PS4

Pooja Sadarangani

2022-11-06

collaborators: "Pranali Oza"

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

#1. Compare differently distributed data

1.1 Human Bodies

1. You'll work about human heights. What kind of measure is this? (nominal, ordered, difference, ratio)? How should it be measured (continuous, discrete, positive. . .)?

Ans. Human height is a ratio measure as this is a numerical quantity that has a well-defined zero. It can compute difference as well as ratio. For instance, we can say that A is twice as tall as B after looking at their heights.

Human height is measured as positive and continuous.

2. Load the fatherson.csvDownload fatherson.csv data. It contains two columns, father's height and son's height (in cm). Let's focus on fathers here (variable fheight) and ignore the sons. Provide the basic descriptives: how many observations do we have? Do we have any missings? Any unreasonable values?

```
## The number of observations are: 1078
```

```
## [1] "There are no missing values"
```

```
## [1] "There are no unreasonable values"
```

3. Compute mean, median, standard deviation and range of the heights. Discuss the relationship between these numbers. Is mean larger than median? By how much (in relative terms)? How does standard deviation compare to mean?

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
## The mean is:  171.9252
```

```
## The median is:  172.1
```

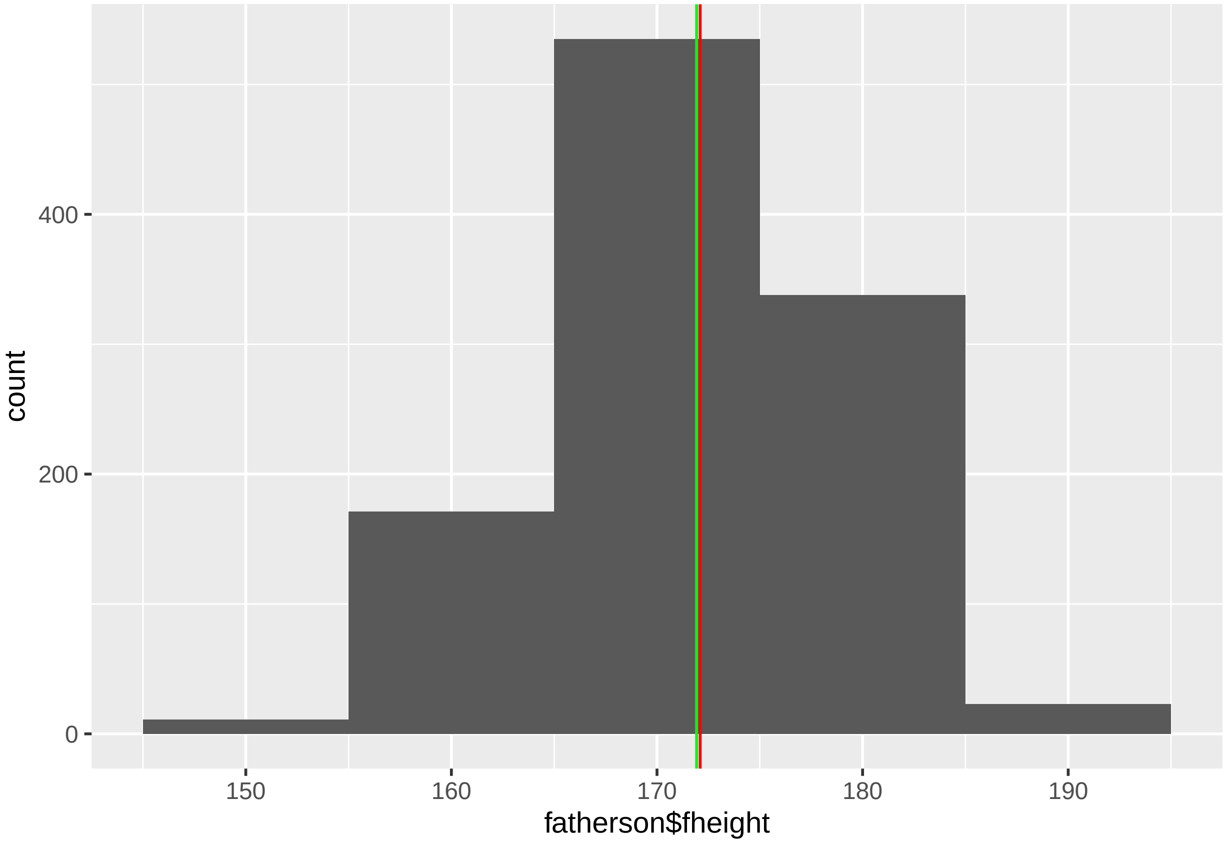
```
## The standard deviation is:  6.972346
```

