COMPSCIX 415.2 Homework 6

Rajat Jain

July 11th, 2018

Contents

Exercises

Exercise 1

Load the `Whickham` dataset (`data(Whickham)`). You will need to load the `mosaicData` package first, but I also included the data as a csv file on Canvas if you would rather download it there and load it with the `readr` package.

Look at the help file on this dataset to learn a bit about it. *Note that the help file only exists if you are using the dataset from the `mosaicData` package. If you are loading the dataset from the csv file, do a Google search of this dataset and package name to help answer the first two questions below.*

1. **What variables are in this data set?**

The `Whickham` data set contains the following variables.

- *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

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

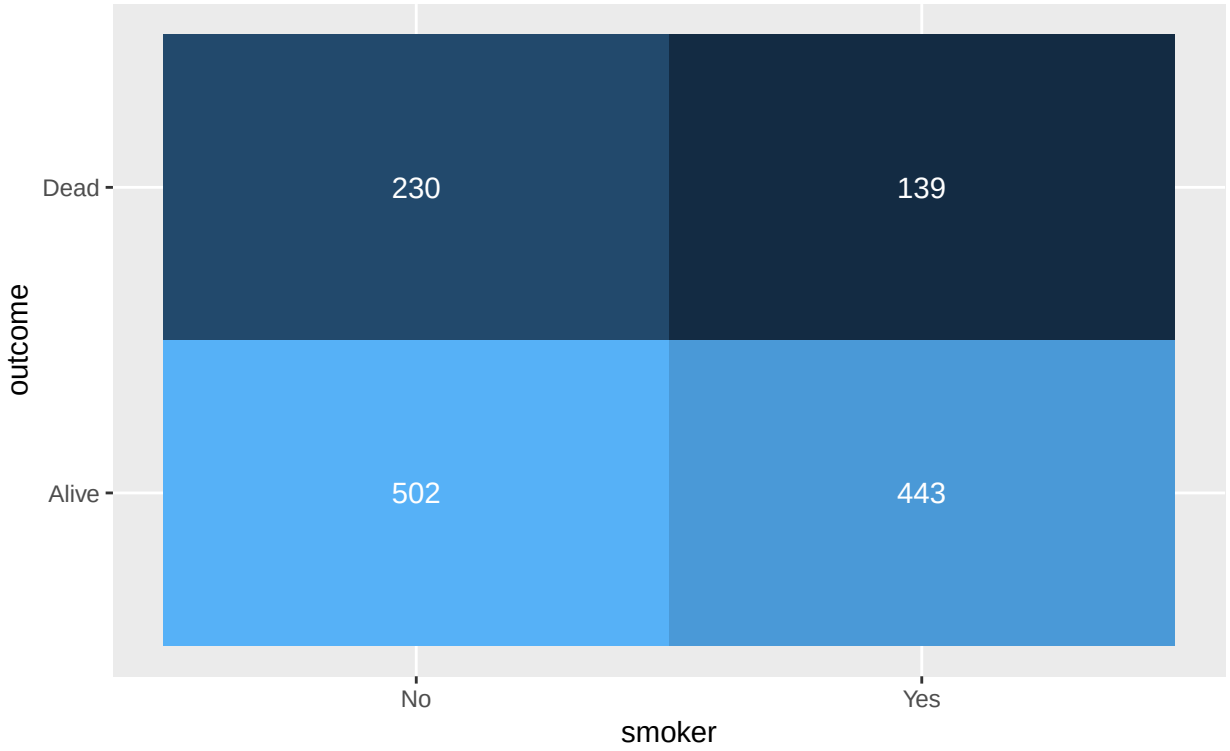
There are 1314 observations. Each observation represents the survey response from women of the electoral roll in Whickham, a mixed urban and rural district near Newcastle upon Tyne, in the UK. The survey was conducted in 1972-1974 to study heart disease and thyroid disease.

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?**

```
(smoking_outcome <- 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
```

```r
smoking_outcome %>%
  ggplot(aes(x = smoker, y = outcome)) +
  geom_tile(aes(fill = n), show.legend = FALSE) +
  geom_text(aes(label = n), color = 'white')
```



Looking at this visualization, one could conclude that the majority of dead people were non smokers which means that smoking does not have any adverse effects on health. It does NOT make sense.

