ECO518 Assignment 1

Tom

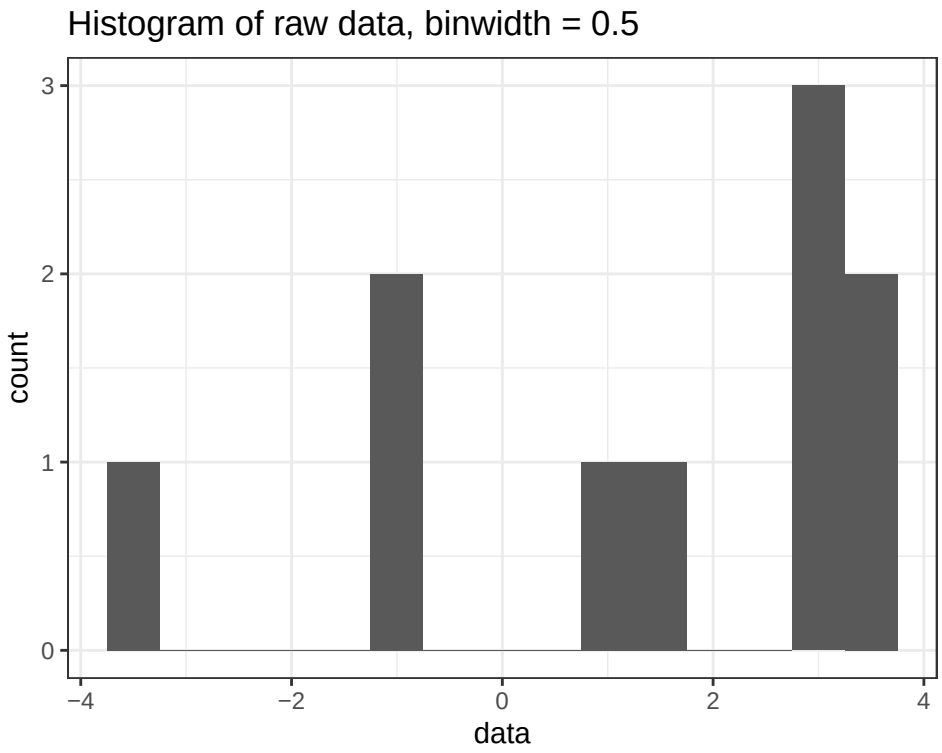
Firstly, we set up our coding environment, and loading packages...

```
knitr::opts_chunk$set(fig.width=5, fig.height=4)
rm(list = ls())
library(dplyr, warn.conflicts = FALSE)   # data manipulation, piping
library(ggplot2, warn.conflicts = FALSE) # nice plots
library(purrr, warn.conflicts = FALSE)   # nice loops
library(tidyr, warn.conflicts = FALSE)   # reshaping function
theme_set(theme_bw()) # ggplot theme
set.seed(123)
```

Problem 1

Before beginning this problem, we load in the data, and present a histogram.

```
data <- c(3.66, 1.00, -0.87, 2.90, -0.80, 3.20, 1.69, -3.53, 3.22, 3.53)
N <- length(data)
ggplot(data.frame(data = data)) +
  geom_histogram(aes(x = data), binwidth = 0.5) +
  ggtitle("Histogram of raw data, binwidth = 0.5")
```



From the histogram, we can see some evidence of skew. However, given our very small sample, this evidence is pretty weak.

1. a)

Make 1000 bootstrap draws from this sample of size 10 and use them to construct an estimate of the pdf of the sample mean and the standard deviation of the sample mean.

