Lab7

Vivian Yeh

A Simple Simulation of Waiting for a Bus

1.
```
set.seed(12345)
box<- seq(from = 0, to = 12, by = 0.01)
draws <- replicate(100000, sample(box, 1, replace = TRUE))
```

2. **Create a histogram of the approximate probability distribution of arrival time. What is the mean arrival time? The SD of arrival time?**

   The mean arrival time is 6.00mins.

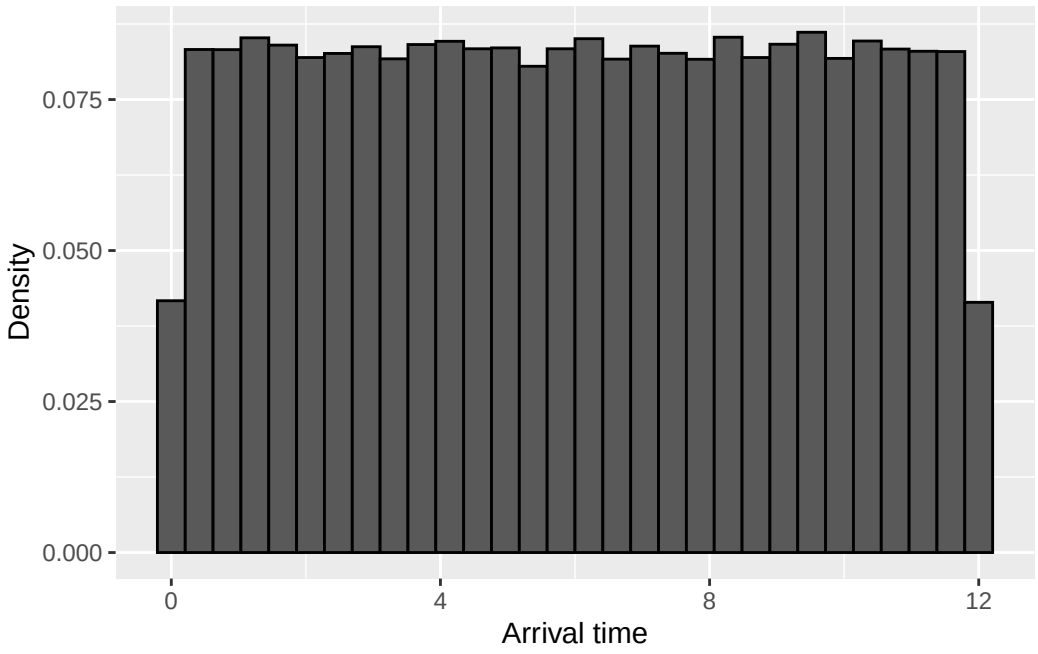
   The SD of arrival time is 3.47 mins.

```
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v ggplot2 3.3.6      v purrr   0.3.4
v tibble  3.1.8      v dplyr   1.0.10
v tidyr   1.2.1      v stringr 1.4.1
v readr   2.1.2      v forcats 0.5.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

```
ggplot(as.data.frame(draws), aes(x=draws, y=..density..)) +
  geom_histogram(color = 'black')+
  ylab("Density") +
  xlab("Arrival time")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
summarize(as.data.frame(draws), mean_arr = mean(draws), sd_arr = sd(draws))
```

