rmarkdown_day3

Lea Briard

May 16, 2018

Statistically literate programming with R Markdown

Literate programming is a programming paradigm due to Linux founder Donald Knuth in which natural langauge explanations of a program's logic are interspersed with the code snippets that actually perform the computation. Statistically literate programming applies this paradigm to data analysis. Statistically literate programming is the idea that the thought process of the data analyst can be captured in a report that contains explanation and interpretation, the code used to perform an analysis, and the products of that analysis such as tables of data, quantities (e.g. p-values) and graphs. There are a handful of ways that one can do statistically literate programming with R/RStudio. In this exercise, we will use R/Markdown and knitr. Markdown is a lightweight markup language that allows documents to be rendered in html and other formats (e.g. pdf) with a minimum of special formatting. Knitr is a system for dynamic report generation in R. Both should already be installed on your computer. In general, a project developed with R/Markdown will consist of a markdown document (with extension .Rmd) and the compiled report. You should learn a little more about the functions of R/Markdown and knitr, but first we will engage in a little learning-by-doing.

```
### packages
library(ggplot2)
library(tidyverse)
library(magrittr)

mers <- read.csv('cases.csv')
head(mers)  #to check out the dataset
class(mers$onset)   # dates are factors but we want them as dates. use package 'lubridate'
```

fix some mistakes in the data

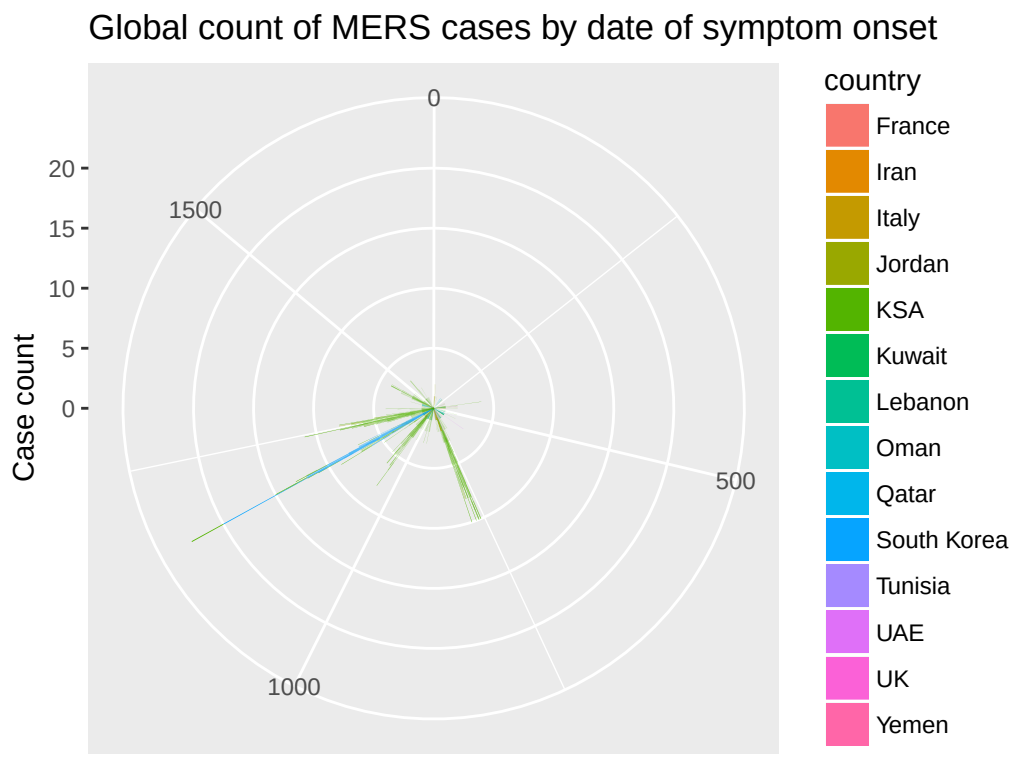
```
mers$hospitalized[890] <- c('2015-02-20') # change the date on row 890
mers <- mers[-471,] # omit row 471 from the dataset (bc it has conflicting dates?)
```

