r_intro PS 1 solutions

Nicholus Tint Zaw

2022-11-06

Exploring the Central Limit Theorem

The CLT is the backbone for the sample survey method and states the following main concepts.

- The sampling distribution of a sample mean is approximately normal if the sample size is large enough, even if the population distribution is not normal.
- The mean of the sampling distribution will equal the mean of the population distribution.
- The standard deviation of the sampling distribution will be equal to the standard deviation of the population distribution divided by the sample size.

In this problem set, we will test those assumptions using some basic R commands.

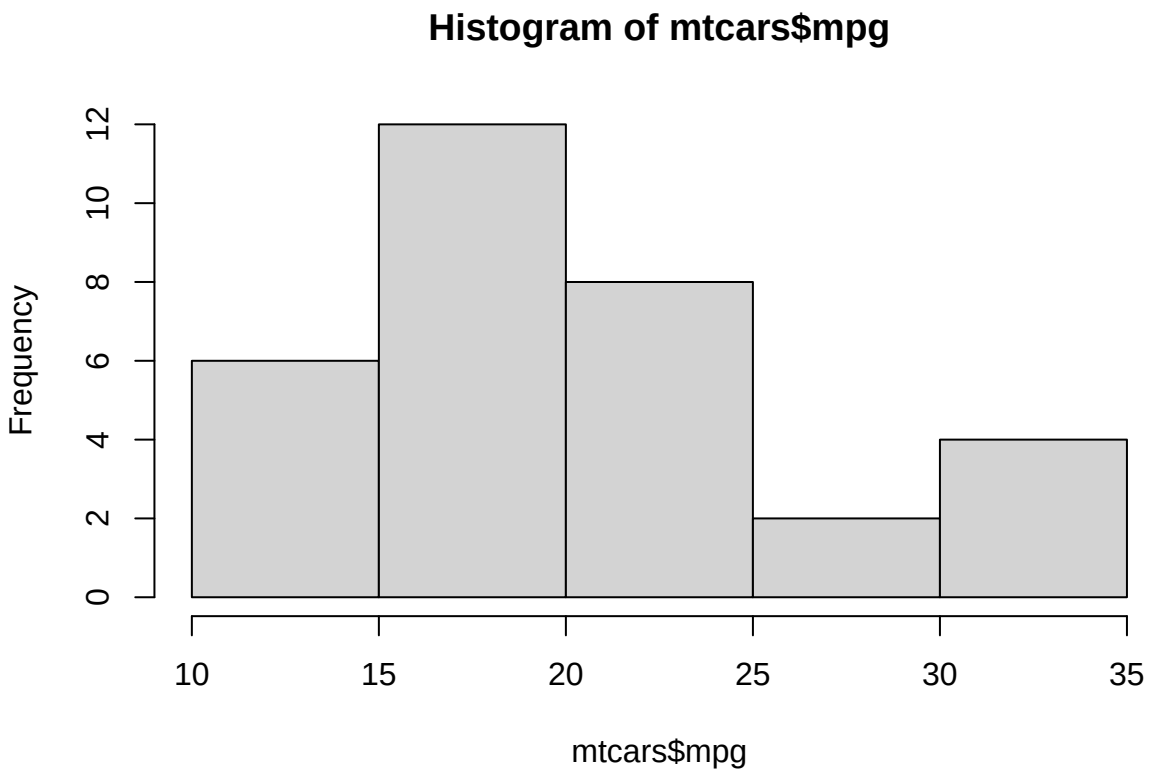
Prerequisite reading

You must finish reading this Facebook post to get a general idea about CLT and sampling distribution. Then, go to this website and perform this website to perform some tests on the concept you learned from that post. As we will implement some of those tests in the R, your understanding of those concepts is critical. So, digest it carefully.

