

The School of Mathematics



THE UNIVERSITY  
*of* EDINBURGH

# Risk Factors for Leptospirosis Infection in a Kenyan Pastoral Landscape

by

Aditya Prabaswara Mardjikoan, s2264710

Dissertation Presented for the Degree of  
MSc in Statistics with Data Science

July 2022

Supervised by  
Dr Gail Robertson and Dr Amy Wilson



## Executive Summary

Due to the enormous impact an infection can have on human and animal morbidity and death, leptospirosis has been classified as a disease of worldwide public health importance [9]. These epidemics have increased in frequency during the past few decades, especially in developing nations [16]. In particular, numerous studies have shown that this zoonotic illness is becoming more common due to various factors, but little has been done to pinpoint individual who are most at risk, particularly in low-income rural areas of developing nations. As a result, the purpose of this report will be to identify individual risk factors for leptospirosis infections in a Kenyan rural area.

We used the International Livestock Research Institute data on household members at villages in Tana River County, Kenya, who undergo ELISA(enzyme-linked immunosorbent assay) test in 2013 and 2014. The main methods we employed in this reports are Binomial generalized linear mixed model (GLMM) [24] and Odds Ratio (OR) [10]. We discover that individual risk factors for leptospirosis infections include increasing age, female gender, living in a large family, living or having contact with their household head who works as a pastoralist, living in a lower-altitude village, and living or having contact with a younger household head.

## Acknowledgments

I am grateful to the supervisors of this project, Dr Gail Robertson and Dr Amy Wilson, for their support and valuable advice. I would also like to thank Bernard Bett and the International Livestock Research Institute for providing the background material and data for this project.

# University of Edinburgh – Own Work Declaration

Name: Aditya Prabaswara Mardjikoen

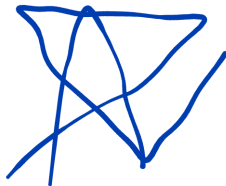
Matriculation Number: s2264710

Title of work: Risk Factors for Leptospirosis Infection in a Kenyan Pastoral Landscape

I confirm that all this work is my own except where indicated, and that I have:

- Clearly referenced/listed all sources as appropriate
- Referenced and put in inverted commas all quoted text (from books, web, etc)
- Given the sources of all pictures, data etc. that are not my own
- Not made any use of the report(s) or essay(s) of any other student(s) either past or present
- Not sought or used the help of any external professional academic agencies for the work
- Acknowledged in appropriate places any help that I have received from others (e.g. fellow students, technicians, statisticians, external sources)
- Complied with any other plagiarism criteria specified in the Course handbook

I understand that any false claim for this work will be penalised in accordance with the University regulations (<https://teaching.maths.ed.ac.uk/main/msc-students/msc-programmes/statistics/data-science/assessment/academic-misconduct>).

A handwritten signature in blue ink, consisting of several overlapping loops and lines, positioned in the lower-left area of the page.

Edinburgh, 29 July 2022

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Background and Motivation . . . . .	1
1.2	Objective . . . . .	1
1.3	Literature Review . . . . .	1
1.4	Data . . . . .	1
<b>2</b>	<b>Exploratory Data Analysis and Data Preparation</b>	<b>2</b>
2.1	Missing Data Analysis . . . . .	2
2.2	ELISA Test Result Analysis . . . . .	3
2.3	Leptospirosis Prevalence Based on Demographic Profile . . . . .	3
2.4	Leptospirosis Prevalence Based on Livestock Ownership and Geographical Location . .	6
2.5	Leptospirosis Prevalence Based on Household Head Occupation . . . . .	10
<b>3</b>	<b>Models</b>	<b>12</b>
3.1	Fixed and Random Effects . . . . .	12
3.2	Generalized Linear Mixed Model . . . . .	12
3.3	Modelling Considerations . . . . .	12
3.4	Leptospirosis Test Results Prediction Model . . . . .	13
3.5	Odds Ratio . . . . .	14
3.6	Model Selection . . . . .	14
3.7	Model Fitting and Diagnostics . . . . .	14
<b>4</b>	<b>Results</b>	<b>16</b>
<b>5</b>	<b>Conclusion</b>	<b>17</b>
	<b>Appendices</b>	<b>20</b>
<b>A</b>	<b>Acronyms</b>	<b>20</b>
<b>B</b>	<b>R Package</b>	<b>20</b>

## List of Tables

1	Link functions for GLM [5]. . . . .	12
2	Results of model comparisons. . . . .	14
3	Results of Binomial GLMM analysis for leptospirosis seropositivity. . . . .	16

## List of Figures

1	Proportion of missing data in each variable of interest and patterns of missing data. .	2
2	Proportion of person from specific household in villages in Tana River County, Kenya who exposed and not exposed to leptospirosis according to ELISA test result. . . . .	3
3	Number of individuals in the data based on their occupations. . . . .	4
4	Leptospirosis prevalence based on age, gender and occupation (with above or 20 worker sampled in the data). . . . .	5
5	Leptospirosis prevalence based on livestock existence at household and sampling site characterization based on land use. . . . .	6
6	Leptospirosis prevalence based on constituency, village altitude and hospital distance. .	7
7	Number of individuals in the data based on villages and locations. . . . .	8
8	Leptospirosis prevalence based on location and village with above or 20 individuals sampled in the data. . . . .	9
9	The number of household heads who work in the given household head occupation in the data. . . . .	10
10	Leptospirosis prevalence based on household head occupation with above or 20 household head works in that occupation in the data. . . . .	11
11	Some model diagnostic plots for leptospirosis test results prediction model. . . . .	15

# 1 Introduction

## 1.1 Background and Motivation

Leptospirosis is a neglected but re-emerging zoonotic disease which has been identified as a disease of global public health importance due to the significant effect infection can have on human and animal morbidity and mortality [9]. Acute fever of unknown cause is the disease's most early sign. *Leptospira* is the most typical bacterium that causes this illness [16]. It is a water-borne bacteria that can cause febrile illness with a chance of serious complications if it is left untreated.

Over the past few decades, leptospirosis outbreaks have become more frequent, especially in developing nations [16]. While outbreaks are often reported in urban areas after flooding events, rural communities are highly vulnerable to leptospirosis outbreaks due to possible risk factors such as outdoor working, contact with animals, and poor sanitation. Various studies have documented a rise in the incidence of this zoonotic infection due to various factors including climate change [12], but less work has been done on identifying individuals at risk, especially in low-income rural areas in developing countries. Therefore, it is important to understand and identify individual risk factor of leptospirosis in order to reduce the disease transmission rate.

## 1.2 Objective

The goals of this report is to identify individual risk factors for leptospirosis infection in a rural area of Kenya.

## 1.3 Literature Review

We now give a brief review regarding the literature about risk factors of leptospirosis. The earliest study we consider is [12]. They find that climate change, flooding, population growth and urbanisation can lead to an escalation of leptospirosis outbreaks. Another study conducted by [9] suggest that leptospirosis is an occupational disease associated with freshwater or animal exposure.

Some study, such as [2], manage to identify individual risk factors of leptospirosis in western Kenya, however, they only consider individual that belongs into slaughterhouse workers group. They discover that personal hygiene factor appear to have the most influence on the risk of leptospirosis transmission. Over the past few years, study conducted by [16] discover that rats is the primary reservoir of leptospirosis. In this report, we will use some methods in [2] to identify individual risk factors for leptospirosis in rural area of Kenya.

## 1.4 Data

The International Livestock Research Institute tested household members for leptospirosis using ELISA (enzyme-linked immunosorbent assay) in 2013 and 2014 in villages in Tana River County, Kenya, and recorded the results in the data used for this report. Each observation in the data represents a leptospirosis test result from one sample taken from a person on a given date from a specific household. Additionally, questionnaire responses from test subjects are documented. It contains demographic data (age and sex) as well as information about the respondent's and their head of household's behaviours.

The location of the respondent's residence as well as details about their work and animal contact behaviour are also gathered. Besides that, we also have data regarding village altitude (measured in metres) and hospital distance (measured in kilometres). The hospital distance is calculated from the household to the local hospital, whereas the village altitude is recorded via GPS.



# 2 Exploratory Data Analysis and Data Preparation

## 2.1 Missing Data Analysis

We discover that there exist 12 duplicated observations in the data and remove them. Figure 1 (see Appendix A for acronyms descriptions) show that around 36% of the data was missing, primarily the occupation of the person sampled from a specific household. Overall, there are around 48% observations with missing data.

