Statistics 2: Computer Practical 1

Jake Ireland (1908320) Isaac Rawcliffe (1603871)

1 Covid-19 incubation period

Statistics is often used in situations where a complicated phenomenon is not understood from first principles, but there is a need to identify some key characteristics. For example, substantial portions of our understanding of the social sciences and medicine is based statistical experiments and analysis.

We consider here an attempt to approximate the distribution of the incubation period for Covid-19, i.e. the time between infection and when symptoms appear. This is of public health interest, as understanding the incubation period can assist in defining isolation periods that effectively manage the competing interests of allowing a return to normal life and reducing the probability of further infections.

This is a reasonable example of the point made above: although an "initial genome" was published by researchers in early January, there is no way to "read off" from this important macroscopic quantities such as the incubation period or epidemiological quantities such as the rate of reproduction (which is also heavily influenced by social behaviour and environmental characteristics).

The modified data contains 25 observations of incubation times (in days).

2 Data

The data we analyze here is a modified version of the data used in the article

> Backer, Jantien A., Don Klinkenberg, and Jacco Wallinga. Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020. Eurosurveillance 25, no. 5 (2020)

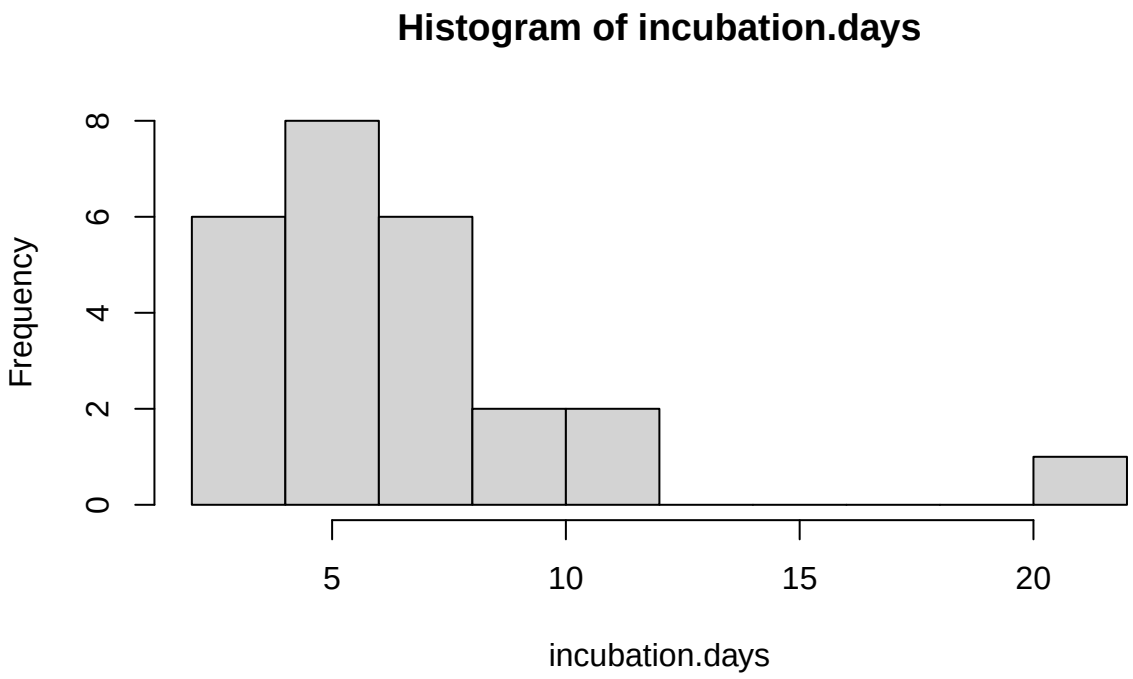
which was accepted for publication on February 6, 2020. The actual data used would require a more sophisticated model to fit than we can use for Statistics 2 (see the Epilogue below).

We can read in the data as follows (make sure it is in the same directory as your Rmd file).

```
covid.incubation <- read.csv("covid-incubation.csv")
```

You can inspect the data by clicking on the data frame in the Environment pane in RStudio. You can see that it contains some demographic information and a description of each patient, with 25 patients in total. For us, the only important data is the `incubation.days` column, the incubation period in days for each patient. We can extract this data and plot a simple histogram to visualize the spread of the values.

```
incubation.days <- covid.incubation$incubation.days
hist(incubation.days,breaks = 10)
```



3 Log-normal incubation period model

One possible model for the incubation period is that it follows a log-normal distribution, i.e.

$$Y = e^X, \qquad X \sim N(\mu, \sigma^2),$$

where $\theta = (\mu, \sigma^2)$ is the statistical parameter. Sartwell (1950) proposed this model after assessing fit for several infectious diseases.

