University of Edinburgh
School of Mathematics
Incomplete Data Analysis

Supporting Materials for Lecture 6

In this file I show how to reproduce the results from the simulation studies presented in the notes.

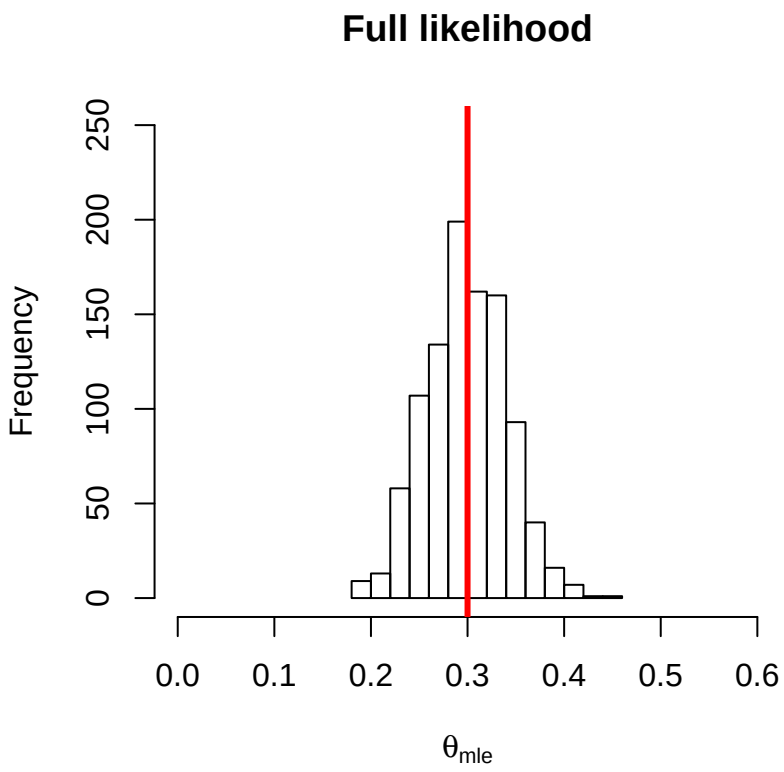
Non distinct parameters

As always, I start by cleaning the workspace and fixing the seed.

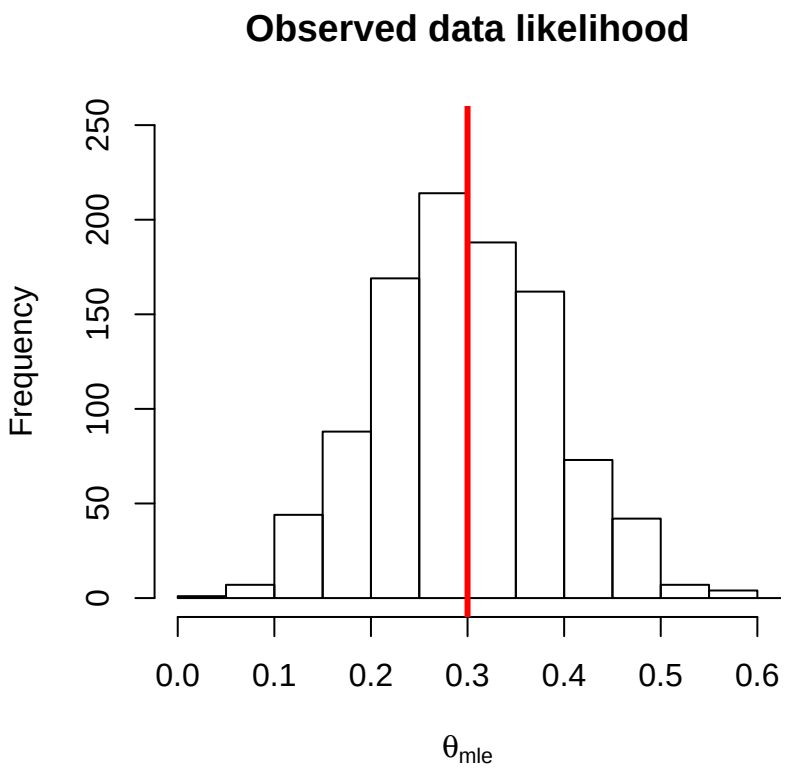
```
rm(list=ls())
set.seed(1)
```

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)
mle1=mle2=numeric(nsim)

for(l in 1:nsim){
y[,l]=rbinom(n,1,theta)
r[,l]=rbinom(n,1,theta)

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

mle1[l]=(m+sum(y[r[,l]==1,l]))/(m+n)
mle2[l]=sum(y[r[,l]==1,l])/m
}

mean(mle1); var(mle1)

## [1] 0.2990487
## [1] 0.00169832

mean(mle2); var(mle2)

## [1] 0.3001491
## [1] 0.008249281

hist(mle1,ylim=c(0,250),xlim=c(0,0.6),xlab=expression(theta[mle]),main="Full likelihood")
abline(v=0.3,col="red",lwd=3)
```




```
hist(mle2,ylim=c(0,250),xlim=c(0,0.6),xlab=expression(theta[mle]),main="Observed data likelihood")
abline(v=0.3,col="red",lwd=3)
```



Violating 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

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

The code follows below.

```
rm(list=ls())
set.seed(1)

nsim=1000; n=100; theta=0.3
y=r=matrix(0,nrow=n,ncol=nsim)
mle=numeric(nsim)

for(l in 1:nsim){
y[,l]=rbinom(n,1,theta)
r[,l]=rbinom(n,1,exp(y[,l])/(1+exp(y[,l])))

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

```r
mle[l]=sum(y[r[,l]==1,l])/m
}

mean(mle); var(mle)

## [1] 0.3844768
## [1] 0.00407711

hist(mle,ylim=c(0,400),xlim=c(0,0.6),xlab=expression(theta[mle]),main="MNAR data")
abline(v=0.3,col="red",lwd=3)
```