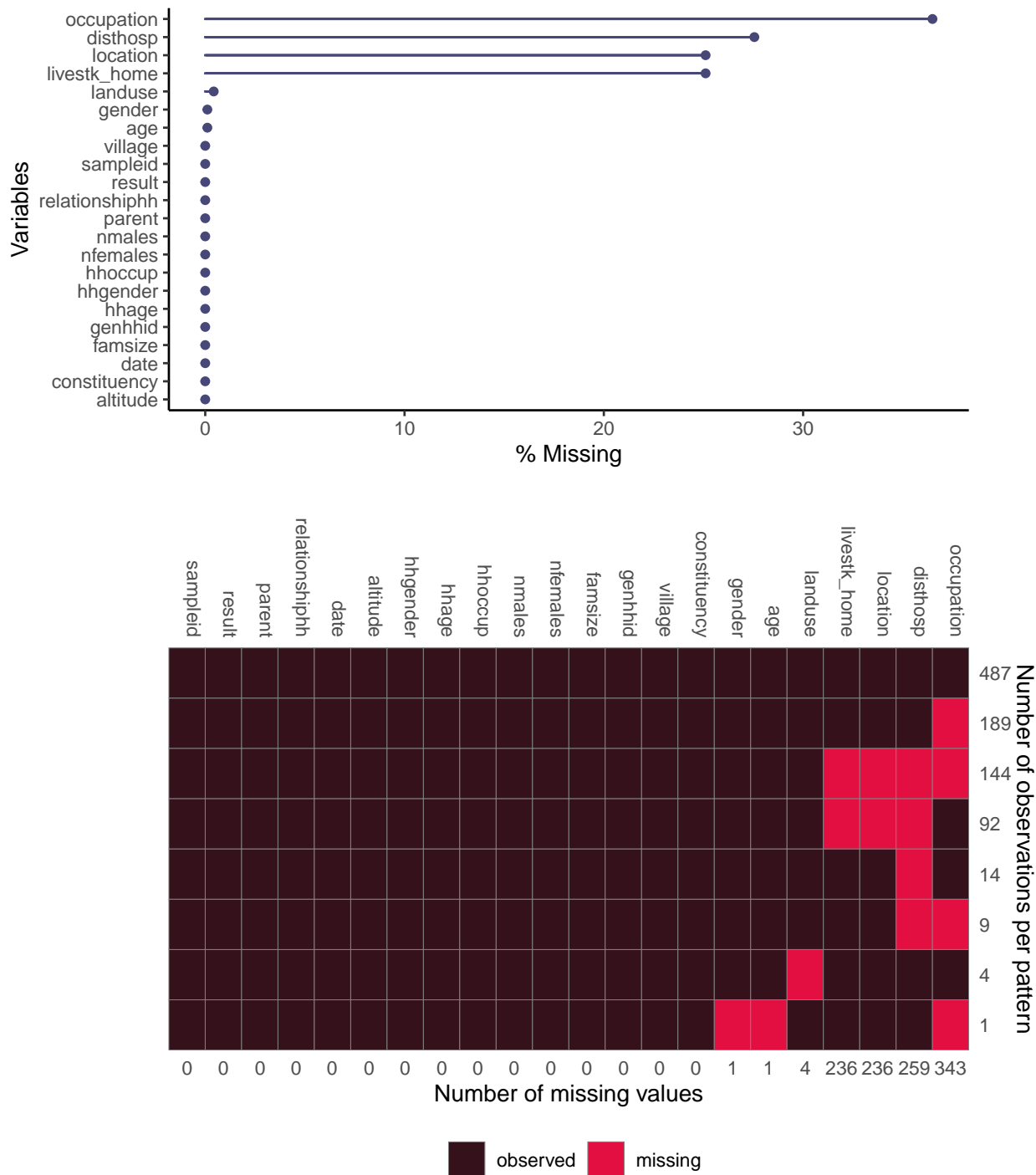


Figure 1: Proportion of missing data in each variable of interest and patterns of missing data.

## 2.2 ELISA Test Result Analysis

We start by looking at the descriptive statistics for the ELISA results for the leptospirosis test. Figure 2 displays the proportions of participants who underwent this immunological assay and were exposed to leptospirosis compared to those who were not. The findings of the ELISA test show that leptospirosis was detected in around 26% of the participants.

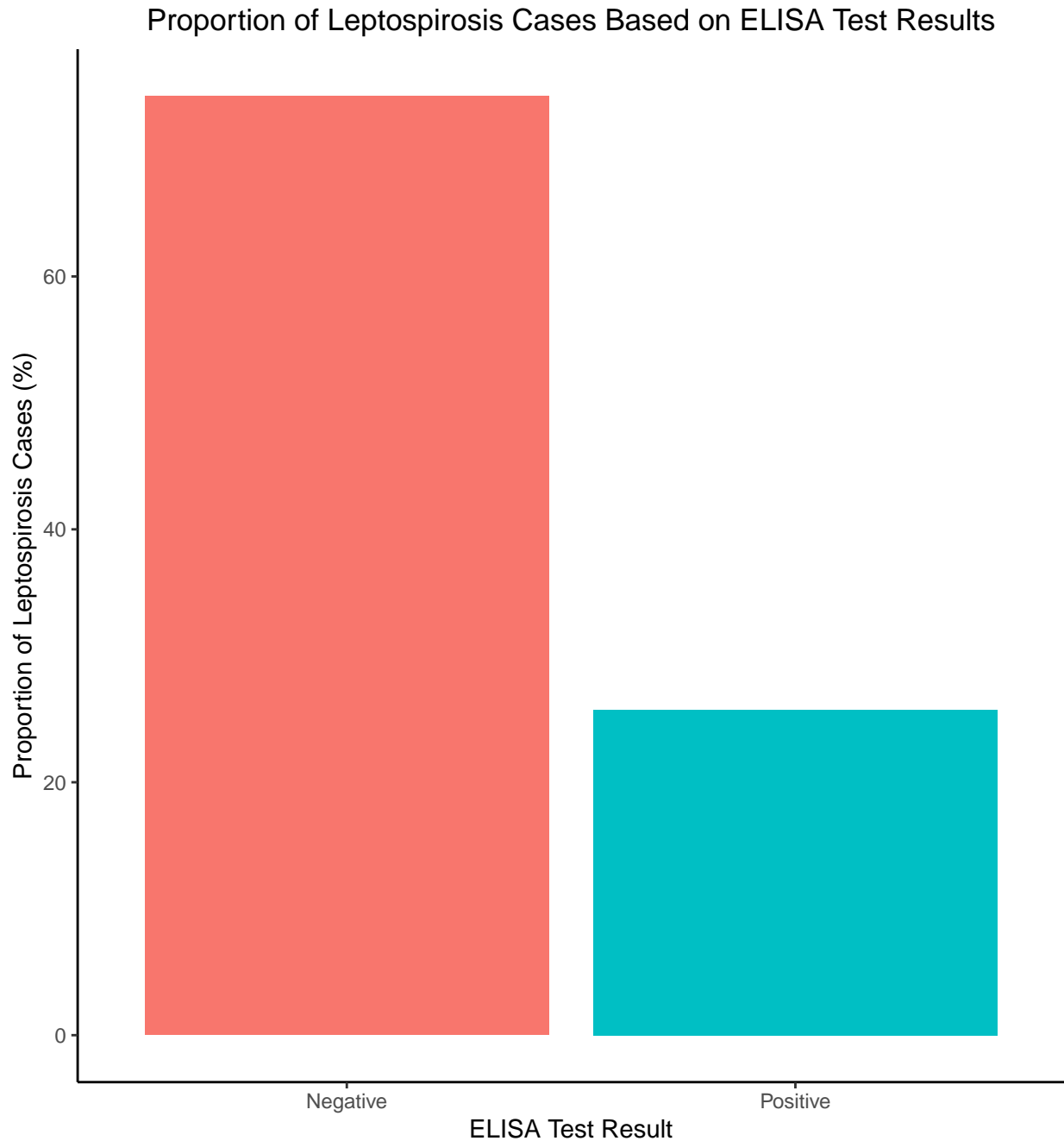


Figure 2: Proportion of person from specific household in villages in Tana River County, Kenya who exposed and not exposed to leptospirosis according to ELISA test result.

## 2.3 Leptospirosis Prevalence Based on Demographic Profile

Prevalence [10] is the proportion of a particular population found to be affected by a disease at a certain time. Because the data was collected between 2013 and 2014, we define the prevalence of leptospirosis cases as the proportion of a population in rural area of Kenya that tested positive for leptospirosis using ELISA between those years.

We started by investigating the number of worker at each occupation that was sampled in the data. Figure 3 depicts that most individual in the data worked as a pastoralist. We decided not calculating the prevalence of leptospirosis cases for jobs with less than 20 workers sampled in the data since doing so would have indicated that certain of those occupations had a larger prevalence of leptospirosis cases than other occupations in the data.

