CLO	Description	PLO mapping	Percentage	Marks	Score
CLO3	Provide solution to the related problems of real life data in various disciplines.	PLO2: Cognitive Skills and Functional work skills with focus on Numeracy skills C5: Synthesis	10%	50	

BACKGROUND

To make inferences about population parameters, we need to know the probability distributions for certain statistics, functions of the observable random variables in the sample (or samples). These probability distributions provide models for the relative frequency behaviour of the statistics in repeated sampling; consequently, they are referred to as **sampling distribution**. We have seen that the normal, χ^2, t and F distributions provide models for the sampling distributions of statistics used to make inferences about the parameters associated with normal distributions.

PROBLEM STATEMENT

Sampling Distributions and the Central Limit Theorem

1. Basic: Illustrates that sampling distribution of the means from a variety of distributions has a normal distribution.
2. 2. Sample Size and Statistic: Allows comparison of the effects of sample size on the distribution of the mean and other sample statistics from various population distributions.

Thus, there is need to determine and investigate the impact of sampling distribution and sample size on representing the population.

CASE STUDY 1 (25 Marks)

A balanced die is tossed three times. Let Y_1, Y_2 and Y_3 denote the number of spots observed on the upper face for tosses 1, 2 and 3, respectively. Suppose we are interested in

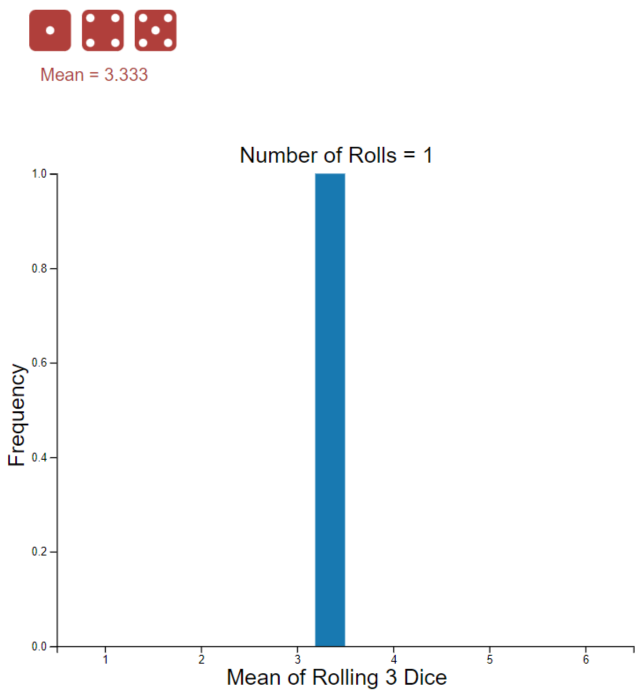
$$\bar{Y} = \frac{Y_1 + Y_2 + Y_3}{3},$$

where Y denotes the number of spots observed on the upper face on a single toss of a balanced die. The following information are given:

$$P(Y = i) = \frac{1}{6}, \qquad i = 1,2,\ldots,6$$

i) Click on the "Rolls of 3 Fair Dice". Use the button "1 Roll" to take a sample of size 3 from the die-tossing population. Discuss the observation on the output by considering the sample mean and the range of the value obtained.

(6 Marks)



Viz 1

Statistical Report on Dice Rolls

Pop Prob: (1) 0.167 (2) 0.167 (3) 0.167 (4) 0.167 (5) 0.167 (6) 0.167

Population: Mean = 3.500 StDev = 1.708

Samples = 1 of size 3

Samples: Mean = 3.333 StDev = NaN

Mean ± 1 StDev: 0.000

Mean ± 2 StDev: 0.000

Mean ± 3 StDev: 0.000

Close

Figure 1

Based on the Figure 1, we know that

$$\text{Population mean, } \mu = 3.500$$

$$\text{Population variance, } \sigma^2 = 2.9173$$

$$\text{Population standard deviation, } \sigma = 1.708$$

$$\text{Sample Mean, } \bar{Y} = 3.333$$

A statistic is a function of the observable random variables in a sample and known constants:

$$\text{Sample mean, } \bar{Y} = 3.500$$

$$\text{Sample variance, } S^2 = \frac{\sigma^2}{n} = \frac{2.9173}{3} = 0.9724$$

$$\text{Sample standard deviation, } S = \frac{\sigma}{\sqrt{n}} = \frac{1.708}{\sqrt{3}} = 0.9861$$