```
  mean_arr   sd_arr
1 6.000291 3.465902
```

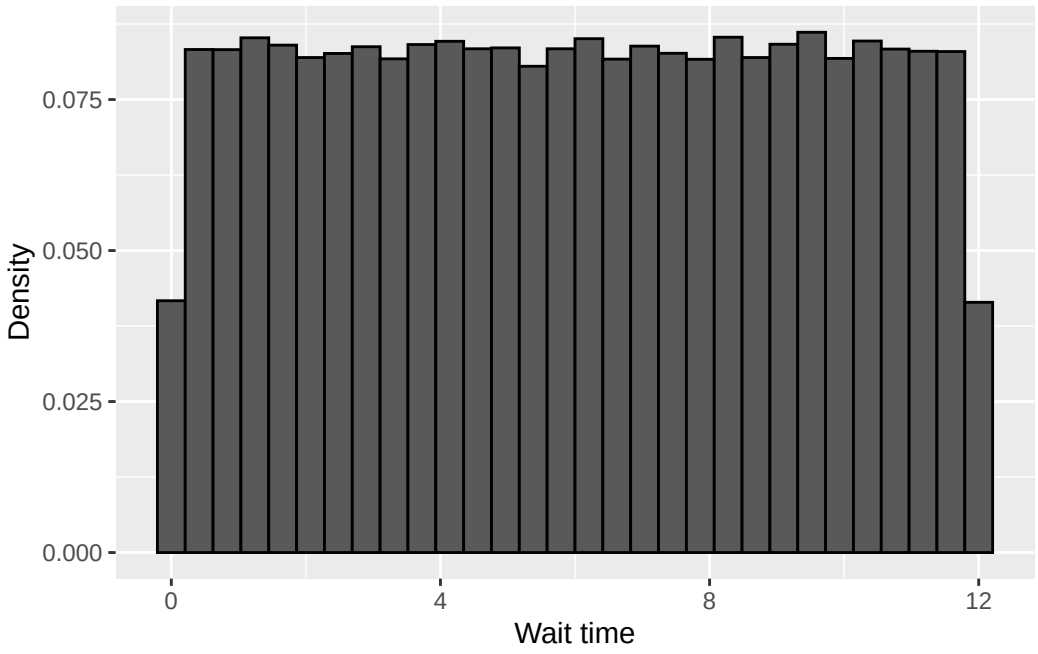
3. **Convert the simulated arrival times into wait times (recall the bus comes every 12 minutes). Repeat the computations from above to find the average wait time, standard deviation of wait time, and approximate probability distribution of the wait time.**

   The mean arrival time is 6.00mins.

   The SD of arrival time is 3.47 mins.

```
wait_times <- ggplot(as.data.frame(draws), aes(x=draws, y=..density..)) +
  geom_histogram(color = 'black')+
  ylab("Density") +
  xlab("Wait time")

wait_times
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
summarize(as.data.frame(draws), mean_arr = mean(draws), sd_arr = sd(draws))
```

```
  mean_arr   sd_arr
1 6.000291 3.465902
```

4. **Are you surprised by the size of the standard deviation? Why do you think this probability distribution is called "uniform"?**

   I am surprised by the size of the standard deviation because I thought the size will be smaller than the actual standard deviation, and the standard deviation is relatively large to the mean. I think this probability distribution is called "uniform" since at most of the time, passengers have to wait 0.1 to 11.9 minutes for the bus, and the wait time at 0 and 12 minutes are visibly less happen.

Understanding the Data for Real Buses

```
bus <- read_csv("https://tinyurl.com/bus-late-data")
```

```
Rows: 1434 Columns: 1
-- Column specification --------------------------------------------------------
Delimiter: ","
```

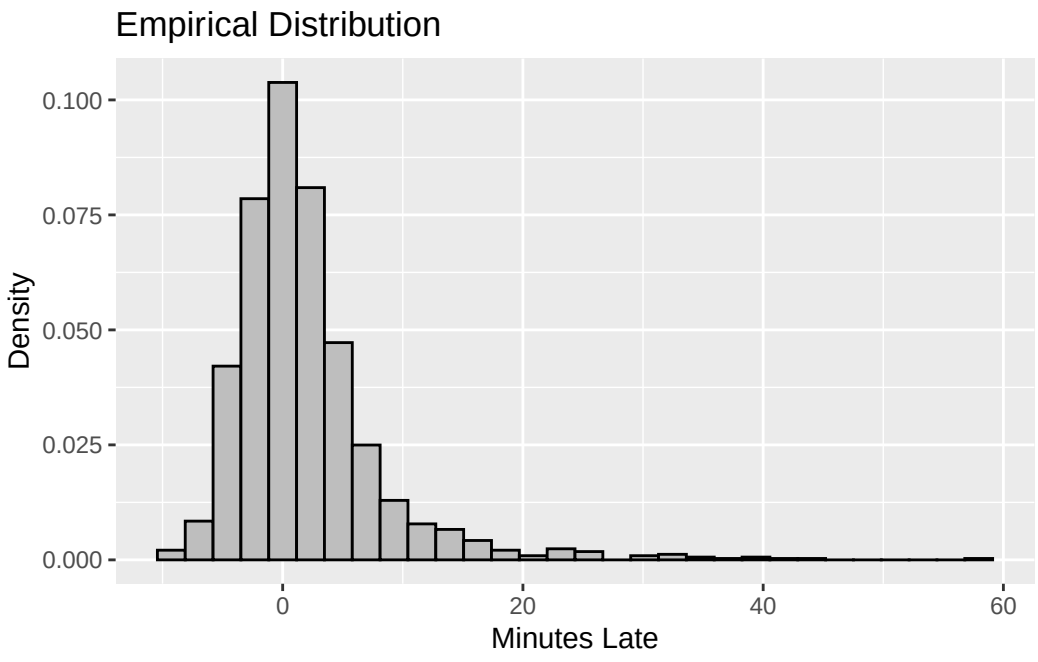
```
dbl (1): minutes_late

