Experiment No.: 8
Title: Statistics using spreadsheet

Batch: B2 **Roll No.: 16010421059** **Experiment No.: 8**

Aim: 1. To generate random numbers and draw samples from the data set using MS Excel
2. Hypothesis testing for mean

Resources needed: MS Excel

Theory

Problem Statement:
Generate random numbers using rand() / randbetween() / Data Analysis Toolpack and draw simple random samples from the dataset.

Concepts

Sample and Sampling

A Sample is a part of the total population. It can be an individual element or a group of elements selected from the population. Although it is a subset, it is representative of the population and suitable for research in terms of cost, convenience, andtime.

A good sample is one which satisfies all or few of the following conditions:

Representativeness: Good samples are those who accurately represent the population. On measurement terms, the sample must be valid. The validity of a sample depends upon its accuracy.

Accuracy: An accurate (unbiased) sample is one which exactly represents the population. It is free from any influence that causes any differences between sample value and population value.

Size: The sample size should be such that the inferences drawn from the sample are accurate to a given level of confidence to represent the entire population under study.

Sampling is the act, process, or technique of selecting a representative part of a population for the purpose of determining the characteristics of the whole population. Sampling is that part of statistical practice concerned with the selection of an unbiased or random subset of individual observations within a population of individuals intended to yield some knowledge about the population of concern, especially for the purposes of making predictions based on statistical inference. Sampling is an important aspect of data collection.

Population OR Universe: The entire aggregation of items from which samples can be drawn is known as a population. Population, contrary to its general notion as a nation's entire population has a much broader meaning in sampling. "N" represents the size of the population.

An operational sampling process can be divided into seven steps as given below:

1. Defining the target population.
2. Specifying the sampling frame.
3. Specifying the sampling unit.
4. Selection of the sampling method.
5. Determination of sample size.

6. Specifying the sampling plan.
7. Selecting the sample.

There are two basic approaches to sampling:

1. Probabilistic Sampling
2. Non-probabilistic sampling.

A Probabilistic sampling scheme is one in which every unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined.

Types of Probabilistic Sampling

- Simple random sampling
- Systematic sampling
- Stratified sampling
- Multistage cluster sampling

Non-probabilistic Sampling It involves the selection of units based on factors other than random chance. It is also known as deliberate sampling and purposive sampling.

Types of Non-Probabilistic Sampling

- Convenience sampling
- Quota sampling
- Judgment sampling
- Snowball sampling

Simple Random Sampling:

A sampling process where each element in the target population has an equal chance or probability of inclusion in the sample is known as Simple Random Sampling. For ex, if a sample of 15000 names is to be drawn from the telephone directory, then there is equal chance for each number in the directory to be selected. These numbers (serial no of name) could be randomly generated by the computer or picked out of a box. These numbers could be later matched with the corresponding names thus fulfilling the list. In small populations random sampling is done without replacement to avoid the instance of a unit being sampled more thanonce.

Hypothesis Testing for mean:

hypothesis test of a mean can be conducted, when the following conditions are met:

- The sampling method is simple random sampling.
- The sampling distribution is normal or nearly normal.

Generally, the sampling distribution will be approximately normally distributed if any of the following conditions apply.

- The population distribution is normal.
- The population distribution is symmetric, unimodal, without outliers, and the sample size is 15 or less.
- The population distribution is moderately skewed, unimodal, without outliers, and the sample size is between 16 and 40.
- The sample size is greater than 40, without outliers.

This approach consists of four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results.

State the Hypotheses

Every hypothesis test requires the analyst to state a null hypothesis and an alternative hypothesis. The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false; and vice versa.

The table below shows three sets of hypotheses. Each makes a statement about how the population mean μ is related to a specified value M. (In the table, the symbol ≠ means " not equal to ".)

Set	Null hypothesis	Alternative hypothesis	Number of tails
1	μ = M	μ ≠ M	2
2	μ >= M	μ < M	1
3	μ ≤ M	μ > M	1

The first set of hypotheses (Set 1) is an example of a two-tailed test, since an extreme value on either side of the sampling distribution would cause a researcher to reject the null hypothesis. The other two sets of hypotheses (Sets 2 and 3) are one-tailed tests, since an extreme value on only one side of the sampling distribution would cause a researcher to reject the null hypothesis.

