Exercise on Mixture Models

STA 325 (exercise from BMLR)

```
library(gridExtra)
library(knitr)
library(kableExtra)
library(tidyverse)
```

1. **Mixture of two normal distributions**

Sometimes, a value may be best modeled by a mixture of two normal distributions. We would have 5 parameters in this case— $\mu_1, \sigma_1, \mu_2, \sigma_2, \alpha$, where $0 < \alpha < 1$ is a mixing parameter determining the probability that an observation comes from the first distribution. We would then have $f(y) = \alpha\ f_1(y) + (1 - \alpha)\ f_2(y)$ (where $f_i(y)$ is the pdf of the normal distribution with μ_i, σ_i).

One phenomenon which could be modeled this way would be the waiting times between eruptions of Old Faithful geyser in Yellowstone National Park. The data can be accessed in R through `faithful`, and a histogram of wait times can be found the figure below. The MLEs of our 5 parameters would be the combination of values that produces the maximum probability of our observed data. We will need to approximate the MLE's using the EM algorithm. Find a combination of $\mu_1, \sigma_1, \mu_2, \sigma_2, \alpha$ for this distribution such that the logged likelihood is above -1050. (The command `dnorm(x, mean, sd)`, which outputs $f(y)$ assuming $Y \sim \mathrm{N}(\mu, \sigma)$, will be helpful in calculating likelihoods.)

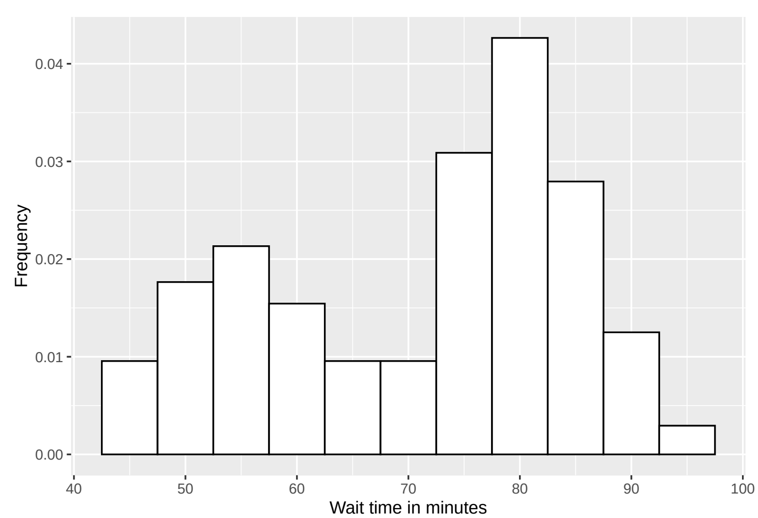

Figure 1: Waiting time between eruptions of Old Faithful.

Hint: Use the `normalmixEM()` from the `packagemixtools` package to estimate the MLE's.

2. **Beta-binomial distribution.** We can generate more distributions by mixing two random variables. Beta-binomial random variables are binomial random variables with fixed n whose parameter p follows a beta distribution with fixed parameters α, β. In more detail, we would first draw p_1 from our beta distribution, and then generate our first observation y_1, a random number of successes from a binomial (n, p_1) distribution. Then, we would generate a new p_2 from our beta distribution, and use a binomial distribution with parameters n, p_2 to generate our second observation y_2. We would continue this process until desired.

Note that all of the observations y_i will be integer values from $0, 1, \ldots, n$. With this in mind, use `rbinom()` to simulate 1,000 observations from a plain old vanilla binomial random variable with $n = 10$ and $p = 0.8$. Plot a histogram of these binomial observations. Then, do the following to generate a beta-binomial distribution:

a. Draw p_i from the beta distribution with $\alpha = 4$ and $\beta = 1$.
b. Generate an observation y_i from a binomial distribution with $n = 10$ and $p = p_i$.
c. Repeat (a) and (b) 1,000 times ($i = 1, \ldots, 1000$).
d. Plot a histogram of these beta-binomial observations.

Compare the histograms of the "plain old" binomial and beta-binomial distributions. How do their shapes, standard deviations, means, possible values, etc. compare?