University of Edinburgh, School of Mathematics
Incomplete Data Analysis, 2020/2021
Workshop 3 – Solutions

Vanda Inácio

1.

(a) We have

$$S(y)=\int_y^\infty \theta e^{-\theta u}\mathrm{d}u=[-e^{-\theta u}]_y^\infty=e^{-\theta y}.$$

Note that $F(y)+S(y)=1$, where $F(y)=\Pr(Y\leq y)$ is the cumulative distribution function.

(b) The contribution to the likelihood from a non-censored observation is $f(y;\theta)$. For the censored observations, all we know is that $Y>C$ and so its contribution to the likelihood is $\Pr(Y>C;\theta)=S(C;\theta)$. Since all ovservations are assumed independent, we can therefore write the likelihood as

$$\begin{aligned}L(\theta)&=\prod_{i=1}^n\left\{[f(y_i;\theta)]^{r_i}[S(C;\theta)]^{1-r_i}\right\}\\&=\prod_{i=1}^n\left\{[\theta e^{-\theta y_i}]^{r_i}[e^{-\theta C}]^{1-r_i}\right\}\\&=\theta^{\sum_{i=1}^n r_i}e^{-\theta\sum_{i=1}^n y_ir_i}e^{-\theta\sum_{i=1}^n C(1-r_i)}\\&=\theta^{\sum_{i=1}^n r_i}e^{-\theta\sum_{i=1}^n y_ir_i+C(1-r_i)},\end{aligned}$$

and noting that $x_i=y_ir_i+C(1-r_i)$, we have

$$L(\theta)=\theta^{\sum_{i=1}^n r_i}e^{-\theta\sum_{i=1}^n x_i}.$$

Note that equivalently we could have also wrote

$$\begin{aligned}L(\theta)&=\prod_{i=1}^n\left\{[f(x_i;\theta)]^{r_i}[S(x_i;\theta)]^{1-r_i}\right\}\\&=\theta^{\sum_{i=1}^n r_i}e^{-\theta\sum_{i=1}^n x_i}.\end{aligned}$$

The corresponding loglikelihood is

$$\log L(\theta)=\log\theta\sum_{i=1}^n r_i-\theta\sum_{i=1}^n x_i.$$

We thus have

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\log L(\theta)=\frac{\sum_{i=1}^n r_i}{\theta}-\sum_{i=1}^n x_i,$$

leading to

$$\widehat{\theta}_{\text{MLE}}=\frac{\sum_{i=1}^n R_i}{\sum_{i=1}^n X_i}.$$

(c) We have

$$\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log L(\theta) = -\frac{\sum_{i=1}^n r_i}{\theta^2}.$$

The expected information is

$$I(\theta) = -E\left(-\frac{\sum_{i=1}^n R_i}{\theta^2}\right) = \frac{1}{\theta^2}nE(R).$$

Note that R is a binary random variable and so

$$\begin{aligned} E(R) &= 1 \times \Pr(R = 1) + 0 \times \Pr(R = 0) \\ &= \Pr(R = 1) \\ &= \Pr(Y \leq C) \\ &= F(C; \theta) \\ &= 1 - e^{-\theta C}. \end{aligned}$$

Therefore,

$$I(\theta) = \frac{n(1 - e^{-\theta C})}{\theta^2}.$$

Let us discuss these results to emphasize the meaning of information.

- Suppose that the censoring point C is very large (e.g., in the context of a medical study/trial, we observe the subjects (or equivalently, we run the study) for a very long time). Then, as $C \to \infty$

  $$I(\theta) \to \frac{n}{\theta^2}.$$

  Note that $\frac{n}{\theta^2}$ is the expected Fisher information for an uncensored exponential sample. That is, for large C we do not lose much *information* compared with the case without censoring.

- Conversely, suppose that C is very small (again, in the context of a medical study, the trial/study is only run for a very short time, possibly due to cost reasons). For $C \to 0$, we have

  $$I(\theta) \to 0.$$

  That is, for very small C, the observations contain little or no *information* about the parameter θ.

