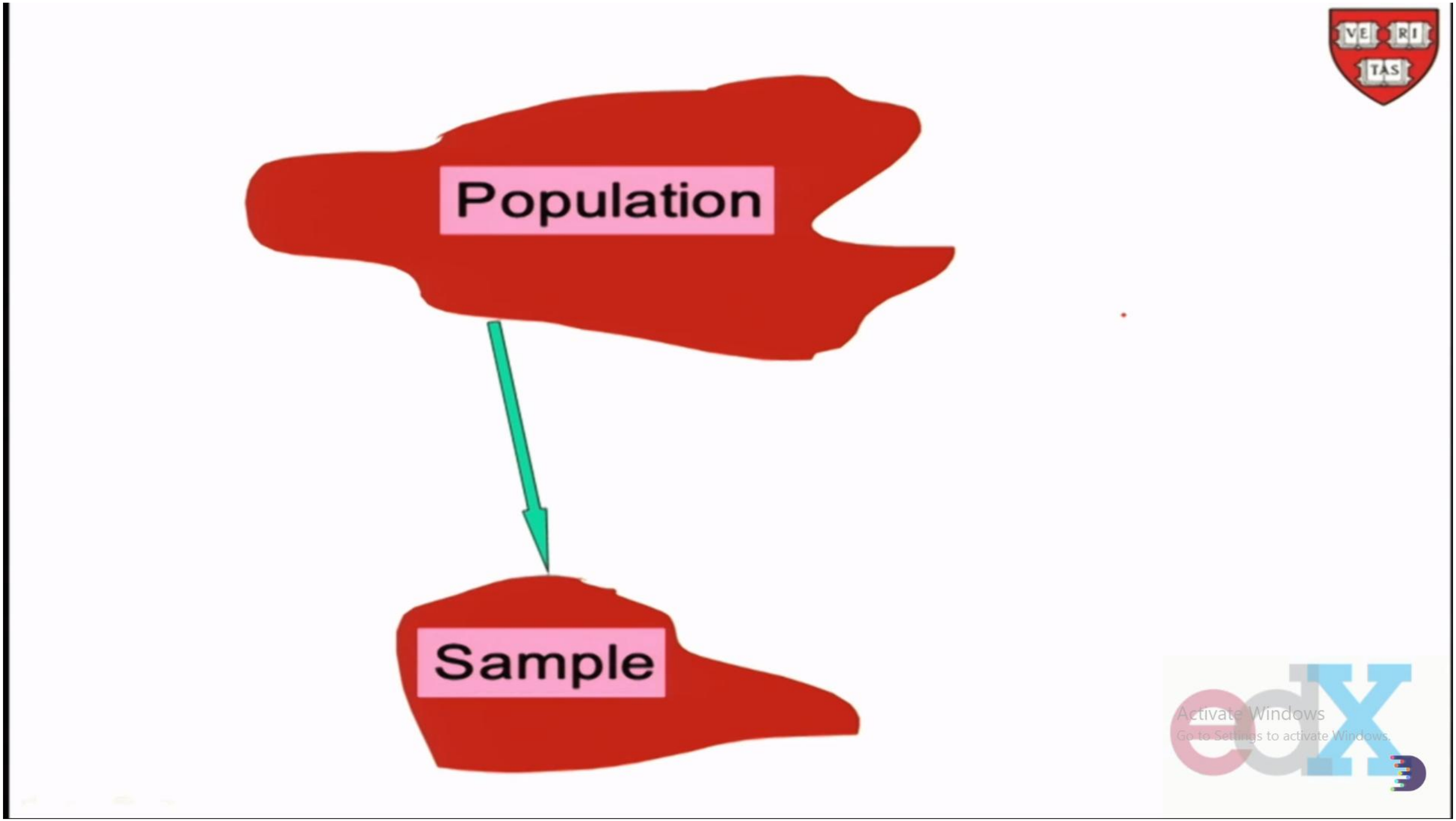
Statistics

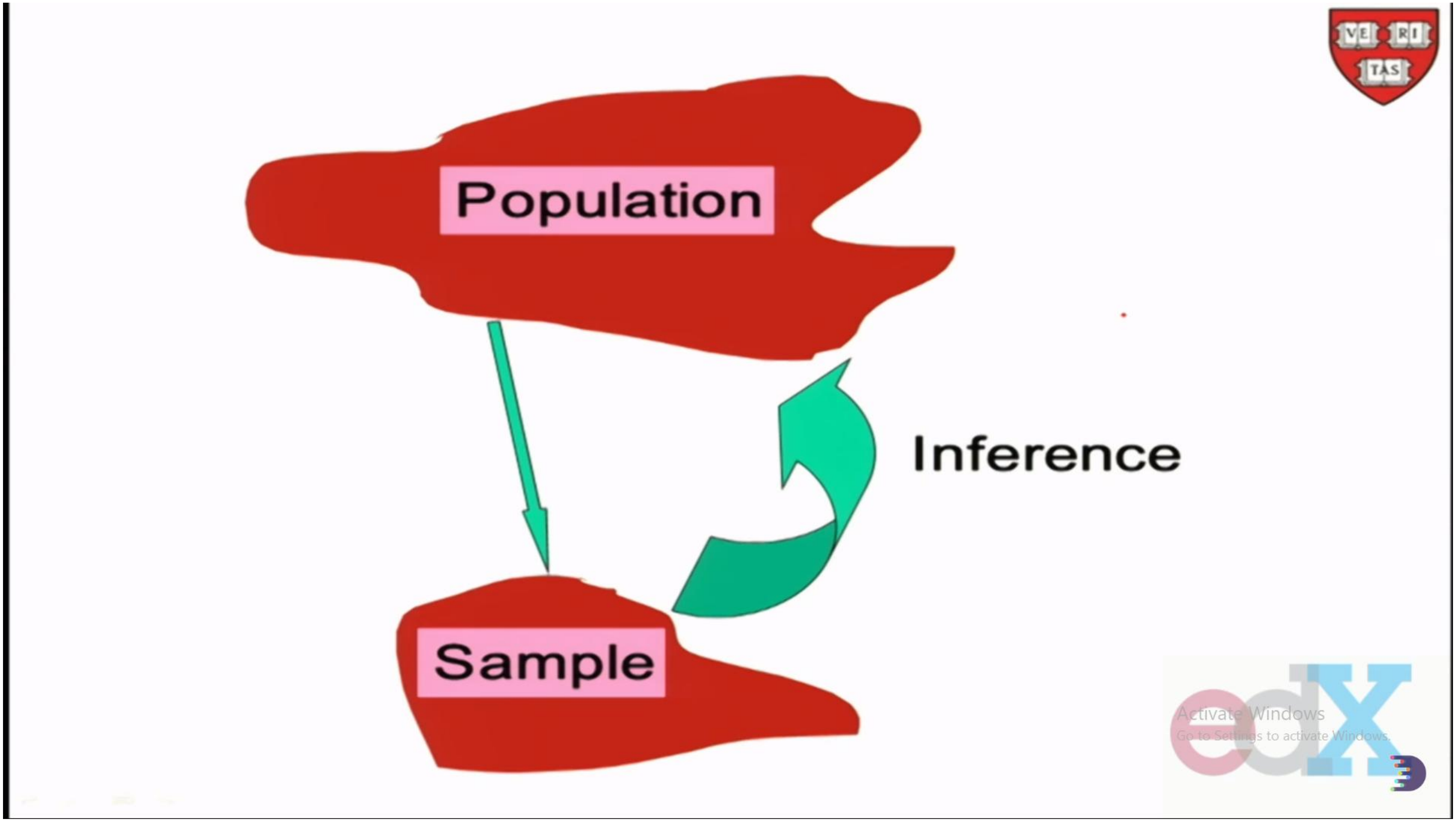
1º. Variability

2º. Inference

3º. Probability


Population


Population
Sample


Population
Inference
Sample

Population

- Can be real, can be conceptual.
- Can be past, current or future.