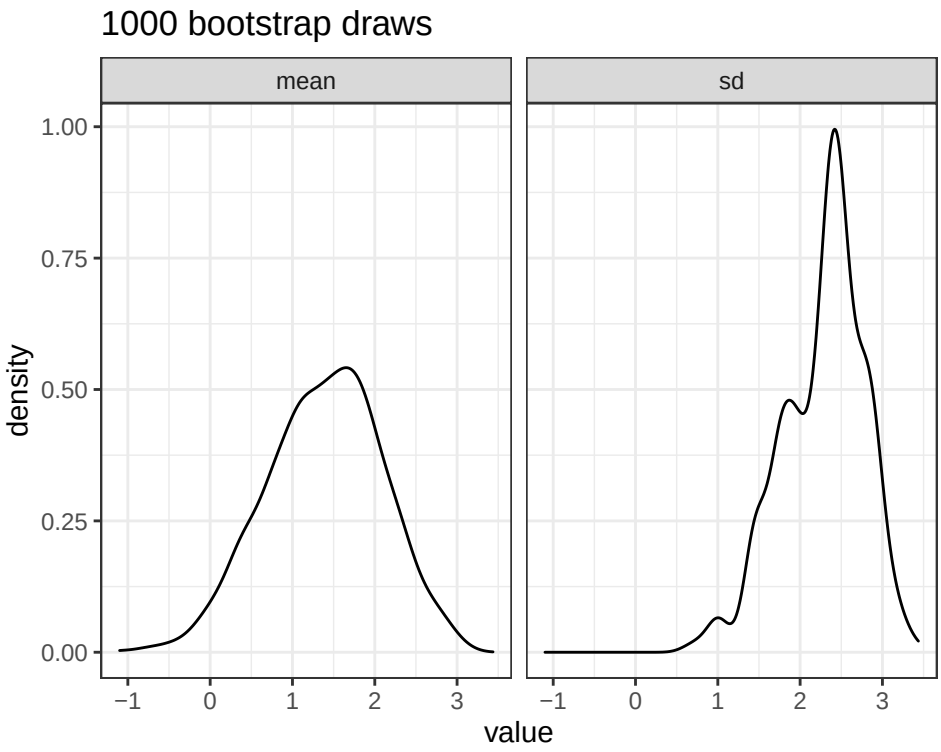
```
# Define a function for taking draws from the data, returning a
# conveniently formated dataframe
take_draw <- function(i, data, N)
  data.frame(value = sample(data, size = N, replace = TRUE), draw = i)

# Take 1000 draws, save them in a dataframe for convenience.
df <- map_dfr(seq(1:1000), take_draw, data = data, N = N)

# Use this to get estimates of the pdf for the mean and SD
stat_df <- df %>%
  group_by(draw) %>%
  summarise(mean = mean(value), sd = sd(value))
```

1. b)

Plot the estimated pdf, which you can do in R with ***plot****(bkde()) using the KernelSmooth package or with the* ***hist****() function.*

```
# plot (using ggplot, my preferred plotting option)
stat_df %>%
```

```
  pivot_longer(cols = -draw, names_to = "statistic") %>%
  ggplot() +
  geom_density(aes(x = value)) +
  facet_wrap(~statistic) +
  ggtitle("1000 bootstrap draws")
```



1. c)

On the assumption that when the mean of the population distribution changes, it changes the distribution by a pure location shift, use your bootstrapped sample to construct a 95% confidence interval for the mean (Should it be "flipped"?)

We can construct quantiles of the distribution, by ordering the means of each draw.

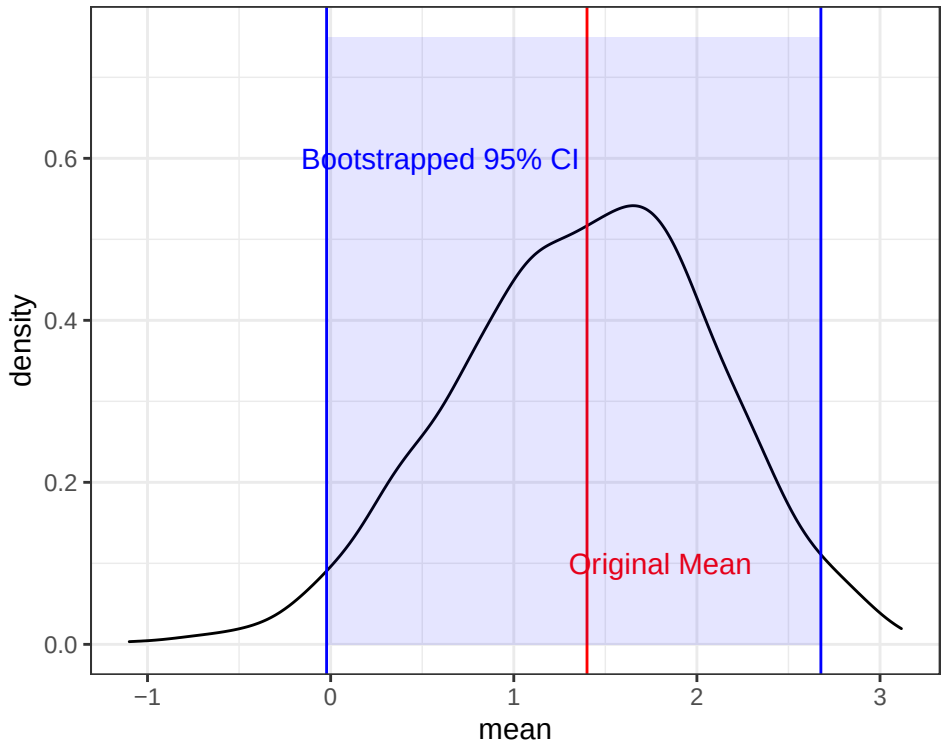
```
CI <- quantile(stat_df$mean, probs = c(0.025, 0.975))
# Show the 95% confidence interval values...
print(CI)
```

```
##    2.5%   97.5%
## -0.0221  2.6771
```

