Experiment No.: 8

Title: Statistics using spreadsheet

Batch: B–1 **Roll No.: 16010422234** **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 data set.

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, and time.

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 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 example, if a sample of 15000 names is to be drawn from the telephone directory, then there is an 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 than once.

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 sampling distribution.
  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 statistics. The test statistic is a t statistic (t) defined by the following equation.
  t = (x - μ) / SE
  where x is the sample mean, μ is the hypothesized population mean in the null 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 computed above.

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.
  Null 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 / sqrt(n) = 20 / sqrt(50) = 20/7.07 = 2.83 DF = n - 1 = 50 - 1 = 49
  t = (x - μ) / 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.
  Null hypothesis: $\mu >= 110$
  Alternative hypothesis: $\mu < 110$
  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 ($\mu < 110$), 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.

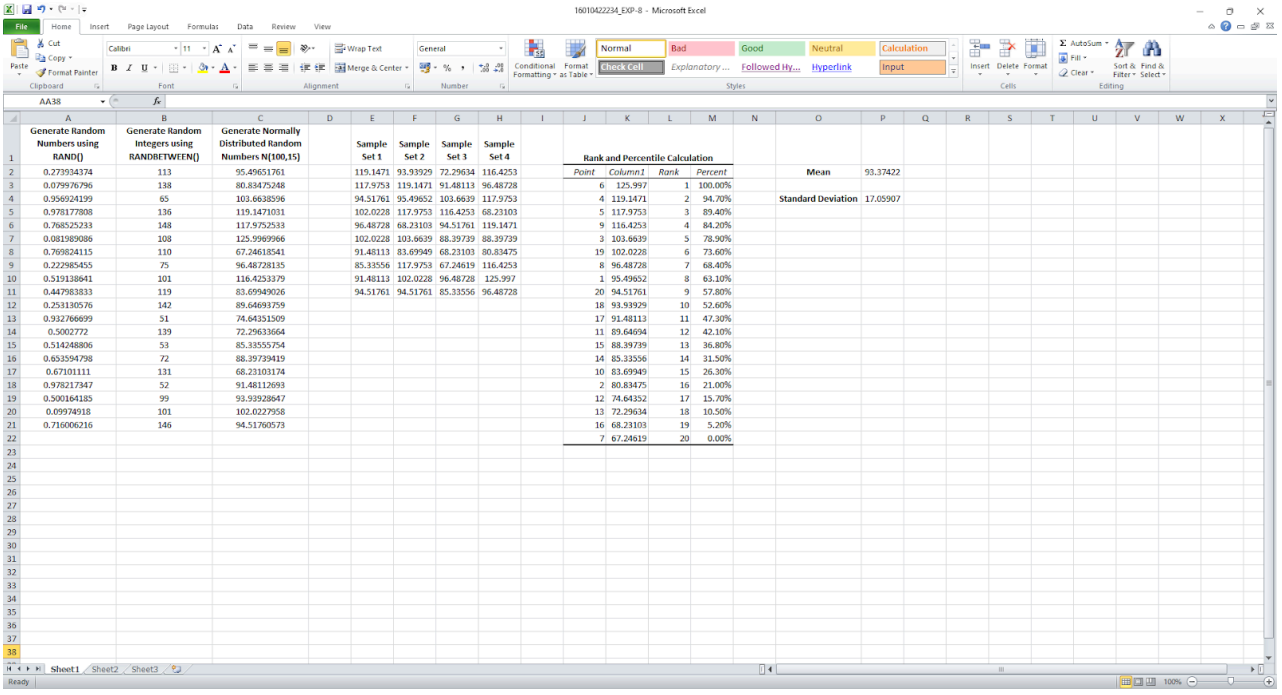
Note: If you use this approach on an exam, you may also want to mention why this approach is appropriate. Specifically, the approach is appropriate because the sampling method was simple random sampling, the population was normally distributed, and the sample size was small relative to the population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

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 standard deviation
- 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: (Screenshot of the excel sheet)