Formulate an Analysis Plan

The analysis plan describes how to use sample data to accept or reject the null hypothesis. It should specify the following elements.

- Significance level. Often, researchers choose significance levels equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
- Test method. Use the one-sample t-test to determine whether the hypothesized mean differs significantly from the observed sample mean.

Analyze Sample Data

Using sample data, conduct a one-sample t-test. This involves finding the standard error, degrees of freedom, test statistic, and the P-value associated with the test statistic.

- Standard error. Compute the standard error (SE) of the samplingdistribution.

  $$SE = s * sqrt\{ (1/n) * [(N - n) / (N - 1)] \}$$

  where s is the standard deviation of the sample, N is the population size, and n is the sample size. When the population size is much larger (at least 20 times larger) than the sample size, the standard error can be approximated by:

  $$SE = s / sqrt(n)$$

- Degrees of freedom. The degrees of freedom (DF) is equal to the sample size (n) minus one. Thus, DF = n - 1.

- Test statistic. The test statistic is a t statistic (t) defined by the following equation.

  $$t = (x - \mu) / SE$$

  where x is the sample mean, μ is the hypothesized population mean in the nul hypothesis, and SE is the standard error.

- P-value. The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a t statistic, use the t Distribution Calculator to assess the probability associated with the t statistic, given the degrees of freedom computedabove.

Interpret Results

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the significance level, and rejecting the null hypothesis when the P-value is less than the significance level.

Test Your Understanding

Two sample problems illustrate how to conduct a hypothesis test of a mean score. The first problem involves a two-tailed test; the second problem, a one-tailed test.

Problem 1: Two-Tailed Test

An inventor has developed a new, energy-efficient lawn mower engine. He claims that the engine will run continuously for 5 hours (300 minutes) on a single gallon of regular gasoline. From his stock of 2000 engines, the inventor selects a simple random sample of 50 engines for testing. The engines run for an average of 295 minutes, with a standard deviation of 20 minutes. Test the null hypothesis that the mean run time is 300 minutes against the alternative hypothesis that the mean run time is not 300 minutes. Use a 0.05 level of significance. (Assume that run times for the population of engines are normally distributed.)

Solution: The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

- State the hypotheses. The first step is to state the null hypothesis and an alternative hypothesis.

  Nul hypothesis: $\mu = 300$

  Alternative hypothesis: $\mu \neq 300$

  Note that these hypotheses constitute a two-tailed test. The null hypothesis will be rejected if the sample mean is too big or if it is too small.
- Formulate an analysis plan. For this analysis, the significance level is 0.05. The test method is a one-sample t-test.
- Analyze sample data. Using sample data, we compute the standard error (SE), degrees of freedom (DF), and the t statistic test statistic (t).

  $$SE = s / \text{sqrt}(n) = 20 / \text{sqrt}(50) = 20/7.07 = 2.83$$

  $$DF = n - 1 = 50 - 1 = 49$$

  $$t = (x - \mu) / SE = (295 - 300)/2.83 = -1.77$$

  where s is the standard deviation of the sample, x is the sample mean, μ is the hypothesized population mean, and n is the sample size.

  Since we have a two-tailed test, the P-value is the probability that the t statistic having 49 degrees of freedom is less than -1.77 or greater than 1.77.

  We use the t Distribution Calculator to find $P(t < -1.77) = 0.04$, and $P(t > 1.77) = 0.04$. Thus, the P-value = 0.04 + 0.04 = 0.08.
- Interpret results. Since the P-value (0.08) is greater than the significance level (0.05), we cannot reject the null hypothesis.

Note: If you use this approach on an exam, you may also want to mention why this approach is appropriate. Specifically, the approach is appropriate because the sampling method was simple random sampling, the population was normally distributed, and the sample size was small relative to the population size (less than 5%).

Problem 2: One-Tailed Test

Bon Air Elementary School has 1000 students. The principal of the school thinks that the average IQ of students at Bon Air is at least 110. To prove her point, she administers an IQ test to 20 randomly selected students. Among the sampled students, the average IQ is 108 with a standard deviation of 10. Based on these results, should the principal accept or reject her original hypothesis? Assume a significance level of 0.01. (Assume that test scores in the population of engines are normally distributed.)

Solution: The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

- State the hypotheses. The first step is to state the null hypothesis and an alternative hypothesis.

