COMPSCIX 415.2 Homework 6

Bryan Hee

March 12, 2018

Exercise 1

Question 1

The dataset Whickam included in the mosaicData library represents 1314 observations of a one-in-six survey of women in Whickham, UK in 1972-1974. The study focused on heart disease and thyroid disease and was followed up 20 years later with a second survey. The variables included are the outcome survival status after 20 years, smoker status during the first survey, and age at the time of the first survey.

```
str(Whickham)
```

```
## 'data.frame':    1314 obs. of  3 variables:
##  $ outcome: Factor w/ 2 levels "Alive","Dead": 1 1 2 1 1 1 1 2 1 1 ...
##  $ smoker : Factor w/ 2 levels "No","Yes": 2 2 2 1 1 2 2 1 1 1 ...
##  $ age    : int  23 18 71 67 64 38 45 76 28 27 ...
```

Question 2

There are 1314 observations in the dataset, and each represents a woman who lived in Whickham in 1972-1974.

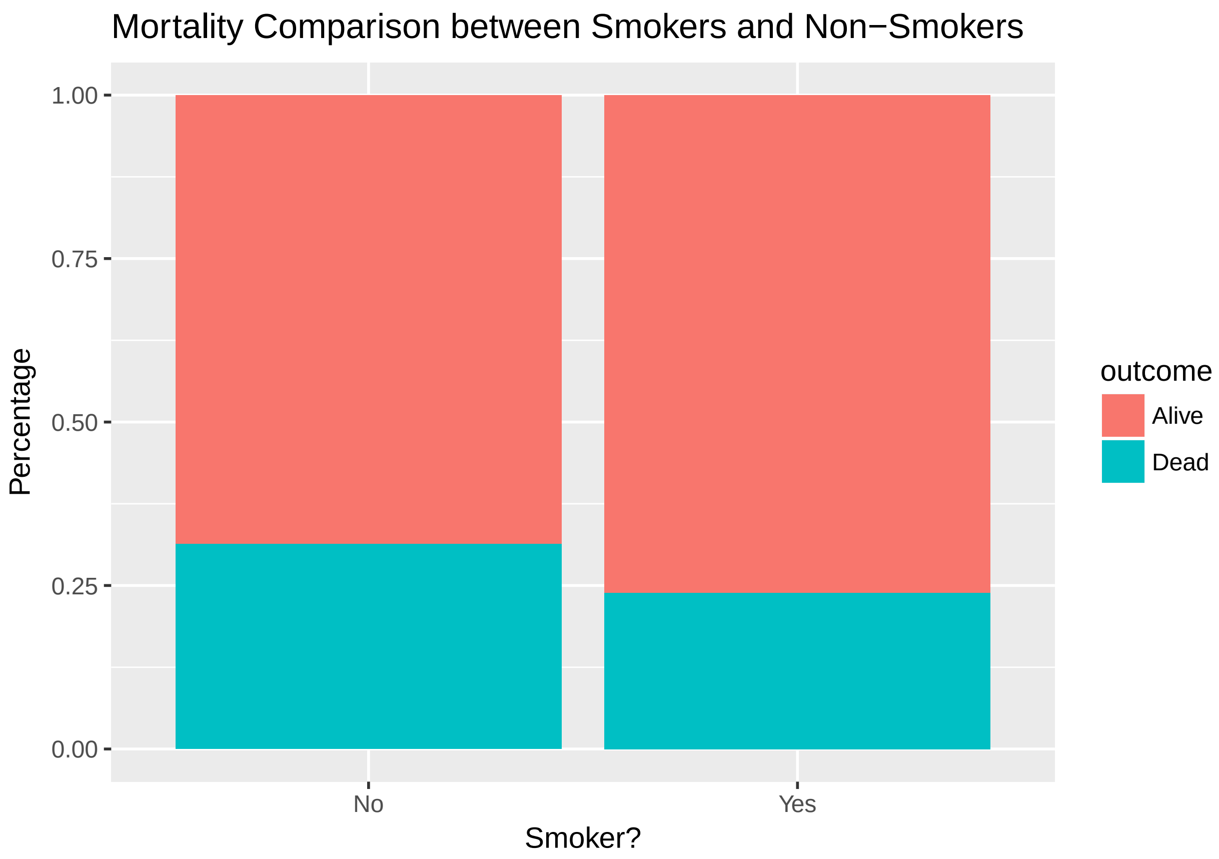
Question 3

The following comparison of the mortality rate between smokers and non-smokers shows a slight correlation between not smoking and dying. This obviously does not make very much sense, as we know that, all else being equal, smoking should increase the mortality rate in a population. This displays Simpson's Paradox. The other information displayed in the graph below is that the number of women alive is much larger than those who are deceased.

```
Whickham %>% count(smoker, outcome) -> Ageless
  colnames(Ageless)[3] <- "total"
  mutate(Ageless, percent = total/sum(total)*100
         ) -> Ageless
Ageless
```

```
## # A tibble: 4 x 4
##   smoker outcome total percent
##   <fct>  <fct>   <int>   <dbl>
## 1 No     Alive     502    38.2
## 2 No     Dead      230    17.5
## 3 Yes    Alive     443    33.7
## 4 Yes    Dead      139    10.6
```

```
  ggplot(Ageless) +
  geom_bar(aes(x = smoker, y = total, fill = outcome), stat = "identity", position = "fill") +
  labs(x = "Smoker?", y = "Percentage", title = "Mortality Comparison between Smokers and Non-Smokers")
```



Question 4

By faceting by age group, the graphs now display the opposite of the trend displayed in the graph above. The new trend shown is that smoking is positively correlated with mortality. This makes more sense given our understanding of the carcinogenic nature of tobacco products.

