Exercise5 - Confidence intervals

Shikhar Dhwaj

29 September 2017

Reading the data file

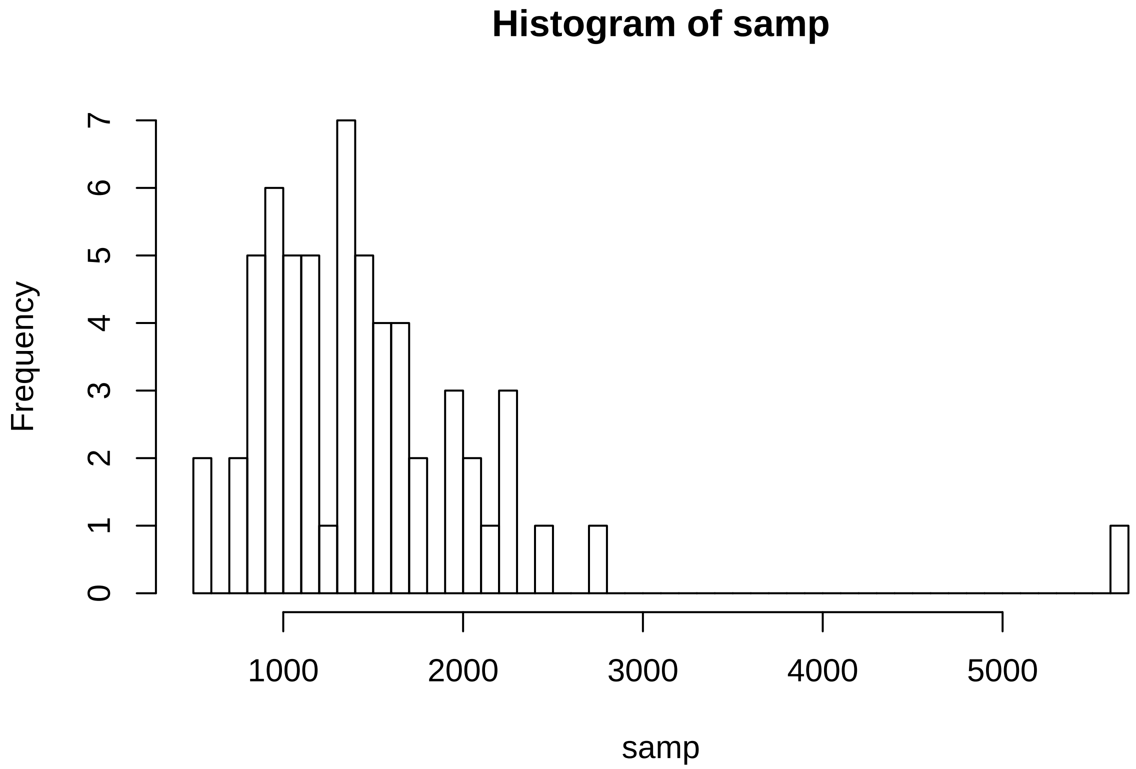
```
download.file("http://www.openintro.org/stat/data/ames.RData", destfile = "ames.RData")
load("ames.RData")
```

Takeing a sample of size 60 of the population

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

Q-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.

```
sample_mean <- mean(samp)
hist(samp,breaks = 50)
```



The plot of the sample is unimodal and skewed towards the right.More of the house areas are concentrated on the lower end of the distribution. The typical area of the houses in Ames falls under 1000 - 1500 range

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

Another student's distribution will be different. The probability is that other student's sample distribution is not similar to mine. That is because of a relative small sample size. The sample size is 60 out of over 2000 cases. Of course, on more rare occasions, the distribution may be similar.

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

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

```
## [1] 1273.407 1644.359
```

The sample consists of at least 30 independent observations and the data are not strongly skewed.

Q-4)What does "95% confidence" mean?

95% of random samples of size 60 will yield confidence intervals that contain the true average area of houses in Ames, Iowa.

Q-5)Does your confidence interval capture the true average size of houses in Ames?

```
summary(population)
```

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

Yes, if you see the mean of the population above, it does fall under lower and upper level of the sample calculated using 95% confidence.

Q-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.

95% of the students will capture the true mean in their interval. By definition, the computation for the control interval will capture the mean 95% of the time - values that are +/- 1.96 times away from the standard error.

Using R, we're going to recreate many samples to learn more about how sample means and confidence intervals vary from one sample to another.

Here is the rough outline:

1) Obtain a random sample.
2) Calculate and store the sample's mean and standard deviation.
3) Repeat steps (1) and (2) 50 times.
4) Use these stored statistics to calculate many confidence intervals.

But before we do all of this, we need to first create empty vectors where we can save the means and standard deviations that will be calculated from each sample. And while we're at it, let's also store the desired sample size as n.

```
samp_mean <- rep(NA, 50)
samp_sd <- rep(NA, 50)
n <- 60
```

Now we're ready for the loop where we calculate the means and standard deviations of 50 random samples.

```
for (i in 1:50) {
    samp <- sample(population, n)
    samp_mean[i] <- mean(samp)
    samp_sd[i] <- sd(samp)
}
```

Lastly, we construct the confidence intervals.