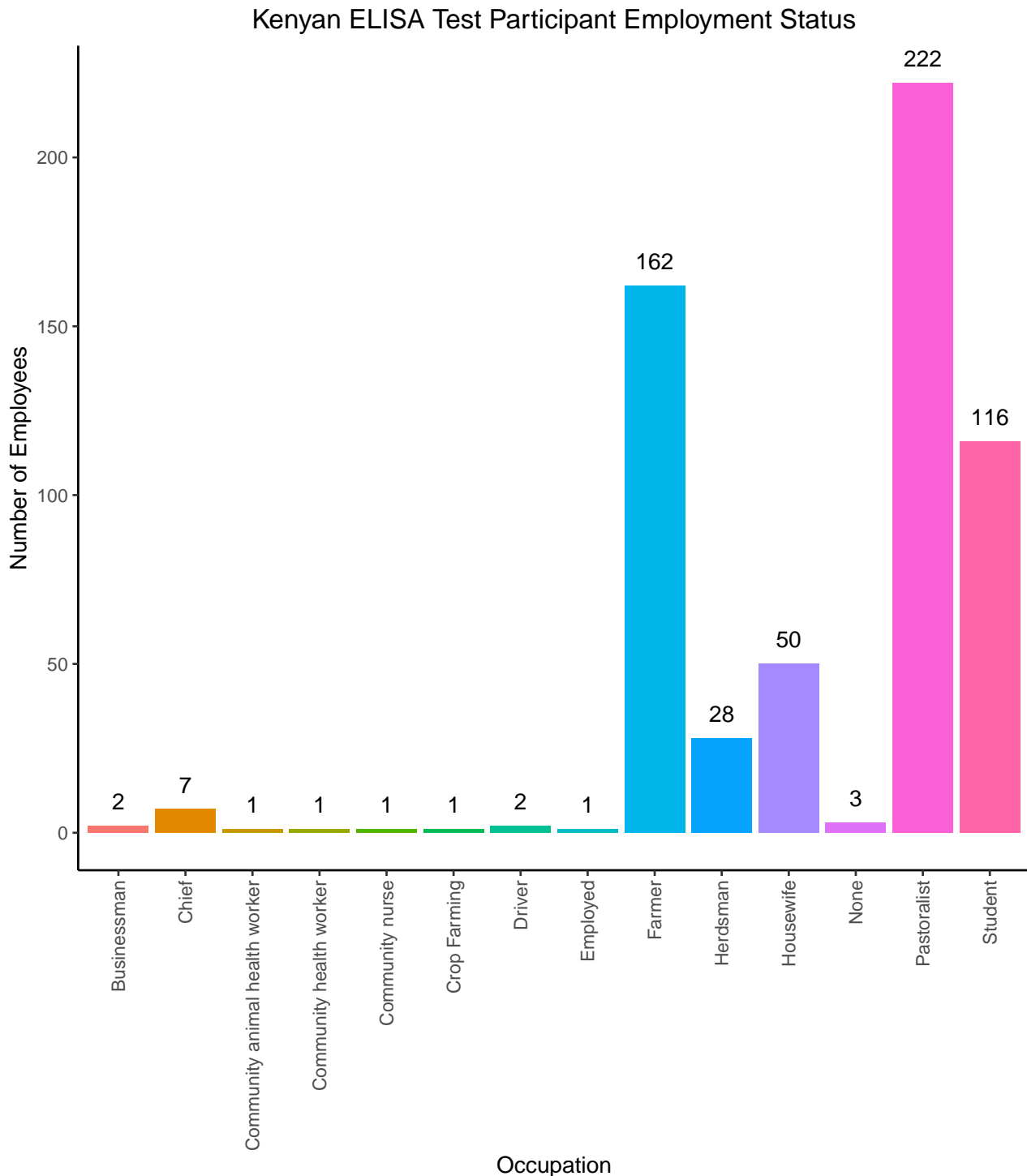


Figure 3: Number of individuals in the data based on their occupations.

We explore the prevalence of leptospirosis cases based on demographic profile (age, gender and occupation). Figure 4 illustrates the prevalence of leptospirosis cases in settlements in Tana River County, Kenya based on the individuals demographic profile. According to the upper left plot, people between the ages of 20 and 49 are most commonly succumb to leptospirosis (the prevalence is around

33.3%). It appears that females are most frequently infected with leptospirosis, as we can see on the upper right plot (the prevalence is around 29%). The lower plot indicates that the majority of leptospirosis cases occur among people who work as pastoralists (the prevalence is around 30%).

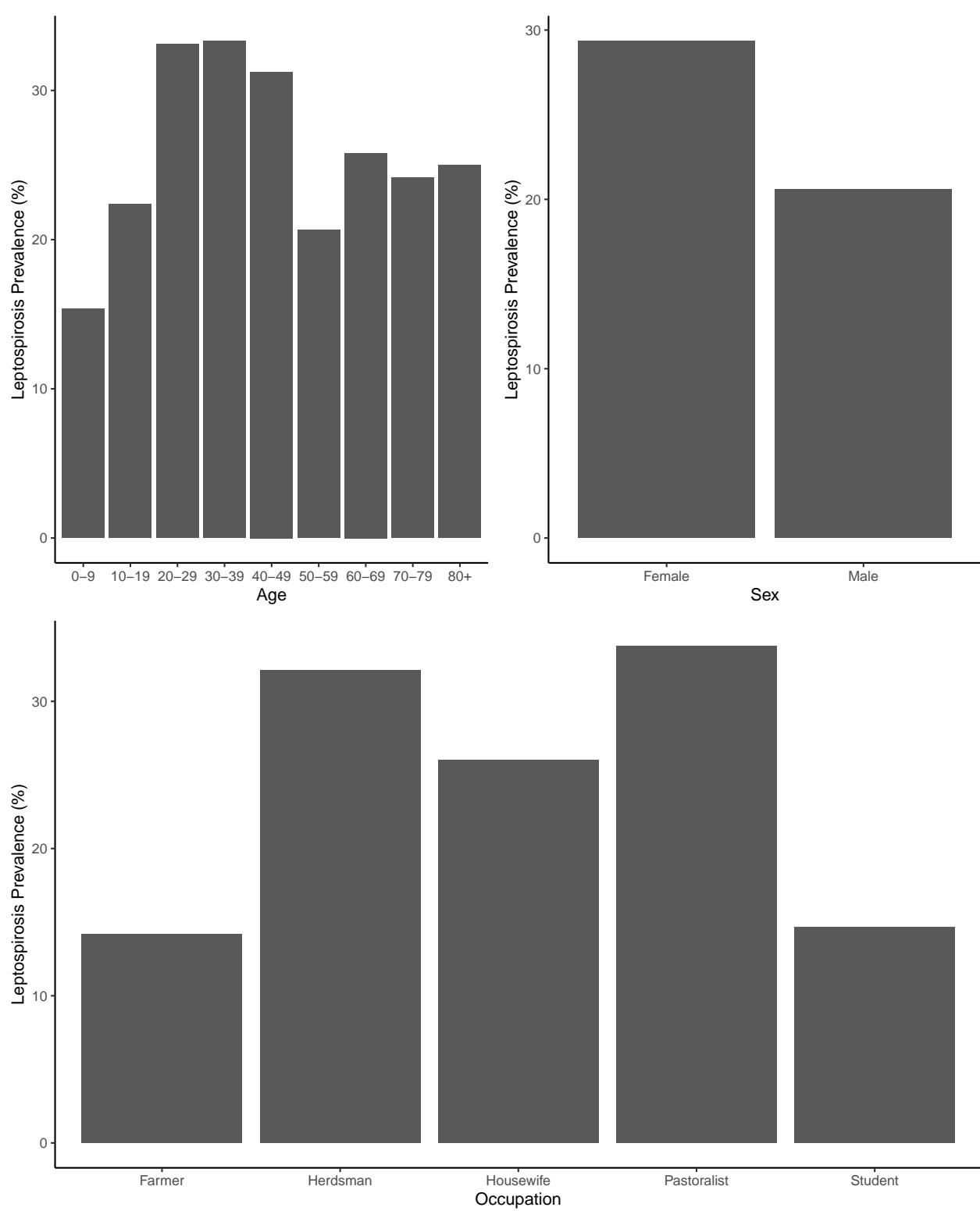


Figure 4: Leptospirosis prevalence based on age, gender and occupation (with above or 20 worker sampled in the data).

## 2.4 Leptospirosis Prevalence Based on Livestock Ownership and Geographical Location

We explore the prevalence of leptospirosis cases based on the existence of livestock in the household of the sampled persons and sampling site characterization based on land use. Figure 5 illustrates that leptospirosis cases are most likely to occur at an individual who kept livestock in the household (the prevalence is around 23%). Moreover, we also discover that leptospirosis cases are most likely to take place in pastoral land use (the prevalence is around 31%).

