Statistics Essentials for Data Science

Assisted Practice: Application of Inferential Statistics

Problem Statement

A simple random sample is a subset of a population chosen at random.

Using Microsoft Excel, generate a random sample of:

Size (N) = 150

Mean = 100

Standard deviation = 10

Variance = 100

To obtain this sample, first obtain a random sample from a uniform distribution in the range [0, 1].

Dataset

The table represents a random sample from a uniform distribution [0, 1].

0.3598	0.8180	0.1952	0.6647	0.7135	0.5984	0.9359	0.6737	0.6848	0.2895
0.4542	0.9266	0.8751	0.7637	0.2027	0.5660	0.5157	0.5127	0.5442	0.3749
0.9356	0.3679	0.2100	0.9368	0.7441	0.5011	0.8756	0.5087	0.1471	0.4873
0.3948	0.2198	0.1011	0.7868	0.5409	0.6742	0.7580	0.3074	0.1384	0.0453
0.2573	0.1077	0.0186	0.7846	0.8538	0.2572	0.3468	0.9058	0.4234	0.0324
0.2806	0.5735	0.8450	0.2450	0.7584	0.7335	0.1165	0.5851	0.9226	0.2910
0.1375	0.3395	0.0170	0.7691	0.4493	0.9857	0.2511	0.3291	0.4391	0.5115

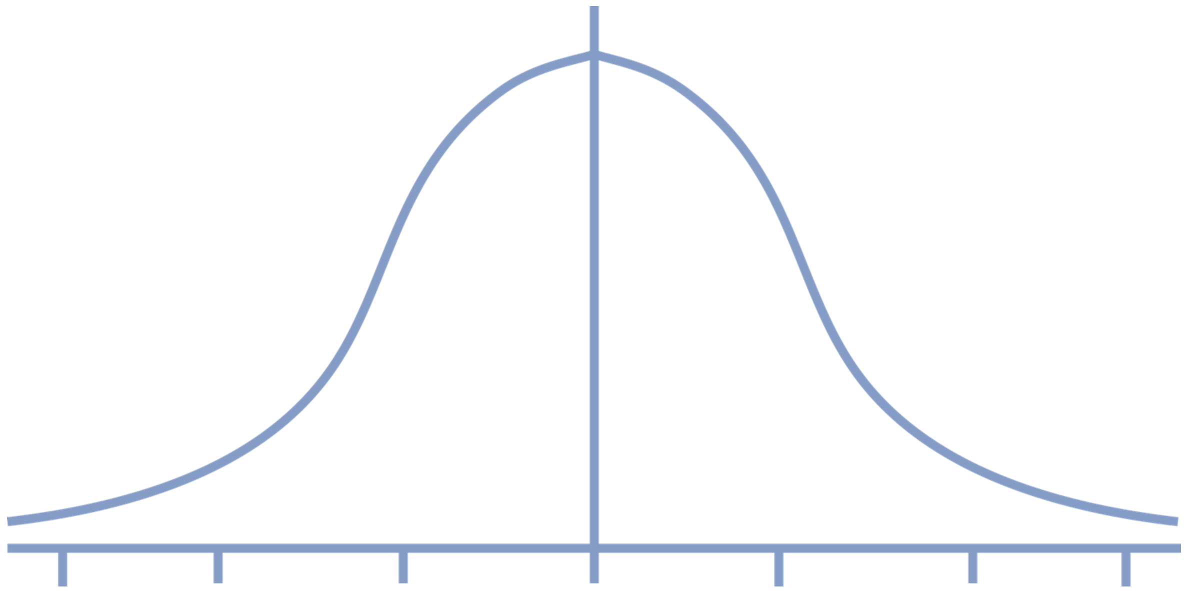
Dataset

The table represents a random sample from a uniform distribution [0, 1].

0.9681	0.6768	0.8466	0.0403	0.4375	0.0519	0.0339	0.4139	0.0569	0.3195
0.5555	0.4571	0.2030	0.1762	0.1571	0.2148	0.8312	0.8833	0.3132	0.4156
0.7849	0.7086	0.8034	0.2164	0.8798	0.7156	0.5037	0.7769	0.4229	0.5184
0.8416	0.7644	0.4125	0.1899	0.9979	0.5510	0.6235	0.7754	0.9087	0.4439
0.2787	0.6668	0.0943	0.9967	0.3901	0.4245	0.9847	0.2857	0.8388	0.0809
0.6216	0.1091	0.5083	0.7062	0.9317	0.8949	0.1526	0.2416	0.4544	0.6265
0.3765	0.7721	0.5282	0.0990	0.2353	0.1911	0.1444	0.9343	0.3735	0.0390
0.3797	0.9394	0.7280	0.0935	0.3321	0.1122	0.9710	0.7004	0.9971	0.1663

Testing Procedure

Let $\bar{x}$ denote the sample mean.

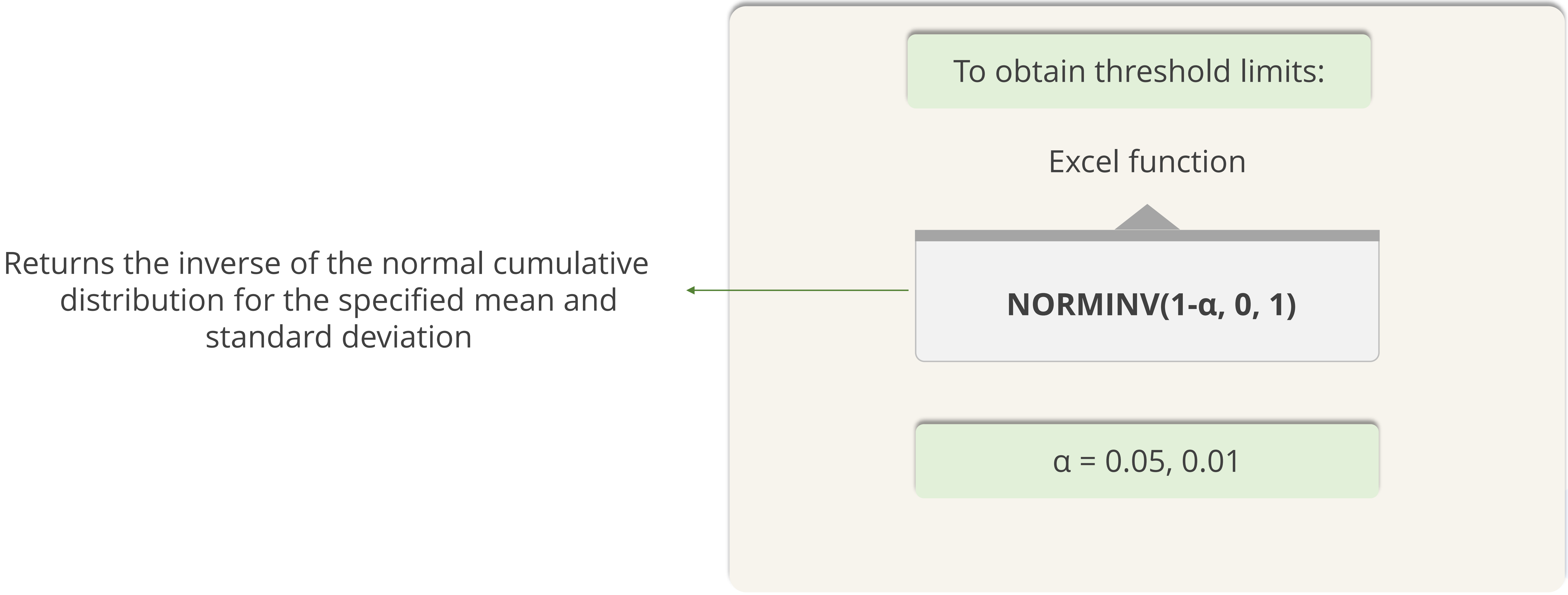


Calculate Z = $((\bar{x} - \mu_0)*(\sqrt{n}/\sigma))$

If Z exceeds a threshold limit, reject H_0.
Otherwise, accept it

Testing Procedure

Z follows a standard normal distribution.



To obtain threshold limits:

Excel function

NORMINV(1-α, 0, 1)

Returns the inverse of the normal cumulative distribution for the specified mean and standard deviation

α = 0.05, 0.01

Presentations of Results

Record the results in the following format:

MEAN (μ_0)	Sample I	Sample II	Sample III	Sample IV	Sample V
95	Reject				
96	Reject				
97	Reject				
98	Reject				
99	Accept				
100	Accept				

Presentations of Results

There will be two such tables as shown in the previous slide, one for each level of significance.

Level of significance

5%

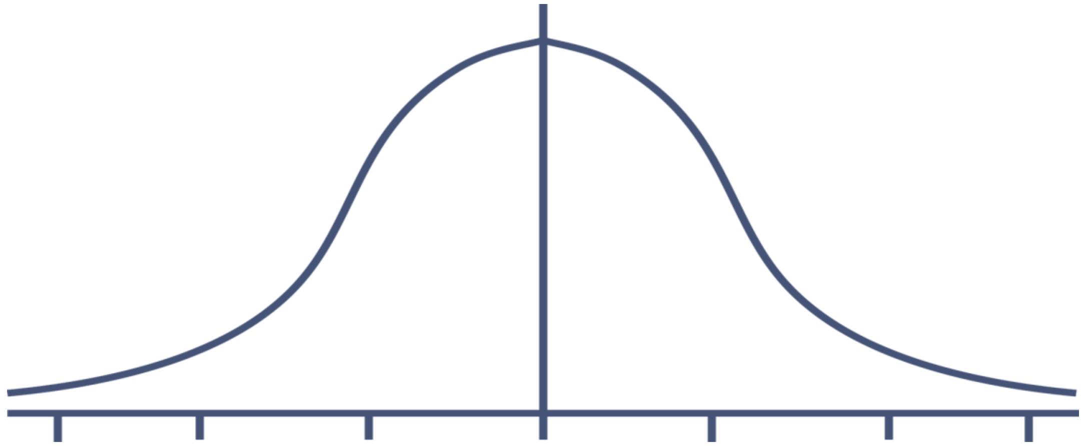
1%

Use the IF statement in Excel to observe the interferences

Inference

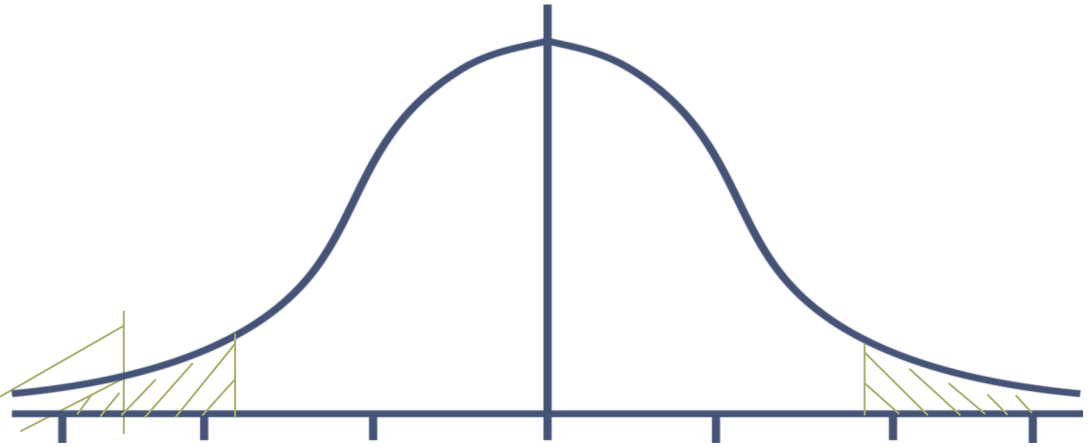
If a given null hypothesis is accepted using a 0.01 level of significance, it will also implicitly be accepted by a higher significance level.

