Foundations for statistical inference - Confidence intervals

Student Name : Sachid Deshmukh

Date : 11/08/2018

- GitHub Location for rmd file
- GitHub Location for pdf file
- RPubs location of published file

Sampling from Ames, Iowa

If you have access to data on an entire population, say the size of every house in Ames, Iowa, it's straight forward to answer questions like, "How big is the typical house in Ames?" and "How much variation is there in sizes of houses?". If you have access to only a sample of the population, as is often the case, the task becomes more complicated. What is your best guess for the typical size if you only know the sizes of several dozen houses? This sort of situation requires that you use your sample to make inference on what your population looks like.

The data

In the previous lab, "Sampling Distributions", we looked at the population data of houses from Ames, Iowa. Let's start by loading that data set.

```
load("more/ames.RData")
```

In this lab we'll start with a simple random sample of size 60 from the population. Specifically, this is a simple random sample of size 60. Note that the data set has information on many housing variables, but for the first portion of the lab we'll focus on the size of the house, represented by the variable `Gr.Liv.Area`.

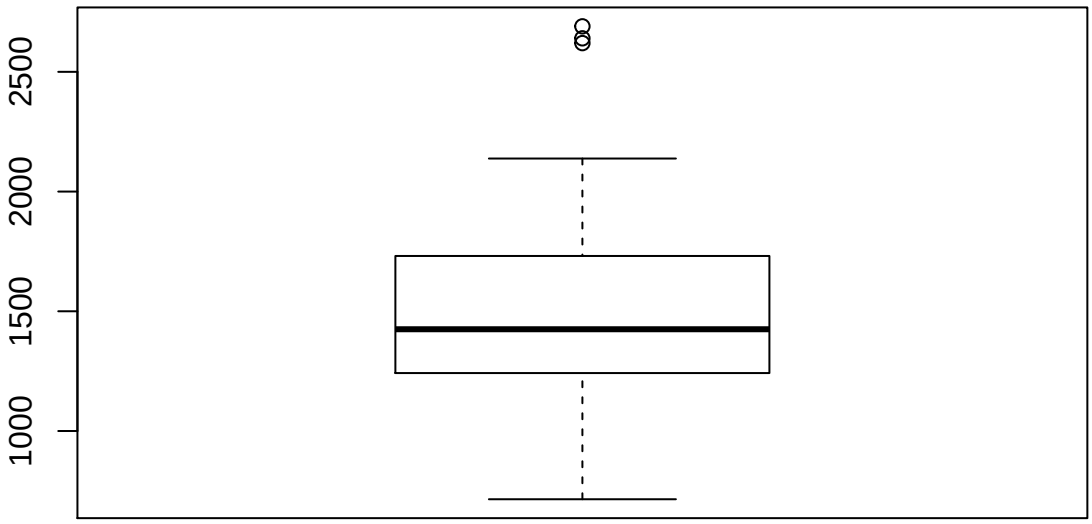
```
population <- ames$Gr.Liv.Area
samp <- sample(population, 60)
```

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.

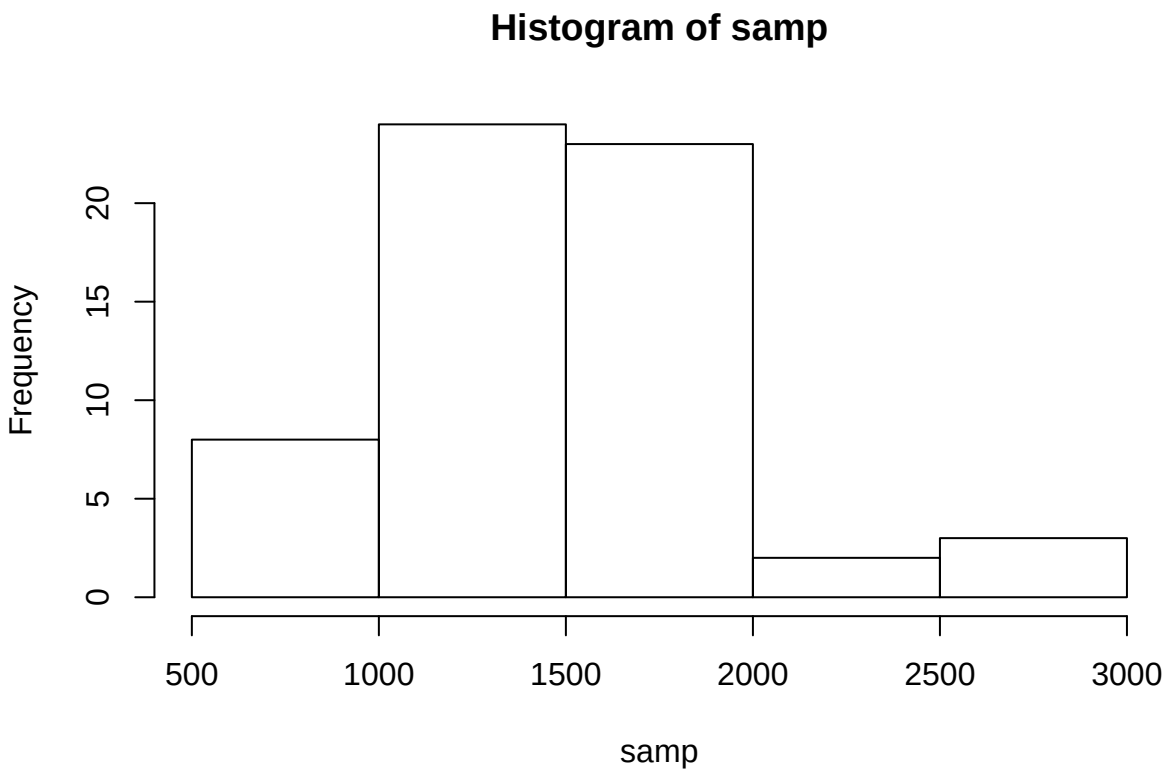
```
summary(samp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     715    1245    1425    1500    1730    2690
```

