“COMPSCIX 415.2 Homework 6”

Santosh Kanutala

7/17/2018

Contents

Github location

My homework assignments can be found at https://github.com/santumagic/compscix-415-2assignments.git

Exercise 1

Question: 1

```
# Load the required packages
library(tidyverse)
library(mdsr)
library(mosaicData)
# glimpse the given dataset
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...
```

Answer:
Below are the three variables from the Whickham dataset.

- outcome
- smoker
- age

Question: 2

Answer:
There are 1314 observations. Each observation represents a person. The data set indicates if the individual is a smoker or not, current age an individual, and if the individual is alive or dead.

Question: 3

```
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
```

Answer:
By looking at the above table, it is difficult to conclude anything, so I calculate the proportions first by using the below code.

```
Whickham_proportions <- Whickham %>% group_by(smoker,outcome) %>%
  summarize(n = n()) %>%
  mutate ( prop = n/sum(n))
 Whickham_proportions
```

```
## # A tibble: 4 x 4
## # Groups:   smoker [2]
##   smoker outcome     n  prop
##   <fct>  <fct>   <int> <dbl>
## 1 No     Alive     502 0.686
## 2 No     Dead      230 0.314
## 3 Yes    Alive     443 0.761
## 4 Yes    Dead      139 0.239
```

By looking at the above proportions it is observed that, 31.4 % of non smokers are dead and 23.8 % of smokers are dead which means more healthy people are dead. So there must be other reasons for deaths or data might be wrong. So the data doesn't make any sense.