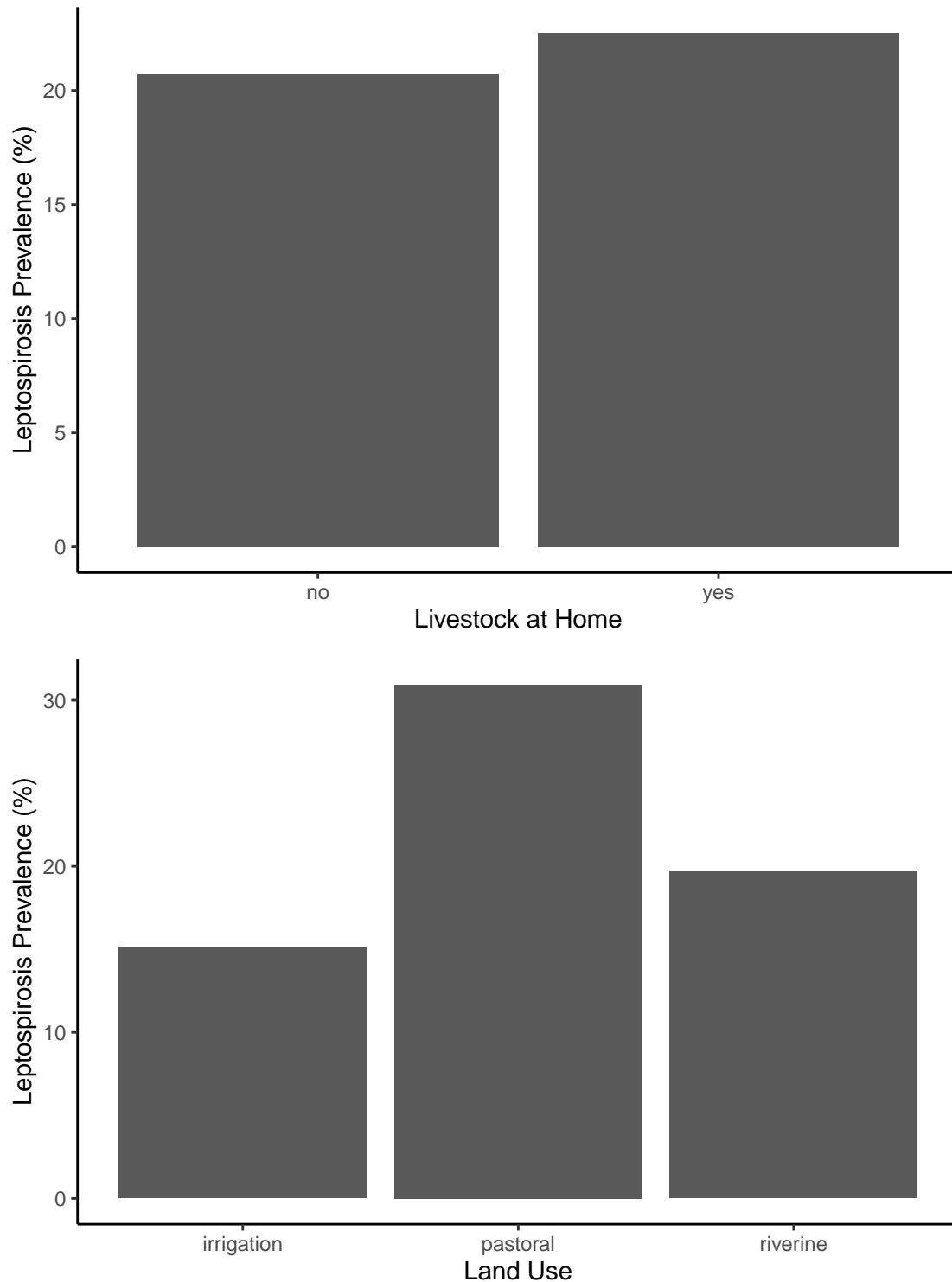


Figure 5: Leptospirosis prevalence based on livestock existence at household and sampling site characterization based on land use.

Next, we explore the prevalence of leptospirosis cases based on constituency, village altitudes and hospital distances. Figure 6 illustrates that leptospirosis cases are most likely to occur in Ijara (the prevalence is around 40%). Moreover, leptospirosis cases are more common in villages with an altitude of 50 metres or less (the prevalence is around 42%). Furthermore, leptospirosis cases are more frequent in an area with a distance from household to local hospital of between 22 and 32 kilometres (the prevalence is around 32%).

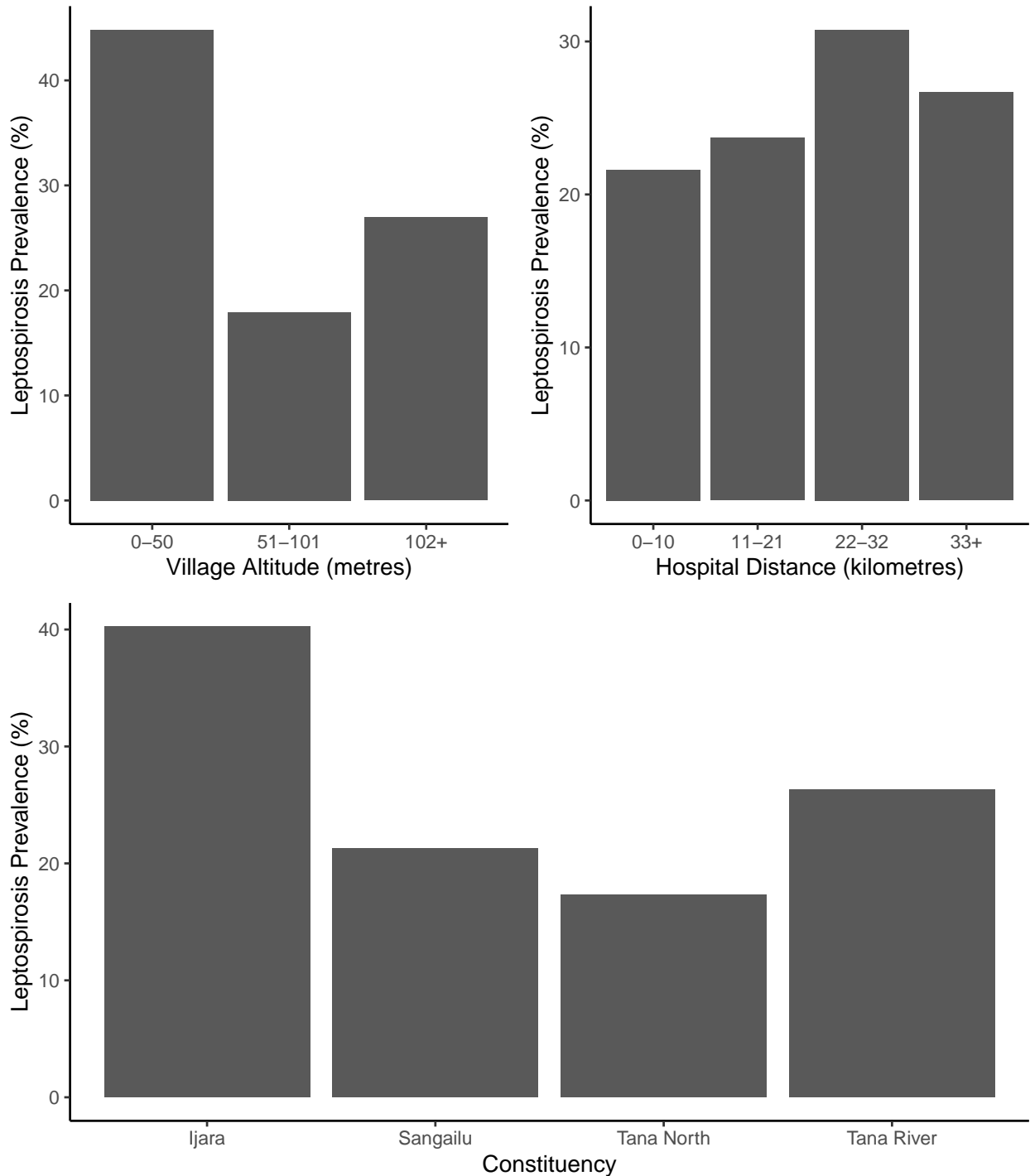


Figure 6: Leptospirosis prevalence based on constituency, village altitude and hospital distance.

Lastly, we investigate the prevalence of leptospirosis cases based on location and village. Figure 7 illustrates that the individual in the data is most likely sampled from village 12. Additionally, it appears that sampling is most frequently done in location 3. We opted not to compute the prevalence

of leptospirosis cases for locations and villages with fewer than 20 individuals sampled in the data since doing so would have shown that specific locations and villages had a higher prevalence of leptospirosis cases than other locations and villages in the data.

