lab5

jason grahn

10/24/2018

# R component

population <- ames$Gr.Liv.Area  
samp <- sample(population, 60)  
  
population.tidy <- ames %>%   
 select(Gr.Liv.Area)  
  
samp.tidy <- population.tidy %>%   
 rep\_sample\_n(size = 60, reps = 1)  
  
summary(samp.tidy)

## replicate Gr.Liv.Area   
## Min. :1 Min. : 773   
## 1st Qu.:1 1st Qu.:1197   
## Median :1 Median :1411   
## Mean :1 Mean :1484   
## 3rd Qu.:1 3rd Qu.:1746   
## Max. :1 Max. :3086

## Exercise 1

### Describe the distribution of your sample. What would you say is the “typical” size within your sample? Also state precisely what you interpreted “typical” to mean.

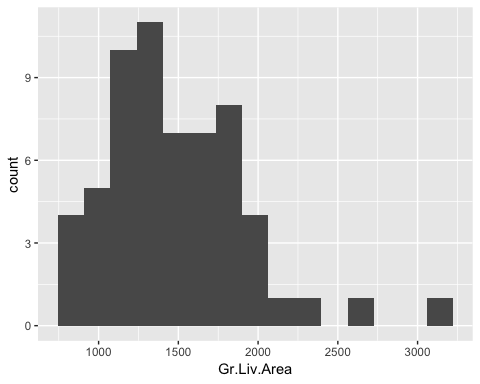
samp.describe <- describe(samp.tidy$Gr.Liv.Area)  
  
samp.histogram <- samp.tidy %>%   
 ggplot() +  
 geom\_histogram(aes(x = Gr.Liv.Area), bins = 15)  
  
samp.qqplot <- samp.tidy %>%   
 ggplot(aes(sample = Gr.Liv.Area)) +   
 stat\_qq() +  
 stat\_qq\_line(color = "red")  
  
Mode <- function(x) {  
 ux <- unique(x)  
 ux[which.max(tabulate(match(x, ux)))]  
 }  
  
Mode( samp.tidy$Gr.Liv.Area)

## [1] 1824

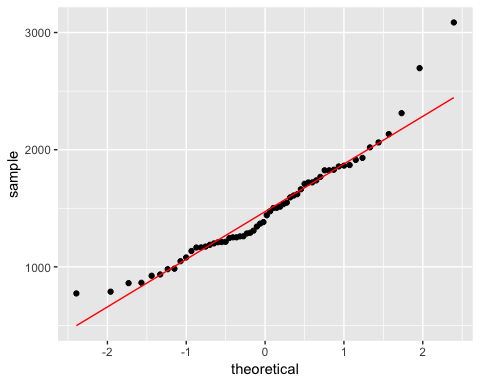
samp.describe

## vars n mean sd median trimmed mad min max range skew  
## X1 1 60 1483.77 446.94 1411 1449.42 390.67 773 3086 2313 1  
## kurtosis se  
## X1 1.62 57.7

samp.histogram



samp.qqplot



The sample appears somewhat normal but with a very strong right skew as we can tell with a skewness of 1.02. The median is 1498 with a mean of 1574.

## Exercise 2

### Would you expect another student’s distribution to be identical to yours? Would you expect it to be similar? Why or why not?

I definitely would *NOT* expect another students distribution to be identical because we pulled the descriptive statistics from a sample of 60 out of a population of 2,930. If the population is truly normal, then their sample would ideed be similar.

## Exercise 3

### For the confidence interval to be valid, the sample mean must be normally distributed and have standard error s/n‾√. What conditions must be met for this to be true?

For the confidence interval of a sample to be true, the sample must be random and independent. The sample size needs to be greater than 30, and there should be no skewness to the distribution.

## Exercise 4

### What does “95% confidence” mean? If you’re not sure, see Section 4.2.2.

The term “95% confidence” means that 95% of the time our point estimates will be within 2 standard errors of the parameter for that statistic.

## Exercise 5

### Does your confidence interval capture the true average size of houses in Ames? If you are working on this lab in a classroom, does your neighbor’s interval capture this value?

pop.mean <- mean(population.tidy$Gr.Liv.Area)   
  
#baseR  
sample\_mean <- mean(samp)  
se <- sd(samp) / sqrt(60)  
lower <- sample\_mean - 1.96 \* se  
upper <- sample\_mean + 1.96 \* se  
c(lower, upper)

## [1] 1374.759 1614.975

The population mean is 1499.6904437. The confidence interval of my sample captures the population mean, yes. I am not working on this within a classroom so I am unable to answer for my neighbors.

## Exercise 6

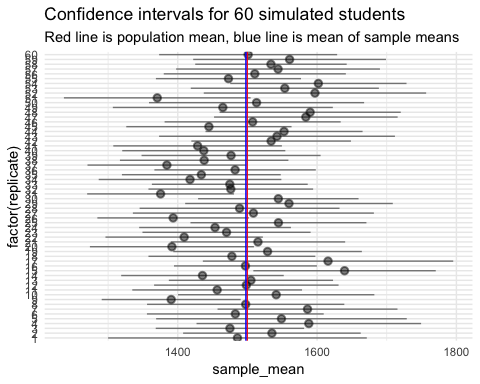
### Each student in your class should have gotten a slightly different confidence interval. What proportion of those intervals would you expect to capture the true population mean? Why? If you are working in this lab in a classroom, collect data on the intervals created by other students in the class and calculate the proportion of intervals that capture the true population mean.

We aren’t in a classroom. I would expect 95% of the intervals to capture the true population mean because that’s how confidence intervals work.

update:

ok so i want to figure out how I can do this for the entire class

#load packages  
library(tidyverse)  
library(broom)  
library(psych)  
  
#download and load file  
download.file("http://www.openintro.org/stat/data/ames.RData", destfile = "ames.RData")  
load("ames.RData")  
  
#get the population into a tidy data frame  
population.tidy <- ames %>%   
 select(Gr.Liv.Area)  
  
#what's the population mean?  
population.mean <- mean(population.tidy$Gr.Liv.Area)  
  
#run the sampling of 60 for 60 students  
class.samp.tidy <- population.tidy %>%   
 rep\_sample\_n(size = 60, reps = 60)  
  
#whats the mean for all the samples taken by the simulated 60 students?  
class.mean <- class.samp.tidy %>%   
 #group by the replicate column  
 group\_by(replicate) %>%   
 #give me an average per replicate  
 summarise(sample\_mean = mean(Gr.Liv.Area)) %>%   
 ungroup() %>%   
 #give me the average for all averages  
 summarise(sample\_mean = mean(sample\_mean))  
  
#ok, lets make a pretty picture with all of this...  
class.samp.tidy %>%   
 group\_by(replicate) %>%   
 #build our confidence intervals  
 summarise(sample\_mean = mean(Gr.Liv.Area),  
 standard\_dev = SD(Gr.Liv.Area),  
 se = standard\_dev/sqrt(60),  
 lower = sample\_mean - 1.96 \* se,  
 upper = sample\_mean + 1.96 \* se) %>%   
 #and now for the pretty picture  
 ggplot() +  
 theme\_minimal() +  
 geom\_pointrange(aes(x = factor(replicate), y = sample\_mean, ymin = lower, ymax = upper), alpha = 0.5) +   
 geom\_hline(yintercept = population.mean, color = "red") +  
 geom\_hline(yintercept = class.mean$sample\_mean, color = "blue") +  
 coord\_flip() +  
 labs(title = "Confidence intervals for 60 simulated students",  
 subtitle = "Red line is population mean, blue line is mean of sample means")



nice.