Aside comment: In this case, it only appeared $E(R)$, but for some other distributions, it might be the case that we also have to compute $E(X)$. If that is the case, we need to remember that X is conditionally defined, and therefore,

$$E(X) = E(Y \mid Y \leq C)\Pr(Y \leq C) + C\Pr(Y > C).$$

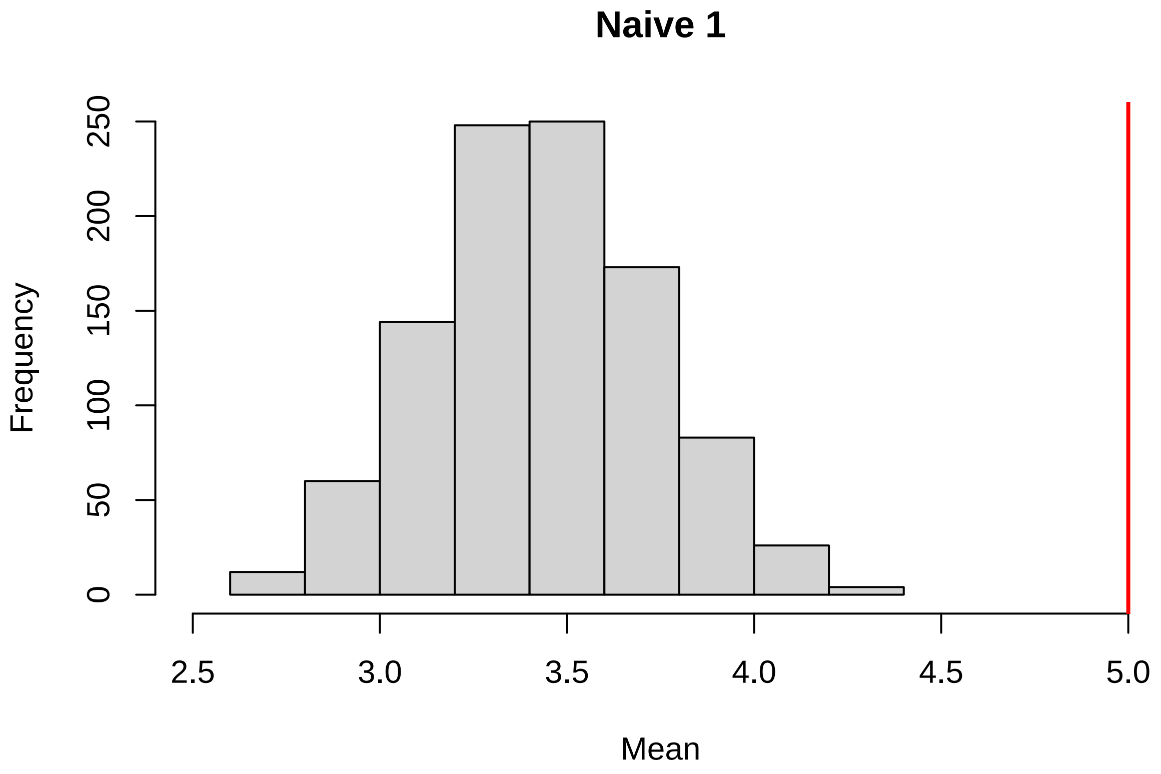
Further note that $E(Y \mid Y \leq C) = \frac{1}{F(C)}\int_0^C yf(y;\theta)\mathrm{d}y$.

(d) MNAR. True value is not observed (i.e., is censored) if the value itself is above some known threshold (C in this case), thus MNAR. We could, however, easily conduct inference because in this case the missingness mechanism is known, in the sense that the censoring point is known.