Calculation 1

But we can see that the result that proceed on Figure 1 is shows that $\bar{Y} = 3.333$, $S^2 = NaN$, $S = NaN$, that is different compared to the theoretical values that calculated at Calculation 1.

When we roll 3 times of dice, we obtained the result is:

$$y_1 = 1, y_2 = 4, y_3 = 5$$

Therefore, mean for "Rolls of 3 Fair Dice" for one roll will be:

$$\bar{Y} = \frac{(y_1 + y_2 + y_3)}{n}$$

$$\bar{Y} = \frac{(1 + 4 + 5)}{3}$$

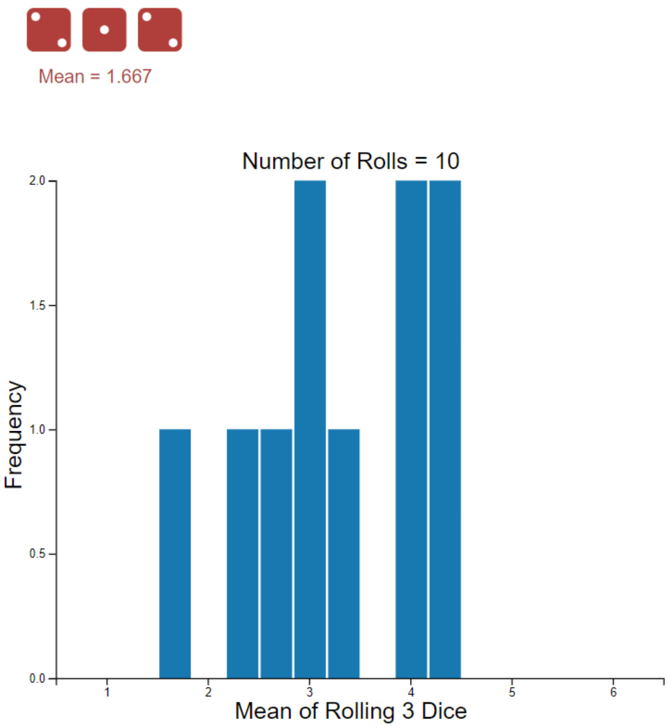
$$\bar{Y} = 3.333$$

The range of the value obtained is from 1 to 5.

As a conclusion, one roll is significantly insufficient to making inferences about a population mean based on the sample mean and make a conclusion it is follow by normal distribution. According to central limit theorem, we can only make a inference when the number of rolls is sufficiently large (usually n > 30).

ii) Use the button "1 Roll" again for nine times to obtain ten samples of size 3 from the die-tossing population. Discuss the observation on the output by considering the sample mean, the range of the value obtained and the shape of data distribution.

(6 Marks)



Viz 2

Statistical Report on Dice Rolls

Pop Prob: (1) 0.167 (2) 0.167 (3) 0.167 (4) 0.167 (5) 0.167 (6) 0.167

Population: Mean = 3.500 StDev = 1.708

Samples = 10 of size 3

Samples: Mean = 3.267 StDev = 0.900

Mean ± 1 StDev: 0.600

Mean ± 2 StDev: 1.000

Mean ± 3 StDev: 1.000

Close

Figure 2

Based on the Figure 2, we know that

$$\text{Population mean, } \mu = 3.500$$

$$\text{Population variance, } \sigma^2 = 2.9173$$

$$\text{Population standard deviation, } \sigma = 1.708$$

$$\text{Sample mean, } \bar{Y} = 3.267$$

$$\text{Sample variance, } S^2 = (0.900)^2 = 0.8100$$

$$\text{Sample standard deviation, } S = 0.900$$

A statistic is a function of the observable random variables in a sample and known constants:

$$\text{Sample mean, } \bar{Y} = 3.500$$

$$\text{Sample variance, } S^2 = \frac{\sigma^2}{n} = \frac{2.9173}{3} = 0.9724$$

$$\text{Sample standard deviation, } S = \frac{\sigma}{\sqrt{n}} = \frac{1.708}{\sqrt{3}} = 0.9861$$