Question: 4

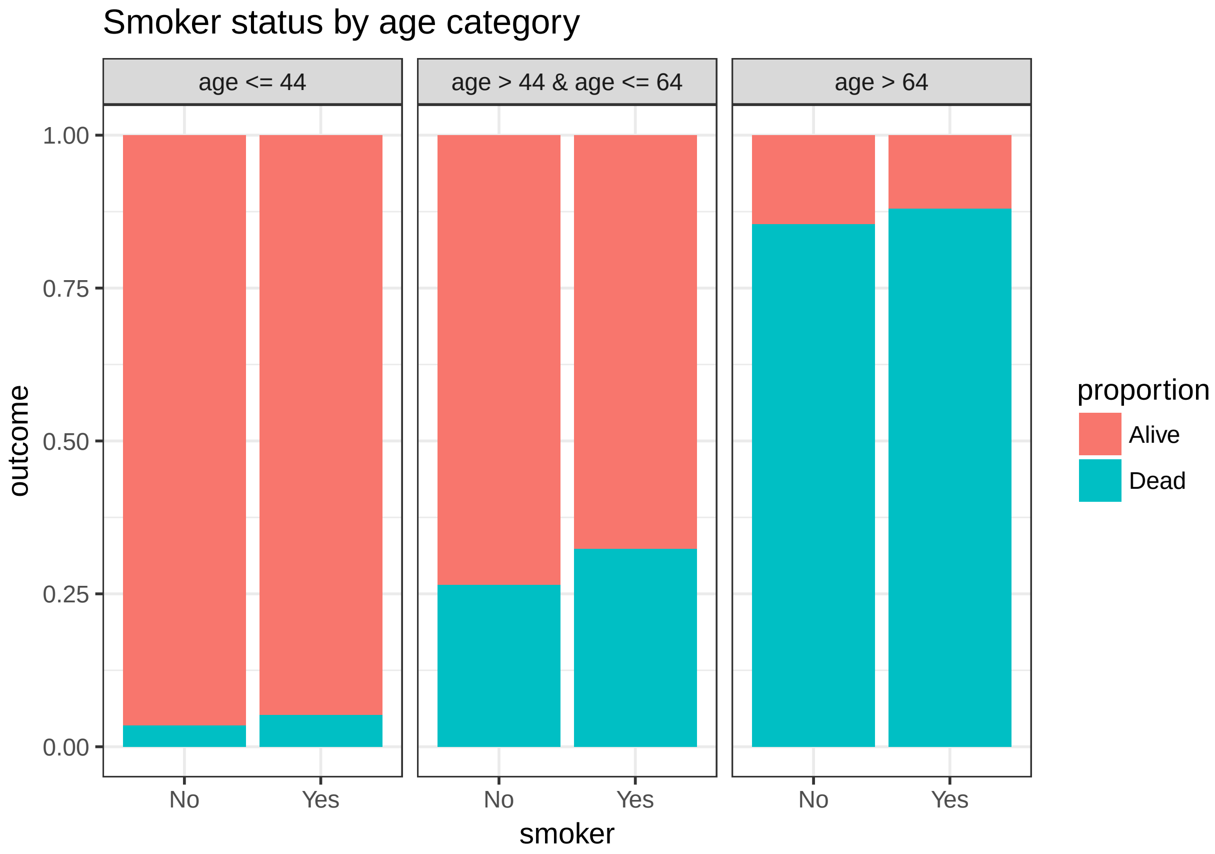
```
# creating the age groups column
Whickham_factor <- Whickham %>% mutate (category =
                                          factor (
case_when(age <= 44 ~ "age <= 44", age > 44 & age <= 64 ~ "age > 44 & age <= 64",
          age > 64 ~ "age > 64")))

head(Whickham_factor) # display the top rows of the result dataset
```

```
##   outcome smoker age             category
## 1   Alive    Yes  23            age <= 44
## 2   Alive    Yes  18            age <= 44
## 3    Dead    Yes  71             age > 64
## 4   Alive     No  67             age > 64
## 5   Alive     No  64 age > 44 & age <= 64
## 6   Alive    Yes  38            age <= 44
```

```
# reorganize the data by grouping,summarise the data and finding the proportions
Whickham_cat <- Whickham_factor %>%
group_by(category,smoker,outcome) %>% summarise( n = n()) %>%
mutate (proportion = n/sum(n))

head(Whickham_cat) # display the top rows of the result dataset
```

```
## # A tibble: 6 x 5
## # Groups:   category, smoker [3]
##   category             smoker outcome     n proportion
##   <fct>                <fct>  <fct>   <int>      <dbl>
## 1 age <= 44            No     Alive     327     0.965
## 2 age <= 44            No     Dead       12     0.0354
## 3 age <= 44            Yes    Alive     270     0.947
## 4 age <= 44            Yes    Dead       15     0.0526
## 5 age > 44 & age <= 64 No     Alive     147     0.735
## 6 age > 44 & age <= 64 No     Dead       53     0.265
```

```r
# create the visualization with the above result set and facet on age categories
Whickham_cat %>%
ggplot() +
  geom_bar(aes(x = smoker, y = proportion, fill = outcome, label = round(proportion,2)),
           stat = 'identity', position = 'fill') +
  labs(x = 'smoker', y = 'outcome', fill = 'proportion',
       title = 'Smoker status by age category') +
  facet_grid(~ category) +
  theme_bw()
```



Answer:
From the above dataset it is observed that untill the age of 44, non-smokers have only a 1% advantage compared to smokers, but this gap increases dramatically between the ages 44 & 64 where 6% more non-smokers are alive than smokers. Beyond 64 year of age, the difference drops to 3% with non-smokers still being alive more often than smokers.

