Homework 5

Tania Ommer, Sneha Augustine, Alisa Prinyarux, Mahek Patel, Sharon Feinleib

2023-11-07

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(moderndive)
library(skimr)
library(gapminder)
library(boot)
```

Use set.seed() and distinct() to set random seed and get rid of duplicates before analysis

0.How many duplicated records were removed from boxoffice data? What is your random seed number?

```
boxoffice <- read.csv("movie_boxoffice.csv", header=T)
glimpse(boxoffice)
```

```
## Rows: 4,969
## Columns: 7
## $ Movie           <chr> "Raise the Titanic", "Flash Gordon", "Popeye", "The Fo~
## $ Month           <chr> "Aug", "Dec", "Dec", "Feb", "Jan", "Jan", "Jan", "Jan"~
## $ Day             <int> 1, 5, 12, 1, 1, 1, 1, 1, 4, 25, 6, 13, 20, 20, 20, 7, ~
## $ Year            <int> 1980, 1980, 1980, 1980, 1980, 1980, 1980, 1980, 1980, ~
## $ Budget          <dbl> 40.00, 35.00, 20.00, 1.00, 6.50, 0.35, 3.50, 35.00, 3.~
## $ Domestic_Gross  <dbl> 7.000000, 27.107960, 49.823037, 21.378361, 31.899000, ~
## $ Worldwide_Gross <dbl> 7.000000, 27.107960, 49.823037, 21.378361, 31.899000, ~
```

```
boxoffice <- boxoffice %>% distinct()
glimpse(boxoffice)
```

```
## Rows: 4,869
## Columns: 7
## $ Movie           <chr> "Raise the Titanic", "Flash Gordon", "Popeye", "The Fo~
## $ Month           <chr> "Aug", "Dec", "Dec", "Feb", "Jan", "Jan", "Jan", "Jan"~
## $ Day             <int> 1, 5, 12, 1, 1, 1, 1, 1, 4, 25, 6, 13, 20, 20, 20, 7, ~
## $ Year            <int> 1980, 1980, 1980, 1980, 1980, 1980, 1980, 1980, 1980, ~
## $ Budget          <dbl> 40.00, 35.00, 20.00, 1.00, 6.50, 0.35, 3.50, 35.00, 3.~
## $ Domestic_Gross  <dbl> 7.000000, 27.107960, 49.823037, 21.378361, 31.899000, ~
## $ Worldwide_Gross <dbl> 7.000000, 27.107960, 49.823037, 21.378361, 31.899000, ~
```

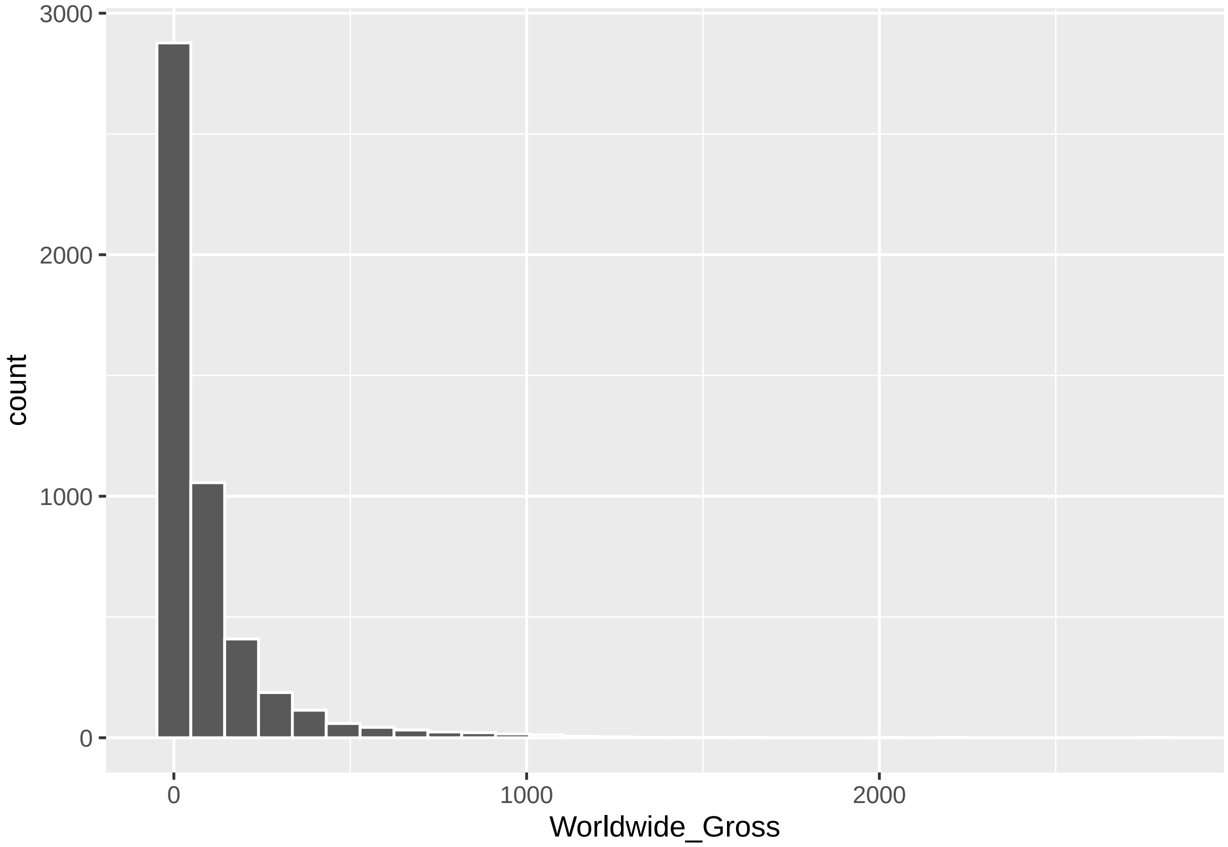
```
set.seed(20009345)
```

4969-4869=100 duplicated records were removed from the boxoffice data. The random seed used is 20009345.

1. Treat the box office data as your population. Have a histogram of the global box office earning. Describe the shape of the distribution