Population

- Can be real, can be conceptual.
- Can be past, current or future.
- The more homogeneous it is, the easier it is to describe.

Population

- Can be real, can be conceptual.
- Can be past, current or future.
- The more homogeneous it is,
  the easier it is to describe.

e.g. Let us think of the Framingham Heart Study participants as our population.

Our Population

```
summ death angina totchol1 sysbp1 diabp1 bmi1 glucose1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
death	4434	.3495715	.4768884	0	1
angina	4434	.1635092	.3698714	0	1
totchol1	4382	236.9843	44.6511	107	696
sysbp1	4434	132.9078	22.4216	83.5	295
diabp1	4434	83.08356	12.056	48	142.5
bmi1	4415	25.84616	4.101821	15.54	56.8
glucose1	4037	82.18578	24.39958	40	394

Sample

- Must be representative

Sample

- Must be representative

  (random).

e.g. Suppose we take a sample of size 49 from our population.

```
. set seed 72576466

. sample 49, count
(4385 observations deleted)

. summ death angina totchol1 sysbp1 diabp1 bmi1 glucose1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       death |         49    .3265306    .4738035          0          1
      angina |         49    .1632653    .3734378          0          1
    totchol1 |         48    233.6667    42.73836        157        410
      sysbp1 |         49    140.3571    22.39745         96        196
      diabp1 |         49    86.53061    12.65669         48        119
-------------+---------------------------------------------------------
        bmi1 |         48    27.21167    4.018565       19.2      36.65
    glucose1 |         43    79.32558     12.2255         60        114
```

```
. set seed 72576466

. sample 49, count
(4385 observations deleted)

. summ death angina totchol1 sysbp1 diabp1 bmi1 glucose1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
death	49	.3265306	.4738035	0	1
angina	49	.1632653	.3734378	0	1
totchol1	48	233.6667	42.73836	157	410
sysbp1	49	140.3571	22.39745	96	196
diabp1	49	86.53061	12.65669	48	119
bmi1	48	27.21167	4.018565	19.2	36.65
glucose1	43	79.32558	12.2255	60	114

```
. summ death angina totchol1 sysbp1 diabp1 bmi1 glucose1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
death	4434	.3495715	.4768884	0	1
angina	4434	.1635092	.3698714	0	1
totchol1	4382	236.9843	44.6511	107	696
sysbp1	4434	132.9078	22.4216	83.5	295
diabp1	4434	83.08356	12.056	48	142.5
bmi1	4415	25.84616	4.101821	15.54	56.8
glucose1	4037	82.18578	24.39958	40	394

```
. summ death angina totchol1 sysbp1 diabp1 bmi1 glucose1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
death	49	.3265306	.4738035	0	1
angina	49	.1632653	.3734378	0	1
totchol1	48	233.6667	42.73836	157	410
sysbp1	49	140.3571	22.39745	96	196
diabp1	49	86.53061	12.65669	48	119
bmi1	48	27.21167	4.018565	19.2	36.65
glucose1	43	79.32558	12.2255	60	114

```
. summ death angina totchol1 sysbp1 diabp1 bmi1 glucose1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
death	4434	.3495715	.4768884	0	1
angina	4434	.1635092	.3698714	0	1
totchol1	4382	236.9843	44.6511	107	696
sysbp1	4434	132.9078	22.4216	83.5	295
diabp1	4434	83.08356	12.056	48	142.5
bmi1	4415	25.84616	4.101821	15.54	56.8
glucose1	4037	82.18578	24.39958	40	394

```
. summ death angina totchol1 sysbp1 diabp1 bmi1 glucose1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
death	49	.3265306	.4738035	0	1
angina	49	.1632653	.3734378	0	1
totchol1	48	233.6667	42.73836	157	410
sysbp1	49	140.3571	22.39745	96	196
diabp1	49	86.53061	12.65669	48	119
bmi1	48	27.21167	4.018565	19.2	36.65
glucose1	43	79.32558	12.2255	60	[illegible]

Variable	Obs	Mean	Std. Dev.	Obs	Mean	Std. Dev.
death	4434	.3495715	.4768884	49	.3265306	.4738035
angina	4434	.1635092	.3698714	49	.1632653	.3734378
totchol1	4382	236.9843	44.6511	48	233.6667	42.73836
sysbp1	4434	132.9078	22.4216	49	140.3571	22.39745
diabp1	4434	83.08356	12.056	49	86.53061	12.65669
bmi1	4415	25.84616	4.101821	48	27.21167	4.018565
glucose1	4037	82.18578	24.39958	43	79.32558	12.2255

