 (no chance of the outcome occurring) and 1 (outcome will certainly occur). As such, it would be a useful indicator of how well the model works if these observed and predicted values were compared.

The Wald statistics has a chi-square distribution and thus allows testing for significance. In cases where the test is significantly different from zero, it is assumed that the predictor makes a significant contribution to the prediction outcome.

Lee and Peters (2016) states that the most important thing to look at is the 'Exp (B)' value for each predictor. This value indicates the change in offs (i.e. probability) resulting from a unit change in each predictor variable. It is far more useful than the b coefficient in interpreting a logistic regression.

In order to calculate the change in odds, first calculate the odds of the event happening given an unknown set of values for the predictors, using the equation (3.33) in Chapter 3. Then calculate the odds of the events happening given a single unit change in the value of the predictors, and then use the values obtained to calculate the percentage change in the odds.

2.7.8. Spearman's rank correlation coefficient

The Spearman's rank correlation is appropriate when one or both of the variable of interest are on an ordinal scale. This replaces the Pearson's correlation test when the data are not collected on an interval or ratio scale but are at least ordinal (Ha and Ha, 2012). Blalock (1979) describes the principle behind Spearman's measure by finding the difference between the two sets of score and squaring these differences and summing them up, and lastly manipulate the measure so that value will be +1.0 whenever the rankings are in the same, -1.0 when they are not the same, and zero if there is no relationship between the rankings. According to Bluman (2009), further attest that computing the ranks correlations is not as difficult as computing the ranks for the Pearson coefficient. r_s is computed by using these differences in the ranks. In cases where both data sets have the same ranks, r_s will be +1, when the data sets are ranked differently, r_s will be -1. For all the instances where there is no relationship between the ranks, r_s is always nearer to 0. The formula for r_s is derived from the formula of a product-moment correlation, and it is applied to the ranks, rather than raw scores (Blalock, 1979).

2.8. Conclusion

The focus of this study is to use bivariate techniques to analyse the relationship between education and causes of death of the youth in South Africa. Other socio-demographic variables (age, gender and province of death) will be analysed using the same bivariate techniques. The chapter focused on the education levels, causes of death and the relationships between the education levels and causes of death from different study around the globe. Several studies revealed that there is a relationship between education levels and causes of death, also inferred that age, gender, race/ethnic and province of death does have an impact on the causes of death. The study further considered methods which Stats SA together with DHA have approved as the methods for ascertaining death in South Africa, as well literature on bivariate analysis used in the study. The following chapter presents the different bivariate techniques used to investigate the relationship between education levels and causes of death, and the socio-demographic variables as well.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter discusses the methodologies applied to determine whether a relationship exists or not between education and causes of death and other socio-demographic variables such as gender, age, and death province. Due to the unavailability of the race category, it was not included in the analysis. The study uses bar charts and bivariate techniques to meet the object of the study. Data is analysed using SAS and SPSS software.

The rest of the chapter is organised as: Section 3.2 outlines the objectives and questions of the study, Section 3.3 is the data source, Section 3.4 presents graphical presentation of the data and section 3.5 entails bivariate analysis that will be performed to answer the questions posed by the study. The main objective of the study is to determine whether there is a relationship between causes of death and education level of the youth in South Africa. The other objective is to determine whether the socio-demographic variables (age, gender and province of death) have any link to the causes of death.

3.2 Research objectives and questions

3.2.1. Objectives of the study

The objectives of this study were stated as follows:

- To determine the relationship between causes of death and education level of the youth in South Africa.
- To determine the relationship between causes of death and other socio-demographic variable (age, gender and province of death) of the youth in South Africa.
- To examine the extent of the relationship between causes of death and education level, age, gender and province of death.

3.2.2. Research questions

Death is one predicaments the human race cannot escape. The youth of the country are either affected or infected by natural ailments or non-natural causes. The questions posed by this study are as follows:

- Is there a relationship between causes of death and education level of South African youth?
- Is there a relationship between socio-demographic variables (age, gender and province of death) and causes of death?
- To what extent is the relationship between the education level and causes of death?

3.3 Data source

The study utilises secondary data set sourced from Statistics South Africa (Stats SA). Stats SA sourced this data from the administrative records at the Department of Home Affairs (DHA). The researcher could not gather primary data due to the sensitivity and difficulty of collecting information of this nature. The study uses the data collected in 2014, which was processed (edited and imputed) by Stats SA.

The focus of this study is on the youth of South Africa and therefore the data used is for persons between the ages of 15 and 34 who died in 2014. Variable were coded during analysis, table 3.1 below shows how the variables were coded.

Table 3.1: Recoding and defining variables

Variable	Description
Age	15-19 = 1 20-24 = 2 25-29 = 3 30-34 = 4
Education level	No education = 0 Primary education = 1 (Grade R- Grade 7) Secondary education = 2 (Grade 8 – Grade 12)

	Higher education = 3 (University, Technicon, colleges, etc)

NB. All missing, not applicable, unknown, unspecified, etc. were removed from the data set.

3.4 Graphical presentation of the data

A graphic presentation reveal at a glance the characteristics of the data. This study utilises bar graphs to give better understanding of the data. Mann (2001) defines "a bar graph as a graph made of bars whose heights represents the frequencies of respective categories". Bar charts will be done for the following variables: age, gender, death province, education level and causes of death.

3.5 Bivariate methods

Bivariate statistics refers to the analysis of two variables where the desire is simply to examine the relationship between the two variables (Tabachnick and Fidell, 2007). A number of statistical test such as chi-square test, Wilcoxon-Mann-Whitney test, Wilcoxon signed-rank test, ANOVA, Kruskal-Wallis test, Friedman test, logistics regression and Spearman's rank correlation were performed to test whether a relationship exists between the dependent variable (causes of death) and independent variables (education level, age, gender and death province).

3.5.1. Chi-square test of independence

The chi-square test of independence determines whether there is a statistically significant relationship between categorical variables (Frost, 2017). The null hypothesis is that the relative proportions of education level and other independent variables such as age, gender and provinces of death are independent of the causes of death; and alternatively the proportions are dependent (McDonald, 2015). The hypothesis for χ^2 is as follows:

H_o: there is no relationship between socio-demographic variables and causes of death

H_a: there is a relationship between socio-demographic variables and causes of death

Note: Socio-demographic variable are education level, age, gender and province of death

A p-value less than or equal to the significant level which might me at either 0.05 or 0.01, indicates that there is sufficient evidence to conclude that the observed distribution is not the same as the expected distribution, concluding that there a relationship between the variables (Frost, 2017). The degree of freedom (df) for a test of independence are:

$$df = (R-1)(C-1) \tag{3.1}$$

where R and C are number of rows and columns, respectively (Mann, 2001).

Everitt (1977) suggests that the following formula be used for the χ^2 statistic:

$$\chi^2 = \Sigma \frac{(f_0 - f_e)^2}{f_e} \tag{3.2}$$

where f_0 is observed frequencies and f_e is expected frequencies.

Rice (2007) defines "statistical analysis of a sample size of *n* cross-classified in a table with *I* rows and *J* columns. Such a configuration is called a contingency table. The joint distribution of column n_{ij}, where $i = 1, \dots, I \; and \; j = 1, \dots, J$, is multinomial with cell probabilities denoted as π_{ij}. Let

$$\pi_{i.} = \sum_{j=1}^{J} \pi_{ij} \tag{3.3}$$

$$\pi_{.j} = \sum_{i=1}^{I} \pi_{ij} \tag{3.4}$$

Denote the marginal probabilities that an observation will fall in the i^{th} row and j^{th} column, respectively. If the row and column classifications are independent of each other.

$$\pi_{ij} = \pi_{i.}\pi_{.j} \tag{3.5}$$

Considering the following null hypothesis:

$$H_0: \pi_{ij} = \pi_{i.}\pi_{.j}, \quad i = 1, \dots, I \; and \; j = 1, \dots, J$$

versus the alternative that the π_{ij} are free. Under H_0, the mle of π_{ij} is

$$\hat{\pi}_{ij} = \hat{\pi}_{i.}\hat{\pi}_{.j}$$

$$= \frac{n_{i.}}{n} * \frac{n_{.j}}{n} \tag{3.6}$$

under the alternative, the mle of π_{ij} is

$$\tilde{\pi}_{ij} = \frac{n_{ij}}{n} \tag{3.7}$$

The estimates can be used to form a likelihood ratio test or an asymptotically equivalent Pearson's chi-square test,

$$X^2 = \sum_{i=1}^{I} \sum_{J=1}^{J} \frac{(O_{ij} - E_{ij})^2}{Eij} \tag{3.8}$$

Here the O_{ij} are the observed counts (n_{ij}). The expected counts, the E_{ij}, are the fitted counts:

$$X^2 = \sum_{i=1}^{I} \sum_{J=1}^{J} \frac{(n_{ij} - n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n} \tag{3.9}"$$

If there is a significant relationship between education level and causes of death; as well as other socio-demographic variables; age, gender and death province, the null hypothesis will therefore be rejected and the researcher can concluded that the variables are related. After finding the association between variables, it is important to test how strong is the relationship between those variable, and for this study the Cramer' V test is used. According to van den Berg (2016) Cramer' V is a number between 0 and 1, and it is used to indicate how strong the association is between two categorical variables. The following formula suggested by van den Berg (2016):

$$\phi_c = \sqrt{\frac{\chi^2}{N(k-1)}} \tag{3.10}$$

where

ϕ_c denotes the Cramer's V, χ^2 denotes the Chi-Square test, N is the sample size, and k is the lesser number of categories of either variable.

3.5.2. Wilcoxon-Mann-Whitney (U) test

The null hypothesis for the Mann-Whitney test states that there is no symmetric difference between the two populations from which the samples are selected. If the value of *U* is

very small, near zero, it is evidence that the two samples are very different and the null hypothesis will be rejected. In cases where *U* value is relatively large, the most likely outcome is that the two samples are similar. When sample data produce a *U* that is less than or equal to the critical value of α = 0.05 and α = 0.01, H_0 (null hypothesis) is rejected (Gravetter and Wallnau, 2004).