```
## The range is:  149.9 191.6
```

The values of mean and median are very close to each other. The median is larger than the mean by 0.1748. The standard deviation is much lesser than the mean which suggests that the data points are close to the mean.

4. Plot a histogram of the data. Add to this histogram mean and median. You can use vertical lines of different color along the lines. What did you find? Which distribution does the result resemble?

```
## Warning: Use of `fatherson$fheight` is discouraged. Use `fheight` instead.
```



Ans: From the graph, we can tell that the mean and median values are very close to each other. The result resembles a normal distribution.

1.2 Human influence

1. What kind of measure is this? What kind of valid figures would you expect to see (continuous, discrete, positive, . . .)

Ans: Paper id -> This variable would be nominal measure as the paper id's will not necessarily have ordered values. I expect to see discrete and positive values in this column.

Citations (Number of times the paper has been cited)-> This variable would be ratio as we can find the difference between values and define them as a ratio. I expect to see discrete and positive (and zeros) values in this field.

2. Read the mag-in-citations.csvDownload mag-in-citations.csv data. Provide the basic descriptives: how many observations do we have? Do we have any missings? Do we have implausible or wrong values? What is the range of the citations?

```
##    paperId citations
## 1  4090687         2
## 2  6537979         2
## 3  7484482         4
## 4  9444380         3
## 5 14056478         5
## 6 14498457         2
```

```
## The number of observations are: 388258
```

```
## [1] " There are no missing values in any of the columns"
```

```
## [1] "There are no unreasonable values in any of the columns"
```

3. Compute mean, median, mode (the most frequent value), standard deviation and range of the number of citations. Discuss the relationship between these numbers. Is mean larger than median? Than mode? By how much (in relative terms)? How does standard deviation compare to mean?

