SMU Data Science Review

Volume 2 | Number 1 Article 15

2019

KadAfrica: Survey Analysis to Support Research for Smallholder Farmers

Gregory Asamoah
gasamoah@smu.edu

Robert Gill
Southern Methodist University, gillr@smu.edu

Frank Sclafani
fsclafani@smu.edu

Bivin Sadler
Southern Methodist University, bsadler@smu.edu

KadAfrica: Survey Analysis to Support Research for Smallholder Farmers

Gregory Asamoah[1], Robert Gill[2], and Frank Sclafani[3], Advisor: Bivin Sadler PhD[4]

[1] KNUST, Ghana PMB KNUST Kumasi, Ghana
gasamoah@mail.smu.edu
[2] Baylor University, 76706, USA
gillr@mail.smu.edu
[3] Daytona Beach, FL 32720, USA
fsclafani@mail.smu.edu
[4] Southern Methodist University, 75205, USA
bsadler@mail.smu.edu

Abstract. In this paper, we present an analysis of survey data with the goal of determining if the KadAfrica training program, a social organization in Uganda, has a significant effect on the lives of the girls who participate in the program. This is done through an observational study of girl's responses to several pre-program and post-program questions. These questions include topics such as the girl's access to hygiene materials and their personal views on family finances. In addition to providing an analysis of historical data, we established a data platform in which future data can be stored and analyzed in an effective manner. In this study we provided statistical analysis of several survey questions for KadAfrica. Finally, we provided KadAfrica with a RShiny dashboard application whereby they may conduct their own future analyses. Based on the results of the questions we analyzed, we conclude that KadAfrica is having a positive impact on the lives of the participants.

1 Introduction

In East Africa, there are many non-governmental organizations (NGOs) that are engaged to act for different countries. In late 2012, Eric Kaduru, a successful businessman, decided to move back to the birth place of his father in Fort Portal, Uganda where he went on to form his own NGO called KadAfrica. The founders of KadAfrica describe it as a social organization in Uganda that, "envisions a world where out-of-school girls are economic drivers of their communities" [1]. It assists the girls in starting their own cooperatives and trains them in basic business management skills and agriculture. In rural Uganda, girls face economic hardships and many young girls are unable to access education. KadAfrica participants are provided with the support of a facilitator who trains and mentors, along with the service of a devoted agriculture extension worker. By participating in the "KadAfrica Experience" program, girls receive a very broad curriculum that includes, family and financial planning, personal hygiene, hands-on

training, access to land, agriculture basics, access to seedlings, and other means of support.

This paper addresses the problem posed by KadAfrica. KadAfrica has requested that the authors answer several questions using historical data that has been collected with the purpose of gauging the effect that the program has on the livelihood of those that participated in the program. Additionally, the organization requires a platform that may be used to store and analyze future data. The authors deliver an analysis of the data that attempts to answer several questions that are of interest to the KadAfrica team. Through answering these questions, the authors gain insight into how the program affects the lives of girls in Western Uganda. The questions that are of interest to KadAfrica are asked with the intent of determining the impact of the program on the lives of the participants both financially and on their daily lives.

The surveys are given to girls both before the program and after the program as well as bi-annually after completion of the program. The questions cover a wide range of topics and have numerous types of responses. Most of the questions that we are investigating are questions that have categorical responses. These categorical responses can be further broken down into three structures. The first response is structured in binary: yes, no responses. The second structure is ordinal: never, sometimes, always. The third is a level agreement scale. This structure gives the participant the option to choose a level of agreement ranging from one to five depending on the structure of the response. For the first two structures, the Fisher's Exact Test is used. For the third structure, the Wilcoxon Ranked Sum Test is used. The objective of both tests is to determine if there is a significant difference in the responses between the pre-program survey and the post-program survey.

The cost of menstrual supplies in Uganda is high relative to other parts of the world. Often, girls are unable to afford these supplies. KadAfrica teaches the girls who participate in the program how to craft homemade supplies. By doing so, the program allows women to gain access to hygiene necessities that they might otherwise be unable to afford. As a result, this is one of the questions that is investigated in this study. Similarly, there are questions that pertain to access to clean water and other hygiene concerns.

Often in Uganda, girls are subject to pressure to adhere to a subordinate role in society. In opposition to this, KadAfrica, through its program, attempts to increase the confidence and independence of the young women and girls who participate in the program. By using several questions asked in the surveys, we are able to provide an analysis of the program's impact in this regard.

In addition to analyzing historical data provided by KadAfrica, this team has the goal of developing a reporting platform whereby KadAfrica is able to effectively and easily store and analyze data. This includes the setup of a back-end data storage structure as well as a front-end reporting platform. Given the limited financing of the company, it is advantageous to pursue little to no cost solutions. This eliminates most already developed data platforms.

For importing the data into the analysis application, we explored a few options including connection to KadAfrica's current data warehouse via a REST API [5]. This

option was appealing as it ensured a clean and updated dataset for each run of the analysis. However, after the customer expressed concerns exposing the privacy and security of the data. That concern plus the unreliability of the internet connections in the part of the world, they conduct their work, means that the requirements for the tool was adjusted to accept CSV files stored locally as the main method of data ingestion.

The front-end platform is an R Shiny application that provides KadAfrica with a dashboard that enables them to load data efficiently and build custom reports. This application runs on top of the proven R code and utilizes many libraries to allow the user to select the data that they wish to analyze [6].