```
boxoffice <- boxoffice %>% mutate(highglobal=ifelse(Worldwide_Gross>=Budget, 1, 0))
ggplot(boxoffice, aes(x=Worldwide_Gross)) + geom_histogram(color="white")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
summary(boxoffice$Worldwide_Gross)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.000    5.778   31.152   95.832  102.333 2783.919
```

The distribution is heavily right skewed, the mean global box office earning is about $96 million, more than 25% of data have more than $100 million, and the max is $2784 million.

2. In the population, what is the average global box office earning? What is the standard deviation? What is the proportion of movies whose global box office earning exceeds budget?

```
boxoffice %>% summarise(avgbox=mean(Worldwide_Gross), stdbox=sd(Worldwide_Gross),numhigh=sum(highglobal
```

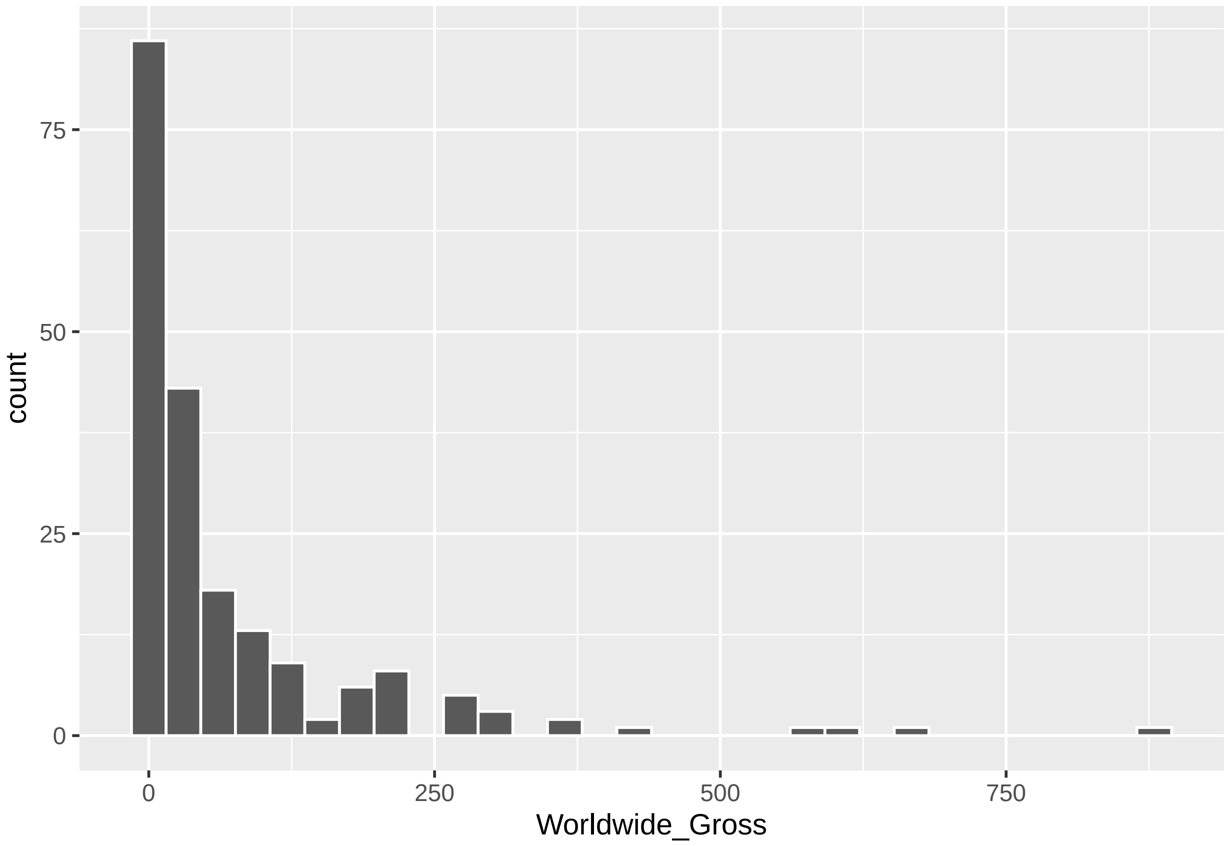
```
##     avgbox   stdbox numhigh nobs prop_high
## 1 95.83149 177.4594    3146 4869 0.6461286
```

The average global box office earning is $95.83 million. The standard deviaiton is $177.46 million. 64.6% of moives have global box office earnings exceed budgets.

##3. Take a random sample of 200 movies from the population, get the histogram of global box office earning, and describe the shape of the distribution. You are going to use this sample in part II and future homework.