```
## The mean is:  15.61223
```

```
## The median is:  3
```

```
## The mode is:  0
```

```
## The standard deviation is:  78.39079
```

```
## The range is:  0 18682
```

Mean is greater than the Median by 12.61223

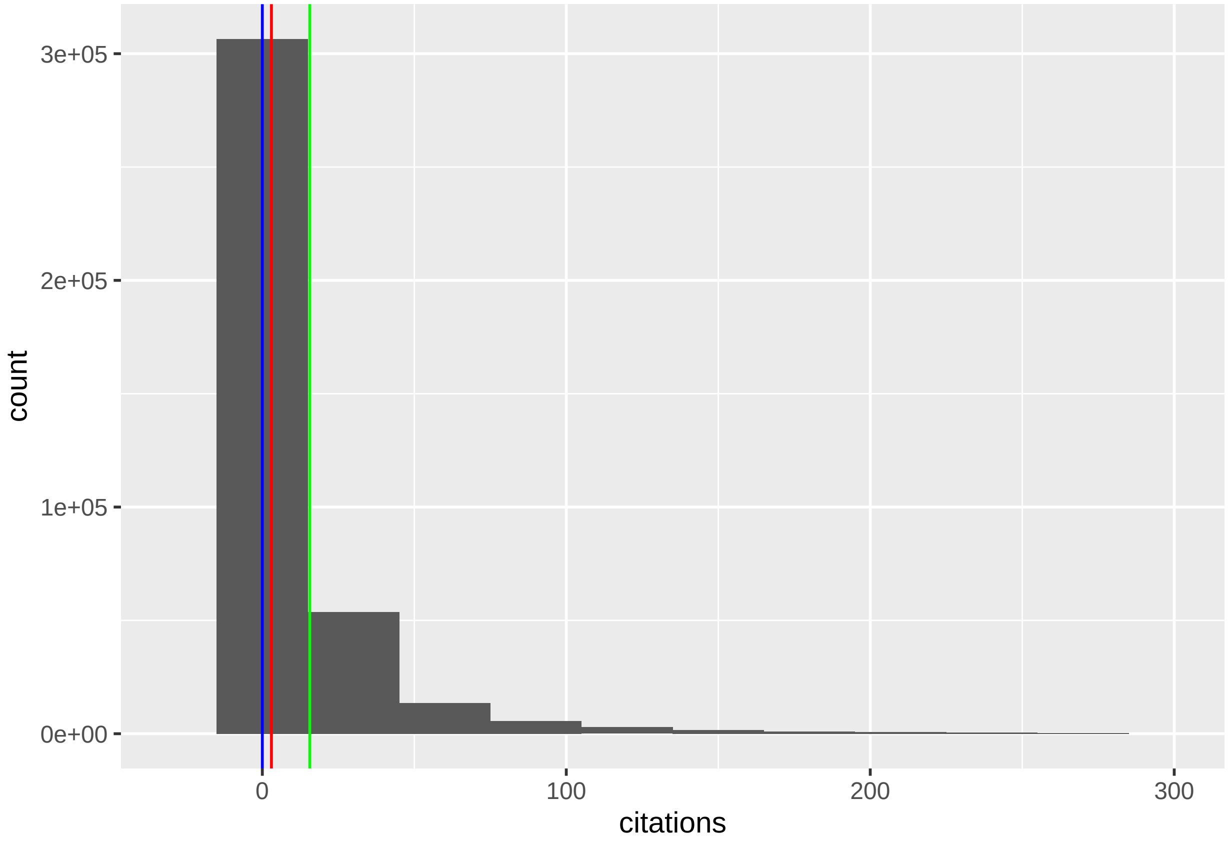
Mean is greater than the Mode by 15.61223

Median is geater than the Mode by 3

Standard Deviation is much greater than the mean by 62.77856 (i.e Standard Deviation is almost 4 times the Mean). This means that the data points widely spread and further away from the mean.

4. Plot a histogram of the data. Add to this histogram mean, median, and mode. You can use vertical lines of different color. How does the histogram look like? Which distribution does it resemble? Can you get it to be a nice and easy to grasp image?