Sample

- Must be representative (random).
- The bigger the sample, the better our inference.

```
. set seed 72576466

. sample 10, count
(4424 observations deleted)

. summ death angina totchol1 sysbp1 diabp1 bmi1 glucose1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
death	10	.3	.4830459	0	1
angina	10	.1	.3162278	0	1
totchol1	10	238.9	72.50969	157	410
sysbp1	10	138.35	18.19043	106	173
diabp1	10	83.3	14.06374	48	98
bmi1	10	25.642	3.375404	19.2	30.91
glucose1	10	81.2	12.05358	60	97

Variable	Obs	Mean	Std.D.	Obs	Mean	Std.D.	Obs	Mean	Std.D.
death	4434	.350	.48	49	.326	.47	10	.3	.48
angina	4434	.164	.37	49	.163	.37	10	.1	.32
totchol1	4382	237.0	44.7	48	233.7	42.7	10	238.9	72.5
sysbp1	4434	133.0	22.4	49	140.4	22.4	10	138.4	18.2
diabp1	4434	83.1	12.1	49	86.5	12.7	10	83.3	14.1
bmi1	4415	25.8	4.10	48	27.2	4.02	10	25.6	3.38
glucose1	4037	82.2	24.4	43	79.3	12.2	10	81.2	12.1

Variable	Obs	Mean	Std.D.	Obs	Mean	Std.D.	Obs	Mean	Std.D.
death	4434	.350	.48	49	.326	.47	10	.3	.48
angina	4434	.164	.37	49	.163	.37	10	.1	.32
totchol1	4382	237.0	44.7	48	233.7	42.7	10	238.9	72.5
sysbp1	4434	133.0	22.4	49	140.4	22.4	10	138.4	18.2
diabp1	4434	83.1	12.1	49	86.5	12.7	10	83.3	14.1
bmi1	4415	25.8	4.10	48	27.2	4.02	10	25.6	3.38
glucose1	4037	82.2	24.4	43	79.3	12.2	10	81.2	12.1

REAL CLEAR POLITICS

General Election: McCain vs. Obama

Poll	Date	Sample	MoE	Obama (D)	McCain (R)	Spread
Final Results	--	--	--	**52.9**	**45.6**	**Obama +7.3**
RCP Average	**10/29 - 11/3**	--	--	**52.1**	**44.5**	**Obama +7.6**
Marist	11/03 - 11/03	804 LV	4.0	52	43	Obama +9
Battleground (Lake)*	11/02 - 11/03	800 LV	3.5	52	47	Obama +5
Battleground (Tarrance)*	11/02 - 11/03	800 LV	3.5	50	48	Obama +2