Blalock (1979) suggests that to obtain *U* when the number of cases is relatively large or if ties occur, the ranks of the separate samples be summed, and these sums of ranks be called R_1 and R_2. Then using the following formulas:

$$U = N_1N_2 + \frac{N_2(N_2+1)}{2} - R_2 \tag{3.11}$$

or

$$U' = N_1N_2 + \frac{N_1(N_1+1)}{2} - R_1 \tag{3.12}$$

3.5.3. Wilcoxon signed-rank test

Rice (2007) explains that if the sample size is greater than 20, a normal approximation to the null distribution can be used. The null hypothesis is that D_i are independent and symmetrically distributed about zero. The mean and variance of W_+ is then calculated.

$$E(W_+) = \frac{n(n+1)}{4} \tag{3.13}$$

$$Var(W_+) = \frac{n(n+1)(2n+1)}{24} \tag{3.14}$$

to facilitate the calculation, W_+ is presented in the following way

$$W_+ = \sum_{k=1}^{n} kI_k \tag{3.15}$$

where $I_k = \begin{cases} 1, \\ 0, \end{cases}$ if the *k*th largest $|D_i|$ has $D_i > 0$ or otherwise

Under H_0, the I_k are independent Bernoulli random variables with $p = \frac{1}{2}$, so

$$E(I_k) = \frac{1}{2}$$

$$Var(I_k) = \frac{1}{4}$$

therefore

$$E(W_+) = \frac{1}{2}\sum_{k=1}^{n} k = \frac{n(n+1)}{4} \quad (3.16)$$

$$Var(W_+) = \frac{1}{4}\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{24} \quad (3.17)$$

3.5.4. Analysis of Variance (ANOVA)

ANOVA "is used to test the null hypothesis that the means of three or more populations are the same against the alternative hypothesis that not all population means are the same" (Mann 2001). Blalock (1979) explains that in order to obtain a value of *F*, the ratio of the between and the within estimates, it is necessary to first calculate the total, between and within sum of squares. The formula for the total sum of squares (TSS) is same as the formula for the variance.

$$TSS = \sum_i \sum_j \left(X_{ij} - \bar{X}_{..}\right)^2 = \sum_i \sum_j X_{ij}^2 - \frac{\left(\sum_i \sum_j X_{ij}\right)^2}{N} \quad (3.18)$$

The formula for the between variation (BSS) is as follows:

$$BSS = \sum_j \frac{\left(\sum_i X_{ij}\right)^2}{N_j} - \frac{\left(\sum_i \sum_j X_{ij}\right)^2}{N} = \left[\frac{(\sum_i X_{i1})^2}{N_1} + \frac{(\sum_i X_{i2})^2}{N_2} + \cdots + \frac{(\sum_i X_{ik})^2}{N_k}\right] - \frac{\left(\sum_i \sum_j X_{ij}\right)^2}{N} \quad (3.19)$$

BSS is then subtracted from TSS, to obtain the within sum of squares (WSS)

$$WSS = TSS - BSS \quad (3.20)$$

Frost (2013) explains "R-squared as a statistical measure of how close the data are to the fitted regression line. It is also known as coefficient of determination. Frost (2013) further defines R-squared as:

$$R^2 = \frac{explained\ variation}{total\ variation} * 100\% \quad (3.21)$$

Levine et al. (2013) states that the "coefficient of determination is equal to the regression sum of squares (i.e. explained variation) divided by the total sum of squares (i.e. total variation)

$$r^2 = \frac{Regression\ sum\ of\ square}{total\ sum\ of\ square} = \frac{SSR}{TSS} \quad (3.22)$$

where

Computational formula for TSS is

$$TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} \quad (3.23)$$

Computational formula for SSR is

$$SSR = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 = b_0 \sum_{i=1}^{n} Y_i + b_1 \sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} \quad (3.24)$$

Computational formula for SSE is

$$SSE = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} Y_i^2 - b_0 \sum_{i=1}^{n} Y_i - b_1 \sum_{i=1}^{n} X_i Y_i \quad (3.25)"$$

3.5.5. Kruskal-Wallis (H) test

Mann (2001) suggests that when the populations being sampled are not normally distributed, ANOVA cannot be applied instead Kruskal-Wallis (*H)* test can be used. This test is not based on any assumption since it is nonparametric. The *H* test is always right-tailed. According to Rice (2007) the observation are assumed to be independent, but there are no assumptions that the distribution is of a particular form, for example normal distribution. Let

R_{ij} = the rank of Y_{ij} in the combined sample

Let

$$\overline{R_{i.}} = \frac{1}{J_i}\sum_{J=1}^{J_i} R_{ij} \quad (3.26)$$

be the average rank in the *i*th group. Let

$$\overline{R_{..}} = \frac{1}{N}\sum_{i=1}^{I}\sum_{J=1}^{J_i} R_{ij} \quad (3.27)$$

where *N* is the total number of observations. As in ANOVA, let

$$SS_B = \sum_{i=1}^{I} J_i \left(\overline{R_{i.}} - \overline{R_{..}}\right)^2 \quad (3.28)$$

be a measure of the dispersion of the $\overline{R_{i.}}$. SS_B may be used to test the null hypothesis that the probability distributions generating the observations are identical. Under the null hypothesis that the probability distributions of the *I* groups are identical, the test statistics

$$k = \frac{12}{n(n+1)} SS_B \tag{3.29}$$

is approximately distributed, as a chi-square random variable with $I - 1$ degrees of freedom. *K* can also be expressed as:

$$k = \frac{12}{n(n+1)} \left(\sum_{i=1}^{I} J_i \bar{R}_{i.}^{\ 2} \right) - 3(n+1) \tag{3.30}$$

3.5.6. Friedman test

Privitera (2012) defines this test as a test used to determine whether the total ranks in two or more groups are significantly different when the same participants are observed in each group. Friedman test is a nonparametric test alternative to one-way within-subjects ANOVA. The null hypothesis for the Friedman test is that the sum of ranks in each group does not differ. The alternative hypothesis is that the sum of ranks in each group differs. When the null hypothesis is true and n is greater than or equal to 5 per group, the test statistic for the Friedman test is approximately distributed as a chi-square distribution with k-1 degree of freedom, where k is the number of groups. According to Privitera (2012) the following formula can be used to perform the Friedman test:

$$\chi_R^2 = \frac{12}{nk(k+1)} \sum R^2 - 3N(K+1) \tag{3.31}$$

where *R* is total ranks in each group, *n* is the sample size in each group, and k is the number of groups.

3.5.7. Logistic regression

Dowdy et al. (2004) explains that one or more independent variables can also be used to predict a dependent variable that is nominal rather than numerical. According to Agresti (2009) the relationships between π (x) and x are usually non linear rather than linear. A constant change in the value of *x* may not necessarily have a high impact on π when it is

near 0 or 1 than when π is near the middle of its range. The corresponding logistic regression model formula is:

$$\log \frac{\pi(x)}{1-\pi(x)} = \alpha + \beta_x \tag{3.32}$$

The random component for the (success, failure) outcomes has a binomial distribution. The link function is the logit function log $[\frac{\pi}{1-\pi}]$ of π, symbolized by "logit" (π). The parameter β above determines the rate of increase or decrease of the curve.

Freund (2004) state that likelihood ratio tests are based on the generalisation of the Neyman-Pearson Lemma, which provide a means for constructing most powerful critical regions for testing a simple null hypothesis against a simple alternative hypothesis,

$H_0: \theta \epsilon \omega$

$H_1: \theta \epsilon \omega'$

where ω is a subset of Ω and $\omega'$is the complement of ω with respect to Ω. Thus, the parameter space for θ is partitioned into the disjoint sets ω and ω'; according to the null hypothesis, θ is an element of the first set, and according to the alternative hypothesis, it is an element of the second set. In most problems Ω is either the set of all real numbers, the set of all positive real numbers, some interval of real numbers, or a discrete set of real numbers. In general case, where at least one of two hypothesis is composite, instead two quantities $\max L_0$ and $\max L$, are compared, where $\max L_0$ is the maximum value of the likelihood function for all values of θ in ω, and $\max L$ is the maximum value of the likelihood function for all values of θ in Ω. In other words, if a random sample of size n from a population whose density at x is $f(x;\theta)$, then $\hat{\theta}$ is the maximum likelihood estimate of θ subject to the restriction that θ must be an element of ω, and $\hat{\hat{\theta}}$ is the maximum likelihood estimate of θ for all values of θ in Ω, then

$$\max L_0 = \prod_{i=1}^{n} f(x_i; \hat{\theta}) \tag{3.33}$$

and

$$\max L = \prod_{i=1}^{n} f\left(x_i; \hat{\hat{\theta}}\right) \tag{3.34}$$

These quantities are both values of random variables, since they depend on the observed values of $x_1, x_2, \ldots, x_n$ and their ratio

$$\lambda = \frac{\max L_0}{max\ L}$$

is referred to as a value of the likelihood ration statistics Λ (capital Greek *lambda*).

Since $\max L_0$ and $\max L$ are both values of a likelihood function and therefore are never negative, it follows that $\lambda \geq 0$; also, since ω is a subset of the parameter space Ω, it follows that $\lambda \leq 1$. When the null hypothesis is false, $\max L_0$ is expected to be small compared to $max\ L$, in which case λ would be close to zero. On the other hand, when the null hypothesis is true and $\theta\ \epsilon\ \omega$, it is expected that $\max L_0$ be close to $max\ L$, in which case λ would be close to 1. A likelihood ratio test states, therefore, that the null hypothesis H_0 is rejected if and only if λ falls in a critical region of the form $\lambda \leqq k$, where $0 < k < 1$.

Lee and Peters (2016) suggest that instead of predicting specific values of y from the predictors, predict the probability that y will occur given known values of the predictor variables. From the many number of way of arranging the logistic regression equation, the following equation is recommended:

$$P(y) = \frac{1}{1+e^{-(b_0+b_1x_1+b_2x_2+\cdots+b_ix_i)}} \tag{3.35}$$

The term e refers to the natural logarithm.

The log-likelihood is the measure used in logistic regression, and the formula for this is shown below:

$$log - likelihood = \sum_{i=1}^{N}\{y_i ln(P(y_i)) + (1 - y_i) ln[1 - P(y_i)]\} \tag{3.36}$$

Larger log-likelihood values indicate worse-fitting models. The best way to use a log-likelihood is to compare the value for the model with some baseline model.

