

Lecture 29: Introducing the EM algorithm

Ciaran Evans

Plan for next week

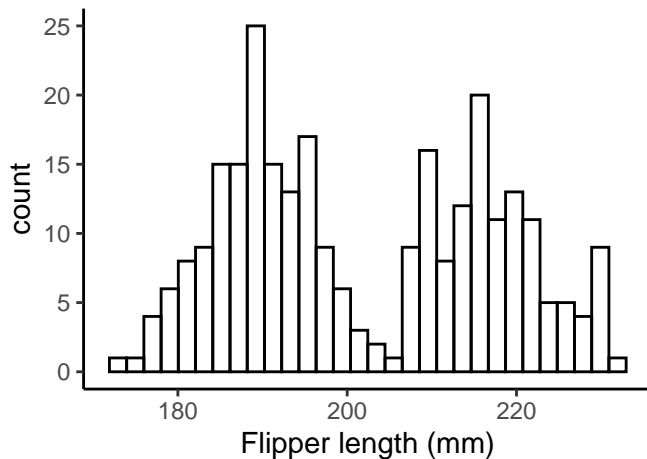
- ▶ Monday: continue EM algorithm
- ▶ Wednesday and Friday: project work days
- ▶ Extra office hours on Tuesday, Wednesday, and Thursday

Motivation: penguins data

Data on 276 penguins (Adelie or Gentoo) on three different islands (Torgersen, Biscoe, Dream) near Antarctica. Variables include

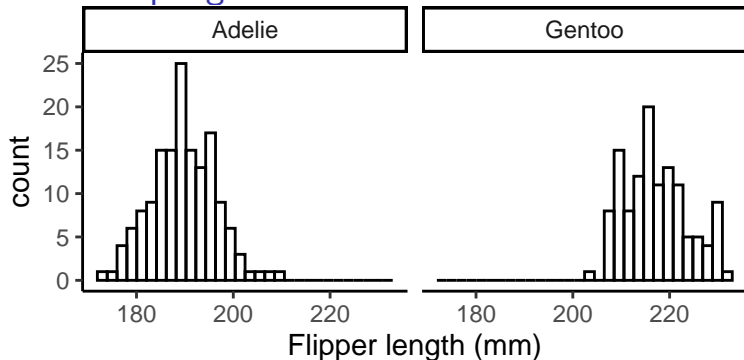
- ▶ species
- ▶ island
- ▶ characteristics like bill length, flipper length, etc.

Motivation: penguins data



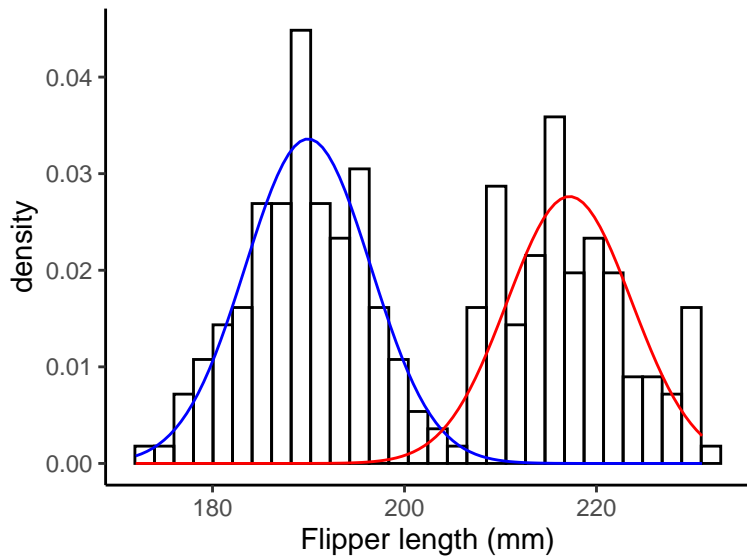
Question: What do you notice about the distribution of flipper length? Why might this be the case?

Motivation: penguins data

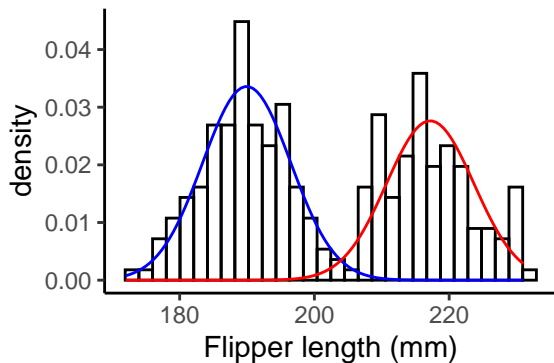


Question: How could I model the distribution of flipper length in each group? What parameters would I estimate?

