**Zachary Shin**

**October 8, 2020**

**Sampling Distribution (of Mean) & The CLT: Rolling Dice Exercises**

**As a result of completing this exercise you should be able to:**

* Explain what a *sampling distribution* is
* Distinguish between *distribution of a sample* and a *sample distribution*
* Simulate sampling distribution and the Central Limit Theorem with R
* Explore the Central Limit Theorem
* Explain what the Central Limit Theorem is

**Instructions**

There are **three section**s in this Lab Exercise. Each student will perform his/her own experiment with real or virtual dice in section one, conduct simulation using R in section two, and answer the questions about the Central Limit Theorem (CLT) in section three.

**Section I: Rolling Dice Experiment by Hand using real dice or virtual dice**

In this section, please do the experiment by “hand”. If you have real fair dice (we need three), you can use and roll them to complete this section. If you don’t have dice or prefer do the experiment virtually, you can roll virtual dice online at <https://www.random.org/dice/>

1. Roll *one* fair die 12 times. Record the outcome of the die. Then use the data here to create tables and plots as well as to answer the following questions.

|  |  |  |  |
| --- | --- | --- | --- |
| Try | Outcome (e.g., 6) | Try | Outcome (e.g., 6) |
| 1st time | 3 | 7th time | 5 |
| 2nd time | 4 | 8nd time | 6 |
| 3rd time | 1 | 9rd time | 5 |
| 4th time | 2 | 10th time | 4 |
| 5th time | 4 | 11th time | 6 |
| 6th time | 3 | 12th time | 4 |

1. Create the **frequency table** and draw a bar plot to show the frequency of the outcome in question 6. (The plot shows the distribution of a sample). Then, *show the plot and describe what the distribution looks like below (see instructions).*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Die Toss Value | 1 | 2 | 3 | 4 | 5 | 6 |
| # of Tosses/Tries | 1 | 1 | 2 | 4 | 2 | 2 |

Now, using R to draw the plot based on the frequency table you just created:

* Create a vector and manually insert the six values in the above frequency table into a vector called dice.value.table. Put the six values within the “c” function. **Please replace <value #> with your own values.** For example, if some student has 2,2,2,2,2,2 in the above table, his/her R code will be dice.value.table <- c(2,2,2,2,2,2)

# create a vector to contain the frequency. Please replace <value #> with your own values from above frequency table

dice.value.table <- c(<value 1>, <value2>,……,<value5>, <value6>)

# name the vector with the categories

names(dice.value.table) <- c("1","2","3","4","5","6")

* **Plot a bar plot to show the distribution of the outcome of rolling dice** with the frequency table above using the vector (dice.value.table) you created. (Label the plots with proper title, x label, y label, and units if applicable)

**Snapshot or show your R codes here**

Chart, scatter chart

Description automatically generated

**Snapshot or insert your plot here**

Chart, bar chart

Description automatically generated

***What does the distribution look like (e.g., shape, modality, …etc.)?***

The distribution is possibly unimodal because the high frequency of number four.

1. This time let roll *three fair dice* together for 20 times. Each time record the outcome of each die in the corresponding row and in one of the first three columns. Then, calculate the average, the mean of those three outcomes and type that in the last column *(Please keep at least 2 digits after the decimal point.)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Try | Outcomes of the three dice (e.g., 3, 5, 2) | | | Mean of the Outcomes (e.g., (3+5+2)/3) |
| Example: | 3 | 5 | 2 | 3.33 |
| 1st time | 2 | 4 | 4 | 3.33 |
| 2nd time | 5 | 2 | 1 | 2.67 |
| 3rd time | 5 | 6 | 1 | 4 |
| 4th time | 3 | 4 | 6 | 4.33 |
| 5th time | 2 | 4 | 2 | 2.67 |
| 6th time | 2 | 6 | 2 | 3.33 |
| 7th time | 4 | 3 | 5 | 4 |
| 8nd time | 3 | 2 | 5 | 3.33 |
| 9rd time | 5 | 2 | 3 | 3.33 |
| 10th time | 3 | 4 | 1 | 2.67 |
| 11th time | 3 | 1 | 5 | 3 |
| 12th time | 1 | 6 | 1 | 2.67 |
| 13th time | 1 | 3 | 1 | 1.67 |
| 14th time | 6 | 1 | 1 | 2.67 |
| 15th time | 2 | 2 | 6 | 3.33 |
| 16th time | 3 | 2 | 1 | 2 |
| 17th time | 6 | 1 | 1 | 2.67 |
| 18th time | 5 | 1 | 5 | 3.67 |
| 19th time | 2 | 5 | 3 | 3.33 |
| 20th time | 4 | 5 | 6 | 5 |