As an NGO, KadAfrica relies on income from grants from organizations like the Bill & Melinda Gates Foundation. During the application process, it is common that these foundations require statistical analysis to prove that their financial contribution will have a positive impact. The analysis that we provided to KadAfrica was used in a grant application to the Bill & Melinda Gates Foundation.

According to Glassdoor Local Pay Reports, Uganda is one of the countries in Central Africa where the salaries of statisticians, mathematicians and data analysts earn more than many other professions. Top statisticians, especially those working for NGOs, can earn 5 million Ugandan Shillings (Ush) per month (current conversion, as of Feb 2018, ~ $1,360.25 USD), on the higher side [8].

A small organization like KadAfrica does not have the financial resources required to hire a professional with the ability to analyze the survey data results. Our initial analysis and the tool that we could provide a huge cost savings to our client as it allows the KadAfrica team to conduct statistically sound data analysis on a “budget” without the need to pay for an enterprise level tool or to hire a fulltime statistician. The result of most of the questions analyzed, demonstrated that KadAfrica is having a positive impact on the lives of the participants. Additionally, the results from our first round of analysis on cohorts 1-4 was used to help create a grant proposal from Bill & Melinda Gates Foundation. This analysis is currently being used in grant proposals and could potentially allow KadAfrica to secure funding with the purpose of furthering their program’s success.

The rest of this paper will discuss the statistical tests that were used to conduct this observational study. It will also explain the data preparation, initial analysis that was requested by the client, a description of the tool that was created, and a discussion of the ethical considerations for dealing with survey data.

2 Statistical Tests

2.1 Fisher’s Exact Test

Fisher’s exact test [9] is a test that can be used when the sample size is small. The test is used to determine whether there is a significant difference in the proportion of a variable. In the case of this study, it is used to determine if there is a difference between the proportions of a given response between the pre-survey responses and the post-

survey responses. This is an ideal test for the given data as is does not make any assumptions about the size of the data set and is particularly well suited to smaller datasets.

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{\binom{a+b}{b}\binom{c+d}{d}}{\binom{n}{b+d}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!\,b!\,c!\,d!} \quad (1)$$

A set of data with n rows containing two groups and two factor levels may be represented by a two by two frequency table. The cells in this table containing frequency counts may be represented by the letter designations a, b, c, and d. The above equation, called the hypergeometric distribution, is used to calculate the probability for any such arrangement of frequency counts. The p-value for the Fisher's Exact test is calculated by summing the probabilities of all arrangements having a probability equal to or smaller than the probability of the observed arrangement.

2.2 Wilcoxon Signed Rank Test

The Wilcoxon Rank Test [10] is a non-parametric test used to compare the distribution of two related samples. In the case of this study, the samples are the pre and post survey responses. The test is used to determine if the mean ranks differ between the two samples. This is done by looking at the distribution of the ordinal responses to the survey questions. This test is used to determine if the mean responses to the questions that have scale responses differ between the pre-program responses and the post-program responses. This test uses the test statistic W for which the calculation is show below.

$$W = \sum_{i=1}^{N_r}\left[sgn\left(x_{2,i} - x_{1,i}\right) \times R_i\right] \quad (2)$$

Where N is the sample size, x_1, x_2 represent the observations, and R_i represents the rank. This W statistic is then compared to a chosen critical value where the null hypothesis is rejected if the value is less than or equal to the critical value. The associated p-value may be calculated by determining the probability a rank assignment that result in a test statistic as extreme or more extreme than the arrangement observed.

3 Data Preparation

3.1 Cleaning

The current historical dataset that is available for analysis, is real-world survey data that was collected by KadAfrica pre-program and post-program. As received, the data is not complete and contains missing values. The first problem to overcome is to identify the deviancies in the data and to prepare the data for analysis. One step that we took to clean this data was that we only conducted analysis on values that contained a complete sample of data from the pre-program all the way to the post-program.

Many of the KadAfrica participants were unable to finish the program and therefore had values in the pre-survey but null values in the post-survey. In this case, we did not include those participants in the summary statistics or pre-post program comparisons. Only girls who completed both the pre and post-survey were included in our analysis. According to the client there are several reasons why the girls were unable to complete the program. Most were due to various health or family concerns. We did not conduct analysis on the number of girls who did not complete the program and explore why they did not finish. However, this would be an excellent topic to explore for any future work.

Nevertheless, on a question-by-question basis, we took steps to verify that the individual provided a valid response to the same question in the pre and post surveys. We had to take this into consideration when we noticed that some participants either did not complete the corresponding question in the post-survey or their response to that question was not in a valid format [11].

3.2 Storage

The KadAfrica team currently collects data from the program participants and stores them in flat files such as excel sheets and CSVs. The data stored in the excel sheets are not consistent between pre and post cohort's programs. Columns names are not consistent between pre and post cohort's programs, making extraction of the data for analysis difficult. Our initial analysis extracted the historical data stored in the excel sheets into R for exploratory data analysis. Additionally, the KadAfrica team stores their survey results in a survey collection application called KoBoToolbox.