use lubridate to change the columns containing dates from "Factor" to "Dates"

```
library(lubridate)
mers$onset2 <- ymd(mers$onset) # create new sister columns for 'onset' and 'hospitalised' columns
mers$hospitalized2 <- ymd(mers$hospitalized) # and use 'ymd' function to turn them into dates
class(mers$onset2) #check the class of these new colums to see if they have changed from "Factor" to "D
class(mers$hospitalized2)

day0 <- min(na.omit(mers$onset2))
#use na.omit because if there are NAs R will be confused about what the minimum is

mers$epi.day <- as.numeric(mers$onset2 - day0) #creating a new numeric value for the epidemic day for e
```

making plots

```
ggplot(data = mers) +
  geom_bar(mapping=aes(x=epi.day, fill=country), na.rm = TRUE) +
  coord_polar() +
  labs(x='Epidemic day', y='Case count', title= 'Global count of MERS cases by date of symptom onset',
       caption="Data from: https://github.com/rambaut/MERS-Cases/blob/gh-pages/data/cases.csv")
```



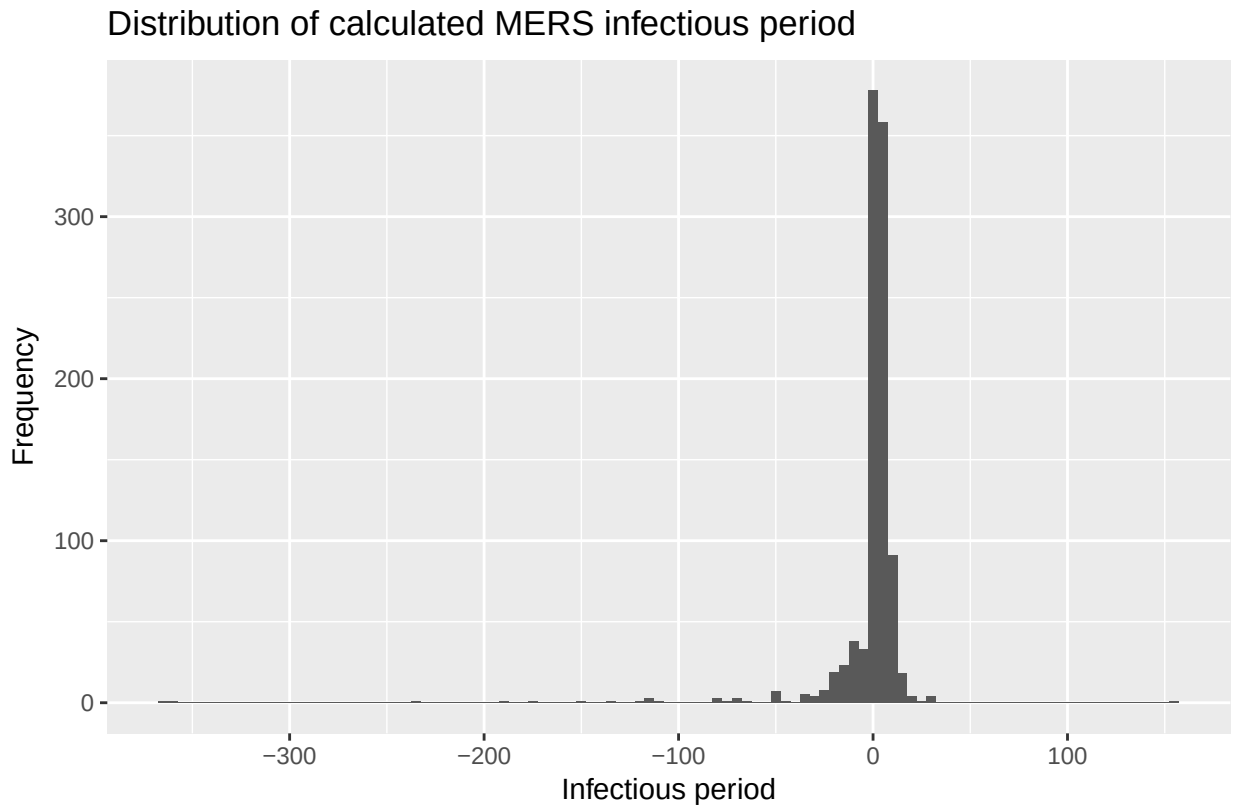
n: https://github.com/rambaut/MERS−Cases/blob/gh−pages/data/cases.csv

