Exercise 2 - Confidence Intervals

Karlo Angelo C. Lazaro

2023-06-03

```
download.file("http://www.openintro.org/stat/data/ames.RData", destfile = "ames.RData")
load("ames.RData")
```

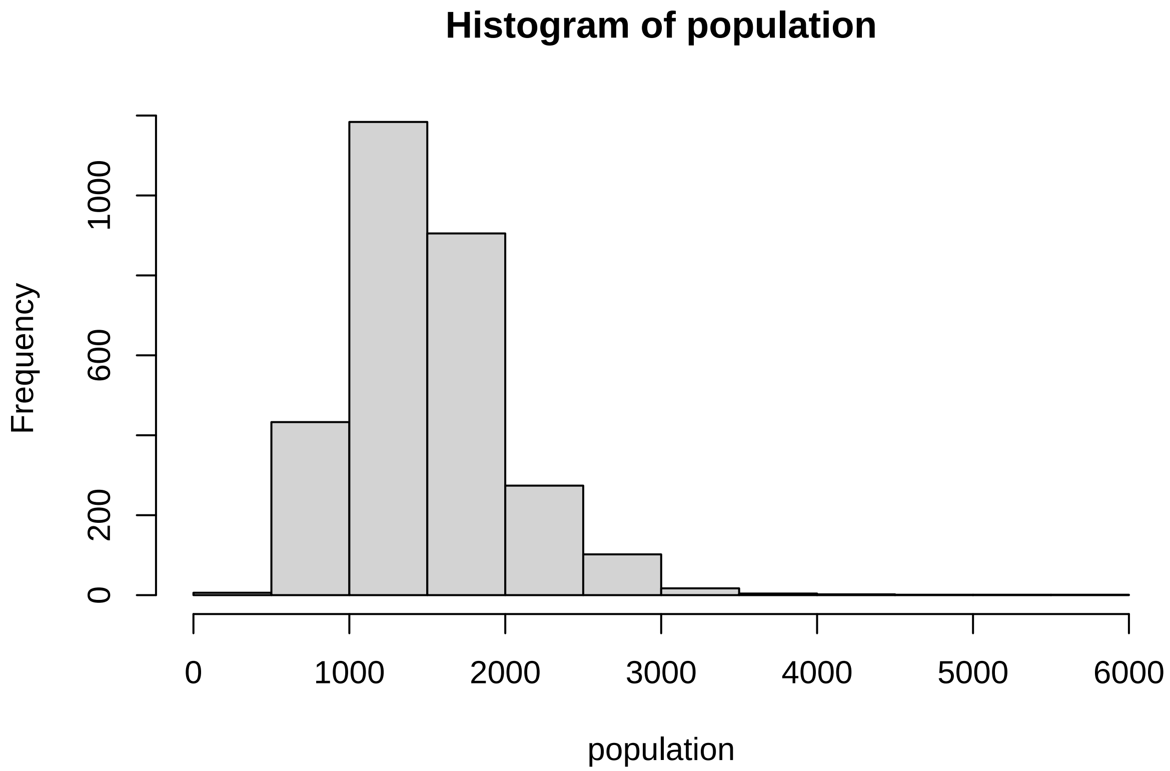
```
set.seed(13)
population <- ames$Gr.Liv.Area
samp <- sample(population, 60)
summary(population)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     334    1126    1442    1500    1743    5642
```