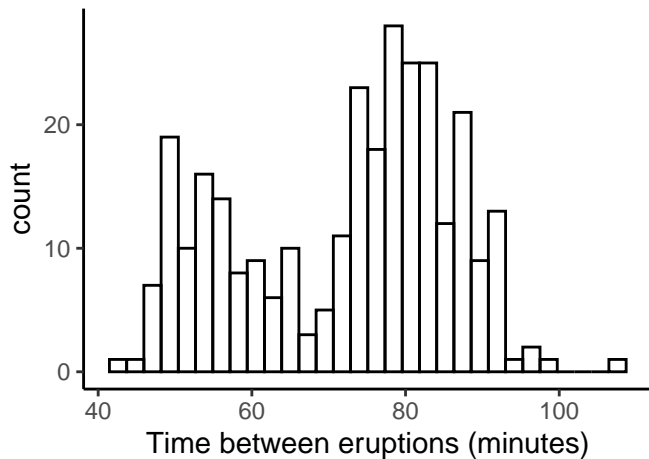
Motivation: penguins data



Writing down a model

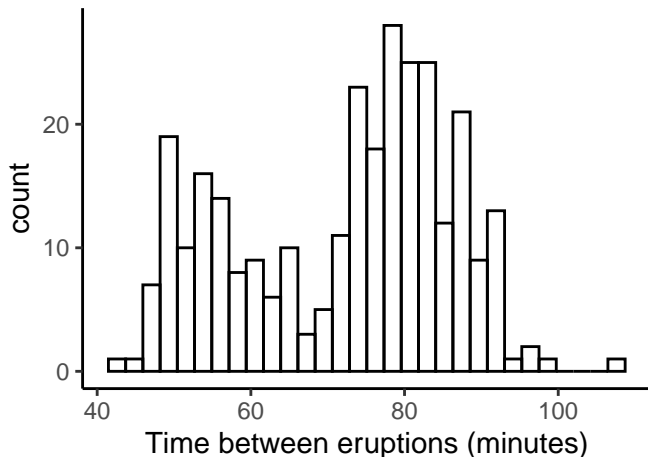


Time between Old Faithful geyser eruptions



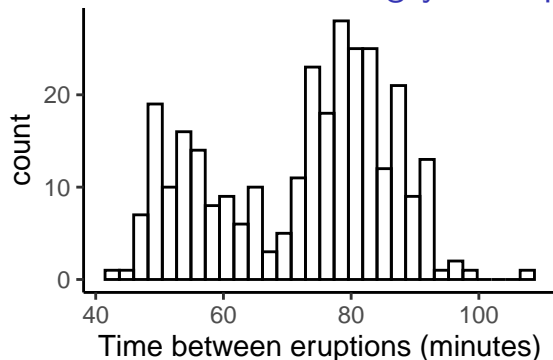
Question: What do you notice about the distribution of waiting times? Why might this be the case?

Time between Old Faithful geyser eruptions



Question: It seems like there are two groups here, but we don't know what they are. What should we do to estimate both the groups and their distributions?

Time between Old Faithful geyser eruptions



Model:

Gaussian mixture model

- ▶ Observe data X_1, \dots, X_n
- ▶ Assume each observation i comes from one of k groups. Let $Z_i \in \{1, \dots, k\}$ denote the group assignment
 - ▶ The group Z is an unobserved (**latent**) variable

Model:

Estimating model parameters: EM algorithm

The **EM algorithm** allows us to estimate both the unknown group assignments, *and* the parameters for each group's distribution (we will discuss the details later). In R:

```
library(mixtools)
```

```
normalmixEM(geyser$waiting, lambda = c(0.5, 0.5), k=2)
```

- ▶ `normalmixEM`: function for estimating parameters in a mixture of normal distributions
- ▶ `lambda`: initial guess at the proportion of data in each group
- ▶ `k`: number of groups

Estimating model parameters: EM algorithm

```
library(mixtools)
```

```
em_res <- normalmixEM(geyser$waiting, lambda = c(0.5, 0.5),  
                      k=2)
```

```
## number of iterations= 28
```

```
em_res$lambda
```

```
## [1] 0.3075953 0.6924047
```

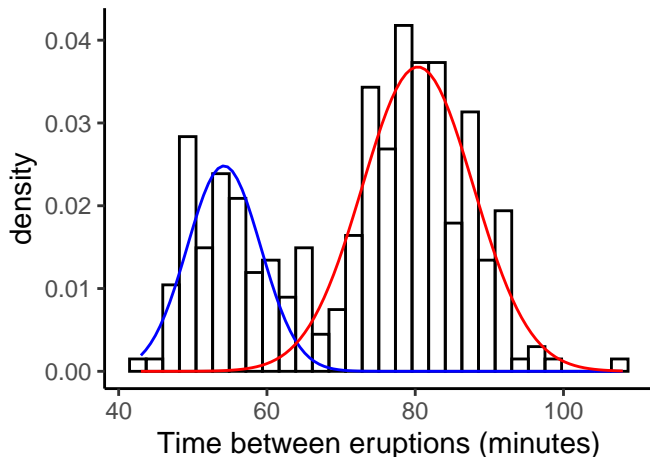
```
em_res$mu
```

```
## [1] 54.20271 80.36036
```

```
em_res$sigma
```

```
## [1] 4.952044 7.507597
```

Fitted parameters



- ▶ Estimated proportion of data in each group: 0.308, 0.692
- ▶ Estimated group means: $\hat{\mu}_1 = 54.203$, $\hat{\mu}_2 = 80.360$
- ▶ Estimated group sd: $\hat{\sigma}_1 = 4.951$, $\hat{\sigma}_2 = 7.508$

Your turn

Simulate data from a Gaussian mixture and explore parameter estimation:

https://sta379-s25.github.io/practice_questions/pq_29.html

- ▶ Start in class
- ▶ Welcome to work with a neighbor
- ▶ Solutions will be posted later on the course website