univariate plots

```
mers$infectious.period <- mers$hospitalized2-mers$onset2 # calculate "raw" infectious period
class(mers$infectious.period) # these data are class "difftime"
```

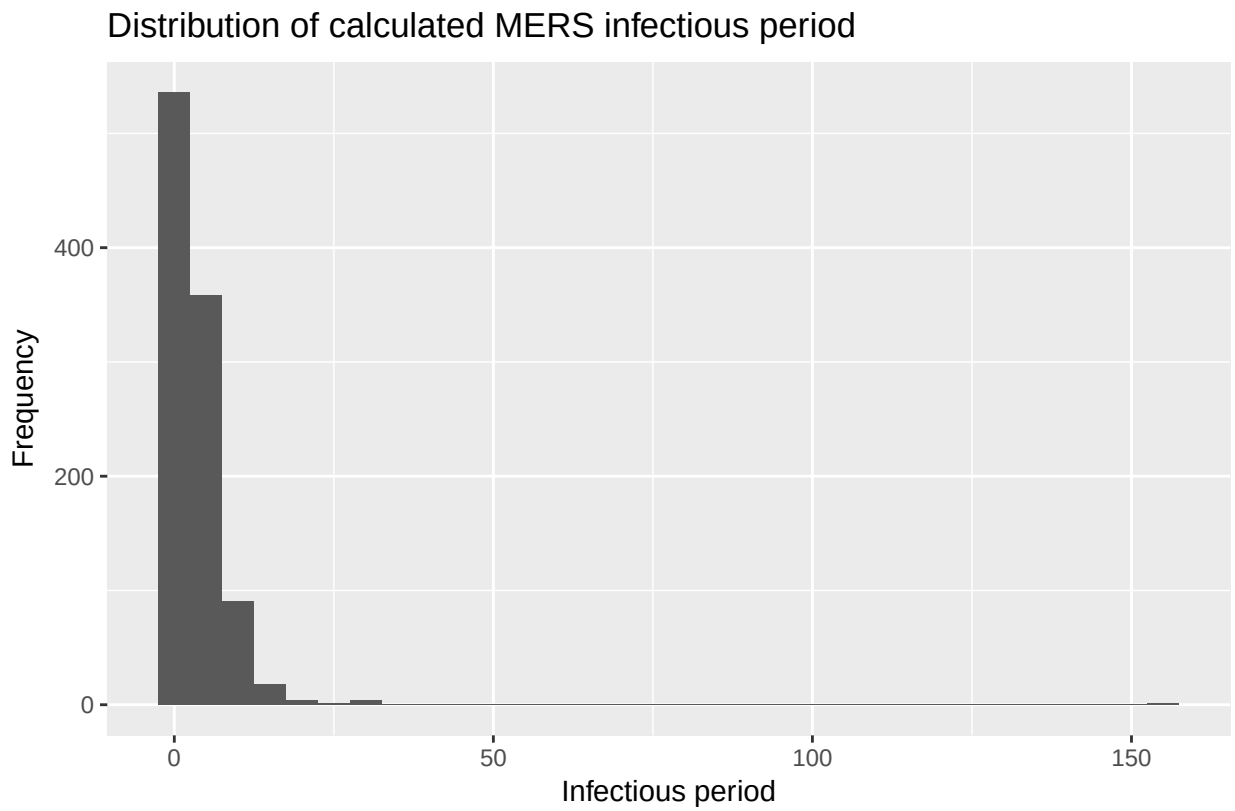
```
## [1] "difftime"
```

```
mers$infectious.period <- as.numeric(mers$infectious.period, units = "days") #convert to days

ggplot(data=mers) +
  geom_histogram(aes(x=infectious.period), na.rm = TRUE, binwidth = 5) +
  labs(x='Infectious period', y='Frequency', title='Distribution of calculated MERS infectious period',
       caption="Data from: https://github.com/rambaut/MERS-Cases/blob/gh-pages/data/cases.csv")
```

