XXX

xxx

Xiao Bai Yichun Zhang

10/04/2022

Abstract

xxx

Contents

Code and data are available at[1]

```
## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion
```

[1] https://github.com/XiaoBai-blip/STA304-Paper-3

1. Introduction

1.1 Background Information

It was in 1992 that the United Republic of Tanzania modified its Constitution to become a multiparty state, and it was in late 1995 that the country held its first multiparty general elections for president and parliament in more than three decades. CCM, the dominant party in Zimbabwe, maintained its grip on power by capturing 186 of the 232 seats up for election in the country's legislature. The minimum age for admission into contractual labour in vocations that have been authorised is set at 15 years. The legislation prevents a young person from working in any activity that is harmful to their health, is risky, or is otherwise improper for their age or experience. Industrial labour is permitted for young people between the ages of 12 and 15, but only between the hours of 6 a.m and 6 p.m., with a few exceptions, according to the law. The Ministry of Labor and Youth Development is in charge of enforcing the law, but the number of inspectors available is insufficient to keep up with the pace of change. According to reports, growing privatisation has resulted in a decrease in the efficiency of government enforcement. Plantations that grow sisal, tea, tobacco, and coffee employ around 3,000 to 5,000 youngsters for seasonal work throughout the growing season. Children working on plantations often get fewer compensation than their adult peers, despite the fact that they may be doing similar tasks to their elders. It is especially dangerous and damaging to youngsters to work on sisal plantations. A sisal plantation had a child labour force that accounted for 30% of the total labour force, with barely half of the youngsters having finished basic education. They suffered from a high incidence of skin and respiratory ailments, were not given with protective clothes, and were deprived of proper nutrition and accommodation, among other things. Additional minors working in unlicensed gemstone mines range from 1,500 to 3,000 in number. Children labour with their parents in the informal economy, which is uncontrolled piecework manufacturing.

Tanzanian Primary School, which is taught in the students' native language of Kiswahili, is meant to be free, but the prices of necessary school uniforms, school supplies, and modest school overhead are considerably above the financial resources of many of the students. Students begin in standard one when they are seven years old and begin studying English in standard three when they are nine years old. Many pupils are unable to attend primary school due to the considerable distance they must go to school (the majority are far further away than the minimum 3 to 5 kilometres), duties at home, bad health, and insufficient money, among other factors. Starting with standard 4, students must pass national tests in order to progress, and a passing score on the standard 7 exams determines where they will be put in secondary school. Kids with the highest test scores and financial aid may be admitted to boarding schools, which are often located far away from their homes, while students with lower test scores may be admitted to local day secondary schools, which are also located far away but less costly.

file:///Users/yichunzhang/Downloads/multi0page.pdf https://www.asantesanaforeducation.com/tanzania-education-system- worldbank, tanzania

1.2 Our work

We conducted a Linear Regression Model between feelings about life as a whole and age group of respondent (groups of 10), number of respondent's children in household - any age/marital status, full-time/part-time job, income of respondent - total (before tax), dwelling - owned or rented, self rated health, self rated mental health, number of weeks employed - past 12 months, average number of hours worked per week, province of residence of the respondent, marital status of the respondent, education - highest certificate, diploma or degree, and living arrangement of respondent's household (12 categories), where feelings about life as a whole is the dependent variable which we want to dig deep into and the rest are the independent variables which we supposed that they might be influential to the dependent variable feelings about life as a whole. Then we conducted a Hypothesis Test on if the self-rated mental health is related to whether dwelling of the respondents is owned or rented, and conducted a Confidence Interval Analysis using the bootstrap method with a confidence interval of 90% of these two groups. After the first Linear Regression Model, we found that age group of respondent (groups of 10), dwelling - owned or rented, self rated health, self rated mental health, and education - highest certificate, diploma or degree are the influential factors to feelings about life as a whole. Using these variables, we conduct a new Linear Regression Model, especially focused on the variables of age groups of 15 to 24 years and 75 years and older, owned and rented a household, self rated health and self rated mental health, and education level less than high school diploma or its equivalent. The result of the Hypothesis Test tells that the p-value is less than 0.05, so we reject the null hypothesis, and believe that there is di