Accept H_0



1%

Accept H_0

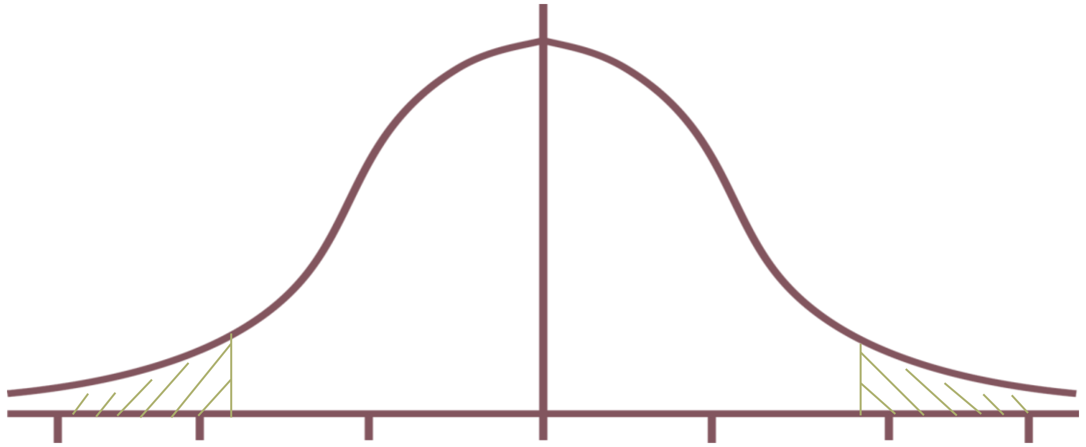


5%

Inference

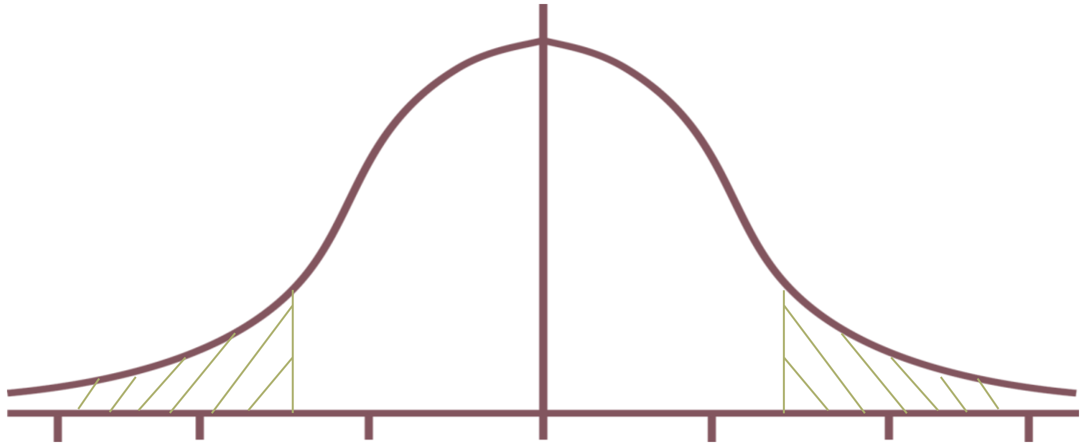
When one rejects H_0 for a 1% level of significance, reject H_0 for a 5% level of significance as well

Reject H_0



1%

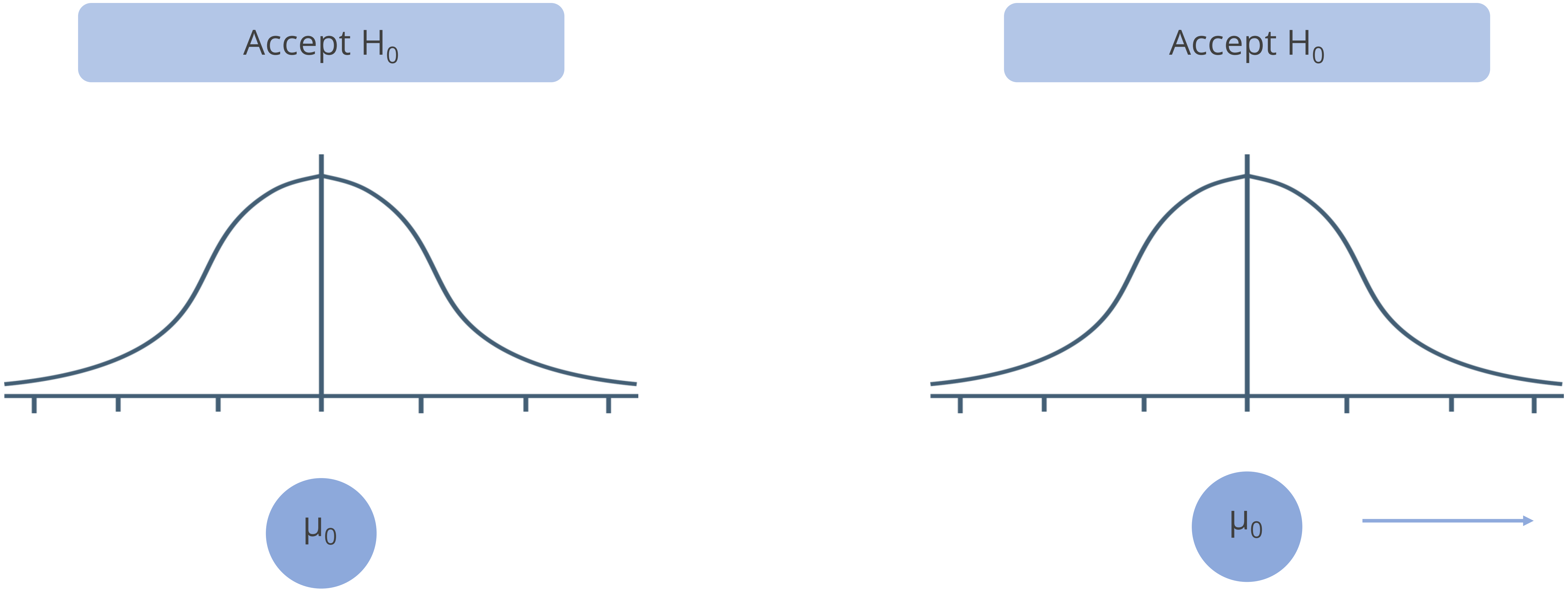
Reject H_0



5%

Inference

Accepting H_0 for one value of μ_0 does not imply acceptance for a higher value; each hypothesis should be evaluated independently based on the specific conditions and evidence.