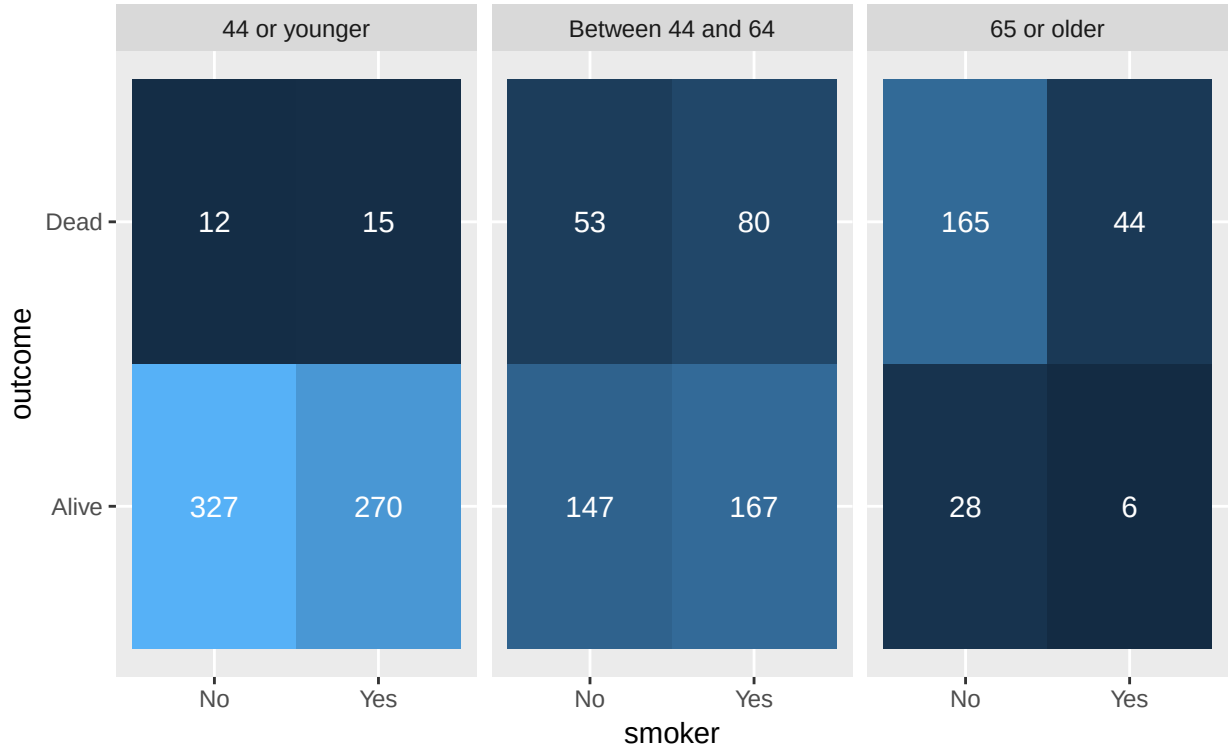
4. **Recode the age variable into an ordered factor with three categories: age <= 44, age > 44 & age <= 64, and age > 64. Now, recreate visualization from above, but facet on your new age factor. What do you see? Does it make sense?**

```r
Whickham_new <- Whickham %>%
  mutate(
    age_group = fct_relevel(
                  factor(
                    case_when(
                      age <= 44 ~ '44 or younger',
                      age > 44 & age <= 64 ~ 'Between 44 and 64',
                      age > 64 ~ '65 or older'
                    ),
                    ordered = TRUE
                  ),
                  '44 or younger', 'Between 44 and 64', '65 or older'
                )
  )

levels(Whickham_new$age_group)
```

```
## [1] "44 or younger"     "Between 44 and 64" "65 or older"
```

```
Whickham_new %>%
  group_by(age_group) %>%
  count(smoker , outcome) %>%
  ggplot(aes(x = smoker, y = outcome)) +
  geom_tile(aes(fill = n), show.legend = FALSE) +
  geom_text(aes(label = n), color = 'white') +
  facet_wrap(~age_group)
```



Visualizing by `age_group` does make more sense. Here are the three things we can observe from this plot:

- Looking at the `age_group 44 or younger` it is evident that the effects of smoking are not visible until much later age.
- From the `age_group Between 44 and 64` at survey, we can observe that smokers are more likely to die in next 20 years or so.
- Majority of women from `age_group 65 or older` at the time of survey, who survied after 20 years were non-smokers.

So we can conclude that smoking has adverse effects on health and eventually on life expectancy.

Exercise 2

The Central Limit Theorem states that the sampling distribution of sample means is approximately Normal, regardless of the distribution of your population. For this exercise our population distribution will be a Gamma(1,2) distribution, and we'll show that the sampling distribution of the mean is in fact normally distributed.