You should just ignore the fact that the incubation period as measured takes discrete values. There is an art to statistical modelling!

4 Questions

Question 1. [2 marks] Derive the maximum likelihood estimators of μ and σ^2, and report the estimate for this dataset.

I suggest that you write a function `ml.estimate` that takes as input some data and returns a vector containing the estimated values of μ and σ^2. For example, something like this:

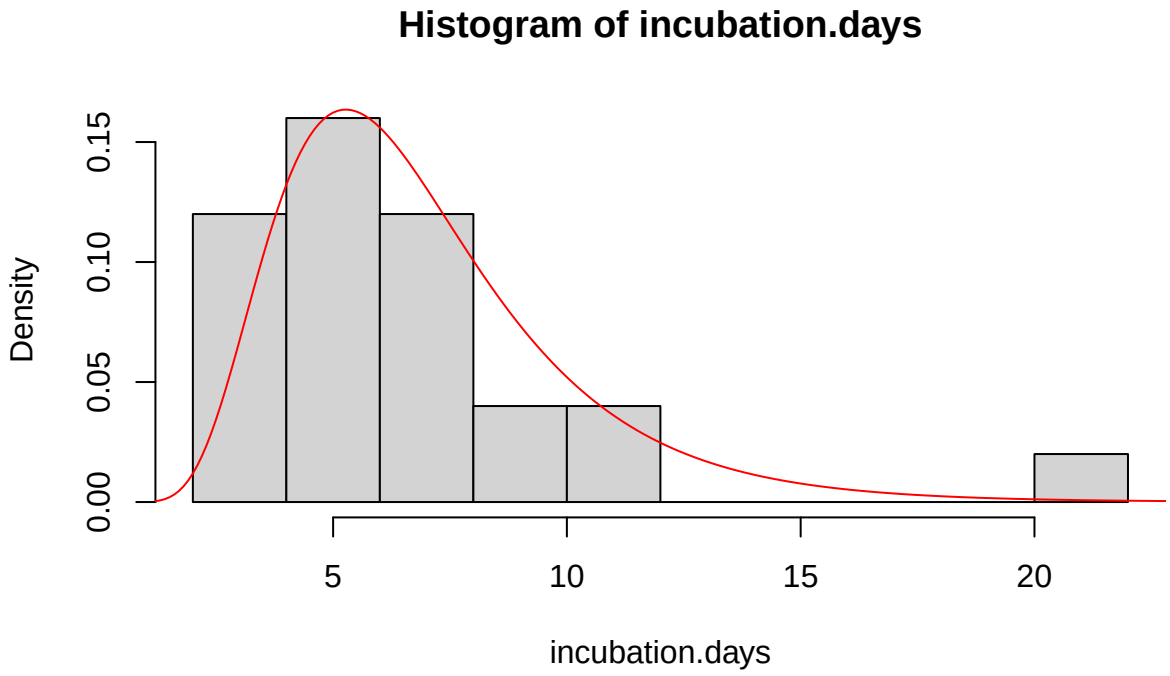
— **Solution** —

```
ml.estimate <- function(ys) {
  mu.hat <- (1/length(ys))*sum(log(ys))
  sigmaSq.hat <- (1/length(ys))*sum((log(ys)-mu.hat)^2)
  c(mu.hat, sigmaSq.hat)
}
```

— **End of Solution** —

If you have written your code correctly, you can visualize the fitted distribution's PDF alongside the histogram of the data.

```
theta.ml <- ml.estimate(incubation.days)
hist(incubation.days,breaks=10,freq=FALSE)
vs <- seq(0,30,0.1)
lines(vs,dlnorm(vs,meanlog=theta.ml[1],sdlog=sqrt(theta.ml[2])),col="red")
```



Question 2. [2 marks] Derive the method of moments estimators of μ and σ^2, and report the estimate for this dataset.

