COMPSCIX 415.2 Homework 6

Michelle Gomez

7/17/2018

My Github repository for my assignments can be found at this URL:[(https://github.com/michelle-gomez/compscix-415-2-assignments)] (https://github.com/michelle-gomez/compscix-415-2-assignments)

Exercise 1

1. What variables are in this data set?
   There are 2 categorical variables (outcome, smoker) and one quantitative variable (age).

```
glimpse(Whickham)
```

```
## Observations: 1,314
## Variables: 3
## $ outcome <fct> Alive, Alive, Dead, Alive, Alive, Alive, Alive, Dead, ...
## $ smoker  <fct> Yes, Yes, Yes, No, No, Yes, Yes, No, No, No, No, Yes, ...
## $ age     <int> 23, 18, 71, 67, 64, 38, 45, 76, 28, 27, 28, 34, 20, 72...
```

2. How many observations are there and what does each represent?
   There are 1,314 observations/rows and each represents a different subject.

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?

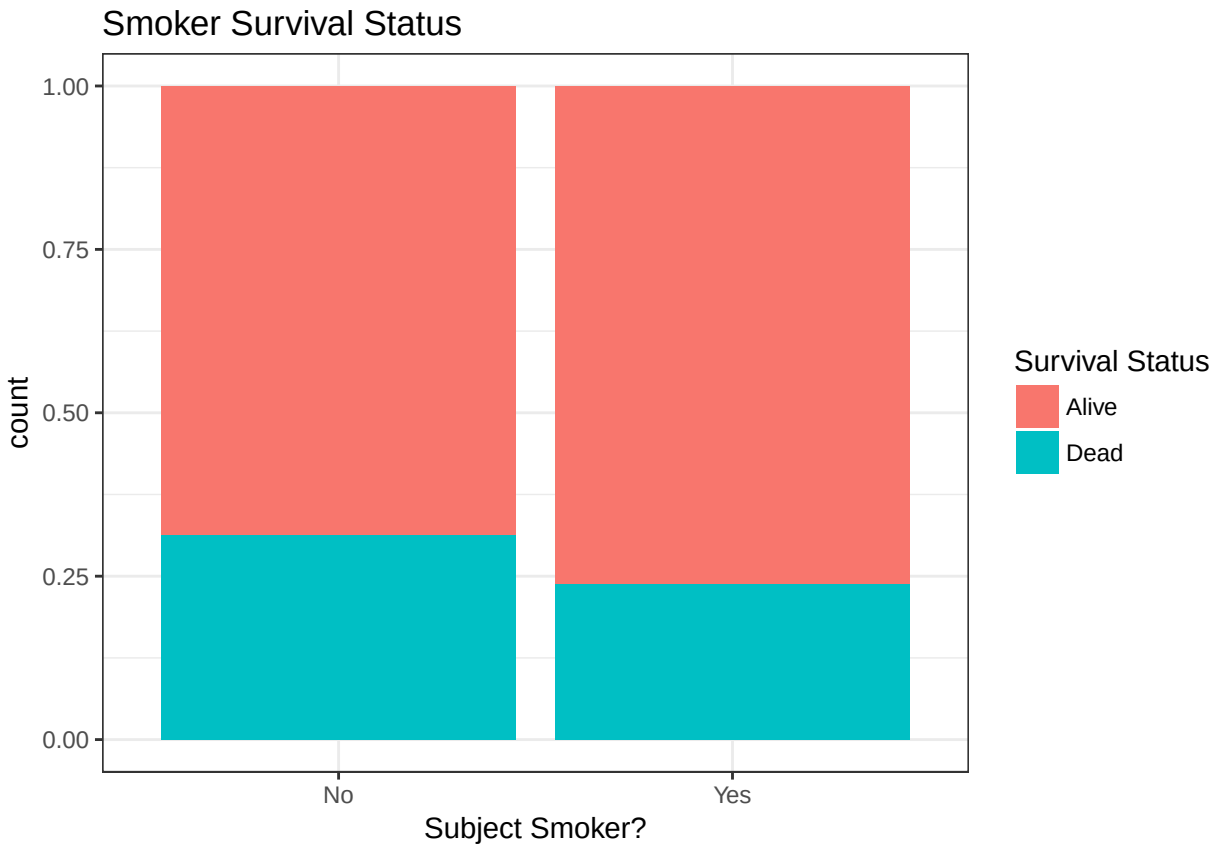
```
library(mosaicData)
library(tidyverse)
Whickham2 <-Whickham %>% count(smoker, outcome) %>%
  arrange_all() %>% mutate(n/1314)
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```
Whickham2
```

```
## # A tibble: 4 x 4
##   smoker outcome     n `n/1314`
##   <fct>  <fct>   <int>    <dbl>
## 1 No     Alive     502    0.382
## 2 No     Dead      230    0.175
## 3 Yes    Alive     443    0.337
## 4 Yes    Dead      139    0.106
```

```
Whickham2 %>%
ggplot() + geom_bar(aes(x = smoker, y = n, fill = outcome), stat = 'identity', position = 'fill') + labs
theme_bw()
```