We can visualize these quantiles on a plot. We also present the mean of the original data, so we can analyze the skew.

```
ggplot() +
  geom_density(data = stat_df, aes(x = mean)) +
  geom_vline(xintercept = mean(data), color = "red") +
  geom_text(aes(x = mean(data) + 0.4, y = 0.1, label = "Original Mean"), color = "red") +
```

```
  annotate("rect", xmin = CI[1], xmax = CI[2], ymin = 0, ymax = 0.75,
        alpha = .1, fill = "blue") +
  geom_vline(xintercept = CI[1], color = "blue") +
  geom_vline(xintercept = CI[2], color = "blue") +
  geom_text(aes(x = mean(data) - 0.8, y = 0.6,
                label = "Bootstrapped 95% CI"), color = "blue")
```



We can see from this plot that the bootstrapped estimates do not appear to be particularly skewed. This suggests that we don't need to flip our estimates. We don't have reason to think that our mean is biased.

1. d)

The data were actually generated from an equally weighted mixture of two normals, one with mean 0, standard deviation 3, one with mean 3, standard deviation 0.2.

Suppose we knew that the distribution had this form, except for the means of the two components. If the first component has mean − 1.5 and the second mean + 1.5, with unknown, we have a pure location shift model of the data.

Plot the likelihood for this sample as a function of . Does it imply a 95% credible set (under a flat prior) similar to what you got from the bootstrap?

Firstly, lets write down the likelihood.

$$L(Y|\mu) = \left(\frac{1}{\sigma_1\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x_i-\mu_1}{\sigma_1}\right)^2}\right)^{0.5N} * \left(\frac{1}{\sigma_2\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x_i-\mu_2}{\sigma_2}\right)^2}\right)^{0.5N}$$

Subbing in our known parameters, we can write…

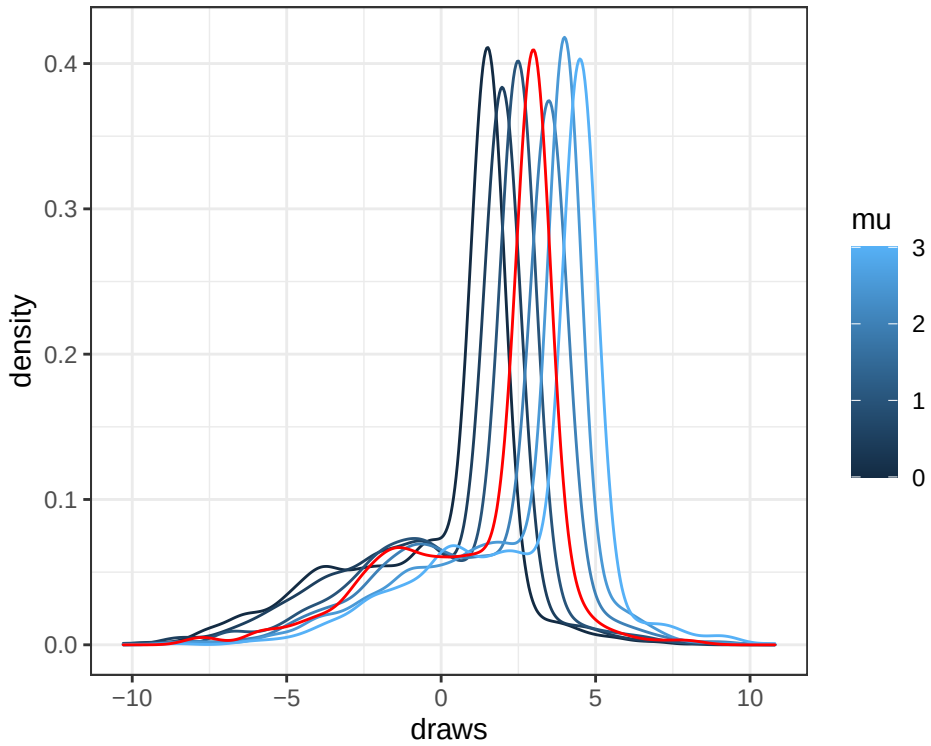
$$L(Y|\mu) = \left(\frac{1}{3\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-(\mu-1.5)}{3}\right)^2}\right)^4 * \left(\frac{1}{0.2\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-(\mu+1.5}{0.2}\right)^2}\right)^4$$

Since we know mu is 1.5, we can plot this likelihood function:

- I don't know how to allocate x vales from the sample to each normal and hence am stuck.
- I know that we want to put four observations in each one. But how do we know which ones?
- What does it mean to plot the likelihood as a function of mu ? Different function for each?

