COMPSCIX 415.2 Homework 6

Ganesh Saravanan

7/13/2018

My Github repository for all of my assignments can be found at this URL below https://github.com/gsaravanan1/rstudiodemo.git

```r
library(mdsr)
library(tidyverse)
library(ggplot2)
library(nycflights13)
library(tibble)
library(mosaicData)
```

Exercise 1

1. What variables are in this data set?

Outcome, Smoker, age

2. How many observations are there and what does each represent?

A data frame with 1314 observations on women.

- outcome:survival status after 20 years: a factor with levels Alive Dead
- smoker:smoking status at baseline: a factor with levels No Yes
- age: age (in years) at the time of the first survey

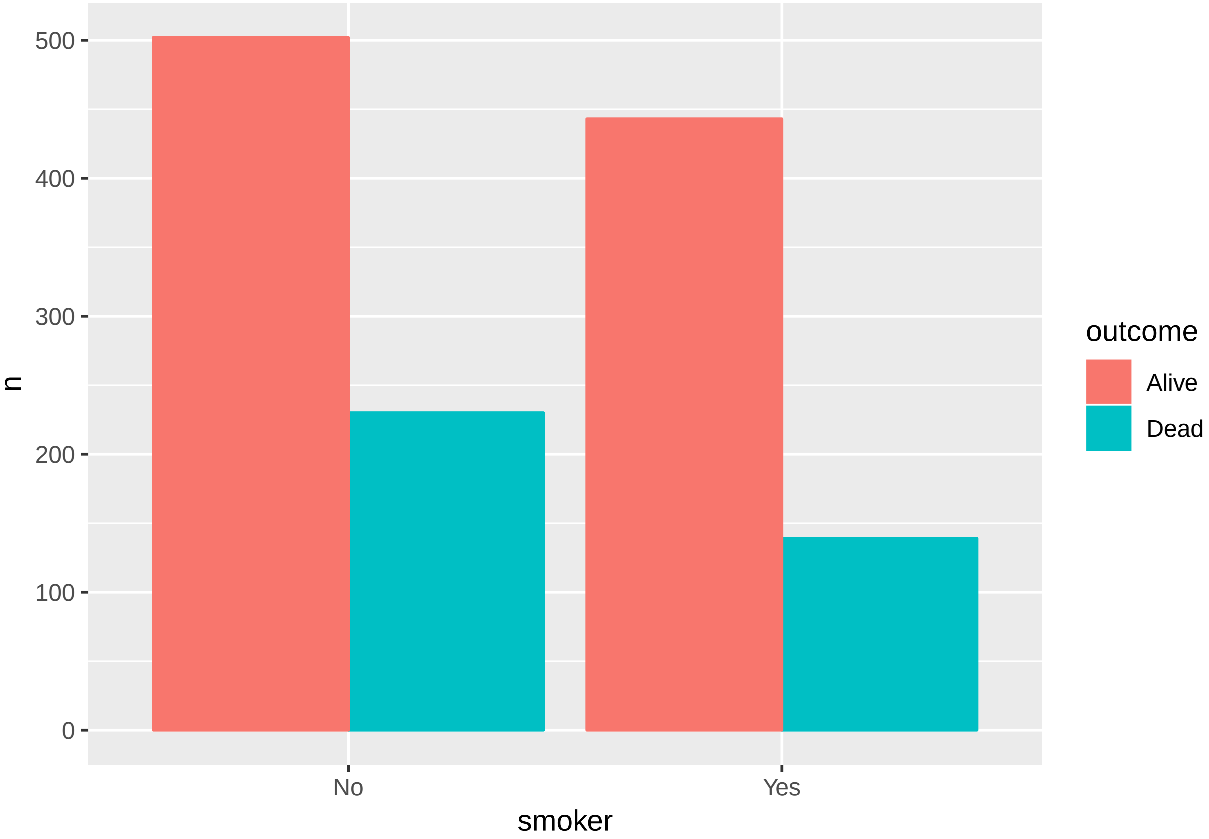
```r
library(mosaicData)
data(Whickham)
library(tidyverse)
help("Whickham")
```

3. Create a table (use the R code below as a guide)and a visualization of the relationship between smoking status and outcome, ignoring age. What do you see? Does it make sense?

```r
library(mosaicData)
library(tidyverse)
Whickham %>% count(smoker,outcome)
```

```
## # A tibble: 4 x 3
##   smoker outcome     n
##   <fct>  <fct>   <int>
## 1 No     Alive     502
## 2 No     Dead      230
## 3 Yes    Alive     443
## 4 Yes    Dead      139
```