  Nul hypothesis: μ >= 110

  Alternative hypothesis: μ < 110

  Note that these hypotheses constitute a one-tailed test. The null hypothesis will be rejected if the sample mean is too small.

- Formulate an analysis plan. For this analysis, the significance level is 0.01. The test method is a one-sample t-test.

- Analyze sample data. Using sample data, we compute the standard error (SE), degrees of freedom (DF), and the t statistic test statistic (t).

  SE = s / sqrt(n) = 10 / sqrt(20) = 10/4.472 = 2.236

  DF = n - 1 = 20 - 1 = 19

  t = (x - μ) / SE = (108 - 110)/2.236 = -0.894

  where s is the standard deviation of the sample, x is the sample mean, μ is the hypothesized population mean, and n is the sample size.

  Here is the logic of the analysis: Given the alternative hypothesis (μ < 110), we want to know whether the observed sample mean is small enough to cause us to reject the null hypothesis.

  The observed sample mean produced a t statistic test statistic of -0.894. We use the t Distribution Calculator to find P(t < -0.894) = 0.19. This means we would expect to find a sample mean of 108 or smaller in 19 percent of our samples, if the true population IQ were 110. Thus the P-value in this analysis is 0.19.

- Interpret results.

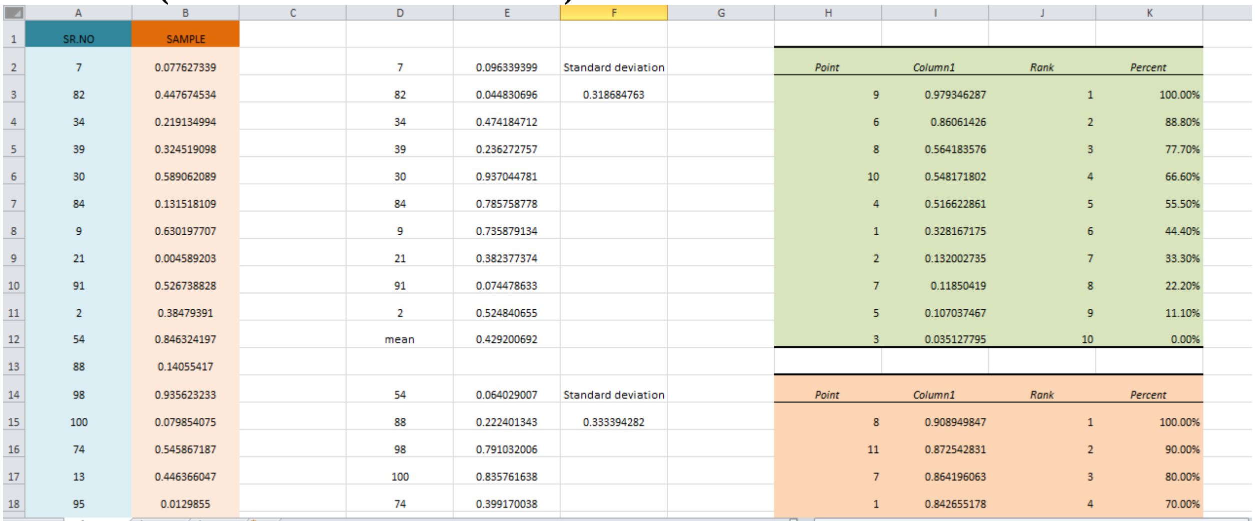
- Since the P-value (0.19) is greater than the significance level (0.01), we cannot reject the null hypothesis.

Note: If you use this approach on an exam, you may also want to mention why this approach is appropriate. Specifically, the approach is appropriate because the sampling method was simple random sampling, the population was normally distributed, and the sample size was small relative to the population size (less than5%)