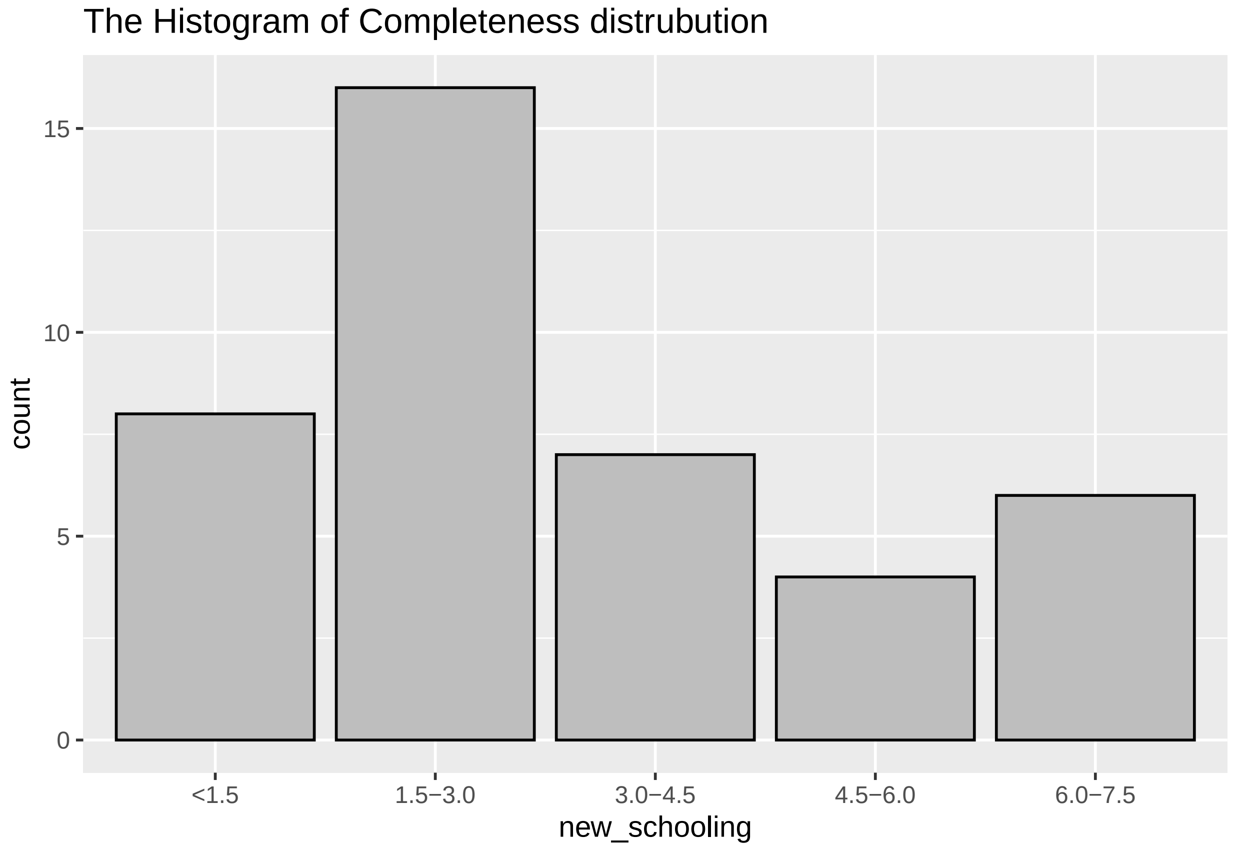
2. Data

2.1 Data Sources

2.3 Data Summary

To clean the data, first we selected relevant variables. It makes sense that age, number of children, job type, income, dwelling, health, mental health, employment situation and marital status of the respondent a

```
## Warning: Ignoring unknown parameters: bins
```



```
##
##  Pearson’s product-moment correlation
##
## data:  data_age$no_education and data_age$median_year_of_schooling
## t = -11.616, df = 7, p-value = 7.905e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9949094 -0.8821046
## sample estimates:
##        cor
## -0.9750297
```