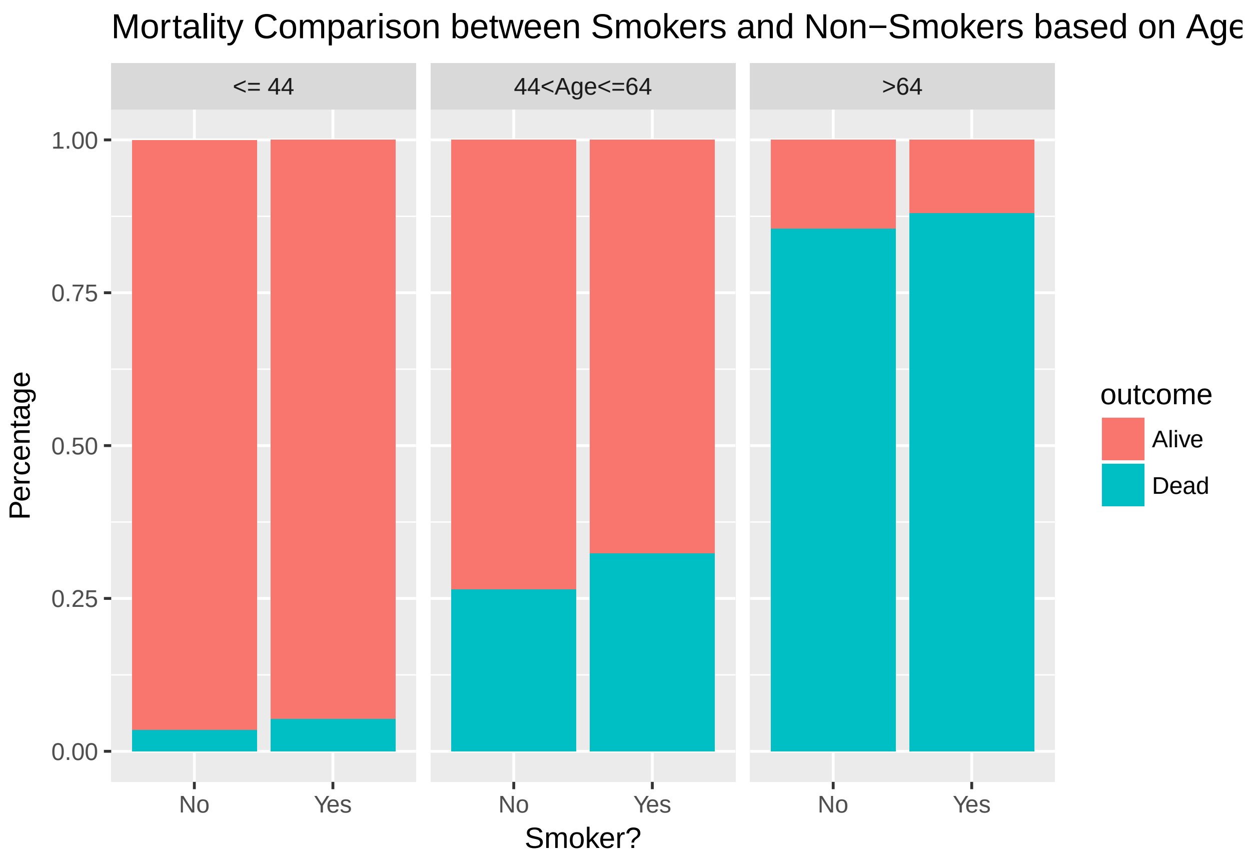
```
Whickham_fct <- Whickham
Whickham_fct %>%
mutate(age_group = factor(case_when(Whickham_fct$age <= 44 ~ '<= 44',
                                    Whickham_fct$age > 44 & Whickham_fct$age <= 64~'44<Age<=64',
                                    Whickham_fct$age > 64 ~ '>64'),
                                    levels = c('<= 44', '44<Age<=64', '>64'), ordered = TRUE)) -> Whick

Whickham_fct %>% count(smoker, outcome, age_group) -> with_age
colnames(with_age)[4] <- "total"
mutate(with_age, percent = total/sum(total)*100) -> with_age

with_age %>% arrange(age_group) -> with_age
with_age
```

```
## # A tibble: 12 x 5
##    smoker outcome age_group  total percent
##    <fct>  <fct>   <ord>      <int>   <dbl>
##  1 No     Alive   <= 44        327  24.9
##  2 No     Dead    <= 44         12   0.913
##  3 Yes    Alive   <= 44        270  20.5
##    4 Yes    Dead     <= 44          15   1.14
##    5 No     Alive    44<Age<=64    147  11.2
##    6 No     Dead     44<Age<=64     53   4.03
##    7 Yes    Alive    44<Age<=64    167  12.7
##    8 Yes    Dead     44<Age<=64     80   6.09
##    9 No     Alive    >64            28   2.13
##   10 No     Dead     >64           165  12.6
##   11 Yes    Alive    >64             6   0.457
##   12 Yes    Dead     >64            44   3.35
```