The table makes sense but the visualization could be better. I think the issue is that there are more non-smoker subjects so this sample may not be a true representation of the population.

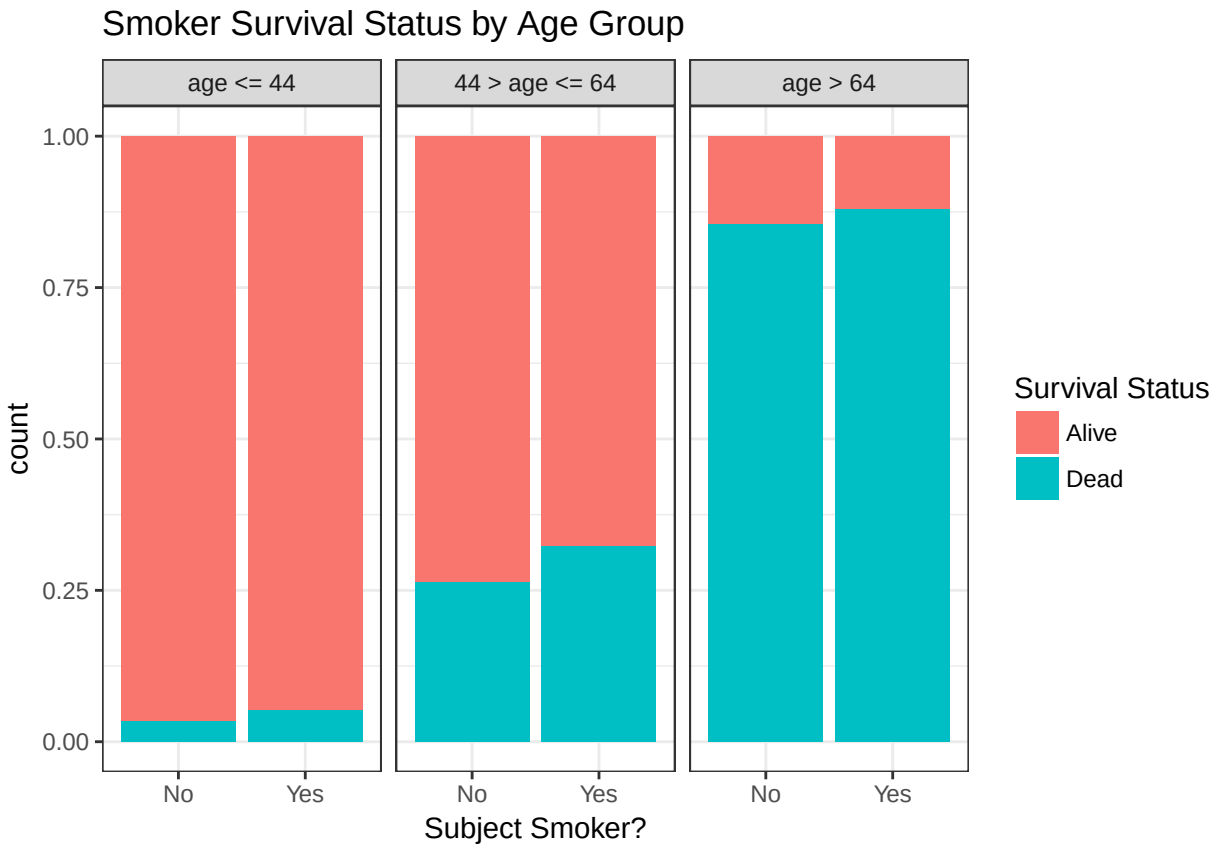
4. Recode the age variable into an ordered factor with three categories: age <= 44, age > 44 & age <= 64, and age > 64. Now, recreate visualization from above, but facet on your new age factor. What do you see? Does it make sense?
   I see 3 facets by age group, segmented my smoker status, and filled in by survival outcome. Overall, it makes sense that the older you are the less likely you are to survive, and also that if you smoke you are more likeley to die (though it doesn't seem statisticaly significant).

```
library(mosaicData)
library(dplyr)
x <- Whickham$age
cate_age <- case_when(x <= 44 ~ 'age <= 44',
                      x > 44 &  x <= 64 ~ '44 > age <= 64',
                      x > 64 ~ 'age > 64')
age_fct <- factor(cate_age, ordered = TRUE)

age_fct <- fct_relevel(age_fct, 'age <= 44', '44 > age <= 64', 'age > 64')
levels(age_fct)
```

```
## [1] "age <= 44"      "44 > age <= 64" "age > 64"
```

```
Whickham3 <- Whickham %>% mutate(age_fct) %>% count(smoker, outcome, age_fct) %>%
  arrange_all()

Whickham3 %>%
ggplot() + geom_bar(aes(x = smoker, y = n, fill = outcome), stat = 'identity', position = 'fill') +
  facet_grid(~ age_fct) +
```

