EPI 569 – Exercise 1: Natural history of infectious diseases

Ben Lopman, Elizabeth Rogawski McQuade - Student: Victoria Ngo

Background

- In December 2009, an outbreak occurred following a restaurant exposure to a foodborne pathogen over many days. More information is in the paper reporting the outbreak, if you are interested, but not necessary that you read the article. https://pubmed.ncbi.nlm.nih.gov/21524343/
- CDC and the local health departments conducted a study of those who fell ill, as well as their household contacts. They found that this outbreak was a result of oysters consumed at the restaurant that were contaminated with norovirus.
- This dataset presents a good opportunity to study natural history, because there is a *known time of exposure* (in hours) for the cases who dined at the restaurant. In addition, there are data on illnesses among the household contacts who did not dine at the restaurant. That way, we can also look at the natural history among household contacts.

The spreadsheet 'NaturalHistoryExercise.csv' contains the data, which you can save from Canvas or import directly into R from the web as follows:

```
data <- read.csv(url("https://raw.githubusercontent.com/blopman/epi569/master/NaturalHistory
view(data)
```

- You can see all the data on times of exposure, onset of symptoms and time of last symptom episode. The **HOUSEHOLD__** variable indicated which household the person was a member; and **HOUSEHOLD_INDEX** indicated if the person got ill at the restaurant. The other fields should be self-explanatory.

- We have calculated the key values for each person for you. For each, there is a value for the linear scale (DAYS) and log scale (LOG DAYS). You can explore the rest of the dataset, but most wont be needed for this exercise.
  - **INDEX_INCUBATION_DAYS** – The time from dining at the restaurant to onset of symptoms in index cases
  - **SERIAL_INTERVAL_DAYS** – The time from onset of illness in the index case to the onset in the household contact
  - **DURATION_DAYS** – The time from the onset of symptoms to the end of illness in all cases.

It is recommended that you use R

- If you use **R**, the following functions will be useful to you:
  - median(data$**VARIABLE_NAME**, na.rm=TRUE) – Median
  - sd(data$**VARIABLE_NAME**, na.rm=TRUE) – Standard deviation
  - log(data$**VARIABLE_NAME**, na.rm=TRUE) – Natural log
  - exp(data$**VARIABLE_NAME**, na.rm=TRUE) – Exponentiate

Now for your work. For this outbreak:

Question 1 (1pt)

- Calculate *the incubation period*:
  - the median of the incubation period
  - the median of the log incubation period
  - the standard deviation of the log incubation period
  - the dispersion, which is exp(sd)

ANSWER

```r
#incubation period = index_incubation_days
median(data$INDEX_INCUBATION_DAYS, na.rm=TRUE) #Median of the incubation period
```

```
[1] 1.79
```

```r
median(data$INDEX_INCUBATION_LOG_DAYS, na.rm=TRUE) #Median of the log incbbation period
```

```
[1] 0.58
```

```r
sd(data$INDEX_INCUBATION_LOG_DAYS, na.rm=TRUE) #Standard deviation of the log incubation days
```

```
[1] 0.4895947
```

```r
sd_index_incubation_log_days <- sd(data$INDEX_INCUBATION_LOG_DAYS, na.rm=TRUE) #creating the

exp(sd_index_incubation_log_days) #Dispersion
```

```
[1] 1.631655
```

- Because incubation periods are log-normally distributed, 66% of cases should fall within the median/exp(sd) and median*exp(sd). Does that appear to be the case?

ANSWER

```r
median_incubation_log_period <- median(data$INDEX_INCUBATION_LOG_DAYS, na.rm=TRUE) #Median of
dispersion <- exp(sd_index_incubation_log_days) #Dispersion

table(is.na(data$INDEX_INCUBATION_LOG_DAYS)) # there are some NA values!
```

```

FALSE  TRUE 
   76   126 
```

```r
lower_bound = (median_incubation_log_period/dispersion)
upper_bound = (median_incubation_log_period*dispersion)

round(100 *
        nrow(data %>% filter(!is.na(INDEX_INCUBATION_LOG_DAYS), INDEX_INCUBATION_LOG_DAYS >=
        nrow(data %>% filter(!is.na(INDEX_INCUBATION_LOG_DAYS)))
      )
```

```
[1] 68
```