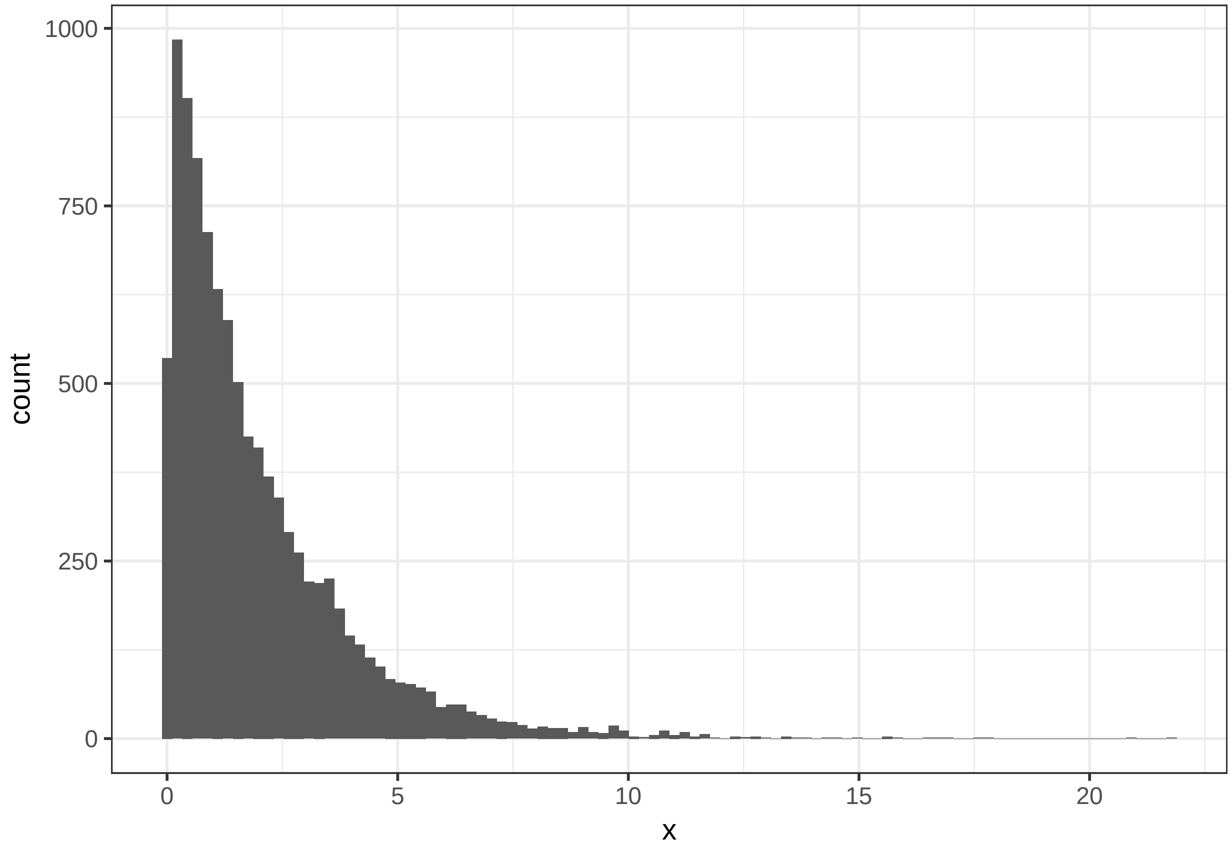
Exercise 2

Question: 1

```r
# given sample code
library(tidyverse)
n <- 10000
# look at ?rgamma to read about this function
gamma_samp <- tibble(x = rgamma(n, shape = 1, scale = 2))
gamma_samp
```

```
## # A tibble: 10,000 x 1
##        x
##    <dbl>
##  1 2.35
##  2 6.23
##  3 3.68
##  4 2.10
##  5 3.35
##  6 3.19
##  7 2.39
##  8 0.795
##  9 1.88
## 10 4.79
## # ... with 9,990 more rows
```

```r
# histogram for the above sample gamma
ggplot(data = gamma_samp) +
  geom_histogram(aes(x=x), bins=100) +
  theme_bw()
```



Question: 2

```
# finding mean
sprintf ("Mean = %f",sapply(gamma_samp, mean, na.rm = TRUE))
```

```
## [1] "Mean = 1.980503"
```

```
# finding standard deviation
sprintf ("Standard deviation = %f", sapply(gamma_samp, sd, na.rm = TRUE))
```

```
## [1] "Standard deviation = 2.017767"
```