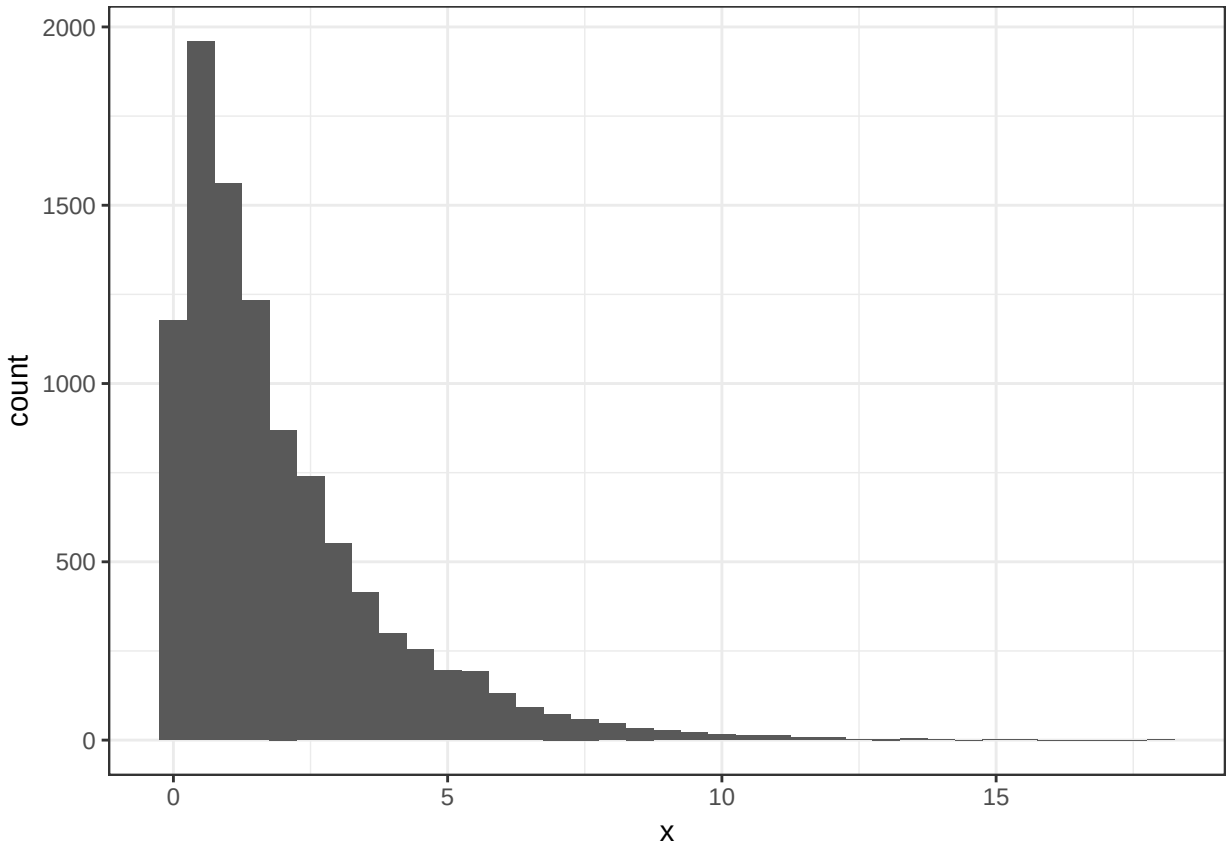
```
  labs(x = 'Subject Smoker?', y = 'count' , fill = 'Survival Status', title = 'Smoker Survival Status by
theme_bw()
```