According to Lee and Peters (2016) each predictor in logistic regression has a b coefficient and a standard error, both of which are used to calculate the Wald statistics, analogous to the t in the linear regression. The Wald statistics is calculated by:

$$Wald = \frac{b}{SE_b} \tag{3.37}$$

According to Hosmer and Lemeshow (2000), the Wald tests is used to estimate the confidence interval of the slope and intercept. The endpoints of a 100 (1-α)% confidence interval for the slope coefficient are:

$$\hat{\beta}_1 \pm z_{1-\alpha/2}\widehat{SE}(\hat{\beta}_1) \tag{3.38}$$

And for the intercept

$$\hat{\beta}_0 \pm z_{1-\alpha/2}\widehat{SE}(\hat{\beta}_0) \tag{3.39}$$

where $1-\alpha/2$ is upper $100(1-\alpha/2)\%$ points from the standard normal distribution and $\widehat{SE}(\cdot)$ denotes a model-based estimator of the standard error of the respective parameter estimator.

A test for the significance of a variable which does not require these computation is the Score test. The test is based on the distribution theory of the derivatives of the log-likelihood (Hosmer and Lemeshow, 2000). The test statistics for the Score test (ST) is

$$ST = \frac{\sum_{i=1}^{n} x_i(y_i-\bar{y})}{\sqrt{\bar{y}(1-\bar{y})\sum_{i=1}^{n}(x_i-\bar{x})^2}} \tag{3.40}$$

The odds of an event happening are formally defined as the probability of an event happening, divided by the probability of that event is not happening or in equation form:

$$odds = \frac{P(event)}{P(no\ event)} \tag{3.41}$$

Or the probability for no event is 1- the probability for an event or more specifically:

$$P(y) = \frac{1}{1+e^{-(b_0+b_1x_1+b_2x_2+\cdots+b_ix_i)}}$$

$$P(no\ y) = 1 - P(y) \tag{3.42}$$

Rice (2007) further states that if an event B has probability P(B) of occurring, the odds of B occurring are defined as:

$$odds(A) = \frac{P(A)}{1-P(A)} \tag{3.43}$$

And can be translated into the following:

$$P(A) = \frac{odds\ (A)}{1-odds(A)} \tag{3.44}$$

3.5.8. Spearman's rank correlation coefficient

The study uses the rank correlation coefficient to determine the degree of relationship between causes of death and education level, as well as age, gender and province of death. Runyon et al. (2000) states that the test requires that the differences in the ranks, squares and sum of squared differences be obtained, and the resulting values be substituted into the formula below:

$$r_s = 1 - \frac{6\,\Sigma D^2}{N(N^2-1)} \tag{3.45}$$

where: D^2 = differences between ranks, squared and N = number of pairs of data.

Huysamen (1976) explains that if the measurements for each two variable consist of the first *N* consecutive untied ranks (ranging from 1 through *N*, the formula for the Product moment correlation:

$$r_{XY} = \frac{N\sum XY-(\sum X)(\sum Y)}{\sqrt{[N\sum X^2-(\sum X)^2][N\sum Y^2-(\sum Y)^2]}} \tag{3.46}$$

simplifies to

$$r_s = \frac{3}{N-1}\left\{\frac{4\sum XY-N(N+1)^2}{N(N+1)}\right. \tag{3.47}$$

Where *XY* is the product of the ranks assigned to any given individual for the two variables. The data should therefore first be converted to ranks in order to compute the Spearman rank correlation.

3.6 Conclusion

This chapter has outlined and described all the bivariate techniques that will be adopted to determine the relationship between causes of death and socio-demographic variables such as education level, gender, age and province of death of the youth in South Africa in 2014. It further outlined the methods that will assist in revealing the degree and strength of the relationships through application of the Spearman's rank correlation coefficient and logistics regression. The next chapter presents the data analysis and interpretation of results.

CHAPTER 4

DATA ANALYSIS AND INTERPRETATION OF RESULTS

4.1 Introduction

The focus of this chapter is the analysis of data amassed and interpreting the results based on the methodology outlined in the previous chapter. The chapter commences with a graphical presentation of the data which will assist in better comprehension of the data. The main aim of the study has been stated as the attempt to determine whether there is a relationship between causes of death and education level of the youth that died in 2014. The study will further interrogate other socio-demographic variables such as gender, province of death and age in completed years.

Statistical test such as chi-square test, Wilcoxon-Mann-Whitney test, Wilcoxon signed rank test, ANOVA, Kruskal-Wallis test, Friedman test, Spearman's rank correlation and logistic regression were performed to test whether a relationship exist between the dependent variable and independent variables. Chi-Square test is used to evaluate whether or not the frequencies which have been empirically obtained differ significantly from those which would be expected under a certain set of theoretical assumptions (Blalock, 1979). The Wilcoxon-Mann-Whitney and Wilcoxon signed rank test will be performed to put more emphasis on the results of the chi-square test. ANOVA is similar to the Chi-Square test, it is used to determine whether or not one or more of the conditions has different effects than the other conditions. The Kruskal-Wallis and Friedman test are nonparametric tests that were used to further test the independence of the variables. Logistic regression is used to describe the relationship between dependent and independent variable. Lastly, the Spearman's rank correlation coefficient is used to determine the degree of the relationship between the variables.

4.2 Graphical presentation of the data

This section focuses on the graphical presentation of the data. The graphs will assist in understanding the data better, and simplified understanding of the variables used in the study. The following figure presents the age variable.

Figure 4.1 Age in completed years

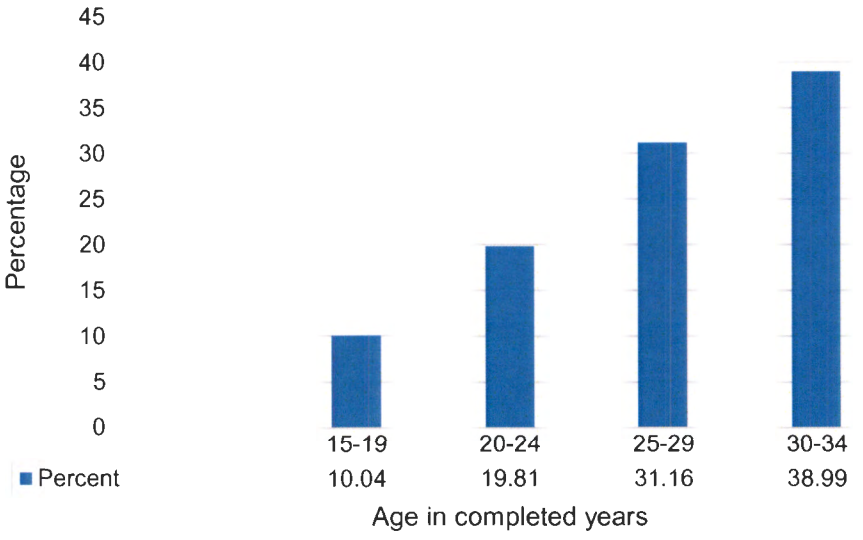
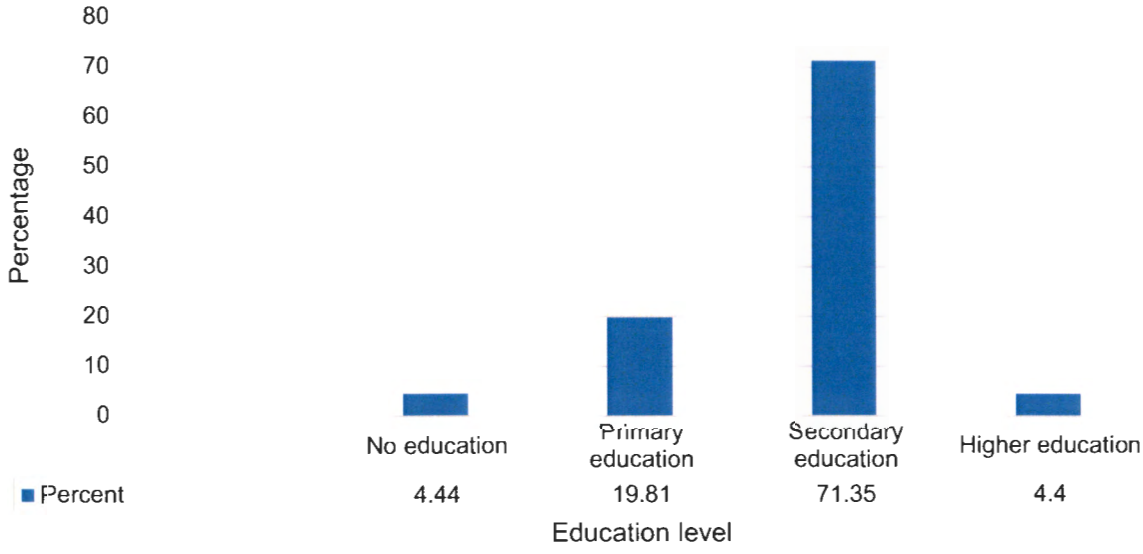


Figure 4.1 above illustrates the age in completed years of the youth at the time of their death. The above table shows a high percentage of 38.99% and 31.16% of the youth aged between 30-34 and 25-29 respectively. The age group that registered a low percentage of the death were the group 15-19 at 10.04%, followed by 20-24 at 19.81%.

Figure 4.2 Education level



Youth with secondary education registered the highest number of deaths in 2014 (71.35%) as shown above by Figure 4.2. The second highest were the youth with primary education at 19.81%. The groups that registered the lowest deaths were those with no education and higher education, at 4.44% and 4.40%, respectively.

Figure 4.3 Gender distribution

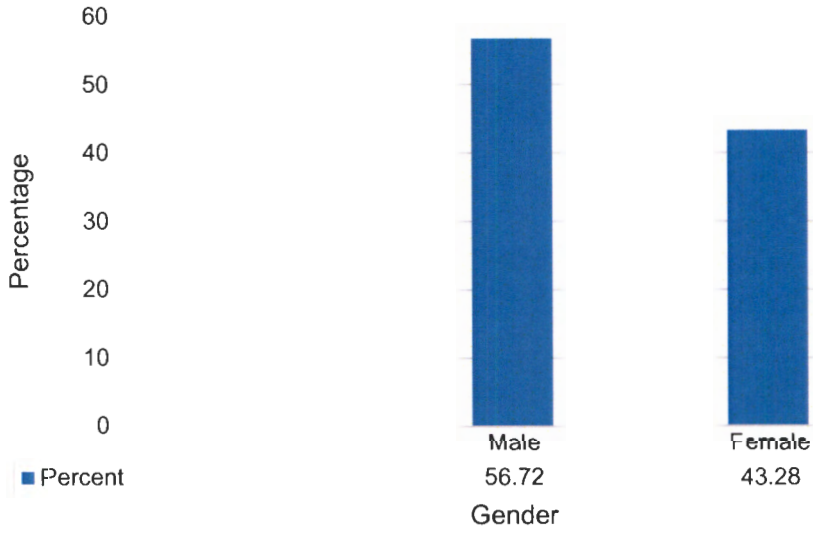


Figure 4.3 above demonstrates that more deaths were registered for the males as compare to female counterparts. The deaths registered for males were 56.72%, while those of females stood at 43.28%.