Rasmussen Reports	11/01 - 11/03	3000 LV	2.0	52	46	Obama +6
Reuters/C-SPAN/Zogby	11/01 - 11/03	1201 LV	2.9	54	43	Obama +11
IBD/TIPP	11/01 - 11/03	981 LV	3.2	52	44	Obama +8
FOX News	11/01 - 11/02	971 LV	3.0	50	43	Obama +7
NBC News/Wall St. Jrnl	11/01 - 11/02	1011 LV	3.1	51	43	Obama +8
Gallup	10/31 - 11/02	2472 LV	2.0	55	44	Obama +11
Diageo/Hotline	10/31 - 11/02	887 LV	3.3	50	45	Obama +5
CBS News	10/31 - 11/02	714 LV	--	51	42	Obama +9
ABC News/Wash Post	10/30 - 11/02	2470 LV	2.5	53	44	Obama +9
Ipsos/McClatchy	10/30 - 11/02	760 LV	3.6	53	46	Obama +7
CNN/Opinion Research	10/30 - 11/01	714 LV	3.5	53	46	Obama +7
Pew Research	10/29 - 11/01	2587 LV	2.0	52	46	Obama +6

REAL CLEAR POLITICS

General Election: McCain vs. Obama

Poll	Date	Sample	MoE	Obama (D)	McCain (R)	Spread
Final Results	--	--	--	**52.9**	**45.6**	**Obama +7.3**
RCP Average	**10/29 - 11/3**	--	--	**52.1**	**44.5**	**Obama +7.6**
Marist	11/03 - 11/03	804 LV	4.0	52	43	Obama +9
Battleground (Lake)*	11/02 - 11/03	800 LV	3.5	52	47	Obama +5
Battleground (Tarrance)*	11/02 - 11/03	800 LV	3.5	50	48	Obama +2
Rasmussen Reports	11/01 - 11/03	3000 LV	2.0	52	46	Obama +6
Reuters/C-SPAN/Zogby	11/01 - 11/03	1201 LV	2.9	54	43	Obama +11
IBD/TIPP	11/01 - 11/03	981 LV	3.2	52	44	Obama +8
FOX News	11/01 - 11/02	971 LV	3.0	50	43	Obama +7
NBC News/Wall St. Jrnl	11/01 - 11/02	1011 LV	3.1	51	43	Obama +8
Gallup	10/31 - 11/02	2472 LV	2.0	55	44	Obama +11
Diageo/Hotline	10/31 - 11/02	887 LV	3.3	50	45	Obama +5
CBS News	10/31 - 11/02	714 LV	--	51	42	Obama +9
ABC News/Wash Post	10/30 - 11/02	2470 LV	2.5	53	44	Obama +9
Ipsos/McClatchy	10/30 - 11/02	760 LV	3.6	53	46	Obama +7
