Lecture 29: Introducing the EM algorithm

Ciaran Evans

Plan for next week

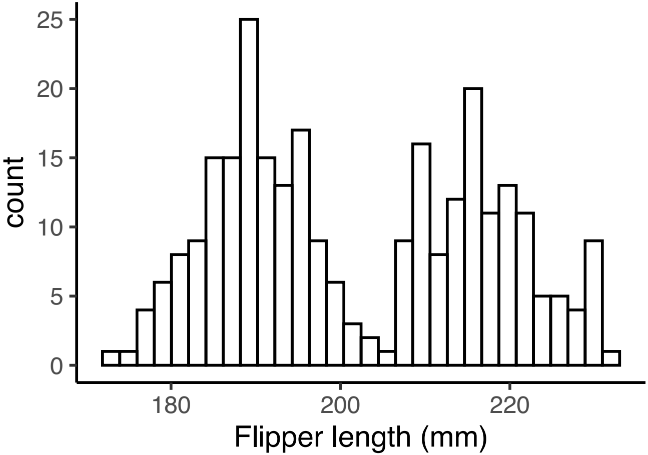
- Monday: continue EM algorithm
- Wednesday and Friday: project work days
- Extra office hours on Tuesday, Wednesday, and Thursday

Motivation: penguins data

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

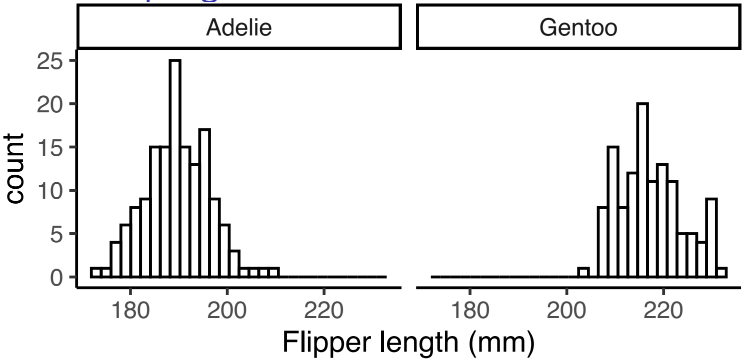
- species
- island
- characteristics like bill length, flipper length, etc.

Motivation: penguins data



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

Motivation: penguins data



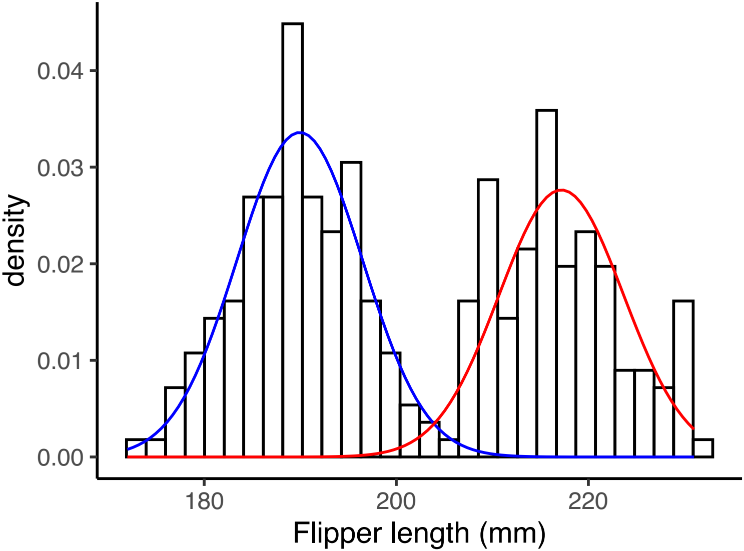
Question: How could I model the distribution of flipper length in each group? What parameters would I estimate?

$$\text{Length} \mid (\text{species} = \text{Adelie}) \sim N(\mu_1, \sigma_1^2)$$

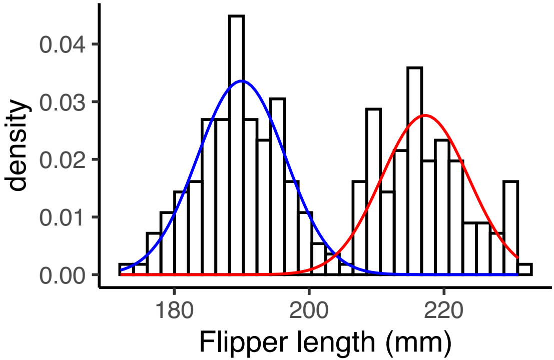
$$\text{Length} \mid (\text{species} = \text{Gentoo}) \sim N(\mu_2, \sigma_2^2)$$

want to estimate $\mu_1, \sigma_1^2, \mu_2, \sigma_2^2,$
$P(\text{Species} = \text{Adelie})$ (or $P(\text{Species} = \text{Gentoo})$)

Motivation: penguins data



Writing down a model



x_i = Flipper length for i^{th} penguin

z_i = species (1 = Adelie, 2 = Gentoo)