Question: 3

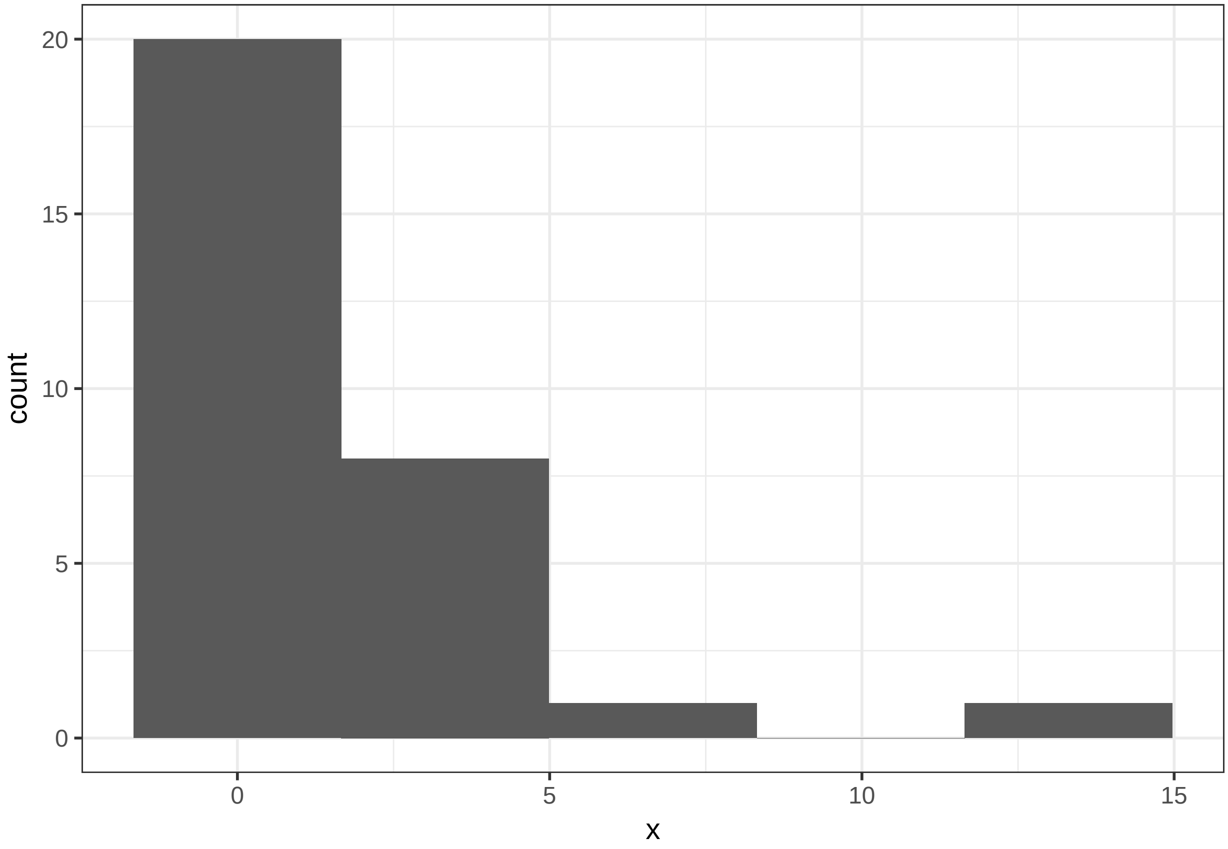
```
# sample of size n = 30
sample_30 <- gamma_samp %>% sample_n(30, replace = TRUE)

# finding mean
sprintf("Mean = %f", sapply(sample_30, mean, na.rm = TRUE))
```

```
## [1] "Mean = 1.954935"
```

```
# finding standard deviation
sprintf ("Standard deviation = %f", sapply(sample_30, sd, na.rm = TRUE))
```

```
## [1] "Standard deviation = 2.625018"
```

```
# Plot the histogram
ggplot(data =sample_30, mapping = aes(x=x)) +
  geom_histogram(bins=5) +
  theme_bw()
```