CNN/Opinion Research	10/30 - 11/01	714 LV	3.5	53	46	Obama +7
Pew Research	10/29 - 11/01	2587 LV	2.0	52	46	Obama +6

REAL CLEAR POLITICS

General Election: McCain vs. Obama

Poll	Date	Sample	MoE	Obama (D)	McCain (R)	Spread
Final Results	--	--	--	**52.9**	**45.6**	**Obama +7.3**
RCP Average	**10/29 - 11/3**	--	--	**52.1**	**44.5**	**Obama +7.6**
Marist	11/03 - 11/03	804 LV	4.0	52	43	Obama +9
Battleground (Lake)*	11/02 - 11/03	800 LV	3.5	52	47	Obama +5
Battleground (Tarrance)*	11/02 - 11/03	800 LV	3.5	50	48	Obama +2
Rasmussen Reports	11/01 - 11/03	3000 LV	2.0	52	46	Obama +6
Reuters/C-SPAN/Zogby	11/01 - 11/03	1201 LV	2.9	54	43	Obama +11
IBD/TIPP	11/01 - 11/03	981 LV	3.2	52	44	Obama +8
FOX News	11/01 - 11/02	971 LV	3.0	50	43	Obama +7
NBC News/Wall St. Jrnl	11/01 - 11/02	1011 LV	3.1	51	43	Obama +8
Gallup	10/31 - 11/02	2472 LV	2.0	55	44	Obama +11
Diageo/Hotline	10/31 - 11/02	887 LV	3.3	50	45	Obama +5
CBS News	10/31 - 11/02	714 LV	--	51	42	Obama +9
ABC News/Wash Post	10/30 - 11/02	2470 LV	2.5	53	44	Obama +9
Ipsos/McClatchy	10/30 - 11/02	760 LV	3.6	53	46	Obama +7
CNN/Opinion Research	10/30 - 11/01	714 LV	3.5	53	46	Obama +7
Pew Research	10/29 - 11/01	2587 LV	2.0	52	46	Obama +6

REAL CLEAR POLITICS

General Election: McCain vs. Obama

Poll	Date	Sample	MoE	Obama (D)	McCain (R)	Spread
Final Results	--	--	--	52.9	45.6	Obama +7.3
RCP Average	10/29 - 11/3	--	--	52.1	44.5	Obama +7.6
Marist	11/03 - 11/03	804 LV	4.0	52	43	Obama +9
Battleground (Lake)*	11/02 - 11/03	800 LV	3.5	52	47	Obama +5
Battleground (Tarrance)*	11/02 - 11/03	800 LV	3.5	50	48	Obama +2
Rasmussen Reports	11/01 - 11/03	3000 LV	2.0	52	46	Obama +6
Reuters/C-SPAN/Zogby	11/01 - 11/03	1201 LV	2.9	54	43	Obama +11
