Home Work Assignment 6

Pradeep Sahoo

7/14/2018

My Github repository for my assignments can be found at this URL: My Github

```
library(tidyverse)
library(mosaicData)
```

Exercise 1

```
data(Whickham)
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...
```

What variables are in this data set? There are 3 variables. Outcome, Smoker, Age.

How many observations are there and what does each represent? There are 1,314 observations. Each represents that the person is alive or dead, the person is a smoker or non-smoker and the age of the person

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?

```
Whickham %>% count( outcome , smoker  )
```

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

```
Whickham_mod <- Whickham %>%as_tibble() %>% group_by(outcome , smoker, age) %>% mutate(count = n())
```

```
Whickham_mod %>% ggplot() + geom_point(aes(x = age, y = count, fill = smoker, color = outcome))
```


```
age_cat <- case_when(Whickham$age <= 44 ~ 'Young',
                     Whickham$age > 44 & Whickham$age <= 64 ~ 'Mid',
                     Whickham$age > 64 ~ 'Old')
class(age_cat)
```

```
## [1] "character"
```

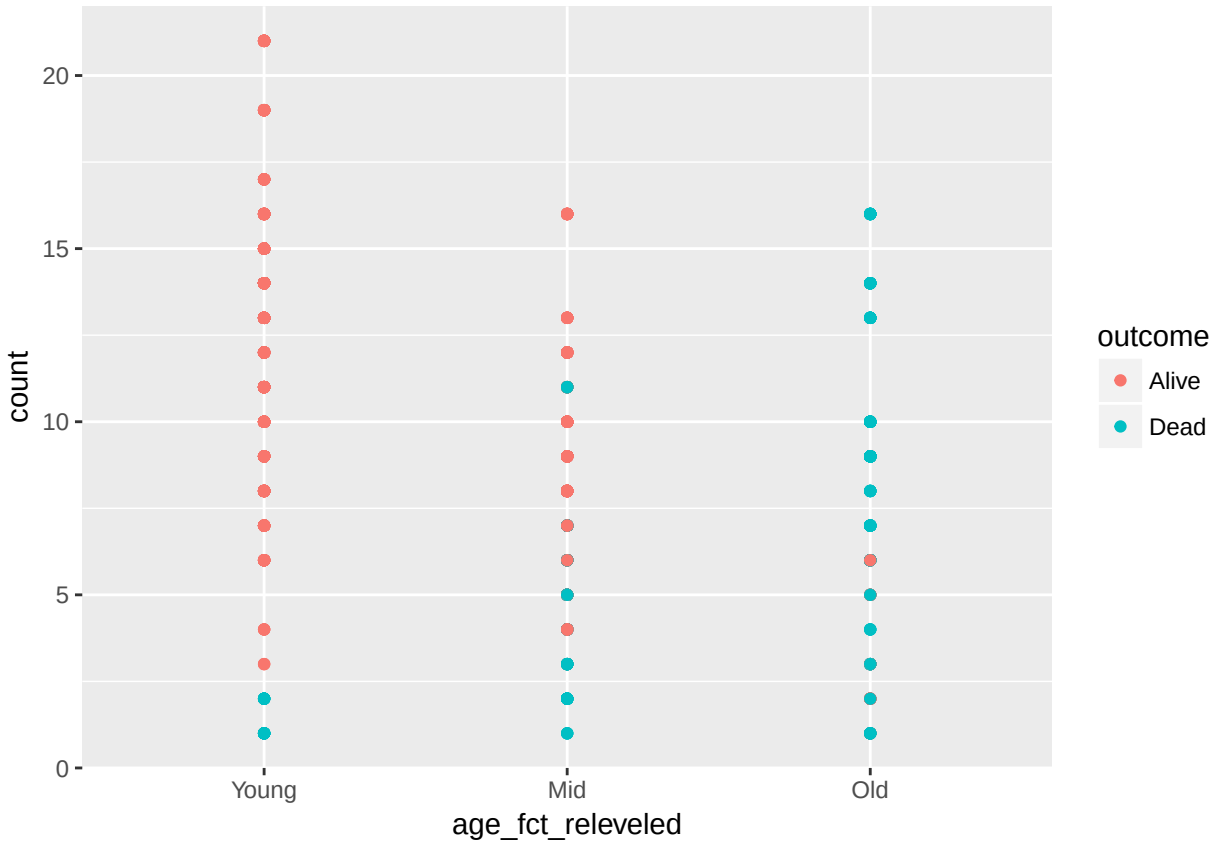
```
age_fct <- factor(age_cat,ordered = TRUE)
levels(age_fct)
```

```
## [1] "Mid"   "Old"   "Young"
```

```
age_fct_releveled <- fct_relevel(age_fct, 'Young','Mid','Old' )
levels(age_fct_releveled)
```

```
## [1] "Young" "Mid"   "Old"
```

```
Whickham_mod %>% ggplot() + geom_point(mapping = aes(x = age_fct_releveled, y = count, color = outcome))
```



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.

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

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

```
gamma_samp%>%
summarise(mean = mean(x),sd = sd(x))
```

```
## # A tibble: 1 x 2
##    mean    sd
##   <dbl> <dbl>
## 1  2.02  2.00
```

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.

```
n <- 30
gamma_samp <- (rgamma(n, shape = 1, scale = 2))
m_samp <- mean(gamma_samp)
m_samp
```

```
## [1] 3.045761
```

```
sd_samp <- sd(gamma_samp)
sd_samp
```

```
## [1] 3.087704
```

```
gamma_samp <- (x <- rgamma(n, shape = 1, scale = 2))
#gamma_samp %>% ggplot() + geom_histogram(aes(x = x))
```

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).

```
# Need to check with Robert
#g_samp %>% ggplot() + geom_histogram(aes(x = mean_samp))
```

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

```
mean_of_mean_samp <- mean(mean_samp)
mean_of_mean_samp
```

```
## [1] 1.99387
```

```
sd_of_mean_samp <- sd(mean_samp)
sd_of_mean_samp
```

```
## [1] 0.3671658
```

7. Did anything surprise you about your answers to #6? result is very close to .365
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 0.365. Repeat #4-#6, but now with a sample of size n = 300 instead. Do your results match up well with the theorem?

```
# 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)
```

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

```
# Need to check with Robert
#g_samp %>% ggplot() + geom_histogram(aes(x = mean_samp))
```

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

```r
mean_of_mean_samp <- mean(mean_samp)
mean_of_mean_samp
```

```
## [1] 1.99879
```

```r
sd_of_mean_samp <- sd(mean_samp)
sd_of_mean_samp
```

```
## [1] 0.1141242
```

standard deviation of your sampling distribution should be close to should be very close to 0.1154700538. My result is also matching to the expected result.