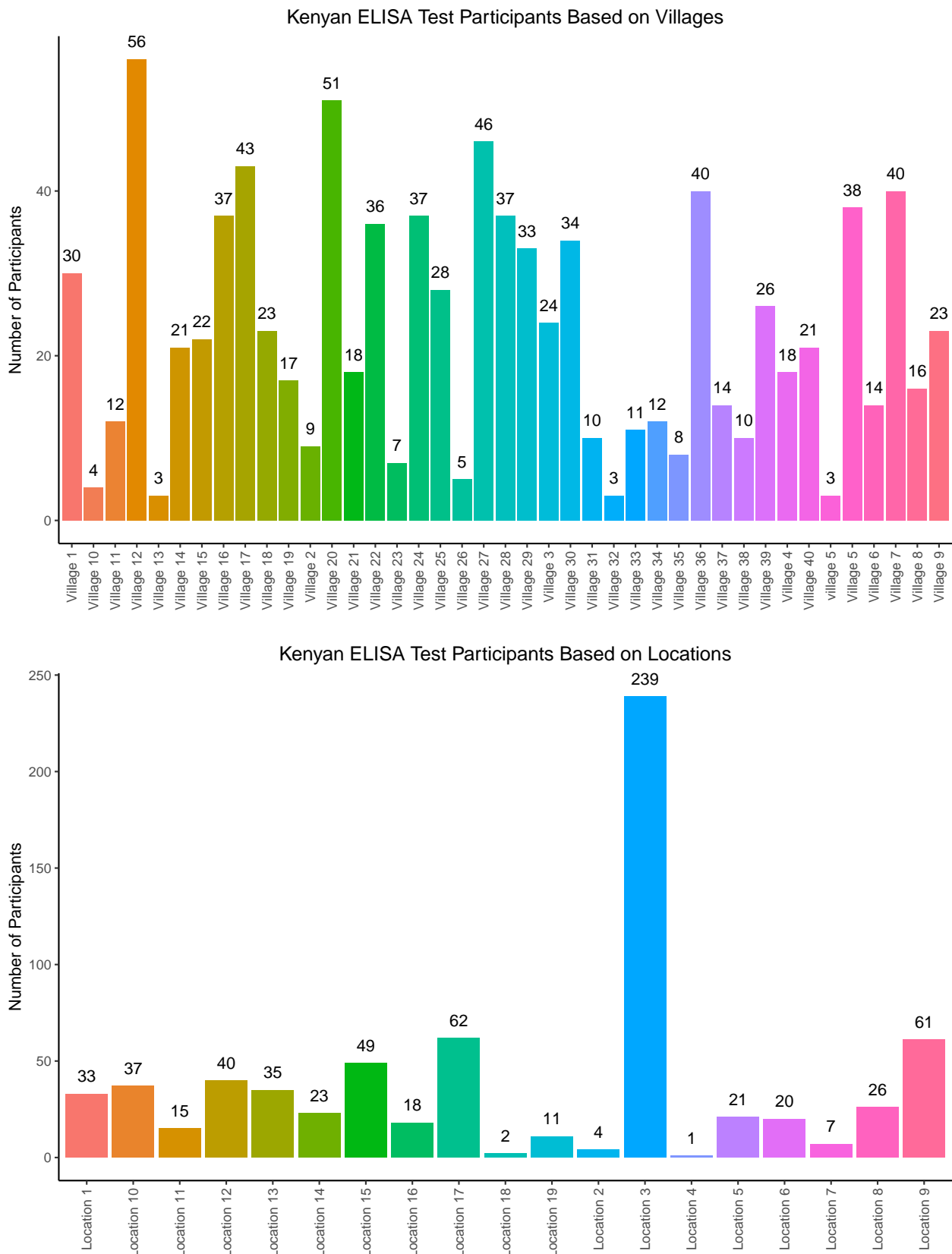


Figure 7: Number of individuals in the data based on villages and locations.

Figure 8 illustrates that leptospirosis cases are most likely to occurred at village 25 (the prevalence is around 59%). It appears that leptospirosis cases are most likely to appeared in location 1 (the prevalence is around 62.5%).

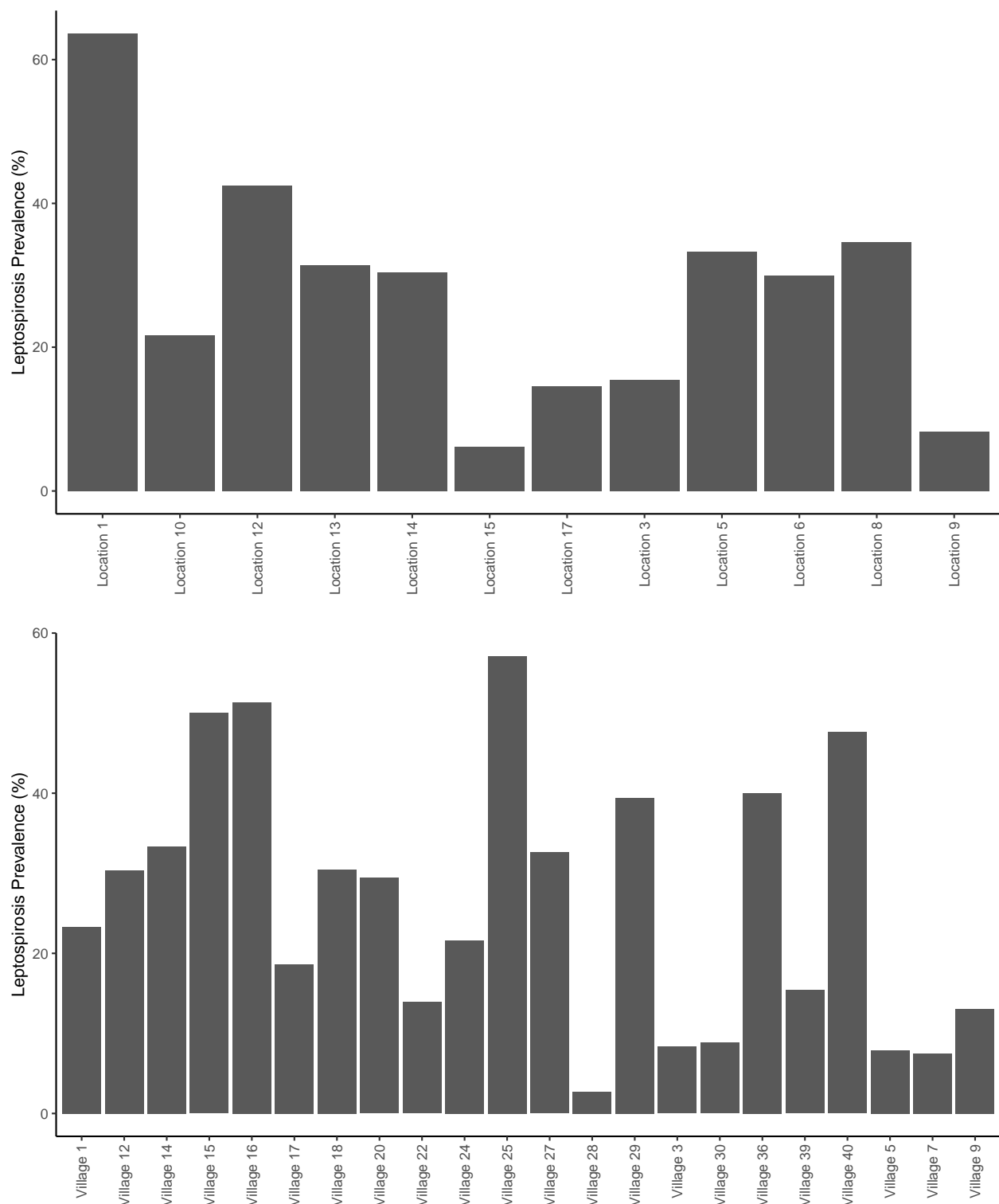


Figure 8: Leptospirosis prevalence based on location and village with above or 20 individuals sampled in the data.



## 2.5 Leptospirosis Prevalence Based on Household Head Occupation

Figure 9 depicts that most household head works as a pastoralist. We decided not calculating the prevalence of leptospirosis cases for household head jobs with less than 20 household head works in these occupations since doing so would have indicated that certain of those household head occupations had a larger prevalence of leptospirosis cases than other household occupations in the data. According to Figure 10, leptospirosis cases is most likely to occur at an individual where their household head works as a pastoralist (the prevalence is around 32%).