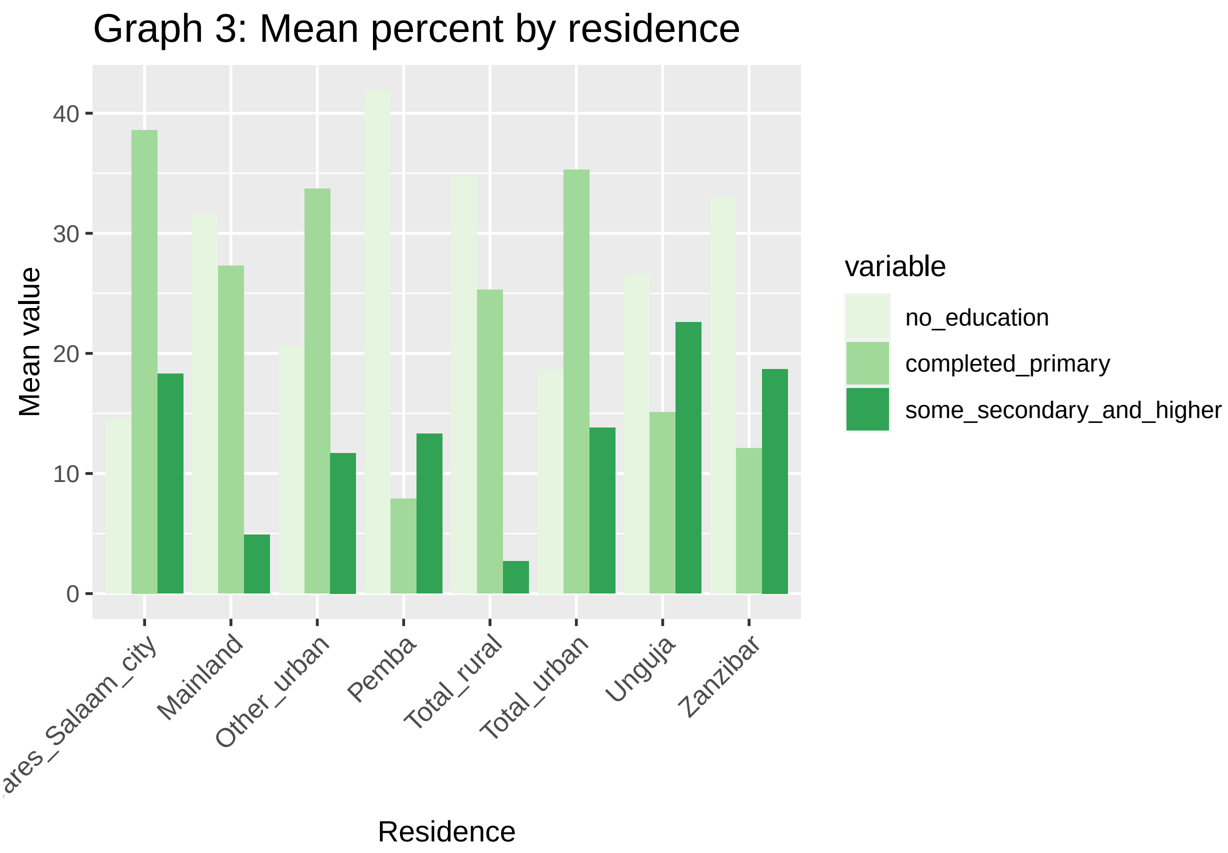
Describe table:

background_characteristic	no_education	primary_incomplete
20-24	9.3	15.4
25-29	10.8	11.2
30-34	11.2	12.0
35-39	16.6	17.3
40-44	21.7	31.4
45-49	26.9	34.2
50-54	29.8	45.8
55-59	39.8	41.9
65+	64.7	27.5

Describe table:

background_characteristic	no_education	primary_incomplete
Mainland	31.6	35.2
Total_urban	18.7	31.2
Dares_Salaam_city	14.6	26.3
Other_urban	20.6	33.5
Total_rural	34.8	36.2