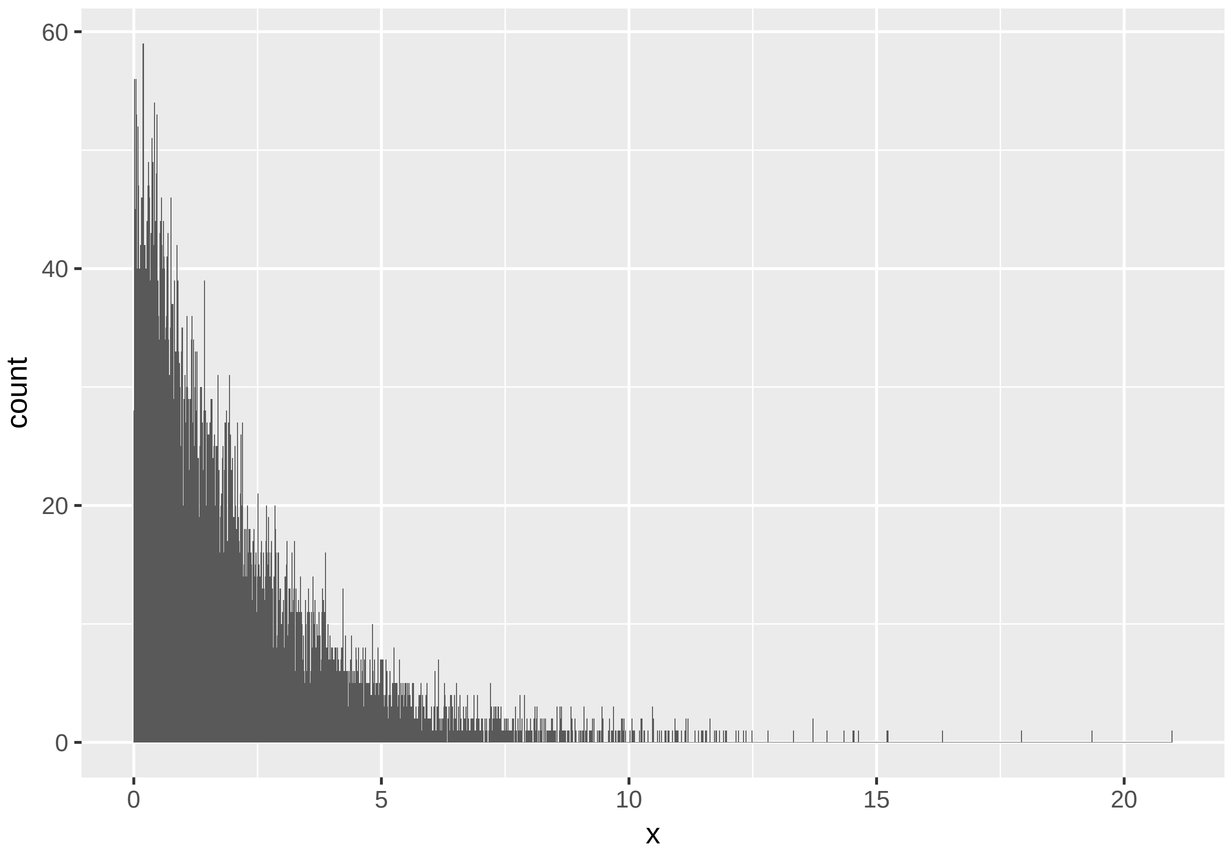
```
with_age %>%
ggplot() +
geom_bar(aes(x = smoker, y = total, fill = outcome), stat = "identity", position = "fill") +
  facet_wrap(~age_group) +
  labs(x = "Smoker?", y = "Percentage", title = "Mortality Comparison between Smokers and Non-Smokers ba
```



Exercise 2

Question 1

```
n <- 10000
gamma_samp <- tibble(x = rgamma(n, shape = 1, scale = 2))
ggplot(gamma_samp, mapping = aes(x = x)) +
geom_histogram(binwidth = 0.01)
```