```
## Warning: Removed 1604 rows containing non-finite values (stat_bin).
## Warning: Removed 2 rows containing missing values (geom_bar).
```



Ans: The Histogram looks right-skewed. It represents a right-skewed normal distribution.Yes, I personally think that it is a nice and an easy to grasp image.

4. Finally, comment on your findings about human bodies and influence.

Ans: Human heights are normally distributed but the shape of the citiations graph seems right skewed as most of the papers have not been citied (Mode = 0).

2. Explore Central Limit Theorem

1. Calculate the expected value and variance of this random variable.

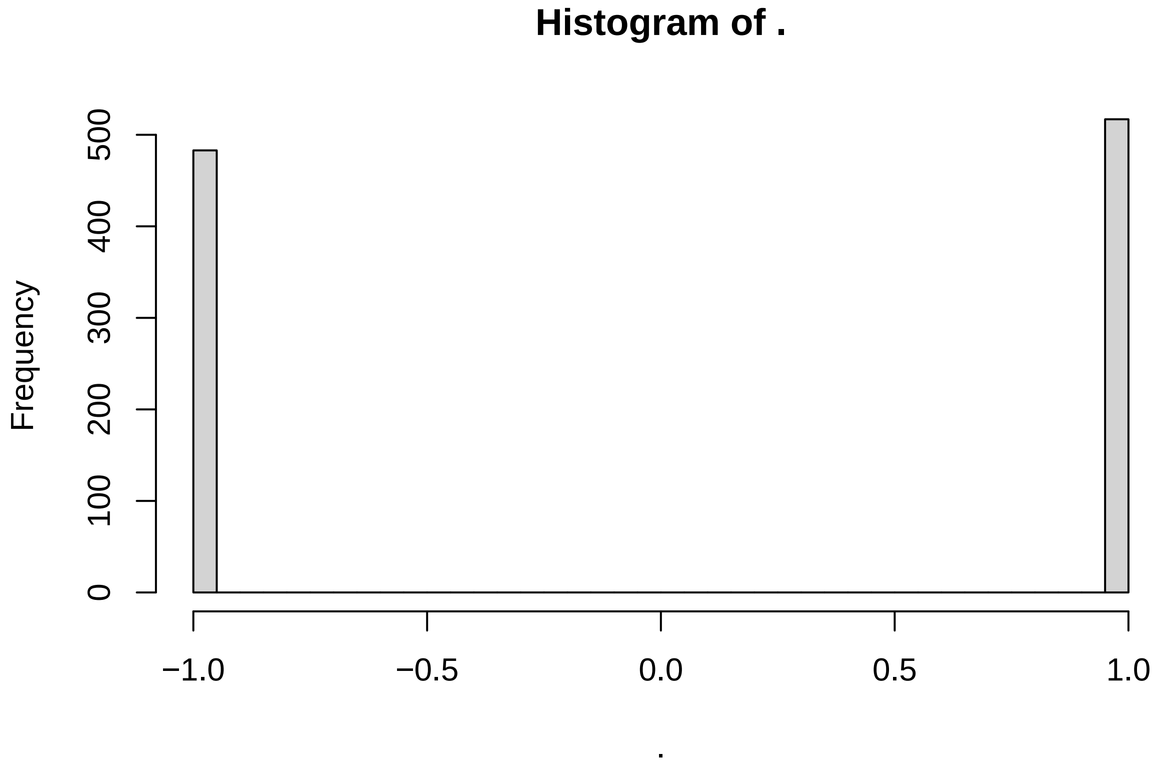
```