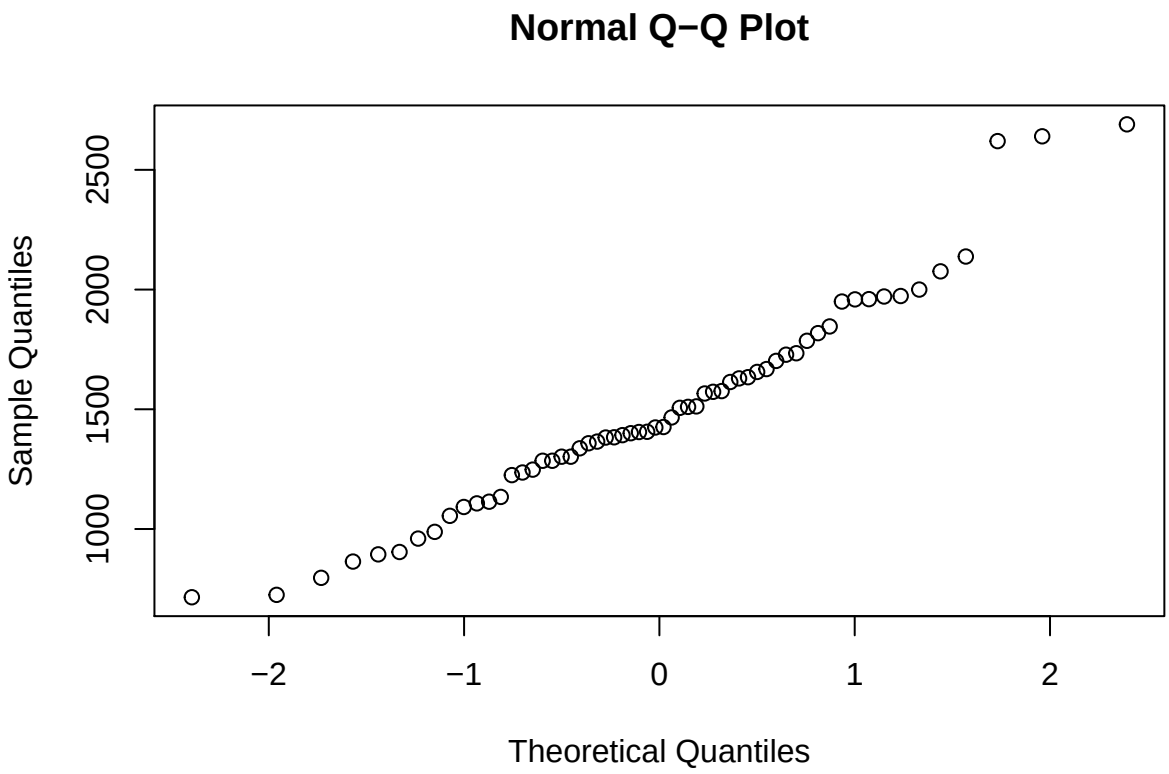
```
boxplot(samp)
```




```
hist(samp)
```


Histogram of samp
Frequency
0
5
10
15
20
500
1000
1500
2000
2500
3000
samp

```
qqnorm(samp)
```




```
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

getmode(samp)
```

```
## [1] 1302
```

Answer : Above ground living area has range of 698 to 2646 SFT. Median value is 1572 and Mean value is 1515. Mean is slightly lesser than median which indicate lightly left skewed data but skewness is not very extreme

Boxplot indicates that most of the houses have above ground living area either small (1st and 2nd quartile) or big (greater than 3rd quartile) Very few observations are sitting between (2nd and 3rd quartile) indicating smaller homes are built with medium size above ground living area.