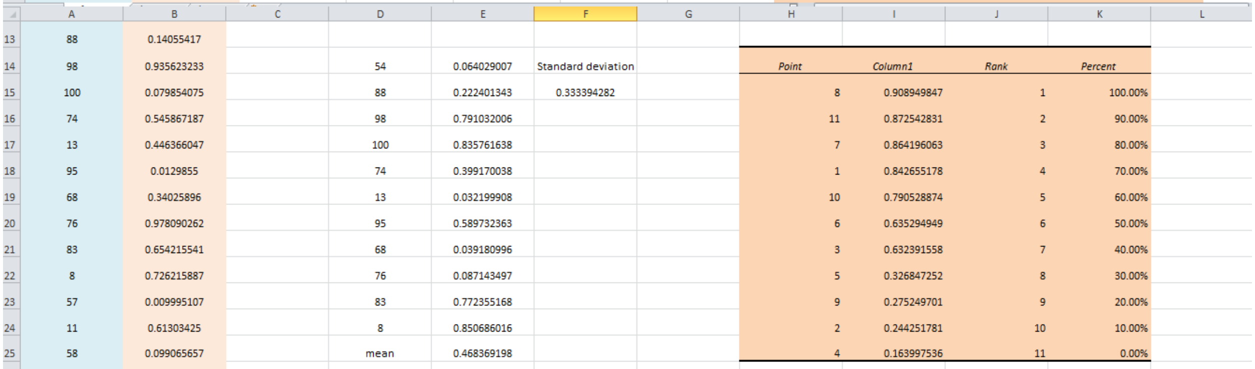
Procedure:

- Draw random numbers in MS Excel using Rand() / Rand between() and using Data Analysis Tool pack draw N(100, 15) random numbers
- Generate 4 (2 each) sample sets (Each set consisting of 10 random numbers) from the data set generated in the previous step using the Sampling feature of the Data Analysis Tool pack
- Use the Rank and Percentile feature of the Data Analysis Tool Pack
- Compute the mean and standard deviation of the samples from the Normal random number set and compare it with the given mean and standarddeviation
- Consider your performance in last 6 semesters
- Make an hypothesis regarding mean score
- Use the t.test () function in excel to compute p value
- Compare the p value with the level of significance
- Take a decision.
- Use the tdist() function of excel to compute p value and compare it with the p value computed using the t.test () function

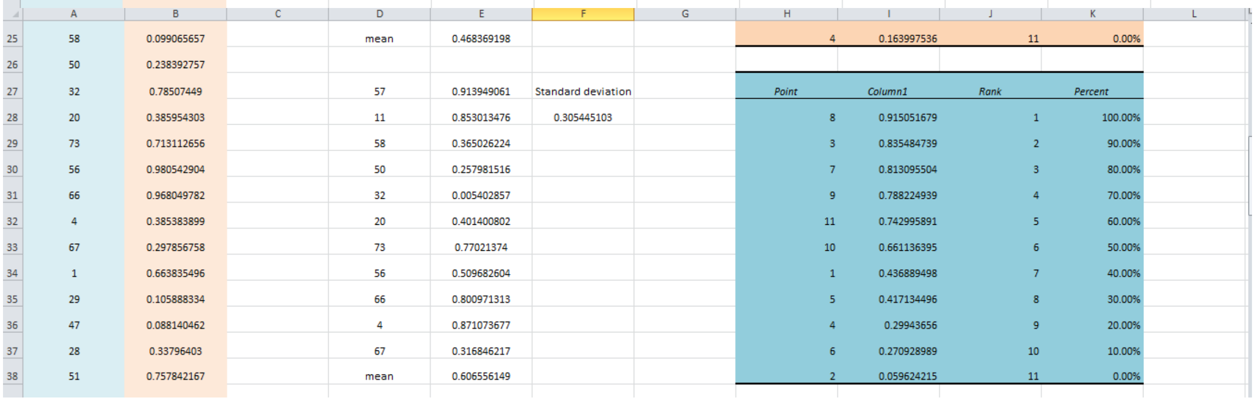
Results: (Screen shot of the excel sheet)



