Fundamentals of Statistical Analysis using R (FSTAR)

19th April 2022 - 13th May 2022

Project

Instructions

The project covers the topics of data importation and wrangling in R. The project should be submitted to the email address info@cema.africa.

- Kindly submit an RMD file that contains your name.

Introduction

A dataset has provided for you - saved as a csv file and available on the url link https://raw.githubusercontent.com/cema-uonbi/F_staR.github.io/main/Week%204/Data/phone_surveillance.csv

The data is from a longitudinal dataset on surveillance for infectious diseases in animals. The publication can be found here - http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0120761

The dataset provided is a subset of the dataset used in making a publication in-press whose abstract is provided below (in order to provide context and ideas on the analysis possible)

Manuscript title: Mobile phone based surveillance for animal disease in rural communities: implications for detection of zoonoses spillover

Abstract:

Improving the speed of outbreak detection and reporting at the community level is critical in managing the threat of emerging infectious diseases, many of which are zoonotic. The widespread use of mobile phones including in rural areas has presented potentially effective tools for real-time surveillance of infectious diseases. Using longitudinal data from a disease surveillance system implemented in 1500 households in rural Kenya, we test the effectiveness of mobile-phone animal syndromic surveillance by comparing it with routine household animal health surveys, determine the individual and household correlates of its use, and examine the broader implications for surveillance of zoonotic diseases. A total of 20,340 animal and death events were reported from the community through the two surveillance systems, half of which were confirmed as valid disease events. The probability of an event being valid was 2.1 times greater for the phone-based system, compared with the household visits. Illness events were 15 times (95% CI 12.8, 17.1) more likely to be reported through the phone-system compared to routine household visits, but not death events [OR 0.1 (95% CI 0.09, 0.11)]. Disease syndromes with severe presentations were more likely to be reported through the phone system. While controlling for herd and flock sizes owned, phone ownership was not a determinant of using phone-based surveillance system, but the lack of a formal education, and having additional source of income besides farming were associated with decreased likelihood of reporting through the phone system. Our study suggests a phone-based surveillance system will sufficiently detect outbreaks of diseases such as Rift Valley Fever that present with severe clinical signs in animal pop-

ulations but in the absence of additional reporting incentives, it may miss early outbreaks of diseases such as avian influenza that present primarily with mortality.

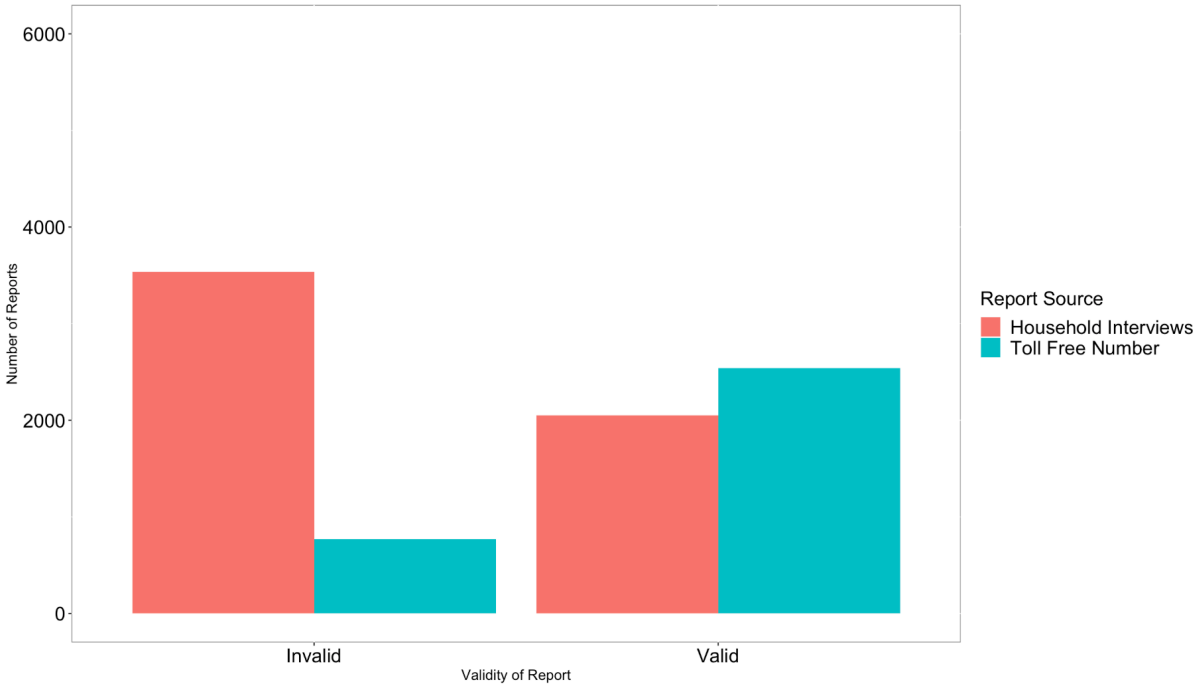
The dataset provided for you contains a 2-year subset of surveillance data used in the publication above and which has the following variables:

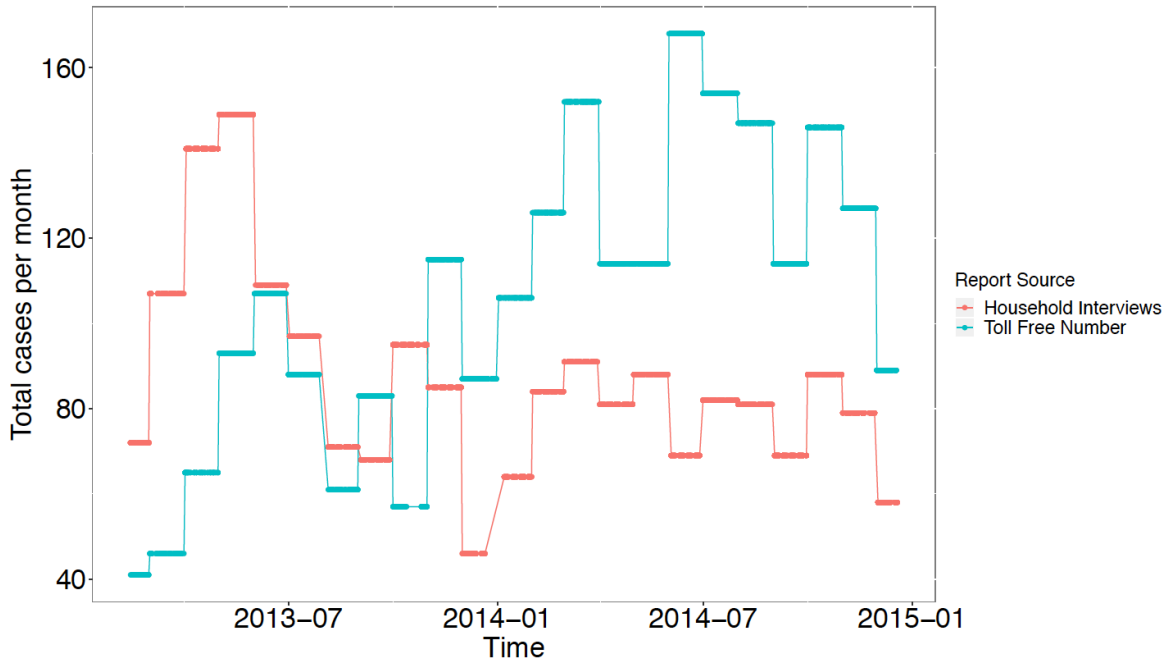
1) IntDate - “InterviewDate” the date when the study household is interviewed
2) HHID - “HouseholdID” which represents a unique ID for each of the households in the study
3) RprtSrc - “ReportSource” - a factor variable representing one of two methods by which study households report cases of illness among their livestock. 1 represents “Household Interviews” and 2 represents “Toll Free Number”
4) Valid - “CaseValid” - is a factor variable indicating whether the case reported by either of the two reporting methods (Household interviews or Toll Free Number) was confirmed as a valid case or not.1 - represents a “Valid” case and 2 represents an “Invalid” case
5) AnimalsDead - variable capturing information if the case reported was a “death case” or not. 1 - represents a “death case”, 0 - “not a death case”, and 99 - “Not applicable” meaning the case was “Invalid”
6) AnimalsSick - variable capturing information if the case reported was an “illness case” or not. 1 - represents an “illness case”, 0 - “not an illness case”, and 99 - “Not applicable” meaning the reported case was “Invalid”
7) Columns/variables number 8 to 12 (CowsDied to ChicksDied) represents a “Yes” - 1 or “No” - 0 if the case reported was a death case for the specific animal species.
8) Columns/variables number 13 to 15 (CowSick to SheepSick) represents a “Yes” - 1 or “No” - 0 if the case reported was an illness case involving the specific animal species.

Project tasks

1. Import the data from the url provided
2. Explore the data to understand the variables and the levels associated with each variable
3. Provide the appropriate summary/descriptive statistics for the variables in the dataset

4 Using the package “ggplot2” - reproduce the two graphics provided (description provided below) and explain what you observe from the graphics.


6000
4000
2000
0
Number of Reports
Invalid
Valid
Validity of Report
Report Source
Household Interviews
Toll Free Number


160
120
80
40
Total cases per month
2013–07
2014–01
2014–07
2015–01
Time
Report Source
Household Interviews
Toll Free Number