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

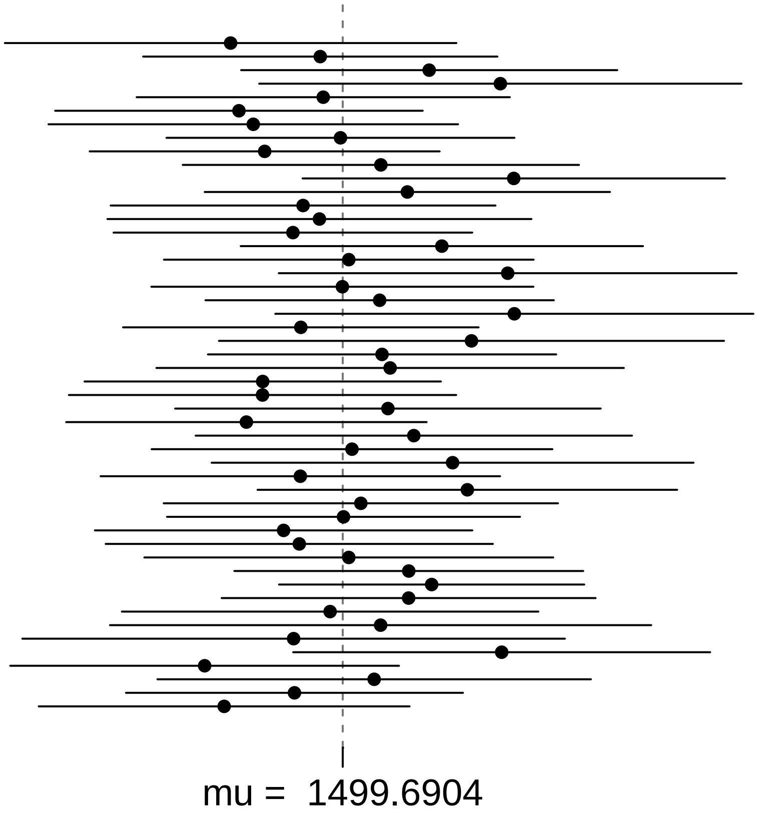
Lower bounds of these 50 confidence intervals are stored in lower_vector, and the upper bounds are in upper_vector. Let's view the first interval.

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

```
## [1] 1306.443 1542.191
```

Q-7)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.

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



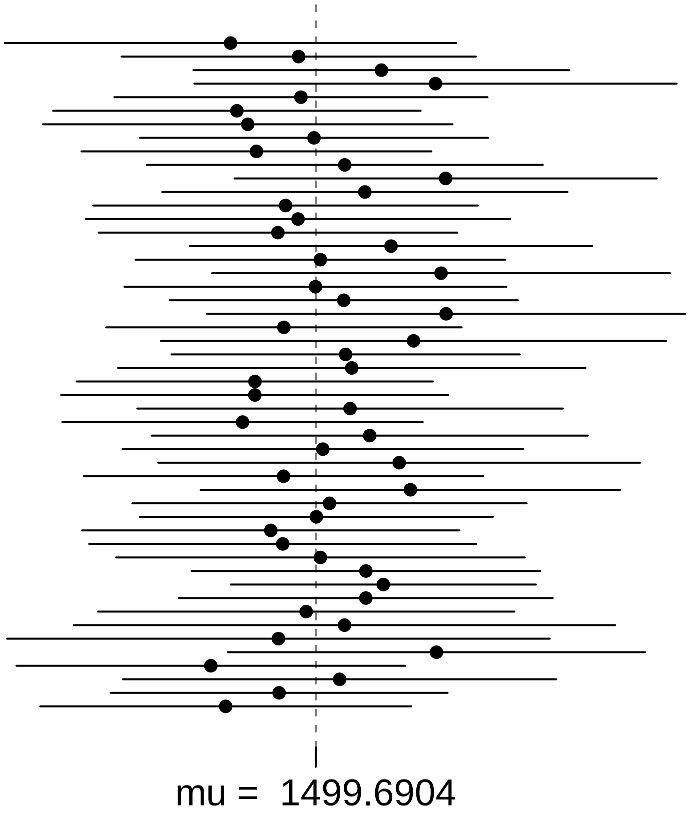
In the first run I got 44/50 i.e. 95%. The 95% confidence level does not mean that every run would have poopulation mean lying between the sample mean interval.

Q-8)Pick a confidence level of your choosing, provided it is not 95%. What is the appropriate critical value?

We can choose the confidence level of 99% by following the given formula: (Sample Mean + (+/- 2.58 * SE)).

Q-9)Calculate 50 confidence intervals at the confidence level you chose in the previous question. You do not need to obtain new samples, simply calculate new intervals based on the sample means and standard deviations you have already collected. Using the plot__ci function, plot all intervals and calculate the proportion of intervals that include the true population mean. How does this percentage compare to the confidence level selected for the intervals?

```
lower_vector <- samp_mean - 2.58 * samp_sd / sqrt(n)
upper_vector <- samp_mean + 2.58 * samp_sd / sqrt(n)
plot_ci(lower_vector, upper_vector, mean(population))
```



50/50 or 100% of the control intervals included the true mean.It is within the 99% confidence interval level