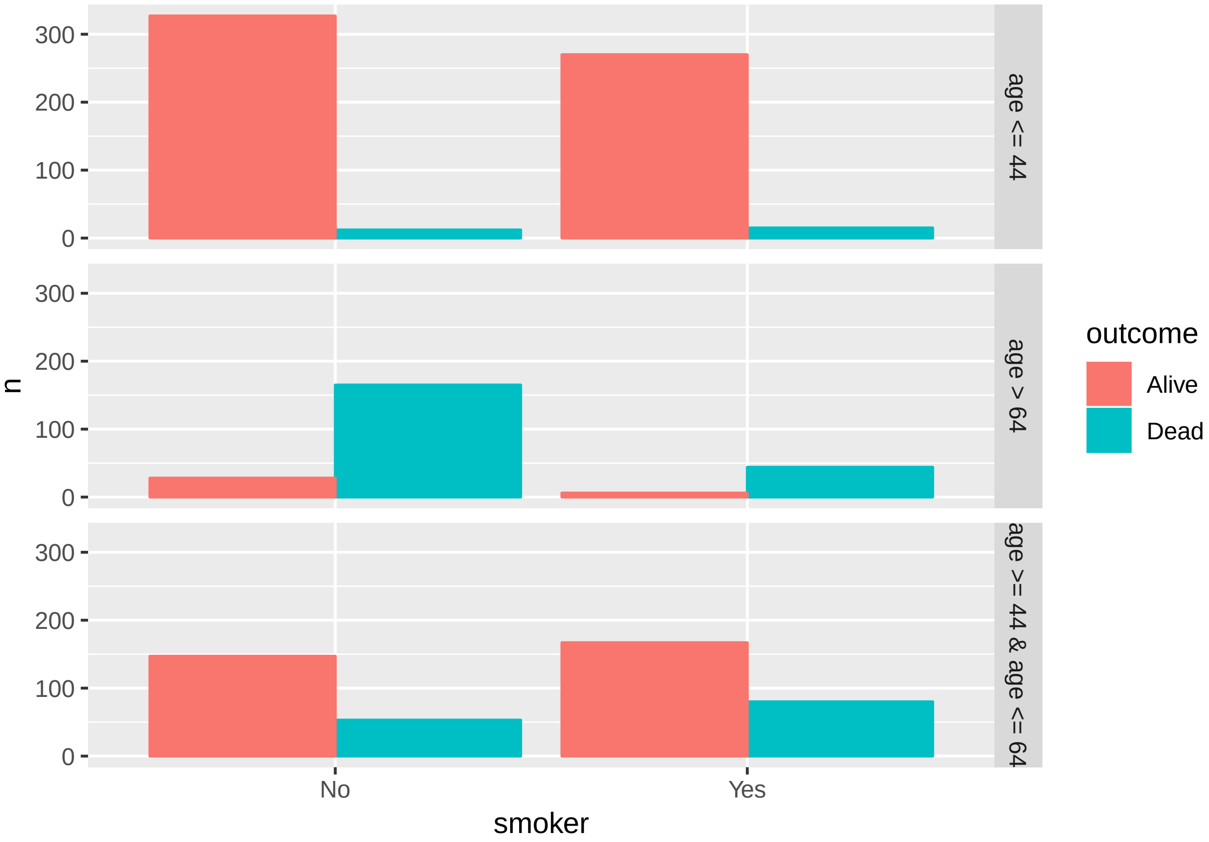
```
ggplot(data=  Whickham %>% count(smoker,outcome),
       aes(x= smoker,y = n)) + geom_bar(aes(color = outcome, fill = outcome),
                                        stat = "identity", position="dodge")
```



We can see that there are more people alive than dead regardless of the smoke status. It seems there is not much relationship between smoking status and outcome. If we look at the ratio of Dead in each group, we can see for nonsmokers, percent of Dead = 230/(230 + 502) = 31.4% and for smokers, percent of Dead = 139/(139 + 443) = 23.8%. Nonsmokers actually have higher percentage of death than smokers, which does not seem make sense.