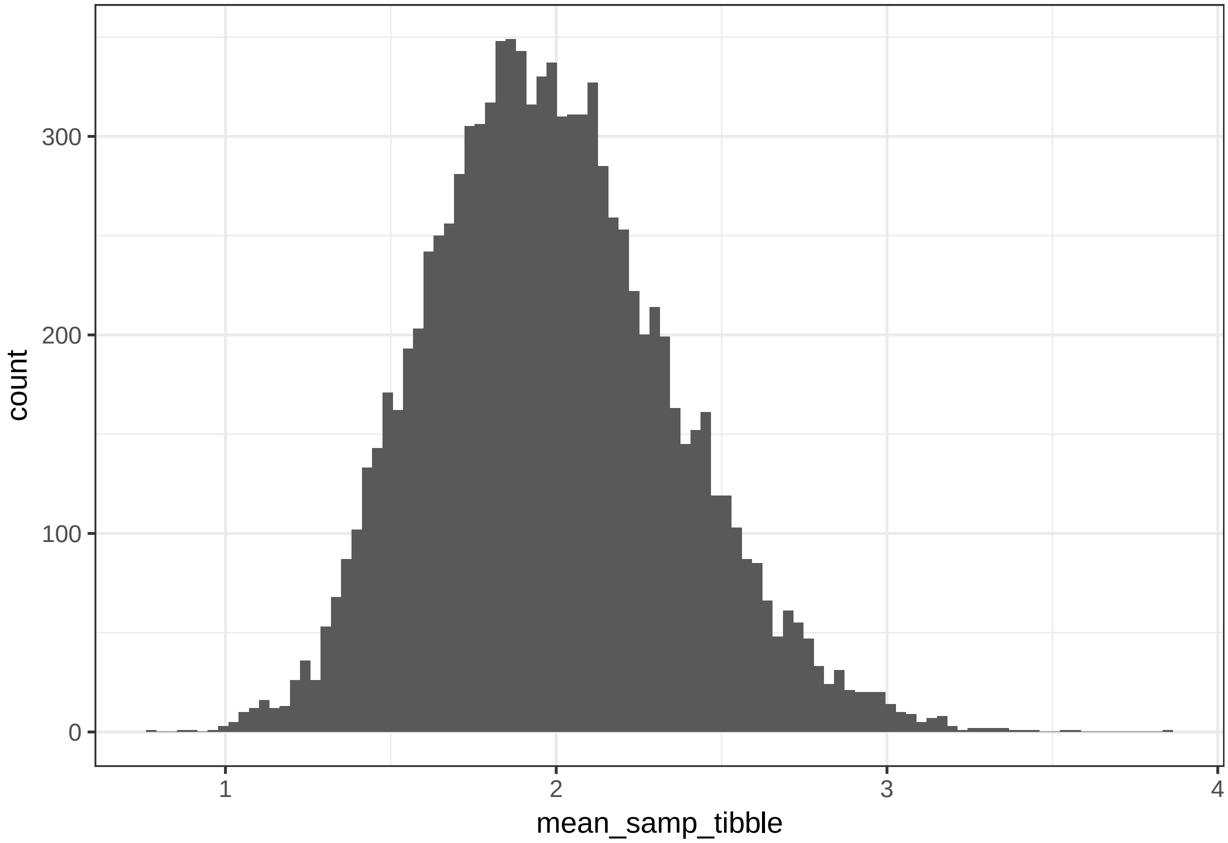
Question: 4

```
mean_samp <- rep(NA, 10000)

mean_sd <- rep(NA, 10000)
for (i in 1:10000) {
g_samp <- gamma_samp %>%
  sample_n(30, replace = TRUE)
mean_samp[i] <- mean(g_samp$x)
mean_sd[i] <- sd(g_samp$x)
}
# tibbles for mean_samp and mead_sd
mean_samp_tibble <- tibble(mean_samp)
mean_sd_tibble <- tibble(mean_sd)
mean_dist <- bind_cols(mean_samp_tibble,mean_sd_tibble)

# display the sample of final vector
head(mean_dist)
```

```
## # A tibble: 6 x 2
##   mean_samp mean_sd
##        <dbl>   <dbl>
## 1       2.40    2.22
## 2       2.25    2.31
## 3       1.05    1.43
## 4       1.39    1.66
## 5       1.48    1.63
## 6       1.73    1.45
```

Question: 5

```
# plot for the means
ggplot(data = mean_dist, mapping = aes(x=mean_samp_tibble)) +
  geom_histogram(bins=100) +
  theme_bw()
```