IBD/TIPP	11/01 - 11/03	981 LV	3.2	52	44	Obama +8
FOX News	11/01 - 11/02	971 LV	3.0	50	43	Obama +7
NBC News/Wall St. Jrnl	11/01 - 11/02	1011 LV	3.1	51	43	Obama +8
Gallup	10/31 - 11/02	2472 LV	2.0	55	44	Obama +11
Diageo/Hotline	10/31 - 11/02	887 LV	3.3	50	45	Obama +5
CBS News	10/31 - 11/02	714 LV	--	51	42	Obama +9
ABC News/Wash Post	10/30 - 11/02	2470 LV	2.5	53	44	Obama +9
Ipsos/McClatchy	10/30 - 11/02	760 LV	3.6	53	46	Obama +7
CNN/Opinion Research	10/30 - 11/01	714 LV	3.5	53	46	Obama +7
Pew Research	10/29 - 11/01	2587 LV	2.0	52	46	Obama +6

REAL CLEAR POLITICS

General Election: McCain vs. Obama

Poll	Date	Sample	MoE	Obama (D)	McCain (R)	Spread
Final Results	--	--	--	52.9	45.6	Obama +7.3
RCP Average	10/29 - 11/3	--	--	52.1	44.5	Obama +7.6
Marist	11/03 - 11/03	804 LV	4.0	52	43	Obama +9
Battleground (Lake)*	11/02 - 11/03	800 LV	3.5	52	47	Obama +5
Battleground (Tarrance)*	11/02 - 11/03	800 LV	3.5	50	48	Obama +2
Rasmussen Reports	11/01 - 11/03	3000 LV	2.0	52	46	Obama +6
Reuters/C-SPAN/Zogby	11/01 - 11/03	1201 LV	2.9	54	43	Obama +11
IBD/TIPP	11/01 - 11/03	981 LV	3.2	52	44	Obama +8
FOX News	11/01 - 11/02	971 LV	3.0	50	43	Obama +7
NBC News/Wall St. Jrnl	11/01 - 11/02	1011 LV	3.1	51	43	Obama +8
Gallup	10/31 - 11/02	2472 LV	2.0	55	44	Obama +11
Diageo/Hotline	10/31 - 11/02	887 LV	3.3	50	45	Obama +5
CBS News	10/31 - 11/02	714 LV	--	51	42	Obama +9
ABC News/Wash Post	10/30 - 11/02	2470 LV	2.5	53	44	Obama +9
Ipsos/McClatchy	10/30 - 11/02	760 LV	3.6	53	46	Obama +7
CNN/Opinion Research	10/30 - 11/01	714 LV	3.5	53	46	Obama +7
Pew Research	10/29 - 11/01	2587 LV	2.0	52	46	Obama +6

REAL CLEAR POLITICS

General Election: McCain vs. Obama