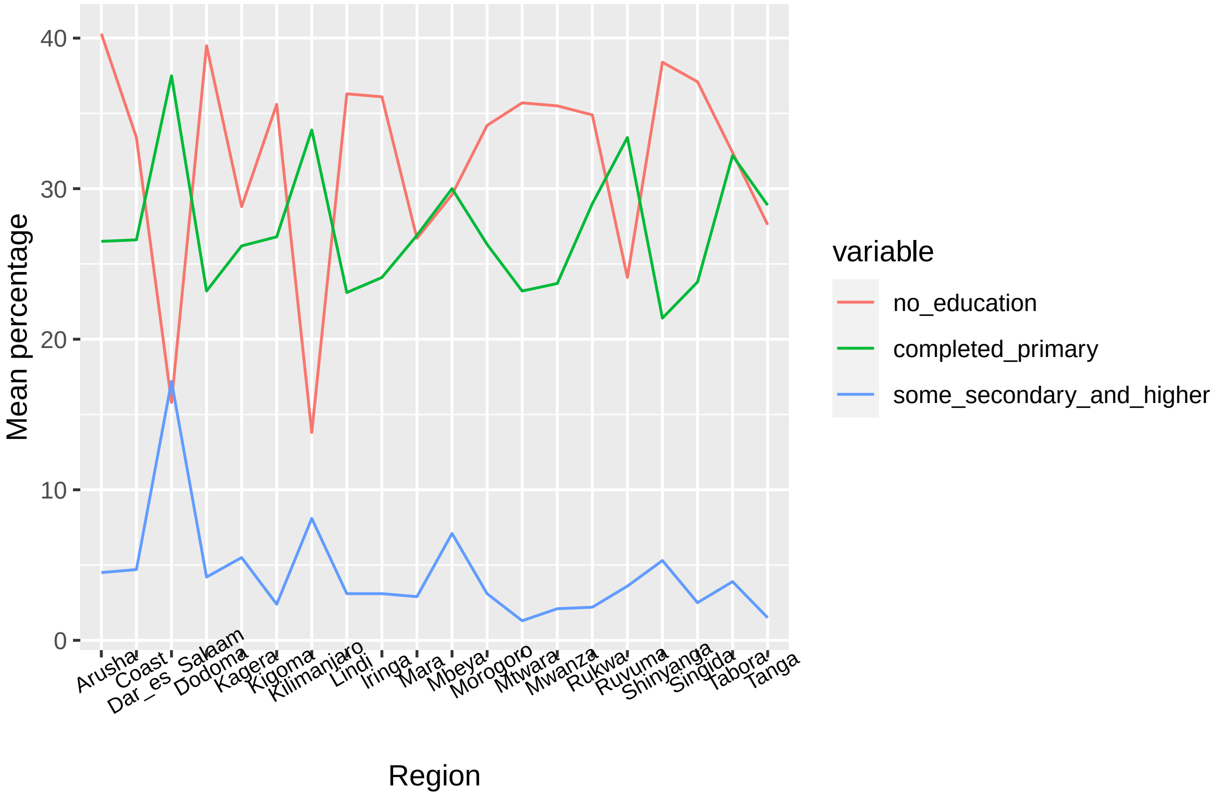
```
## Warning in pal_name(palette, type): Unknown palette green
```



The bar plot shows the percent distribution of the male household population by three different levels of education attended according to different residence types. The x-axis shows eight main residences in Tanzania and the total male population is classified based on their highest education level. In other words, we group male residence according to: people who had no education experience, have completed primary school, and had attended secondary or higher education schools.

```
## Warning in pal_name(palette, type): Unknown palette green
```

Graph 3: Mean percent by region



3. Method and Model

Statistical methods are applied in this paper to help observing and interpreting the data in an alternative way. In this paper, we will use _____ statistical methods to explore deeply about our data. These methods are simple linear regression model, confidence interval and hypothesis test.

When conducting an estimate in statistics, there is always uncertainty around the estimate because the number is based on a sample of the population. Therefore, using the confidence interval method is also essential to a statistical research because it measures the degree of uncertainty or certainty in a sampling method. More specifically, it provides an approximate set of values that is likely to contain a true parameter that is uncertain. A true parameter can be a true mean, true proportion or standard deviation. Besides, each confidence interval has a percentage associated with it, called a confidence level. This percentage represents how confident we are that the results will capture the true population parameter, relying on the bond's luck together with your random sample. In surveys, confidence levels of 90%/95%/99% are frequently used.

Hypothesis testing refers to the procedures to accept or reject statistical hypotheses. We apply this method as we expected do a strict comparison with a pre-specified hypothesis and significance level. In common word, we know that the best way to determine whether a statistical hypothesis is true would be to examine the entire population. However, this is almost impossible. Therefore, we only examine a random sample from the population to see if the sample data is consistent with the statistical hypothesis. There are two types of hypothesis, null hypothesis and alternative hypothesis. Null hypothesis (H0) assumes that the difference between the chosen characteristics in a set of data is due to chance, and alternative hypothesis (Ha) is the opposite of null hypothesis.

In the following sections, we will explain more about how we conduct these methods to provide a better understanding of our dataset.

Variable	Sample mean	Standard deviation
No education experience	25.644	17.81
Secondary or higher education experience	6.82	5.59

3.1 Confidence interval:

To analyze our data more deeply, we narrow down our topic to be more focusing on the average percentage of Tanzania communities within all background characteristics that had no education experience. We calculated the estimated mean percentage above using this dataset, which is 25.644. However, since the dataset is just one sample, there might be a problem about how we obtain a measure of precision and confidence about our estimate. Therefore, in order to describes the uncertainty surrounding an estimate, we will perform statistical inference and apply bootstrap method to get the confidence interval in this section.