Exercise 2

1. Generate a random sample of size n = 10000 from a gamma(1,2) distribution and plot a histogram or density curve. Use the code below to help you get your sample.

```
library(tidyverse)
n <- 10000

?rgamma
gamma_samp <- tibble(x = rgamma(n, shape = 1, scale = 2))
gamma_samp %>% sample_n(n) %>% ggplot() + geom_histogram(aes(x = x), binwidth = .5) +
  theme_bw()
```



2. What is the mean and standard deviation of your sample? They should both be close to 2 because for a gamma distribution:

```
mean_samp <- gamma_samp %>% .[['x']] %>% mean()
sd_samp <- gamma_samp %>% .[['x']] %>% sd()
mean_samp
```

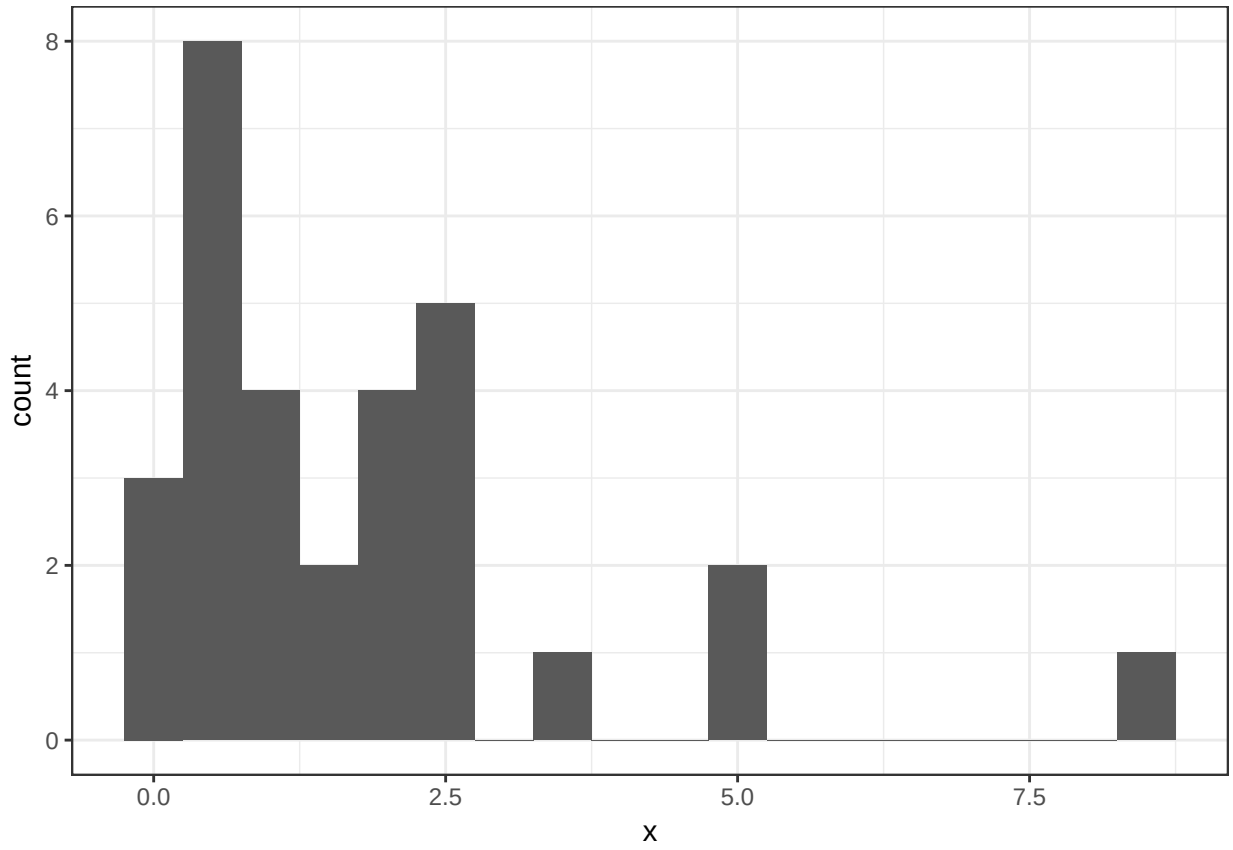
```
## [1] 1.975879
```

```
sd_samp
```

```
## [1] 1.974424
```

3. Pretend the distribution of our population of data looks like the plot above. Now take a sample of size n = 30 from a Gamma(1,2) distribution, plot the histogram or density curve, and calculate the mean and standard deviation.

```
gamma_samp %>% sample_n(30) %>% ggplot() + geom_histogram(aes(x = x), binwidth = .5) +
  theme_bw()
```




```
mean_samp <- gamma_samp %>% sample_n(30) %>% .[['x']] %>% mean()
sd_samp <- gamma_samp %>% sample_n(30) %>% .[['x']] %>% sd()
mean_samp
```

```
## [1] 2.08165
```

