University of Edinburgh, School of Mathematics

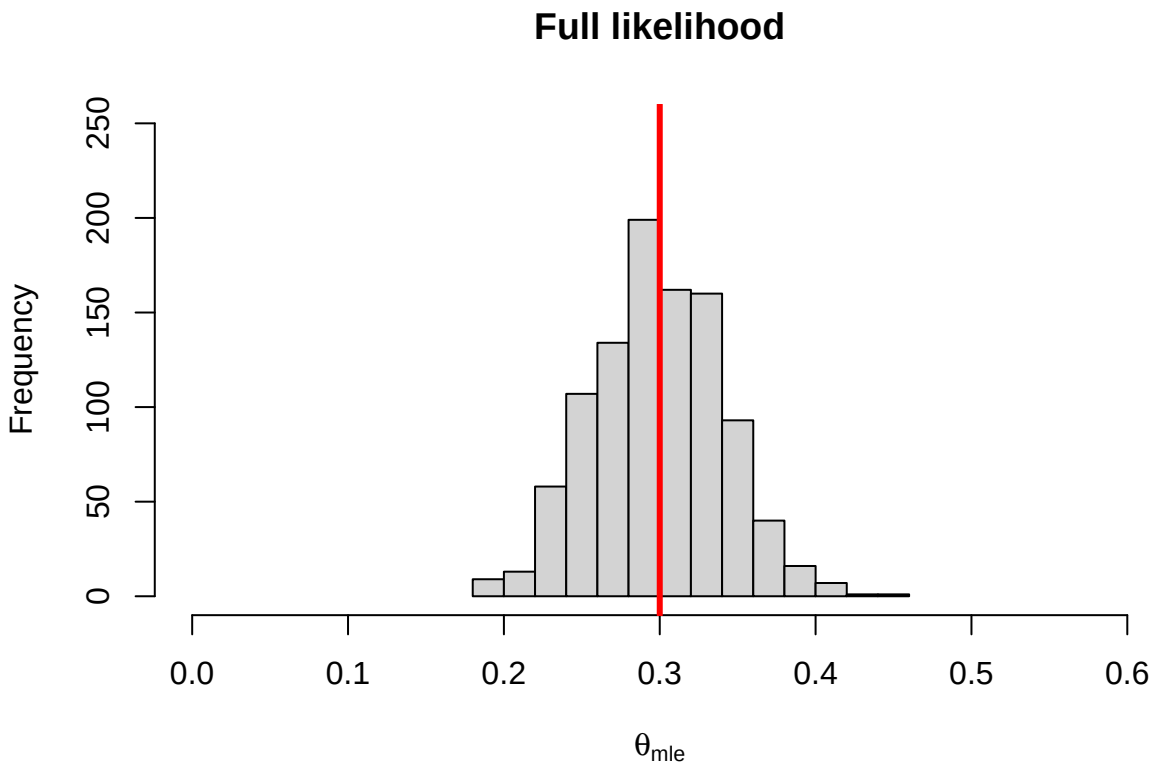
Incomplete Data Analysis, 2020/2021

Ignorability – simulation study

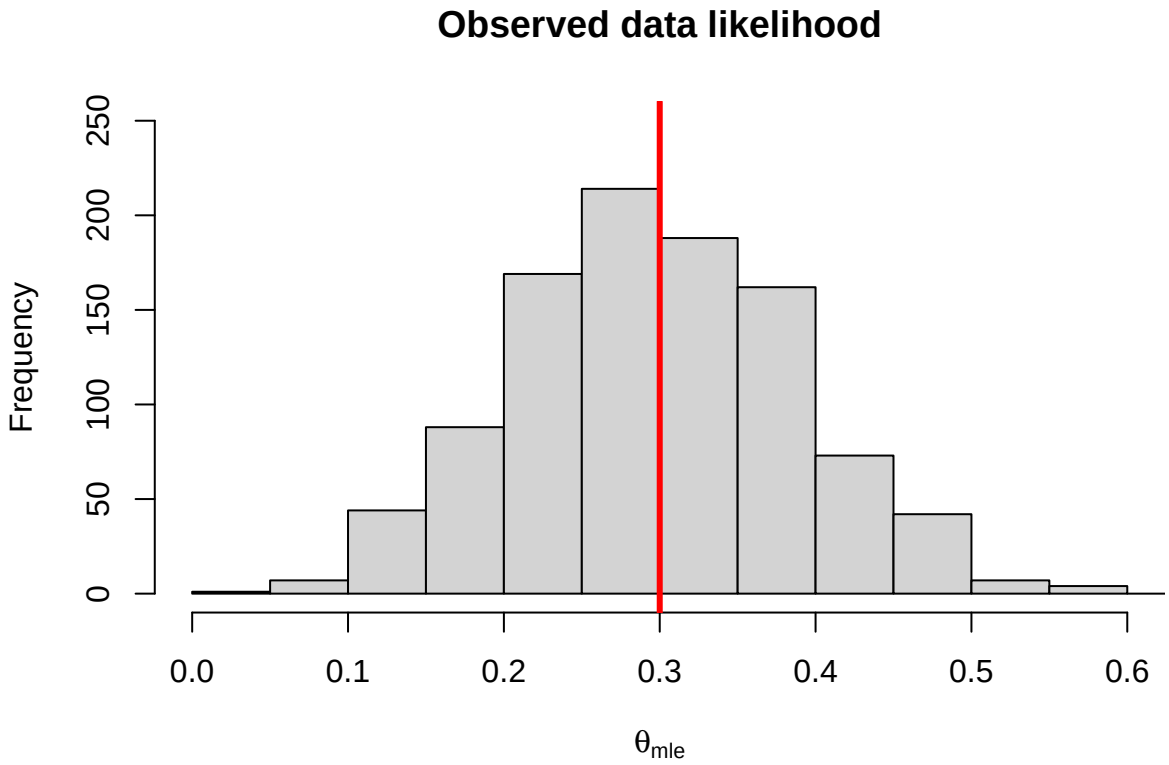
Vanda Inácio

In this supplementary file I show how to reproduce the results from the simulation study presented in the slides. We are assuming that $Y_i \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$ and $R_i \mid Y_i \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$. I will simulate $nsim = 1000$ datasets of sample size $n = 100$ and consider $\theta = 0.3$. I will store the generated data and corresponding missing data indicators in a $n \times nsim$ matrix. The maximum likelihood estimates (from the 1000 simulated datasets) based on both the full and observed data likelihood will be stored in two separate vectors.

```
nsim <- 1000; n <- 100; theta <- 0.3
y <- r <- matrix(0, nrow = n, ncol = nsim)
mle_full <- mle_observed <- numeric(nsim)

set.seed(1)
for(l in 1:nsim){
y[, l] <- rbinom(n, 1, theta)
r[, l] <- rbinom(n, 1, theta)

m <- length(which(r[,l] == 1))

mle_full[l] <- (m + sum(y[r[, l] == 1, l]))/(m + n)