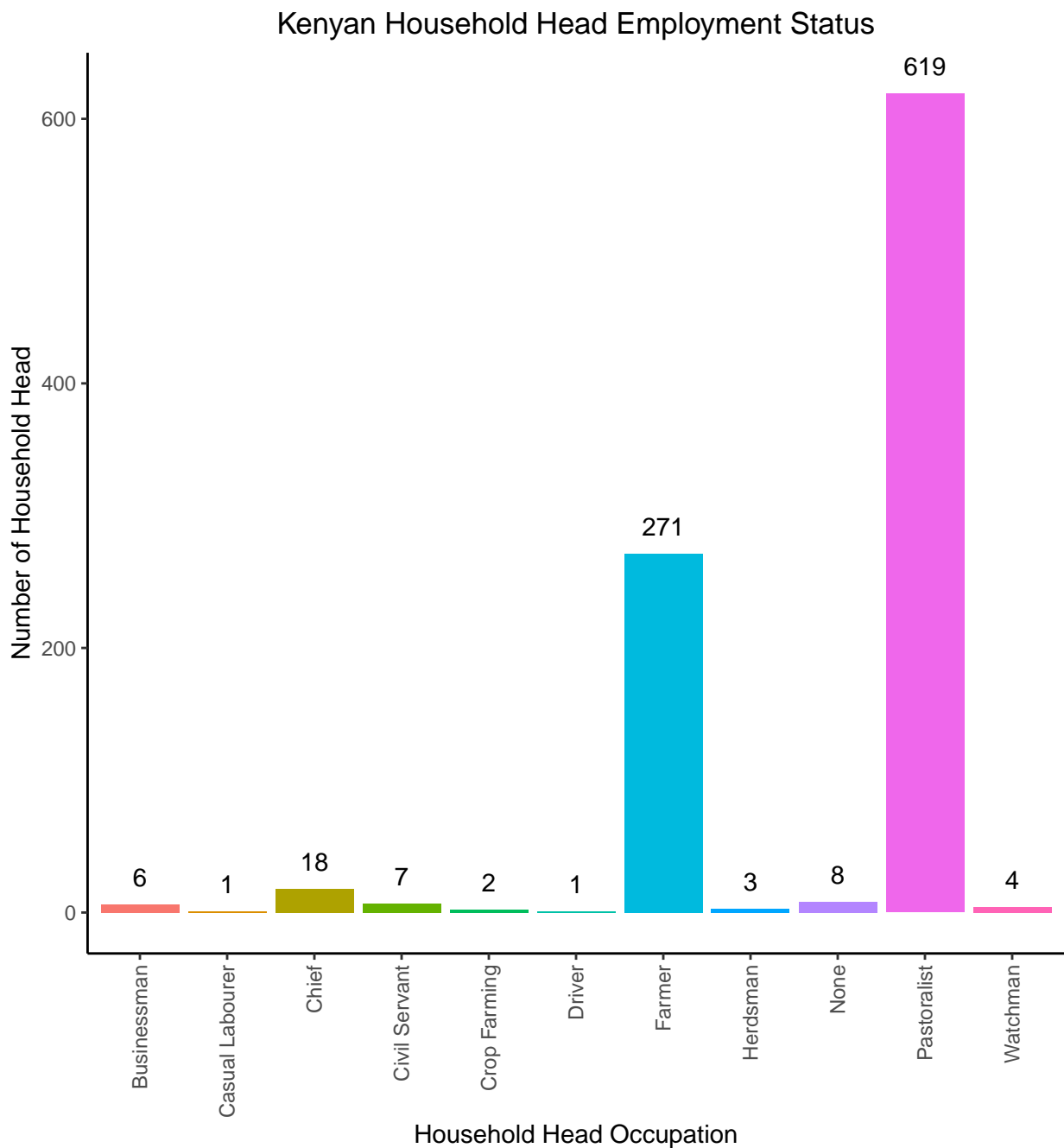


Figure 9: The number of household heads who work in the given household head occupation in the data.

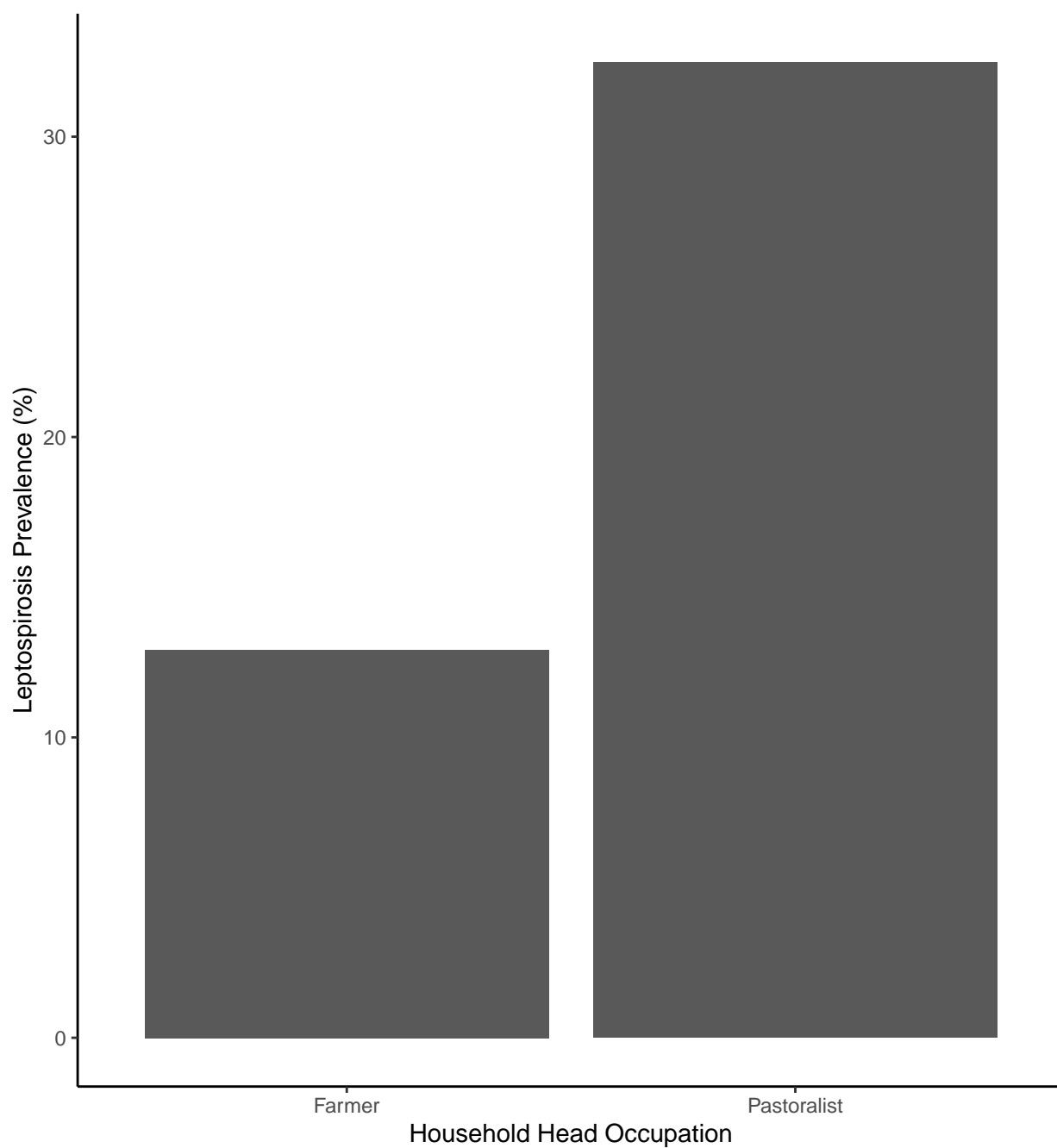


Figure 10: Leptospirosis prevalence based on household head occupation with above or 20 household head works in that occupation in the data.

### 3 Models

#### 3.1 Fixed and Random Effects

Mixed effects model or mixed model is a regression models that contain both fixed and random effects [5]. An example of this model is discussed in Section 3.2. We define effects (or coefficients) in the mixed effects model as constant if they are similar for all groups in a population and varying if they are possible to vary between groups to groups [6]. For instance, the model  $y_{ij} = \alpha_j + \beta x_{ij} + \epsilon_{ij}$  (of the  $i$ th observations in groups  $j$ ) has a constant slope and varying intercept, and  $y_{ij} = \alpha_j + \beta_j x_{ij} + \epsilon_{ij}$  has a varying slope and intercept.

Fixed effects are effects in which the population elements are fixed, whereas random effects are effects in which the population elements can change [6]. Thus, we can consider random effects as a random variables [5]. The distinguishing feature of fixed and random effects is that fixed effects assumed that observations are independent, whereas random effects presume that some observations have a relationship [6]. For instance, gender is a fixed effect variable since the only conceivable values of that variable are males and females, and those values are independent of one another. A study on food pricing at restaurants in different locations can consider location as a random effects because prices can vary depending on location.

#### 3.2 Generalized Linear Mixed Model

Generalized linear mixed model (GLMM) is a generalized linear model (GLM) with random effects terms [5]. In matrix form, we can formulate this model as follows:

$$\boldsymbol{\eta} = g(E[\mathbf{y}]) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}$$

where  $\mathbf{X}$  is the design matrix for the fixed effects observations (predictors),  $\boldsymbol{\beta}$  is the fixed effects coefficients column vectors,  $\mathbf{Z}$  is the design matrix for the random effects observations,  $\boldsymbol{\gamma}$  is the random effects,  $\boldsymbol{\eta}$  is the linear predictors column vectors, and  $\mathbf{y}$  is the column vector of response variables with mean  $E[\mathbf{y}]$ . The link function  $g(\cdot)$  relates the outcome  $\mathbf{y}$  with the linear predictors  $\boldsymbol{\eta}$ . Table 1 provide some examples of link functions based on the distribution of the response variables.

Table 1: Link functions for GLM [5].

Family	Link Function
Normal	$g(\mu) = \mu$
Poisson	$g(\mu) = \log \mu$
Binomial	$g(\mu) = \log \left( \frac{\mu}{1 - \mu} \right)$
Gamma	$g(\mu) = \mu^{-1}$
Inverse Gaussian	$g(\mu) = \mu^{-2}$

#### 3.3 Modelling Considerations

After analysing our data in Section 2, we decided not to include occupation, presence of livestock at home, location and hospital distance because these variables contain 30% or more missing data, which would cause some variables that should have a significant effect on leptospirosis infection to become insignificant because we reduced around 48% of the data if we include these variables. As a result, we decide to consider the occupation of the household head instead, because this variable is the only variable related with occupations which is fully-observed and leptospirosis can occur through human-to-human transmission [9].

We decide to omit observations where the household head occupation had fewer than 20 household head as the workers recorded in the data, as keeping these observations before fitting our mixed effects

model would increase the uncertainty of the estimated and true risk of leptospirosis infections in people who live with the household head that works in these occupations and grouping these occupations is difficult due to some occupation have different characteristic with others. Thus, we only consider two type of household head occupation in our model: farmer and pastoralist. Due to we only consider this two types of household head occupation, we treat this variable as a fixed effects variable.

According to [12], the frequency of leptospirosis patients varies significantly by region due to factors including climate and environmental change. Thus, we would consider village, household and constituency as a random effects in our model. However, we will treat land use as a fixed effects variable since we only consider 3 land use in this report: irrigation, pastoral and riverine. As for gender, it is a fixed effects variables based on explanation in Section 3.1.