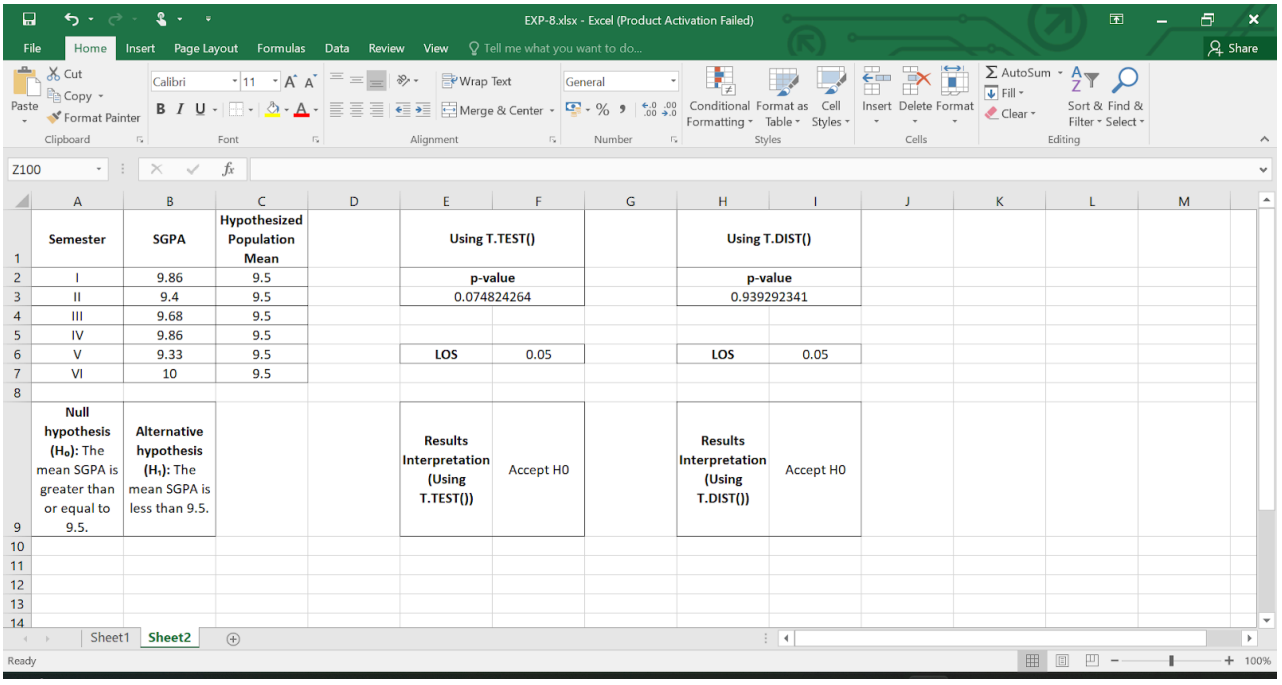


Generate Random Numbers using RAND()	Generate Random Integers using RANDBETWEEN()	Generate Normally Distributed Random Numbers N(100,15)		Sample Set 1	Sample Set 2	Sample Set 3	Sample Set 4
0.273934374	113	95.49651761		119.1471	93.93929	72.29634	116.4253
0.079976796	138	80.83475248		117.9753	119.1471	91.48113	96.48728
0.956924199	65	103.6638596		94.51761	95.49652	103.6639	117.9753
0.978177808	136	119.1471031		102.0228	117.9753	116.4253	68.23103
0.768525233	148	117.9752533		96.48728	68.23103	94.51761	119.1471
0.081989086	108	125.9969966		102.0228	103.6639	88.39739	88.39739
0.769824115	110	67.24618541		91.48113	83.69949	68.23103	80.83475
0.222985455	75	96.48728135		85.33556	117.9753	67.24619	116.4253
0.519138641	101	116.4253379		91.48113	102.0228	96.48728	125.997
0.447983833	119	83.69949026		94.51761	94.51761	85.33556	96.48728
0.253130576	142	89.64693759					
0.932766699	51	74.64351509					
0.5002772	139	72.29633664					
0.514248806	53	85.33555754					
0.653594798	72	88.39739419					
0.67101111	131	68.23103174					
0.978217347	52	91.48112693					
0.500164185	99	93.93928647					
0.09974918	101	102.0227958					
0.716006216	146	94.51760573					