4 Initial Analysis

In the following sections, we provide an analysis of historical data provided by KadAfrica. The topics analyzed were chosen at the request of KadAfrica. These topics were chosen with the intent to determine the impact of the program on the lives of the women and girls who have participated in the program. The data provided spans five cohorts of participants. Due to data consistency and volume issues, the primary focus was placed on analyzing data from cohorts four and five. At the request of KadAfrica, the data was analyzed as an aggregate across the cohorts. The data was analyzed as a response to the eight questions asked by KadAfrica. These questions are listed below:

1. Is there any significant change in the access to menstrual hygiene products from pre to post?
2. Is there any significant change in the use of mosquito net from pre to post?
3. Is there any significant change in the access to clean drinking water?
4. Is there any significant change in the access to toilets?
5. Is there any significant change in illnesses within the last 6 months?
6. Changed perspectives on gender equity / gender?
7. Relevant findings in savings behavior, business creation, personal investment.
8. Description of questions asked solely in the post survey.

4.1 Access to Menstrual Hygiene Products

According to a 2010 United Nations study in various African countries, half of the girl pupils in the study reported missing 1-3 days of primary school per month due to a lack of access to proper menstrual hygiene products. Additionally, over 50% of the senior women teachers confirmed that there is no provision for menstrual pads to school girls [12]. This lack of access not only causes missed learning due to menstrual period; it can be potentially detrimental to their reproductive health.

In analyzing the girl's access to menstrual hygiene, we look at the responses to the following question asked by the KadAfrica pre-program and post-program surveys: "How often do you have access to menstrual hygiene products?" The girls submitted one of three possible responses: sometimes, always, never. We then analyzed the responses to this question using the Fisher's Exact Test for difference of proportions. The "never" and "sometimes" responses were grouped together and opposed to the always responses. The alternative hypothesis for this test is that the proportion of those that always have access to menstrual hygiene supplies after the KadAfrica program is greater than before the program.

With a p-value of 1.272 x 10-6 at an alpha level of 0.05, we can reject the null hypothesis and conclude that there is a statistically significant higher proportion of girls always having access to menstrual hygiene supplies after participating in the KadAfrica program for girls who participated in cohorts four and five of the program.

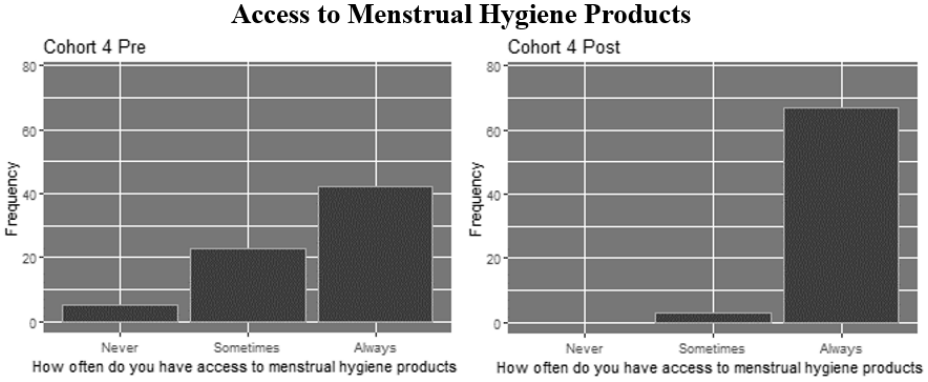

Fig. 1: An on a scale comparison from Sometimes, Always, Never for the question: … How often do you have access to menstrual hygiene products?

As seen in Figure 1, this result is in line with expectations as one of the focuses of the program is to instruct the women and girls on how craft homemade menstrual hygiene supplies. The program includes this instructional material because it recognizes the high financial costs of these supplies in their region of the world. With only two full historical surveys, cohort 3 and 4, we were unable to conduct time series analysis to determine if there is a trend. After more full surveys are collected, we would like to include a time series analysis as a suggestion for any future investigation of this dataset.

4.2 Access to Treated Mosquito Nets

Long-lasting insecticidal nets (LLINs) are core malaria prevention tool that is recommended by the World Health Organization (WHO) for use by people at risk of contracting malaria especially in malaria endemic countries, such as Uganda [6]. An indicator of economic success can be determined by the participants' ability to afford and gain access to LLINs.

In order to determine access to mosquito nets, the participants are asked the following question: "How often do you sleep under a treated mosquito net?" The sometimes and never responses are grouped into a single category. The distribution of these responses is shown below in Figure 2.

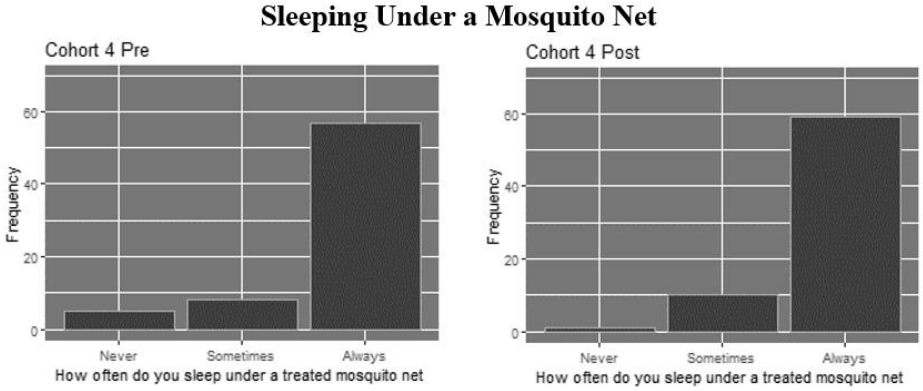

Fig. 2: An on a scale comparison from Sometimes, Always, Never for the question: **….** How often do you sleep under a treated mosquito net?