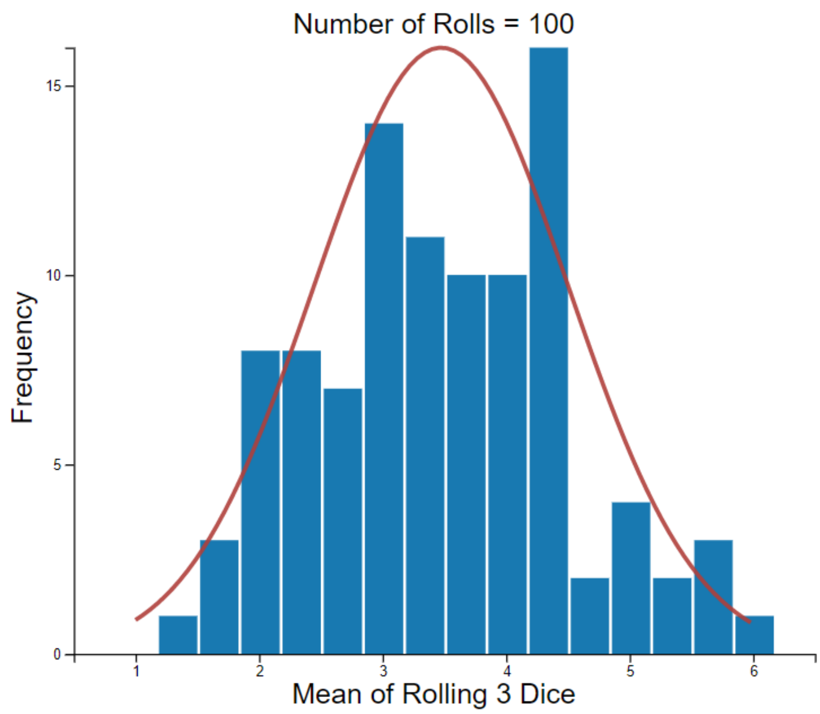
Calculation 2

From Figure 2, we can see that the result that proceed on Figure 2 is shows that $\bar{Y} = 3.267$, $S^2 = (0.900)^2 = 0.8100$, $S = 0.900$, Even thought that is different compared to the theoretical values that calculated at Calculation 2, but we can see that all the value of $\bar{Y}$, S^2 and S is close to the result that proceed on Figure 2. While the shape of data distribution in Viz 2 is bimodal distribution.

iii) Use the button “10 rolls” for nine more times until you have obtained and plotted 100 realized values for the sample mean, $\bar{Y}$. Click on the button “Stats” to see the mean and standard deviation of the 100 values of $\bar{Y}$. Discuss the observation on the histogram and the statistics by comparing with 10 samples of $\bar{Y}$ from part **(ii)**.

(8 Marks)

Means: 4.333, 4.333, 5.667, 5.000, 3.000, 2.000, 5.333, 4.000, 3.000, 3.667



Viz 3

Statistical Report on Dice Rolls

Pop Prob: (1) 0.167 (2) 0.167 (3) 0.167 (4) 0.167 (5) 0.167 (6) 0.167

Population: Mean = 3.500   StDev = 1.708

Samples = 100 of size 3

Samples: Mean = 3.473   StDev = 1.030

Mean ± 1 StDev: 0.680

Mean ± 2 StDev: 0.950

Mean ± 3 StDev: 1.000

Close

Figure 3

Based on the Figure 3, we know that

$$\text{Population mean, } \mu = 3.500$$

$$\text{Population variance, } \sigma^2 = 2.9173$$

$$\text{Population standard deviation, } \sigma = 1.708$$

$$\text{Sample mean, } \bar{Y} = 3.473$$

$$\text{Sample variance, } S^2 = (1.030)^2 = 1.0609$$

$$\text{Sample standard deviation, } S = 1.030$$

A statistic is a function of the observable random variables in a sample and known constants:

$$\text{Sample mean, } \bar{Y} = 3.500$$

$$\text{Sample variance, } S^2 = \frac{\sigma^2}{n} = \frac{2.9173}{3} = 0.9724$$

$$\text{Sample standard deviation, } S = \frac{\sigma}{\sqrt{n}} = \frac{1.708}{\sqrt{3}} = 0.9861$$