```
movie_sample <- boxoffice %>% rep_sample_n(size=200)
ggplot(movie_sample, aes(x=Worldwide_Gross)) + geom_histogram(color="white")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
summary(movie_sample$Worldwide_Gross)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   3.136  20.703  70.465  85.400 879.621
```

The distribution is still heavily right skewed, with outlier points towards the very right of the graph.

##4. In your sample, what is the average global box office earning? What is the standard deviation? What is the proportion of movies whose global box office earning exceeds budget? Are these summary statistics from the sample close to those population parameters?

```
movie_sample %>% summarise(avgbox=mean(Worldwide_Gross), stdbox=sd(Worldwide_Gross), numhigh=sum(highgl
```

```
## # A tibble: 1 x 6
##   replicate avgbox stdbox numhigh  nobs prop_high
##       <int>  <dbl>  <dbl>   <dbl> <int>     <dbl>
## 1         1   70.5   122.     123   200     0.615
```

The average global box office earning is $107 million, and the standard deviaiton is $211.2 million. 61.5% of movies have global box office earnings exceed budgets. They are close to those population parameters.

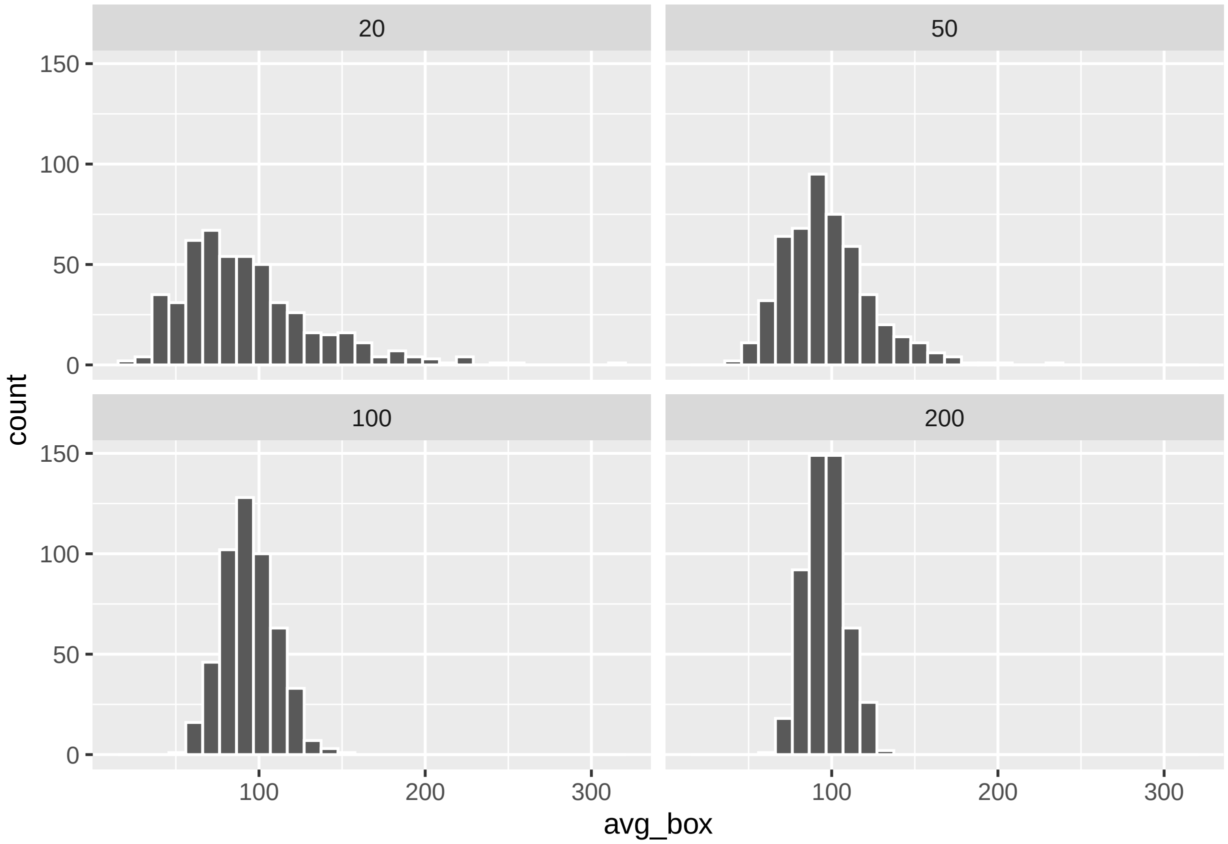
##5. Take a random sample of n movies from the population, calculate the average global box office earning, and the proportion of movies whose global box office earning exceeds budget. Repeat 500 times. (Do this step for n=20, 50, 100, 200 respectively)

