cestode toy data1

brian avery

July 7, 2016

Intro

Making some test/toy data for Syd to use to start learning/doing some data analysis in R while she gathers data. It's not the best toy data ever, but it's a start.

Important considerations:

we have different sample collection sites

data was collected over several years

each site should have a mix of infected and uninfected results

Cestode data code book

(this test data is also a nice test of the format of the dataset so for now this should be considered a draft)
in general in a code book, each column corresponds to a variable, the code book has the following info for each column.
this is the format we'll use:

Column label: the name of the varialbe in the dataset

Variable: an explanation of what the variable is/measures

Variable type: numeric, integer, character, logical (Boolean), factor, date, etc.

Allowable values: list of possible values

Comments: anything else

year

Column label: year

Variable: year individual was collected

Variable type: integer

Allowable values: 2010-2016

Comments:

site

Column label: site

Variable: site from whence individual was collected

Variable type: character

Allowable values: blackrock, statemarina, antelopeisland, sprialjetty

Comments:

ID

Column label: ID

Variable: indentifier of individual, a few letters that identify the site

Variable type: character

Allowable values: BR, SM, AI, or SJ followed by a number from 1-200

Comments: BR=blackrock, SM=statemarina, AI=antelopeisland, SJ=sprialjetty. numbers are individual number within that site

infected

Column label: infected

Variable: infection status of individual

Variable type: logical

Allowable values: TRUE, FALSE

Comments: TRUE=infected, FALSE=uninfected. this is of course our best call based on the PCR

Generate the toy data

First we have a lot of numbers etc. to make, then we want to format it into a data frame to make it easy to work with.

```r
# this ensures that the random numbers generated are the same if we re-run the code later
set.seed(84105)

## build each column, the ID col is complicated so it gets built in 2 pieces here and combined later
years <- rep(c(2012, 2013), each = 120)
sites <- rep(c("blackrock", "statemarina", "antelopeisland", "sprialjetty"), each = 30, times = 2)
IDL <- rep(c("BR", "SM", "AI", "SJ"), each = 30, times = 2)
IDno <- rep(c(1:30), 8)

## get some infection data, generated from four differently weighted distributions
## so the different sites might have different outcomes
inf1a <- sample(c(TRUE, FALSE, NA), 30, replace = TRUE, prob = c(.49, .49, 0.02))
inf1b <- sample(c(TRUE, FALSE, NA), 30, replace = TRUE, prob = c(.49, .49, 0.02))
inf1c <- sample(c(TRUE, FALSE, NA), 30, replace = TRUE, prob = c(.49, .49, 0.02))
inf1d <- sample(c(TRUE, FALSE, NA), 30, replace = TRUE, prob = c(.49, .49, 0.02))
inf2 <- sample(c(TRUE, FALSE, NA), 30, replace = TRUE, prob = c(.74, .24, 0.02))
inf3a <- sample(c(TRUE, FALSE, NA), 30, replace = TRUE, prob = c(.44, .54, 0.02))
inf3b <- sample(c(TRUE, FALSE, NA), 30, replace = TRUE, prob = c(.44, .54, 0.02))
inf4 <- sample(c(TRUE, FALSE, NA), 30, replace = TRUE, prob = c(.94, .04, 0.02))

## combine into one vector for the infected column
infected <- c(inf1a, inf1b, inf3a, inf2, inf1c, inf1d, inf3b, inf4)
```

The `dplyr` and `tidyr` packages are really good for working with data frames. Look them up. Here I'm going to use them to format the numbers generated above into a nice, tidy data frame.

```r
# loads the packages so we can use them,
# you need to use install.packages() if you don't have them yet
library(dplyr)
library(tidyr)
```

```
## build the start of the dataframe out of the first few simulated columns of data
cestodetoy <- data.frame(year=years, site=sites, IDL=IDL, IDno=IDno)

## combine the 2 pieces of the ID column into 1 piece and remove the old halves
cestodetoy <- unite(cestodetoy, ID, IDL, IDno, sep = '', remove = TRUE)

## add the infected column
cestodetoy$infected <- infected
```