Question 2

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

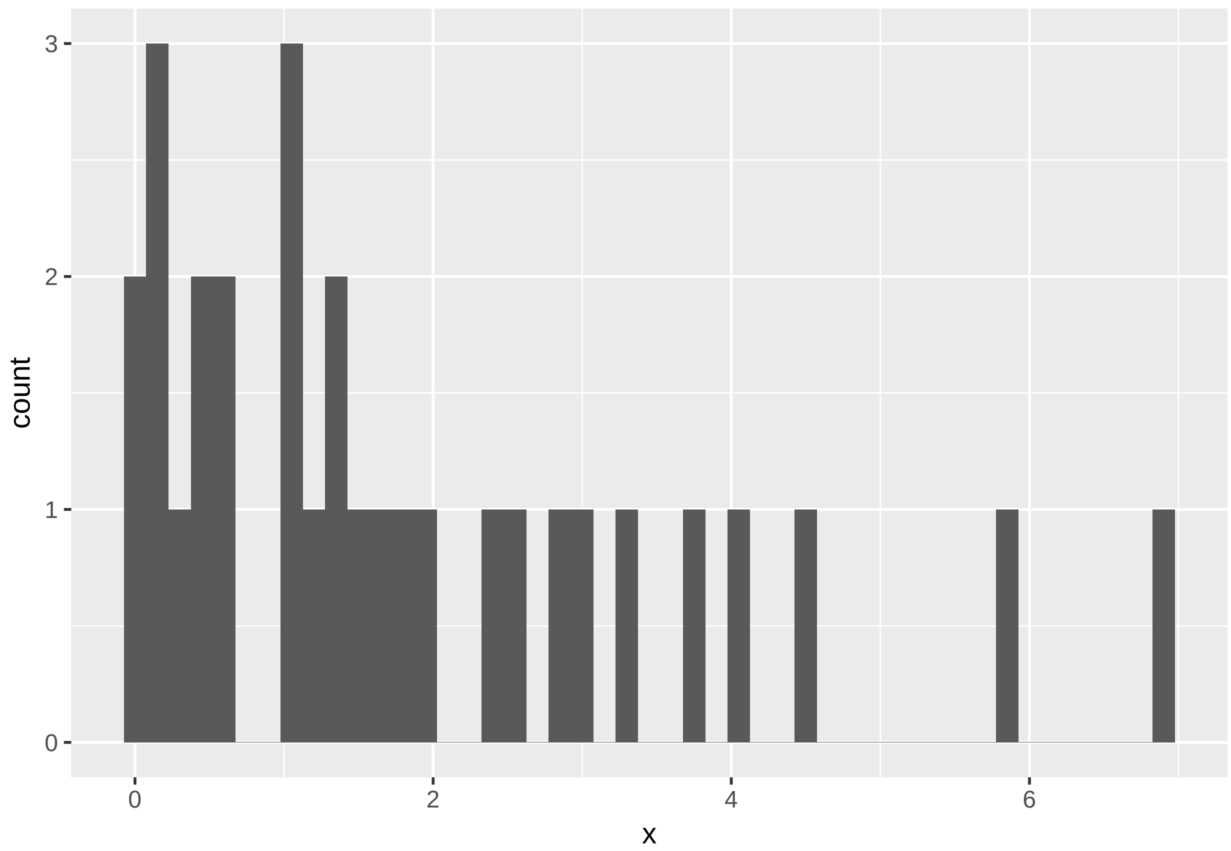
```
## [1] 2.01715
```

```
variance_samp <- gamma_samp %>% .[['x']] %>% var()
variance_samp
```

```
## [1] 4.048672
```

Question 3

```
n = 30
gamma_samp_3 <- tibble(x = rgamma(n, shape = 1, scale = 2))
ggplot(gamma_samp_3, mapping = aes(x = x)) +
geom_histogram(binwidth = .15)
```



```
mean_samp_3 <- gamma_samp_3 %>% .[['x']] %>% mean()
mean_samp_3
```

```
## [1] 1.866469
```

```
variance_samp_3 <- gamma_samp_3 %>% .[['x']] %>% var()
variance_samp_3
```

