BACS HW2

109062710

March 17, 2021

Question 1

Here is the helper functions for Q1

```r
standardize <- function(data) {
  standardized <- (data - mean(data)) / sd(data)
  return(standardized)
}

create_density <- function(data, title) {
  mean <- mean(data)
  ggplot(mapping = aes(data)) +
    geom_density(
      fill="#69b3a2",
      color="#e9ecef",
    ) +
    geom_vline(xintercept = mean, col="black") +
    geom_vline(xintercept = c(sd(data) * -1, sd(data)), col="red") +
    ggtitle(title)
}

create_histogram <- function(data, title) {
  n = length(data)

  # Freidman-Darconis' Binwidth Rule
  binwidth <- (2 * IQR(data)) / n^(1/3)
  bins <- ceiling(max(data) - min(data)) + binwidth

  ggplot(mapping = aes(data)) +
    geom_histogram(
      fill="#69b3a2",
      color="#e9ecef",
      bins = bins,
      binwidth = binwidth
    ) +
    ggtitle(title)
}
```

A. create a normal distribution (`mean = 940, sd = 190`) and standardize it

```
rnorm <- rnorm(1000, mean = 940, sd = 190)
rnorm_std <- standardize(rnorm)
```

i) What should we expect the mean and standard deviation of rnorm_std to be, and why?

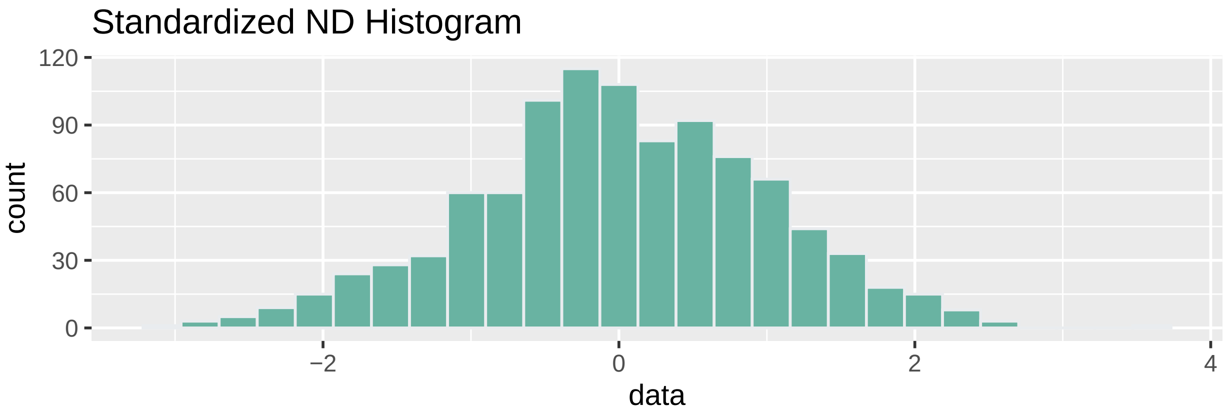
```
glue(
  "The mean of rnorm is {nonstd_rnorm_mean},
  and its standard deviation is {nonstd_rnorm_sd}."
)
```

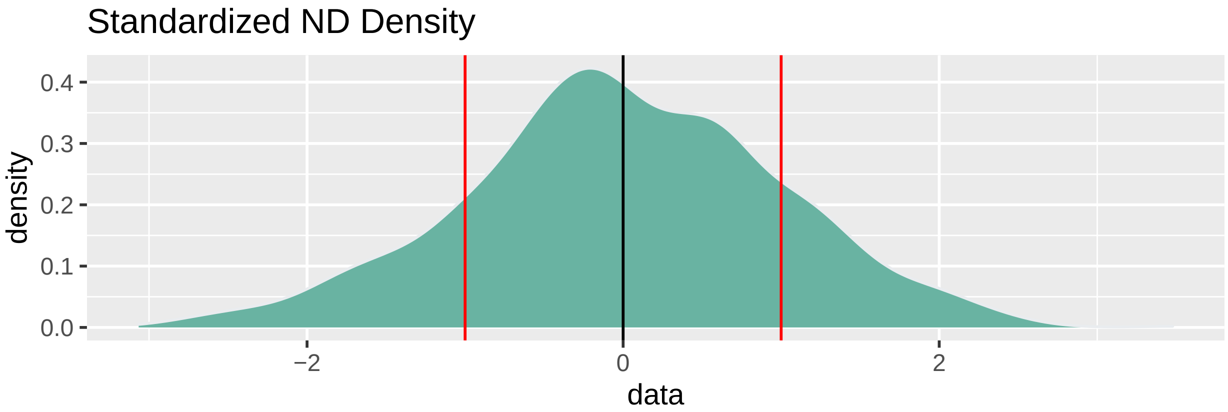
```
## The mean of rnorm is 941.968873509524,
## and its standard deviation is 195.679048005594.
```

```
glue(
  "The mean of rnorm_std is {std_rnorm_mean},
  and its standard deviation is {std_rnorm_sd}."
)
```

```
## The mean of rnorm_std is -6.83080978933215e-17,
## and its standard deviation is 1.
```

```
grid.arrange(
  std_rnorm_hist,
  std_rnorm_density,
  ncol=1,
  nrow=2
)
```