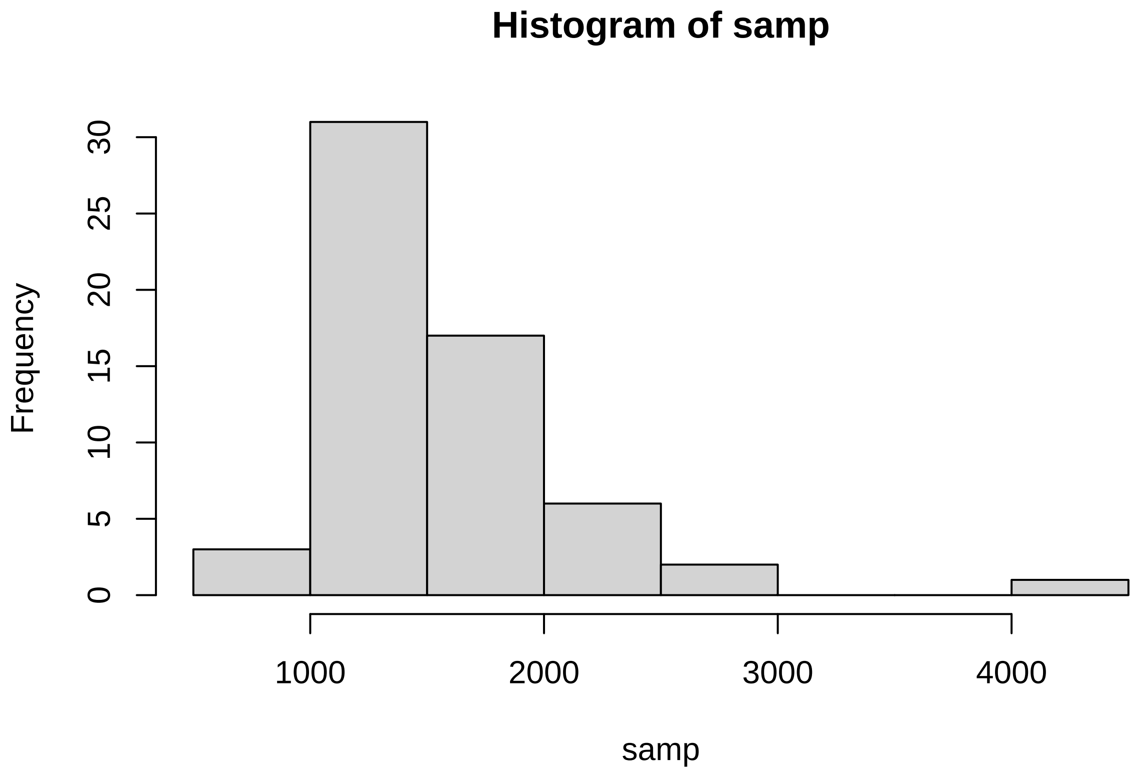
```
summary(samp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     572    1264    1430    1576    1754    4316
```

```
hist(population)
```


Histogram of population
Frequency
0
200
600
1000
0
1000
2000
3000
4000
5000
6000
population

```
hist(samp)
```



```r
getmode <- function(samp) {
   uniqv <- unique(samp)
   uniqv[which.max(tabulate(match(samp, uniqv)))]
}
result <- getmode(samp)
print(result)
```

```
## [1] 1398
```

Question 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.

Answer 1 The sample distribution is a non-symmetric unimodal distribution with a positively skewed curve. I would say that the "typical" size of the above ground living area of the house in square feet is 1398 square feet. In this instance, I interpreted "typical" as the sample mode.

Question 2 Would you expect another student's distribution to be identical to yours? Would you expect it to be similar? Why or why not?

Answer 2 I wouldn't expect another student's distribution to be exactly identical to mine, even if we are sampling from the same population since the sampling is done randomly there is always the possibility that my and my classmates' samples will differ.

However, I do expect that our sample distributions will be similar since we did draw the sample from the same population. The similarity would just depend on the size of the samples and the homogeneity of the population.

As we are only taking 60 samples from a 2000+ population, there is a chance that the samples we draw will not be representative of the population. In the case of homogeneity, if the samples we draw are not homogeneous, then it's also possible that our samples will contain different sets of house sizes.

```
sample_mean <- mean(samp)
print(sample_mean)
```

```
## [1] 1576.033
```

```
se <- sd(samp) / sqrt(60)
lower <- sample_mean - 1.96 * se
upper <- sample_mean + 1.96 * se
c(lower, upper)
```

```
## [1] 1432.896 1719.171
```

Question 3 For the confidence interval to be valid, the sample mean must be normally distributed and have standard error s/sq(n). What conditions must be met for this to be true?

Answer 3 For it to be true, the following conditions must be met: 1. Sample must be randomly selected from the population. 2. Data must be independent. 3. The population must be normally distributed. 4. Sample size is large enough.

Question 4 What does "95% confidence" mean?