### 3.4 Leptospirosis Test Results Prediction Model

In this report, we used a Binomial GLMM [24] model since the response variable is a binary variable (takes only two values, in this case, negative and positive test results) and it enables us to account for both hierarchical data structures and correlations between observations (such as geographical correlations). Let  $Y_{ijk}$  be the ELISA test results for person  $i$  at village  $j$  and household  $k$ .  $Y_{ijk}$  is 1 if the ELISA test results for person  $i$  from village  $j$  and household  $k$  is positive and 0 if negative. We denoted  $p_{ijk}$  as the probability that person  $i$  from village  $j$  and household  $k$  is tested positive for leptospirosis. The Binomial GLMM model that we will used in this report is formulated as follows:

$$\begin{aligned}
Y_{ijk} &\sim \text{Bernoulli}(p_{ijk}) \\
\log\left(\frac{p_{ijk}}{1-p_{ijk}}\right) &= \alpha + \beta_1 \text{Age}_{ijk} + \beta_2 \text{Altitude}_{ijk} + \beta_3 \text{Gender}_{ijk} \\
&\quad + \beta_4 \text{HHOccupationPastoralist}_{ijk} + \beta_5 \text{HHAge}_{ijk} + \beta_6 \text{FamilySize}_{ijk} \\
&\quad + \beta_7 \text{LandusePastoral}_{ijk} + \beta_8 \text{LanduseRiverine}_{ijk} + \gamma_j + \lambda_k \\
\gamma_j &\sim N(0, \sigma_\gamma^2) \\
\lambda_k &\sim N(0, \sigma_\lambda^2)
\end{aligned} \tag{3.1}$$

where:

- $\text{Age}_{ijk}$  is the sampled person age;
- $\text{Gender}_{ijk}$  is the sampled person gender (1 = female and 0 = male);
- $\text{Altitude}_{ijk}$  is the altitude of the village where the person was sampled;
- $\text{HHOccupationPastoralist}_{ijk}$  is 1 if person  $i$  from village  $j$  and household  $k$  live with household head with occupation as a pastoralist and 0 if the household head work as a farmer;
- $\text{HHAge}_{ijk}$  is the household head age in household  $k$ ;
- $\text{FamilySize}_{ijk}$  is the number of family member who lived in household  $k$ ;
- $\text{LandusePastoral}_{ijk}$  is 1 if person  $i$  from village  $j$  and household  $k$  was sampled in pastoral land use and 0 if otherwise;
- $\text{LanduseRiverine}_{ijk}$  is 1 if person  $i$  from village  $j$  and household  $k$  was sampled in riverine land use and 0 if otherwise.

Model (3.1) is an example of logistic regression model [10]. We assume that  $Y_{ijk}$  has a Bernoulli (single-trial Binomial) distribution with probability  $p_{ijk}$ .  $\gamma_j$  is the random effect for villages and is assumed to be normally distributed with mean 0 and variance  $\sigma_\gamma^2$ . As for  $\lambda_k$ , it is the random effect for households and is assumed to be normally distributed with mean 0 and variance  $\sigma_\lambda^2$ . The parameter  $\alpha$  and  $\beta_i$  in Model (3.1) represents the fixed effects intercept and slope respectively.

### 3.5 Odds Ratio

We define the odds [10] of outcome  $D$  as the probability of  $D$  will occur divided by the probability that it will not occur. The Odds Ratio (OR) [10] compare the odds of  $D$  in the exposed and unexposed subgroups to examine their association. Let  $\alpha$  and  $\beta_i$  respectively be the slope and intercept for variables  $X_i$  in a logistic regression model with binary response variable  $D$ . We can interpret  $e^{\beta_i}$  as the OR of  $D$  associated with  $X_i$ , whereas  $e^\alpha$  represent the odds of  $D$  when all observed values for the predictors is zero. According to [10], the OR values for  $D$  with binary risk factor  $E$  can be interpret as follows:

1. OR = 1 suggest independence of  $D$  and  $E$ .
2. OR > 1 suggest there is a greater risk of  $D$  when  $E$  is present.
3. OR < 1 suggest there is a lower risk of  $D$  when  $E$  is present.

To estimate the precision of the OR, we used a 95% confidence interval (CI) [10, 5]. The interpretation of this intervals is if we repeat the study 100 times (100 sampling from the same population), we would expect the true values of the OR to lie within the derived intervals 95 times.

### 3.6 Model Selection

Model (3.1) was selected using Akaike information criterion (AIC) and Bayesian information criterion (BIC). The literature we use to give brief explanation about AIC and BIC is [5]. Let  $p$  and  $l(\hat{\theta})$  be the number of paramterers and maximised log-likelihood in a particular regression model with parameter of interest  $\theta$  respectively. We denoted  $n$  as the number of observations that we use to fit into this model. The AIC and BIC are defined as follows:

$$\begin{aligned} \text{AIC} &= -2l(\hat{\theta}) + 2p \\ \text{BIC} &= -2l(\hat{\theta}) + p \log n \end{aligned}$$

When we compare multiple model using AIC and BIC, we pick the model with the lowest AIC and BIC score. Initially, our first model is Model (3.1) with an addition of constituency as random effects intercept and household head gender as fixed effects variable. Table 2 shows that Model (3.1) have lower AIC and BIC scores than our first model, which indicates that Model (3.1) is better than our first model.

Table 2: Results of model comparisons.

Model	AIC	BIC
I	982.91	1045.18
II	981.11	1033.80

### 3.7 Model Fitting and Diagnostics

We fitted the data in Section 1.4 to Model (3.1) after some data preparation in Section 2 and 3.3. The middle and lower plot in Figure 11 exhibits the diagnostic plots to check the normality assumptions for the random effects terms in Model (3.1). The lower plot shows that the distribution of the random effects of village is approximating a normal distribution. As for the middle plot, it suggest that the distribution of the random effects of household is approximating a normal distribution although it is slightly right-skewed as the histogram sugessted.

We used the binned residuals plot [7] to examine the residual patterns in our fitted values for Model (3.1). The fitted values in this report represent the predicted probability of testing positive for leptospirosis. As [7] suggests, we divide the data in this report into categories (bins) based on their fitted values and then plot the average fitted values versus the average residuals for each bin, which

can be seen at the upper plot in Figure 11. The grey line represents the  $\pm 2$  standard error bounds, within which approximately 95% of the binned residuals would fall if Model (3.1) were true. Overall, most of the fitted values lies within the standard error bounds, which indicates that Model (3.1) looks reasonable.

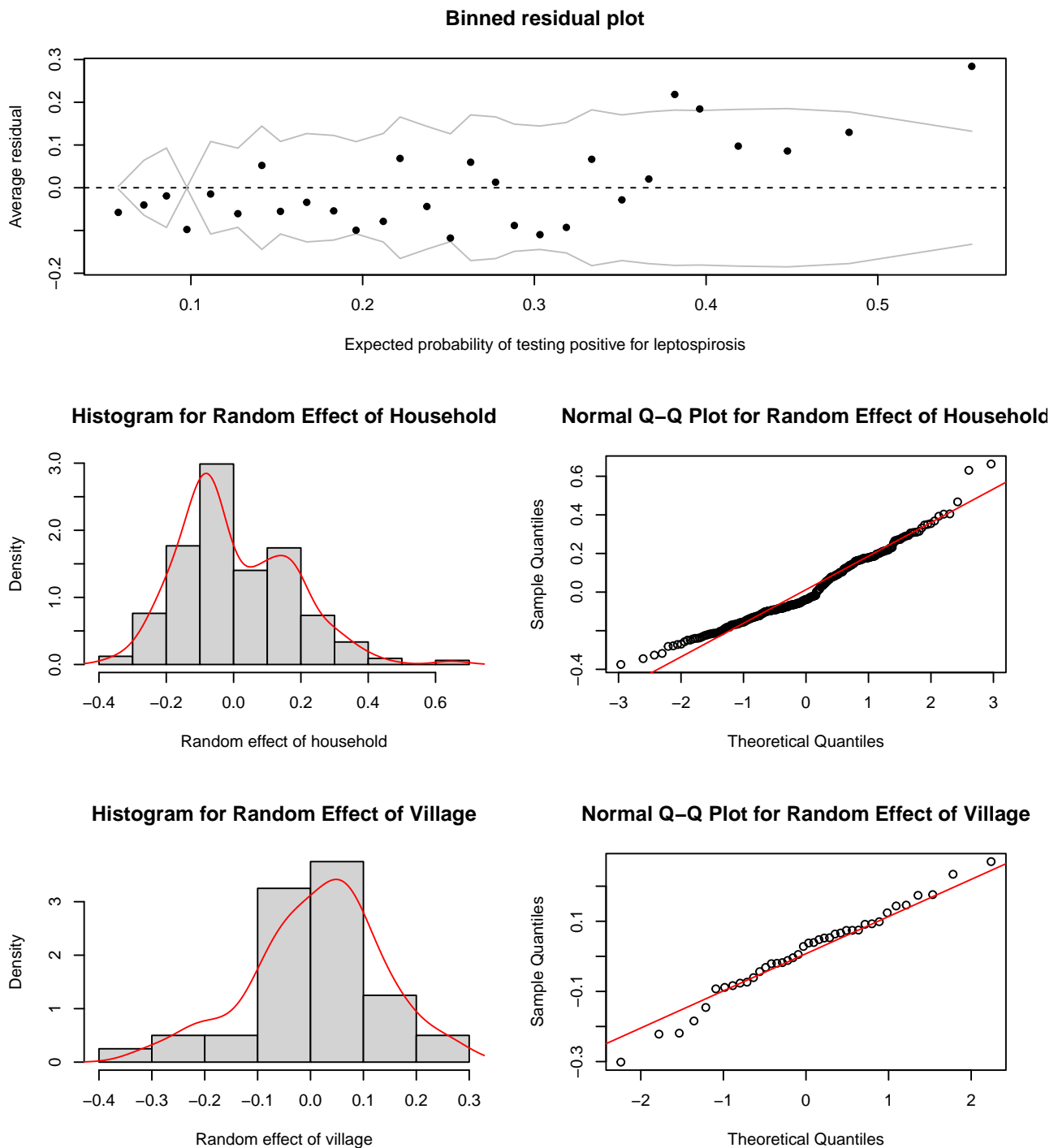


Figure 11: Some model diagnostic plots for leptospirosis test results prediction model.