mle_observed[l] <- sum(y[r[, l] == 1, l])/m
}

hist(mle_full, ylim = c(0, 250), xlim = c(0, 0.6),
     xlab = expression(theta[mle]), main = "Full likelihood")
abline(v = theta, col = "red", lwd = 3)
```




```
hist(mle_observed, ylim = c(0, 250), xlim = c(0, 0.6),
     xlab = expression(theta[mle]), main = "Observed data likelihood")
abline(v = theta, col="red", lwd = 3)
```



We will now violate the MAR assumption. We repeat a similar exercise but now, instead of violating the

non-distinctness of parameters assumption, we violate the MAR assumption. In particular, we assume

$$Y_i \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta), \quad \text{and} \quad \Pr(R_i = 1 \mid Y_i) = \frac{e^{Y_i}}{1 + e^{Y_i}}.$$

The code follows below.

```
r <- matrix(0, nrow = n, ncol = nsim)
mle_observed_mnar <- numeric(nsim)

set.seed(1)
for(l in 1:nsim){
r[, l] <- rbinom(n, 1, exp(y[, l])/(1+exp(y[, l])))

m <- length(which(r[,l] == 1))

mle_observed_mnar[l] <- sum(y[r[, l] == 1, l])/m
}

hist(mle_observed_mnar, ylim = c(0, 400), xlim = c(0, 0.6),
     xlab = expression(theta[mle]), main = "MNAR data")
abline(v = theta, col = "red", lwd = 3)
```