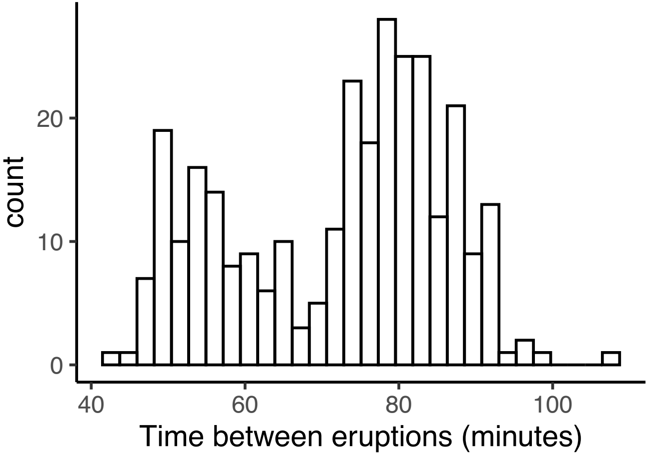
$x_i \mid (z_i = 1) \sim N(\mu_1, \sigma_1^2)$ $\quad$ $x_i \mid (z_i = 2) \sim N(\mu_2, \sigma_2^2)$

$\hat{p}(z_i = 2) = 0.449$

$\hat{\mu}_1 = 189.95$ $\quad$ $\hat{\sigma}_1 = 6.54$

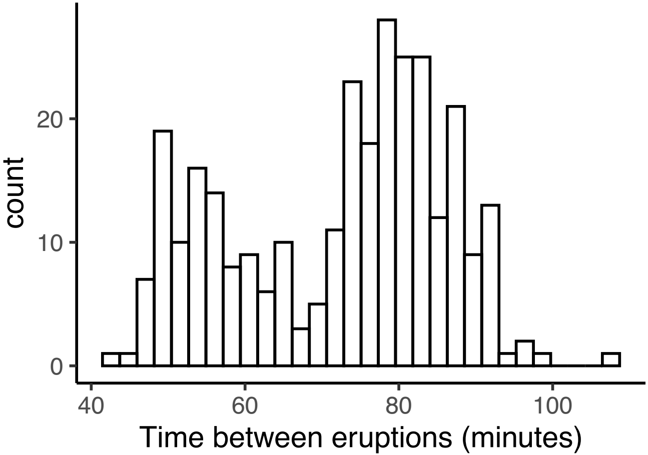
$\hat{\mu}_2 = 217.19$ $\quad$ $\hat{\sigma}_2 = 6.48$

Time between Old Faithful geyser eruptions



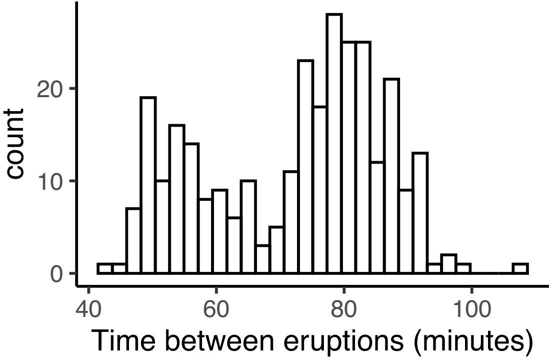
Question: What do you notice about the distribution of waiting times? Why might this be the case?

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:

X_i = time between eruptions

Z_i = group (unobserved, i.e. latent)

$$X_i \mid (Z_i = 1) \sim N(\mu_1, \sigma_1^2)$$

$$X_i \mid (Z_i = 2) \sim N(\mu_2, \sigma_2^2)$$

Gaussian mixture model

- Observe data $X_1, \ldots, X_n$

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

Model:

$P(Z_i = j) = \lambda_j$ (probability of belonging to group j)

$\sum_{j=1}^{k} \lambda_j = 1$

$X_i \mid (Z_i = j) \sim N(\mu_j, \sigma_j^2)$

Normal distribution for each group

Marginal distribution of X is a mixture of the Normal distributions for each group

Estimating model parameters: EM algorithm

(Expectation-Maximization)

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:

data to model

```
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

λ_j guesses

Estimating model parameters: EM algorithm

```
library(mixtools)

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

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

← iterative estimation algorithm

```
em_res$lambda
```

$\hat{\lambda}_1$ $\hat{\lambda}_2$

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

```
em_res$mu
```

$\hat{\mu}_1$ $\hat{\mu}_2$

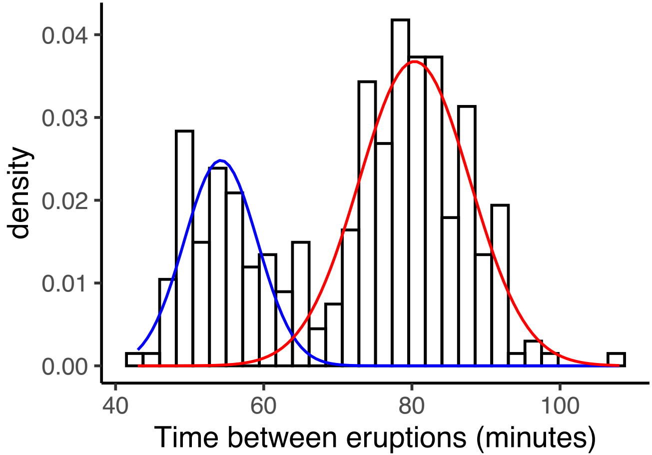
```
## [1] 54.20271 80.36036
```

```
em_res$sigma
```

$\hat{\sigma}_1$ $\hat{\sigma}_2$

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

Fitted parameters



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

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