Alternative approach: draw from these two normals?

```
N = 1000
# Function for drawing from the likelihood, for a given mu
sim_normals <- function(mu, N)
  data.frame(mu = mu, draws = c(rnorm(N/2, mean = mu - 1.5, sd = 3),
              rnorm(N/2, mean = mu + 1.5, sd = 0.2)))

sim_data <- map_dfr(seq(0,3,0.5), sim_normals, N = N)

ggplot() +
  geom_density(data = filter(sim_data, mu != 1.5), aes(x = draws, color = mu, group = mu), alpha = .001
  geom_density(data = filter(sim_data, mu == 1.5), aes(x = draws), color = "red", alpha = 1)
```



This doesn't look very similar to what we got from the bootstrap. However, lets take a look at the 95% credible set...

- this doesn't seem right - we want to look at the posterior??

```
# quantile(filter(sim_data, mu == 1.5) , 0.5)
```

1. e)

Now suppose instead that we know the first component has mean 0, but don't know the mean of the second component. The mean of the second component is 2 , so that is still the mean of the distribution, but no longer produces a pure location shift. Plot the likelihood and find a 95% credible set for for this case.

- We can just estimate mu from the sample? ??

Problem 2: The lady tasting tea

Part 1

A lady claims to be able to tell whether she has been served tea with the milk put in the cup before, or after, the tea. She is given a sequence of cups of tea, some of which (or perhaps all, or none of which) have had the milk put in first. She announces her guesses, we find out how many of them were correct. How do we assess the evidence for her abilities?

This is a classic example, probably the simplest and most appealing case for randomization inference. There is no need to appeal to hypothetical repetitions of the lady's test, as in the usual frequentist framework. Here there 2^N possible sequences of milk-first/milk-second, where N is the number of trials. So if she is given, say, 5 cups, there are 32 possibilities. If we choose the sequence of cups she is presented with "at random" from these 32 possibilities, the probability that she gets them all right if she is "just guessing" is 1/32, and that she gets at least 4 out of 5 right is 6/32. This lets us construct an exact, finite-sample test of "H0: she's guessing" at a .03125 or .1875 significance level.

But what is the alternative hypothesis? If it is that her guessing is i.i.d. across cups, with probability p of being right on each cup, we can construct a likelihood function for p and ask what is the flat-prior probability of $p > .5$ given that she has made 4, or 5, correct guesses. Do that. (The posterior is a Beta distribution.)

Let $Y = \{y_1, y_2, ...y_5\}$ denote the observed sample of guesses, where $y_i = 1$ if the guess is correct, and $y_i = 0$ if the guess is incorrect. Let $n = \sum_i y_i$ denote the total number of correct guesses, and let $N = 5$ be the sample size.

Then, we can write the likelihood function for p as follows:

$$L(Y|p) = \prod_{i=1}^{N} p^{y_i}(1-p)^{1-y_i} = p^n(1-p)^{N-n}$$

Given we have a flat prior, we can derive the posterior, $q(p|Y)$ as follows:

$$q(p|Y) \propto L(Y|p) = p^n(1-p)^{N-n} \propto Beta(n+1, N-n+1)$$

Now, we can use this to estimate the probability that $p > 0.5$, given she made 4 or five correct guesses. To do this, we just take a lot of draws from the relevant *Beta* distribtuion, and count the number where $p > 0.5$.

Firstly, define a function for doing this simulation.

```
post_prob_greater_than_.5 <- function(N,n){
  draws <- rbeta(10000, n + 1, N-n + 1)
  sum(draws>0.5) / length(draws)
}
N <- 5
```

Now, run it for n = 4...

```
post_prob_greater_than_.5(N, 4)
```

```
## [1] 0.8948
```

And for n = 5...

```
post_prob_greater_than_.5(N, 5)
```

```
## [1] 0.9861
```

Part 2

Suppose it happens that every one of the 5 test cups of tea has milk first, and she guesses milk first on every one of them. Does this seem like less strong evidence of her ability than if the cups she had been presented had at least some of each type? How might one justify a claim that the evidence is less strong in this case?

We would need to incorporate some kind of prior, suggesting that it is easier to guess them all correctly if they are all of the same type.

Prior, $\pi(p)$ could be decreasing in p, for example.