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

5. **Visualize and describe the empirical distribution of the lateness of the northbound C buses stopping at Third and Pike in this two-month period. Is it unimodal? Is it skewed left, skewed right, or symmetric? Are there any surprising values?**

   The empirical distribution is a right-skewed distribution. There is a surprising value at about 1.3 mins, and this means that the most bus delayed 1.3 mins.

```
emp_dist <- ggplot(bus, aes(x=minutes_late, y=..density..)) +
      geom_histogram(color="black", fill="gray") +
      xlab("Minutes Late") +
      ylab("Density") +
      ggtitle("Empirical Distribution")

emp_dist
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



6. **Summarize the typical amount of time the bus is late and compute a measure of spread to accompany your typical time.**

   The typical amount of time the bus is late is 0.74 minutes, and the standard deviation is 6.33 minutes.

```
summarize(bus, median(minutes_late), sd(minutes_late))
```

```
# A tibble: 1 x 2
  `median(minutes_late)` `sd(minutes_late)`
                   <dbl>              <dbl>
1                  0.742               6.33
```

7. **How might the distribution of the actual arrival times at the bus stop impact the waiting time experience by the blogger?**

   According to the information from the previous questions, I think the distribution will not be as flat as the distribution from the first part.

   It will be similar to the histogram of the lateness of the bus, but shifted to the left.

   I think the SD will be larger than the SD from first simulation because the actual minutes late has the larger median, and the SD will be affected.

8. **First we will create a sequence of bus arrival times. We start with times that are exactly 12-minutes apart, and then adjust them by using the observed lateness of each bus.**

```
num_stops <- nrow(bus)

sched_times <- seq(from = 12, by = 12, length = num_stops)

act_times <- sched_times + bus$minutes_late
act_times <- sort(act_times)
```

9. **Next, you will generate 100,000 arrival times of passengers at the bus stop. They still arrive at times that are equally likely and to the nearest 1/100 of a minute. But now the passengers arrive at times along the timeline from 0 to the last bus.**

```
# Set up your box
box_big <- seq(from = 0, to = max(act_times), by = 0.01)
# HINT: now your box will have over 1.7 million tickets in it!

# set the seed of the random generator so that you get the same
```

```r
# sequence of "random" values each time you run this code.
set.seed(11252022)
sample_size <- 100000

# Make draws from your box
pass_arrivals <- sample(box_big, size = sample_size, replace = TRUE)
pass_arrivals <- sort(pass_arrivals)
```