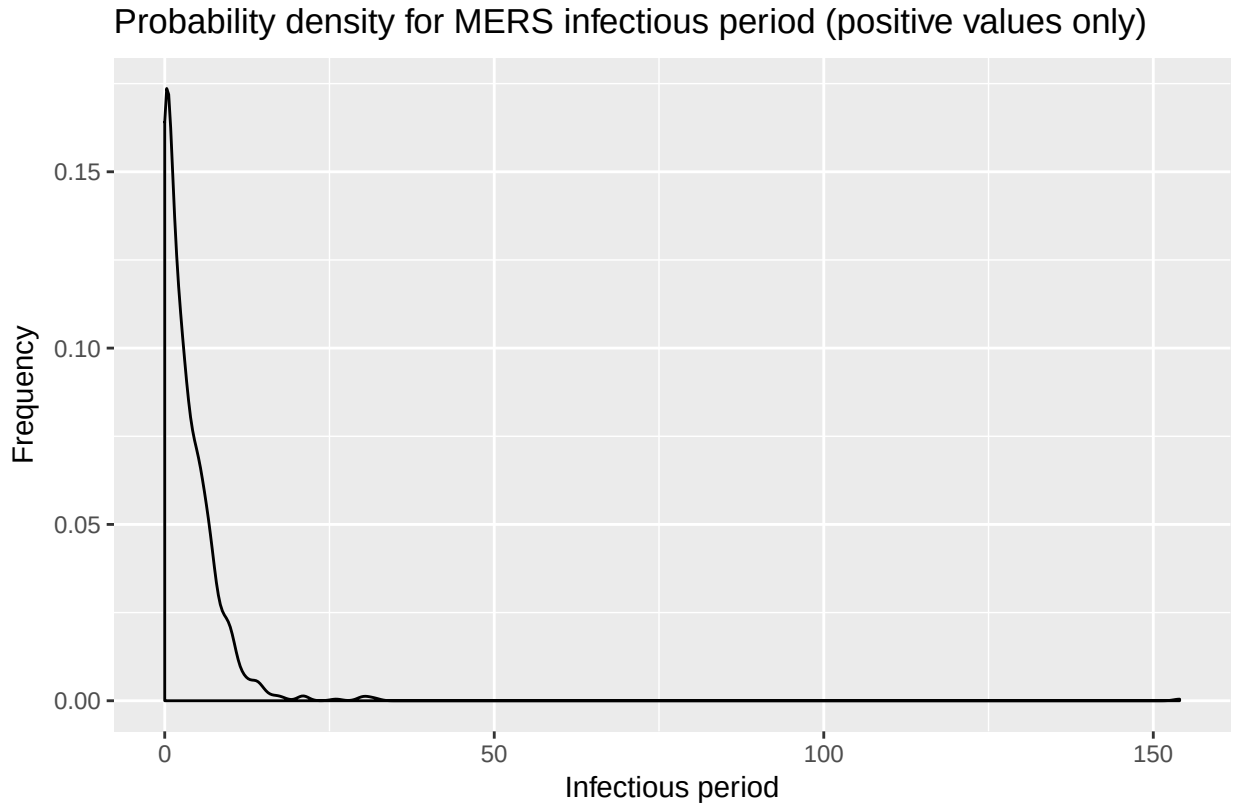



```
#problem: there are negative infectious periods because of nosocomical infections
```