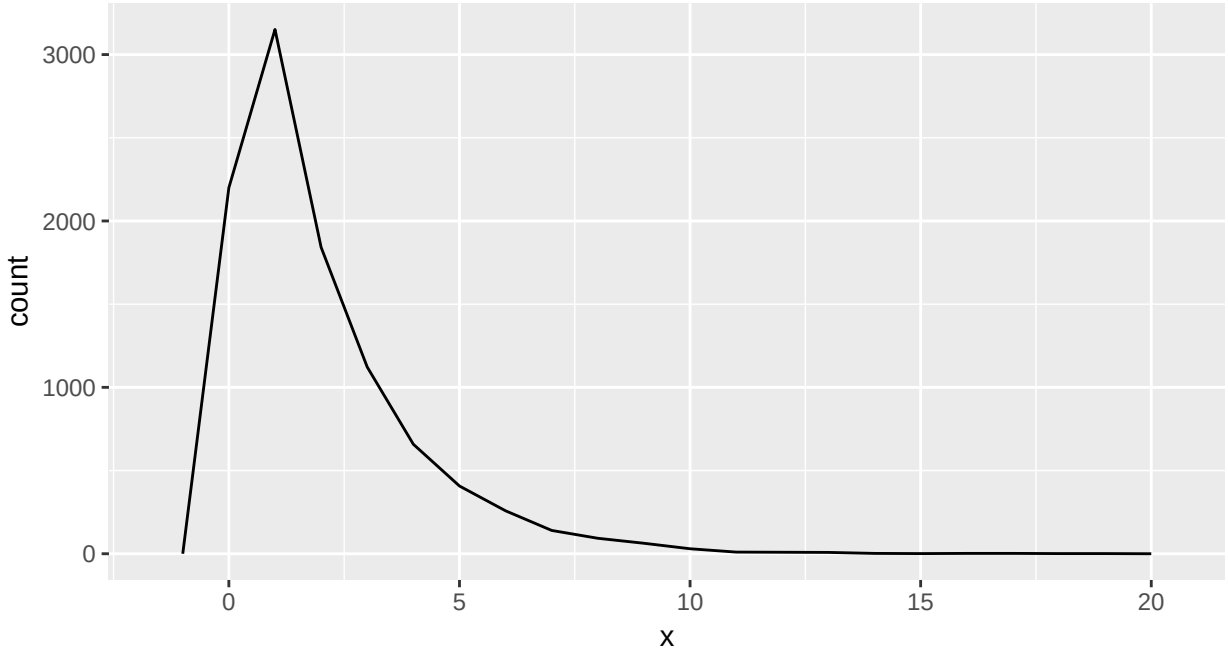
1. **Generate a random sample of size n = 10000 from a gamma(1,2) distribution and plot a histogram or density curve. Use the code below to help you get your sample.**

```
library(tidyverse)
n <- 10000

# look at ?rgamma to read about this function
gamma_samp <- tibble(x = rgamma(n, shape = 1, scale = 2))
```

```
gamma_samp <- tibble(x = rgamma(n = 10000, shape = 1, scale = 2))

ggplot(gamma_samp, aes(x = x)) +
  geom_freqpoly(binwidth = 1)
```



2. **What is the mean and standard deviation of your sample? They should both be close to 2 because for a gamma distribution:**
   mean = shape x scale
   variance = shape x scale^2