## 4 Results

Model (3.1) was fitted with data from Section 3.7, and Table 3 displays the estimated odds ratios for leptospirosis seropositivity (tested positive for leptospirosis) [2] associated with the fixed effects predictor variables in Model (3.1), along with their 95% CI and p-value [10]. We consider the variable with p-value  $< 0.05$  were significantly associated with leptospirosis seropositivity. Overall, only variable related with land use that is not significantly associated with leptospirosis seropositivity.

Table 3: Results of Binomial GLMM analysis for leptospirosis seropositivity.

Variable	OR <sup>1</sup>	95% CI <sup>1</sup>	p-value
Age of the sampled person	1.0064	1.0033, 1.0096	$<0.001$
Village altitude	0.989	0.986, 0.992	$<0.001$
Gender of the sampled person			
Female	—	—	
Male	0.6173	0.4348, 0.8766	0.007
Age of the household head	0.991	0.988, 0.995	$<0.001$
Family size	1.0366	1.0332, 1.0401	$<0.001$
Household head occupation			
Farmer	—	—	
Pastoralist	5.8191	1.8675, 18.132	0.002
Land use			
irrigation	—	—	
pastoral	0.3719	0.1163, 1.1889	0.10
riverine	0.4766	0.1645, 1.3809	0.2

<sup>1</sup>OR = Odds Ratio, CI = Confidence Interval

Based on Table 3, risk factors that were significant for leptospirosis seropositivity were: increasing individual age (OR 1.0064; 95%CI 1.0033 to 1.0096), increasing family size (OR 1.0366; 95%CI 1.0332 to 1.0401), and household head working as a pastoralist (OR 5.8191; 95%CI 1.8675 to 18.132). Protective factors that were significant for leptospirosis seropositivity were: male individual (OR 0.6173; 95% CI 0.4348 to 0.8766), increasing village altitude (OR 0.989; 95% CI 0.986 to 0.992), and increasing household head age (OR 0.991; 95% CI 0.988 to 0.995).

## 5 Conclusion

Based on the findings in Section 4, we can conclude that individuals who live in a large family, live or have contact with their household head who works as a pastoralist, are older, belong to a female individual, live in a lower-altitude village, and live or have contact with a younger household head, particularly the household head who is still of working age, are more likely to be seropositive to leptospirosis.

This report has two shortcomings. The first limitation of this analysis is that we were unable to include variables with more missing data in our model, such as location, presence of livestock at home, distance from household to local hospital, and sampled individual's occupation. Despite our efforts to account for this variable, the model produces results in which almost all of the variables in the model are not significantly associated with leptospirosis seropositivity.

The second limitation of this analysis is that we were unable to compare the risk factors associated with all different occupation variables (individual or household head occupations) because some occupations had fewer workers recorded in the data. Although we attempted to include almost all occupations in the data, this caused the model to produce results indicating that almost all occupations had no significant effect on leptospirosis seropositivity due to lower sample sizes in some occupation categories. As a result of reducing the number of occupation variables, we can only compare the risk factors for leptospirosis seropositivity between small occupational categories.



## References

- [1] D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.
- [2] E. A. J. Cook, W. A. de Glanville, L. F. Thomas, S. Kariuki, B. M. d. C. Bronsvoort, and E. M. Fèvre. Risk factors for leptospirosis seropositivity in slaughterhouse workers in western kenya. *Occupational and Environmental Medicine*, 74(5):357–365, 2017.
- [3] M. Dowle and A. Srinivasan. *data.table: Extension of ‘data.frame’*, 2021. R package version 1.14.2.
- [4] N. S. Erler, D. Rizopoulos, and E. M. E. H. Lesaffre. JointAI: Joint analysis and imputation of incomplete data in R. *Journal of Statistical Software*, 100(20):1–56, 2021.
- [5] J. Faraway. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. A Chapman & Hall Book. CRC Press, Taylor & Francis Group, second edition, 2016.
- [6] A. Gelman. Analysis of variance—why it is more important than ever. *The annals of statistics*, 33(1):1–53, 2005.
- [7] A. Gelman, P. Gelman, and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press, 2007.
- [8] A. Gelman and Y.-S. Su. *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models*, 2021. R package version 1.12-2.
- [9] C. Goarant. Leptospirosis: risk factors and management challenges in developing countries. *Research and Reports in Tropical Medicine*, Volume 7:49–62, 2016.
- [10] N. Jewell. *Statistics for Epidemiology*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press, Boca Rotan, 2003.
- [11] A. Kassambara. *ggpubr: ‘ggplot2’ Based Publication Ready Plots*, 2020. R package version 0.4.0.
- [12] C. L. Lau, L. D. Smythe, S. B. Craig, and P. Weinstein. Climate change, flooding, urbanisation and leptospirosis: fuelling the fire? *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 104(10):631–638, 2010.
- [13] T. L. Pedersen. *patchwork: The Composer of Plots*, 2020. R package version 1.1.1.
- [14] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.
- [15] D. D. Sjoberg, K. Whiting, M. Curry, J. A. Lavery, and J. Larmarange. Reproducible summary tables with the gtsummary package. *The R Journal*, 13:570–580, 2021.
- [16] Z. M. P. Soo, N. A. Khan, and R. Siddiqui. Leptospirosis: Increasing importance in developing countries. *Acta Tropica*, 201:105183, 2020.
- [17] N. Tierney, D. Cook, M. McBain, and C. Fay. *nanianr: Data Structures, Summaries, and Visualisations for Missing Data*, 2021. R package version 0.6.1.
- [18] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.
- [19] H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Golem, A. Haye, L. Henr, J. Hest, M. Kuh, T. L. Pederse, E. Mille, S. M. Bach, K. Müll, J. Oo, D. Robins, D. P. Seid, V. Spi, K. Takahas, D. Vaugh, C. Wil, K. W, and H. Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019.
- [20] H. Wickham and J. Bryan. *readxl: Read Excel Files*, 2022. R package version 1.4.0.
- [21] H. Wickham, R. Francois, L. Henry, and K. Muller. *dplyr: A Grammar of Data Manipulation*, 2022. R package version 1.0.9.

- [22] Y. Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2022. R package version 1.39.
- [23] H. Zhu. *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*, 2021. R package version 1.3.4.
- [24] A. Zuur, E. N. Ieno, N. Walker, A. A. Saveiliev, and G. M. Smith. *Mixed Effects Models and Extensions in Ecology with R*. Statistics for Biology and Health. Springer, New York, 2009.

# Appendices

## A Acronyms

We use the following acronyms in the missing data pattern plot in Section 2.1:

- sampleid: id of the aliquote tested generated at biorepository
- result: ELISA test results (positive or negative) of the person sampled
- parent: id of the tube which the aliquote was taken from
- relationshipphh: relationship of the person sampled with household head
- gender: gender of the person sampled
- occupation: occupation of the person sampled
- age: age of the person sampled
- landuse: characterization of sampling site based on land use
- date: date of sampling
- altitude: altitude of the village where the sample was collected
- hhgender: gender of the household head
- hhage: age of the household head
- hhoccup: occupation of the household head
- nmales: number of males in the household sampled
- n females: number of females in the household sampled
- famsize: number of people in the household sampled
- disthosp: distance from household to local hospital
- livestk\_home: livestock ownership (yes or no) at the sampled person household
- genhhid: generated household id
- village: village (anonymised) where the person sampled comes from
- location: location (anonymised) where the person sampled comes from
- constituency: constituency where the sampling was collected

## B R Package

We provide a brief explanation about the R [14] packages that was used in this reports. For reading the data into R, we used `readxl` [20] package. The data cleaning and manipulation process was done using the `dplyr` [21] and `data.table` [3] packages. Plots were created with `arm` [8], `ggplot2` [18], `naniar` [17], `JointAI` [4], `ggpubr` [11] and `patchwork` [13] packages. Binomial GLMM models were fitted using `lme4` [1]. Regression tables were created using `gtsummary` [15], `knitr` [22] and `kableExtra` [23]. Note that `ggplot2` and `dplyr` are available in `tidyverse` [19] package.