## The expected value should be 0

## The mean is  -0.4

## The variance is:  1
```

2. Choose your number of repetitions R. 1000 is a good number. Be clear the number of repetitions R is not the same as sample size S below!

3. Create a vector of R random numbers as explained above. Make a histogram of those. Comment the shape of the histogram.



Ans: Shape of the histogram is not very clear as the bars are present for only 2 values (-1 and 1)

4. Compute and report mean and variance of the random numbers you created (just use mean and var functions). Compare these numbers with the theoretical values computed in question 1.

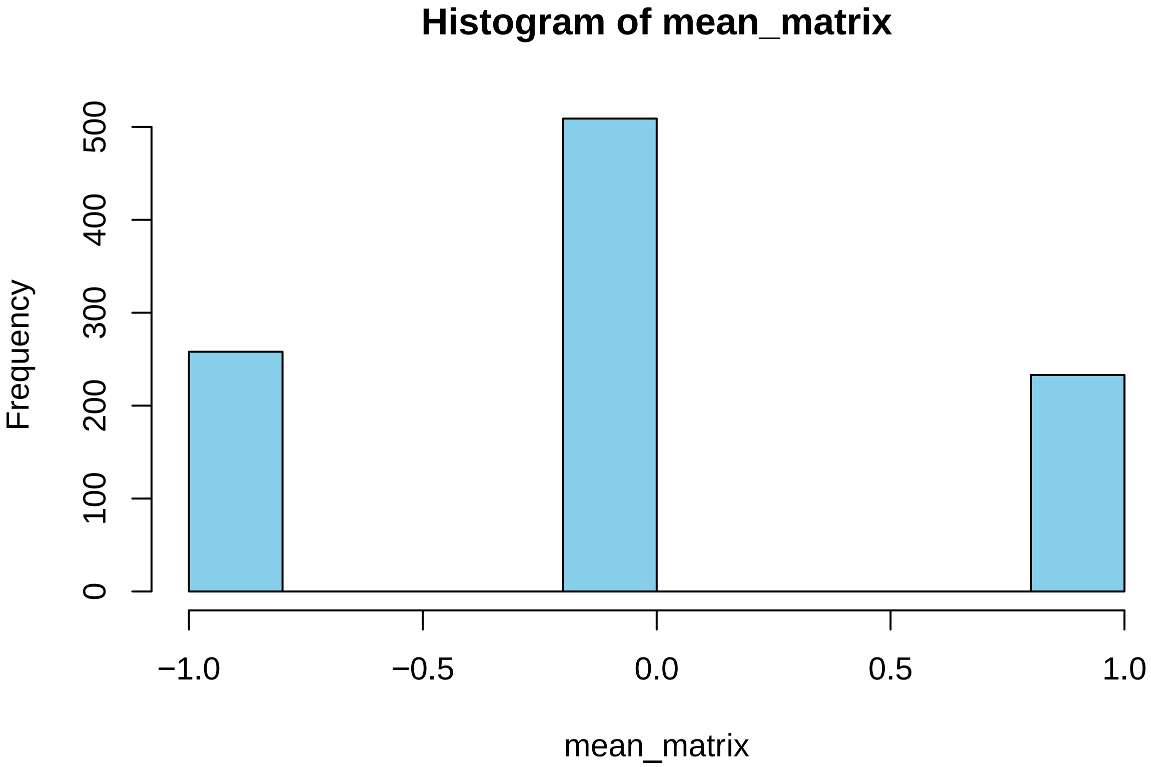
```
## The theoretical mean is: -0.4
```