4. Recode the age variable into an ordered factor with three categories: age <= 44, age > 44 & age <= 64, and age > 64. Now, recreate visualization from above, but facet on your new age factor. What do you see? Does it make sense?

```
Whickham <- Whickham %>% mutate(age = as.factor(ifelse(age <= 44, "age <= 44", ifelse(
  age <= 64, "age >= 44 & age <= 64", "age > 64"
))))
ggplot(data=  Whickham %>% count(smoker,outcome, age),
       aes(x= smoker,y = n)) + geom_bar(aes(color = outcome, fill = outcome),
                                        stat = "identity", position="dodge") + facet_grid(rows = vars(a
```

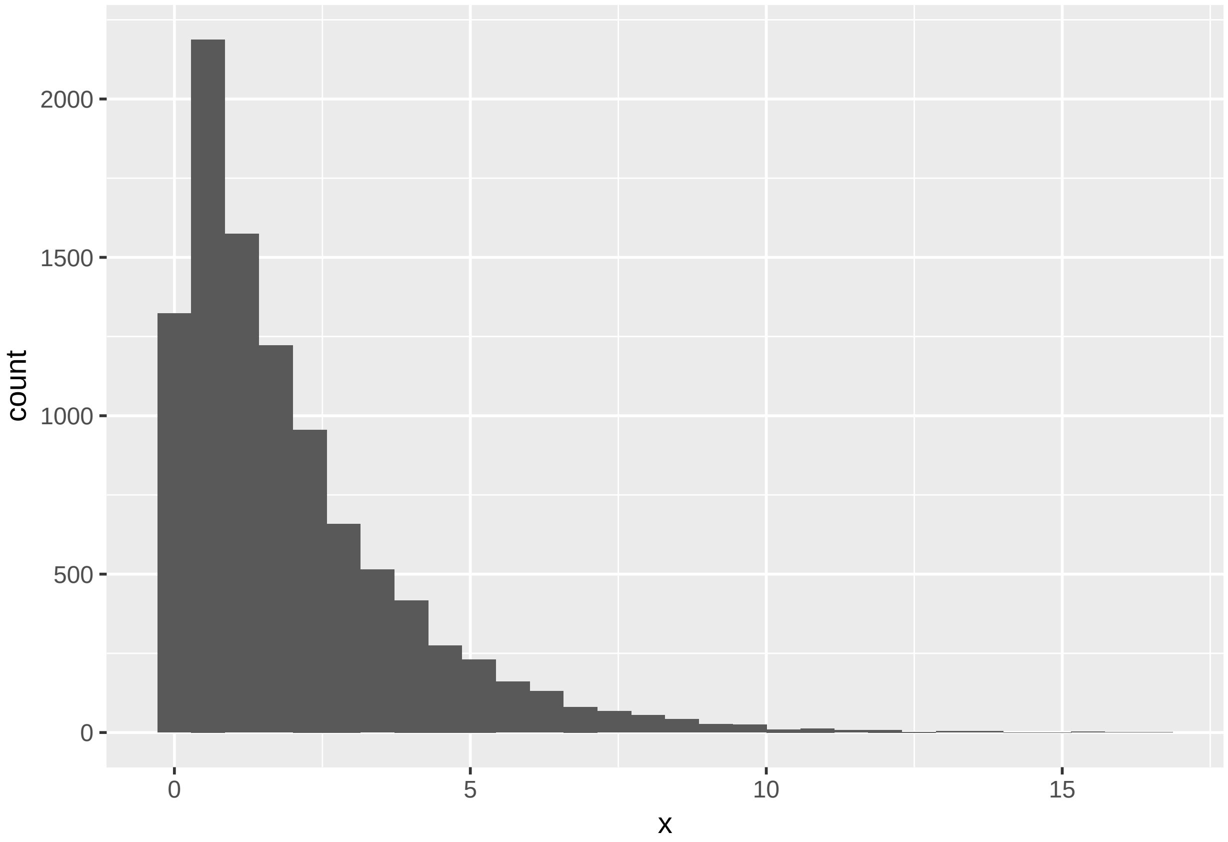


Now we can see that for people older than 64, the percentage of death is higher for smokers compared to the other 2 age groups. Also for people older than 64, percent of death for smokers = 44/(44+6) = 88% and percent of death for nonsmokers = 85%, which makes more sense.