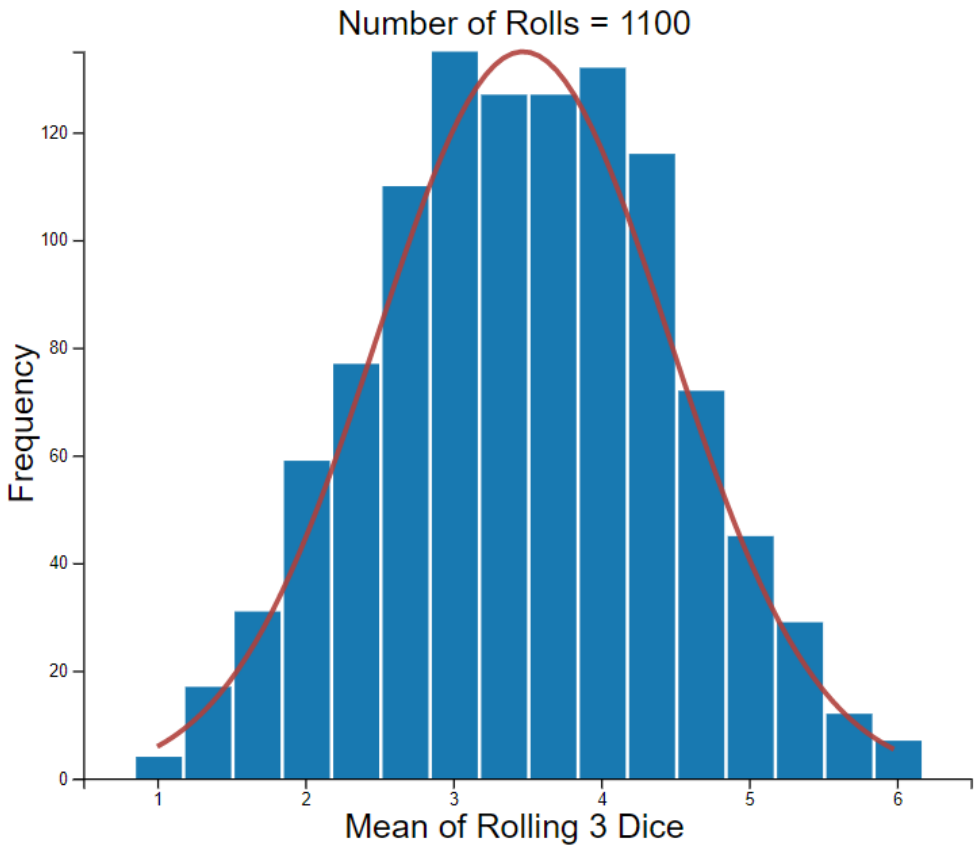
Calculation 3

From Figure 3, we can see that the histogram have two peaks which is bimodal histogram and that the result that proceed on Figure 3 is shows that $\bar{Y} = 3.473$, $S^2 = (1.030)^2 = 1.0609$, $S = 1.030$, Even thought that is different compared to the theoretical values that calculated at Calculation 3, but we can see that all the value of $\bar{Y}$, S^2 and S is close to the result that proceed on Figure 3 as compared to Figure 2.

From the statistics show in Figure 3, we can see that the sample mean, $\bar{Y} = 3.473$ was greater than the sample mean in Figure 2, 10 samples of $\bar{Y}$ from part (ii) which is $\bar{Y} = 3.267$. So, we conclude that when the number of rolls of dice increase, the sample mean, $\bar{Y}$ will be closer to the theoretical sample mean $\bar{Y}$.

iv) Click the button “1000 rolls” to generate 1100 samples of $\bar{Y}$, then click on the button “Normal”. What can you conclude based on the output?