With a p-value of 0.4115 at an alpha level of 0.05, we are unable to reject the null hypothesis and conclude that there is not a statistically significant higher proportion of girls always having access to sleep under a treated mosquito net after participating in the KadAfrica program. While this test is not significant, in Figure 2 above it is observed that there is a small change from girls with Never responses to girls with Sometimes or Always responses. While this change is not statistically significant, it is still a positive improvement that will have an important impact on the girl's health and safety.

4.3 Access to Clean Drinking Water

In countries like Uganda, women and children used to spend massive amounts of time collecting water for drinking, cooking, and washing from sources that may be miles away. Returning with them full water jugs, weighing up to 40 pounds, this daily time burden can equate to losing out on educational and work opportunities. Additionally, harmful bacteria, viruses, and other pollutants are an equally heavy concern [3].

The following question was asked on the surveys in order to gauge access to clean drinking water: "How often does your family home have safe purified drinking water?" In accordance with the previous tests, the sometimes and never responses were grouped

together. The alternative hypothesis is that the proportion of those that always have access to clean drinking water after the KadAfrica program is greater than before the program.

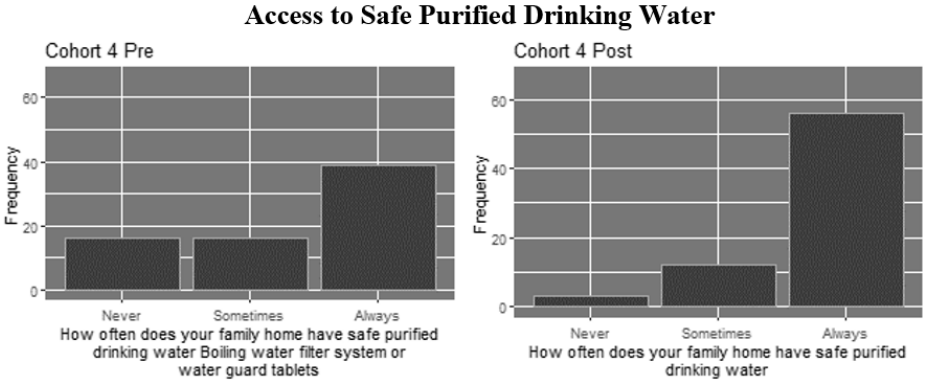

Fig. 3: An on a scale comparison from Sometimes, Always, Never for the question: … How often does your family have access to safe purified drinking water?

With a p-value of 1.27 x 10-8 at an alpha level of 0.05, we reject the null hypothesis and conclude that there is a statistically significant higher proportion of girls always having access to clean drinking water after participating in the KadAfrica program for the girls who participated in cohort 4 of the program. The high significance of this test is clearly visible in the distribution of the responses shown above in Figure 3.

Of course, the difference between the pre and post responses cannot be directly contributed the program, the lesson "safe water management/ clean water" would have an impact on the final result. This means that the girls are taught how they can handle water, which would increase their access to safe water. Additionally, future work may help to identify if a relationship exists between increased income and increased access to clean water.

4.4 Significant change in the access to toilets

In Uganda, many families do not have access to toilets within their household. While this is considered a necessity in many countries, there are places across the globe where toilets are considered a luxury and are expensive to install in households. In Africa, it is estimated that only six percent of the population has access to a flush toilet [2]. The KadAfrica program works to help provide the women and girls who participate in this program with access to this sanitary necessity.

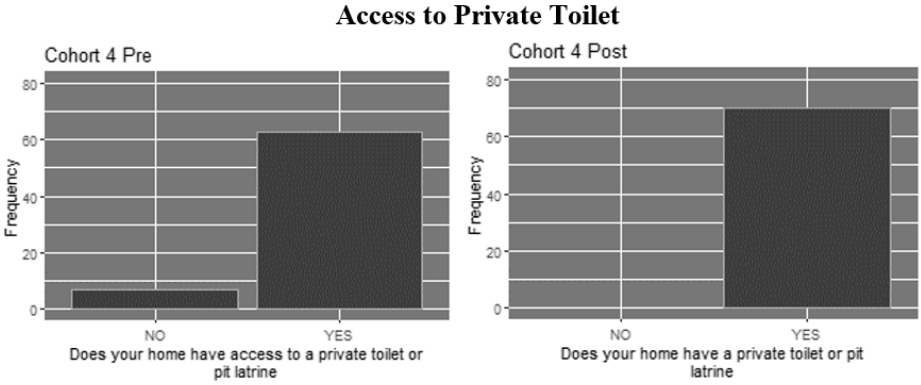

Fig. 4: An on a scale comparison for the question: … Does your home have access to a private toilet or pit latrine?

With a p-value of 0.0067 at an alpha level of 0.05, we can conclude that for cohort 4, the program was able to improve the access to a toilet for the girls who participated in the program. Additionally, Figure 4 above demonstrates that after participating in the program, all the girls in cohort 4 had access to a private toilet or pit latrine. This result may be attributed to KadAfrica's focus on personal hygiene during the program.

4.5 Illnesses within the last 6 months

This question could not be analyzed due to the nature of the wording difference between the pre-survey and post-survey. In the pre-survey, the question was asked with the wording: "Have you fallen sick in the past six months?" In the post-survey, the question was asked with the wording: "Have you fallen sick in the past three months?" The distinct difference between six- and three-month periods means that we could not compare the responses statistically. KadAfrica was advised of the importance of questions being asked in the exact same manner between the pre and post surveys.