Poll	Date	Sample	MoE	Obama (D)	McCain (R)	Spread
Final Results	--	--	--	52.9	45.6	Obama +7.3
RCP Average	10/29 - 11/3	--	--	52.1	44.5	Obama +7.6
Marist	11/03 - 11/03	804 LV	4.0	52	43	Obama +9
Battleground (Lake)*	11/02 - 11/03	800 LV	3.5	52	47	Obama +5
Battleground (Tarrance)*	11/02 - 11/03	800 LV	3.5	50	48	Obama +2
Rasmussen Reports	11/01 - 11/03	3000 LV	2.0	52	46	Obama +6
Reuters/C-SPAN/Zogby	11/01 - 11/03	1201 LV	2.9	54	43	Obama +11
IBD/TIPP	11/01 - 11/03	981 LV	3.2	52	44	Obama +8
FOX News	11/01 - 11/02	971 LV	3.0	50	43	Obama +7
NBC News/Wall St. Jrnl	11/01 - 11/02	1011 LV	3.1	51	43	Obama +8
Gallup	10/31 - 11/02	2472 LV	2.0	55	44	Obama +11
Diageo/Hotline	10/31 - 11/02	887 LV	3.3	50	45	Obama +5
CBS News	10/31 - 11/02	714 LV	--	51	42	Obama +9
ABC News/Wash Post	10/30 - 11/02	2470 LV	2.5	53	44	Obama +9
Ipsos/McClatchy	10/30 - 11/02	760 LV	3.6	53	46	Obama +7
CNN/Opinion Research	10/30 - 11/01	714 LV	3.5	53	46	Obama +7
Pew Research	10/29 - 11/01	2587 LV	2.0	52	46	Obama +6

©Fir0002/
Flagstaffotos

Inference

From part infer about the whole.

Inference

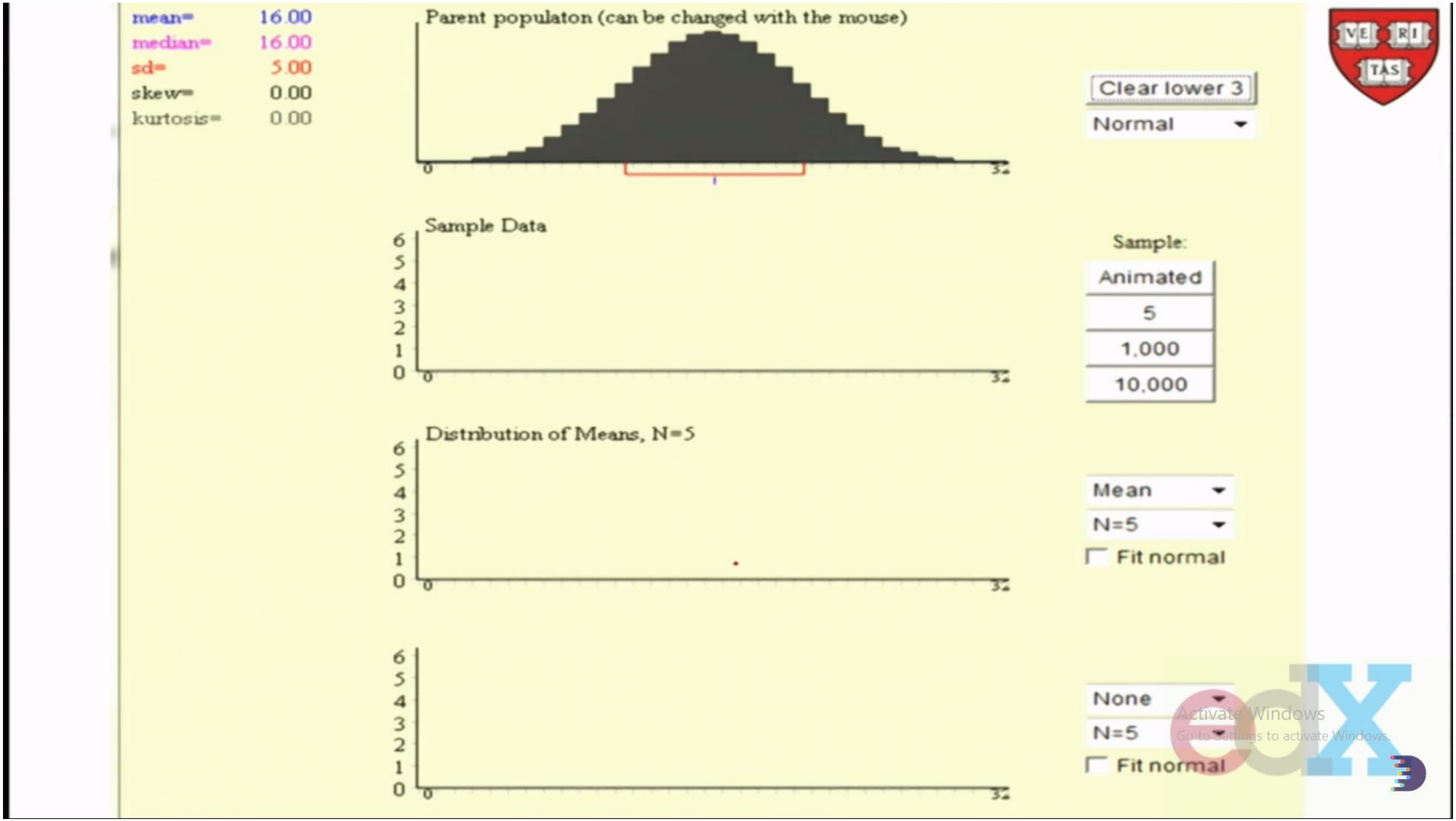
From part infer about the whole.

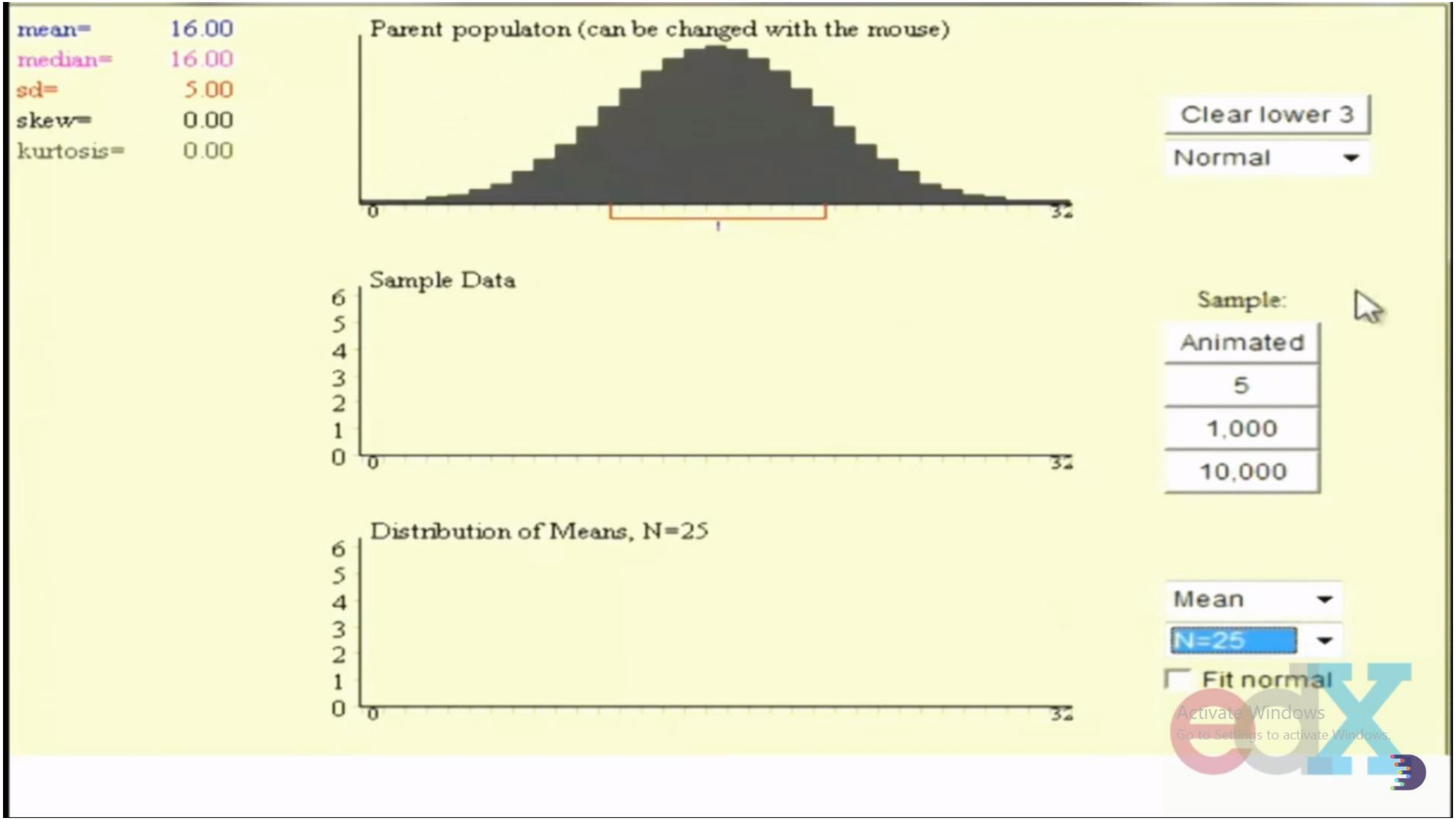
- Uncertainty
- Probability

Sampling Distribution

Rice University

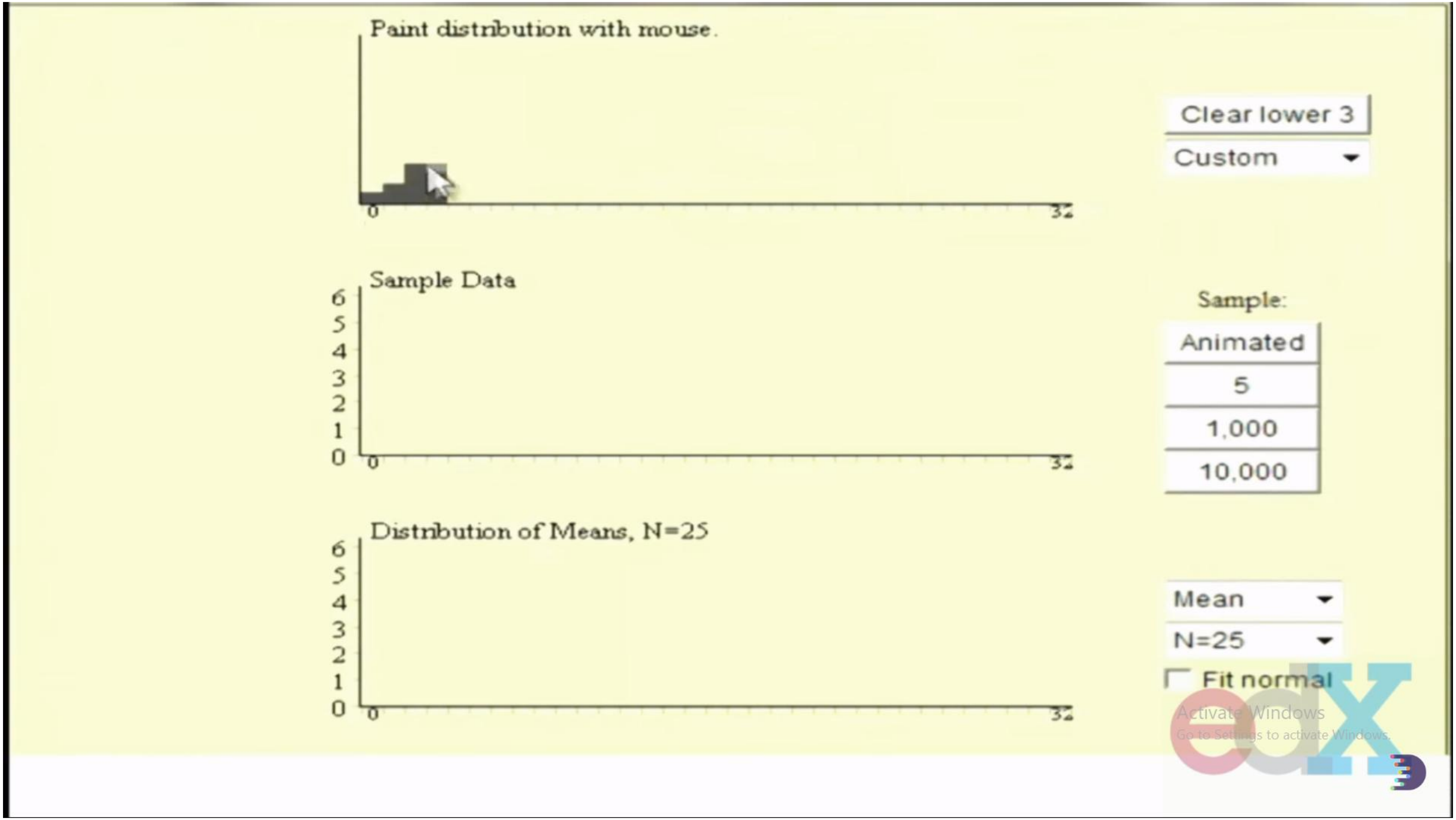
http://onlinestatbook.com/stat_sim/sampling_dist/index.html

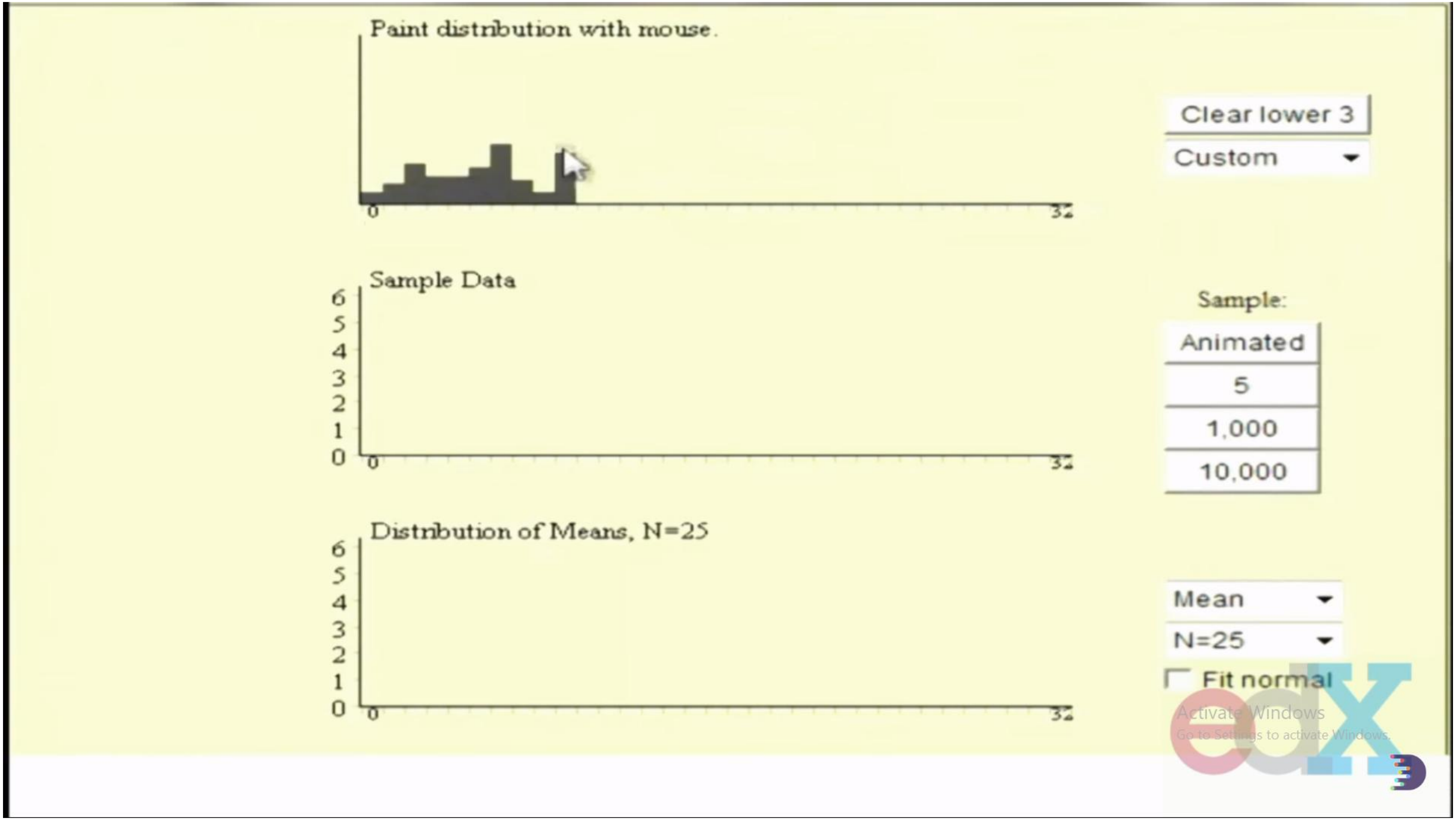

mean= 16.00
median= 16.00
sd= 5.00
skew= 0.00
kurtosis= 0.00
Parent populaton (can be changed with the mouse)
Clear lower 3
Normal
Sample Data
Sample:
Animated
5
1,000
10,000
Distribution of Means, N=5
Mean
N=5
Fit normal
None
N=5
Fit normal