```
samplesum1 <- boxoffice %>% rep_sample_n(size=20, reps=500) %>% group_by(replicate) %>%
  summarise(numhigh=sum(highglobal), avg_box=mean(Worldwide_Gross)) %>% mutate(prop_high=numhigh/20, si


samplesum2 <- boxoffice %>% rep_sample_n(size=50, reps=500) %>% group_by(replicate) %>%
  summarise(numhigh=sum(highglobal), avg_box=mean(Worldwide_Gross)) %>% mutate(prop_high=numhigh/50, si

samplesum3 <- boxoffice %>% rep_sample_n(size=100, reps=500) %>% group_by(replicate) %>%
  summarise(numhigh=sum(highglobal), avg_box=mean(Worldwide_Gross)) %>% mutate(prop_high=numhigh/100, s


samplesum4 <- boxoffice %>% rep_sample_n(size=200, reps=500) %>% group_by(replicate) %>%
  summarise(numhigh=sum(highglobal), avg_box=mean(Worldwide_Gross)) %>% mutate(prop_high=numhigh/200, s
sample_stat <-rbind(samplesum1, samplesum2, samplesum3, samplesum4)
```

6. For each n=20, 50, 100 and 200, get the histogram of the average global box office earning. Have histograms using facet_wrap() so the four histograms are in the same picture for easy comparison. Also get the mean and standard error of the average global box office earning.

```
ggplot(data=sample_stat, aes(x=avg_box )) + geom_histogram(color="white") + facet_wrap(~size, nrow=2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
sample_stat %>% group_by(size) %>% summarise(mean_avg=mean(avg_box), sd_avg=sd(avg_box))
```

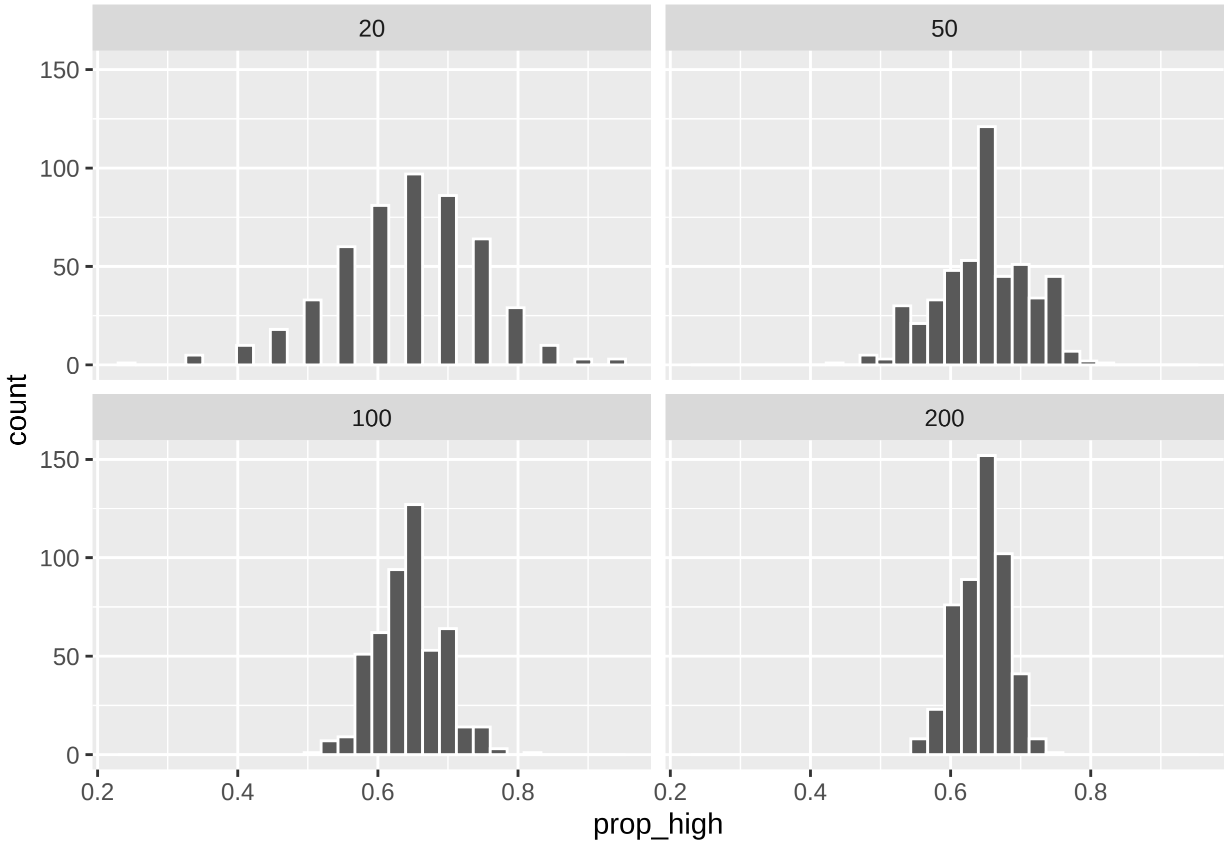
```
## # A tibble: 4 x 3
##    size mean_avg sd_avg
##   <dbl>    <dbl>  <dbl>
## 1    20     93.0   40.7
## 2    50     97.2   26.6
## 3   100     94.3   16.3
## 4   200     96.1   12.2
```

##7. Compare the distributions among different n values, and also compare them to the distribution from the population in Q1 1.The average of the sample mean is very close to the mean in the population, regardless of what the sample size is. 2. The standard error of the sample mean decreases as the sample size increases, but all the standard errors of the mean are smaller than the standard deviation in the population. 3. Having a sample size is small (ex. 20), The shape of the distribution of the sample mean is still right skewed, but as the sample size increases, the distribution is more like normal distribution, a bell shaped symmetry.

##8. For each n=20, 50, 100 and 200, get the histogram of the proportion of movies whose global box office earning exceeds budget. Have histograms using facet_wrap() so the four histograms are in the same picture for easy comparison. Also get the mean and standard error of the proportion of movies whose global box office earnings exceed budget.