Accept H_0

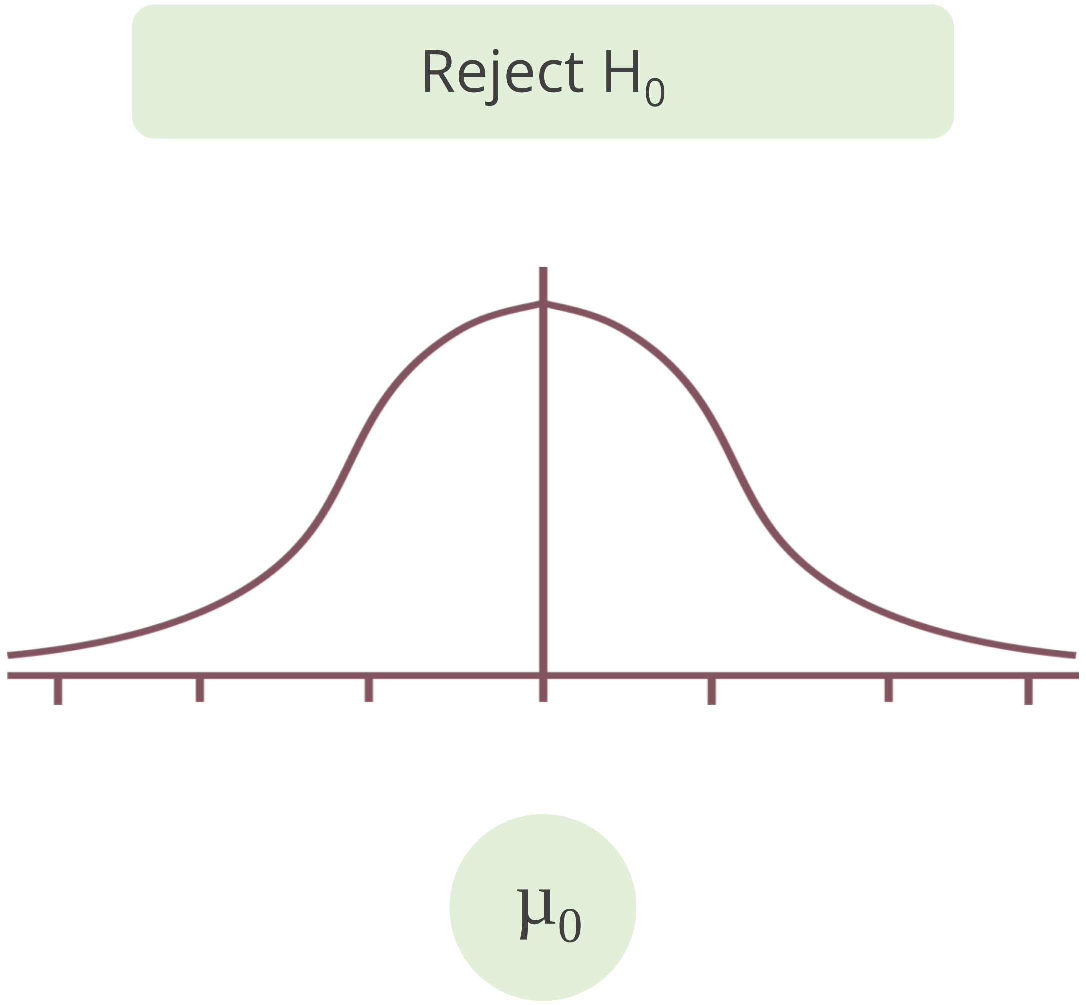
Accept H_0

μ_0

μ_0

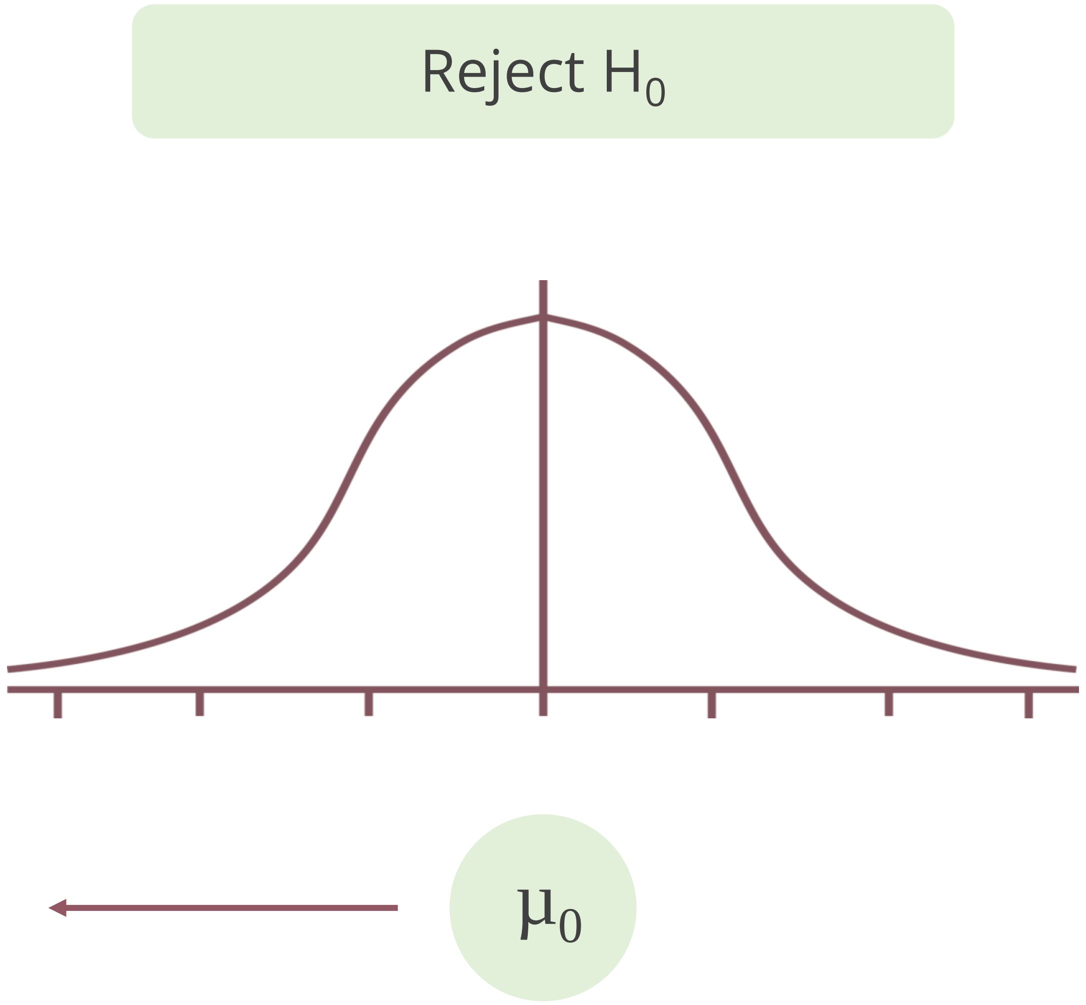
Inference

When one rejects H_0 for a value of μ0, reject H_0 for lower value of μ0



Reject H_0

μ_0

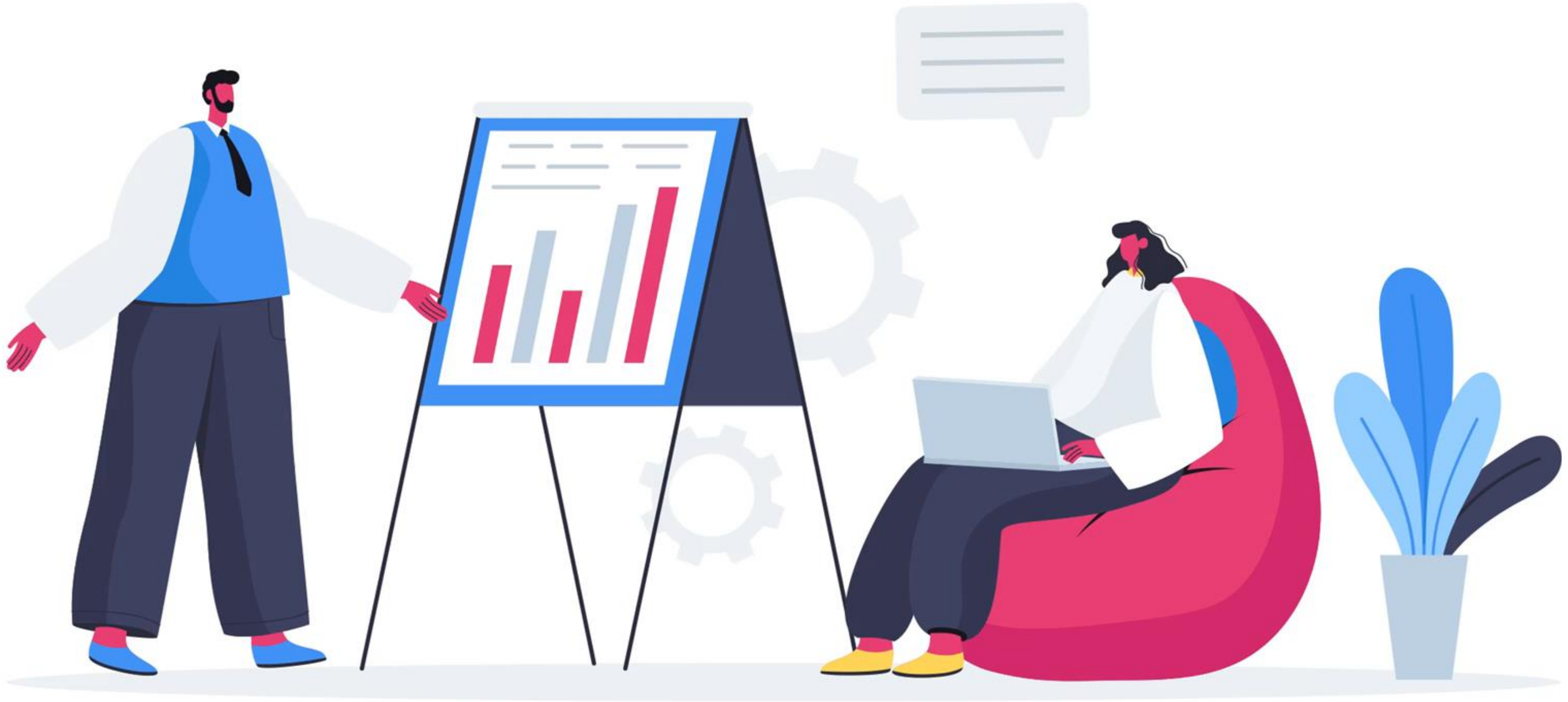


Reject H_0

μ_0

Observations

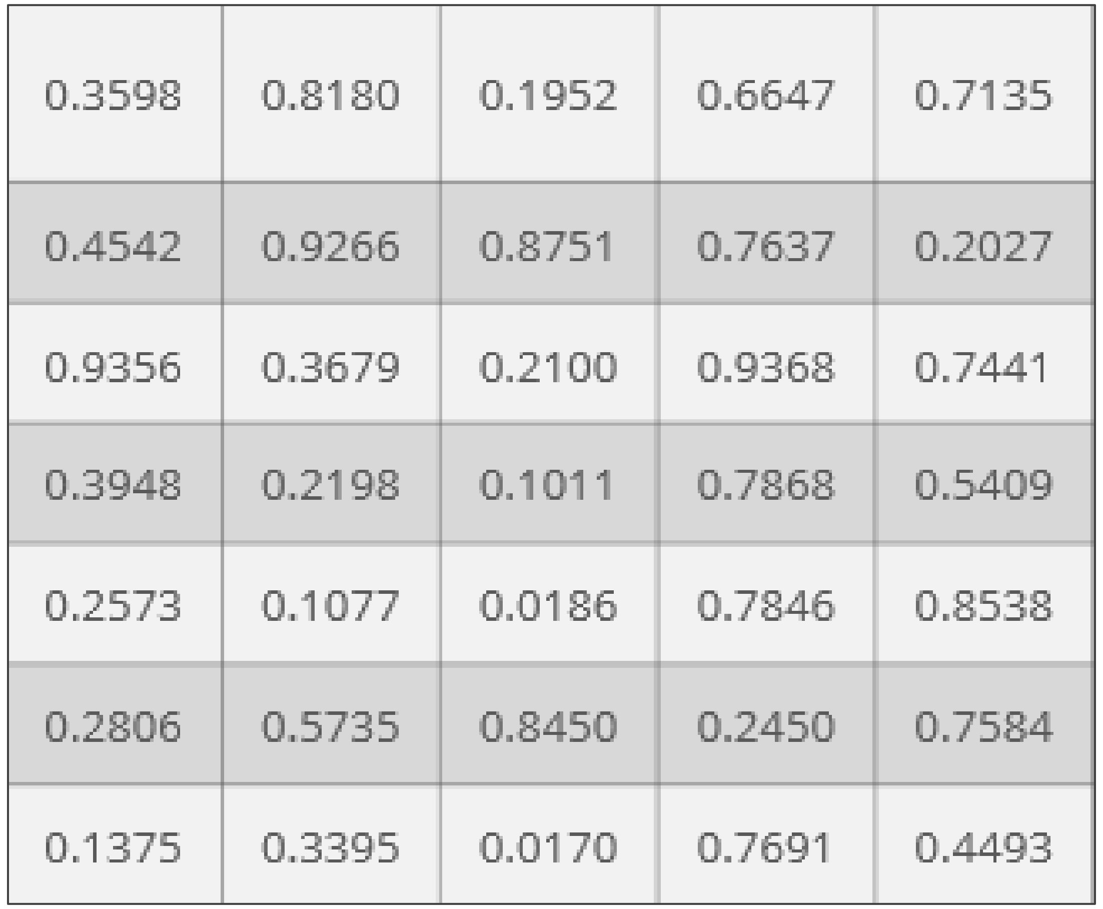
Record observations, if any



Solution

Generate Random Numbers

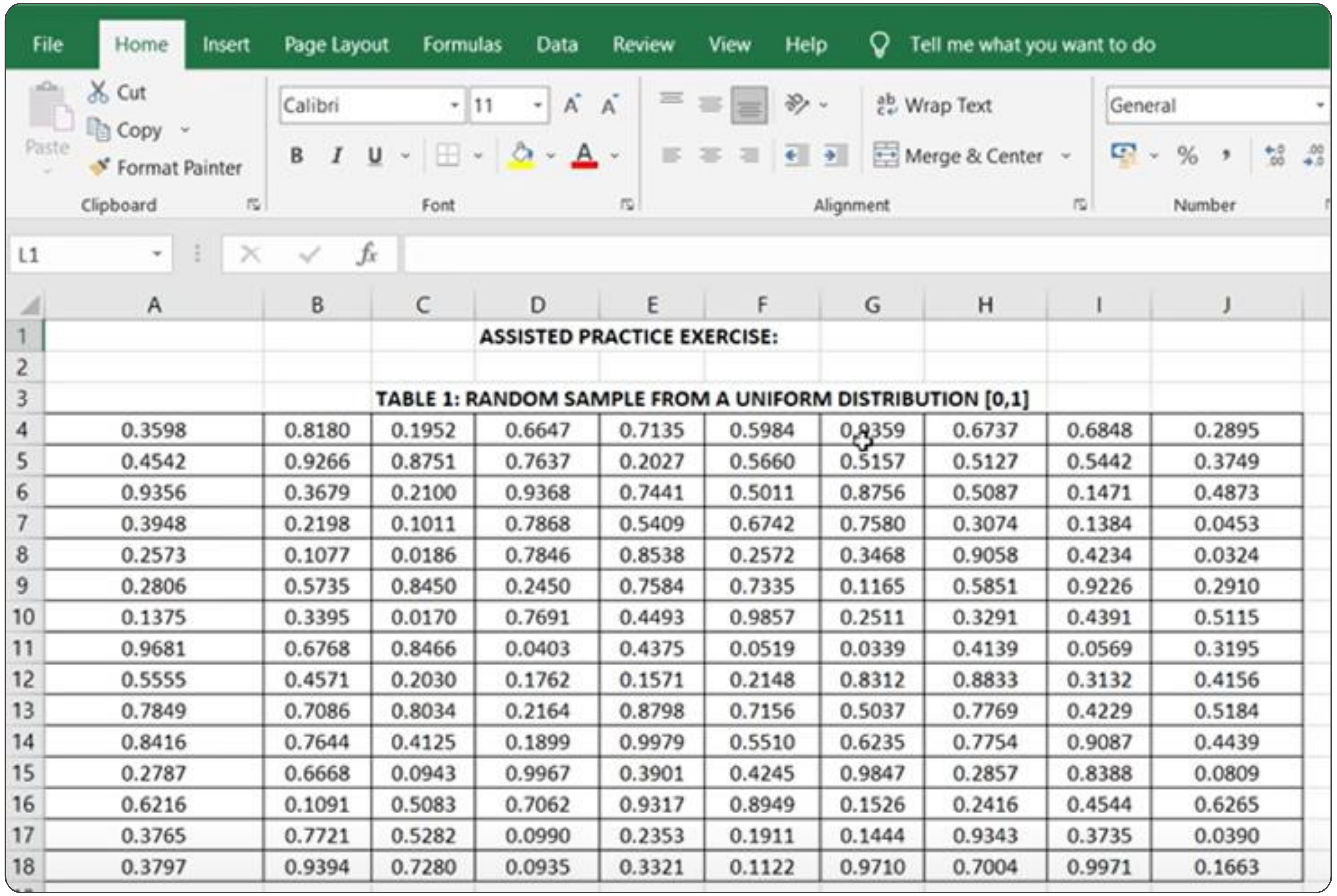
Step 1: Generate data from a uniform distribution generated using the RAND() function



0.3598	0.8180	0.1952	0.6647	0.7135
0.4542	0.9266	0.8751	0.7637	0.2027
0.9356	0.3679	0.2100	0.9368	0.7441
0.3948	0.2198	0.1011	0.7868	0.5409
0.2573	0.1077	0.0186	0.7846	0.8538
0.2806	0.5735	0.8450	0.2450	0.7584
0.1375	0.3395	0.0170	0.7691	0.4493

Generate Random Numbers

Use Excel to perform the task with the four decimal point numbers



ASSISTED PRACTICE EXERCISE:

TABLE 1: RANDOM SAMPLE FROM A UNIFORM DISTRIBUTION [0,1]

	A	B	C	D	E	F	G	H	I	J
4	0.3598	0.8180	0.1952	0.6647	0.7135	0.5984	0.9359	0.6737	0.6848	0.2895
5	0.4542	0.9266	0.8751	0.7637	0.2027	0.5660	0.5157	0.5127	0.5442	0.3749