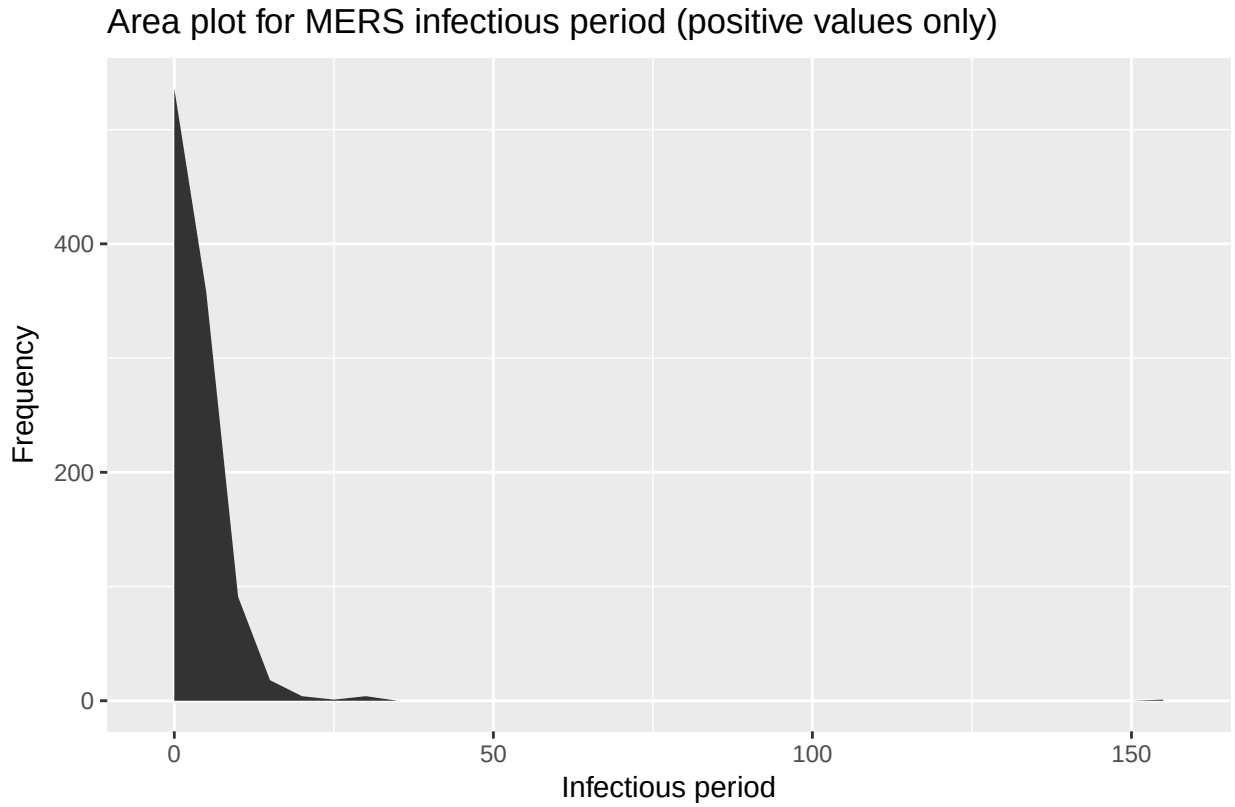
```
mers$infectious.period2 <- ifelse(mers$infectious.period<0, 0, mers$infectious.period)

ggplot(data=mers) +
  geom_histogram(aes(x=infectious.period2), na.rm = TRUE, binwidth = 5) +
  labs(x='Infectious period', y='Frequency', title='Distribution of calculated MERS infectious period',
       caption="Data from: https://github.com/rambaut/MERS-Cases/blob/gh-pages/data/cases.csv")
```