Some useful Base-R commands for this exercise

Use `hist()` for plot histogram.

```
hist(mtcars$mpg)
```



For random sampling, use `sample()`.

```
sample(1:50, 2)
```

```
## [1]  1 14
```

For multiple-time random selection processes (replication), use this command - `replicate()`. You can type `?replicate()` to study the general description and usage in the help file.

```
# select 2 numbers from 1 to 50 and calculate mean
# and repeat that process for 100 times

replicate(n = 100, expr = mean(sample(1:50, size = 2, replace=TRUE)))
```

```
##   [1] 21.0  6.0 23.0 22.5 38.5 15.0 17.5  3.0 19.0 14.0 36.5 28.5 16.5 32.5 24.0
##  [16] 12.0 26.0 24.5 21.5 39.0 25.5 28.0 23.0 24.5 26.5 42.0 34.5 25.5 27.5 34.5
##  [31] 28.0 29.5 32.0  9.0 24.0 46.5 15.0 41.0 24.0 21.0 12.5 34.0 22.0 31.5 14.5
##  [46] 25.0 32.5 23.0 26.0 31.0 31.0 37.0 40.0  6.0  6.0 27.0 28.5 30.0 18.5 10.5
##  [61] 24.5 20.0 25.5 21.0 14.0 43.0 28.5 20.5 23.0 34.5  7.0 19.0 24.0  9.5 34.0
##  [76] 34.5 32.5 12.0 10.0 35.0 12.5 15.0 29.5 21.0 43.5  3.5 39.0 41.0 47.5 36.5
##  [91] 33.5 18.5 49.0 25.0 20.0 18.5 18.5 25.5 22.0  6.0
```

OK. Now, we are landing on the actual problem set question. We will test the claims of CLT using the `iris` dataset (built-in dataset from R).

- Please load the `iris` dataset and assign it as `df`.
- Use `Sepal.Length` variable and calculate the mean value - assigned as `pop_mean`.
- Then, select the 30 sample data points from `Sepal.Length` column and calculate the mean. This time assign the calculation result as `sample_mean`.
- Compare the `pop_mean` and `sample_mean`, and explain what you observed.
- Before moving to test the sampling distribution of means, please plot the `Sepal.Length` as a histogram plot and observe its distribution. Is it normal distribution? Please explain your answer.

Answer:

```
# load iris
df <- iris

# mean calculation
pop_mean <- mean(df$Sepal.Length)

set.seed(3432)
sample_mean <- mean(sample(df$Sepal.Length, 30))

# compare the two means
c(pop_mean, sample_mean)
```

```
## [1] 5.843333 5.963333
```

```
pop_mean - sample_mean
```

```
## [1] -0.12
```

Not normal, left skewed.

Finally, we are going to implement the sampling distribution of the means.

1. Construct the list of 100 means values from the `Sepal.Length` and each mean value should construct from 10 sample sizes. (sample size = 10, replication 100 times)
2. Assigned the result (vector with 100 numbers) as `means_list_1`.
3. Plot histogram using `mean_list_1`.
4. Calculate the mean value of `mean_list_1` and assign the result as `means_clt_1`.

Repeat the same process from numbers 1 to 4, but using different sample size and replication number this time.

- for `means_list_2`, using a sample size 30 and replication time 200. Save its mean value as `means_clt_2`.
- for `means_list_3`, using a sample size 50 and replication time 1,000. Save its mean value as `means_clt_3`.

Answer:

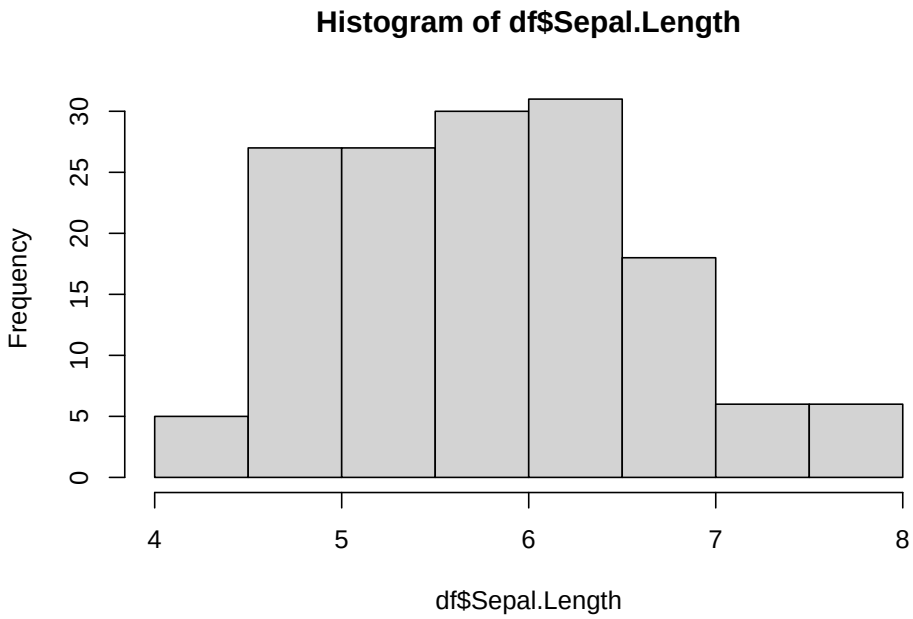


Figure 1: Sepal Length Distribution

```
set.seed(2435)
means_list_1 <- replicate(n = 100,
                          expr = mean(
                            sample(df$Sepal.Length,
                                   size = 10,
                                   replace=TRUE)))


means_list_2 <- replicate(n = 200,
                          expr = mean(
                            sample(df$Sepal.Length,
                                   size = 30,
                                   replace=TRUE)))


means_list_3 <- replicate(n = 1000,
                          expr = mean(
                            sample(df$Sepal.Length,
                                   size = 50,
                                   replace=TRUE)))
```

Compare the 4 histograms and 4 different mean values and explain whether or not the CLT's claims are working well.

What I mean by 4 is the histograms and mean values from the following 4 distributions.

1. all `Sepal.Length` observations from iris dataset - as population

2. means_list_1
3. means_list_2
4. means_list_3

Answer:

```
means_compare <- c("population mean" = pop_mean,
                   "means_list_1" = mean(means_list_1),
                   "means_list_2" = mean(means_list_2),
                   "means_list_3" = mean(means_list_3))
means_compare
```

```
## population mean    means_list_1    means_list_2    means_list_3
##        5.843333        5.848500        5.845133        5.846978
```

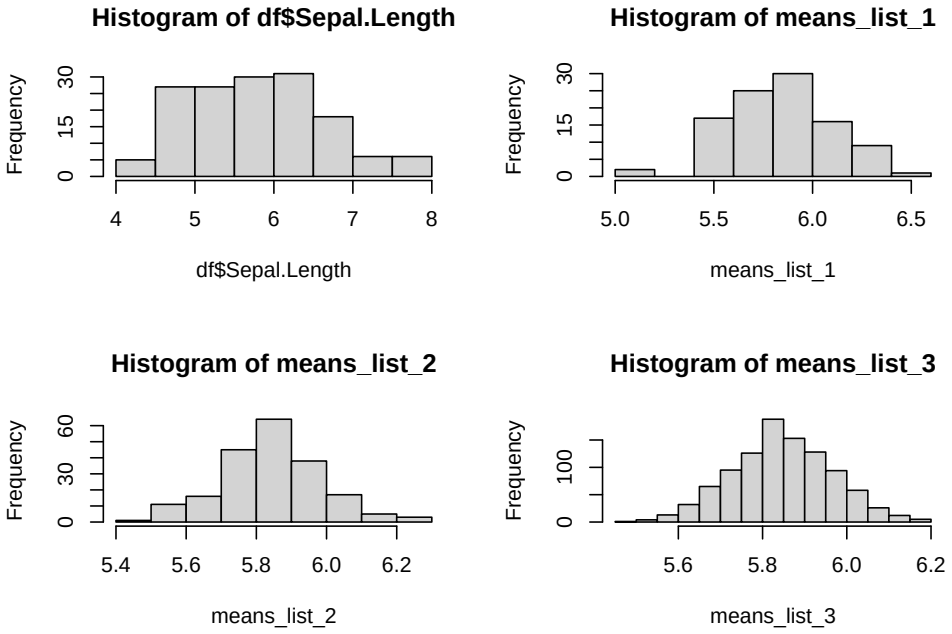


Figure 2: Sepal Length Comparisions