6	0.9356	0.3679	0.2100	0.9368	0.7441	0.5011	0.8756	0.5087	0.1471	0.4873
7	0.3948	0.2198	0.1011	0.7868	0.5409	0.6742	0.7580	0.3074	0.1384	0.0453
8	0.2573	0.1077	0.0186	0.7846	0.8538	0.2572	0.3468	0.9058	0.4234	0.0324
9	0.2806	0.5735	0.8450	0.2450	0.7584	0.7335	0.1165	0.5851	0.9226	0.2910
10	0.1375	0.3395	0.0170	0.7691	0.4493	0.9857	0.2511	0.3291	0.4391	0.5115
11	0.9681	0.6768	0.8466	0.0403	0.4375	0.0519	0.0339	0.4139	0.0569	0.3195
12	0.5555	0.4571	0.2030	0.1762	0.1571	0.2148	0.8312	0.8833	0.3132	0.4156
13	0.7849	0.7086	0.8034	0.2164	0.8798	0.7156	0.5037	0.7769	0.4229	0.5184
14	0.8416	0.7644	0.4125	0.1899	0.9979	0.5510	0.6235	0.7754	0.9087	0.4439
15	0.2787	0.6668	0.0943	0.9967	0.3901	0.4245	0.9847	0.2857	0.8388	0.0809
16	0.6216	0.1091	0.5083	0.7062	0.9317	0.8949	0.1526	0.2416	0.4544	0.6265
17	0.3765	0.7721	0.5282	0.0990	0.2353	0.1911	0.1444	0.9343	0.3735	0.0390
18	0.3797	0.9394	0.7280	0.0935	0.3321	0.1122	0.9710	0.7004	0.9971	0.1663

Generate Data Using the Random Data

The next step is to generate data from a normal distribution with a mean of 100 and a standard deviation of 10.