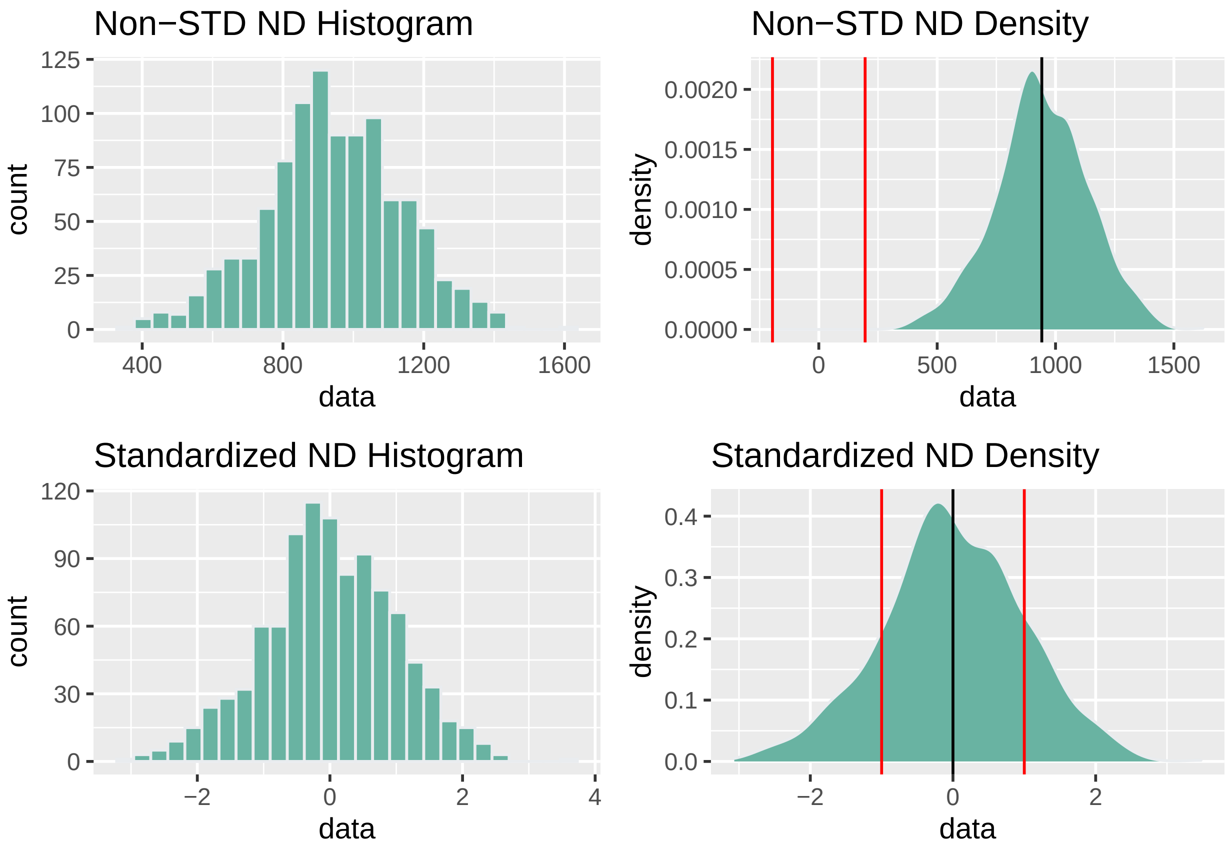




As we can see from the result above and the graph above, mean value and standard deviation value are concentrated around `-3` to `3`, instead of `0` to `1600` which before standardization. After standardization, each `x_value` in the graph represents how far each instance from the mean in STD unit. This happens because standardization scales down everything to STD unit scale.

