Lab 1: Key

Q1. Prior Distributions (4 pts)

Summarize a prior distribution in a way that someone with a minimal statistical background, such as a parent or sibling, could understand.

Note that while you are taking a statistics class, writing, and speaking, will be essential to how we convey our results. Please use proper grammar, sentence structure, and complete paragraphs.

The goal of statistical modeling is to use parameterized models to learn about the state of nature. The parameters in the those statistical model represent characteristics, such as the average response, about a scientific process. Bayesian inference requires that a prior distribution that encodes existing beliefs about those parameters.

Q2. Beta Distribution

a. (2 pts)

Write out probability distribution for a beta distribution. Hint, you can use LaTex…

$$p(x|\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{1-\beta}}{\beta(\alpha,\beta)}$$

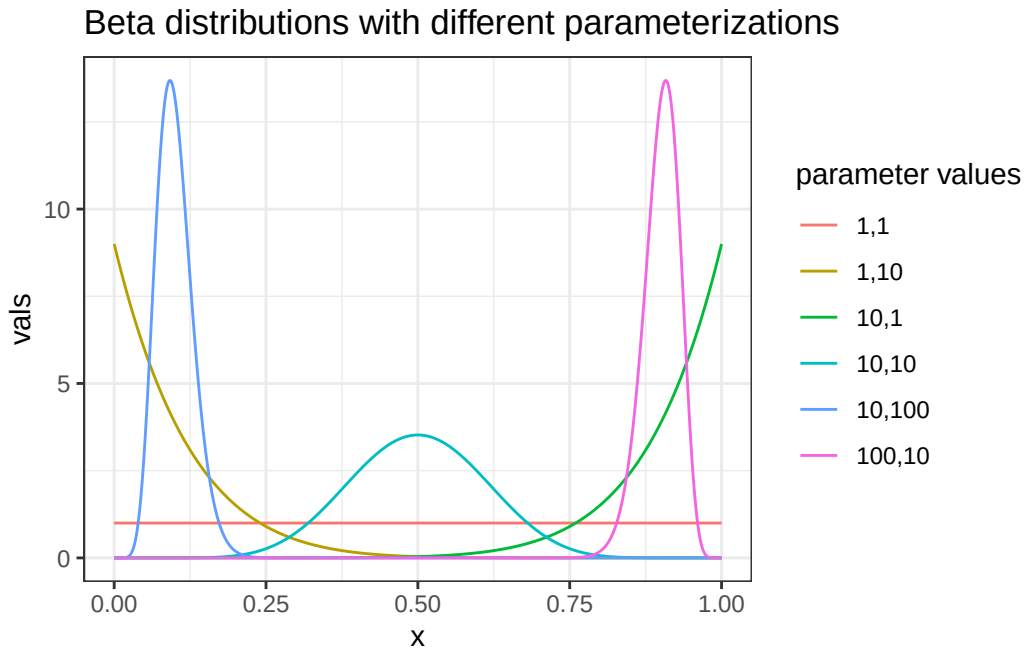
b. (4 pts)

Using `ggplot2` and `rbeta()` or `dbeta()` investigate the impact of α and β on the shape of the resultant probability distribution function. In particular, create or overlay at least 4 different curves that correspond to alternative specifications of α and β.