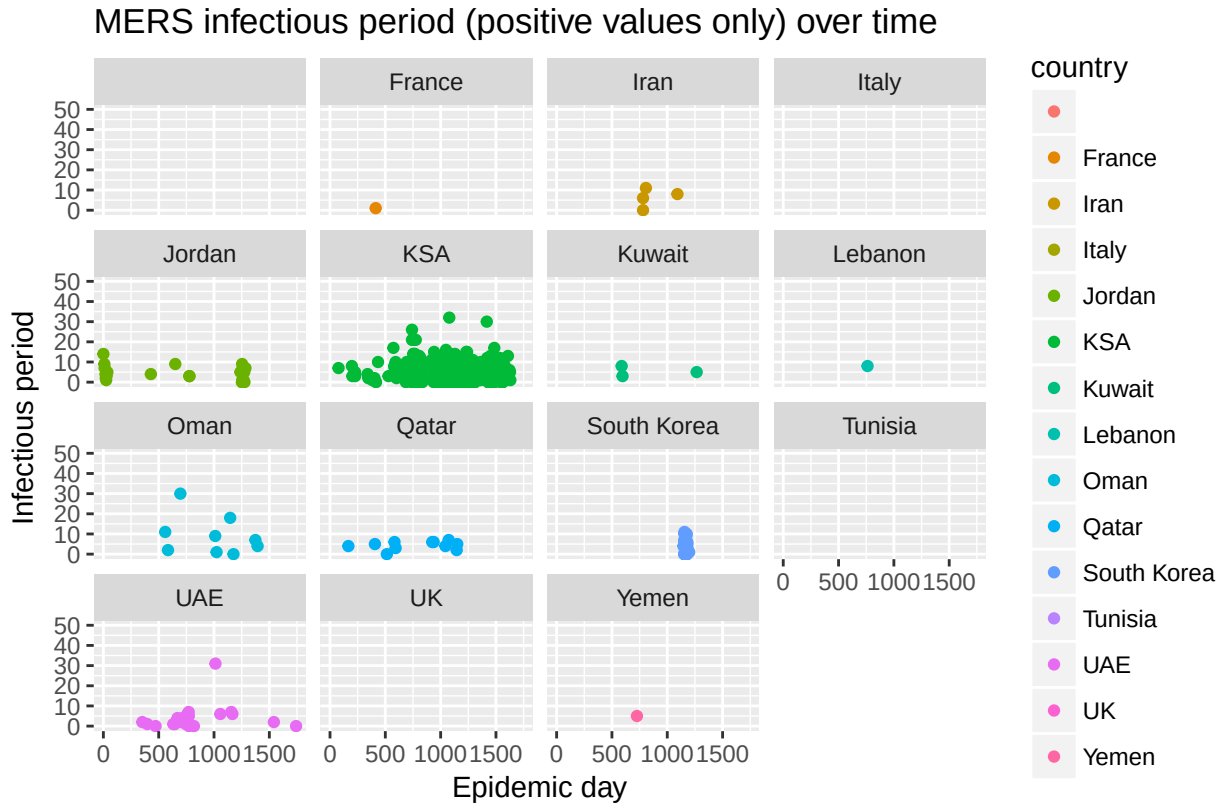

```
ggplot(data=mers) +
  geom_density(mapping=aes(x=infectious.period2), na.rm = TRUE) +
  labs(x='Infectious period', y='Frequency',
       title='Probability density for MERS infectious period (positive values only)',
       caption="Data from: https://github.com/rambaut/MERS-Cases/blob/gh-pages/data/cases.csv")
```




```
ggplot(data=mers) +
  geom_area(stat = 'bin', mapping=aes(x=infectious.period2), na.rm = TRUE, binwidth =5) +
  labs(x='Infectious period', y='Frequency',
       title='Area plot for MERS infectious period (positive values only)',
       caption="Data from: https://github.com/rambaut/MERS-Cases/blob/gh-pages/data/cases.csv")
```




```
#bivariate plots
ggplot(data=mers, mapping=aes(x=epi.day, y=infectious.period2)) +
  geom_point(mapping = aes(color=country), na.rm = TRUE) +
  facet_wrap(~ country) +
  scale_y_continuous(limits = c(0, 50)) +
  labs(x='Epidemic day', y='Infectious period',
       title='MERS infectious period (positive values only) over time',
       caption="Data from: https://github.com/rambaut/MERS-Cases/blob/gh-pages/data/cases.csv")
```



Data from: https://github.com/rambaut/MERS−Cases/blob/gh−pages/data/cases.csv