ii) What should the distribution (shape) of rnorm_std look like, and why?

```
grid.arrange(
  nonstd_rnorm_hist,
  nonstd_rnorm_density,
  std_rnorm_hist,
  std_rnorm_density,
  ncol=2,
  nrow=2
)
```



Basically, `rnorm_std` plots should look entirely the same compared to `rnorm` plots. Let's take the graph above as our main reference. However, there are two key points worth mentioning here: 1. Non-standardized and standardized histograms look almost the same, but there is a slight difference if you take a close look. 2. The Standard Deviation lines are located in unusual location in non-standardized histogram. Unlike in standardized histogram plot, the SD lines are located at the expected locations.

iii) What do we generally call distributions that are normal and standardized?

It's called **bell-shaped curved** distribution.

B. Create a standardized version of `minday` from the earlier question (let's call it `minday_std`)

```
minday_std <- standardize(minday)
```

i) What should we expect the mean and standard deviation of `minday_std` to be, and why?

```
glue("The mean of minday_std {minday_std_mean}, while its SD is {minday_std_sd}.")
```

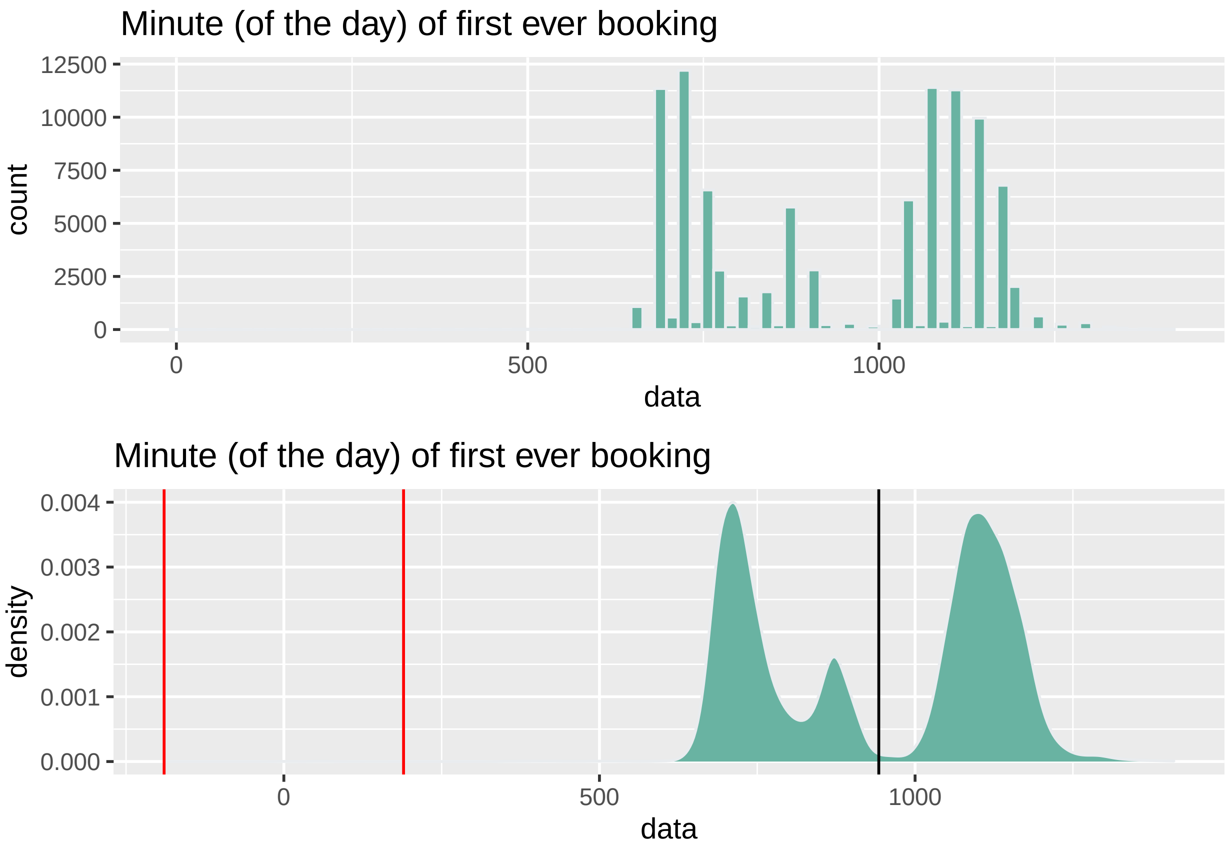
```
## The mean of minday_std -4.25589034500073e-17, while its SD is 1.
```

we expect the mean and the STD values to be really small which are within `-2.5` to `2.5` range after standardization. This happens because standardization scales down everything to STD unit scale.

ii) What should the distribution of `minday_std` look like compared to `minday`, and why?