I suggest that you write a function `mom.estimate` that takes as input some data and returns a vector containing the estimated values of μ and σ^2. For example, something like this:

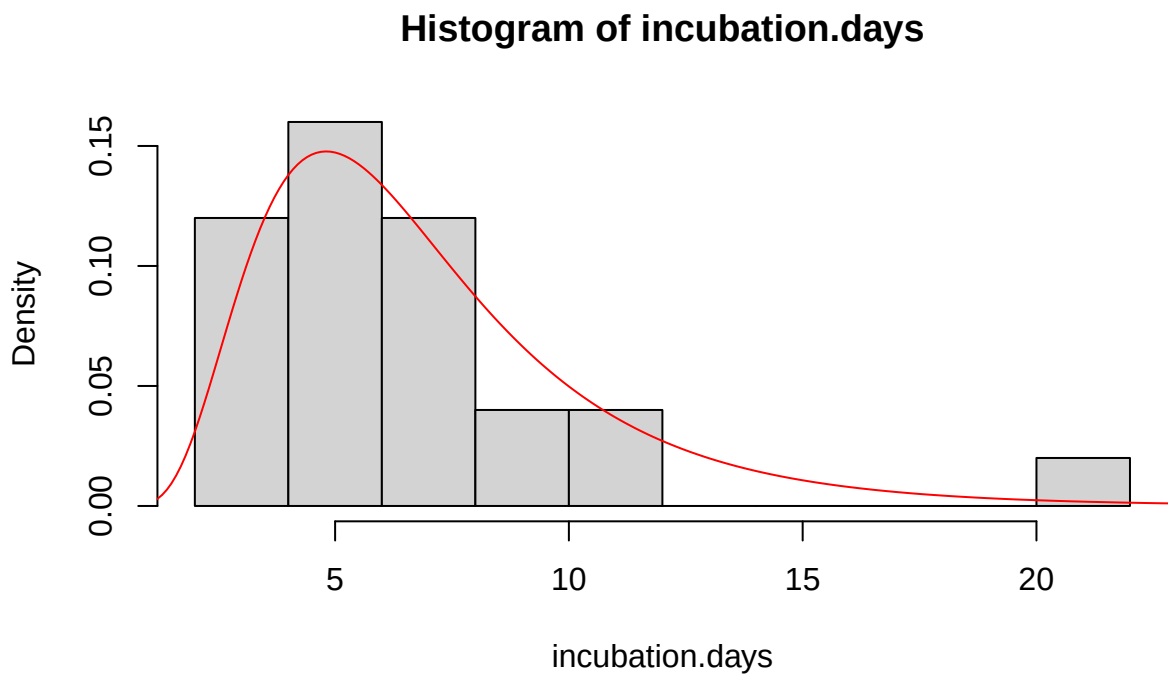
— **Solution** —

```
mom.estimate <- function(ys) {
  sample.mean = mean(ys)
  sample.variance = var(ys)
  mu.hat <- log((sample.mean^2) / sqrt(sample.variance + sample.mean^2))
  sigmaSq.hat <- log((sample.variance / (sample.mean^2)) + 1)
  c(mu.hat, sigmaSq.hat)
}
```

— **End of Solution** —

Again, you can visualize the fit.

```r
theta.mom <- mom.estimate(incubation.days)
hist(incubation.days,breaks=10,freq=FALSE)
vs <- seq(0,30,0.1)
lines(vs,dlnorm(vs,meanlog=theta.mom[1],sdlog=sqrt(theta.mom[2])),col="red")
```



Question 3. [2 marks] Using simulations, compare empirically the mean-squared errors of the maximum likelihood estimators and the method of moments estimators when $\theta = (1.8, 0.2)$.

— **Solution** —

```r
B <- 1000 # number of samples
n <- 100 # size of sample
theta <- c(1.8, 0.2)

simulate <- function(n, B, theta, estimator) {
  # generate nB random numbers. Note rlnorm takes a stdev not a variance.
  big.sample <- rlnorm(n*B, theta[1], sqrt(theta[2]))
  # arranged random numbers in a Bxn matrix, each row represents a different sample
  samples <- matrix(big.sample, nrow=B)

  # apply the given estimator function to each row and replace with the estimates
  # Note output is 2xB matrix, first row are mu estimates, second is sigmaSq estimates
  estimates <- apply(samples, 1, estimator)

  # calculate MSEs
  mu_MSE <- mean((estimates[1,] - theta[1])^2)
  sigmaSq_MSE <- mean((estimates[2,] - theta[2])^2)

  c(mu_MSE, sigmaSq_MSE)
```

```
}

ml_mse <- simulate(n, B, theta, ml.estimate)
mom_mse <- simulate(n, B, theta, mom.estimate)

cat("MSE for ML estimate of mu = ", ml_mse[1])
```

```
## MSE for ML estimate of mu =  0.001905652
```

```
cat("MSE for ML estimate of sigmaSq = ", ml_mse[2])
```

```
## MSE for ML estimate of sigmaSq =  0.0007959919
```

```
cat("MSE for MOM estimate of mu = ", mom_mse[1])
```

```
## MSE for MOM estimate of mu =  0.002048484
```

```
cat("MSE for MOM estimate of sigmaSq = ", mom_mse[2])
```

```
## MSE for MOM estimate of sigmaSq =  0.00134922
```

— **End of Solution** —

Question 4. [1 mark] In light of Q3, comment on the model obtained by observing $X = \log(Y)$ rather than Y.