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```r
num_x <- 10000
x_seq <- seq(0,1, length.out = num_x)
tibble(vals = c(dbeta(x_seq, 1, 1), dbeta(x_seq, 9, 1), dbeta(x_seq, 1, 9),
                dbeta(x_seq, 10, 10), dbeta(x_seq, 90, 10), dbeta(x_seq, 10, 90)),
       `parameter values` = rep(c('1,1','10,1', '1,10', '10,10','100,10', '10,100'), each =nu
       x = rep(x_seq, 6)) %>%
  ggplot(aes(x = x, y = vals, color = `parameter values`)) +
  geom_line() + theme_bw() +
  ggtitle('Beta distributions with different parameterizations')
```



c. (4 pts)

Building on Q2b, summarize how $\frac{\alpha}{\alpha+\beta}$ and $\alpha+\beta$ change the shape of the probability distribution.

- $\frac{\alpha}{\alpha+\beta}$ *is the mean of the distribution*
- $\alpha+\beta$ *is related to the concentration, higher values result in more precise distributions*

Q3. Winter Temperature

On Tuesday we considered waiting time for the sunnyside lift on Saturday mornings at Bridger Bowl. Suppose you found the wait times to be too long for your liking. Now let's consider estimating the temperature at Hyalite Canyon (at the reservoir) at 10 AM on Saturdays in February.

We will outline the first three steps of a Bayesian inference:

a. (4 pts) Identify the data relevant to the research question.

Specifically, describe how you design a data collection process to answer the research question (What are the range of expected temperatures at Hyalite Reservoir on Saturdays in February?).

Ideally, we'd have access to historical data, such as https://www.mtavalanche.com/weather/stations/hyalite-weather-station that would allow us to use historical weather records. In the absence of that, we'd look for volunteers, perphaps a citizen science initiative, to record temperatures in February. We likely aren't concerned exclusively about Saturday whether - not believing there are atmospheric conditions that lead to different weather patterns on Saturday.

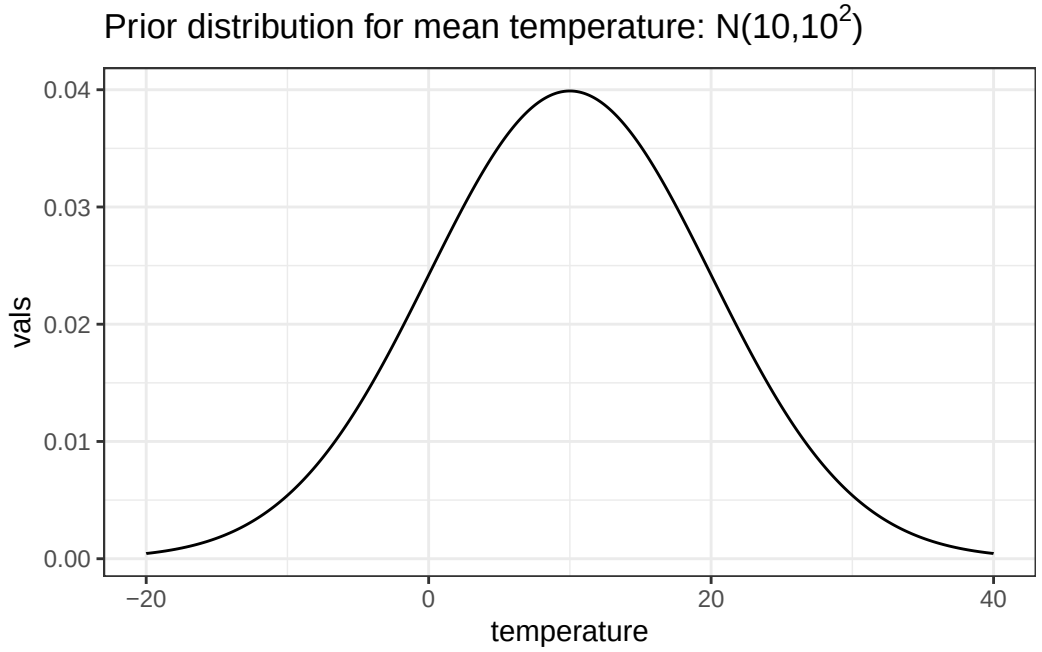
b. (4 pts) Identify a descriptive statistical model for the relevant data. Then interpret the statistical parameters in that model.

A normal distribution seems reasonable to use in this case. Temperature can go below zero, so a distribution restricting responses to positive values is not necessary. The normal distribution, commonly expressed through the idea of a bell curve, has two parameters: a mean and variance (or standard deviation). The mean temperature will give us the average temperature and Hyalite in February; whereas, the standard deviation will encode how much variability is present in temperatures.

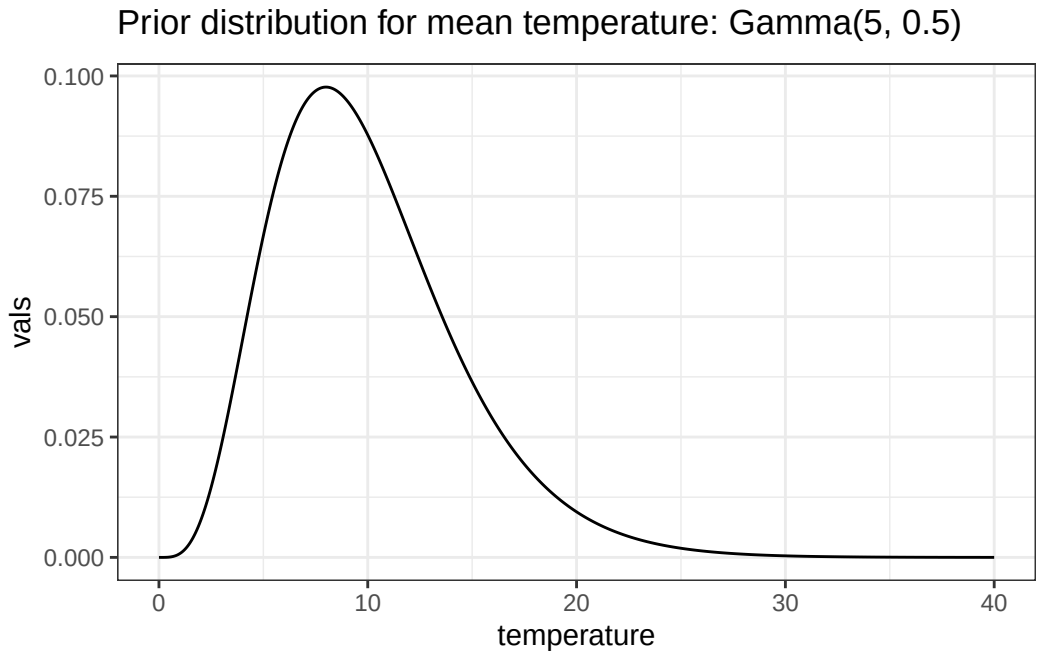
c. (4 pts) Specify a prior distribution for all parameters in the model.

Plot your distribution using ggplot2 and state the parameters in that model. Note you may have multiple parameters that require prior distributions.

```
temp_seq <- seq(-20,40, length.out = num_x)
tibble(vals = c(dnorm(temp_seq, 10, 10)),
       temperature = temp_seq) %>%
  ggplot(aes(x = temperature, y = vals)) +
  geom_line() + theme_bw() +
  ggtitle(expression(paste('Prior distribution for mean temperature: N(10,',10^2,')')))
```




```
sd_seq <- seq(0,40, length.out = num_x)
tibble(vals = c(dgamma(sd_seq, 5, .5)),
       temperature = sd_seq) %>%
  ggplot(aes(x = temperature, y = vals)) +
  geom_line() + theme_bw() +
  ggtitle(expression(paste('Prior distribution for mean temperature: Gamma(5, 0.5)')))
```



Q4. Optional (ungraded)

This class will have one project that spans the entire course. Recall you are allowed to work in groups of up to size 2 for this project. If you have any remaining course time, give some thought to identifying a dataset that is interesting to you along with an associated research question.

The project will be scaffolded over the course of the semester and all necessary statistical tools (such as regression) will be taught in this class.