Figure 4.4 Causes of death

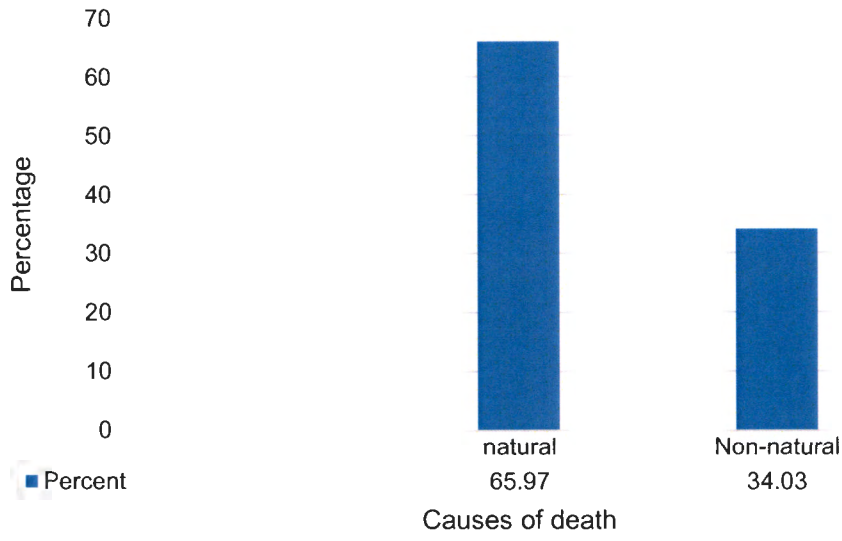


Figure 4.4 above reflects that a considerable number of the youth died from natural causes of death. The death registered for natural causes of death were 65.97% while those of non-natural death were 34.03.

Figure 4.5 Province of death

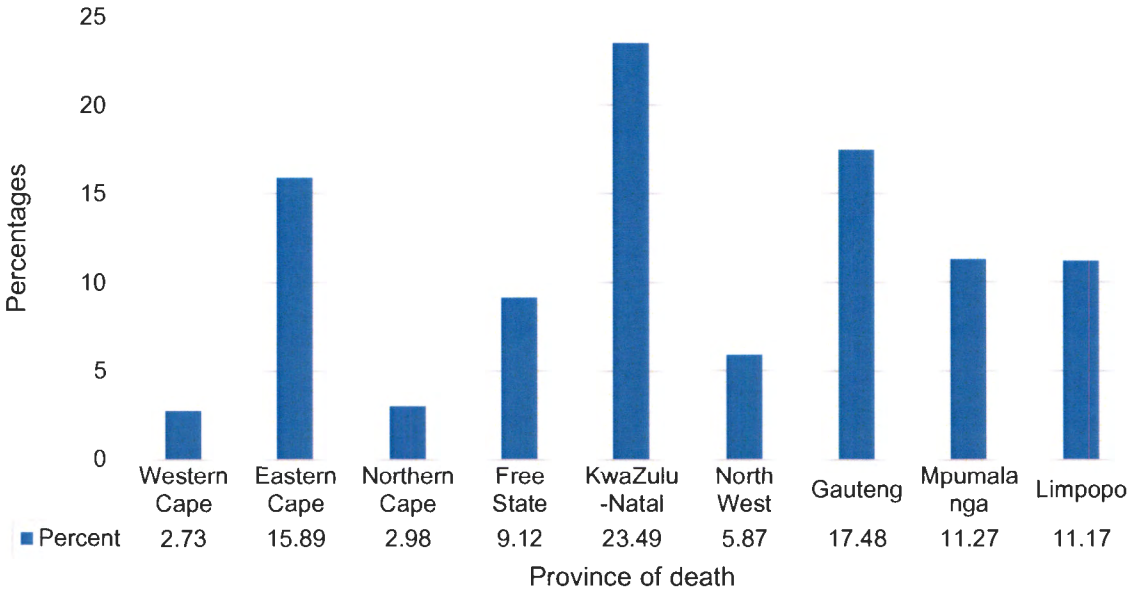


Figure 4.5 indicates the provinces where the deaths occurred. KwaZulu-Natal, Gauteng and Eastern Cape registered the highest deaths of the youth in 2014. Provinces such as Western Cape and Northern Cape registered the lowest. In Mpumalanga and Limpopo almost the same percentage of deaths were registered. Free State and North West registered 9.12% and 5.87%, respectively.

4.3 Bivariate techniques results

Bivariate data refers to a process where two variables are studied. It is used to explore if there is a relationship between two sets of values, usually denoted as variables X and Y (Andale, 2015). The following test will be carried out to determine whether a relationship exists between causes of death and education level of the youth. Other socio-demographic variable such as Gender, Province of death and age of the youth will be tested as well. This section will commence by performing the chi-square test, Wilcoxon-Mann-Whitney test, Wilcoxon signed rank test, ANOVA, Kruskal-Wallis test, Friedman test, logistic regression and Spearman's rank correlation in that order.

4.3.1. Chi-Square test of independence

This is a test procedure followed to derive the significance test in statistics. The following tests for each independent variable will be based on a hypothesis test. The hypothesis for χ^2 is such that:

H_o: There is no relationship between socio-demographic variable and causes of death

H_1: There is relationship between socio-demographic variable and causes of death

Table 4.1: Chi-Square test results for education and causes of death

Statistics	DF	Value	Prob
Chi-Square	3	351.3378	<0.0001
Likelihood Ratio Chi-Square	3	364.0498	<0.0001
Mantel-Haenszel Chi-Square	1	347.0032	<0.0001
Phi Coefficient		0.0886	
Contingency Coefficient		0.0882	
Cramer's V		0.0886	

The result obtained in Table 4.1 shows that there is a statistically significant relationship between education and causes of death. The χ^2 is 351.338 with df = 3 and the corresponding p-value is 0.0001, which indicates that there is evidence that education

and causes of death are related. The Cramer's V coefficient which is 0.089 indicate a very weak positive relationship between education level and causes of death of the youth.

Table 4.2: Chi-Square test results for gender and causes of death

Statistics	DF	Value	Prob
Chi-Square	1	6541.8548	<0.0001
Likelihood Ratio Chi-Square	1	7023.0126	<0.0001
Continuity Adj. Chi-Square	1	6540.2270	<0.0001
Mantel-Haenszel Chi-Square	1	6541.7087	<0.0001
Phi Coefficient		-0.3822	
Contingency Coefficient		0.3570	
Cramer's V		-0.3822	

Table 4.2 above displays that there is a statistically significant relationship between gender and causes of death of the youth. The χ^2 is 6541.855 with 1 as a df and the corresponding p-value is 0.0001 indicating that there is a relationship between gender and causes of death. The Cramer's V coefficient is -0.382 which indicates a moderate negative relationship between the gender and causes of death.

Table 4.3: Chi-Square test results for age and causes of death

Statistics	DF	Value	Prob
Chi-Square	3	2005.5227	<0.0001
Likelihood Ratio Chi-Square	3	2009.9532	<0.0001
Mantel-Haenszel Chi-Square	1	1856.9525	<0.0001
Phi Coefficient		0.2116	
Contingency Coefficient		0.2070	
Cramer's V		0.2116	

The results in the above table, indicates that there is a statistically significant relationship between age and causes of deaths of the young people in South Africa. The χ^2 (3) is 2005. 523, $p < 0.0001$ provides evidence that the relationship does exist between age

and causes of death of the youth. However, a weak positive associated is shown by the Cramer's V coefficient (0.212) between the age and causes of death.

Table 4.4: Chi-Square test results for province of death and causes of death

Statistics	DF	Value	Prob
Chi-Square	8	695.3291	<0.0001
Likelihood Ratio Chi-Square	8	681.6803	<0.0001
Mantel-Haenszel Chi-Square	1	5.3096	0.0212
Phi Coefficient		0.1246	
Contingency Coefficient		0.1236	
Cramer's V		0.1246	

Table 4.4 above shows that the provinces were the youth died and their causes of deaths are statistically significantly associated. The χ^2 of 695.329 with df of 8 and the corresponding p-value of 0.0001 indicates that indeed the province of death and causes of deaths are related. The Cramer's V coefficient (0.125) indicates a very weak positive relationship between the two variables but further tests will be done to determine the strength of the relationship.

4.3.2. Wilcoxon-Mann-Whitney test

This test is used to determine whether the total ranks in two independent groups are significantly different. The tests will be based on the following null hypothesis:

H_o: ranks in two groups are equally dispersed

H_1: ranks in two groups are not equally dispersed

Table 4.5: Wilcoxon-Mann-Whitney test results for education and causes of death

Ranks

	Causes of death	N	Mean Rank	Sum of Ranks
Education	1	29550	21778.92	6.44×10^8
	2	15244	23596.59	3.60×10^8
	Total	44794		

Test statistics

	Education
Mann-Whitney U	2.07×10^8
Wilcoxon W	6.44×10^8
Z	-17.776
Asymp. Sig. (2-tailed)	0.000

From the result in Table 4.5 above, it can be concluded that education and causes of death have a significant relationship (U = 0.0000002, p = 0.000). The obtained *Z* value of -17.776 which exceeds the critical value of -1.96 ($\alpha = 0.05$) since this is two tailed. The decision is therefore to reject the null hypothesis. Education is vital for the survival of the youth from the causes of death.