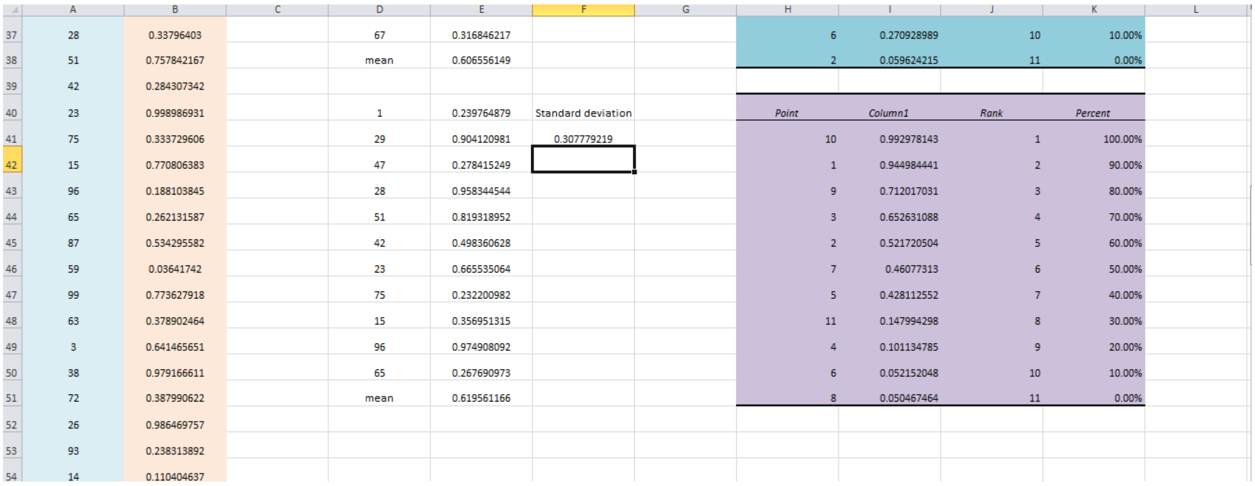
	A	B	C	D	E	F	G	H	I	J	K
1	SR.NO	SAMPLE									
2	7	0.077627339		7	0.096339399	Standard deviation		Point	Column1	Rank	Percent
3	82	0.447674534		82	0.044830696	0.318684763		9	0.979346287	1	100.00%
4	34	0.219134994		34	0.474184712			6	0.86061426	2	88.80%
5	39	0.324519098		39	0.236272757			8	0.564183576	3	77.70%
6	30	0.589062089		30	0.937044781			10	0.548171802	4	66.60%
7	84	0.131518109		84	0.785758778			4	0.516622861	5	55.50%
8	9	0.630197707		9	0.735879134			1	0.328167175	6	44.40%
9	21	0.004589203		21	0.382377374			2	0.132002735	7	33.30%
10	91	0.526738828		91	0.074478633			7	0.11850419	8	22.20%
11	2	0.38479391		2	0.524840655			5	0.107037467	9	11.10%
12	54	0.846324197		mean	0.429200692			3	0.035127795	10	0.00%
13	88	0.14055417									
14	98	0.935623233		54	0.064029007	Standard deviation		Point	Column1	Rank	Percent
15	100	0.079854075		88	0.222401343	0.333394282		8	0.908949847	1	100.00%
16	74	0.545867187		98	0.791032006			11	0.872542831	2	90.00%
17	13	0.446366047		100	0.835761638			7	0.864196063	3	80.00%
18	95	0.0129855		74	0.399170038			1	0.842655178	4	70.00%



	A	B	C	D	E	F	G	H	I	J	K	L
13	88	0.14055417										
14	98	0.935623233		54	0.064029007	Standard deviation		Point	Column1	Rank	Percent	
15	100	0.079854075		88	0.222401343	0.333394282		8	0.908949847	1	100.00%	
16	74	0.545867187		98	0.791032006			11	0.872542831	2	90.00%	
17	13	0.446366047		100	0.835761638			7	0.864196063	3	80.00%	
18	95	0.0129855		74	0.399170038			1	0.842655178	4	70.00%	
19	68	0.34025896		13	0.032199908			10	0.790528874	5	60.00%	
20	76	0.978090262		95	0.589732363			6	0.635294949	6	50.00%	
21	83	0.654215541		68	0.039180996			3	0.632391558	7	40.00%	
22	8	0.726215887		76	0.087143497			5	0.326847252	8	30.00%	
23	57	0.009995107		83	0.772355168			9	0.275249701	9	20.00%	
24	11	0.61303425		8	0.850686016			2	0.244251781	10	10.00%	
25	58	0.099065657		mean	0.468369198			4	0.163997536	11	0.00%	