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

```
## [1] 1.966149
```

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

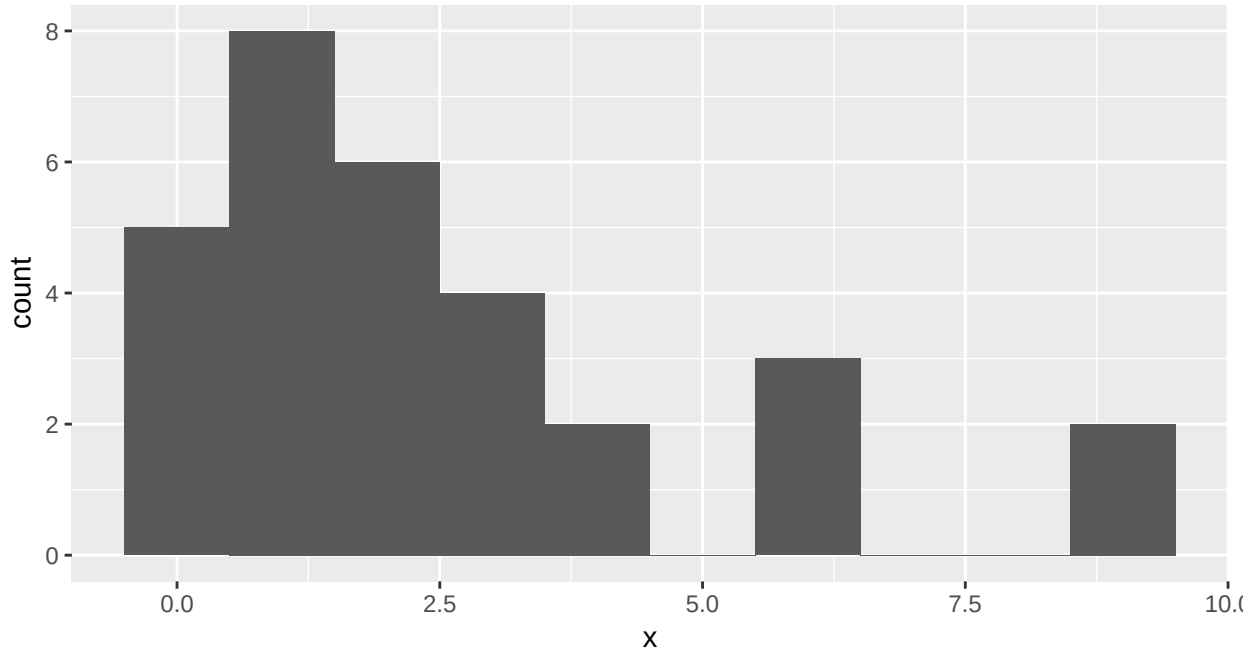
```
## [1] 1.955538
```

The mean of the sample is 1.9661485 and the standard deviation is 1.9555379. Yes both of them are very close to 2.

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_sample <- tibble(x = rgamma(n = 30, shape = 1, scale = 2))

# Distribution Curve
gamma_sample %>%
  ggplot(aes(x = x)) +
  geom_histogram(binwidth = 1)
```




```
#Mean
gamma_sample %>% .[['x']] %>% mean()
```

```
## [1] 2.568364
```

```
#Standard deviation
gamma_sample %>% .[['x']] %>% sd()
```

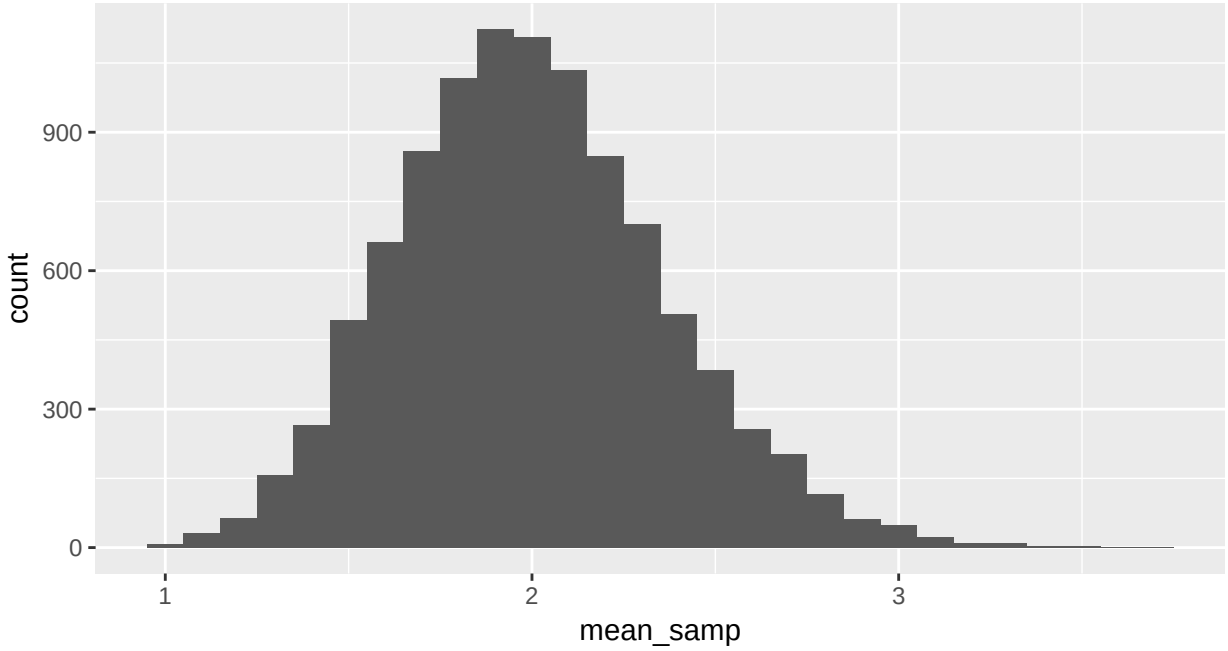
```
## [1] 2.466683
```

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.**