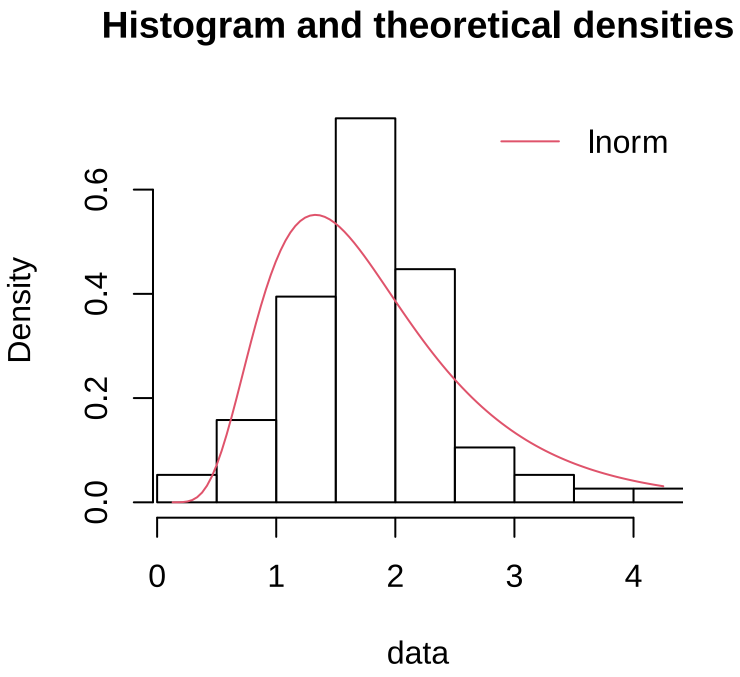
68% of cases fall within the calculated upper and lower range. This percentage is higher than the 66% hypothesized proportion.

- Now, *use the following commands* to plot the distribution of this variable and a fitted log-normal distribution as follows:
  - Create a variable without all the missing data
  - Fit a log normal distribution
  - Look at the parameters of this log normal distribution
  - Plot the log normal distribution along with the data (histogram)

```
incubation<-data$INDEX_INCUBATION_DAYS[!is.na(data$INDEX_INCUBATION_DAYS)]
incubation_fit_ln <- fitdist(incubation, "lnorm")
summary(incubation_fit_ln)
```

```
Fitting of the distribution ' lnorm ' by maximum likelihood
Parameters :
         estimate Std. Error
meanlog 0.5187953 0.05552546
sdlog   0.4840598 0.03926168
Loglikelihood:  -92.12621   AIC:  188.2524   BIC:  192.9139
Correlation matrix:
        meanlog sdlog
meanlog       1     0
sdlog         0     1
```

```
denscomp(incubation_fit_ln)
```



Question 2 (1pt)

- Calculate the same four values (median of the incubation period, the median of the log incubation period, the standard deviation of the log incubation period, and the dispersion) and generate the same plot, this time for *the serial interval.*
  - Did you find a wide distribution of serial intervals (i.e. observed serial intervals of many different lengths)? If so, why do think that could be?

ANSWER

```
#serial interval = serial_interval_days
median(data$SERIAL_INTERVAL_DAYS, na.rm=TRUE) #Median of the serial interval
```

```
[1] 2.42
```

```r
median(data$SERIAL_INTERVAL_LOG_DAYS, na.rm=TRUE) #Median of the log serial interval
```

```
[1] 0.9
```

```r
sd(data$SERIAL_INTERVAL_LOG_DAYS, na.rm=TRUE) #Standard deviation of the log serial interval
```

```
[1] 0.7061295
```

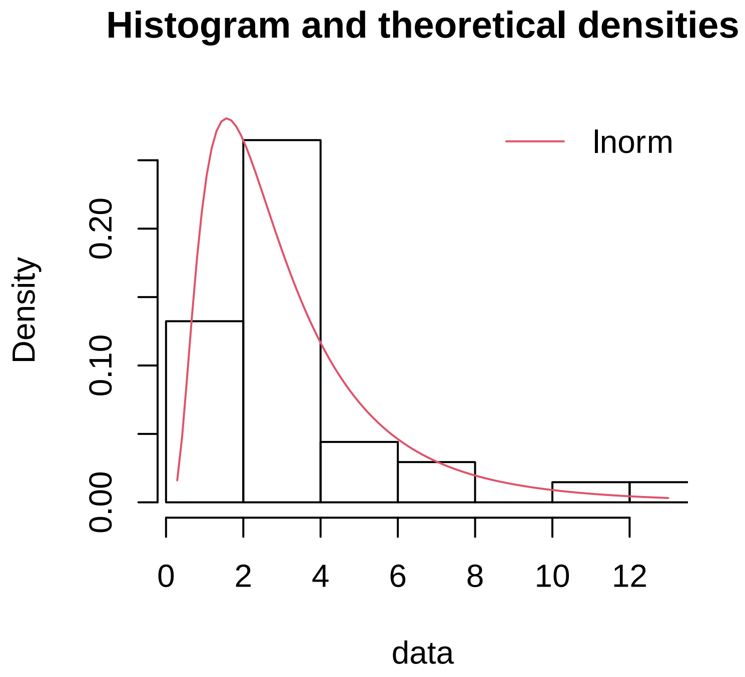
```r
sd_index_incubation_log_days <- sd(data$SERIAL_INTERVAL_LOG_DAYS, na.rm=TRUE) #creating the s
exp(sd_index_incubation_log_days) #Dispersion
```

```
[1] 2.026134
```

```r
serial_int<-data$SERIAL_INTERVAL_DAYS[!is.na(data$SERIAL_INTERVAL_DAYS)]
serial_int_fit_ln <- fitdist(serial_int, "lnorm")
summary(serial_int_fit_ln)
```

```
Fitting of the distribution ' lnorm ' by maximum likelihood
Parameters :
         estimate Std. Error
meanlog 0.9490885 0.12105946
sdlog   0.7058919 0.08560119
Loglikelihood:  -68.67095   AIC:  141.3419   BIC:  144.3946
Correlation matrix:
        meanlog sdlog
meanlog       1     0
sdlog         0     1
```

```r
denscomp(serial_int_fit_ln)
```



Yes, there was a wide distribution of serial intervals due to having a high sdlog value.