```r
k <- 1
pass_wait <- vector(mode = "numeric", length = sample_size)
for (i in (1:sample_size)) {
  for (j in k:length(act_times)) {
    if (act_times[j] >= pass_arrivals[i]) break
    }
  k <- j
  pass_wait[i] <- act_times[j] - pass_arrivals[i]
}
```

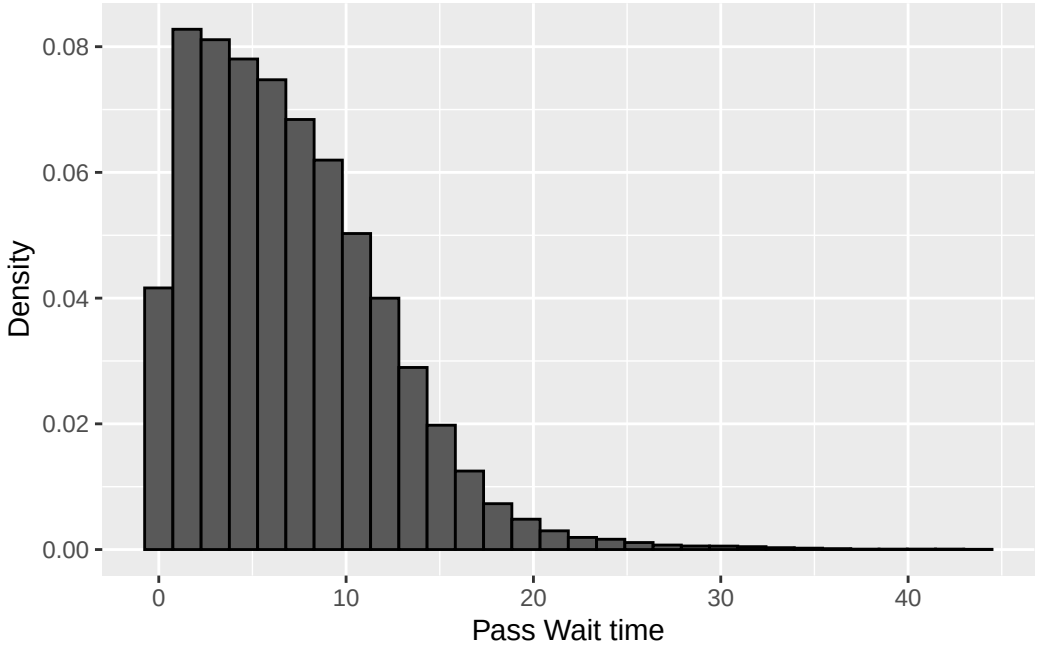
10. **What does the simulated distribution of wait times look like? Compare the distribution of pass_wait to the distribution of wait times from the first part of this lab. Also, compare the distribution to that of the observed lateness of the buses.**

    The distribution of pass_wait is extremely different from the distribution of wait times from the first part of this lab. The distributio of pass_wait is a right-skewed distribution, and the distribution of wait times is a uniform distribution.

    The shape of the distribution of pass_wait is similar to the observed lateness of the buses, both of them are right-skewed distribution.

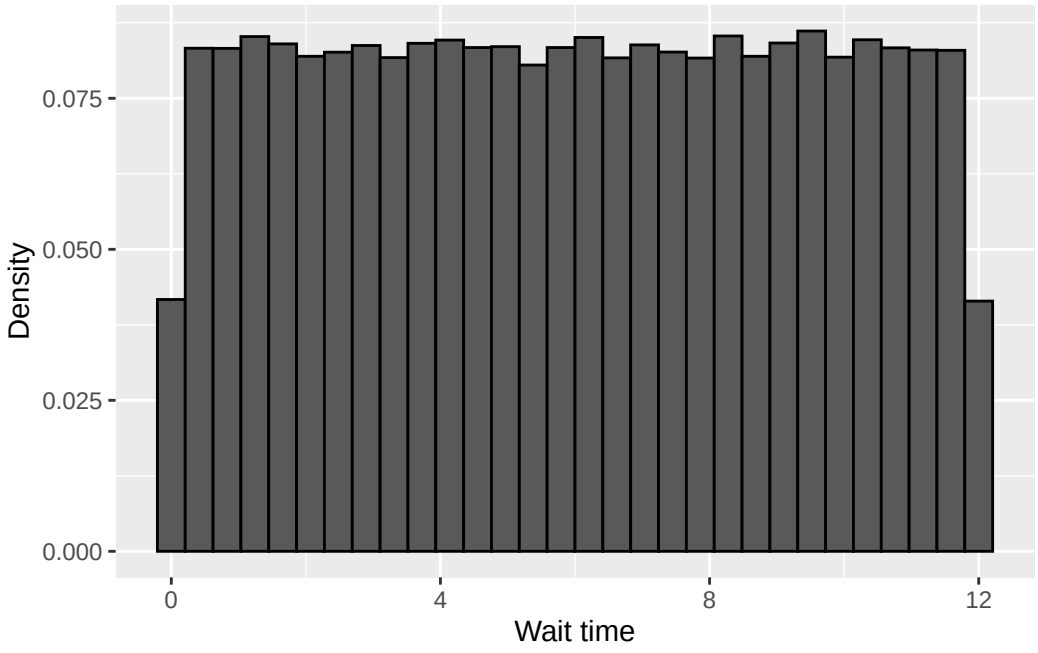
```r
library(ggplot2)
pass_wait_plot <- ggplot(as.data.frame(pass_wait), aes(x=pass_wait, y=..density..)) +
  geom_histogram(color = 'black')+
  ylab("Density") +
  xlab("Pass Wait time")