4.6 View on Women's Opinion in the Community

The following question uses a level of agreement Likert scale for responses. The scale used allows the individual to provide an answer ranging from one to five with one being strongly disagree and five being strongly agree. The individual was then asked to what extent they agree with the following statement: "It is important for girls (and women) to express their opinions in the community." The alternative hypothesis is that the girls' level of agreement with the above statement is higher after the KadAfrica program than before the program. In order to analyze the responses to the question, the Wilcoxon Signed Rank Test is used to evaluate if there is a difference in the mean ranks between the pre-program and post-program responses.

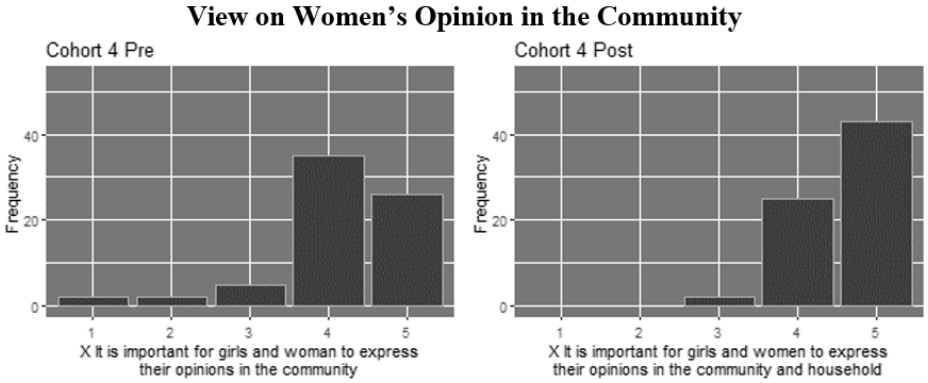

Fig. 5: An on a scale comparison from 1 to 5 for the question: … It's important for girls and women to express their opinions in the community?

With a p-value of 0.000122 at an alpha level of 0.05, we can reject the null hypothesis and conclude that there is a statistically significant higher level of agreement with the above statement. This represents a significant change in the participants' views on the value of women's opinions. As demonstrated above in Figure 5, there is a statistically significant increase in the number of participants who believe that it's important for girls and women to express their opinions in the community.

4.7 View on Opportunities for Girls

As in the previous question, the following question uses a level of agreement Likert scale for responses. The scale used allows the individual to provide an answer ranging from one to five with one being strongly disagree and five being strongly agree. The individual was then asked to what extent they agree with the following statement: "Girls have the same opportunities as boys." Again, a Wilcoxon Signed Rank Test is used to evaluate the responses.

Fig. 6: An on a scale comparison from 1 to 5 for the question: … Girls have the same opportunities as boys?

With a p-value of 0.0346 at an alpha level of 0.05, we can reject the null hypothesis and conclude that there is a statistically significant higher level of agreement with the above statement. From the distribution of responses in Figure 6, after participating in the program, the girls have changed their opinion about the statement that girls have the same opportunities as boys.

4.8 Economic Impact

The analysis of the pre, post, and the biannual historical survey data provided new insights on how the KadAfrica program is impacting the economic status of girls in the local communities in Uganda. This section presents findings from survey conducted with the pre-cohort, post-cohort, and the biannual data collected in various rural communities in Uganda by the KadAfrica team, highlighting the economic impact assessment pre and post of girls participating in the program.

The KadAfrica team created the following questions in their pre and post surveys to gauge the economic impact after participating in the program:

1. "Since harvesting your passion fruit, has your economic status in your household increased?"
2. "Have you personally ever accessed a loan from KadAfrica savings group or another institution?"
3. "I can take care of my family using money earned from Agriculture?"

At the time we conducted our analysis the most recent cohort that had post survey data was cohort 4. We therefore were able to compare that cohort's results to the bi-annual survey results, which contains data from all patriciates that response to the survey.

For the first question, "Since harvesting your passion fruit, has your economic status in your household increased?", of the 72 respondents for the post cohort 4, 58 responded No, 14 responded Yes. Of the 107 respondents for the Biannual Survey, 16 responded No, 91 responded Yes. After conducting our analysis, we were able to conclude that there is a statistically significant higher proportion of participants who believe that their economic status in their household increased since harvesting their passion fruit.

Table 1. Increase in Economic Status in the Post Cohort 4 and Biannual Survey

Question	**Post Responses**		**Biannual Survey**	
	No	**Yes**	**No**	**Yes**
Since harvesting your passion fruit, has your economic status in your household increased?	58	14	16	91

The response to this question was conducted post program using the biannual survey data. The analysis was important to the KadAfrica team as it can indicate an overall economic impact of the program on the girl's participating in the KadAfrica program.

The second question posed the question, "Have you personally ever accessed a loan from KadAfrica savings group or another institution?". Of the 70 respondents for the pre cohort 4, 53 responded No, 17 responded Yes. For the post survey, 18 responded No, 52 responded Yes. After conducting our proportion test, we were able to conclude that there is a statistically significant higher proportion of participants who have personally accessed a loan from KadAfrica savings group or an-other institution. This is indicative of better financial position as an institution was willing to extend them a line of credit. This is supported by a p-value of 2.208-8.