(5 Marks)



Viz 4

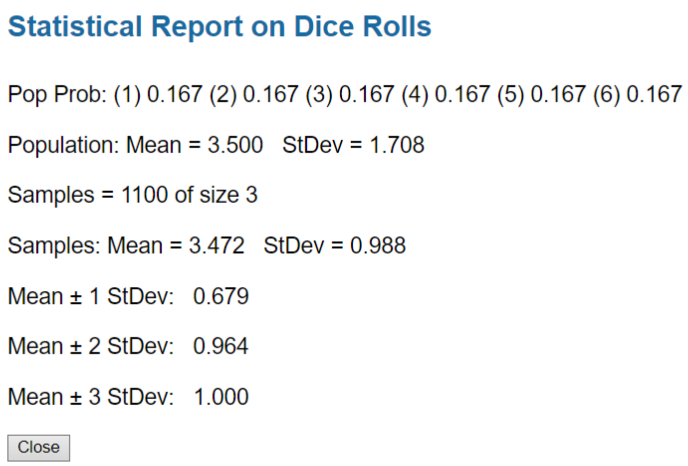


Figure 4

From the Viz 4, we can see that the histogram is almost symmetric distribution or unimodal distribution, which is only has one peak in the distribution. From the Figure 4, we obtained the results from 1100 times of size 3 dices, we can see that the sample mean, $\bar{Y} = 3.472$, sample variance, $S^2 = (0.988)^2 = 0.9861$ and, sample standard deviation $S = 0.988$, there is very close to the Calculation 3, which is the $\bar{Y} = 3.500$, sample variance, $S^2 = 0.9724$ and, sample standard deviation $S = 0.9861$. So, we can conclude again that when the increase in number of rolls of dice, the more closer the result proceed with theoretical values.

CASE STUDY 2 (25 Marks)

A balanced die is tossed twelve times. Let Y_1, Y_2 and Y_3 denote the number of spots observed on the upper face for tosses 1, 2 and 3, respectively. Suppose we are interested in

$$\bar{Y} = \frac{Y_1 + Y_2 + \ldots + Y_{12}}{12},$$

where Y denotes the number of spots observed on the upper face on a single toss of a balanced die. The following information is given:

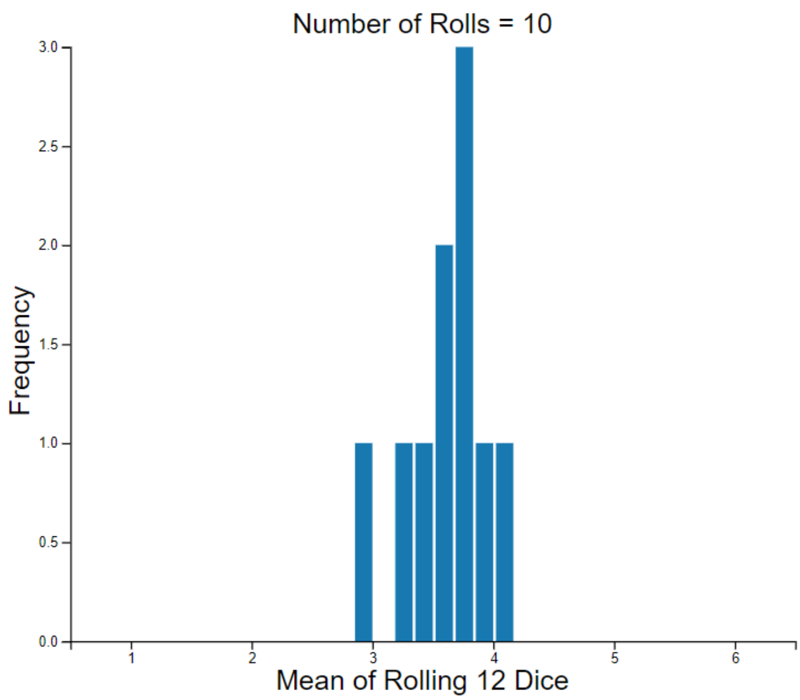
$$P(Y = i) = \frac{1}{6}, \qquad i = 1,2,\ldots,6$$