Table 4.6: Wilcoxon-Mann-Whitney test results for gender and causes of death

Ranks

	Causes of death	N	Mean Rank	Sum of Ranks
Gender	1	29550	25443.38	7.52×10^8
	2	15244	23596.59	2.51×10^8
	Total	44794		

Test statistics

	Gender
Mann-Whitney U	1.35×10^8
Wilcoxon W	2.51×10^8
Z	-80.881
Asymp. Sig. (2-tailed)	0.000

The preceding results indicate that the gender of the youth counts where causes of death are concerned. It can therefore be concluded that gender and causes of death have a significant relationship (U = 0.0000001, p = 0.000). The table above shows that the obtained *Z* value of -80.881 exceeds the critical value of -1.96 (α = 0.05) since this is two tailed. The decision is therefore to reject the null hypothesis.

Table 4.7: Wilcoxon-Mann-Whitney test results for age and causes of death

Ranks

	Causes of death	N	Mean Rank	Sum of Ranks
Age	1	29550	24232.73	7.16×10^8
	2	15244	18839.97	2.87×10^8
	Total	44794		

Test statistics

	Age
Mann-Whitney U	1.71×10^8
Wilcoxon W	2.87×10^8
Z	-44.042
Asymp. Sig. (2-tailed)	0.000

Table 4.7 above shows the results that *U* is 0.0000002, p = 0.000 which indicates that age and causes of death are associated. The obtained *Z* value of -44.042 exceeds the critical value of -1.96 (α = 0.05). The decision is therefore to reject the null hypothesis. The analysis shows that age of the youth has a significant impact on the causes of death of the youth.

Table 4.8: Wilcoxon-Mann-Whitney test results for province of death and causes of death

Ranks

	Causes of death	N	Mean Rank	Sum of Ranks
Province of death	1	29550	22498.61	6.65×10^8
	2	15244	22201.50	3.38×10^8
	Total	44794		

Test statistics

	Province of death
Mann-Whitney U	2.22×10^8
Wilcoxon W	3.38×10^8
Z	-2.335
Asymp. Sig. (2-tailed)	0.000

The obtained Z value of -2.335 exceeds the critical value of -1.96 ($\alpha = 0.05$); it reached the rejection region. The decision is therefore to reject the null hypothesis. The analysis shows that provinces of death had a major influence on the causes of death of the youth. *U* is 0.0000002, p = 0.000 which indicates that age and causes of death are associated.

4.3.3. Wilcoxon signed-rank test

This test is used to determine whether the total ranks in two related groups are significant (Privitera. 2012). The tests will be based on the following hypothesis:

H_0: there is no difference in ranks between groups

H_1: there is a difference in ranks between groups

Table 4.9: Wilcoxon signed-rank test results for education and causes of death

Ranks

		N	Mean Rank	Sum of Ranks
Education-causes of death	Negative	4604[a]	14041.18	64645602.00
	Positive	22576[b]	13498.59	3.05×10^8
	Ties	17614[c]		
	Total	44794		

Note a. education < causes of death

b. education > causes of death

c. education = causes of death

Test Statistics

	Education-causes of death
Z	-104.430[a]
Asymp. Sig. (2-tailed)	0.000

a. Based on negative ranks

Findings in Table 4.9 shows there is a significant difference between education and causes of death of the youth. The obtained *z* value is – 104.430, p is 0.000. The decision is to reject the null hypothesis and conclude that education plays a significantly role reducing the causes of death among the youth.

Table 4.10: Wilcoxon signed-rank test results for gender and causes of death

Ranks

		N	Mean Rank	Sum of Ranks
causes of death –gender	Negative	16808[a]	14737.00	2.48×10^8
	Positive	12665[b]	14737.00	1.87×10^8
	Ties	15321[c]		
	Total	44794		

a. causes of death < gender

b. causes of death > gender

c. causes of death = gender

Test Statistics

	Causes of death - gender
Z	-24.133[a]
Asymp. Sig. (2-tailed)	0.000

a. Based on positive ranks

Table 4.10 show that the *z* value is – 24.133 which exceeds the critical value. The decision is to reject the null hypothesis. The analysis shows that gender has a significant impact on the causes of death amongst the youth.

Table 4.11: Wilcoxon signed-rank test results for age and causes of death

Ranks

		N	Mean Rank	Sum of Ranks
Age – causes of death	Negative	2132[a]	5769.00	12299508.00
	Positive	36047[b]	19877.87	7.17×10^8
	Ties	6615[c]		
	Total	44794		

a. age < causes of death

b. age > causes of death

c. age = causes of death

Test Statistics

	Age - causes of death
Z	-165.856[a]
Asymp. Sig. (2-tailed)	0.000

a. Based on negative ranks

Table 4.11 show that the obtained value of *z* is – 165.856 which exceeds the critical value. The analysis shows that age of the youth significantly has an impact in the causes of death. The decision is therefore to reject the null hypothesis.

Table 4.12: Wilcoxon signed-rank test results for province of death and causes of death

Ranks

		N	Mean Rank	Sum of Ranks
Causes of death – province of death	Negative	41332[a]	21185.12	8.76×10^8
	Positive	555[b]	2987.00	1657785.00
	Ties	2907[c]		
	Total	44794		

a. causes of death < province of death

b. causes of death > province of death

c. causes of death = province of death

Test Statistics

	Province of death - causes of death
Z	-177.015[a]
Asymp. Sig. (2-tailed)	0.000

a. Based on positive ranks

Table 4.12 show that the obtained value of *z* (– 177.015) exceeds the critical value. The decision is to reject the null hypothesis. Province were the youth died from significantly has an impact in the causes of death of the youth.

4.3.4. ANOVA

Anova is utilised to test whether there is a relationship between mean causes of death and socio-demographic variable. The following are results after analysis. The hypothesis test for ANOVA is as follows:

H_o: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ (all socio-demographic variable are related to causes of death)
H_1: not all socio-demographic variables are related to causes of death

Table 4.13: ANOVA results for education and causes of death

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	78.87542	26.29181	118.03	<0.0001
Error	44790	9977.38658	0.22276		
Corrected Total	44793	10056.26200			

R-Square	Coeff Var	Root MSE	Causes of death Mean
0.007843	35.21369	0.471974	1.340313

Table 4.13 above shows the p value is .0001 which is below significance level of 0.05 indicating that there is a statistically significant relationship between education and causes of death. The F-value is 118.03. The null hypothesis is therefore rejected

indicating that there is a relationship between causes of death and education level of the youth in South Africa and conclude that our model explains a statistically significant proportion of the variance between causes of death and education level of the youth in South Africa. The proportion of total variance explained by the model is 0.008.

Table 4.14: ANOVA results for gender and causes of death

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1468.64771	1468.64771	7660.30	<0.0001
Error	44792	8587.61429	0.19172		
Corrected Total	44793	10056.26200			

R-Square	Coeff Var	Root MSE	Causes of death Mean
0.146043	32.66854	0.437861	1.340313

The above table indicates that there is a statistically significant relationship between gender and causes of death of the youth that died in 2014. The probability value of 0.0001 is below 0.05 level of significance indicating a relationship do exist between the gender and causes of death. The F-value is 7660.30, the null hypothesis is therefore reject and conclude that a relationship does exist between the gender of the youth and their causes of death. The R-Square which explains the proportion of the total variance is 0.146.

Table 4.15: ANOVA results for age and causes of death

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	450.24026	150.08009	699.78	<0.0001
Error	44790	9606.02174	0.21447		
Corrected Total	44793	10056.26200			

R-Square	Coeff Var	Root MSE	Causes of death Mean
0.044772	34.55213	0.463107	1.340313

It is indicated from the above table that the p value is 0.0001 which is below significance level of 0.05 indicating that there is a statistically significant relationship between age and causes of death. The F-value is 699.78, therefore the null hypothesis is rejected and conclude that there is a relationship between causes of death and age of the youth. It is therefore concluded that our model explains a statistically significant proportion of the variance of 0.045 between causes of death and age of the youth in South Africa.

Table 4.16: ANOVA results for province of death and causes of death

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	156.10152	19.51269	88.27	<0.0001
Error	44785	9900.16048	0.22106		
Corrected Total	44793	10056.26200			

R-Square	Coeff Var	Root MSE	Causes of death Mean
0.15523	35.07910	0.470170	1.340313

Table 4.16 shows that the F-value is 88.27 and the p-value is found to be 0.0001 which is below 5% level of significance. This therefore indicates that there is a statistically significant relationship between province of death and the causes of deaths of the youth. The total percentage of the total variance is 0.155. It is then concluded that a relationship does exist between the province of death and causes of death of the youth.

4.3.5. Kruskal-Wallis test

Privitera (2012) describes the Kruskal-Wallis as a test used to determine whether the total ranks in two or more independent groups are significantly different. The test is based on the following null hypothesis:

H_o: sum ranks in each group does not differ

H_1: sum ranks in each group differs

Table 4.17: Kruskal-Wallis test for education and causes of death

Ranks

Causes of death		N	Mean Rank
Education	1	29550	21778.92
	2	15244	23596.59
	Total	44794	

Test Statistics

	Education
Chi-Square	315.991
Df	1
Asymp. Sig.	0.000

Kruskal-Wallis *H* test in Table 4.17 above shows that there is a statistically significant differences in education between the different causes of death, χ^2 (1 df) is 315.991, p = 0.000 with mean rank education of 21778.92 for natural causes of death and 23596.59 for non-natural causes of death.

Table 4.18: Kruskal-Wallis test for gender and causes of death

Ranks

	Causes of death	N	Mean Rank
Gender	1	29550	25443.38
	2	15244	16493.15
	Total	44794	

Test Statistics

	Gender
Chi-Square	6.542E3
Df	1
Asymp. Sig.	0.000

Table 4.18 above shows that there is a statistically significant differences in gender of the youth between the different causes of death, χ^2 (1 df) is 0.006, p = 0.000 with mean rank gender of 25443.38 for natural causes of death and 16493.5 for non-natural causes of death.

Table 4.19: Kruskal-Wallis test for age and causes of death

Ranks

	Causes of death	N	Mean Rank
Age	1	29550	24232.73
	2	15244	18839.97
	Total	44794	

Test Statistics

	Age
Chi-Square	1.940E3
Df	1
Asymp. Sig.	0.000

Kruskal-Wallis *H* test in Table 4.19 above is χ^2 (1 df) is 0.002, p = 0.000 which shows that there is a statistically significant differences in age of the youth between the different causes of death, with mean rank age of 24232.73 for natural causes of death and 18839.97 for non-natural causes of death.