```
ggplot(data=sample_stat, aes(x=prop_high)) + geom_histogram(color="white") + facet_wrap(~size, nrow=2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
sample_stat %>% group_by(size) %>% summarise(mean_prop=mean(prop_high), sd_prop=sd(prop_high))
```

```
## # A tibble: 4 x 3
##    size mean_prop sd_prop
##   <dbl>     <dbl>   <dbl>
## 1    20     0.641  0.108
## 2    50     0.647  0.0654
## 3   100     0.644  0.0462
## 4   200     0.645  0.0336
```

##9. Compare the distributions among different n values Looking at the distributions made above, we can see that the mean proportion of movies exceeding the budget increases as the sample size increases in small increments. As the sample size increases from 20 to 200 movies, the mean proportion increases specifically from 0.641 to 0.645. This shows to us that with larger sample sizes, a higher proportion of movies exceed their budget. By looking at the standard error, we can conclude that the standard error (variability) of the proportion exceeding the budget decreases as the sample size decreases as the sample size increases from 20 to 200. This indicates to us that as the sample size increases, the estimates become more precise.

##10. Use data from Q3 as the initial sample, use bootstrapping method to resample once with 200 movies. Are there any duplicated movies in your bootstrap sample? Is this expected or something is wrong?

```
movie_sample <- movie_sample %>% mutate(id=1:n())
bootstrap1 <- movie_sample %>% rep_sample_n(size=200, replace=TRUE)
table(bootstrap1$id)
```

```
##
##   1   2   3   4   8  11  12  13  14  15  17  20  21  22  23  28  29  30  31  33
##   1   1   1   3   1   2   2   1   1   2   2   1   1   1   1   1   1   1   1   2
##  34  35  36  39  41  42  44  45  46  47  50  51  52  53  56  58  59  60  61  62
##   1   2   1   2   3   2   2   1   1   1   1   1   1   1   2   1   1   2   1   2
##  63  65  66  67  69  70  73  75  76  77  78  79  81  83  85  86  90  91  94  95
##   2   3   2   1   2   1   2   1   1   3   3   2   2   1   1   2   2   2   2   2
##  97 100 101 102 103 104 106 107 108 109 110 111 112 119 122 123 124 126 127 128
##   1   2   4   1   1   2   3   1   2   1   3   2   1   2   1   2   1   2   3   1
## 129 130 132 133 134 135 136 137 138 139 140 141 144 146 147 148 149 150 151 154
##   2   1   2   1   1   1   1   1   2   2   1   1   3   1   1   1   1   2   2   1
## 160 162 163 164 166 167 168 175 176 177 178 179 180 182 183 184 188 189 190 191
##   2   2   1   1   1   1   4   3   2   1   2   1   2   4   1   1   1   1   1   3
## 192 196 197 198 199 200
##   1   3   1   1   1   1
```