Exercise 2

1

```
library(tidyverse)
n <- 10000
# look at ?rgamma to read about this function
gamma_samp <- tibble(x = rgamma(n, shape = 1, scale = 2))
ggplot(gamma_samp, aes(x=x)) +
 geom_histogram()
```



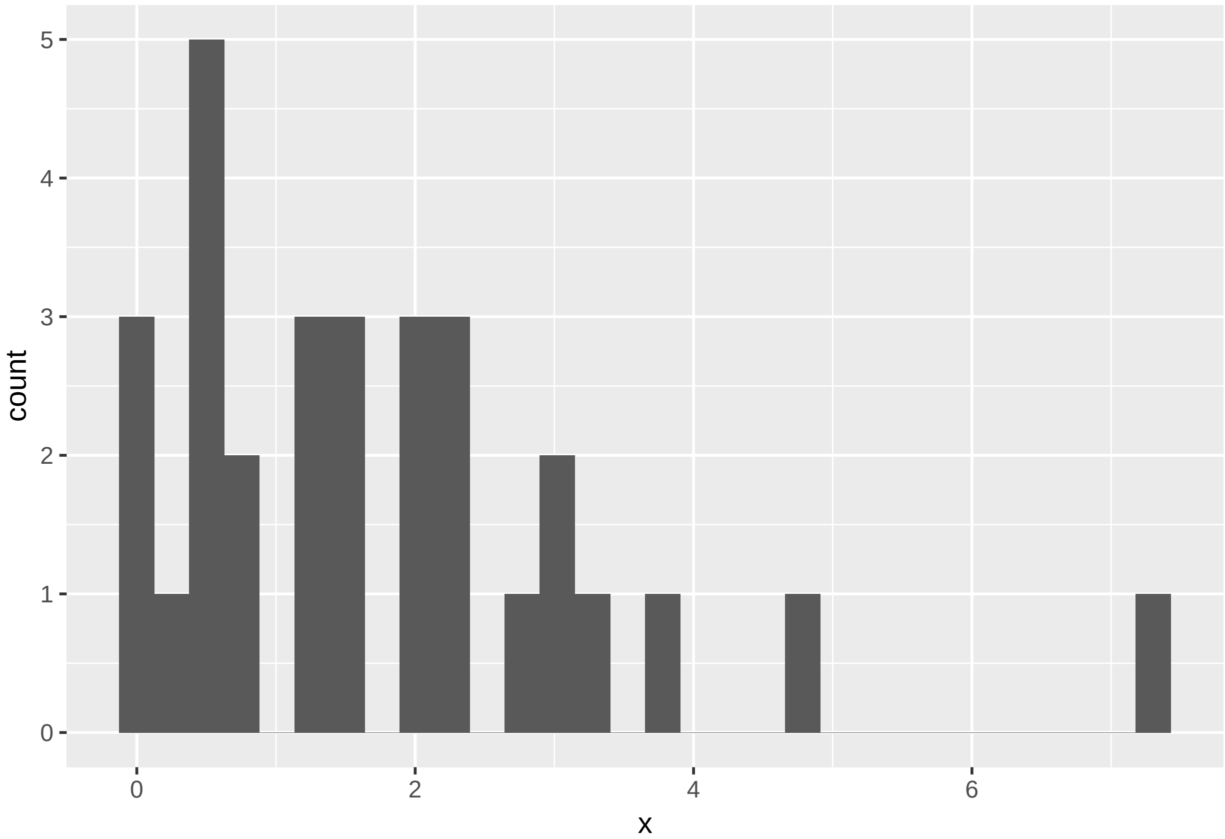
2.

```
mean_samp <- gamma_samp %>% .[['x']] %>% mean()
var_samp <- gamma_samp %>% .[['x']] %>% var() %>% sqrt()
print(paste("mean of the sample: ", round(mean_samp,3), "standard deviation of the sample: ", round(var
```

```
## [1] "mean of the sample:  1.988 standard deviation of the sample:  1.961"
```

3.

```
g_samp <- tibble(x = rgamma(30, shape = 1, scale = 2))
ggplot(g_samp, aes(x=x)) +
 geom_histogram()
```



```
mean_samp <- g_samp %>% .[['x']] %>% mean()
var_samp <- g_samp %>% .[['x']] %>% var() %>% sqrt()
print(paste("mean of the sample: ", round(mean_samp,3), "standard deviation of the sample: ", round(var_
```

```
## [1] "mean of the sample:  1.792 standard deviation of the sample:  1.594"
```