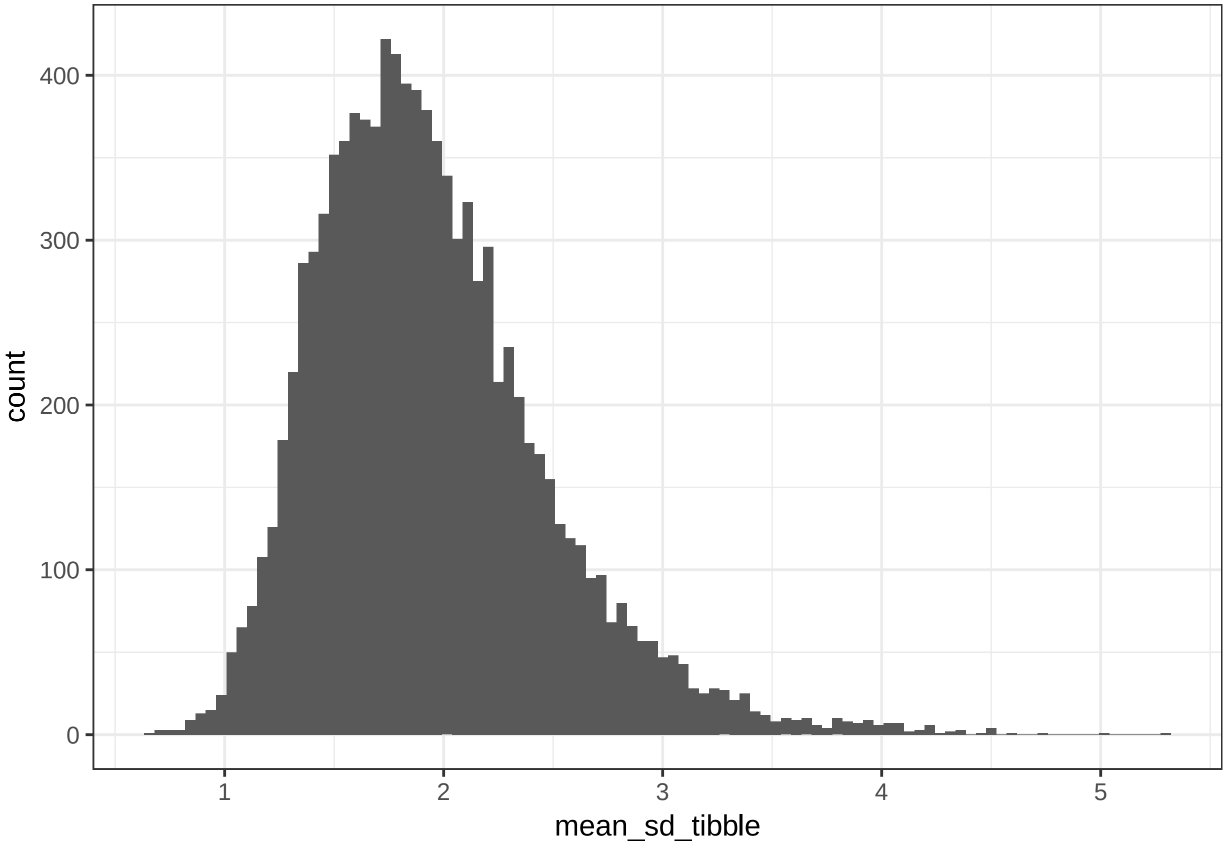


Question: 6

```
# from the code of Question 4 : below are sample of the standard deviations of the population
head(mean_sd_tibble)
```

```
## # A tibble: 6 x 1
##   mean_sd
##     <dbl>
## 1    2.22
## 2    2.31
## 3    1.43
## 4    1.66
## 5    1.63
## 6    1.45
```

```
# plot for the standard deviation
ggplot(data = mean_dist, mapping = aes(x=mean_sd_tibble)) +
  geom_histogram(bins=100) +
  theme_bw()
```



Question: 7

Answer: Both the mean and sd plots of the population looks normally distributed.