Question 3 (1pt)

- Finally, calculate the same four values (median of the incubation period, the median of the log incubation period, the standard deviation of the log incubation period, and the dispersion) and generate the same plot, this time for *the duration of illness*.

ANSWER

```
#duration = duration_days
median(data$DURATION_DAYS, na.rm=TRUE) #Median of the duration
```

```
[1] 0.75
```

```
median(data$DURATION_LOG_DAYS, na.rm=TRUE) #Median of the log duration
```

```
[1] -0.26
```

```
sd(data$DURATION_LOG_DAYS, na.rm=TRUE) #Standard deviation of the duration
```

```
[1] 1.345742
```

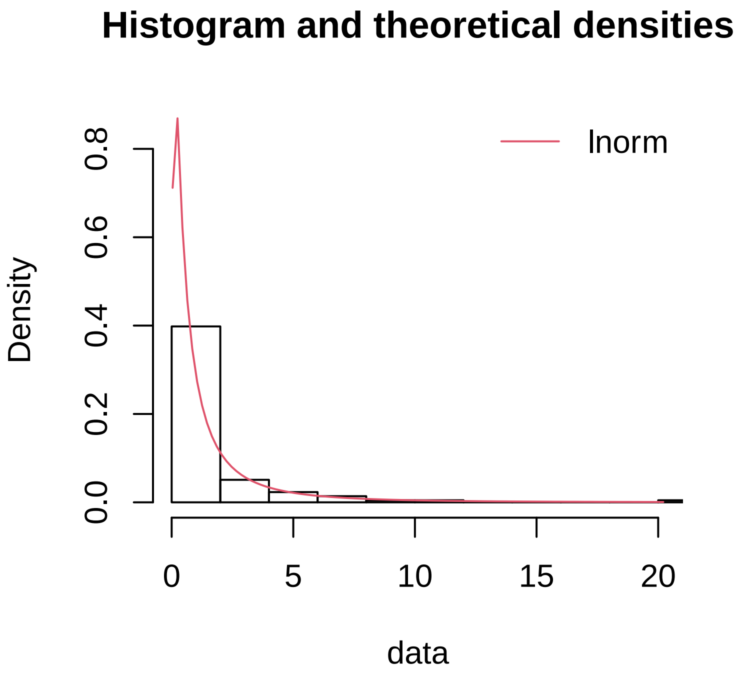
```
sd_index_incubation_log_days <- sd(data$DURATION_LOG_DAYS, na.rm=TRUE) #creating the sd of du
exp(sd_index_incubation_log_days) #Dispersion
```

```
[1] 3.841036
```

```
duration<-data$DURATION_DAYS[!is.na(data$DURATION_DAYS)]
duration<-duration[duration > 0]
duration_fit_ln <- fitdist(duration, "lnorm")
summary(duration_fit_ln)
```

```
Fitting of the distribution ' lnorm ' by maximum likelihood
Parameters :
         estimate Std. Error
meanlog -0.303703 0.12961209
sdlog    1.346968 0.09164936
Loglikelihood:  -152.6139   AIC:  309.2279   BIC:  314.5921
Correlation matrix:
        meanlog sdlog
meanlog       1     0
sdlog         0     1
```

```
denscomp(duration_fit_ln)
```


Histogram and theoretical densities

Question 4 (1pt)

- Are the values you calculated for (1-3) consistent with norovirus as the cause of the outbreak? Research online for the natural history parameters for norovirus to justify your answer.

ANSWER

In questions 1 -3, the median incubation period, serial interval, and duration of illness were calculated. According to Robilotti et al. (2015), the Kaplan;s criteria for a norovirus outbreak is a mean (or median) illness duration of 12-60 hours and a mean (or median) incubation period of 24 to 48 hours. Both of which align with this investigation's findings of a median duration of 0.75 days and a median incubation period of 1.79 days.

Robilotti, Elizabeth, Stan Deresinski, and Benjamin A. Pinsky. "Norovirus." *Clinical microbiology reviews* 28.1 (2015): 134-164.

Question 5 (1pt)

- Calculate the secondary attack rate among household contacts.
  - *(number ill /number exposed)*

ANSWER

```r
#creating objects
number_ill <- sum(data$ILL == "Y", na.rm = TRUE)

number_exposed <- nrow(data)

# Calculate the secondary attack rate
secondary_attack_rate <- number_ill / number_exposed

# Print result
secondary_attack_rate*100
```

```
[1] 54.45545
```

Question 6 [Extra Credit] (0.5pt)

- Calculate and plot the distribution of secondary attack rates by household.

ANSWER