pass_wait_plot
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



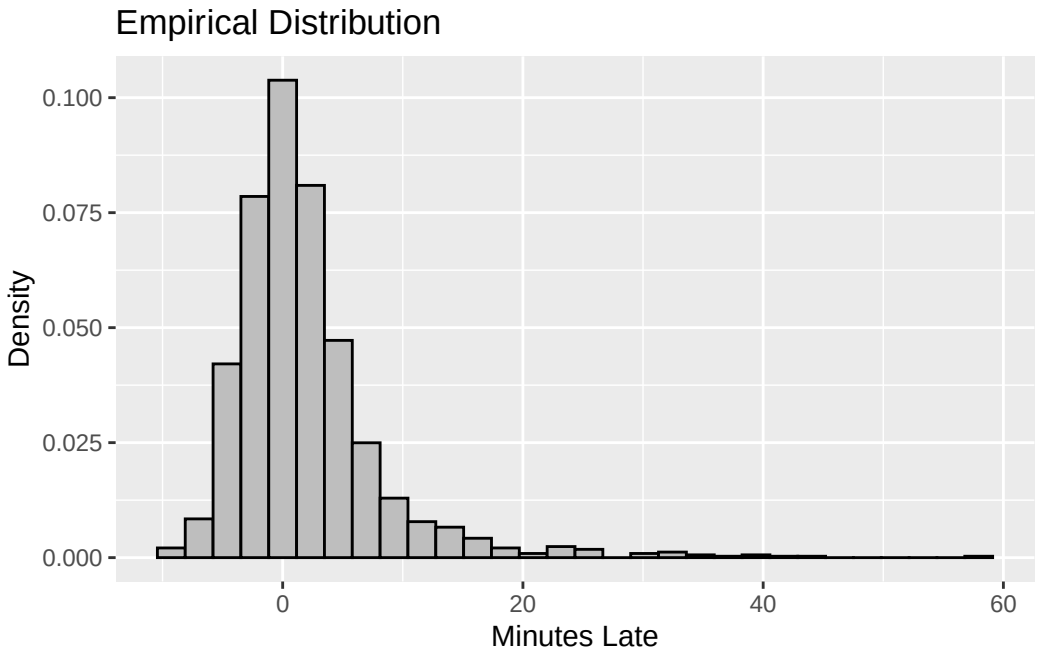

```
wait_times
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
emp_dist
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



11. **Compute the means and SDs of these three distribution. Also, compute the lower quartile, median, and upper quartile of these three distributions. Compare these statistics. Make a statement about the typical wait times, based on this final simulation. Include a statement about the deviations from this typical value.**

    For the distribution of pass_wait, the mean is 7.04, the standard deviation is 5.12, the median is 6.23, the lower quantile is 3.02, and the upper quantile is 10.04.

    For the distribution of draws, the mean is 6.00, the standard deviation is 3.47, the median is 6.01, the lower quantile is 3, and the upper quantile is 9.01.

    For the distribution of bus, the mean is 1.92, the standard deviation is 6.33, the median is 0.74, the lower quantile is -1.81, and the upper quantile is 3.81.

    Based on the final simulation, the typical wait times(the median) of the distribution of pass_wait and draws are similar which are 6.01 and 6.23. Additionally, the standard deviation of the distribution of pass_wait and draws are not equal; the sd of pass_wait is 5.12 and the sd of draws is 3.47. Although the two distribution have the similar median

and different standard deviation, the lower and upper quartiles are verly similar, which is about 3 to 10.

```
as.data.frame(pass_wait)%>%
  summarize(mean_pass_wait = mean(pass_wait),
            sd_pass_wait = sd(pass_wait),
            median_pass_wait = median(pass_wait),
            lower_pass = quantile(x = pass_wait, probs = c(0.25)),
            upper_pass = quantile(x = pass_wait, probs = c(0.75)))
```

```
  mean_pass_wait sd_pass_wait median_pass_wait lower_pass upper_pass
1       7.039072     5.118212         6.231667       3.02   10.04333
```

```
as.data.frame(draws)%>%
  summarize(mean_wait = mean(draws),
            sd_wait = sd(draws),
            median_wait = median(draws),
            lower_wait = quantile(x = draws, probs = c(0.25)),
            upper_wait = quantile(x = draws, probs = c(0.75)))
```

```
  mean_wait  sd_wait median_wait lower_wait upper_wait
1  6.000291 3.465902        6.01          3       9.01
```

```
bus%>%
  summarize(mean_bus_late = mean(minutes_late),
            sd_bus_late = sd(minutes_late),
            median_bus_late = median(minutes_late),
            lower_bus = quantile(x = minutes_late, probs = c(0.25)),
            upper_bus = quantile(x = minutes_late, probs = c(0.75)))
```

```
# A tibble: 1 x 5
  mean_bus_late sd_bus_late median_bus_late lower_bus upper_bus
          <dbl>       <dbl>           <dbl>     <dbl>     <dbl>
1          1.92        6.33           0.742     -1.81      3.81
```