Sample size (n) = 150

Normal distribution

Mean = 100

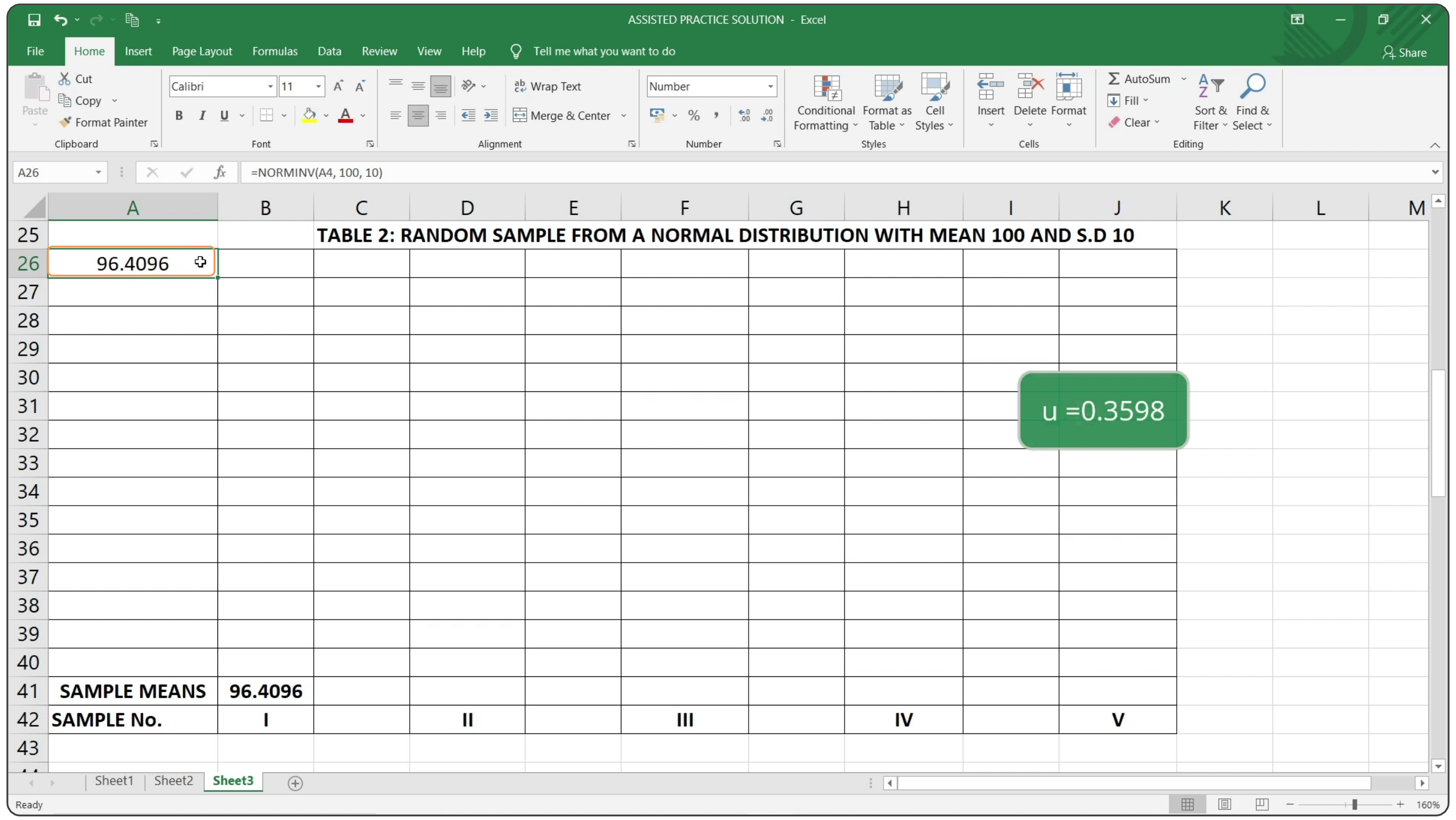
Standard deviation = 10

Use the formula **= NORMINV(u, 100, 10)** to generate the data

Where u is the random number in the corresponding cell

Generate Data Using the Random Data

Step 2: Calculate the number where u = 0.3598 and the obtained value is 96.4096



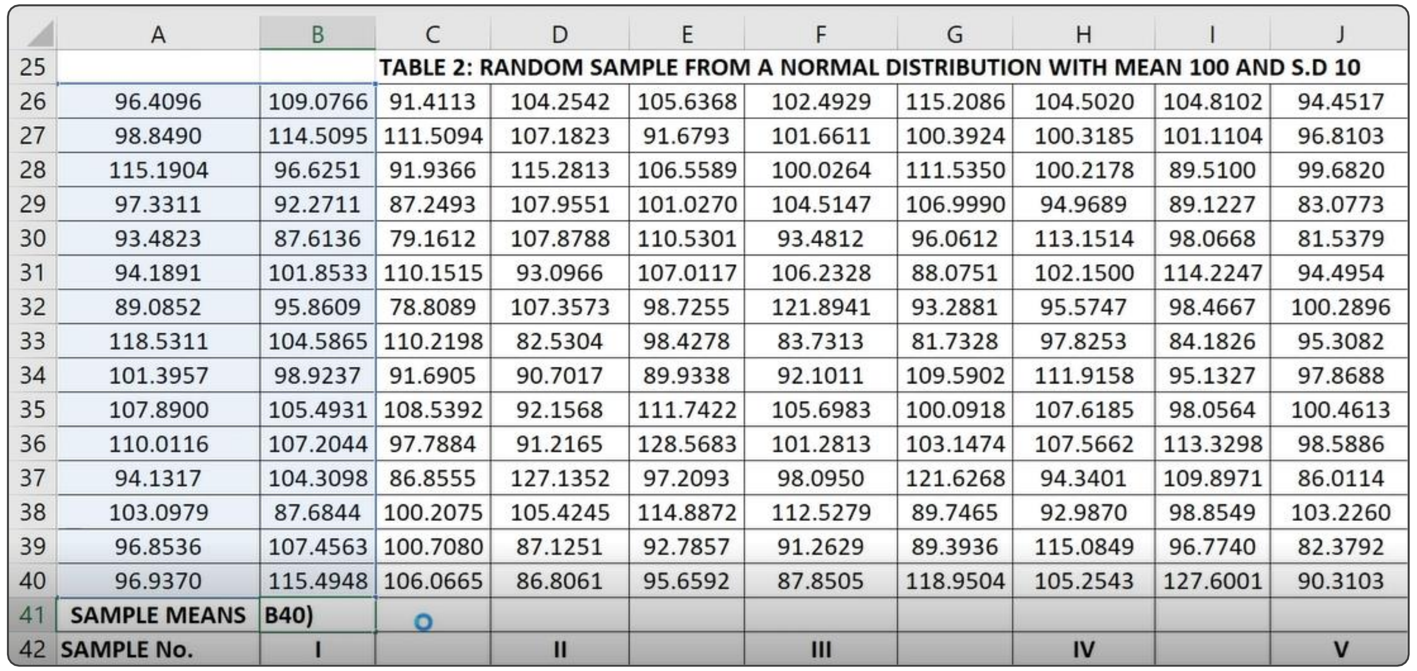
A26 =NORMINV(A4, 100, 10)

	A	B	C	D	E	F	G	H	I	J
25		TABLE 2: RANDOM SAMPLE FROM A NORMAL DISTRIBUTION WITH MEAN 100 AND S.D 10								
26	96.4096									
27										
28										
29										
30										
31										
32										
33										
34										
35										
36										
37										
38										
39										
40										
41	SAMPLE MEANS	96.4096								
42	SAMPLE No.	I		II		III		IV		V
43										

Step 3: To determine the number in Table 2 for every random number in Table 1, drag the original cell to the rest of the cells

Compute the Statistic Values

The next step is to compute the test statistic values for each of the five samples of size n = 30

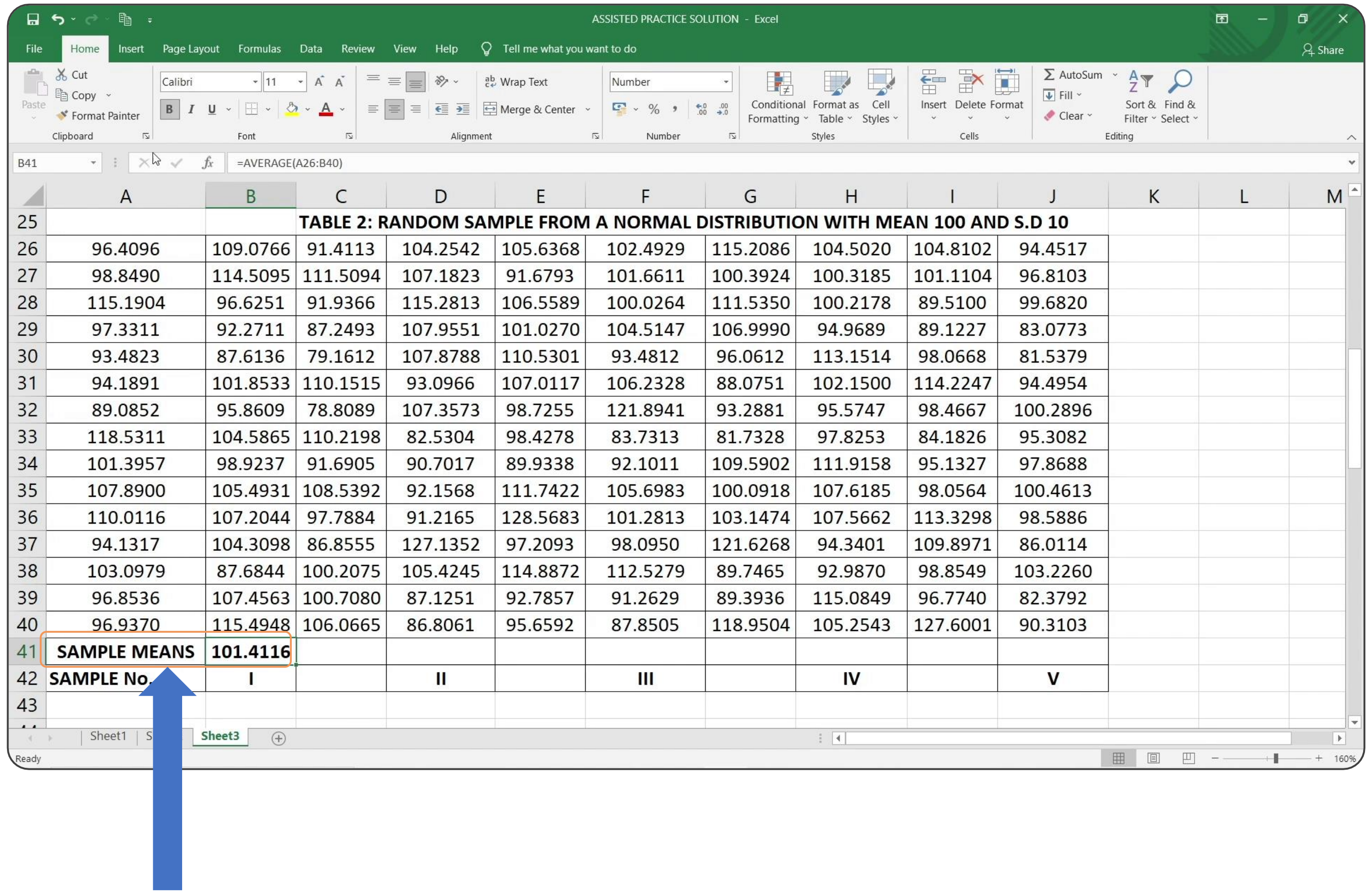


	A	B	C	D	E	F	G	H	I	J
25			TABLE 2: RANDOM SAMPLE FROM A NORMAL DISTRIBUTION WITH MEAN 100 AND S.D 10							
26	96.4096	109.0766	91.4113	104.2542	105.6368	102.4929	115.2086	104.5020	104.8102	94.4517
27	98.8490	114.5095	111.5094	107.1823	91.6793	101.6611	100.3924	100.3185	101.1104	96.8103
28	115.1904	96.6251	91.9366	115.2813	106.5589	100.0264	111.5350	100.2178	89.5100	99.6820
29	97.3311	92.2711	87.2493	107.9551	101.0270	104.5147	106.9990	94.9689	89.1227	83.0773
30	93.4823	87.6136	79.1612	107.8788	110.5301	93.4812	96.0612	113.1514	98.0668	81.5379
31	94.1891	101.8533	110.1515	93.0966	107.0117	106.2328	88.0751	102.1500	114.2247	94.4954
32	89.0852	95.8609	78.8089	107.3573	98.7255	121.8941	93.2881	95.5747	98.4667	100.2896
33	118.5311	104.5865	110.2198	82.5304	98.4278	83.7313	81.7328	97.8253	84.1826	95.3082
34	101.3957	98.9237	91.6905	90.7017	89.9338	92.1011	109.5902	111.9158	95.1327	97.8688
35	107.8900	105.4931	108.5392	92.1568	111.7422	105.6983	100.0918	107.6185	98.0564	100.4613
36	110.0116	107.2044	97.7884	91.2165	128.5683	101.2813	103.1474	107.5662	113.3298	98.5886
37	94.1317	104.3098	86.8555	127.1352	97.2093	98.0950	121.6268	94.3401	109.8971	86.0114
38	103.0979	87.6844	100.2075	105.4245	114.8872	112.5279	89.7465	92.9870	98.8549	103.2260
39	96.8536	107.4563	100.7080	87.1251	92.7857	91.2629	89.3936	115.0849	96.7740	82.3792
40	96.9370	115.4948	106.0665	86.8061	95.6592	87.8505	118.9504	105.2543	127.6001	90.3103
41	SAMPLE MEANS	B40)								
42	SAMPLE No.	I		II		III		IV		V

- Calculate the mean $\bar{x}$ for each sample
- In the Excel sheet, values of the first sample are indicated in the cells A26 to A40 and B26 to B40.

Compute the Statistic Values

The values of Sample I are listed in columns A and B.