```
duplicated(bootstrap1)
```

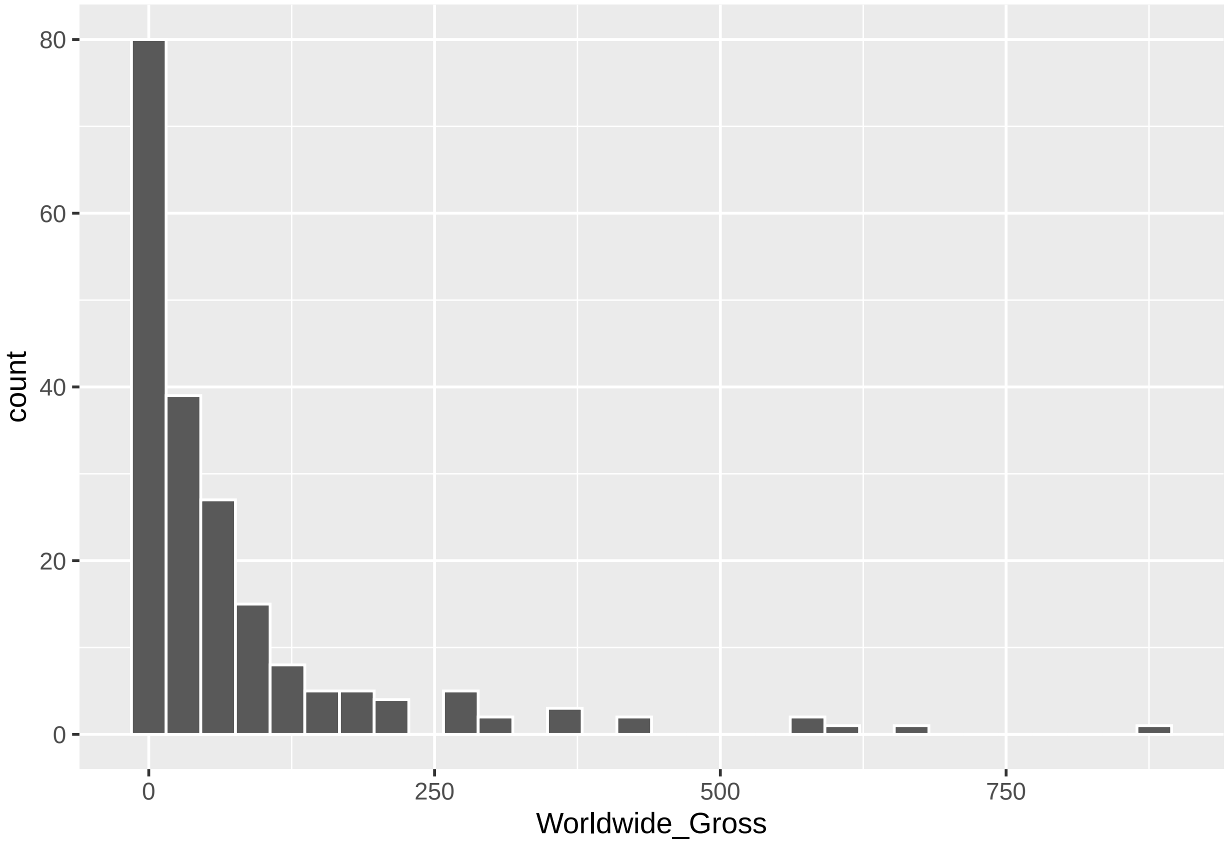
```
##   [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [25] FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE
##  [49]  TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE  TRUE
##  [61] FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
##  [73] FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
##  [85]  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE
##  [97]  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE  TRUE
## [109] FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE  TRUE FALSE  TRUE  TRUE
## [121] FALSE  TRUE FALSE  TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
## [133] FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE
## [145]  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE
## [157] FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
## [169]  TRUE  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE FALSE  TRUE
## [181]  TRUE FALSE  TRUE FALSE  TRUE FALSE  TRUE FALSE  TRUE  TRUE FALSE  TRUE
## [193] FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE  TRUE
```

There are a few duplicated movies. This is to be expected. Bootstrapping involves sampling WITH replacement, so it is normal to get some repetitions within a bootstrap sample. There are duplicates because of drawing with replacement in bootstrapping, so it is OK, and there's nothing wrong.

##11. Get the histogram of global box office earing in the bootstrap sample. Describe the shape of the distribution, and compare it to Q3

```
ggplot(bootstrap1, aes(x=Worldwide_Gross)) + geom_histogram(color="white")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The shape of the Histogram is similar to that in Q3, quite right skewed, with an outlier towards the very right of the graph.