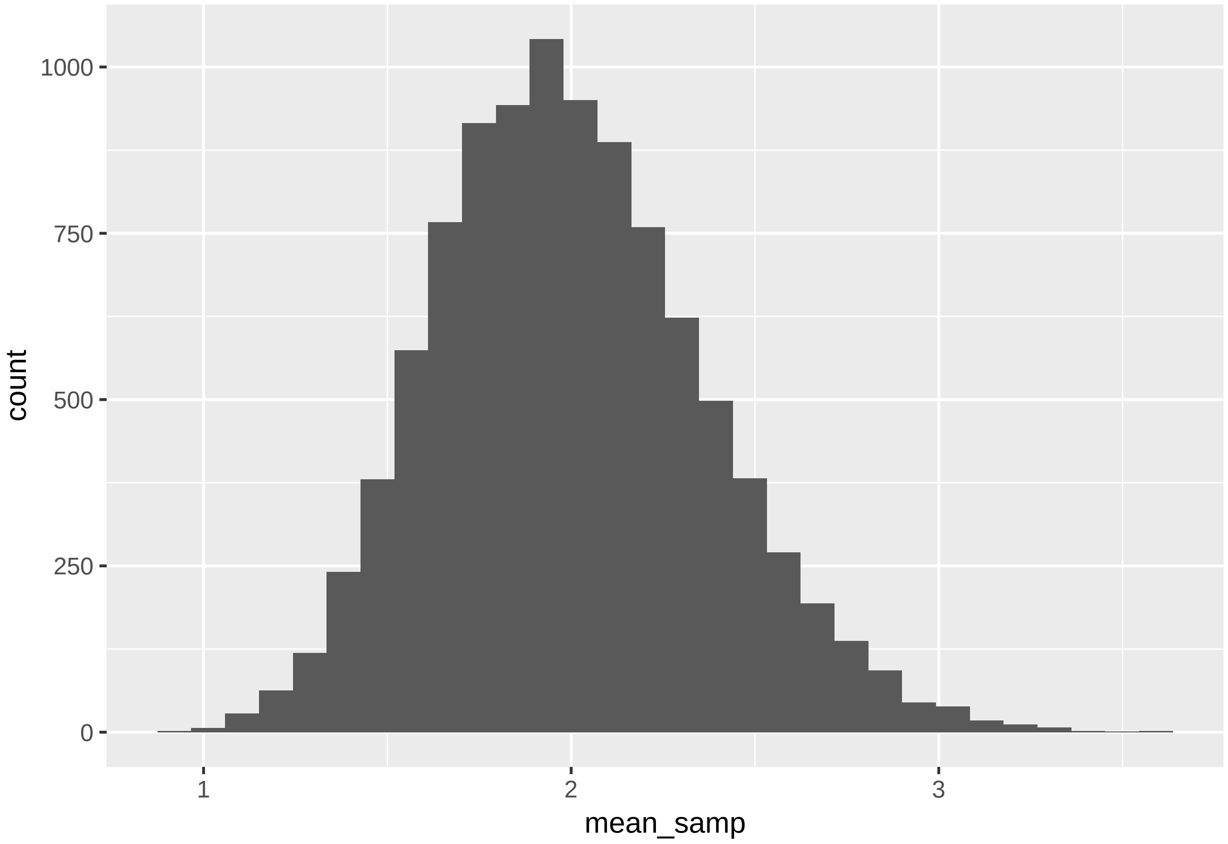
4.

```
# create a vector with 10000 NAs
mean_samp <- rep(NA, 10000)
# start a loop
for(i in 1:10000) {
  g_samp <- rgamma(30, shape = 1, scale = 2)
  mean_samp[i] <- mean(g_samp)
}
# Convert vector to a tibble
mean_samp <- tibble(mean_samp)
mean_samp
```

```
## # A tibble: 10,000 x 1
##    mean_samp
##        <dbl>
##  1      1.91
##  2      1.71
##  3      1.84
##  4      1.71
##  5      2.90
##  6      1.67
##  7      1.65
##  8      2.11
##  9      2.13
## 10      1.88
## # ... with 9,990 more rows
```

5.

```
ggplot(mean_samp, aes(x=mean_samp)) +
 geom_histogram()
```



6.

```
mean_samp_mean <- mean_samp %>% .[['mean_samp']] %>% mean()
var_samp <- mean_samp %>% .[['mean_samp']] %>% var() %>% sqrt()
print(paste("mean of the sample: ", round(mean_samp_mean,3), "standard deviation of the sample: ", roun
```