=AVERAGE(A26:B40)

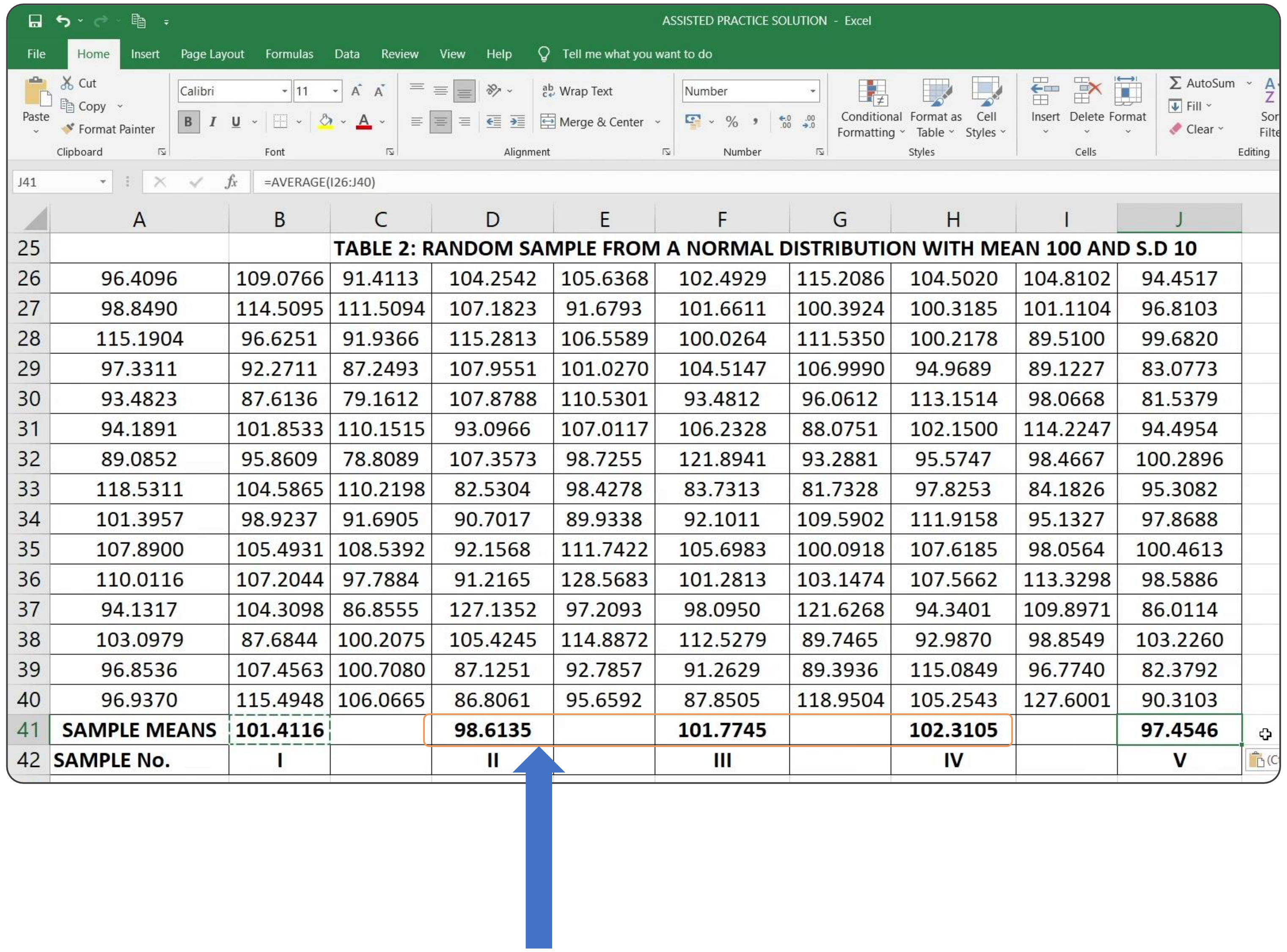
	A	B	C	D	E	F	G	H	I	J
25			TABLE 2: RANDOM SAMPLE FROM A NORMAL DISTRIBUTION WITH MEAN 100 AND S.D 10							
26	96.4096	109.0766	91.4113	104.2542	105.6368	102.4929	115.2086	104.5020	104.8102	94.4517
27	98.8490	114.5095	111.5094	107.1823	91.6793	101.6611	100.3924	100.3185	101.1104	96.8103
28	115.1904	96.6251	91.9366	115.2813	106.5589	100.0264	111.5350	100.2178	89.5100	99.6820
29	97.3311	92.2711	87.2493	107.9551	101.0270	104.5147	106.9990	94.9689	89.1227	83.0773
30	93.4823	87.6136	79.1612	107.8788	110.5301	93.4812	96.0612	113.1514	98.0668	81.5379
31	94.1891	101.8533	110.1515	93.0966	107.0117	106.2328	88.0751	102.1500	114.2247	94.4954
32	89.0852	95.8609	78.8089	107.3573	98.7255	121.8941	93.2881	95.5747	98.4667	100.2896
33	118.5311	104.5865	110.2198	82.5304	98.4278	83.7313	81.7328	97.8253	84.1826	95.3082
34	101.3957	98.9237	91.6905	90.7017	89.9338	92.1011	109.5902	111.9158	95.1327	97.8688
35	107.8900	105.4931	108.5392	92.1568	111.7422	105.6983	100.0918	107.6185	98.0564	100.4613
36	110.0116	107.2044	97.7884	91.2165	128.5683	101.2813	103.1474	107.5662	113.3298	98.5886
37	94.1317	104.3098	86.8555	127.1352	97.2093	98.0950	121.6268	94.3401	109.8971	86.0114
38	103.0979	87.6844	100.2075	105.4245	114.8872	112.5279	89.7465	92.9870	98.8549	103.2260
39	96.8536	107.4563	100.7080	87.1251	92.7857	91.2629	89.3936	115.0849	96.7740	82.3792
40	96.9370	115.4948	106.0665	86.8061	95.6592	87.8505	118.9504	105.2543	127.6001	90.3103
41	SAMPLE MEANS	101.4116								
42	SAMPLE No.	I		II		III		IV		V
43										

Step 4:

- Select the cell where the average of the first mean should be displayed
- Use the formula **=AVERAGE(A26:B40)** to calculate the mean for the first sample

Compute the Statistic Values

Step 5: Copy and paste the mean of Sample I to the rest of the cells where the mean of other samples needs to be displayed

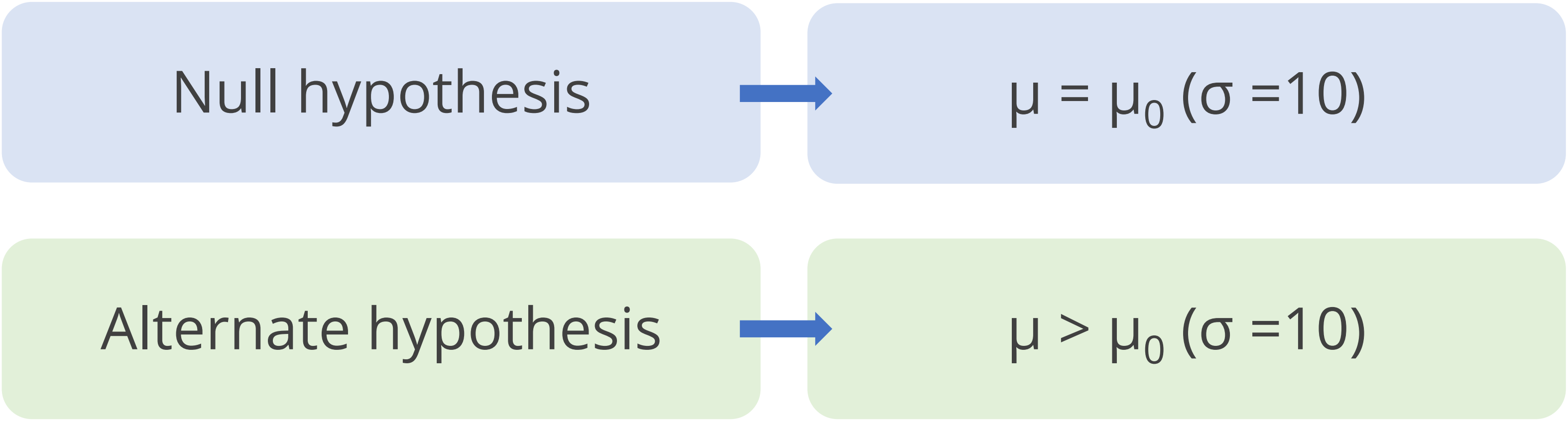


J41 | =AVERAGE(I26:J40)

	A	B	C	D	E	F	G	H	I	J
25			TABLE 2: RANDOM SAMPLE FROM A NORMAL DISTRIBUTION WITH MEAN 100 AND S.D 10							
26	96.4096	109.0766	91.4113	104.2542	105.6368	102.4929	115.2086	104.5020	104.8102	94.4517
27	98.8490	114.5095	111.5094	107.1823	91.6793	101.6611	100.3924	100.3185	101.1104	96.8103
28	115.1904	96.6251	91.9366	115.2813	106.5589	100.0264	111.5350	100.2178	89.5100	99.6820
29	97.3311	92.2711	87.2493	107.9551	101.0270	104.5147	106.9990	94.9689	89.1227	83.0773
30	93.4823	87.6136	79.1612	107.8788	110.5301	93.4812	96.0612	113.1514	98.0668	81.5379
31	94.1891	101.8533	110.1515	93.0966	107.0117	106.2328	88.0751	102.1500	114.2247	94.4954
32	89.0852	95.8609	78.8089	107.3573	98.7255	121.8941	93.2881	95.5747	98.4667	100.2896
33	118.5311	104.5865	110.2198	82.5304	98.4278	83.7313	81.7328	97.8253	84.1826	95.3082
34	101.3957	98.9237	91.6905	90.7017	89.9338	92.1011	109.5902	111.9158	95.1327	97.8688
35	107.8900	105.4931	108.5392	92.1568	111.7422	105.6983	100.0918	107.6185	98.0564	100.4613
36	110.0116	107.2044	97.7884	91.2165	128.5683	101.2813	103.1474	107.5662	113.3298	98.5886
37	94.1317	104.3098	86.8555	127.1352	97.2093	98.0950	121.6268	94.3401	109.8971	86.0114
38	103.0979	87.6844	100.2075	105.4245	114.8872	112.5279	89.7465	92.9870	98.8549	103.2260
39	96.8536	107.4563	100.7080	87.1251	92.7857	91.2629	89.3936	115.0849	96.7740	82.3792
40	96.9370	115.4948	106.0665	86.8061	95.6592	87.8505	118.9504	105.2543	127.6001	90.3103
41	**SAMPLE MEANS**	**101.4116**		**98.6135**		**101.7745**		**102.3105**		**97.4546**
42	**SAMPLE No.**	**I**		**II**		**III**		**IV**		**V**

Execute the Hypothesis Test

The next step is to perform the hypothesis test.



Null hypothesis → $\mu = \mu_0$ (σ =10)

Alternate hypothesis → $\mu > \mu_0$ (σ =10)

Values of μ_0 should be varied to take the values 95, 96, 97, 98, 99, and 100.