Now that the entire data set is built, let's look at it.

```
# 3 really good, quick ways to see what the data looks like.
head(cestodetoy)
```

```
##   year      site  ID infected
## 1 2012 blackrock BR1    FALSE
## 2 2012 blackrock BR2     TRUE
## 3 2012 blackrock BR3    FALSE
## 4 2012 blackrock BR4    FALSE
## 5 2012 blackrock BR5     TRUE
## 6 2012 blackrock BR6    FALSE
```

```
str(cestodetoy)
```

```
## 'data.frame':    240 obs. of  4 variables:
##  $ year    : num  2012 2012 2012 2012 2012 ...
##  $ site    : Factor w/ 4 levels "antelopeisland",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ ID      : chr  "BR1" "BR2" "BR3" "BR4" ...
##  $ infected: logi  FALSE TRUE FALSE FALSE TRUE FALSE ...
```

```
summary(cestodetoy)
```

```
##       year                   site         ID              infected
##  Min.   :2012   antelopeisland:60   Length:240         Mode :logical
##  1st Qu.:2012   blackrock     :60   Class :character   FALSE:95
##  Median :2012   sprialjetty   :60   Mode  :character   TRUE :139
##  Mean   :2012   statemarina   :60                      NA's :6
##  3rd Qu.:2013
##  Max.   :2013
```

`dplyr` is your new and best-est friend for summarzing data!
the pipe symbol `%>%` is super helpful in using `dplyr` for summarizing data since you can string together multiple commands. another interesting/weird thing is that `TRUE=1` and `FALSE=0` so you can sum and mean them!

```
# grouped by site (not years) the "na.rm=TRUE" removes NAs if there are any
cestodetoy %>% group_by(site) %>% summarise(sum(infected, na.rm=TRUE))
```

```
## # A tibble: 4 x 2
##             site sum(infected, na.rm = TRUE)
##           <fctr>                       <int>
## 1 antelopeisland                          27
## 2      blackrock                          31
## 3    sprialjetty                          51
## 4    statemarina                          30
```

```
## the mean of infected gives you proportion
cestodetoy %>% group_by(site) %>% summarise(mean(infected, na.rm=TRUE))
```

```
## # A tibble: 4 x 2
##             site mean(infected, na.rm = TRUE)
##           <fctr>                        <dbl>
## 1 antelopeisland                    0.4500000
## 2      blackrock                    0.5535714
## 3    sprialjetty                    0.8793103
## 4    statemarina                    0.5000000
```

```
## grouped by site and then year
cestodetoy %>% group_by(site, year) %>% summarise(sum(infected, na.rm=TRUE))
```

```
## Source: local data frame [8 x 3]
## Groups: site [?]
##
##             site  year sum(infected, na.rm = TRUE)
##           <fctr> <dbl>                       <int>
## 1 antelopeisland  2012                          14
## 2 antelopeisland  2013                          13
## 3      blackrock  2012                          15
## 4      blackrock  2013                          16
## 5    sprialjetty  2012                          22
## 6    sprialjetty  2013                          29
## 7    statemarina  2012                          15
## 8    statemarina  2013                          15
```

```
cestodetoy %>% group_by(site, year) %>% summarise(mean(infected, na.rm=TRUE))
```

```
## Source: local data frame [8 x 3]
## Groups: site [?]
##
##             site  year mean(infected, na.rm = TRUE)
##           <fctr> <dbl>                        <dbl>
## 1 antelopeisland  2012                    0.4666667
## 2 antelopeisland  2013                    0.4333333
## 3      blackrock  2012                    0.5357143
## 4      blackrock  2013                    0.5714286
## 5    sprialjetty  2012                    0.7857143
## 6    sprialjetty  2013                    0.9666667
## 7    statemarina  2012                    0.5000000
## 8    statemarina  2013                    0.5000000
```

```
## write to csv file
## uncomment the next line if you want to write the data to a csv file

# write.csv(cestodetoy, file = "cestode_toy_dataset1.csv", row.names=FALSE)
```

Now you can use ggplot2 to make some plots and look at the data. I'll start you off with one simple graph.

```
# loads the package so we can use it,
# you need to use install.packages() if you don't have it yet
library(ggplot2)

# a simple histogram of the number of infected animals per site
# should really be proportion, but in this fake example all of the sample sizes are the same

# recreate the summarized data from above, but this time assign it to allinfected
allinfected <- cestodetoy %>% group_by(site) %>% summarise(sum(infected, na.rm=TRUE))
```