Table 2. Financial Institution Responses in the Post Cohort 4 and Biannual Survey

Question	Post Responses		Biannual Survey	
	No	Yes	No	Yes
Have you personally ever accessed a loan from KadAfrica savings group or another institution?	37	35	30	78

4.9 Descriptive Analysis for Post Program Questions

The survey responses to the following questions were analyzed for the purpose of calculating frequency analysis and descriptive statistics. The following questions were only asked in the post survey. However, KadAfrica considered these questions as very important, therefore we provided descriptions of the responses, in the form of numerical calculations and tables. The intent of this analysis was to enable the KadAfrica team to make inferences and predictions about the population based on the response data taken from the population in the questions below.

Table 3. Results of Important Questions asked in the Cohort 4 Post Survey

Question	Post Responses	
	No	Yes
Do you feel that you accomplished your reason in joining KadAfrica?	7	65
Do you feel more confident in voicing your opinions?	1	71
Have there ever been times when there was no food in the house?	41	31
Can you now plant passion fruit without anyone's help?	7	65
Would you take up farming passion fruit as a business?	0	72

For the question, "Do you feel that you accomplished your reason in joining KadAfrica?", of the 72 respondents, 7 responded No, and 65 responded Yes. The response to this question that was conducted post program is important to the KadAfrica team as

it can indicate an overall feeling of success of the program from the view point of the participants.

Likewise, for the question, "Do you feel more confident in voicing your opinions?", of the 72 respondents, 71 responded Yes, and only 1 responded No. When asked, "if no, why?", the only "No" responder explained that it was because their financial situation had not improved. This response is indicated that all but one participant observed an increase in their confidence voicing their opinion after participating in this program.

For the question, "Have there ever been times when there was no food in the house?", of the 72 respondents, 31 responded Yes, and 41 responded No. Of the 12 post program questions explored in this analysis, this question had the smallest deviation (>10) between the binary responses. This may indicate to the KadAfrica team that more work may needed to address access to food issues.

For the question, "Can you now plant passion fruit without anyone's help?", of the 72 respondents, 65 responded Yes, and 7 responded No. This response was one of the most important to KadAfrica as the main goal of the program was to teach the girls how to farm on their own.

Finally, for the question, "Would you take up farming passion fruit as a business?", of the 72 respondents, all 72 responded Yes. This is important as only a handful of questions ever received a unanimous response.

Self-Health Scaled Question

The following question was prefaced with the instructions below:

On a scale of 1 to 5 please rate these statements, where 1 means that you strongly disagree and 5 means that you strongly agree. Select the number that best describes how you feel. 1. Strongly disagree 2. Disagree 3. Somehow agree 4. Agree 5. Strongly agree. Since joining KadAfrica, I am more often asked for my opinion on the decisions of the household in relation to...

Fig. 7: On a scale from 1 to 5 for the question: …my own health (for example going to the doctor, buying menstrual products, deciding to pay for prescriptions), the mean response was equivalent to 3.542.

The mean response for this question was the lowest of all the post survey questions analyzed. One possible reason that may have contributed to the responses in the 1-2 ratings, shown in Figure 7, are that this response is subject to the individual in terms of how willing their household on openly discussing medical issues.

5 Analysis Tool

5.1 Data Platform Architecture

The data platform tool will include multiple technologies including R Studio, R Shiny, and a locally hosted webserver. All the statistical modeling and analysis will be running in R Studio and results would then be placed into readable format for R Shiny dashboard to present via a public website. Using existing survey data Comma Separated Value format, our system will scrap for relevant data, parse it, and extract the data for exploratory and statistical analysis.

5.2 R and RStudio

R is an open source programing language used for data manipulation, exploratory analysis, statistical calculations and visualization of data. RStudio is a free open source integrated development environment for R programming language.

Setting up the application development environment involves downloading the latest version of R and RStudio. We installed R and RStudio successfully and import various R packages such as shiny, shinydashboard, ggplot2, plotly, gdata and many more into R.

5.3 R Shiny

Shiny is an R package used to build interactive applications that can be deployed over the web with R. The program consists of three main components:

1. User Interface (UI) - nested R functions that assemble an HTML user interface for the app
2. Server - a function with instructions on how to build and rebuild the R objects displayed in the UI
3. ShinyApp - combines UI and server into a functioning app

In the R Shiny application, we organize the UI layout with a drop-down menu for pre and post survey data. We created the UI file input box which enables the user to select a specific cohort dataset from a drop-down menu. The user then selects a pre and post period for the cohort selected for the analysis from a period drop-down menu. Once the

user selects the cohort dataset, reactive functions run the code load the dataset and use the data path associated with the cohort files. The shiny dashboard has another drop–down menu to enable the user to select questions from the pre and post cohort dataset.

After the cohort dataset is loaded, it is passed to the server to populate the selected inputs with a list of variables from the cohort dataset. The variable selections from the drop-down menu are then passed to the server where the summary, the statistical test and various plots are created and passed back to the main section of the UI which displays the output.

5.4 User Interface Design and Application Design

The final deliverable to KadAfrica is in the form of a user dashboard built in RShiny. This dashboard will allow the user to easily and effectively view an analysis of the ten questions that they asked at the beginning of this project. These questions are arranged as tabs on the dashboard. Additionally, if the user would like to look at survey questions that were not covered in the 10 questions, there is a tab for ad-hoc question analysis.