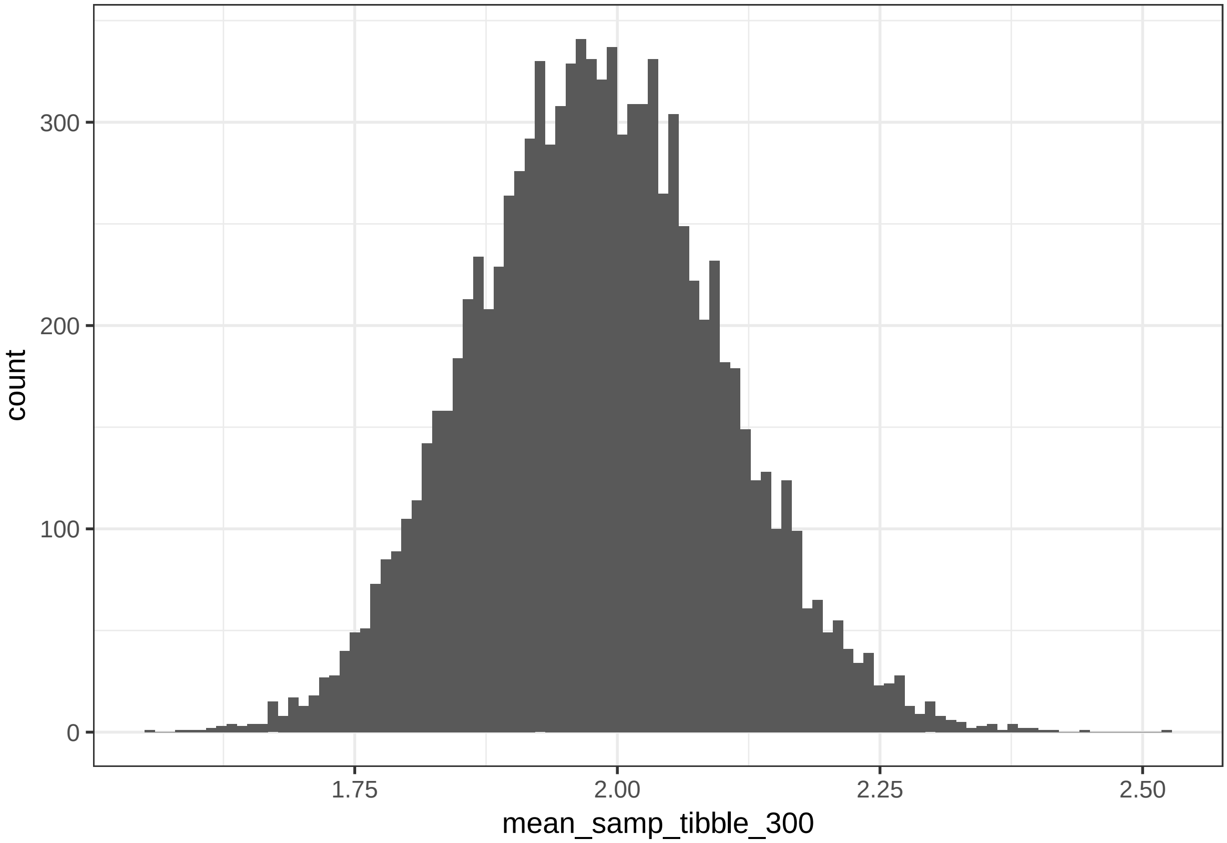
Question: 8

```
# Repeat Question 4 with sample size = 300
mean_samp_300 <- rep(NA, 10000)

mean_sd_300 <- rep(NA, 10000)
for (i in 1:10000) {
g_samp_300 <- gamma_samp %>%
  sample_n(300, replace = TRUE)
mean_samp_300[i] <- mean(g_samp_300$x)
mean_sd_300[i] <- sd(g_samp_300$x)
}
```

```
# tibbles for mean_samp and mead_sd
mean_samp_tibble_300 <- tibble(mean_samp_300)
mean_sd_tibble_300 <- tibble(mean_sd_300)
mean_dist_300 <- bind_cols(mean_samp_tibble_300,mean_sd_tibble_300)
# display the sample of final vector
head(mean_dist_300)
```

```
## # A tibble: 6 x 2
##   mean_samp_300 mean_sd_300
##           <dbl>       <dbl>
## 1          1.91        1.91
## 2          1.99        2.05
## 3          2.12        2.32
## 4          1.89        1.82
## 5          1.68        1.62
## 6          2.05        2.22
```