##12. In the bootstrap sample in Q10, what is the average global box office earning? What is the standard deviation of it? What is the proportion of movies whose global box office earning exceeds budget? Are they close enough to those in the initial sample?

```
bootstrap1 %>% summarise(avgbox=mean(Worldwide_Gross), stdbox=sd(Worldwide_Gross), numhigh=sum(highgloba
```

```
## # A tibble: 1 x 6
##   replicate avgbox stdbox numhigh  nobs prop_high
##       <int>  <dbl>  <dbl>   <dbl> <int>     <dbl>
## 1         1   74.5   128.     128   200      0.64
```

The values in the bootstrap sample are reasonably close enough to those in the original sample, suggesting that the bootstrapping process has provided estimates that are relatively close to the true population values. In general, the proportion of movies exceeding the budget in the bootstrap sample are in a similar range to those in the original sample. The the average global box office earnings and standard deviation in the bootstrap sample is slightly higher since resampling using bootstrapping introduces some variability.

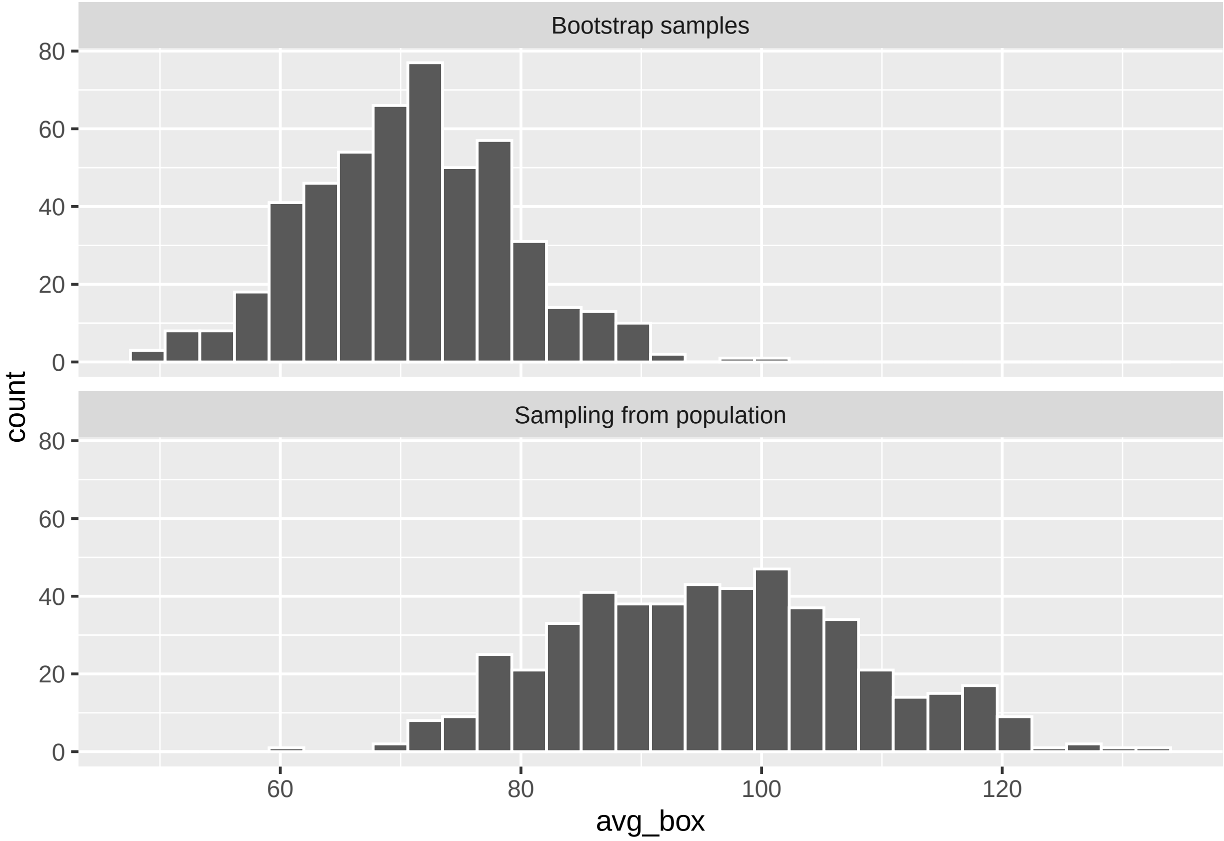
##13. Get the bootstrapping distribution of the average global box office earning, and the proportion of movies whose global box office earning exceeds budget, by resampling 500 times with bootstrapping method. Get the mean and standard error of the average global box office earning. Get the mean and standard error of the proportion of movies whose global box office earning exceeds budget

```
bootstrap_stat <- movie_sample %>% rep_sample_n(size=200, replace=TRUE, reps=500) %>% group_by(replicat
  summarise(numhigh=sum(highglobal), avg_box=mean(Worldwide_Gross)) %>% mutate(prop_high=numhigh/200, s

bootstrap_stat %>% summarise(mean_avg=mean(avg_box), sd_avg=sd(avg_box),mean_prop=mean(prop_high), sd_p
## # A tibble: 1 x 4
##   mean_avg sd_avg mean_prop sd_prop
##      <dbl>  <dbl>     <dbl>   <dbl>
## 1     70.8   8.46     0.614  0.0335
```