Table 4.20: Kruskal-Wallis test for province of death and causes of death

Ranks

	Causes of death	N	Mean Rank
Province of death	1	29550	22498.61
	2	15244	22201.50
	Total	44794	

Test Statistics

	Province of death
Chi-Square	5.451
Df	1
Asymp. Sig.	0.020

Table 4.20 shows results of the Kruskal-Wallis *H* test. The results above shows that there is a statistically significant differences in provinces of death where the youth died, and their different causes of death, χ^2 (1 df) is 5.451, p = 0.020 with mean rank province of death of 22498.61 for natural causes of death and 22201.50 for non-natural causes of death.

4.3.6. Friedman test

Privitera (2012) state that the Friedman test is used to determine whether the total ranks in two or more groups are significantly different when the same participants are observed in each group.

Table 4.21: Friedman test results

Ranks

	Mean Rank
Education	1.92
Gender	1.51
Age	2.89
Province of death	3.68

Test Statistics	
N	44794
Chi-Square	8.79×10^4
Df	3
Asymp. Sig.	0.000

Table 4.21 above shows that the Friedman test is significant; χ_R^2 is 0.0009 with df = 3 indicating a strong differences among causes of death and the socio-demographic variable (education, gender, age and province of death). Since p value is 0.000, we reject the null hypothesis and conclude that there is evidence that the impact of education, gender, age and province of death differs with regards to the causes of death.

4.3.7. Logistic regression

Logistic regression is used to describe the relationship between causes of death and socio-demographic variables. The relationships are explained in the following tables.

Table 4.22: Logistic regression results for education and causes of death

Model fit statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	57449.692	57091.642
SC	57458.402	57126.482
-2 Log L	57447.692	57083.642

Testing Global Null Hypothesis: BETA = 0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	364.0498	3	<0.0001
Score	351.3378	3	<0.0001
Wald	343.4624	3	<0.0001

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Education	3	343.4624	<0.0001

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimates**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-0.2599	0.0454	32.7725	<0.0001
Education	0	1	-1.0847	0.0716	229.4134	<0.0001
Education	1	1	-0.6120	0.0510	143.8155	<0.0001
Education	2	1	-0.3356	0.0469	51.2228	<0.0001

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
Education 0 vs 3	0.338	0.294	0.389
Education 1 vs 3	0.542	0.491	0.599
Education 2 vs 3	0.715	0.652	0.784

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	26.1	Somers' D	0.081
Percent Discordant	18.0	Gamma	0.184
Percent Tied	55.9	Tau-a	0.036
Pairs	450460200	c	0.541

The value of the parameter estimates for no education, primary education and secondary education indicates that the log-odds of youth not dying from the causes of death decreases (because the estimates are negative) by -1.085, -0.612 and -0.336

respectively as compared to those with higher education. The results are significant because of the Wald Chi- Square with 3 df is 343.462 with $p < 0.0001$. The odds ratio indicates that the odds of dying from the causes of death are 33.8, 54.2 and 71.5 times higher for the youth with no education, primary education and secondary education than the youth with higher education. It is evident that youth with higher education have better chances of not dying from the causes of death than the others. The Likelihood ratio chi-square is 364.0498 with p-value of 0.0001 shows that the model fits significantly and that applies to the Score and Wald test. The analysis of effect shows that the education and causes of death are significantly related. This indicates that there is evidence that people with no education are most likely to die from these causes of death.

Table 4.23: Logistic regression results for gender and causes of death

Model fit statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	57449.692	50428.680
SC	57458.402	50446.099
-2 Log L	57447.692	50424.680

Testing Global Null Hypothesis: BETA = 0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	7023.0124	1	<0.0001
Score	6541.8548	1	<0.0001
Wald	5772.3099	1	<0.0001

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Gender	1	5772.3099	<0.0001

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimates**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-1.8742	0.0211	7855.3970	<0.0001
Gender	1	1	1.8681	0.0246	5772.3099	<0.0001

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
1 vs 2	6.476	6.171	6.796

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	47.3	Somers' D	0.400
Percent Discordant	7.3	Gamma	0.733
Percent Tied	45.4	Tau-a	0.179
Pairs	450460200	c	0.700

The odds ratio (64.76) above shows that the odds of dying from the causes of death are 64.76 times higher for males than for females. It is thus evident that there is a strong relationship between males and causes of death than with females. The Likelihood ratio chi-square is 7023.012 with p-value of 0.0001 shows that the model fits significantly. The Score and Wald and Score test also shows that the model fits significantly. The analysis of effect shows that the gender and causes of death are significantly related. The results of the Wald Chi-Square with 1 df is 5772.310, $p < 0.0001$ indicates significance of the estimates. The value of the parameter estimate for gender indicates that the log-odds of dying from the causes of death increase by 1.868 for males compared to females.

Table 4.24: Logistic regression results for age and causes of death

Model fit statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	57449.692	55445.739
SC	57458.402	55480.578
-2 Log L	57447.692	55437.739

Testing Global Null Hypothesis: BETA = 0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	2009.9532	3	<0.0001
Score	2005.5227	3	<0.0001
Wald	1946.5036	3	<0.0001

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Age	3	1946.5036	<0.0001

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**estimates**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-1.1881	0.0179	4414.0020	<0.0001
Age	1	1	1.0848	0.0348	971.0952	<0.0001
Age	2	1	1.1040	0.0278	1580.2188	<0.0001
Age	3	1	0.5362	0.0253	450.6334	<0.0001

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
Age 1 vs 4	2.959	2.764	3.168
Age 2 vs 4	3.016	2.856	3.185
Age 3 vs 4	1.709	1.627	1.796

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	48.4	Somers' D	0.242
Percent Discordant	24.2	Gamma	0.333
Percent Tied	27.3	Tau-a	0.108
Pairs	450460200	c	0.621

The Likelihood ratio chi-square is 2009.953 with p-value of 0.0001 indicates that the model fits significantly, and the results of the Score and Wald test are also supporting that. The analysis of effect shows that the age and causes of death are significantly related. The Wald Chi-Square with 3 df is 1946.504, $p < 0.0001$ is significant. The value of parameter estimates for age shows that the log-odds of the youth dying from causes of death increases by 1.085, 1.104 and 0.536 for ages 15-19, 20-24 and 25-29, respectively as compare to those in their 30s. This results are significant. The odds ratio indicates that the odds of dying from the causes of death are 29.59, 30.16 and 17.09 are higher for the other age categories than those in the 30-34 age category. The results indicate that a relationship does exist between age and causes of death

Table 4.25: Logistic regression results for province of death and causes of death

Model fit statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	57449.692	56784.012
SC	57458.402	56862.400
-2 Log L	57447.692	56766.012

Testing Global Null Hypothesis: BETA = 0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	681.6803	8	<0.0001
Score	695.3291	8	<0.0001
Wald	685.6404	8	<0.0001

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Province of death	8	685.6404	<0.0001

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**estimates**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-0.9199	0.0313	862.8421	<0.0001
Province of death	1	1	0.7331	0.0654	125.6557	<0.0001
Province of death	2	1	0.1403	0.0404	12.0648	0.0005
Province of death	3	1	0.5282	0.0640	68.1111	<0.0001

Province of death	4		0.1685	0.0459	13.4902	0.0002
Province of death	5		0.2210	0.0375	34.6728	<0.0001
Province of death	6		0.0177	0.0532	0.1106	0.7395
Province of death	7		0.7066	0.0387	333.3599	<0.0001
Province of death	8		-0.0118	0.0442	0.0710	0.7898

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
Province of death 1 vs 9	2.082	1.831	2.366
Province of death 2 vs 9	1.151	1.063	1.245
Province of death 3 vs 9	1.696	1.496	1.922
Province of death 4 vs 9	1.184	1.082	1.295
Province of death 5 vs 9	1.247	1.159	1.343
Province of death 6 vs 9	1.018	0.917	1.130
Province of death 7 vs 9	2.027	1.879	2.187
Province of death 8 vs 9	0.988	0.906	1.078

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	49.4	Somers' D	0.135
Percent Discordant	35.8	Gamma	0.159
Percent Tied	14.8	Tau-a	0.061
Pairs	450460200	c	0.568

The odds ratio estimate for province of death 1 vs 3 is 2.082, indicating that the odds of dying from causes of death for the youth in Western Cape (Province 1) are 2.082 of the odds of dying from the causes of death for the youth in Limpopo (Province 9), the same is applicable to Gauteng (Province 7) with 2.027 of the odds. The odds for other provinces were also over 100% indicating that the odds of the youth dying in other province where higher than those dying in Limpopo. Only Mpumalanga (Province 8) showed the odds less than 100%. The Wald Chi-Square with 8 df is 685.640, $p < 0.0001$ which indicates that the results are significant. The result of the Likelihood ratio chi-square is 681.680 with p-value of 0.0001 shows that the model fits significantly. The Score and Wald tests also shows that the model does fit significantly. The analysis of effect shows a significant relationship between province of death and causes of death.

4.3.8. Spearman's rank correlation coefficient

In order to determine the degree of relationship between causes of death and education, Spearman Rank Correlation will be used, which is a non-parametric test used to measure the degree of association between two variables. The test is computed for other socio-demographic variables as well.

Table 4.26: Spearman's rank correlation results for education and causes of death

Spearman Correlation Coefficients, N = 44794 Prob > |r| under H0: Rho=0		
	Causes of death	Education
Causes of death	1.00000	0.08399
		<0.0001
Education	0.08399	1.00000
	<0.0001	

Table 4.26 shows the Spearman ranks correlation coefficient which indicates a very weak positive relationship of 0.084 between education and causes of death of the youth who died in 2014. The p value is 0.0001 which is evidence that education and causes of death of the youth are correlated.

Table 4.27: Spearman's rank correlation results for gender and causes of death

Spearman Correlation Coefficients, N = 44794 Prob > |r| under H0: Rho=0		
	Causes of death	Gender
Causes of death	1.00000	-0.38216
		<0.0001
Gender	-0.38216	1.00000
	<0.0001	

The p value in Table 4.27 indicates that gender and causes of death are correlated. The results above shows a moderate negative relationship of -0.382 between gender and causes of death of the youth who died in 2014.