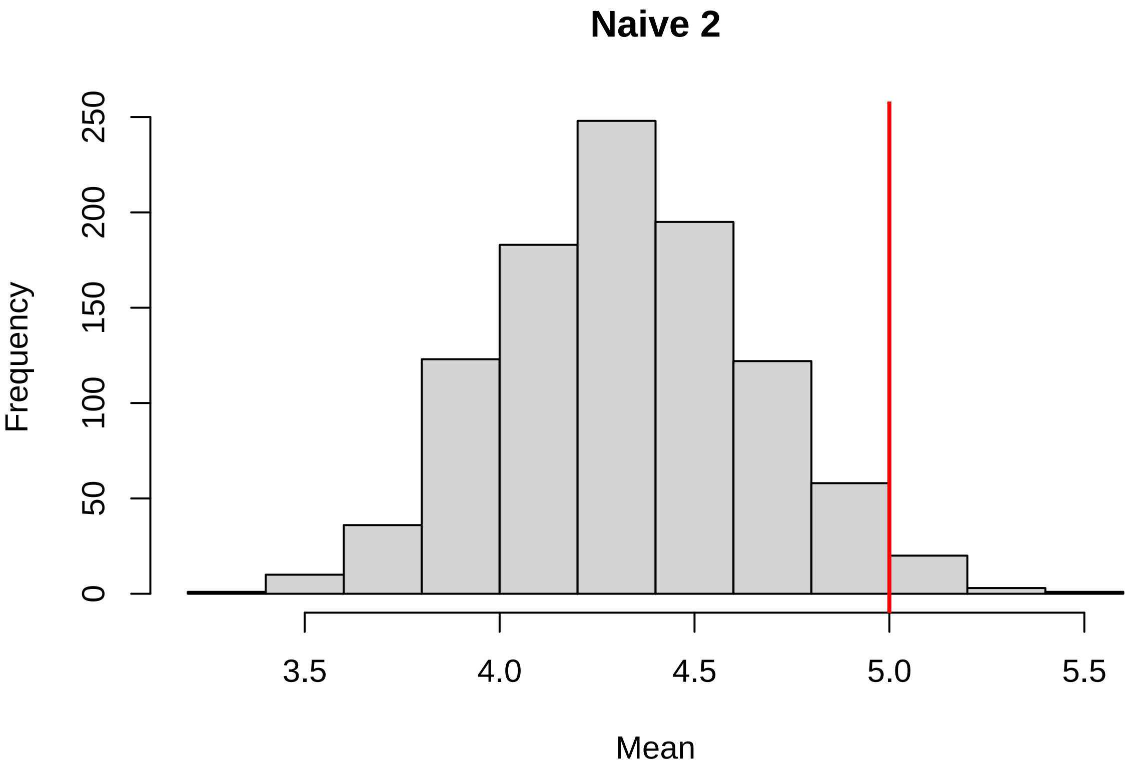
(e)

```
n <- 100
nsim <- 1000
theta <- 1/5

# uncensored samples
y <- matrix(0, nrow = n, ncol = nsim)
set.seed(1)
for(l in 1:nsim){
```