14. Have the histogram of the sample mean and sample proportion from Q13. Have histogram using facet_wrap() to make comparison of the bootstrap distribution with those the in Q6 and Q8 when n=200. Describe the comparison.

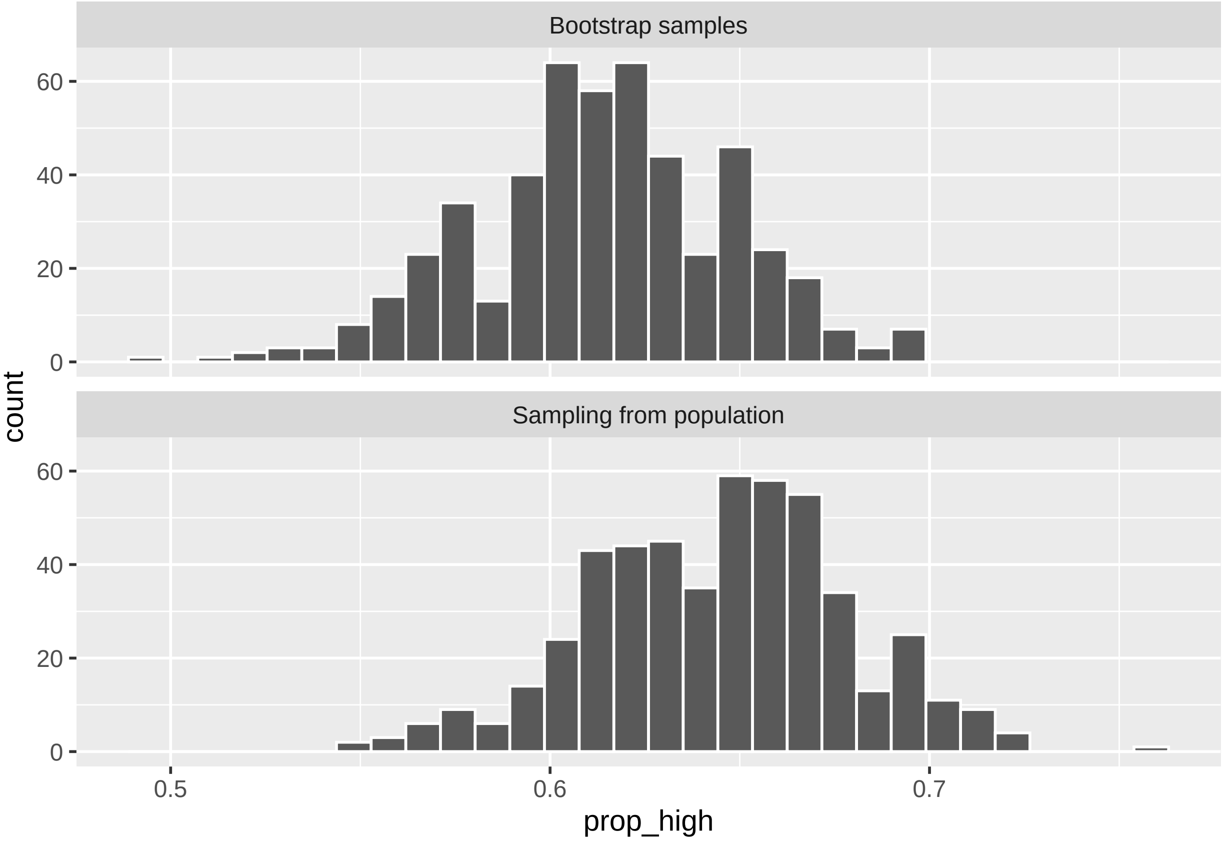
```
sample <- rbind(bootstrap_stat, samplesum4)
ggplot(data=sample, aes(x=avg_box)) + geom_histogram(color="white") + facet_wrap(~method, nrow=2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
ggplot(data=sample, aes(x=prop_high)) + geom_histogram(color="white") + facet_wrap(~method, nrow=2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Regardless of the sample mean or the sample proportion, both bootstrap samples and random samples from the population have normal shaped curve, besides the different centers. The bootstrap samples center around the mean or proportion from the initial sample, while the random samples from the population center around the mean or proportion from the population. The standrad errors are close to each other from the bootstrap samples and random samples from the population.