Rank and Percentile Calculation

Point	Column1	Rank	Percent
6	125.997	1	100.00%
4	119.1471	2	94.70%
5	117.9753	3	89.40%
9	116.4253	4	84.20%
3	103.6639	5	78.90%
19	102.0228	6	73.60%
8	96.48728	7	68.40%
1	95.49652	8	63.10%
20	94.51761	9	57.80%
18	93.93929	10	52.60%
17	91.48113	11	47.30%
11	89.64694	12	42.10%
15	88.39739	13	36.80%
14	85.33556	14	31.50%
10	83.69949	15	26.30%
2	80.83475	16	21.00%
12	74.64352	17	15.70%
13	72.29634	18	10.50%
16	68.23103	19	5.20%
7	67.24619	20	0.00%

Mean	93.37422
Standard Deviation	17.05907



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**Semester**	**SGPA**	**Hypothesized Population Mean**		**Using T.TEST()**			**Using T.DIST()**					
2	I	9.86	9.5		**p-value**			**p-value**					
3	II	9.4	9.5		0.074824264			0.939292341					
4	III	9.68	9.5										
5	IV	9.86	9.5										
6	V	9.33	9.5		**LOS**	0.05		**LOS**	0.05				
7	VI	10	9.5										
8													
9	**Null hypothesis (H_0):** The mean SGPA is greater than or equal to 9.5.	**Alternative hypothesis (H_1):** The mean SGPA is less than 9.5.			**Results Interpretation (Using T.TEST())**	Accept H0		**Results Interpretation (Using T.DIST())**	Accept H0				
10													
11													
12													
13													
14													

Questions:

1) **Define the term sample and sampling with an example?**

   Sample: A sample is a subset of a population selected for measurement or observation. It represents the population in terms of the characteristics of interest. For example, if we want to determine the average height of students in a school, we may take a random sample of 50 students rather than measuring all the students.

   Sampling: Sampling is the process of selecting a group of individuals or elements from a larger population for research or study. An example is drawing 10 names from a list of 1000 students to survey their opinions on a school policy.

2) **Why is it necessary to do sampling during any research study?**

   Sampling is necessary because it is often impractical or impossible to collect data from an entire population. It allows researchers to make inferences about a larger population without the high cost, time, or logistical challenges of gathering data from every individual. Furthermore, sampling, when done correctly, can provide reliable estimates of population parameters.

3) **What is the significance of p value?**

   The p-value is the probability that the observed data (or something more extreme) would occur if the null hypothesis were true. A small p-value (typically ≤ 0.05) indicates that the null hypothesis can be rejected, implying that the observed result is statistically significant. A larger p-value suggests that there is insufficient evidence to reject the null hypothesis.

4) **Joe is the third-string quarterback for the university of lower Alatoona. The probability that Joe gets into any game is 0.40.**

   a) **What is the probability that the first game Joe enters is the fourth game of the season?**

      The probability that the first game Joe enters is the fourth game can be modeled using a geometric distribution:

$P\ (first\ game\ is\ 4th) = (1 - 0.40)^3 \times 0.40$

$P = (0.60)^3 \times 0.40 = 0.216 \times 0.40 = 0.0864$

So, the probability is 0.0864.

b) What is the probability that Joe plays in no more than two of the first five games?

We can use the binomial distribution to calculate the probability of Joe playing in 0, 1, or 2 games out of the first 5 games:

$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$

Where X is the number of games Joe plays in, and the binomial probability formula is:

$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$

For n=5, p=0.40, and k=0,1,2, you would calculate the individual probabilities for k=0,1,2 and sum them.

Outcomes: CO3 — Analyze simulation results to reach an appropriate conclusion.

Conclusion:

In this experiment, we generated random samples and performed hypothesis testing for the mean using T.TEST and T.DIST in MS Excel. By comparing p-values with the significance level, we determined whether to accept or reject the null hypothesis. The experiment helped us understand the process of hypothesis testing, the role of p-values, and how they inform conclusions about population parameters based on sample data. Overall, the experiment demonstrated the practical application of statistical methods to analyze data efficiently.

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

Signature of faculty in-charge with date

References:

Books/ Journals/ Websites:

1) Understanding Hypothesis Testing - GeeksforGeeks
2) Hypothesis Testing | A Step-by-Step Guide with Easy Examples
3) Statistics - Hypothesis Testing
4) Difference Between One-Tailed and Two-Tailed Tests - GeeksforGeeks

---