	A	B	C	D	E	F	G	H	I	J	K	L
25	58	0.099065657		mean	0.468369198			4	0.163997536	11	0.00%	
26	50	0.238392757										
27	32	0.78507449		57	0.913949061	Standard deviation		Point	Column1	Rank	Percent	
28	20	0.385954303		11	0.853013476	0.305445103		8	0.915051679	1	100.00%	
29	73	0.713112656		58	0.365026224			3	0.835484739	2	90.00%	
30	56	0.980542904		50	0.257981516			7	0.813095504	3	80.00%	
31	66	0.968049782		32	0.005402857			9	0.788224939	4	70.00%	
32	4	0.385383899		20	0.401400802			11	0.742995891	5	60.00%	
33	67	0.297856758		73	0.77021374			10	0.661136395	6	50.00%	
34	1	0.663835496		56	0.509682604			1	0.436889498	7	40.00%	
35	29	0.105888334		66	0.800971313			5	0.417134496	8	30.00%	
36	47	0.088140462		4	0.871073677			4	0.29943656	9	20.00%	
37	28	0.33796403		67	0.316846217			6	0.270928989	10	10.00%	
38	51	0.757842167		mean	0.606556149			2	0.059624215	11	0.00%	



	A	B	C	D	E	F	G	H	I	J	K	L
37	28	0.33796403		67	0.316846217			6	0.270928989	10	10.00%	
38	51	0.757842167		mean	0.606556149			2	0.059624215	11	0.00%	
39	42	0.284307342										
40	23	0.998986931		1	0.239764879	Standard deviation		*Point*	*Column1*	*Rank*	*Percent*	
41	75	0.333729606		29	0.904120981	0.307779219		10	0.992978143	1	100.00%	
42	15	0.770806383		47	0.278415249			1	0.944984441	2	90.00%	
43	96	0.188103845		28	0.958344544			9	0.712017031	3	80.00%	
44	65	0.262131587		51	0.819318952			3	0.652631088	4	70.00%	
45	87	0.534295582		42	0.498360628			2	0.521720504	5	60.00%	
46	59	0.03641742		23	0.665535064			7	0.46077313	6	50.00%	
47	99	0.773627918		75	0.232200982			5	0.428112552	7	40.00%	
48	63	0.378902464		15	0.356951315			11	0.147994298	8	30.00%	
49	3	0.641465651		96	0.974908092			4	0.101134785	9	20.00%	
50	38	0.979166611		65	0.267690973			6	0.052152048	10	10.00%	
51	72	0.387990622		mean	0.619561166			8	0.050467464	11	0.00%	
52	26	0.986469757										
53	93	0.238313892										
54	14	0.110404637										

Questions:

1. Definition of Sample and Sampling:

- Sample: In statistics, a sample refers to a subset of individuals or items selected from a larger population. It is used to represent the characteristics of the entire population. For instance, if a researcher wants to study the average income of people in a city, instead of surveying every individual in the city, they might select a sample of, say, 500 individuals and use their data to estimate the average income for the entire population.
- Sampling: Sampling is the process of selecting a subset of individuals or items from a population to estimate characteristics of the whole population. There are various sampling techniques such as simple random sampling, stratified sampling, cluster sampling, etc., each with its own advantages and limitations.

2. Necessity of Sampling in Research Study:

- Sampling is necessary in research studies for several reasons:
  - Resource Constraints: It is often impractical or impossible to study the entire population due to limited resources such as time, money, and manpower.
  - Feasibility: Sometimes, the population size is too large to study comprehensively, making sampling the only feasible option.
  - Accuracy: Properly conducted sampling can provide accurate estimates of population parameters, thus allowing researchers to make inferences about the entire population based on the characteristics of the sample.
  - Practicality: Sampling allows researchers to collect data efficiently, reducing the time and effort required compared to studying the entire population.

3. Significance of p-value:

- The p-value, or probability value, is a measure used in hypothesis testing to determine the strength of evidence against the null hypothesis. It indicates the probability of obtaining an observed result (or more extreme) when the null hypothesis is true.
- In statistical hypothesis testing, a p-value less than a chosen significance level (usually 0.05) suggests that the observed data is statistically significant, meaning that the null hypothesis can be rejected in favor of the alternative hypothesis.
- The significance of the p-value lies in its ability to provide a quantitative measure of the evidence against the null hypothesis, helping researchers make informed decisions about the validity of their hypotheses.