Table 4.28: Spearman's rank correlation for age and causes of death

Spearman Correlation Coefficients, N = 44794 Prob > |r| under H0: Rho=0		
	Causes of death	Age
Causes of death	1.00000	-0.20810
		<0.0001
Age	-0.20810	1.00000
	<0.0001	

The above Table 4.28 shows that age and causes of death have a weak negative relationship. Spearman's rank correlation coefficient is -0.208, which is evident that the two variable do not have a perfect linear relationship. The p value of 0.0001 is evidence that age and causes of death of the youth are correlated.

Table 4.29: Spearman's rank correlation results for province of death and causes of death

Spearman Correlation Coefficients, N = 44794 Prob > |r| under H0: Rho=0		
	Causes of death	Province of death
Causes of death	1.00000	-0.01103
		0.0196
Province of death	-0.01103	1.00000
	0.0196	

Table 4.29 above shows that province of death and causes of death have a very weak negative relationship. Spearman's rank correlation coefficient is -0.011. The p value is 0.0196 which indicate that province of death and causes of death of the youth are correlated.

4.4 Conclusion

In this chapter, the results were presented visually and through several bivariate techniques to investigate the relationship between education level and causes of death. It was established that there is a statistically significant relationship between education levels and causes of death; the result will be discussed thoroughly in the next chapter.

CHAPTER 5

DISCUSSIONS OF THE FINDINGS, CONCLUSIONS AND RECOMMENDATION, AND AREA OF FURTHER STUDY

5.1 Introduction

This chapter discusses the findings of the study based on the results from the previous chapter. The discussion focuses on determining whether there is any link or not between education level and causes of death of the youth based on different bivariate techniques. The study will further present the conclusions and recommendations of areas that may need further research.

5.2 Discussion

The study utilised different bivariate techniques to investigate the relationship between education level and causes of death. The study also investigated the relationship between causes of death and socio-demographic variable such as age, gender and province of death. The results entail graphical presentation of the data and tests such as Chi-Square, Wilcoxon-Mann-Whitney, Wilcoxon signed-rank, ANOVA, Kruskal-Wallis and Friedman, which are used to determine whether or not there is a relationship between causes of death and education, and other socio-demographic variables. Also included in the analysis of the study are the logistic regression and Spearman's rank correlation coefficient which are used to describe and determine the degree of the relationship between the dependent variable and independent variables. The data is based on the deaths that were registered at the DHA in 2014.

The graphical presentation exposed that the deaths for the youth were high at 38.99% for the age category 30-34 compared with other age categories. The education level of the youth that died in 2014 indicated a high percentage of 71.35 for the youth that completed secondary education which was higher than other education levels. More youth males (56.72%) died in 2014 compared to youth females (43.28%). Stats SA (2015) reported that over 80% of death in South Africa were due to natural causes for the past 19 year period,

but in 2010 there was a decline in the percentages of. Conversely, the percentages of deaths due to non-natural causes have increase from 8.7% in 2009 to 11.1% in 2015. Similarly, the results of this study have shown that the natural causes for the youth were higher than non-natural causes in 2014. The provinces that registered most deaths were KwaZulu-Natal, Eastern Cape and Gauteng.

The Chi-Square test is used to investigate whether or not a relationship exists between dependent and independent variables. The results display that there is a statistically significant relationship between causes of death and education, also the results indicates that a statistically significant relationship between causes of death and other socio-demographic variables (age, gender and province of death) exists. Albano et al. (2007) also emphasises that educational attainment is strongly and inversely related to death from all cancers combined, for all black and white men, as well as white women. Literature have shown that educated people have lower death rates from the most common acute and chronic diseases because they are able to afford better health care services (Cutler, 2007). Since there is evidence that there is a relationship between the variables, a test had to be conducted to examine the strength of that relationship and the Cramer's V test was used. The results of Cramer's V for education and causes of death were 0.089 which showed a very weak positive relationship, gender showed a moderate negative relationship (-0.382), while age showed a weak positive relationship (0.212) and lastly, provinces of death were 0.125 which indicated a very weak positive relationship. In order to test the strength of the relationship between variables, further test (Spearman's rank correlation coefficient) was done, this test will be explained further later on in this section. Overall findings of the Chi-Square test confirm that there is a statistically significant relationship between education and causes of death of the youth. The results indicates that education does have an impact/effect on the causes of death, as well as the age, gender and province of death of the youth.

Further tests had to be done to back the Chi-Square test result, and the following; Wilcoxon-Mann-Whitney test and Wilcoxon signed-rank test were performed. The results of both the test indicated that indeed the Chi-Square test results were true because the test showed that there is a significant relationship between education and causes of death of the youth. The result of age, gender and province of death of the youth also indicated

that there is a statistical significant relationship between the socio-demographic variables and causes of death in 2014. The results indicates that education, age, gender and province of death plays a significant role in the causes of death of the youth.

ANOVA and Kruskal-Wallis test were performed to further test the data and determine whether or not a relationship exist between education and causes of death. As expected, the results showed that there is a statistically significant relationship between education and causes of death. The same result were obtained for age, gender and province of death were the variables had a link to the causes of death. For ANOVA, the R-squared was used to indicate how close the data is to the fitted regression line. The results of the R-squared were very low for all variables indicating that variables did not really have much influence on the causes of the death of the youth in 2014. To further investigate the relationship between education and causes of death, and confirm the results of the other tests, the Friedman test was performed. The results of the test indicated that there is a statistically significant difference in the causes of death depending on the education, age, gender and province of death of the youth.

Logistics regression was adopted to describe the relationship between the socio-demographic variables and causes of death. The results indicated a significant relationship between education causes of death and same applies for other variables. For education, the odds of dying from the causes of death were higher for the youth with no education, primary and secondary education. The results for gender indicated that young males had higher odds of dying compared to young females. The results also showed that age of the youth mattered when coming to the causes of death. Western Cape and Gauteng were the only two provinces having higher odds of the youth dying in those provinces compared to other provinces. The strength of these relationship had to be determined, and the Spearman's rank correlation coefficient was employed. The results showed a very weak relationship between education and causes of death, as was shown by the Cramer's V test earlier. Even for the other socio-demographic variables the relationship was very weak, except for the results of gender that showed a moderate relationship.

In conclusion, it is evident that there is a significant relationship between education and causes of death, however the strength of that relationship is considered to be very weak.

5.3 Conclusions and recommendations

The main objective of this study is to determine whether there is a relationship between education and causes of death of South African youth in 2014. The study achieved this objective, it was established that there is a statistically significant relationship between education and causes of death of the youth through application of different bivariate techniques.

Various studies in chapter 2 concurred that there is a relationship between education and causes of death. The study revealed that education is a strong factor of causes of death. Hardarson (2001) revealed that all-cause mortality and coronary artery disease mortality are significantly related to education. This present study has as well, through many tests, corroborated the fact that there is a significant relationship between education and causes of death. The only concern may the strength of the relationship between the two variables, which will necessitate further investigated together with the other suggested focal points in the next section.

5.4 Area of further study

The study recommends that there are other areas that researchers can explore in relation to this study. The following are areas that can be researched:

- The study should be continuous, based on how rapid the economy is changing which increases mortality of the population.
- The study should be carried out including other variables such as race, employment status, and other pertinent ones.
- The study should be expanded and include adults as well; and
- The study should be conducted at provincial level

REFERENCES

African Union. (2006). African Youth Charter. http://www.refworld.org/docid/493fe0b72

Agresti, A. (2007). An introduction to categorical data analysis. Department of statistics, University of Florida. 2nd edition.

Albano, J. D., Ward, E., Jemal, A., Anderson, R., Cokkinides, V. E., Murray, T., Henley, J., Liff, J. and Thun, M. J. (2007). Cancer mortality in the United States by education level and race.

Andale. (2015). Bivariate analysis definitions and examples - Statistics how to. http://www.statisticshowto.com/bivariate-analysis/

Australian Government. (2009). State of Australia's young people. Office of youth.

Blalock, H. M. Jr. (1979). Social statistics. Revised 2nd edition.

Blum, R. W. (2011). Morbidity and mortality among adolescents and young adults in the United States. AstraZeneca Fact sheet.

Bluman, A. G. (2009). Elementary statistics. A step by step approach. 7th edition.

Borode, M. (2011). Higher education and poverty reduction among youth in the Sub-Saharan Africa. European Journal of educational Studies 3 (1).

Bradshaw, D. and Steyn, K. (2001). Poverty and chronic diseases in South Africa. Technical Report. South African Medical Research Council (SAMRC).

Bradshaw, D., Pillay – Van Wyk, V., Laubscher, R., Nojilana, B., Groenewald, P., Runyon, P. R., Coleman, K. A., and Pittenger, D. J. (2000). Fundamentals of behavioural statistics. 9th edition.

Bradshaw, D., Pillay – Van Wyk, V., Laubscher, R., Nojilana, B., Groenewald, P., Nannan, N. and Metcalf, C. (2010). Causes of death statistics for South Africa: challenges and possibilities for improvements. Medical Research Council of South Africa. Burden of disease Research Unit.

Braveman, P. and Gottlieb, L. (2014). The social determinants of health: It's time to consider the causes of the causes. Public Health Report. 2014; 129 (Suppl 2): 19-31

Center for suicide prevention. (2011). Trends in youth suicide.

Child Trends Data Bank. (2016). Educational attainment. Indicators of child and youth well-being.

Cutler, D. and Lleras–Muney, A. (2007). Education and health. National Poverty Center.

De Wet, N and Odimegwu, C. (2012). Levels and causes of adolescent mortality in South Africa. University of the Witwatersrand, Department of Demography and Population studies.

De Wet, N. (2013). Levels, causes and determinants of adolescent mortality in South Africa, 2001 to 2007.

De Wet, N. (2016). Gender differences in AIDS and AIDS-related causes of death among youth with secondary education in South Africa 2009-2011.

Dowdy, S., Wearden, S. and Chilko, D. (2004). Statistics for research. 3rd edition.

Education Policy and Data Center. (2014). South Africa, Sun-Saharan Africa. https://www.epdc.org/sites/default/files/documents/EPDC%20NEP_South%20Africa.pdf

Everett, B. G., Rehkopf, D. H. and Rogers, R. G. (2014). The nonlinear relationship between education and mortality: An examination of Cohort, Race/Ethnic, and Gender Differences.

Everitt, B. S. (1977). The analysis of contingency table.