1. Create a histogram to show the sampling distribution of mean of your twenty dice outcomes. Describe what *the distribution looks like?*

Using R to draw the plot:

* Create a vector and manually insert the **20 means in last column** in the above table into a vector called dice.means. Put the **20 values** within the “c” function. **Please replace <value #> by your own values of the means (the 20 values in the last column) from the above table. In the c function, you will type the 20 numbers and separate them by commas in the function.**

# create a vector to contain the means. Please replace <value #> with your own values by the means (the 20 values in the last column) from the above table

dice.means <- c(<value 1>, <value2>,…… ,<value19>, <value20>)

* **Plot a histogram to show the sampling distribution of the mean** from your experiments of rolling dice using the vector (dice.means) you created. (Label the plots with proper title, x label, y label, and units if applicable)

Snapshot or insert your plot here

**Snapshot or show your R codes here**



**Snapshot or insert your plot here**

*Chart, histogram

Description automatically generated*

***What does the distribution look like (e.g., shape, modality, …etc.)?***

The distribution is unimodal with the one hump in between 2.5-3.0.

**Section II: Simulation of rolling dice**

In this section, we **use R to simulate the sampling distribution of rolling dice** with **large sample size (number of dice per try)** and **the number of tries (# of time to try)**. We will use different sample size in our simulations (i.e., 1, 3, and 100) to explore how the data distribution changes based on the sample size. The R codes to use are as below. In this example, the sample size is set as 1 and tries is set as 1000 (the setting we use in question 4.)

# assign sample size (# of dice per try) and the number of tries (# of times tossing the dice)

sample\_size <- 1

tries <- 1000

# create simulation of outcome of rolling sample\_size dices tries times

rolls <- replicate(tries, sample(c(1:6), sample\_size, replace=T))

# get the mean of the 100 dices in each try

if (sample\_size > 1){

roll.means <- apply(rolls,2,mean)

}else{

roll.means <- rolls

}

# summarize roll.means

summary(roll.means)

1. Let’s first **set the sample size as 1 and the number of tries as 1000**. Calculate the mean of each batch of samples. Present the summary of your means, the roll.means, using summary function.

**Snapshot or show your R codes here**

**Text, letter

Description automatically generated**

**Snapshot and paste the summary of the means here**

A picture containing text

Description automatically generated

1. **Draw a histogram of the means you get from the simulated data** above (the variable called roll.means). (Label the plots with proper title, x label, y label, and units if applicable)

hist(roll.means, main = "Sampling Distribution of Mean: Rolling Dice", xlab = "Mean of Die Outcomes", ylab = "Frequency")

Chart, bar chart

Description automatically generated

1. Use R to simulate the sampling distribution of rolling dice with **sample size (number of dice) as 3 and the number of tries as 1000**. Calculate the mean of each batch of samples. Present the summary of your means, the roll.means, using summary function. (Hint: modify the sample size variable in the simulation codes)

**Snapshot or show your R codes here**

**A picture containing text

Description automatically generated**

**Snapshot and paste the summary of the means here**

A picture containing text

Description automatically generated

1. **Draw a histogram of the means** you get from the simulated data with sample size as 3. (Label the plots with proper title, x label, y label, and units if applicable)

Chart, histogram

Description automatically generated

1. Let’s simulate the sampling distribution of rolling dice **with large sample size (number of dice) as 100** and **the number of tries as 1000**. Calculate the mean of each batch of samples. Present the summary of your means, the roll.means, using summary function. (Hint: modify the sample size variable in the simulation codes)

**Snapshot or show your R codes here**

**A picture containing text

Description automatically generated**

**Snapshot and paste the summary of the means here**

A picture containing text

Description automatically generated

1. **Draw a histogram of the means** you get from the simulated data with large sample size as 100. (Label the plots with proper title, x label, y label, and units if applicable)

Chart, histogram

Description automatically generated

1. Comparing those histograms with different sample sizes (i.e., 1 die, 3 dice, and 100 dice), **what trend do you observe? Explain here.**

The shape of the histogram becomes more structured. With 1 die sample size, the unimodal graph is not cohesively shaped. As there are more dice, the unimodal trend can be seen clearly as shown with the 100 dice histogram.

**Section III: The Central Limit Theorem**

1. How would you explain the CLT to someone who do not know statistics?

CLT states that any collection of samples with an average and a quantity for the whole data will be normally distributed as the sample gets larger.

1. In addition to the mean, does the CLT apply to the sum, the variance, the proportion?

CLT can apply to sum when making larger samples and taking their sums. I don’t think CLT can apply to the variance since CLT must have a distribution with a mean and the variance itself. CLT provides info for the sampling distribution for proportions.