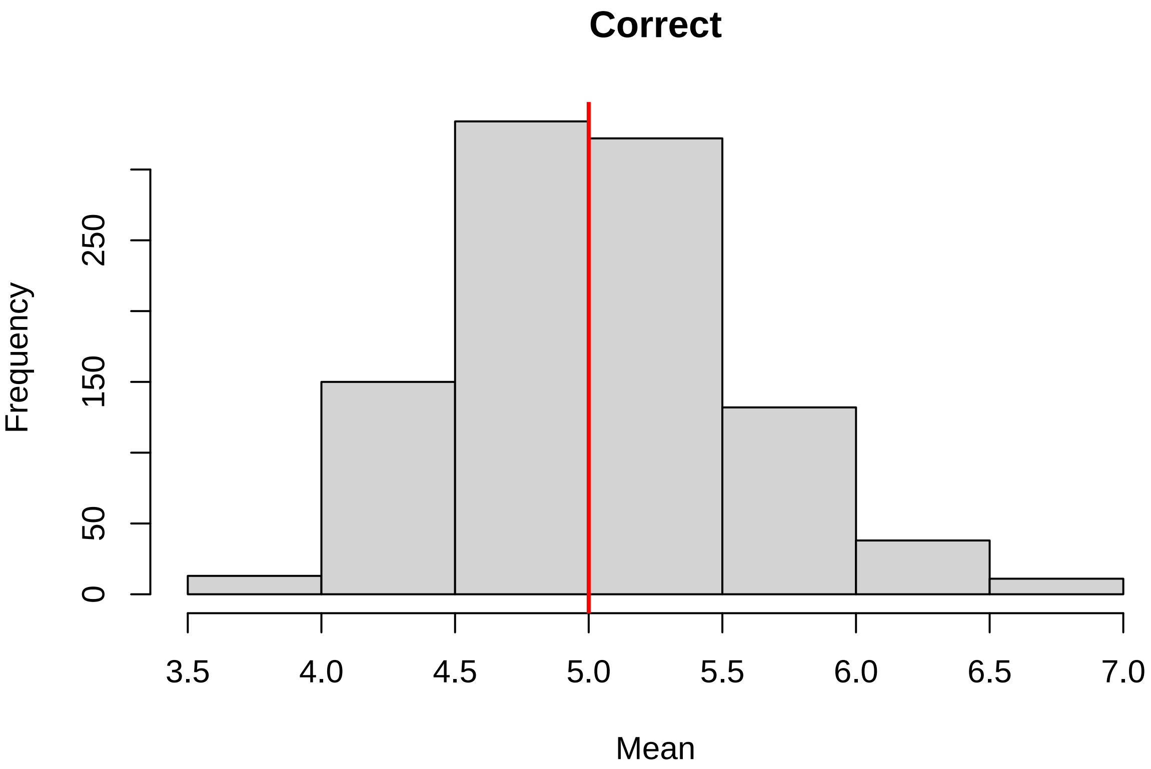
```r
  y[, l] <- rexp(n, theta)
}

# censoring indicators and corresponding observed samples
r <- x <- matrix(0, nrow = n, ncol = nsim)
for(l in 1:nsim){
  r[, l] <- ifelse(y[, l] <= 10, 1, 0)
  x[, l] <- ifelse(r[, l] == 1, y[, l], 10)
}

# estimates
mean_naive_1 <- mean_naive_2 <- mean_correct <- numeric(nsim)
for(l in 1:nsim){
  mean_naive_1[l] <- sum(x[r[, l] == 1, l])/sum(r[, l] == 1)
  mean_naive_2[l] <- sum(x[, l])/n
  mean_correct[l] <- sum(x[, l])/sum(r[, l] == 1)
}

hist(mean_naive_1, xlab = "Mean", xlim = c(2.5, 5), main = "Naive 1")
abline(v = 5, col = "red", lwd = 2)
```



```r
hist(mean_naive_2, xlab = "Mean", main = "Naive 2")
abline(v = 5, col = "red", lwd = 2)
```



```r
hist(mean_correct, xlab = "Mean", main = "Correct")
abline(v = 5, col = "red", lwd = 2)
```



As can be observed from the three histograms above, the two naive approaches provide biased estimates of the mean of the distribution, whereas correctly adjusting the likelihood provide accurate estimates of the mean.

2.

(a) The likelihood for $\boldsymbol{\beta}$ is given by

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{n} \{p_i(\boldsymbol{\beta})^{y_i}[1-p_i(\boldsymbol{\beta})]^{1-y_i}\} \\ &= \prod_{i=1}^{n} \left\{ \left(\frac{e^{\beta_0+x_i\beta_1}}{1+e^{\beta_0+x_i\beta_1}} \right)^{y_i} \left(\frac{1}{1+e^{\beta_0+x_i\beta_1}} \right)^{1-y_i} \right\}. \end{aligned}$$