Execute the Hypothesis Test

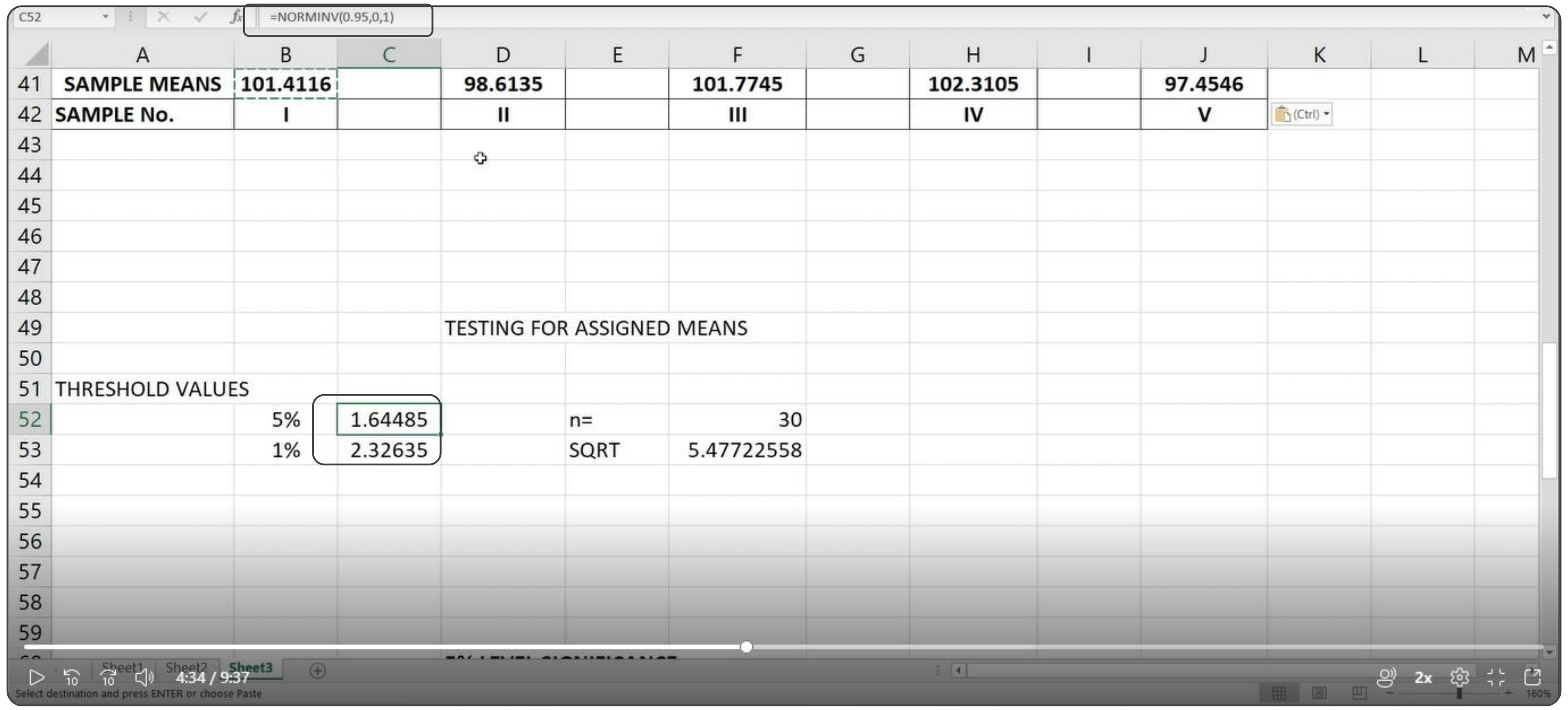
In hypothesis testing:

- Null hypothesis (H_0) can be rejected when $\bar{x} > c$
- c is chosen such that P ($\bar{x} >$ c when H_0 is true) = α
- This is equivalent to:

$P(Z = ((\bar{x} - \mu_0) * (\sqrt{n}/\sigma)) > ((c - \mu_0) * (\sqrt{n}/\sigma))$ when H_0 is true$) = \alpha$

Execute the Hypothesis Test

Use the Excel formula **NORMINV (1-α, 0, 1)** to obtain the threshold value for the standard normal distribution.



C52 =NORMINV(0.95,0,1)

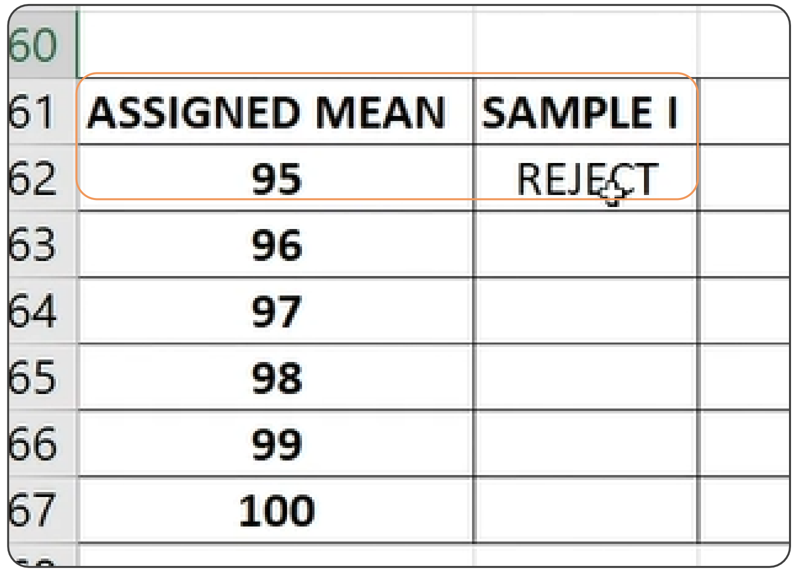
	A	B	C	D	E	F	G	H	I	J	K	L	M
41	**SAMPLE MEANS**	**101.4116**		**98.6135**		**101.7745**		**102.3105**		**97.4546**			
42	**SAMPLE No.**	**I**		**II**		**III**		**IV**		**V**			
43													
44													
45													
46													
47													
48													
49				TESTING FOR ASSIGNED MEANS									
50													
51	THRESHOLD VALUES												
52		5%	1.64485		n=	30							
53		1%	2.32635		SQRT	5.47722558							
54													
55													
56													
57													
58													
59													

α can either be 5% or 1%.

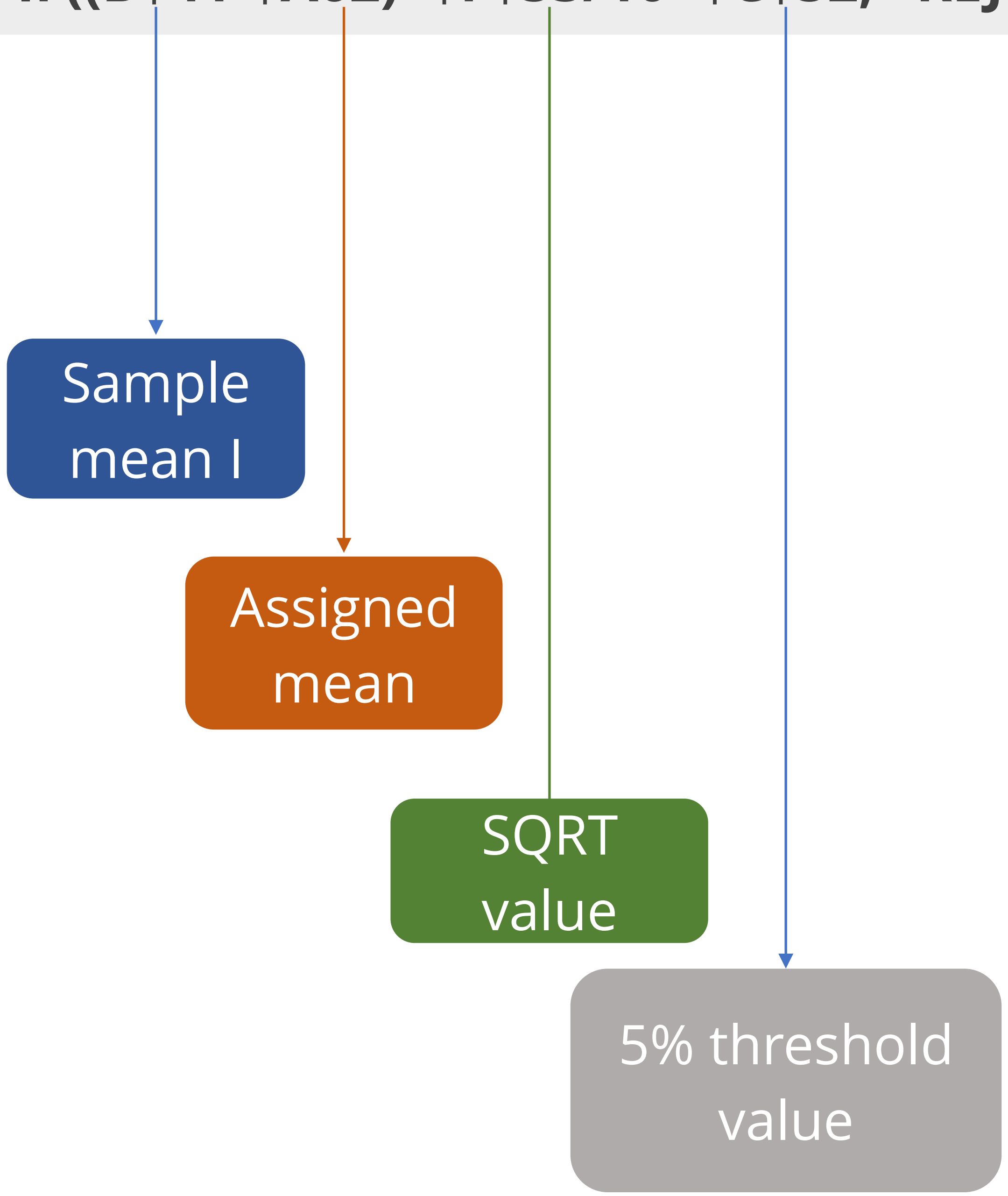
The cumulative distribution values are 1.644853627 and 2.32635 for 5% and 1% levels, respectively.

Execute the Hypothesis Test

Use the Excel formula =**IF((B$41-$A62)*$F$53/10>$C$52, "REJECT", "ACCEPT")**.