Feinstein, L., Sabates, R., Anderson, T. M., Sorhaindo, A. and Hammond, C. (2006). Measuring the effects of education on health and civil engagement: proceeding of the Copenhagen Symposium. Organisation for Economic Co-operation and Development (OECD).

Feinstein, L., Sabates, R., Anderson, T. M., Sorhaindo, A. and Hammond, C. (2006). OECD. What are effects of education on health? Organisation for Economic Co-operation and Development (OECD).

Freud, J. E. (2004). Mathematical statistics with applications. 7th edition.

Frost, J. (2013). Regression Analysis: How do I interpret R-squared and assess the goodness-of-fit.

Frost, J. (2017). Chi-square test of independence and an example. http://statisticsbyjim.com/hypothesis-testing/chi-square-test-independence-example/

Garriga, C., Garcia de Olalla, P., Miró J. M., Ocaña, I., Knobel, H., and Baberá, M. J. (2015). Mortality, causes of death and associated factors related to a large HIV population-based cohort.

Gibbons, J. D. and Chakraborti, S. (2011). Nonparametric statistical inference. 5th edition.

Gravetter, F. J. and Wallnau, L. B. (2004). Statistics for the behavioural science. 6th edition.

Griffiths, D., Stirling, W. D. and Weldon, K L. (1998). Understanding data. Principles and practices of Statistics.

Ha, R. R. and Ha, J. C. (2012). Integrative statistics for social and behavioural science. University of Washington.

Hanushek, E. A. and Wößmann, L. (2007). Education quality and economic growth.

Hardarson, T., Gardarsdóttir, M., Gudmundsson, K. T., Thorgeirsson, G., Sigvaldason, H., and Sigfússon, N. (2001). The relationship between education level and mortality. The Reykjavík study.

Health 24. (2016). South African youth Aids Mortality rate has doubled.

Hosmer, D. W. and Lemeshow, S. (2000). Applied logistic regression. 2nd edition.

http://southafrica.unfpa.org/en/topics/young-people-1#sthash.WhKa0CDv.dpuf

http://www.un.org/esa/socdev/documents/youth/fact-sheets/youth-education.pdf

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2938757/

Hummer, R. A. and Hernandez, E. M. (2013). The effect of educational attainment on adult mortality in the United States

Huysamen, G. K. (1976). Descriptive statistics for the social and behavioural sciences.

Ivančič, A. Dr., Mirčeva, J., Vrečer, N. Dr. (2008). Literature Review Report: Impact of education on health (youth, women, people with disability). Slovenia Iinstitute for Aadult Eeducation.

Jackson, S. L. (2005). Statistics plain and simple.

Krueger, P. M., Tran, M. K., Hummer, R. A. and Chang, V. W. (2015). Mortality attributable to low levels of education in the United States. PLoS ONE 10(7) e0131809. http://doi.org/10.1371/journal.pone.0131809

LaVeist, T. A. (2005). Minority populations and health: An introduction to health disparities in the United States.

Lee, N. and Peters, M. (2016). Business statistics using Excel and SPSS.

Levine, D. M., Krehbiel, T. C. and Berenson, M. L. (2013). Business Statistics. A first course. 6th edition

Liu, P., Li, C., Wang, Y., Zeng, W., Wang, H., Wu, H., Lu, J., Sun, M., Li, X., Chang, F. and Hao, M. (2014). The impact of the major causes of death on life expectancy in China: a 60-year longitudinal study.

Mann, P. S. (2001). Introductory statistics. 4th edition.

McDonald, J. H. (2015). Handbook of biological statistics.

Miniño, A. (2010). Mortality among teenagers aged 12-19 years: United States, 1999-2006.

Mokomane, Z., Mokhele, T., Mathews, C., and Makoae, M. (2017). Availability and accessibility of public health services for adolescent and young people in South Africa. Human Sciences Research Council (HSRC).

Mwamwenda, T. S. (2014). Education level and HIV/AIDS knowledge in Kenya.

Mwaniki, A. (2017). Leading causes of death in Africa. http://www.worldatlas.com/articles/the-leading-causes-of-death-in-the-african-continent.html

National Center for Education Statistics. (2011). America's youth transitions to adulthood.

National Institutes of Health. (2015). Conceptualizing the link between education and health. Office of Behavioral and Social Sciences Research. http://obssr.od.nih.gov/conceptualizing-link-education-health/

News from the Steinhardt School of culture, education and human development. (2015). Education and mortality study finds association between high school and college education and life expectancy.

Organisation for Economic Co-operation and Development (OECD). (2017). Education at a glance 2017: OECD indicators, OECD publishing, Paris. http://dx.doi.org/10.1787/eag-2017-en

Ott, R. L. and Longnecker, M. (2010). An introduction to statistical methods and data analysis. 6th edition.

Privitera, G. J. (2012). Statistics for the behavioural sciences.

Republic of Estonia. Ministry of education and research. (2016). Annual analysis: number of youth with low level of education should be reduced.

Rice, J. A. (2007). Mathematical statistics and Data analysis. 3rd edition.

Ridsdale, B. and Gallop, A. (2010). Mortality by cause of death and by socio-economic and demographic stratification. Paper for ICA 2010.

Sawe, B. E. (2016). Leading causes of death among youth in India. http://www.worldatlas.com/articles/leading-causes-of-death-among-the-youth-of-india.html

Statistics South Africa, Report No. 03 – 09 – 12. (2015). Morbidity and mortality patterns among the youth of South Africa, 2013.

Statistics South Africa, Statistical release P0211.4.2. (2015). National and provincial labour market: youth Q1: 2008 – Q1: 2015

Statistics South Africa. (2017). South Africa Demographic and Health Survey (SADHS) – 2016.

Statistics South Africa. (2012). Cause of Death Certification: A guide for completing the Notice of Death/Stillbirth (DHA-1663).

Statistics South Africa. Statistical release P0309.3. (2015) Mortality and causes of death in South Africa, 2014: Findings from death notification.

Statistics South Africa. Statistical release P0309.3. (2016) Mortality and causes of death in South Africa, 2015: Findings from death notification.

Statistics South Africa. Statistical release. P0302. (2016). Mid-year population estimates.

Szumilas, M. (2010). Explaining the odds ratios. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2938757/

Tabachnick, B. G. and Fidell, L. S. (2007). Using multivariate statistics. 5th edition.

The Center on Society and Health. (2015). Why education matters to health: exploring the causes

United Nation Economic Commission for Africa. (2015). Sustainable development goals for the Southern Africa subregion. Summary report.

United Nations Development Programme. (2016). Sustainable Development Goals

United Nations Youth. Youth education. Fact Sheet. United Nations Education, Scientific and Cultural Organisation (UNESCO).

Van den Berg, R. G. (2016). SPSS Correlation tutorials: Cramer' V – what and why?

World Health Organisation (WHO). (2010). A conceptual framework for action on the Social Determinants of Health. Social Determinants of health Discussion Paper 2. Debates, Policy and Practice, Case Studies.

World Health Statistics. (2016). Monitoring health for the Sustainable Development Goals (SDG).

www.statssa.gov.za

APPENDIX A: CHI SQUARE TEST

Table of education by causes of death				
Level of education		Causes of death		Total
		Natural	Non-natural	
No education	Frequency	1577	411	1988
	Expected	1311.5	676.54	
	Percent	3.52	0.92	4.44
Primary education	Frequency	6256	2616	8872
	Expected	5852.7	3019.3	
	Percent	13.97	5.84	19.81
Secondary education	Frequency	20603	11358	31961
	Expected	21084	10877	
	Percent	45.99	25.36	71.35
Higher education	Frequency	1114	859	1973
	Expected	1301.6	671.44	
	Percent	2.49	1.92	4.40
Total	Frequency	29550	15244	44794
	Percent	65.97	34.03	100.00

Table of gender by causes of death				
Gender		Causes of death		Total
		Natural	Non-natural	
Male	Frequency	12742	12665	25407
	Expected	16761	8646.3	
	Percent	28.45	28.27	56.72
Female	Frequency	16808	2579	19387
	Expected	12789	6597.7	
	Percent	37.52	5.76	43.28
Total	Frequency	29550	15244	44794

	Percent	65.97	34.03	100.00

Table of age by causes of death				
Age		**Causes of death**		**Total**
		Natural	**Non-natural**	
15 – 19	**Frequency**	2364	2132	4496
	Expected	2966	1530	
	Percent	5.28	4.76	10.04
20 – 24	**Frequency**	4624	4251	8875
	Expected	5854.7	3020.3	
	Percent	10.32	9.49	19.81
25 – 29	**Frequency**	9176	4781	13957
	Expected	9207.2	4749.8	
	Percent	20.48	10.67	31.16
30 – 34	**Frequency**	13386	4080	17466
	Expected	11522	5943.9	
	Percent	29.88	9.11	38.99
Total	**Frequency**	29550	15244	44794
	Percent	65.97	34.03	100.00

Table of province of death by causes of death				
Province of death		**Causes of death**		**Total**
		Natural	**Non-natural**	
Western Cape	**Frequency**	669	555	1224
	Expected	807.46	416.54	
	Percent	1.49	1.24	2.73
Eastern Cape	**Frequency**	4880	2238	7118
	Expected	4695.6	2422.4	
	Percent	10.89	5.00	15.89

Northern Cape	**Frequency**	796	538	1334
	Expected	880.02	453.98	
	Percent	1.78	1.20	2.98
Free State	**Frequency**	2775	1309	4084
	Expected	2694.2	1389.8	
	Percent	6.20	2.92	9.12
KwaZulu-Natal	**Frequency**	7028	3494	10522
	Expected	6941.2	3580.8	
	Percent	15.69	7.80	23.49
North-West	**Frequency**	1871	759	2630
	Expected	1735	895.02	
	Percent	4.18	1.69	5.87
Gauteng	**Frequency**	4330	3498	7828
	Expected	5164	2664	
	Percent	9.67	7.81	17.48
Mpumalanga	**Frequency**	3623	1427	5050
	Expected	3331.4	1718.6	
	Percent	8.09	3.19	11.27
Limpopo	**Frequency**	3578	1426	5004
	Expected	3301.0	1702.9	
	Percent	7.99	3.18	11.17
Total	**Frequency**	29550	15244	44794
	Percent	65.97	34.03	100.00