Use the applet **DiceSample** via the following link, in order to complete the assignment.
https://college.cengage.com/nextbook/statistics/wackerly_966371/student/html/index.html

i) Click on the "Rolls of 12 Fair Dice". Use the button "10 Rolls" to generate 10 mean values for the sample of size 12 from the die-tossing population. Discuss the observation on the output by comparing with the sample of size 3 from **Question 1(ii)**.

(8 Marks)

Means: 3.500, 4.083, 3.500, 3.667, 3.417, 3.833, 2.833, 3.250, 3.667, 3.750



Viz 5

Statistical Report on Dice Rolls

Pop Prob: (1) 0.167 (2) 0.167 (3) 0.167 (4) 0.167 (5) 0.167 (6) 0.167

Population: Mean = 3.500 StDev = 1.708

Samples = 10 of size 12

Samples: Mean = 3.550 StDev = 0.343

Mean ± 1 StDev: 0.800

Mean ± 2 StDev: 0.900

Mean ± 3 StDev: 1.000

Close

Figure 5

Based on the Figure 5, we know that

$$\text{Population mean, } \mu = 3.500$$
$$\text{Population variance, } \sigma^2 = 2.9173$$
$$\text{Population standard deviation, } \sigma = 1.708$$
$$\text{Sample mean, } \bar{Y} = 3.550$$
$$\text{Sample variance, } S^2 = (0.343)^2 = 0.1176$$
$$\text{Sample standard deviation, } S = 0.343$$

A statistic is a function of the observable random variables in a sample and known constants:

$$\text{Sample mean, } \bar{Y} = 3.500$$
$$\text{Sample variance, } S^2 = \frac{\sigma^2}{n} = \frac{2.9173}{12} = 0.2431$$
$$\text{Sample standard deviation, } S = \frac{\sigma}{\sqrt{n}} = \frac{1.708}{\sqrt{12}} = 0.4931$$

Calculation 5

From the Viz 5, we can see that the shape of data distribution of the histogram is left skewed distribution which means most of the data are on the right and fewer data showing up on the left side of the histogram. Compared to the sample of size 3 from **Question 1(ii)**, which is the shape of data distribution in Viz 2 is bimodal distribution.

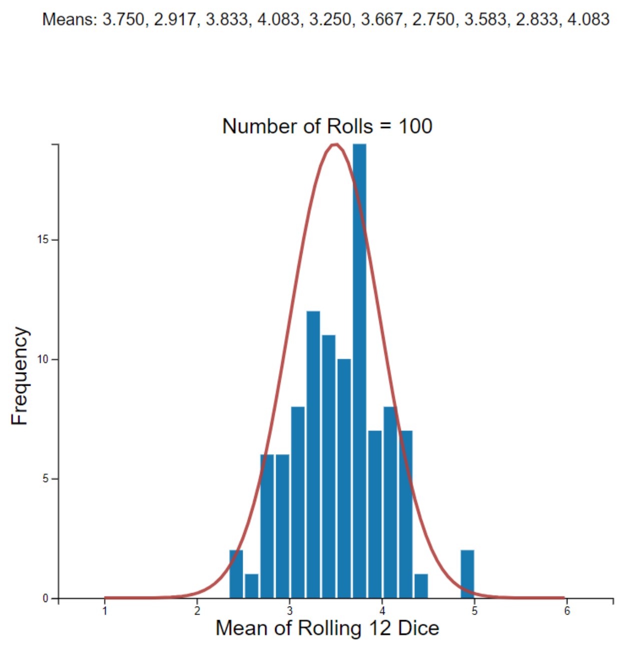
The Sample mean, $\bar{Y}$ value for the rolls of 12 fair dice which is 3.550 is greater than the $\bar{Y}$ value for the rolls of 3 fair dice which is 3.267. In this situation, we can see that the $\bar{Y}$ value for the rolls of 12 fair dice is much closer the value of the theoretical $\bar{Y}$ in Calculation 5 compared to the $\bar{Y}$ value for 3 fair dice.