```
# 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)
```

5. **Make a histogram of your collection of means from above (`mean_samp`).**

```
ggplot(data = mean_samp, aes(x = mean_samp)) +
  geom_histogram(binwidth = 0.1)
```



6. **Calculate the mean and standard deviation of all of your sample means.**

```
paste("Mean: ", mean(mean_samp[['mean_samp']]))
```

```
## [1] "Mean:  1.9988643465569"
```

```
paste("Standard deviation: ", sd(mean_samp[['mean_samp']]))
```

```
## [1] "Standard deviation:  0.363876181059988"
```

7. **Did anything surprise you about your answers to #6?**

Yes! While the mean of the sample means is very close to the expected mean of the original Gamma distribution, the standard deviation of the sample means is very different from the standard deviation of the original distribution.

This demonstrates the applicability of Central Limit Theorem. If we refer to the output in #5, we can clearly see that the sample means is a Normal distribution with the mean and standard deviation above, irrespective of the fact that the original distribution used to generate the samples was a Gamma distribution.

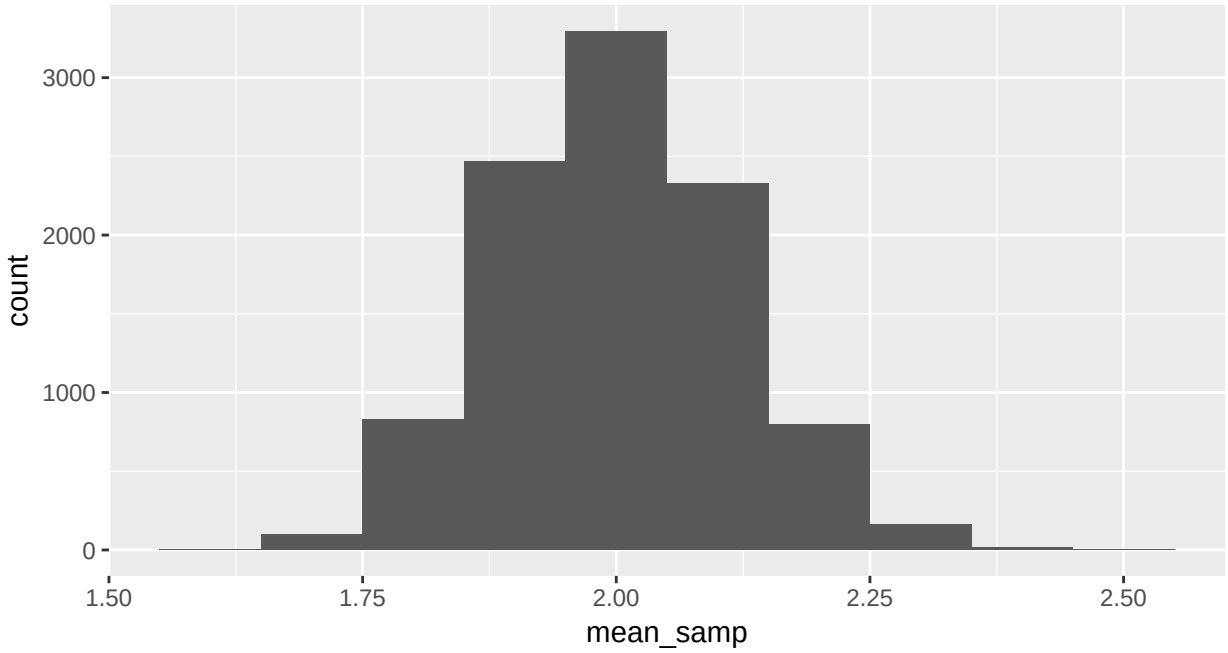
8. **According to the Central Limit Theorem, the mean of your sampling distribution should be very close to 2, and the standard deviation of your sampling distribution should be close to $\frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{30}} = 0.365$. Repeat #4-#6, but now with a sample of size n = 300 instead. Do your results match up well 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)

ggplot(data = mean_samp, aes(x = mean_samp)) +
  geom_histogram(binwidth = 0.1)
```




```
(actual_mean = mean(mean_samp[['mean_samp']]))
```

```
## [1] 2.000612
```

```
(actual_sd = sd(mean_samp[['mean_samp']]))
```

```
## [1] 0.1150713
```

By Central Limit Theoem,

- Expected Mean: 2
- Expected Standard Deviation: $\frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{300}} = 0.115$

As calculated using the code above,

- Actual Mean: 2.000612
- Actual Standard Deviation: 0.1150713

We can clearly see that our results match up closely with the expected results from the theorem.