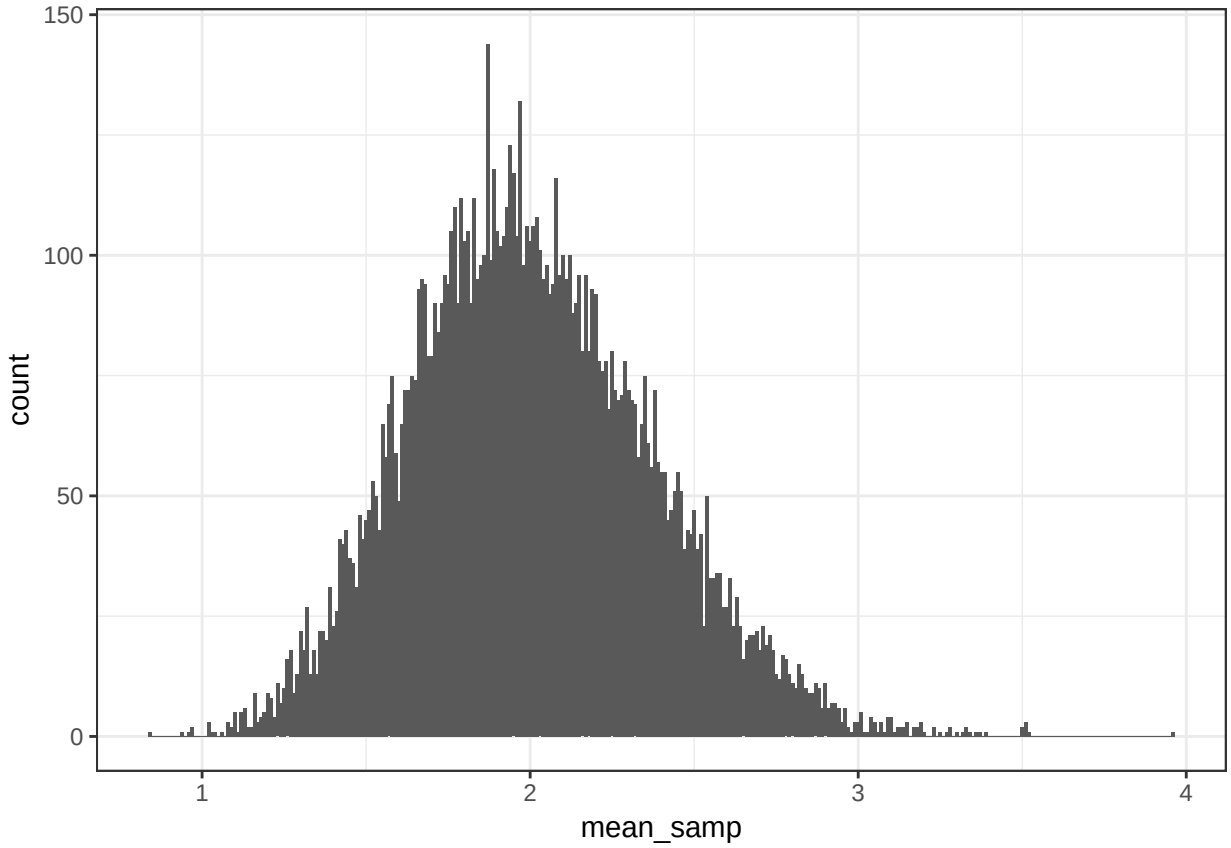
```
sd_samp
```

```
## [1] 2.397276
```

4. Take a sample of size n = 30, again from the Gamma(1,2) distribution, calculate the mean, and assign it to a vector named mean_samp. Repeat this 10000 times!!!! The code below might help.

5.

```
mean_samp %>% ggplot() + geom_histogram(aes(x = mean_samp), binwidth = .01) +
  theme_bw()
```



6.

```
mean_samp2 <- mean_samp %>% .[['mean_samp']] %>% mean()
sd_samp2 <- mean_samp %>% .[['mean_samp']] %>% sd()
mean_samp2
```

```
## [1] 2.003943
```

```
sd_samp2
```

```
## [1] 0.3690197
```