Before standardization,

```
grid.arrange(
  minday_hist,
  minday_density,
  ncol=1,
  nrow=2
)
```

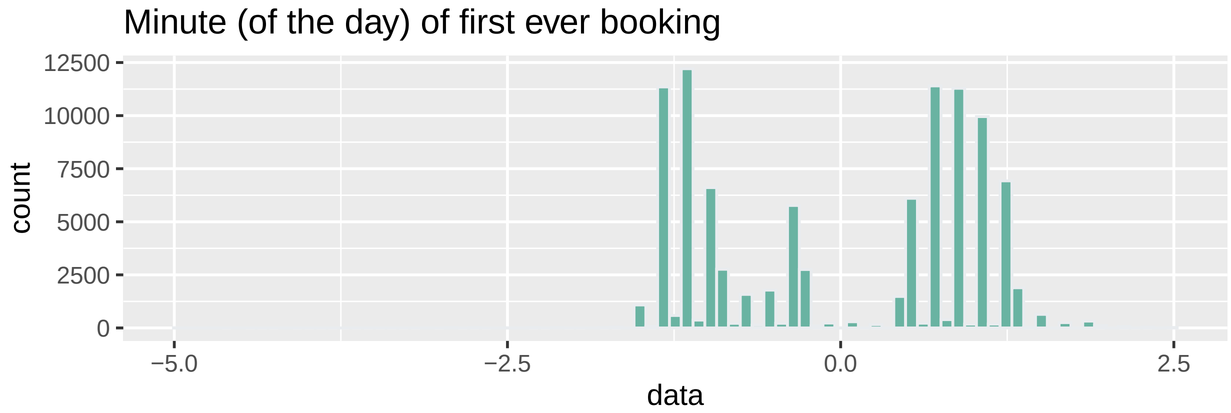


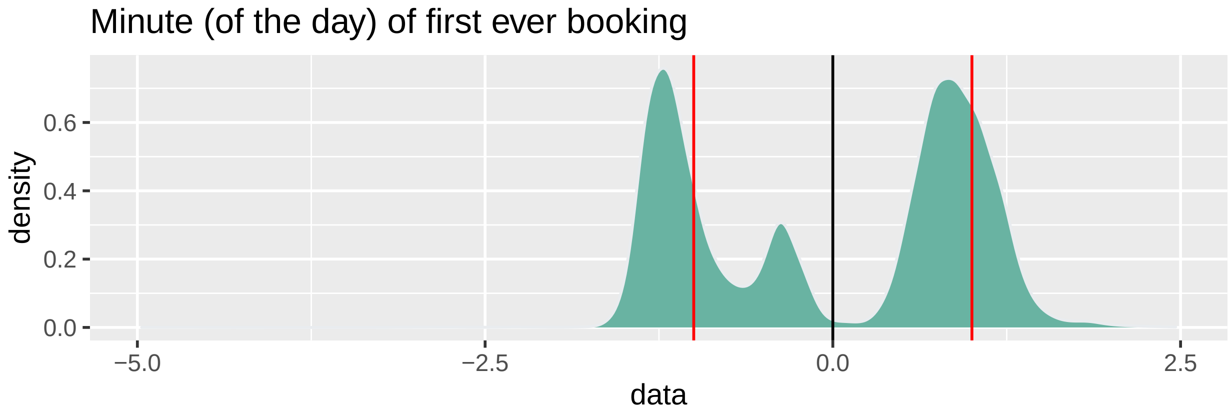
Minute (of the day) of first ever booking

Minute (of the day) of first ever booking

After standardization,

```
grid.arrange(
  minday_std_hist,
  minday_std_density,
  ncol=1,
  nrow=2
)
```





The situation is the similar to the section a, part ii. In the non-standardized data set, the STD lines are far away when we expect them to be. Besides, we have a huge range of `x_value` which is from `0` to `1500`.

However, in the standardized data set, the mean line is exactly in between the STD lines. In addition, we have a smaller range of `x_value` which is from `-4` to `4`.

Question 2

Question 3