— **Solution** —

If $X = \log(Y)$ then we have $X = \log(Y) \sim N(\mu, \sigma^2)$ so X is distributed normally and as the natural logarithm is a bijection and we have invariance under bijective reparameterisation, the estimates of μ and σ^2 will still hold with the same MSE.

— **End of Solution** —

Question 5. [2 marks] It is of interest to determine the probability that the incubation period exceeds 7, 10 and 14 days, since asking people to isolate for this number of days could be a simple public health message. What are the maximum likelihood estimates of these probabilities?

— **Solution** —

```
theta.mom <- mom.estimate(incubation.days)

N <- 1000
Y <- rlnorm(n*B, theta[1], sqrt(theta[2]))
Fy <- function(y) mean(Y <= y)

cat("P(incubation exceeds 7 days) = ", 1-Fy(7))
```

```
## P(incubation exceeds 7 days) =  0.37293
```

```r
cat("P(incubation exceeds 10 days) = ", 1-Fy(10))
```

```
## P(incubation exceeds 10 days) =  0.13186
```

```r
cat("P(incubation exceeds 14 days) = ", 1-Fy(14))
```

```
## P(incubation exceeds 14 days) =  0.03047
```

— **End of Solution** —

Question 6. [1 mark] Assume that you did not know how to derive the maximum likelihood estimator for this model. Use the `optim` function to find and report the maximizer of the log-likelihood.

— **Solution** —

```r
log.liklihood <- function(ys, theta) {
  n <- length(ys)
  mu <- theta[1]
  sSq <- theta[2]

  return(-(n/2)*log(2*pi) - (n/2)*log(sSq) - (1/(2*sSq))*sum((log(ys) - mu)^2) - sum(log(ys)))
}

minus.log.liklihood <- function(theta) return(-log.liklihood(incubation.days, theta))

# Setting the lower bound of sigmaSq to 0 gave error due to
# the log-liklihood function being non-finite
optim.out <- optim(c(0,1), minus.log.liklihood,
                   method="L-BFGS-B", lower=c(-Inf,0.001),
                   upper=c(Inf, Inf))
optim.out$par
```

```
## [1] 1.8412187 0.1791281
```

— **End of Solution** —

5 Epilogue

5.1 Public health consequences

Question 5 and its solution are the most relevant from a public health perspective. In particular, the probability of incubation period exceeding a particular number of days allows one to test the hypothesis that a person has indeed been infected. It should not come as a surprise, therefore, that NHS guidelines after exposure to Covid-19 were to self-isolate for 14 days (unless symptoms appear, in which case one was told to self-isolate for longer).

One should be careful not to make logical errors in interpreting the probabilities above. In particular, the probability that the incubation exceeds 10 days is the probability that someone who is infected has no symptoms at the 10 day mark. But it is **not** the probability that they are infected if they have no symptoms at the 10 day mark: whether someone is infected is not a random variable in this model.

Finally, one should also bear in mind that there are asymptomatic cases of Covid-19, and this is not necessarily reflected in the model used here.

5.2 Actual data

This is a simple analysis of (modified) data that was used early on in the Covid-19 pandemic to infer the incubation period of the virus. Although the data was modified, the ML estimates are not too different to the estimates produced by the original study.

The specific modification of the data is that we pretend that we have observations of the true incubation period for each of the patients. Such observations are, however, not possible to record. In the original dataset the date of symptom onset is recorded but the date of infection is not, since it is not known when a particular patient was infected. Instead, an exposure window of dates was recorded. Part of the more sophisticated statistical analysis involves modelling the date of infection of each patient as an unobserved random variable. Using such a model for this computer practical would not have been appropriate.

The dataset used for this practical is also smaller, as the method used to "impute" the date of infection data involved taking the midpoint of the exposure window, and several of the exposure windows were open-ended (e.g. people who were in Wuhan long before the outbreak began or whose dates of travel were not known). To simplify the analysis, patients with open-ended exposure windows were ignored.

Obviously, this means that the data used here are to some extent "fake". But hopefully you can see the potential impact of being able to infer a distribution for the incubation period based on a small amount of high-quality data.