Bootstrap is a statistical method that is used to estimate the sampling distribution about a given population. It creates multiple resamples (with replacement) from a single set of observations, and then computes the effect size of interest on each of these resamples. This bootstrap resamples of the effect size can then be used to determine the confidence interval. One type of bootstrap is empirical bootstrap, which samples from an estimator's sampling distribution without specifying the data distribution. In this paper, we will use empirical bootstrap. Besides, each confidence interval has a percentage associated with it, called a confidence level. More specificity, if we perform 95% confidence interval, 95% indicates that any such confidence interval will capture the population mean difference 95% of the time. Alternatively, it means that when repeating an experiment or survey over and over again, 95 percent of the time the results will match the results we get from a population. Moreover, with a 95 percent confidence interval, we have a 5 percent chance of being wrong. In addition, for a given dataset, increasing the confidence level of a confidence interval will only result in larger intervals (or at least not smaller). With the small sample, we expect to see that the 95% confidence interval is similar to the range of the data. But only a tiny fraction of the values in the large sample lie within the confidence interval. This is because the 95% confidence interval defines a range of values that you can be 95% certain contains the population mean. With large samples, we know that mean with much more precision than you do with a small sample, so the confidence interval is quite narrow when computed from a large sample. In our dataset, since the sample size is too small (n=43) and I think a wider confidence level might give an accurate result than a narrower one, I will use a relatively wider confidence level (95%) in this report. Before applying the bootstrap, there are some assumptions that need to be concerned. We assume that all samples are independent, and the parameter will be the true mean of percentage of people had no education experience.

3.2 Hypothesis test:

Hypothesis testing is generally used when we want to assess the plausibility of a hypothesis by using sample data. The higher education system of Tanizania is divided into non-university level and university level studies. Higher education is offered at 28 universities, 19 university colleges and various training colleges and institutes [https://www.nuffic.nl/sites/default/files/2020-08/education-system-tanzania.pdf]. In addition, according to (http://www.sussex.ac.uk/wphegt/tanzania), access to primary education has exploded in Tanzania since the year 2000, and the recent government figures indicate that net enrollment in primary education has reached 96.1 percent point. By contrast, access to secondary education is extremely limited in Tanzania. In 2006, net enrollment for secondary school reached only 13.4 percent. This statistic draws our attention as want to analyze if the true average percent of people that have attended secondary school before is actually equal to 13.4 percent or less than 13.4 percent.

In this case, the null hypothesis is that the true mean percentage will be the same as 13.4 ($H_0 : \mu = 13.4$) and the alternative hypothesis is it will be less than 13.4 ($H_a : \mu < 13.4$). The assumption will be all samples are independent, and since the sample size is small, we will assume that percentage of people have attended secondary or higher educational school follows a normal distribution. Taking our dataset as one sample, we calculated the sample mean above which is 6.82 percent.

Test statistics can be obtained with the formula:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

4 Results

In general, educational attainment is greater in urban areas than it is in rural areas, according to the data. For women, the percentage of those without a formal education in urban regions (25 percent) is lower than that in rural areas (46 percent); for males, the proportion of those without a formal education in urban areas is 19 percent, compared to 35 percent in rural areas. Women and men in urban areas are more likely than those in rural areas to have completed elementary and secondary school. Both men and girls with no education are in greater proportion in Zanzibar than on the mainland, which is a result of a combination of factors. Zanzibar, on the other hand, has the greatest percentage of the population with a secondary or higher level of education. This is owing to the fact that obligatory primary education includes three years of secondary school as part of the curriculum. The Dodoma, Arusha, Lindi, Mtwara, lringa, Singida, Kigoma, Shinyanga, and Mwanza areas have the greatest percentage of women with no education (over 40 percent) and males with no education (above 35 percent). The areas of Dar es Salaam and Kilimanjaro have the lowest shares of male and female respondents who have no formal education, respectively (below 20 and 25 percent, respectively).

4.1 Confidence interval

The graph is the result of bootstrap confidence interval for mean percentage of Tanzania that did not have any education experience. Values between the 2 red lines are in the 95% interval. We rounded our result to three significant digits (refer to the table). We are 95% confident that true mean is between 15.233 and

37.301. The confidence interval is meaningful because it is between 0 and 1, and both number (15.233 and 37.301) is close to and bounded around the sample mean we calculated above. Specifically, we are 95% sure that the true average of percentage of Tanzania communities that did not have any education experiences is between 15.233 and 37.301.

Table	2.5%	97.5%	CI
	15.233	37.301	(15.233, 37.301)

4.2 Hypothesis Test:

```
## [1] 1
```

5. Discussion

5.1 Survey Methology

5.2 Findings

5.3 Weakness, Potential and Future

6. Appendix

7. Reference