```
## The mean is 0.034
```

```
## The variance is 0.9998438
```

Ans: Mean is slightly greater than that of the sample created in question 1 and variance is slightly lesser than that of the sample created in question 1. However, the values for both are very close to the theoretical values computed in question 1.

5. Now create R pairs of random numbers. For each pair, compute its mean. You should have R means. Make histogram of the means. How does this look like?



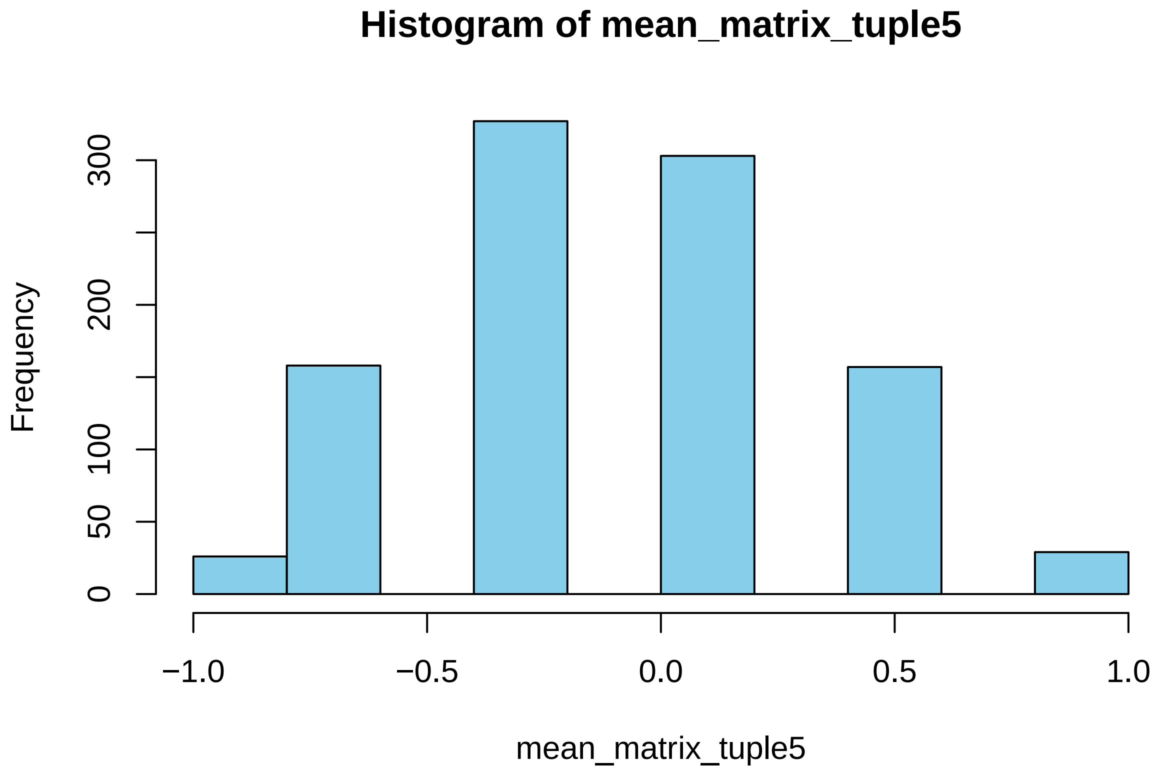
## This histogram is inclining towards a normal distribution.

6. Compute and report mean of the pair means, and variance of the means. Compare these numbers with the theoretical values. You need to adjust the values you computed in question 1 for sample size based on CLT. Remember, CLT tells that the variance now should be just 1/2 of that of X as for pairs S = 2