7. Did anything surprise you about your answers to #6?
   No because I expected the standard deviation to decrease.

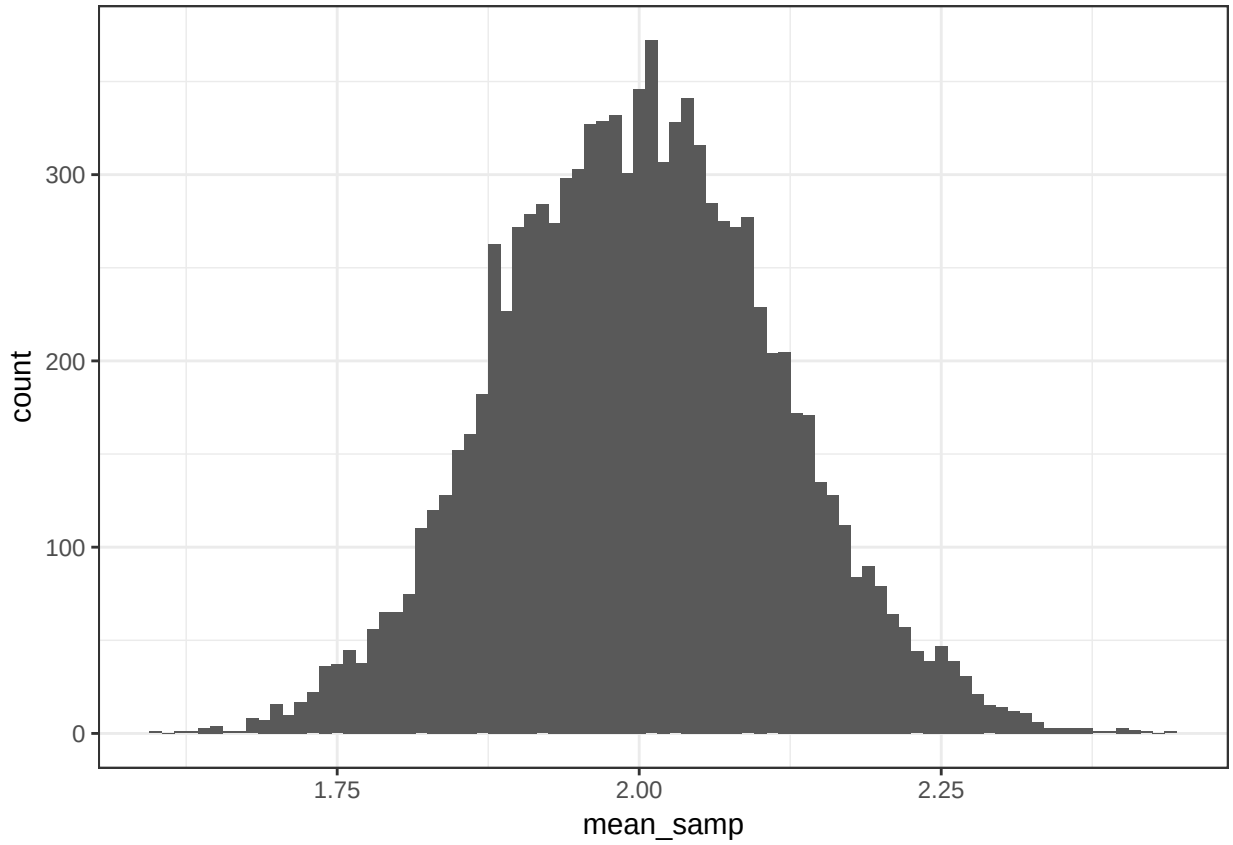
8. The results match up with the theorem.

```
mean_samp <- rep(NA, 10000)
for(i in 1:10000) {
  g_samp <- rgamma(300, shape = 1, scale = 2)
  mean_samp[i] <- mean(g_samp)
}
mean_samp <- tibble(mean_samp)
mean_samp
```

```
## # A tibble: 10,000 x 1
##    mean_samp
##        <dbl>
##  1      1.93
##  2      2.06
```

```
##  3      2.23
##  4      1.91
##  5      1.97
##  6      1.81
##  7      1.91
##  8      1.91
##  9      2.09
## 10      2.19
## # ... with 9,990 more rows
```

```
mean_samp %>% ggplot() + geom_histogram(aes(x = mean_samp), binwidth = .01) +
  theme_bw()
```




```
mean_samp2 <- mean_samp %>% .[['mean_samp']] %>% mean()
sd_samp2 <- mean_samp %>% .[['mean_samp']] %>% sd()
mean_samp2
```

```
## [1] 1.999896
```

```
sd_samp2
```

```
## [1] 0.1161324
```