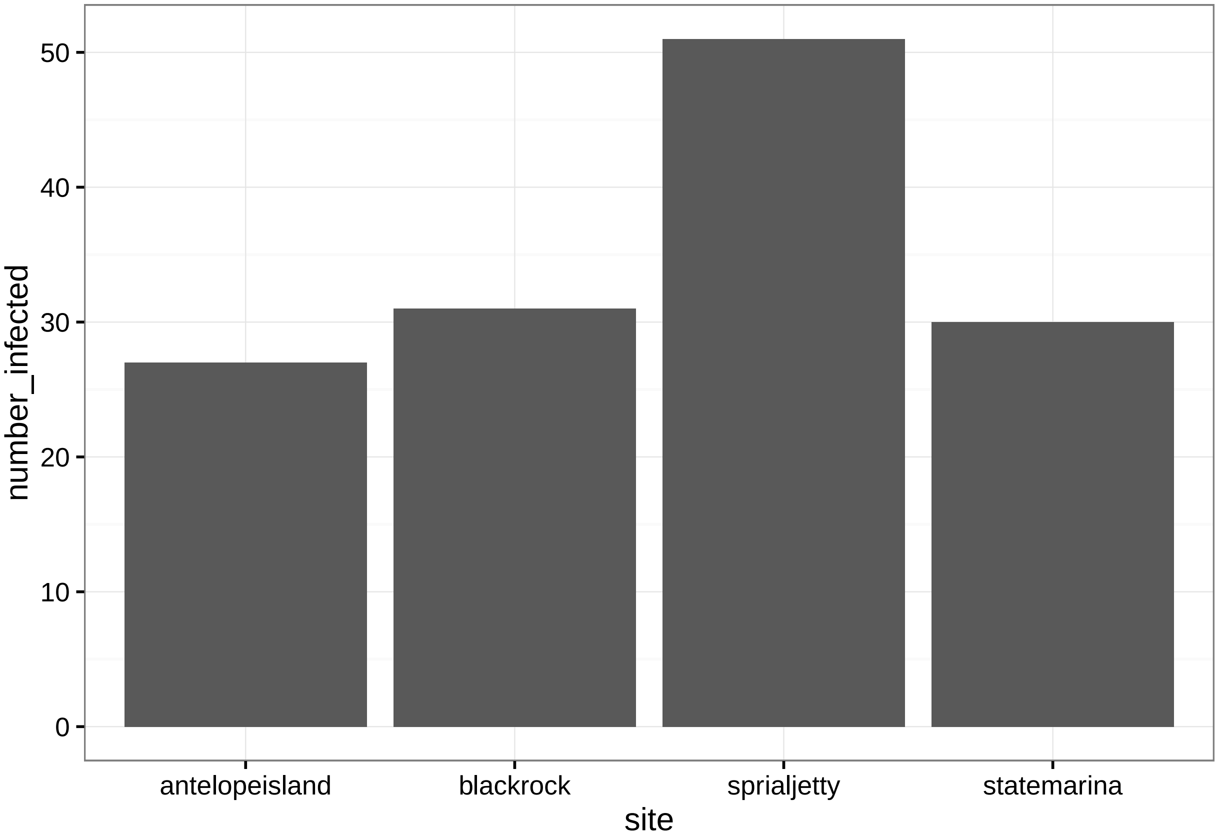
```
allinfected
```

```
## # A tibble: 4 x 2
##              site sum(infected, na.rm = TRUE)
##            <fctr>                       <int>
## 1 antelopeisland                          27
## 2      blackrock                          31
## 3    sprialjetty                          51
## 4    statemarina                          30
```

```
# the title of the second column that contains the sum of the infected has a terrible name
# let's fix that
names(allinfected)[names(allinfected) == "sum(infected, na.rm = TRUE)"] <- "number_infected"
allinfected
```

```
## # A tibble: 4 x 2
##              site number_infected
##            <fctr>           <int>
## 1 antelopeisland              27
## 2      blackrock              31
## 3    sprialjetty              51
## 4    statemarina              30
```

```
# make the histogram
ggplot(data=allinfected, aes(x= site, y=number_infected)) +
  geom_bar(stat="identity") +
  theme_bw()
```



your turn!