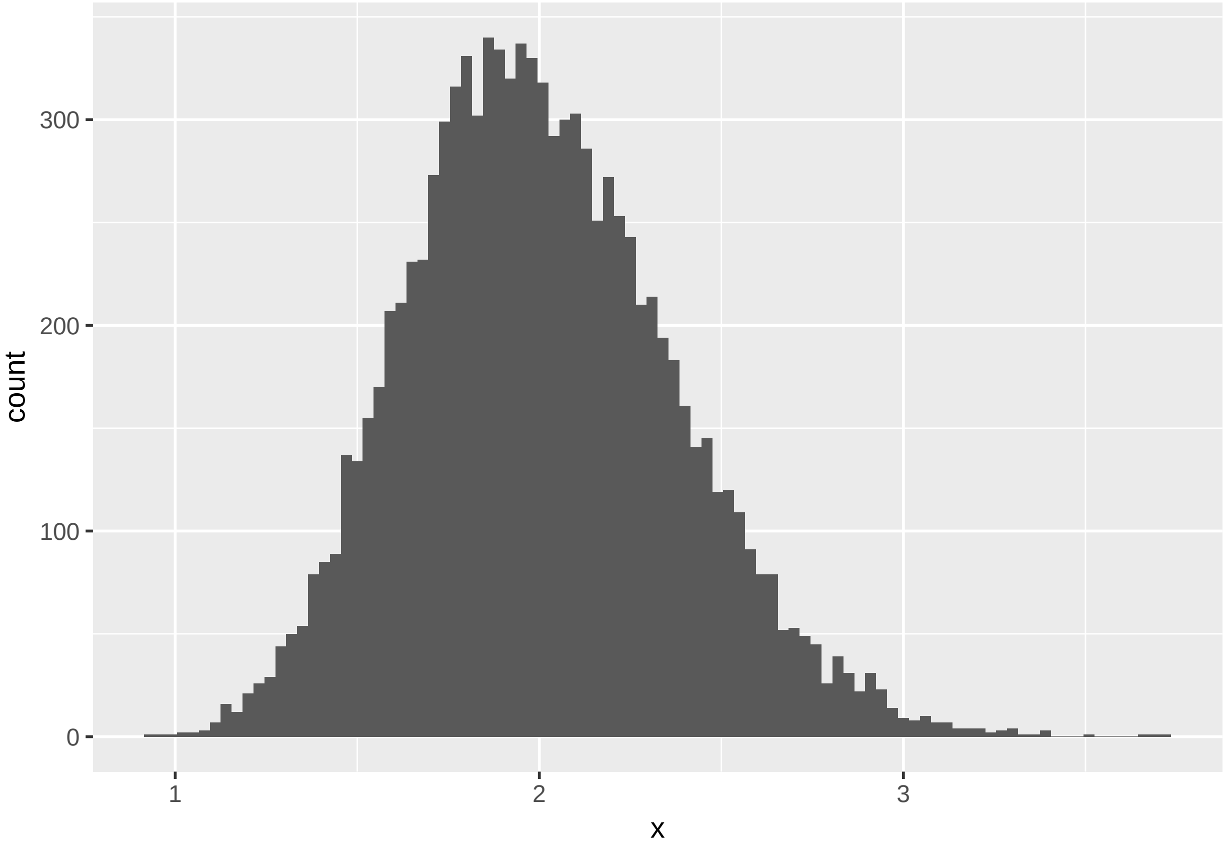
```
## [1] 3.140132
```

Question 4

```
gamma_mean_samp_4 <- rep(NA, 10000)
for(i in 1:10000) {
  g_samp_4 <- rgamma(30, shape =1, scale = 2)
  gamma_mean_samp_4[i] <- mean(g_samp_4)
}
gamma_mean_samp_4 <- tibble (gamma_mean_samp_4)
colnames(gamma_mean_samp_4)[1] <- 'x'
```

Question 5

```
gamma_mean_samp_4 %>%
ggplot(mapping = aes(x=x)) +
  geom_histogram(binwidth = 0.03)
```



Question 6

```
mean_samp_4 <- gamma_mean_samp_4 %>% .[['x']] %>% mean()
mean_samp_4
```

```
## [1] 2.004487
```

```
variance_samp_4 <- gamma_mean_samp_4 %>% .[['x']] %>% var()
variance_samp_4
```

```
## [1] 0.1333627
```