The application works by pulling in survey data from csv files are placed in a designated directory. So long as the files are formatted as csv files with the headers being the individual survey questions. Each file in the directory is read and the data is place into a custom R object called “survey”. Each survey object is then appended to a separate custom R object called “survey collection”.

The survey object contains five fields and three methods. The five fields are as follows: file path, cohort, period, data, questions. The file path field identifies the file path and file name of the file the data was pulled from. The cohort and period fields identify the cohort and period the survey was given. The data field is a data frame object containing the survey data pulled from the data file. The questions field is a list containing questions present on the survey.

The methods for the survey are initialize, search question, and search id. The initialize method is used to parse, clean, and restructure the data file for the given survey. The search question method is used to search for a given question on the survey using key words and the search id method is used to search for a given individuals user id [4].

The survey collection object contains one field, a list of surveys, and eleven methods. The eleven methods consist of an initialize method, an import and append method, as well as the primary methods for combining survey data and running analysis.

Once the data is imported from the designated directory, the application cleans and structures the data to prepare it for analysis. This is done when the user opens the dashboard. After opening the dashboard, the user can view eleven tabs. Ten tabs corresponding the questions of interest as well as an ad-hoc tab for exploratory analysis. A screenshot of the interface is shown below.

Fig. 8: A screenshot of the survey analysis tool that we provided with an example comparison of a question from the cohort 4 dataset.

Shown above is an example of one of the tabs in the dashboard. This tab is for the question regarding participants' access to menstrual hygiene supplies. The first option given to the user is the ability to select a given cohort. The image above shows that cohort one is selected. Next, the user has the ability to select a question that contains the term "menstrual hygiene". The user must select a question from both the pre survey and the post survey. Once this is done, the user is shown several displays.

The first is the question similarity. This is calculated using the Levenshtein Similarity score. This score ranges from zero to one and is calculated by determining the number of edits required to transform the second string into the first string [9]. This is then divided by the length of the first string and subtracted from one. This statistic is shown to inform the user if the wording of the question changed between the pre-survey and post-survey.

Second, are the plots of the data in both the pre and post surveys. For categorical responses, this is done to give the user a visual representation of the data used in the analysis.

In Figure 8 above, below the plots, are two boxed sections. The first section shows the changes in responses from pre to post for the given question. The second section shows the statistical analysis of the changes. Based on the type of categorical responses present in the data, the application chooses the appropriate statistical test to perform and then shows the results of this test to the user. Above, we see that the application identified Fisher's Exact Test as the appropriate analysis and shows the user that this test was highly significant with a p-value well below 0.05.

6 Ethics

There are many ethical codes to consider when performing work for any client, including NGOs. In addition to the technical details of the analysis, we had to be attentive of how the data is presented to the client. We discovered through this process of working with a client how important it was to explain the results of our statistical tests. Due to the limitations of how the data was collected, this analysis was limited to the participants of this program. Therefore, inferences can only be made to the sample set that was used. The ethical issue that needs to be avoided in this scenario is that the client could make incorrect inferences from the analysis. To ensure this was not an issue, our team worked to make sure that the KadAfrica team understood the statistical significance of our tests and their limitations.

By providing the KadAfrica team with a tool that conducts statistical analysis and generates graphs our team had the responsibility to provide training to the user. It would have been unethical to not adequately instruct the KadAfrica team on how to use the tool and what the results mean. The graphs that are generated vary depending on the type of analysis performed. Understanding these differences is key to ensuring that the correct inferences are made, and that the data is represented as accurately as possible.

User privacy is a top concern for the KadAfrica team. They take measures to ensure that their participants data is properly protected. Although the results of these surveys are voluntarily given, medical, financial, and personal responses are recorded need to be protected. During this analysis the names and other identifying information such as birthday and home address were only utilized to ensure the individuality of each record and was then removed as data that was not relevant. All survey analysis was then conducted using the participants ID number which helps to ensure their anonymity.

7 Results

While working with a customer posed its unique challenges, it was a beneficial experience. We are grateful to the KadAfrica team for trusting us with their data and to provide them accurate analysis. The client relationship that we shared with them would not have been possible without their honesty and candid feedback and their desire for detailed statistical analysis. According to one of the members of the KadAfrica team, our

analysis has greatly impacted the application created for the Grand Challenges Canada nonprofit organization. That grant is in progress at the moment and KadAfrica believes that if it goes through, it will be in a large part due to some of the data analysis we provided.

The result is that our team was able to provide an NGO that operates thousands of miles away with sound analysis of survey results. Not every survey question analyzed demonstrated a positive impact. In fact, some showed no significant difference from pre to post. Ultimately, the results of most of the questions we analyzed, demonstrated that that KadAfrica is having a positive impact on the lives of the participants.

Additionally, we were able to provide our client with a tool that replicated the same analysis that we preformed. After providing the KadAfrica team with the tool and the required operating instructions, a non-statistician can utilize this graphical interface tool to create the same results and plots on any future data they collect. Saving the client money on past analysis and creating a tool for future analysis will enable their continued success and financial stability as they make use of our tools to help secure future grants. It is our sincerest hope that the KadAfrica team utilize the tools we provided to help maintain their mission of helping future generations of Ugandan woman.

