hw4

jason grahn

10/19/2018

# *R*

## Question 1:

### Take a random sample of size 50 from price. Using this sample, what is your best point estimate of the population mean?

price <- ames %>%   
 select(SalePrice)   
  
small\_sample <- price %>%   
 rep\_sample\_n(size = 50, reps = 1) %>%   
 group\_by(replicate) %>%   
 #give me an average per replicate  
 summarise(sample\_mean = mean(SalePrice))

At the time of this writing, the point estimate of sample\_mean is 1, 1.656478210^{5}.

## Question 2:

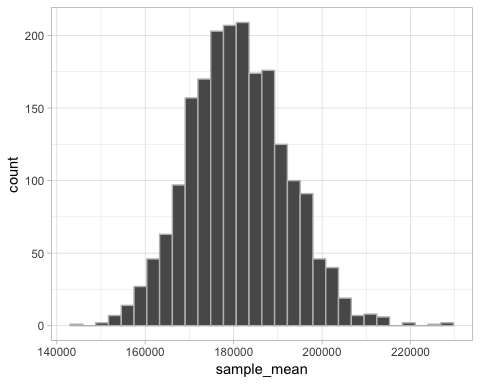
### Since you have access to the population, simulate the sampling distribution for price by taking 2000 samples from the population of size 50 and computing 2000 sample means. Store these means in a vector called sample\_means50.

sample\_means50 <- price %>%   
 rep\_sample\_n(size = 50, reps = 2000) %>%   
 group\_by(replicate) %>%   
 summarise(sample\_mean = mean(SalePrice))

### Plot the data, then describe the shape of this sampling distribution. Based on this sampling distribution, what would you guess the mean home price of the population to be?

sample\_means50 %>%   
 ggplot() +  
 geom\_histogram(aes(x = sample\_mean), color = "grey") +  
 theme\_light()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



mean\_of\_sample\_means50 <- sample\_means50 %>%   
 ungroup() %>%   
 mutate(spread = max(sample\_mean) - min(sample\_mean)) %>%   
 summarise(mean = mean(sample\_mean),  
 median = median(sample\_mean),  
 sd = sd(sample\_mean),  
 spread = mean(spread))

The shape of the sample distribution appears highly normal, though it appears to have a bit of a right skew. The median is 1.806556910^{5} with a standard deviation of 1.10898810^{4} This larger sampling provides a point estimate of 1.811187310^{5}, 1.806556910^{5}, 1.10898810^{4}, 8.37927410^{4}.

### Finally, calculate and report the population mean.

price %>%   
 summarise(mean = mean(SalePrice))

## mean  
## 1 180796.1

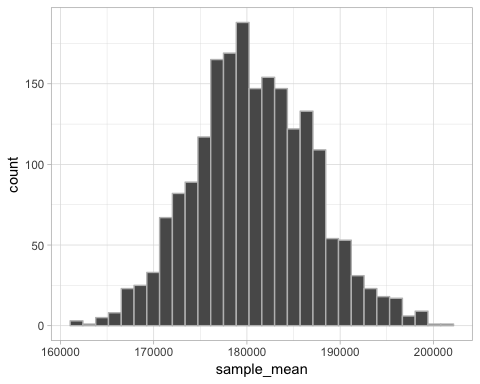
## Question 3:

### Change your sample size from 50 to 150, then compute the sampling distribution using the same method as above, and store these means in a new vector called sample\_means150. Describe the shape of this sampling distribution, and compare it to the sampling distribution for a sample size of 50. Based on this sampling distribution, what would you guess to be the mean sale price of homes in Ames?

sample\_means150 <- price %>%   
 rep\_sample\_n(size = 150, reps = 2000) %>%   
 group\_by(replicate) %>%   
 summarise(sample\_mean = mean(SalePrice))

sample\_means150 %>%   
 ggplot() +  
 geom\_histogram(aes(x = sample\_mean), color = "grey") +  
 theme\_light()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



mean\_of\_sample\_means150 <- sample\_means150 %>%   
 ungroup() %>%   
 mutate(spread = max(sample\_mean) - min(sample\_mean)) %>%   
 summarise(mean = mean(sample\_mean),  
 median = median(sample\_mean),  
 sd = sd(sample\_mean),  
 spread = mean(spread))

The mean of the larger sample size mean\_of\_sample\_means150 is less than the mean of mean\_of\_sample\_means50 by -346.3593567. The median of mean\_of\_sample\_means150 is about 200 higher.

The shape of the sample distribution appears highly normal, though it appears to have a bit of a right skew. The median is 1.804374110^{5} with a standard deviation of 6364.8935567 This larger sampling provides a point estimate of 1.807723710^{5}, 1.804374110^{5}, 6364.8935567, 3.974647310^{4}.

## Question 4:

### Of the sampling distributions from 2 and 3, which has a smaller spread? If we’re concerned with making estimates that are more often close to the true value, would we prefer a sampling distribution with a large or small spread?

#The spread of mean\_of\_sample\_means150 is  
mean\_of\_sample\_means150$spread

## [1] 39746.47

#The spread of mean\_of\_sample\_means50 is  
mean\_of\_sample\_means50$spread

## [1] 83792.74

The spread of the sampling distribution from 3 is smaller.