Question 7

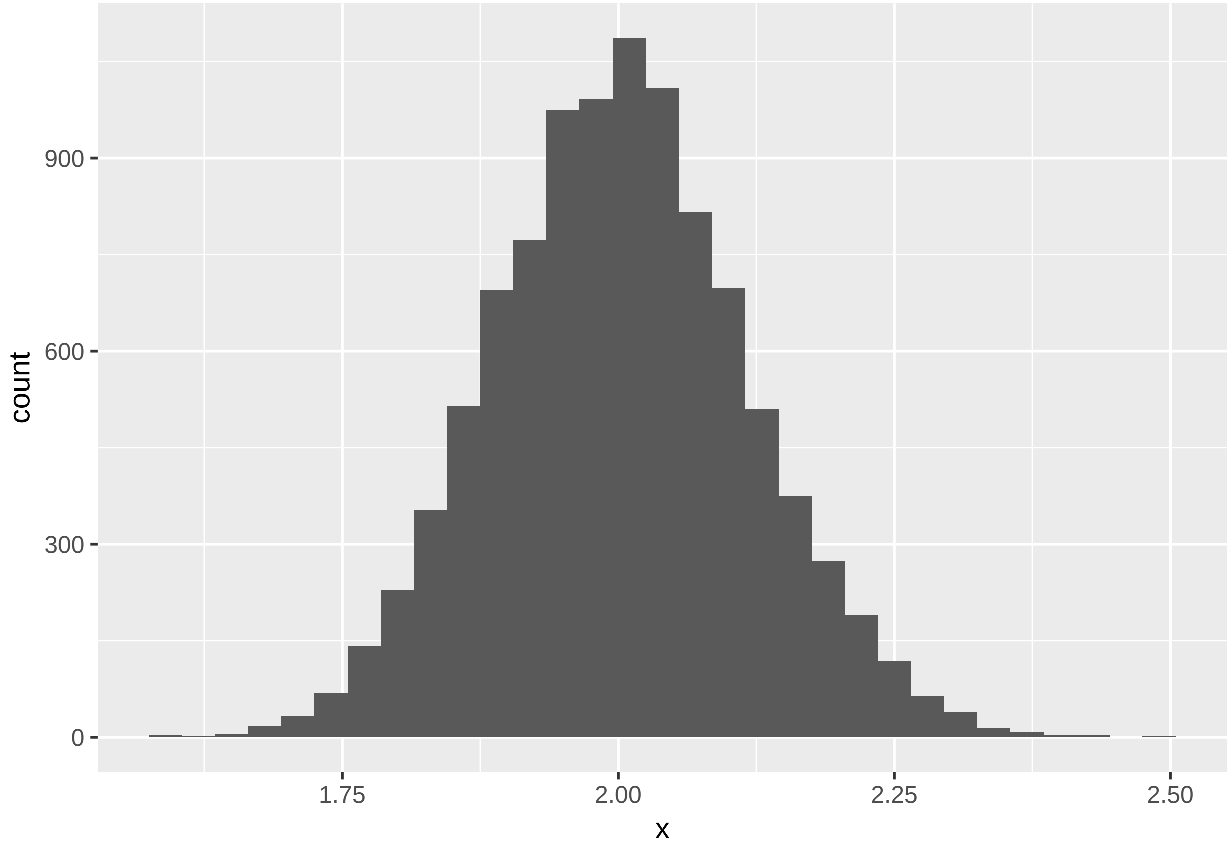
THe small variance of the data within the tibble of 10,000 means in the previous question was surprising to me. In the previous problems, it was much closer to 4. However, according to the Central Limit Theorem, the variance should be 0.365ˆ2, or 0.133. The variance in the previous problem is very close to that.

Question 8

The mean continues to be very close to 2. According to the Central Limit Theorem the standard deviation should be ~0.115 which yields a variance of ~0.013333. When taking a sample of 300, the variance in the following code confirms this theorem with a sample variance close to ~0.0133.

```
gamma_mean_samp_5 <- rep(NA, 10000)
for(i in 1:10000) {
  g_samp_5 <- rgamma(300, shape =1, scale = 2)
  gamma_mean_samp_5[i] <- mean(g_samp_5)
}
gamma_mean_samp_5 <- tibble (gamma_mean_samp_5)
colnames(gamma_mean_samp_5)[1] <- 'x'

gamma_mean_samp_5 %>%
ggplot(mapping = aes(x=x)) +
  geom_histogram(binwidth = 0.03)
```



```
mean_samp_5 <- gamma_mean_samp_5 %>% .[['x']] %>% mean()
mean_samp_5
```

```
## [1] 2.001443
```

```
variance_samp_5 <- var(gamma_mean_samp_5$x)
variance_samp_5
```

```
## [1] 0.01323302
```