```
## The mean of the pair mean is:  -0.025
```

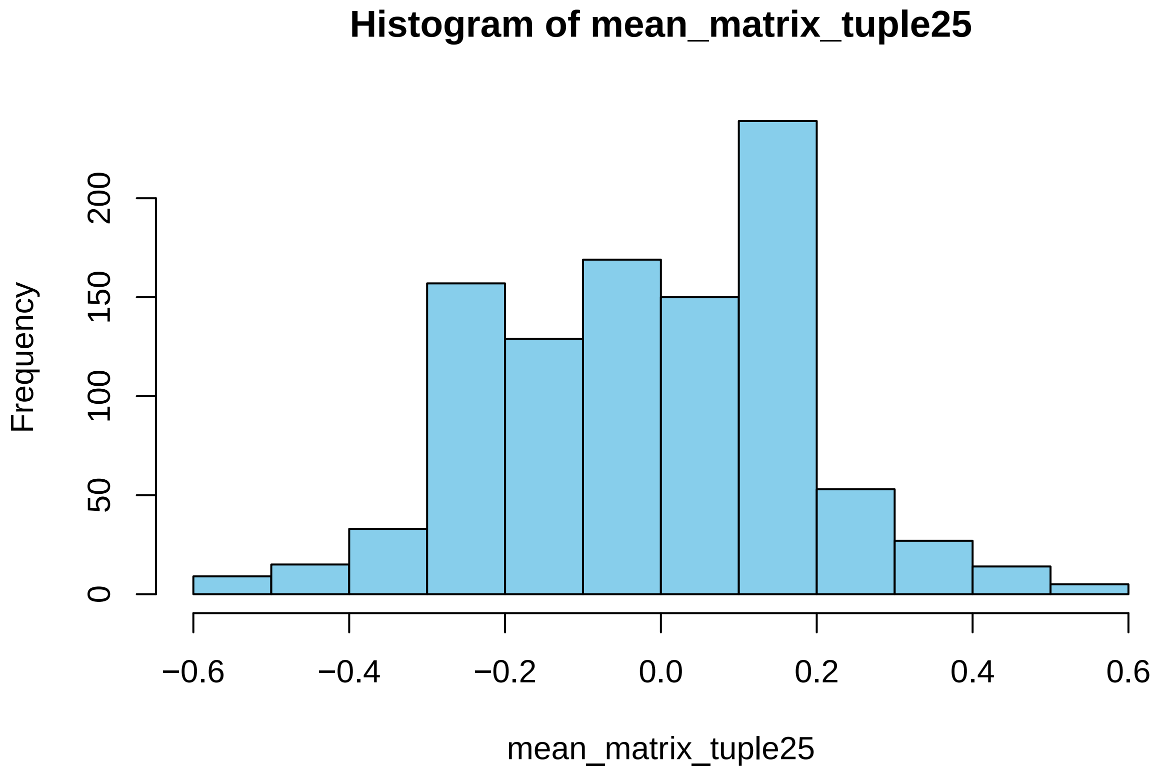
```
## The variance of the pair mean is:  0.4908659
```

7. Now instead of pairs of random variables, repeat this with 5-tuples of random numbers (i.e., 5 random numbers per one observations instead of a pair). Do you spot any noticeable differences in the histogram?

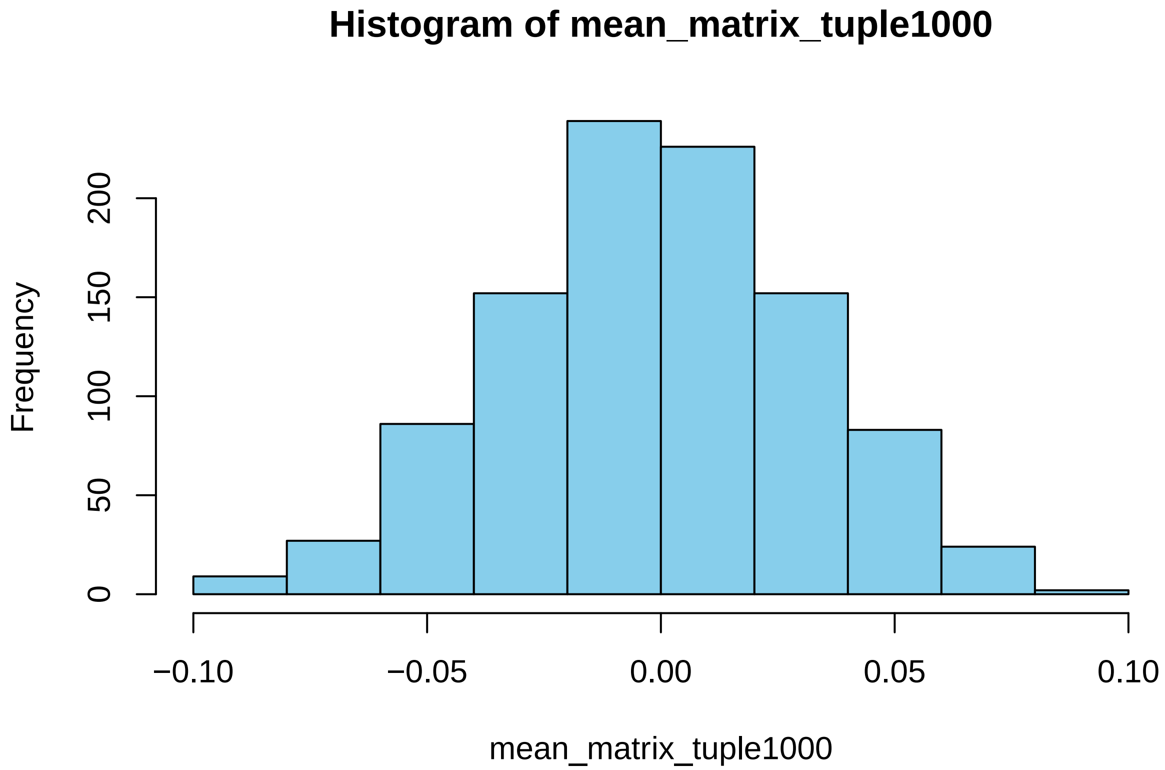


## The normal distribution is more evident in this histogram as compare to the previous ones.

8. Repeat with 25-tuples...



9. . . . and with 1000-tuples.



10. Comment on the tuple size, and the shape of the histogram.

As the tuple size increases, the shape histogram becomes more normal.

11. Explain why do the distribution becomes to look more and more normal as we take mean of a large sample of individual values. In particular, explain what happens when we move from single values S = 1 to pairs S = 2. Why did two equal peaks turn into a the above shaped histogram?

The CLT calculates the probability of an event based on the sample. As the sample size grows, so does the accuracy of this probability. This is independent of the distribution type.Thus, with the increase in sample size, the distribution becomes more normal.