4. Probability Calculation for Joe:

- Given: Probability of Joe getting into any game = 0.40
- (a) Probability that the first game Joe enters is the fourth game of the season:
  - Since Joe is the third-string quarterback, he can only enter the game if both first and second-string quarterbacks are unavailable. So, the probability that Joe enters the fourth game directly is the probability that the first three games he doesn't enter multiplied by the probability that he enters the fourth game:
  - Probability(Joe enters fourth game directly) = (1 - 0.40)^3 * 0.40 = (0.60)^3 * 0.40 = 0.216 * 0.40 = 0.0864

- (b) Probability that Joe plays in no more than two of the first five games:
  - This can be calculated by considering the scenarios where Joe plays in 0, 1, or 2 of the first five games and adding up their probabilities:
  - Probability(Joe plays 0 games) = (1 - 0.40)^5 = (0.60)^5 ≈ 0.07776
  - Probability(Joe plays 1 game) = 5C1 * (0.40)^1 * (1 - 0.40)^4 = 5 * 0.40 * (0.60)^4 ≈ 0.2304
  - Probability(Joe plays 2 games) = 5C2 * (0.40)^2 * (1 - 0.40)^3 = 10 * (0.40)^2 * (0.60)^3 ≈ 0.3456
  - Probability(Joe plays no more than two games) = Probability(Joe plays 0 games) + Probability(Joe plays 1 game) + Probability(Joe plays 2 games)
  - ≈ 0.07776 + 0.2304 + 0.3456 ≈ 0.65376

Outcomes:

CO3: Analyze simulation results to reach an appropriate conclusion.

Conclusion:

Successfully generated random numbers and draw samples from the data set using MS Excel

Grade: AA / AB / BB / BC / CC / CD /DD

Signature of faculty in-charge with date

References:

Books/ Journals/ Websites:

1. "Linear Congruential Generators" by Joe Bolte, Wolfram Demonstrations Project.
2. Severance, Frank (2001). *System Modeling and Simulation*. John Wiley & Sons, Ltd. p. 86. ISBN 0-471-49694-4.
3. The GNU C library's *rand()* in stdlib.h uses a simple (single state) linear congruential generator only in case that the state is declared as 8 bytes. If the state is larger (an array), the generator becomes an additive feedback generator and the period increases. See the simplified code that reproduces the random sequence from this library.
4. "A collection of selected pseudorandom number generators with linear structures, K. Entacher, 1997". Retrieved 16 June 2012.
5. "How Visual Basic Generates Pseudo-Random Numbers for the RND Function". *Microsoft Support*. Microsoft. Retrieved 17 June 2011.
6. In spite of documentation on MSDN, RtlUniform uses LCG, and not Lehmer's algorithm, implementations before Windows Vista are flawed, because the result of multiplication is cut to 32 bits, before modulo is applied
7. GNU Scientific Library: Other random number generators
8. Novice Forth library
9. Matsumoto, Makoto, and Takuji Nishimura (1998) ACM Transactions on Modeling and Computer Simulation
10. S.K. Park and K.W. Miller (1988). "Random Number Generators: Good Ones Are Hard To Find". *Communications of the ACM* **31** (10): 1192–1201. doi:10.1145/63039.63042.
11. D. E. Knuth. *The Art of Computer Programming*, Volume 2: *Seminumerical Algorithms*, Third Edition. Addison-Wesley, 1997. ISBN 0-201-89684-2. Section 3.2.1: The Linear Congruential Method, pp. 10–26.
12. P. L'Ecuyer (1999). "Tables of Linear Congruential Generators of Different Sizes and Good Lattice Structure". *Mathematics of Computation* **68** (225): 249–260. doi:10.1090/S0025-5718-99-00996-5.
13. Press, WH; Teukolsky, SA; Vetterling, WT; Flannery, BP (2007), "Section 7.1.1. Some History", *Numerical Recipes: The Art of Scientific Computing* (3rd ed.), New York: Cambridge University Press, ISBN 978-0-521-88068-8
14. Gentle, James E., (2003). *Random Number Generation and Monte Carlo Methods*, 2nd edition, Springer, ISBN 0-387-00178-6.
15. Joan Boyar (1989). "Inferring sequences produced by pseudo-random number generators". *Journal of the ACM* **36** (1): 129–141. doi:10.1145/58562.59305. (in this paper, efficient algorithms are given for inferring sequences produced by certain pseudo-random number generators).