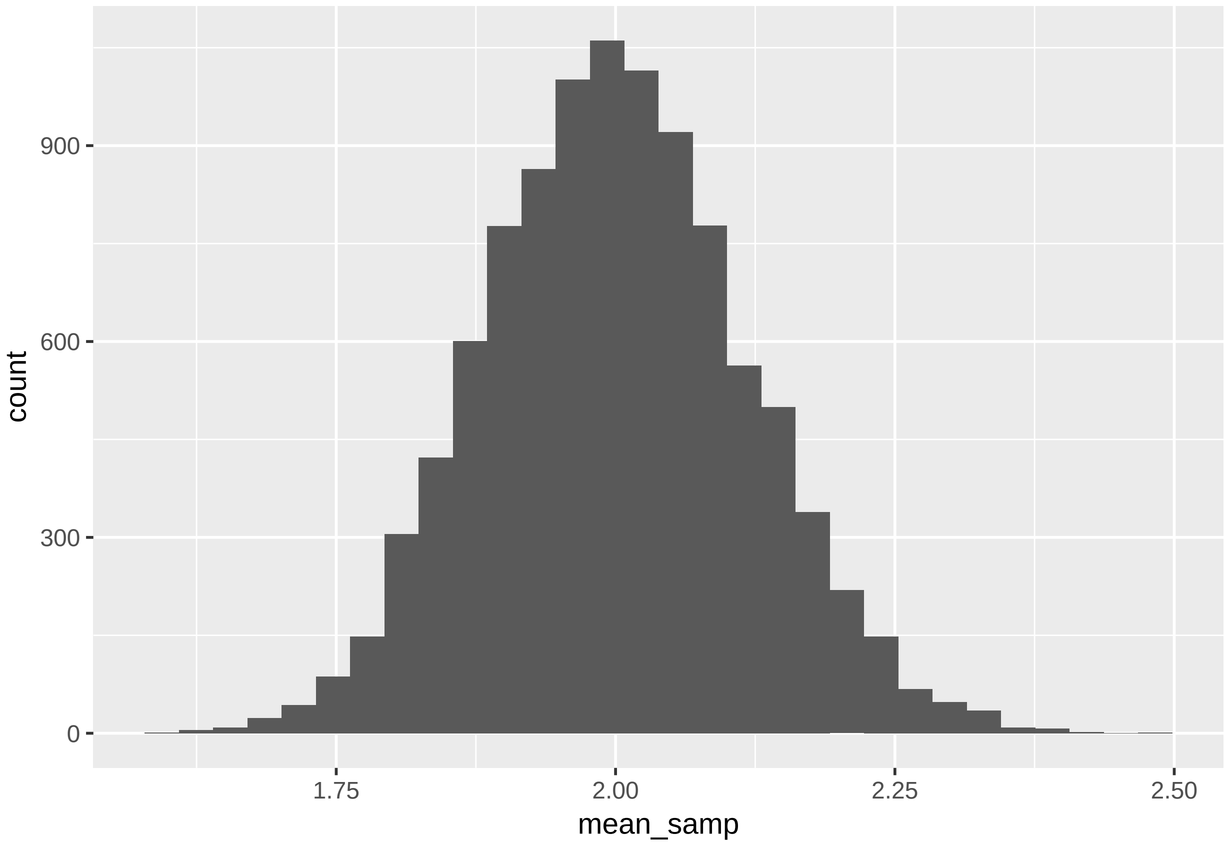
```
## [1] "mean of the sample:  1.999 standard deviation of the sample:  0.367"
```

7.

From #6, we can see that mean is close to 2, but the standard deviation is much smaller than 2.

8.

```r
# create a vector with 10000 NAs
mean_samp <- rep(NA, 10000)
# start a loop
for(i in 1:10000) {
  g_samp <- rgamma(300, shape = 1, scale = 2)
  mean_samp[i] <- mean(g_samp)
}
# Convert vector to a tibble
mean_samp <- tibble(mean_samp)
ggplot(mean_samp, aes(x=mean_samp)) +
 geom_histogram()
```



```r
mean_samp_mean <- mean_samp %>% .[['mean_samp']] %>% mean()
var_samp <- mean_samp %>% .[['mean_samp']] %>% var() %>% sqrt()
print(paste("mean of the sample: ", round(mean_samp_mean,3), "standard deviation of the sample: ", roun
```

```
## [1] "mean of the sample:  2 standard deviation of the sample:  0.117"
```

We can see the mean is close to 2, and the standard deviation is 0.116, which is close to $\sigma/\sqrt{n} = 2/\sqrt{300} =$ 0.1154