Answer 4 It means that if we repeatedly take samples of size 60 from the same population, in this case the house size in Ames, then approximately 95% of these intervals will contain the true average of house size in Ames.

```
pop_mean <- mean(population)
print(pop_mean)
```

```
## [1] 1499.69
```

Question 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?

Answer 5 Yes, the confidence interval I've run captured the true average size of houses in Ames. The lower end of the interval is 1432.896 while the upper end is 1719.171. With 1499.69 as the true average house size, it is well within the CI levels we have.

Question 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?

Answer 6 Around 95% because we are doing the same experiment. We are taking samples of the same size from the same population repeatedly so approximately 95% of these intervals should contain the true population mean.

On Your Own 1 Using the following function (which was downloaded with the data set), plot all intervals. What proportion of your confidence intervals include the true population mean? Is this proportion exactly equal to the confidence level? If not, explain why.

```
set.seed(31)
samp_mean <- rep(NA, 50)
samp_sd <- rep(NA, 50)
n <- 60
```

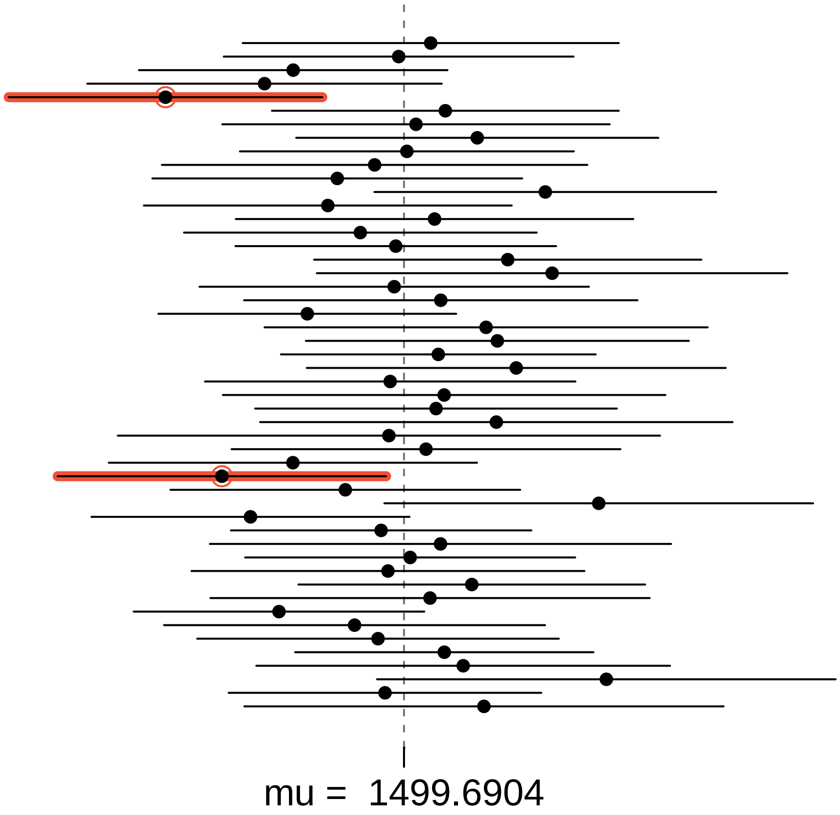
```
set.seed(31)
for(i in 1:50){
  samp <- sample(population, n) # obtain a sample of size n = 60 from the population
  samp_mean[i] <- mean(samp)    # save sample mean in ith element of samp_mean
  samp_sd[i] <- sd(samp)        # save sample sd in ith element of samp_sd
}
```

```
lower_vector <- samp_mean - 1.96 * samp_sd / sqrt(n)
upper_vector <- samp_mean + 1.96 * samp_sd / sqrt(n)
```

```
c(lower_vector[1], upper_vector[1])
```

```
## [1] 1391.372 1716.495
```

```
plot_ci(lower_vector, upper_vector, mean(population))
```



On Your Own 1 - Answer Based on the plot obtained, the proportion of CIs including the true population mean is 48/50 or 96%. In this case the proportion is slightly higher than the confidence level of 95% but quite close already. And we need to remember that the confidence interval is a range of values, and not a single value. This just means that there is a chance that the true population mean may be outside the drawn confidence interval.