8 Conclusion

The KadAfrica pre and post training program is empowering girls in Uganda in the fields of income, employment, personal hygiene, education and many more, based on the results from the analysis of the survey data in the paper above. For potential future work, adding additional functionality to the application would be a top priority. The ability to compare more than just two cohorts would be a highly useful feature. Additionally, more statistical tests (or a statistical test recommendation feature) would be very beneficial to a less statically inclined user. One final future recommendation is to work with the KadAfrica team to make this application available to more organizations.

The analysis conducted on the questions selected by KadAfrica in section 4 should help the client to showcase the positive impact of their pre and post program to the world and to donor agencies to help them raise funds for the training program.

From all the illustrations and statistical analysis discussed in the paper using R and shiny dashboard, the application we provided can be used to analyze future cohort survey data by the KadAfrica team. The application will help review the outputs from different statistical tests such as Fisher test, Wilcoxon rank test, and proportion test on KadAfrica survey data.

The continued use of this application will help KadAfrica standardize their survey data in a structured file formats, review and analyze critical statistical outputs, and generate plots from the survey results.

9 Acknowledgments

Thank you to all who reviewed this paper and our Professor, Dr. Bivin Sadler, who provided valuable inputs and feedback. We would also like to thank Jessica and Eric

Kaduru, Maria Kjaer and the entire KadAfrica team for partnering with us and for sharing our yearning for precise statistical analysis. We are honored to contribute, in our humble way, to a program that is drastically changing the lives of so many.

References

1. About Us. (n.d.). Retrieved from http://www.kadafrica.org/
2. Banerjee, S., Wodon, Q., Diallo, A., Pushak, T., Uddin, E., Tsimpo, C., & Foster, V. (2008). Access, affordability, and alternatives: Modern infrastructure services in Africa.
3. Clean Water for Rural Uganda. (n.d.). Retrieved from https://www.ifc.org/wps/wcm/connect/news_ext_content/ifc_external_corporate_site/news and events/news/clean water for uganda
4. Few, Stephen. “Rich Data, Poor Data: Designing Dashboards to Inform” Available at http://www.perceptualedge.com/articles/Whitepapers/Rich_Data_Poor_Data.pdf
5. Few, Stephen. "Common Pitfalls in Dashboard Design.” Available at http://www.perceptualedge.com/articles/Whitepapers/Common_Pitfalls.pdf
6. Fink, A. (2003). The survey kit: How to manage, analyze, and interpret survey data. Thousand Oaks: Sage.
7. Gamble C, Ekwaru PJ, Garner P, ter Kuile FO, “Insecticide-treated nets for the prevention of malaria in pregnancy: a systematic review of randomised controlled trials. PLoS Med 2007, 4:e107. pmid:17388668
8. “Company Salaries.” Glassdoor, www.glassdoor.com/Salaries/index.htm.
9. Hox, J. J., Moerbeek, M., van de Schoot, R. (2018). Multilevel Analysis. New York: Routledge.
10. Levenshtein, V. I. (1966, February). Binary codes capable of correcting deletions, insertions, and reversals. In Soviet physics doklady (Vol. 10, No. 8, pp. 707-710).
11. Pazzaglia, A. M., Stafford, E. T., & Rodriguez, S. M. (2016). Survey methods for educators: Analysis and reporting of survey data (part 3 of 3) (REL 2016–164). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast & Islands. Retrieved from http://ies.ed.gov/ncee/edlabs
12. Schmidberger, M., & Friedman, A. (2018, September 20). Running R on AWS | Amazon Web Services. Retrieved November 5, 2018, from https://aws.amazon.com/blogs/big-data/running-r-on-aws/
13. UN. 2010. The world’s women 2010. Trends & statistics. Series. No. 19. UN statistics division. N.Y.

10 Appendix

KadAfrica Training Curriculum

Table of Contents

Acknowledgements 3
About KadAfrica 3
POWERFUL BEGINNINGS 5
LIFE SKILLS 7
SELF-AWARENESS 7
SELF-ESTEEM, CONFIDENCE AND VISUALIZATION 11
GOAL SETTING 13
DECISION-MAKING 16
GENDER 18
EFFECTIVE COMMUNICATION SKILLS 21
PUBLIC SPEAKING 24
CONFLICT MANAGEMENT 25
HEALTHY RELATIONSHIPS AND DOMESTIC VIOLENCE 29
NETWORKING 34
ROLE MODELS AND GUEST SPEAKER 37
PLANNING A COMMUNITY EVENT 39
REPRODUCTIVE HEALTH 41
HEALTHY BODIES AND NUTRITION 41
SAFE WATER MANAGEMENT / CLEAN WATER 46
ALCOHOL AND SUBSTANCE ABUSE 50
HYGIENE 52
ADOLESCENT GROWTH AND DEVELOPMENT 54
MENSTRUATION 61
SEX AND LOVE 73
PREGNANCY 78
OBSTETRIC FISTULA 82
ABORTION 83
SEXUALLY TRANSMITTED INFECTIONS 86
HIV/AIDS PREVENTION 91
LIVING HEALTHY WITH HIV 98
ADVOCACY: Creating within your community 101
BUSINESS SKILLS 116
UNDERSTANDING BUSINESS 116
THE MOTIVE OF DOING BUSINESSS 118
BUSINESS ENVIRONMENT 120
SUPPLY AND DEMAND 122
PRODUCTION AND SALES 126
MONEY MANAGEMENT AND BUDGETING 129
MARKETING 133
BOOKKEEPING 135
BUSINESS PLAN 139