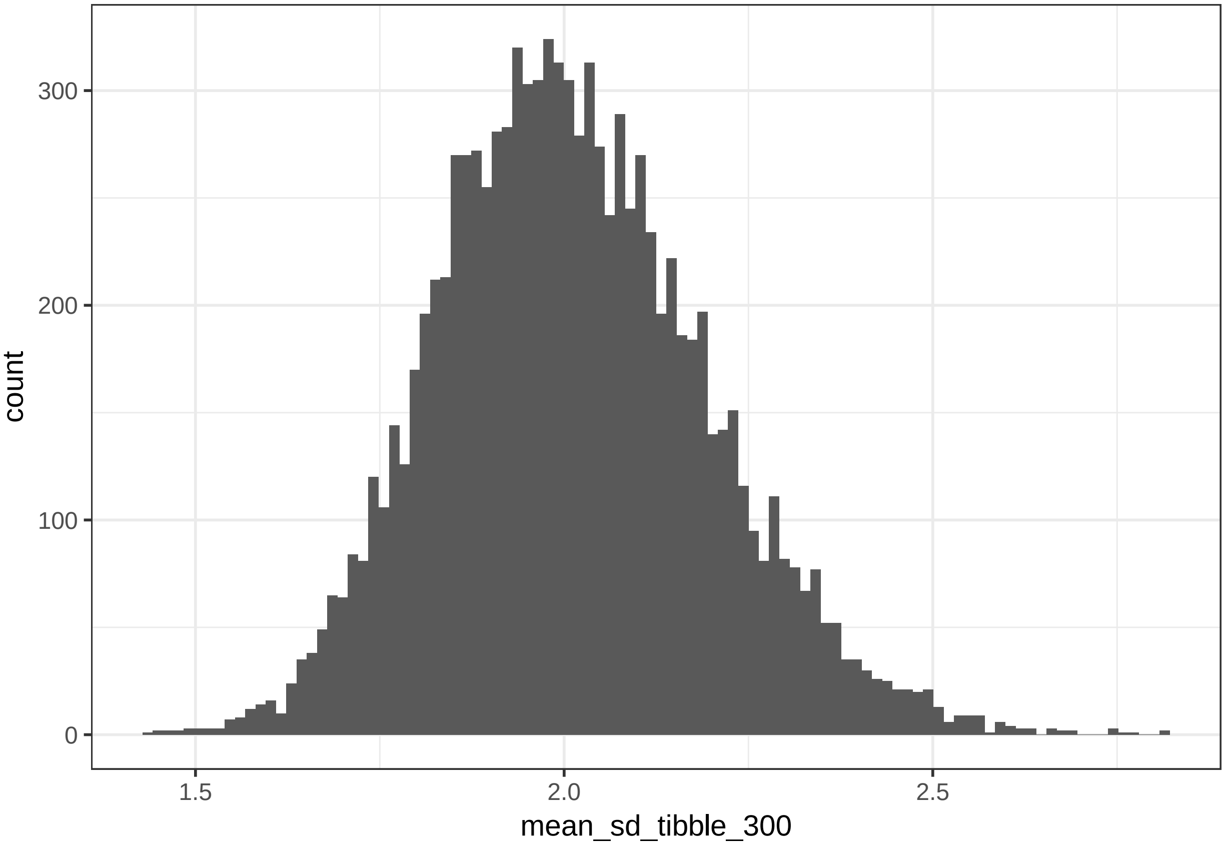
```
# plot for the means_300
ggplot(data = mean_dist_300, mapping = aes(x=mean_samp_tibble_300)) +
  geom_histogram(bins=100) +
  theme_bw()
```



```
# Repeat Question 6 with sample size = 300

# below are sample of the standard deviations of the population
head(mean_sd_tibble_300)
```

```
## # A tibble: 6 x 1
##   mean_sd_300
##         <dbl>
## 1        1.91
## 2        2.05
## 3        2.32
## 4        1.82
## 5        1.62
## 6        2.22
```

```
# plot for the standard deviation
ggplot(data = mean_dist_300, mapping = aes(x=mean_sd_tibble_300)) +
  geom_histogram(bins=100) +
  theme_bw()
```



Answer: Even with the sample size of 300, the distributions of means and sd are looking normally distributed to me.