Histogram and Normality plot suggests that sample is close to the normal distribution

I would say mode of the sample is the typical size within sample. With mode = 1646 and mean = 1515 I would say typical size of the above ground living area is higher than mean. This indicates more homes are having smaller size above ground living area. This in turn suggest left skew in the data which can be seen though histogram as well.

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 : I would expect another sample to be similar in distribution however point estimates would vary. I won't expect exact same mean and median for another sample however left skewed data which is close to

normal distribution is expected

Confidence intervals

One of the most common ways to describe the typical or central value of a distribution is to use the mean. In this case we can calculate the mean of the sample using,

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

Return for a moment to the question that first motivated this lab: based on this sample, what can we infer about the population? Based only on this single sample, the best estimate of the average living area of houses sold in Ames would be the sample mean, usually denoted as $\bar{x}$ (here we're calling it `sample_mean`). That serves as a good *point estimate* but it would be useful to also communicate how uncertain we are of that estimate. This can be captured by using a *confidence interval*.

We can calculate a 95% confidence interval for a sample mean by adding and subtracting 1.96 standard errors to the point estimate (See Section 4.2.3 if you are unfamiliar with this formula).

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

```
## [1] 1389.358 1610.975
```

This is an important inference that we've just made: even though we don't know what the full population looks like, we're 95% confident that the true average size of houses in Ames lies between the values *lower* and *upper*. There are a few conditions that must be met for this interval to be valid.

3. For the confidence interval to be valid, the sample mean must be normally distributed and have standard error $s/\sqrt{n}$. What conditions must be met for this to be true?

Answer : Following two conditions should be satisfied for above eqation to be true

- The population from where sample is drawn should be normally distributed
- Sample size should be sufficiently large for accruate estimation of the Standard Error. Statistical theory suggest that Sample size greater than 30 is desired to achieve accurate results

Confidence levels

4. What does "95% confidence" mean? If you're not sure, see Section 4.2.2.

Answer : A 95% confidence interval is a range of values that you can be 95% certain contains the true mean of the population

In this case we have the luxury of knowing the true population mean since we have data on the entire population. This value can be calculated using the following command:

```
mean(population)
```

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

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 : Mean of the entire population is 1499.69 and 95% confidence interval derived from sample is 1397.989 to 1632.644. From this observation we can conclude that confidence interval captures the true average of the houses in Ames

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.

Answer : 95% of the sample intervals should capture true population mean. Our range is 95% means there is still 5% chance that the sample mean interval will not represent the actual population mean. However this is fine for our analysis since we are aligning with 95% confidence interval and it is understood that it can be wrong 5% of the times

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. *Loops* come in handy here (If you are unfamiliar with loops, review the Sampling Distribution Lab).

Here is the rough outline:

- Obtain a random sample.
- Calculate and store the sample's mean and standard deviation.
- Repeat steps (1) and (2) 50 times.
- 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) # 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
}
```

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] 1354.405 1615.295
```

On your own

- 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.

Answer : From the plot below we can see that 1 out of 50 samples have confidence interval which does not include true population mean. This is expected since we are calculating 95% confidence interval. That means we have at most 5% chance of error. As per this statistics it is expected to have 5% of 50 samples that means

~3 samples to have confidence interval which doesn't represent true population mean. The plot is in line with our expectations.

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

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

Answer : Critical value for 90% confidence interval is c(qnorm(0.05), qnorm(0.95)) which is -1.64 to 1.64

- 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.90 <- samp_mean - 1.64 * samp_sd / sqrt(n)
upper_vector.90 <- samp_mean + 1.64 * samp_sd / sqrt(n)
plot_ci(lower_vector.90, upper_vector.90, mean(population))
```

Answer : We can see that 5 samples out of 50 have interval which doesn't include true population parameter. This is expected at 90% confidence interval. At 90% confidence interval, we have 10% chance of making error. That means we can expect 10% (10% of 50 = 5) samples to have confidence interval which doesn't include true population mean. The above observation is inline with our expectation