By comparing variance and standard deviations, the rolls of 12 fair dice are less disperse and more consistent compared to the rolls of 3 fair dice more variable. Because of low variability, the values are more consistent, so it will be easier to make predictions compared to the high variability (data in **Question 1(ii)**).

Obviously, the data for the rolls of 12 fair dice is much better compared to the data for **Question 1(ii)**. We also can conclude that the larger samples are more closely approximate the population. Because the objective of inferential statistics is to generalize from a sample to a population, it is less in inference if the sample size is large.

ii) Use the button "10 rolls" for nine more times until you have obtained and plotted 100 realized values for the sample mean, $\bar{Y}$. Click on the button "Stats" to see the mean and standard deviation of the 100 values of $\bar{Y}$. Discuss the observation on the histogram and the statistics by comparing with 10 samples of $\bar{Y}$ from part **(i)**.

(8 Marks)



Viz 6

Statistical Report on Dice Rolls

Pop Prob: (1) 0.167 (2) 0.167 (3) 0.167 (4) 0.167 (5) 0.167 (6) 0.167

Population: Mean = 3.500 StDev = 1.708

Samples = 100 of size 12

Samples: Mean = 3.496 StDev = 0.494

Mean ± 1 StDev: 0.630

Mean ± 2 StDev: 0.950

Mean ± 3 StDev: 1.000

Close

Figure 6

Based on the Figure 6, we know that

$$\text{Population mean, } \mu = 3.500$$
$$\text{Population variance, } \sigma^2 = 2.9173$$
$$\text{Population standard deviation, } \sigma = 1.708$$
$$\text{Sample mean, } \bar{Y} = 3.496$$
$$\text{Sample variance, } S^2 = (0.494)^2 = 0.2440$$
$$\text{Sample standard deviation, } S = 0.494$$

A statistic is a function of the observable random variables in a sample and known constants:

$$\text{Sample mean, } \bar{Y} = 3.500$$
$$\text{Sample variance, } S^2 = \frac{\sigma^2}{n} = \frac{2.9173}{12} = 0.2431$$
$$\text{Sample standard deviation, } S = \frac{\sigma}{\sqrt{n}} = \frac{1.708}{\sqrt{12}} = 0.4931$$

Calculation 6

From the histogram shows in Viz 6, we can see that the shape of data distribution is bimodal distribution. While the shape of data distribution of histogram shows in Viz 5 is left skewed distribution.

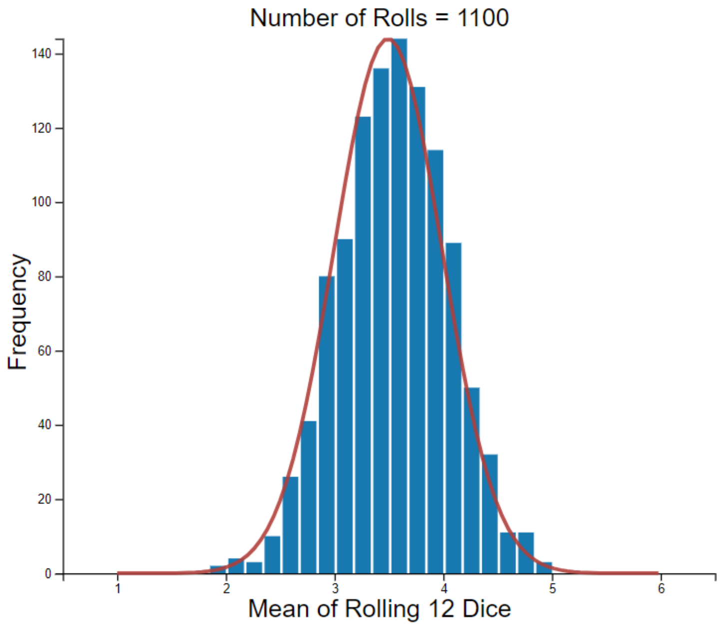
The result obtained in Figure 6, which is the $\bar{Y} = 3.496$, sample variance, $S^2 = (0.494)^2 = 0.2440$ and, sample standard deviation $S = 0.494$. All of the result obtained are very close to the theoretical values the obtained in the Calculation 6 compared to the result that obtained in Figure 5 which is the 10 samples of $\overline{Y}$ from part **(i)**. So, we can conclude that the data for sample 100 of size 12 is better than sample 10 of size 12 because it is more approximately.