60		
61	ASSIGNED MEAN	SAMPLE I
62	95	REJECT
63	96	
64	97	
65	98	
66	99	
67	100	



Sample mean I

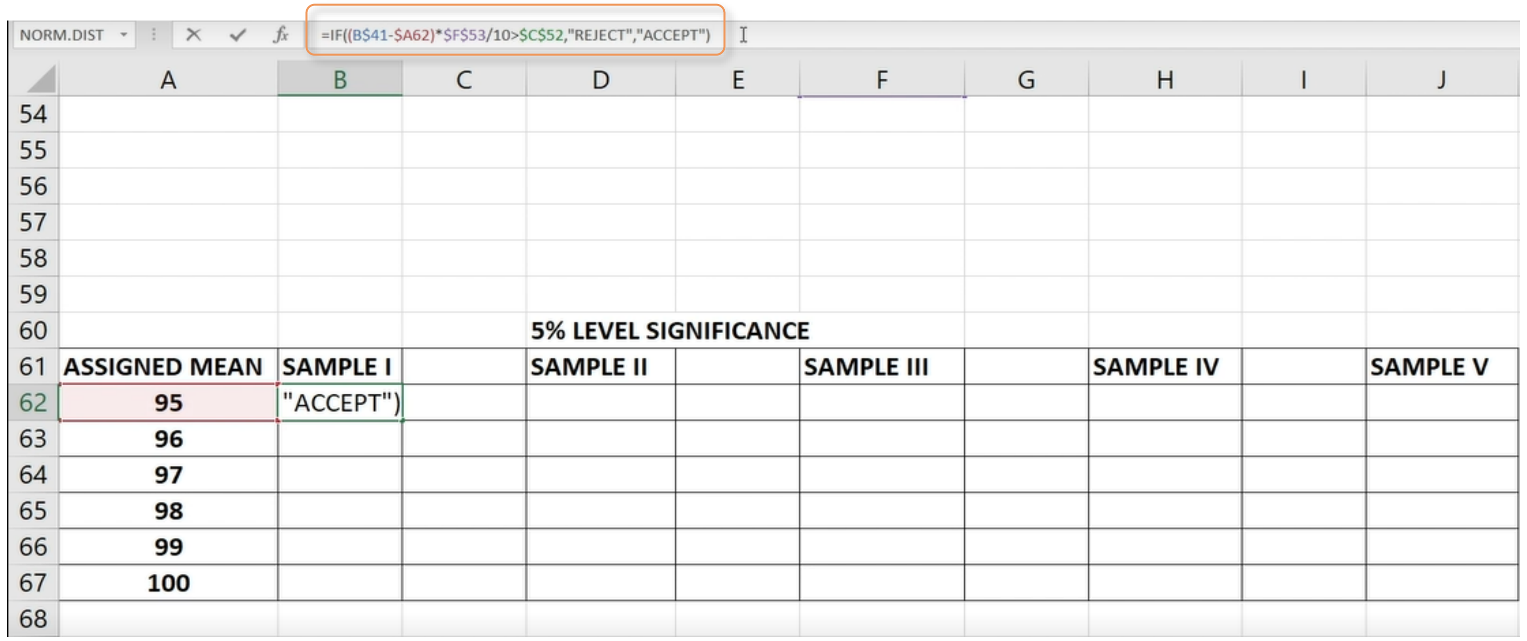
Assigned mean

SQRT value

5% threshold value

Execute the Hypothesis Test

The inference for Sample I when μ_0 = 95 is stated as **REJECT**.



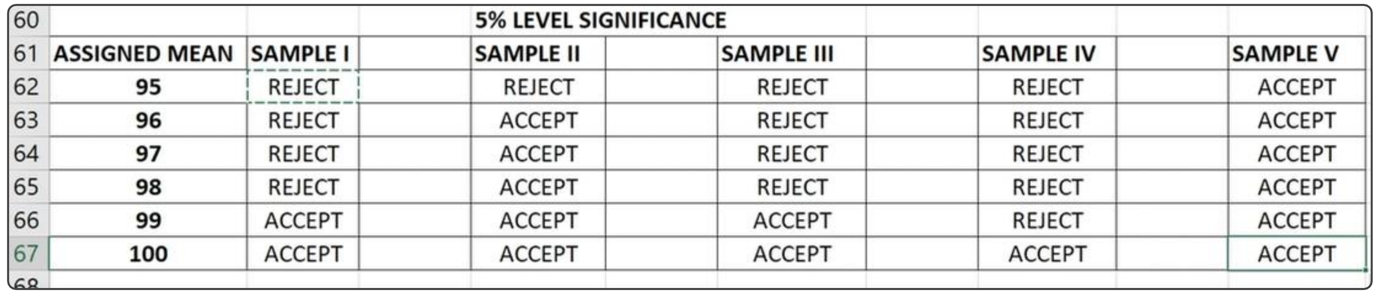
```
=IF((B$41-$A62)*$F$53/10>$C$52,"REJECT","ACCEPT")
```

	A	B	C	D	E	F	G	H	I	J
54										
55										
56										
57										
58										
59										
60				5% LEVEL SIGNIFICANCE						
61	ASSIGNED MEAN	SAMPLE I		SAMPLE II		SAMPLE III		SAMPLE IV		SAMPLE V
62	95	"ACCEPT")								
63	96									
64	97									
65	98									
66	99									
67	100									
68										

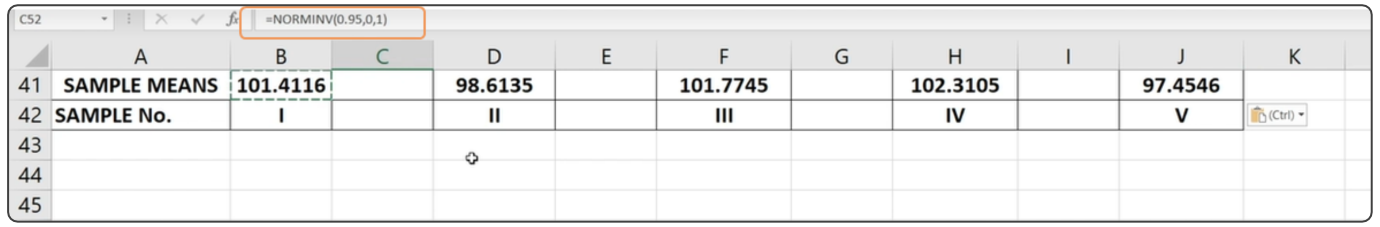
- The sample mean for the first sample: B41
- The threshold limits: C52 and C53 (for 5% and 1%, respectively)
- The value of n: F52 and the square root of n: F53
- The first value of assigned mean (95): A62

The copy and paste command replicates the formula entered in B62 across all tables.

Execute the Hypothesis Test



60		5% LEVEL SIGNIFICANCE				
61	ASSIGNED MEAN	SAMPLE I	SAMPLE II	SAMPLE III	SAMPLE IV	SAMPLE V
62	95	REJECT	REJECT	REJECT	REJECT	ACCEPT
63	96	REJECT	ACCEPT	REJECT	REJECT	ACCEPT
64	97	REJECT	ACCEPT	REJECT	REJECT	ACCEPT
65	98	REJECT	ACCEPT	REJECT	REJECT	ACCEPT
66	99	ACCEPT	ACCEPT	ACCEPT	REJECT	ACCEPT
67	100	ACCEPT	ACCEPT	ACCEPT	ACCEPT	ACCEPT



C52 =NORMINV(0.95,0,1)

	A	B	C	D	E	F	G	H	I	J	K
41	SAMPLE MEANS	101.4116		98.6135		101.7745		102.3105		97.4546	
42	SAMPLE No.	I		II		III		IV		V	
43											
44											
45											

- The sample mean recorded in cell B41 is to be used for all calculations in column B (B62 to B67).
- If, however, B41 and not B$41 is used, the formula would use values in B42, B43, B44 and B45.
- Numerical values cannot be calculated in these cells if they do not appear (as in this case).
- $C$52, is set as a constant value.

Observations

The observations for the five samples are as shown below:

Assigned Mean (μ_0)	Sample I	Sample II	Sample III	Sample IV	Sample V
95	Reject	Reject	Reject	Reject	Accept
96	Reject	Accept	Reject	Reject	Accept
97	Reject	Accept	Reject	Reject	Accept
98	Reject	Accept	Reject	Reject	Accept
99	Accept	Accept	Accept	Reject	Accept
100	Accept	Accept	Accept	Accept	Accept

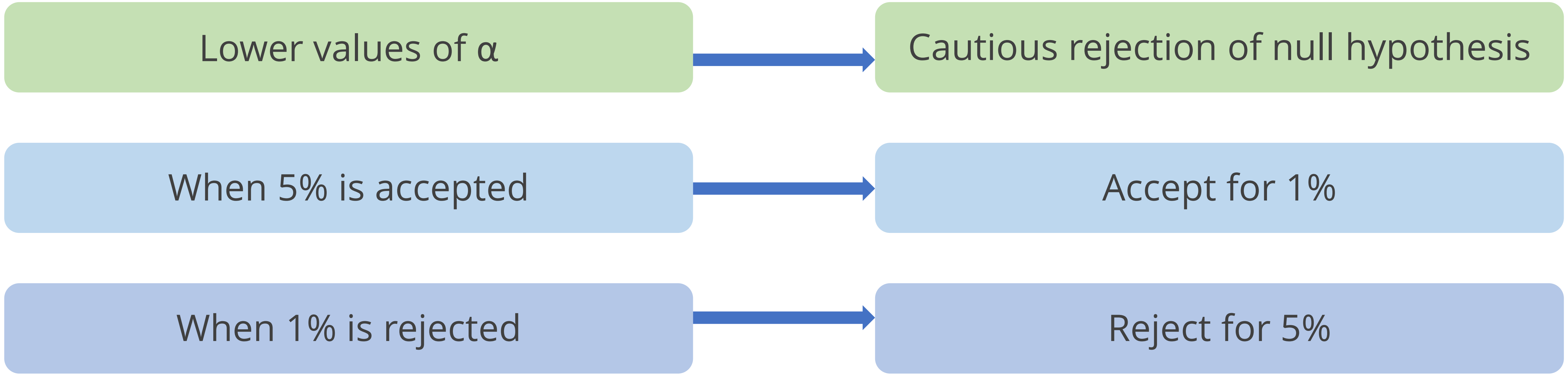
Results for 5% level of significance

Assigned Mean (μ_0)	Sample I	Sample II	Sample III	Sample IV	Sample V
95	Reject	Accept	Reject	Reject	Accept
96	Reject	Accept	Reject	Reject	Accept
97	Reject	Accept	Reject	Reject	Accept
98	Accept	Accept	Accept	Reject	Accept
99	Accept	Accept	Accept	Accept	Accept
100	Accept	Accept	Accept	Accept	Accept

Results for 1% level of significance

Observations

The value of α is the probability of rejecting the null hypothesis when it is true.



Lower values of α → Cautious rejection of null hypothesis

When 5% is accepted → Accept for 1%

When 1% is rejected → Reject for 5%

Observations

The critical region gets narrower as μ_0 increases.

Assigned Mean (μ_0)	Sample I	Sample II	Sample III	Sample IV	Sample V
95	Reject	Reject	Reject	Reject	Accept
96	Reject	Accept	Reject	Reject	Accept
97	Reject	Accept	Reject	Reject	Accept
98	Reject	Accept	Reject	Reject	Accept
99	Accept	Accept	Accept	Reject	Accept
100	Accept	Accept	Accept	Accept	Accept

The value ($\bar{x}$ ($_0\mu$ - decreases as μ_0 increases.

Observations

The risks are controlled by:

Setting the value of α at a certain level

Selecting a reasonably large sample size

Observations

Inferences depend on:

Null and alternate hypothesis

Sample size and sample data

Level of significance

Thank You