The corresponding log likelihood is

$$\begin{aligned} \log L(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \left\{ y_i \log \left(\frac{e^{\beta_0+x_i\beta_1}}{1+e^{\beta_0+x_i\beta_1}} \right) + (1-y_i) \log \left(\frac{1}{1+e^{\beta_0+x_i\beta_1}} \right) \right\} \\ &= \sum_{i=1}^{n} \{ y_i(\beta_0+x_i\beta_1) - \log(1+e^{\beta_0+x_i\beta_1}) \}. \end{aligned}$$

The score function is given by

$$U(\boldsymbol{\beta}) = \begin{bmatrix} \frac{\partial}{\partial\beta_0} \log L(\boldsymbol{\beta}) \\ \frac{\partial}{\partial\beta_1} \log L(\boldsymbol{\beta}) \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} \left(y_i - \frac{e^{\beta_0+x_i\beta_1}}{1+e^{\beta_0+x_i\beta_1}} \right) \\ \sum_{i=1}^{n} \left(y_i x_i - x_i \frac{e^{\beta_0+x_i\beta_1}}{1+e^{\beta_0+x_i\beta_1}} \right) \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} \{y_i - p_i(\boldsymbol{\beta})\} \\ \sum_{i=1}^{n} x_i \{y_i - p_i(\boldsymbol{\beta})\} \end{bmatrix}.$$

(b) The solutions of $U(\boldsymbol{\beta}) = (0,0)'$ have no closed form expression and thus we need to resort to numerical methods. In order to use the recommended functions we need to define the log likelihood function.

```r
rm(list = ls())
load("dataw3.Rdata")
require(maxLik)

log_like <- function(param, data){
x <- data[,1]; y <- data[,2]
beta0 <- param[1]
beta1 <- param[2]
sum((y*(beta0+beta1*x)) - log(1+exp(beta0+beta1*x)))
}

mle <- maxLik(logLik = log_like, data = dataw3, start = c(beta0 = 0, beta1 = 0))
summary(mle)
```

```
## --------------------------------------------
## Maximum Likelihood estimation
## Newton-Raphson maximisation, 5 iterations
## Return code 1: gradient close to zero
## Log-Likelihood: -106.2768
## 2  free parameters
## Estimates:
##       Estimate Std. error t value  Pr(> t)
## beta0   0.7446     0.1797   4.143 3.43e-05 ***
## beta1  -2.3134     0.3690  -6.270 3.61e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## --------------------------------------------
```

The estimates are $\widehat{\beta_0} = 0.745$ (s.e.: 0.1797) and $\widehat{\beta_1} = -2.313$ (s.e.: 0.3690). We will now do the same but using the `optim` function.

```r
res <- optim(c(0, 0), log_like, data = dataw3,
             control = list(fnscale = -1), hessian = TRUE)
res
```

```
## $par
## [1]  0.7445009 -2.3129071
##
## $value
## [1] -106.2768
##
## $counts
## function gradient
##       65       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
##
## $hessian
##            [,1]      [,2]
## [1,] -35.520426 -6.184372
## [2,]  -6.184372 -8.417293
```

```r
# inverse of the observed Fisher information matrix
# (the diagonal extract the standard errors of beta0 and beta1)
sqrt(diag(solve(-1*res$hessian)))
```

```
## [1] 0.179673 0.369093
```

Estimates are very much similar to those obtained before.

3. If $\phi_1 = 0$ then the probability that a Y value is missing only depends on $\mathbf{x}$, and so data are MAR. For the missing data mechanism to be ignorable for likelihood inference we need data to be MAR, and so we need to impose that $\phi_1 = 0$, and we further need that the parameters of the model for the missing data mechanism and of our data model are distinct, i.e., that $\boldsymbol{\phi}_0$ and $\boldsymbol{\beta}$ are distinct. Formally this implies that the parameter space of $(\boldsymbol{\phi}_0, \boldsymbol{\beta})$ is equal to the Cartesian product of their individual product spaces. Informally stated, this means that the model for the missing data mechanism does not contain information about the parameters of the complete data model. For instance, if $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ and $\boldsymbol{\phi}_0 = (\phi_{01}, \beta_1)'$ then distinctness does not hold.