iii) Click the button "1000 rolls" to generate 1100 samples of $\overline{Y}$, then click on the button "Normal" to get a line of normal distribution to data. Discuss the observation on the output based on

(a) the comparison of the statistics for the 1100 samples with mean and variance for Y (Notice that the mean and standard deviation of Y are given on the second line of the "Stat" report pop-up screen)

(b) the conclusion for normality of data.

(9 Marks)



Viz 7

Statistical Report on Dice Rolls

Pop Prob: (1) 0.167 (2) 0.167 (3) 0.167 (4) 0.167 (5) 0.167 (6) 0.167

Population: Mean = 3.500   StDev = 1.708

Samples = 1100 of size 12

Samples: Mean = 3.491   StDev = 0.497

Mean ± 1 StDev: 0.671

Mean ± 2 StDev: 0.960

Mean ± 3 StDev: 0.998

Close

Figure 7

(a) Based on the Figure 7, we know that

$$\text{Population mean, } \mu = 3.500$$

$$\text{Population variance, } \sigma^2 = 2.9173$$

$$\text{Population standard deviation, } \sigma = 1.708$$

$$\text{Sample mean, } \bar{Y} = 3.491$$

$$\text{Sample variance, } S^2 = (0.497)^2 = 0.2470$$

$$\text{Sample standard deviation, } S = 0.497$$

The result obtained in Figure 7, which is the $\bar{Y} = 3.491$, sample variance, $S^2 = (0.497)^2 = 0.2470$ and, sample standard deviation $S = 0.497$ are very close to the theoretical values obtained in Calculation 6. Since the sample increase to 1100 of size 12, the sampling distribution of the mean will approach a normal distribution. Then, as a sample size increases, the sample mean and standard deviation will be closer in value to the population mean μ and standard deviation σ.

(b) The central limit theorem states that if you have a population with a mean and standard deviation and collect sufficiently large random samples from the population with replacement, then the distribution of the sample will be normal before moving on to the conclusion that the data are normal.Means will approximately follow a normal distribution. If the sample size is sufficiently large ($n > 30$), this will be true whether the source population is normal or skewed. Theorem is valid even for samples less than 30 if the population is normal.

We may conclude that the $\bar{y}$ is normal by using central limit theorem based on the solution in section iii(a). This is due to the fact that the values for sample mean, variance, and standard deviation are quite near to the theoretical values obtained using central limit theorem, as shown in (calculated 6). It's also taken sufficiently large random samples from the population with replacement. We can observe from the histogram in viz 7 that the line of normal distribution to data is practically fit with histogram. As the sample size increases to 1100 of size 12, the histogram in viz 7 becomes nearly bell-shaped and symmetric about the mean. According to central limit theorem, the distribution of sample means will be nearly normal regardless of the form of the population distribution.

In a conclusion, we may infer that the data is practically well modelled by a normal distribution. In a nutshell, when the number of dice roll is more increasing, the more probable the shape of the distribution of the means approached a normal distribution graph.

Reference:

1. https://www.albany.edu/~jr853689/CentralLimitTheoremForDice.htm
2. https://builtin.com/data-science/understanding-central-limit-theorem
3. https://statisticsbyjim.com/basics/central-limit-theorem/
4. https://www.statisticshowto.com/probability-and-statistics/normal-distributions/central-limit-theorem-definition-examples/
5. https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_probability/BS704_Probability12.html
6. https://www.investopedia.com/terms/c/central_limit_theorem.asp