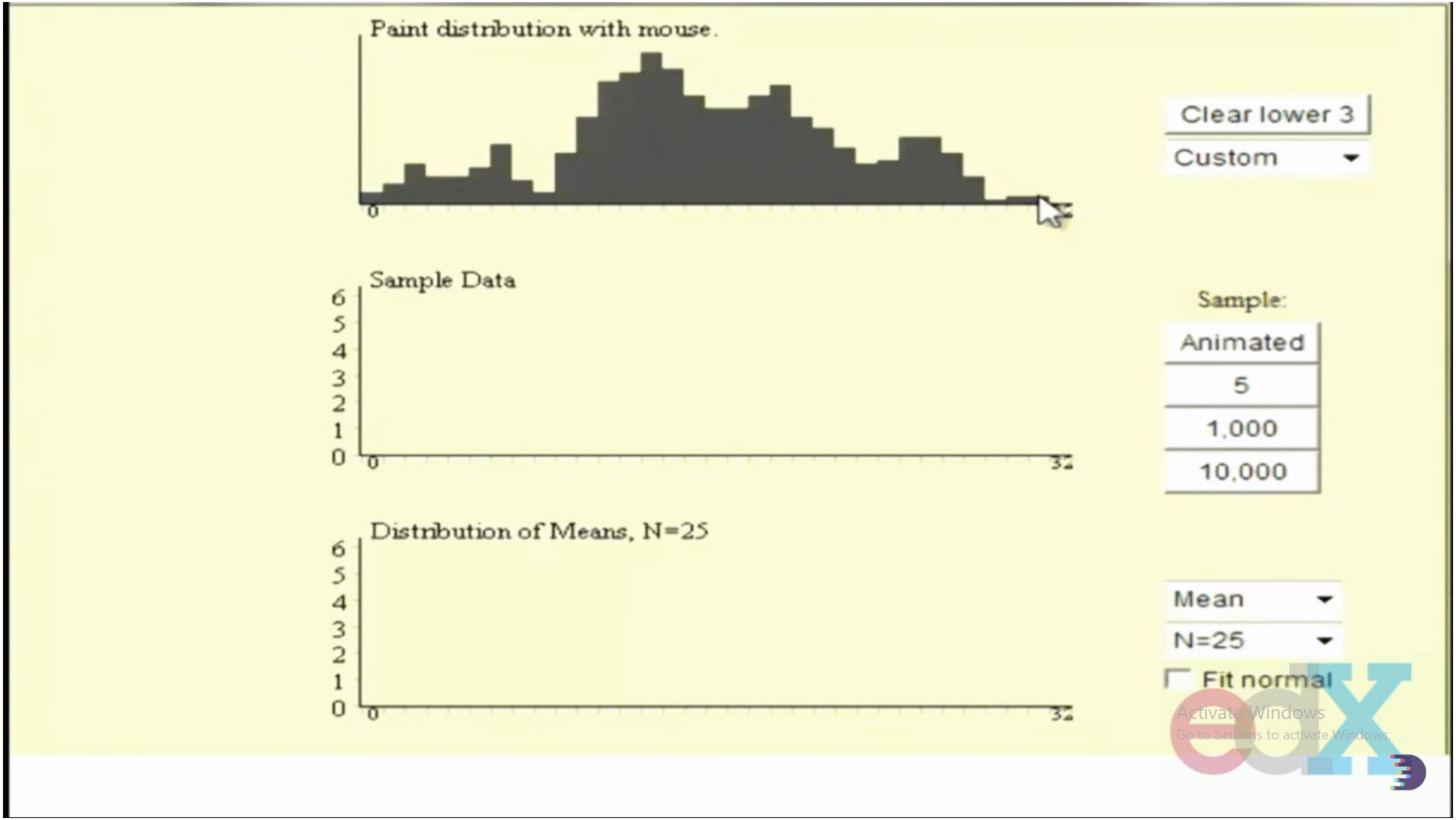
mean= 16.00
median= 16.00
sd= 5.00
skew= 0.00
kurtosis= 0.00
Parent populaton (can be changed with the mouse)
0
32
Clear lower 3
Normal
Sample Data
6
5
4
3
2
1
0
Sample:
Animated
5
1,000
10,000
Distribution of Means, N=25
Mean
N=25
Fit normal
Activate Windows
Go to Settings to activate Windows.

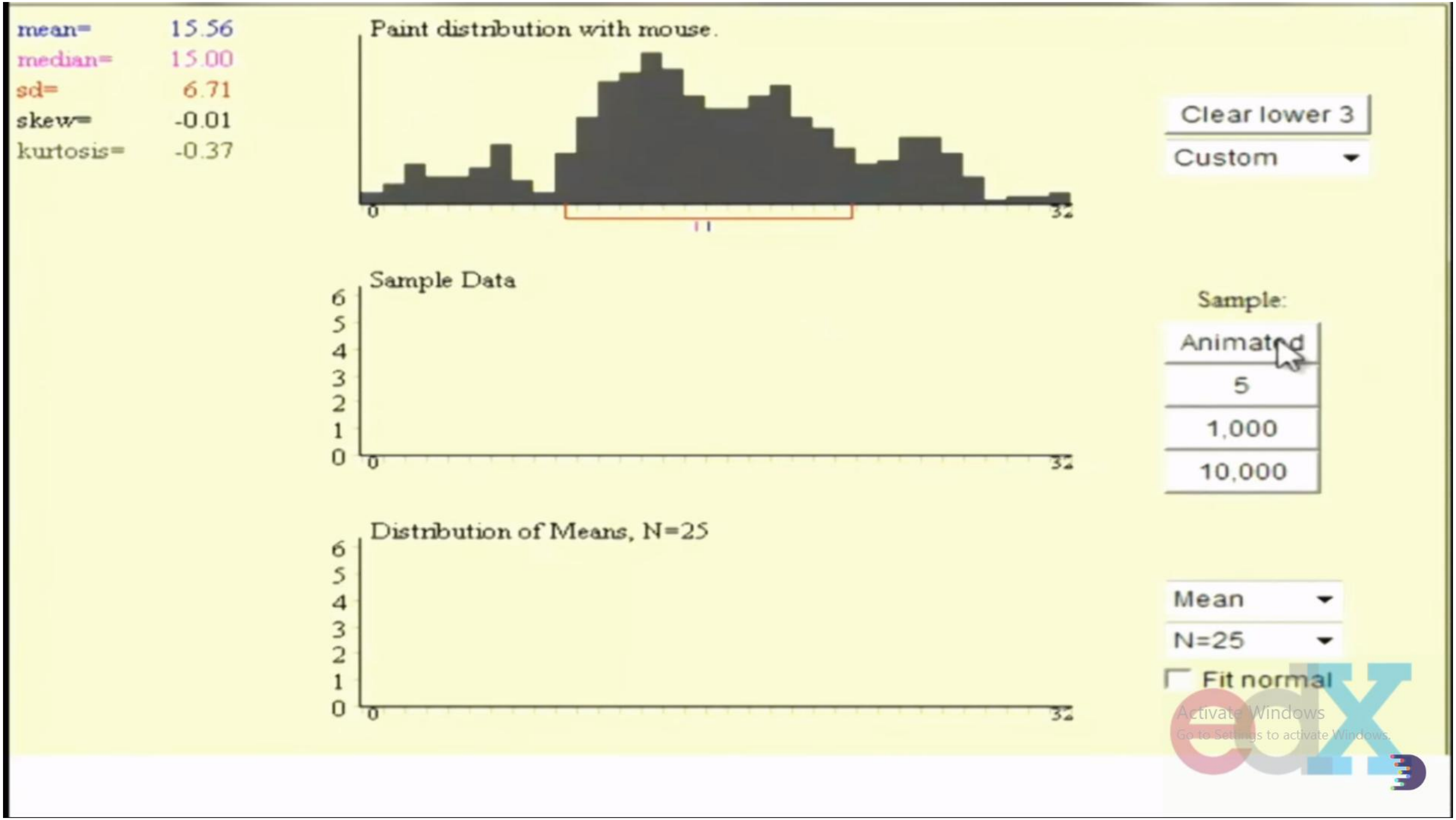

Week 5 : SAMPLING DISTRIBUTIONS
Clear lower 3
Custom
0
32
Sample Data
6
5
4
3
2
1
0
Sample:
Animated
5
1,000
10,000
Distribution of Means, N=25
Mean
N=25
Fit normal
2:52 / 17:12

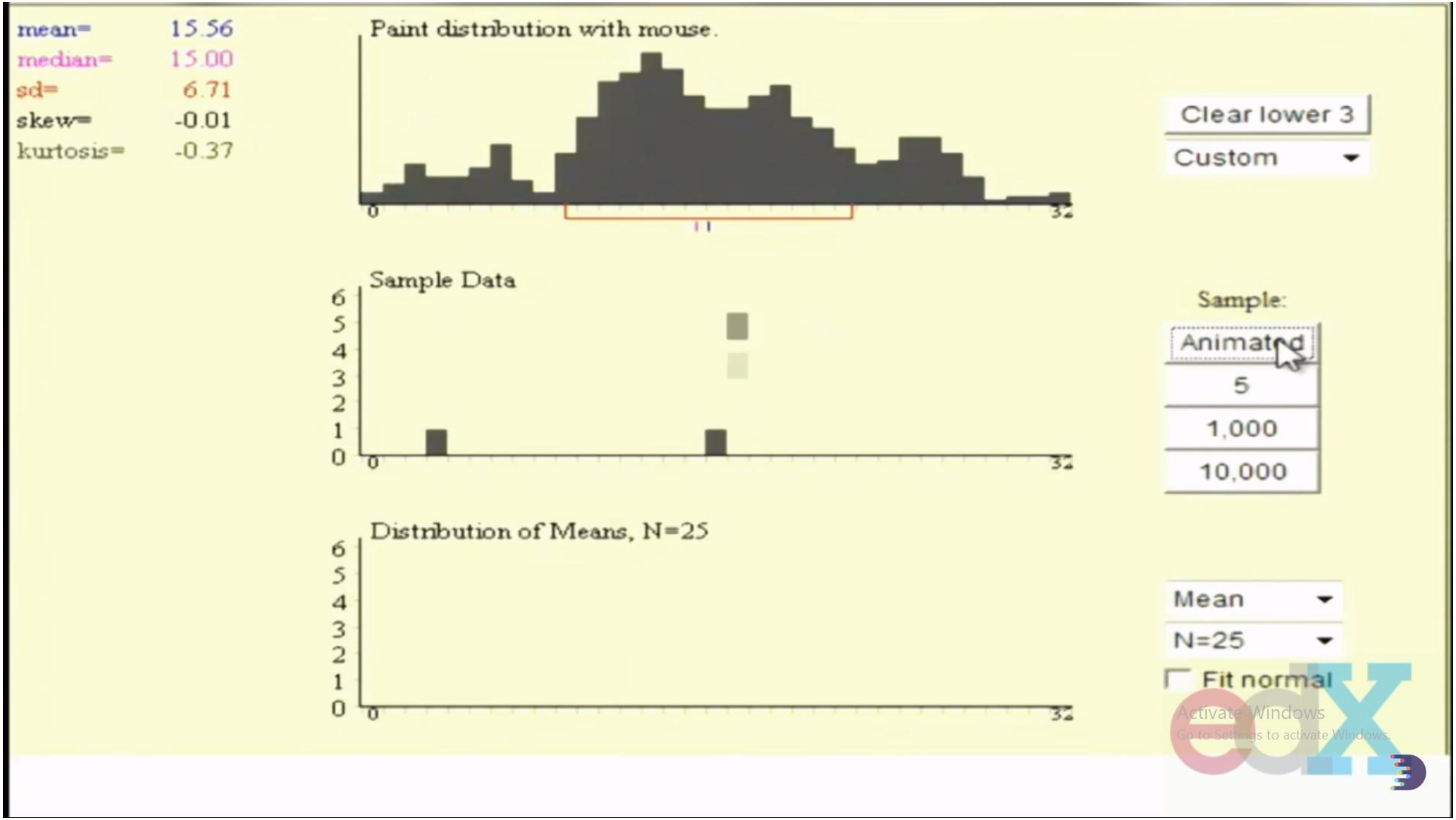

Paint distribution with mouse.
0
32
Sample Data
6
5
4
3
2
1
0
0
32
Distribution of Means, N=25
6
5
4
3
2
1
0
0
32
Clear lower 3
Custom
Sample:
Animated
5
1,000
10,000
Mean
N=25
Fit normal



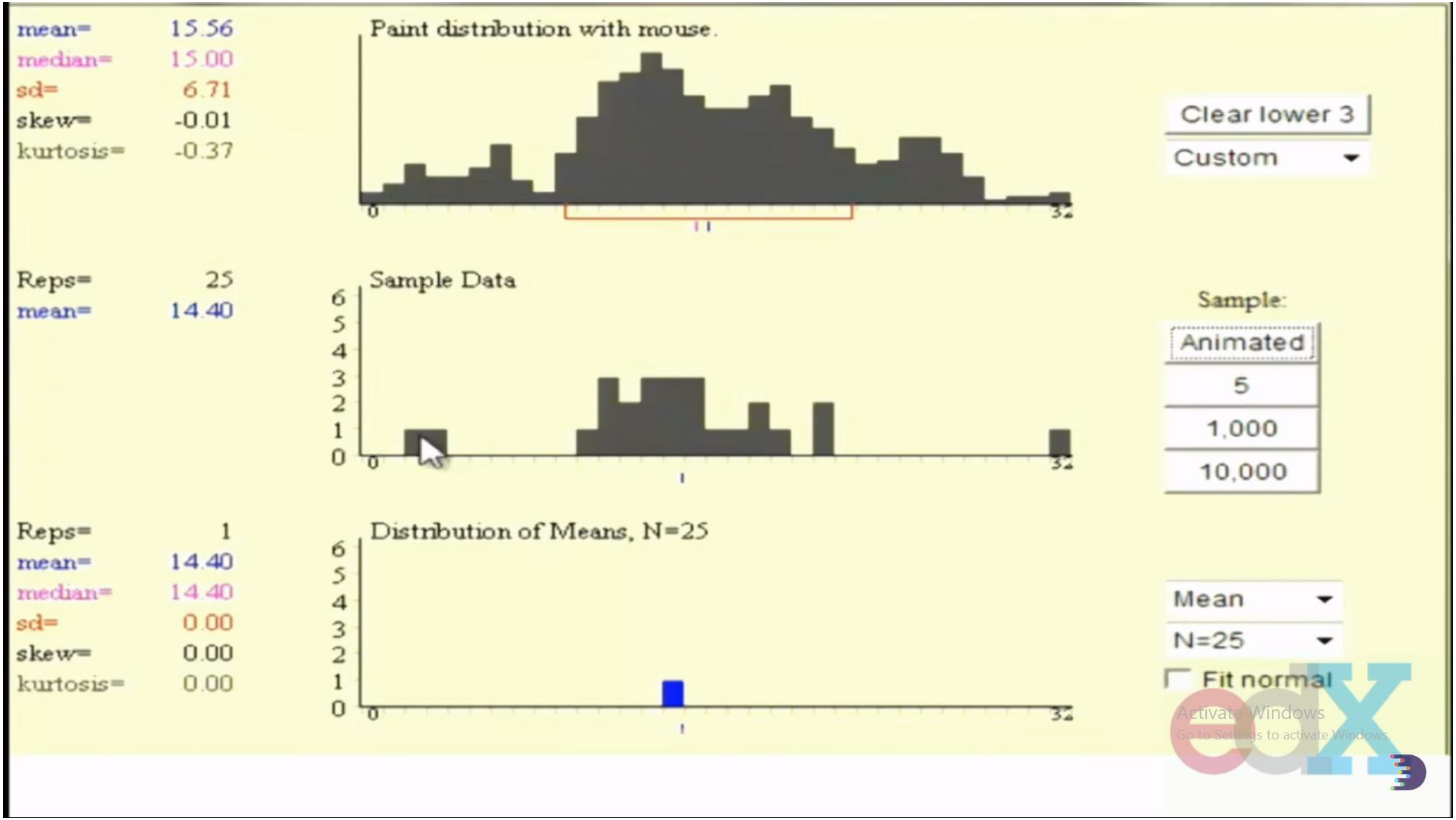
Paint distribution with mouse.
0
32
Clear lower 3
Custom
Sample Data
6
5
4
3
2
1
0
0
32
Sample:
Animated
5
1,000
10,000
Distribution of Means, N=25
6
5
4
3
2
1
0
0
32
Mean
N=25
Fit normal
Activate Windows
Go to Settings to activate Windows.


Paint distribution with mouse.
0
Clear lower 3
Custom
Sample Data
6
5
4
3
2
1
0
0
32
Sample:
Animated
5
1,000
10,000
Distribution of Means, N=25
6
5
4
3
2
1
0
0
32
Mean
N=25
Fit normal
Activate Windows
Go to Settings to activate Windows.

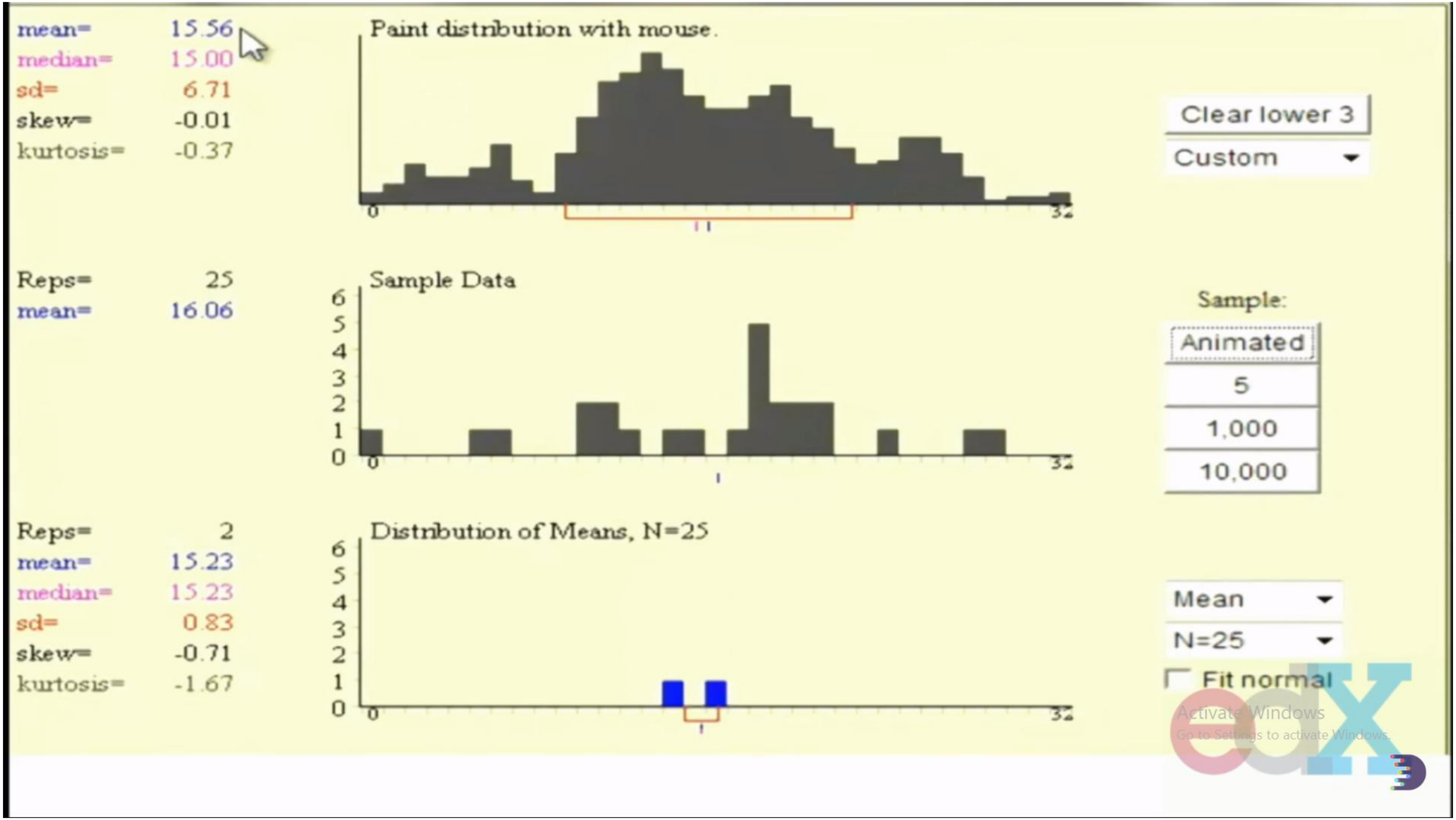

mean= 15.56
median= 15.00
sd= 6.71
skew= -0.01
kurtosis= -0.37
Paint distribution with mouse.
0
32
Clear lower 3
Custom
Sample Data
6
5
4
3
2
1
0
0
32
Sample:
Animated
5
1,000
10,000
Distribution of Means, N=25
6
5
4
3
2
1
0
0
32
Mean
N=25
Fit normal
Activate Windows
Go to Settings to activate Windows.



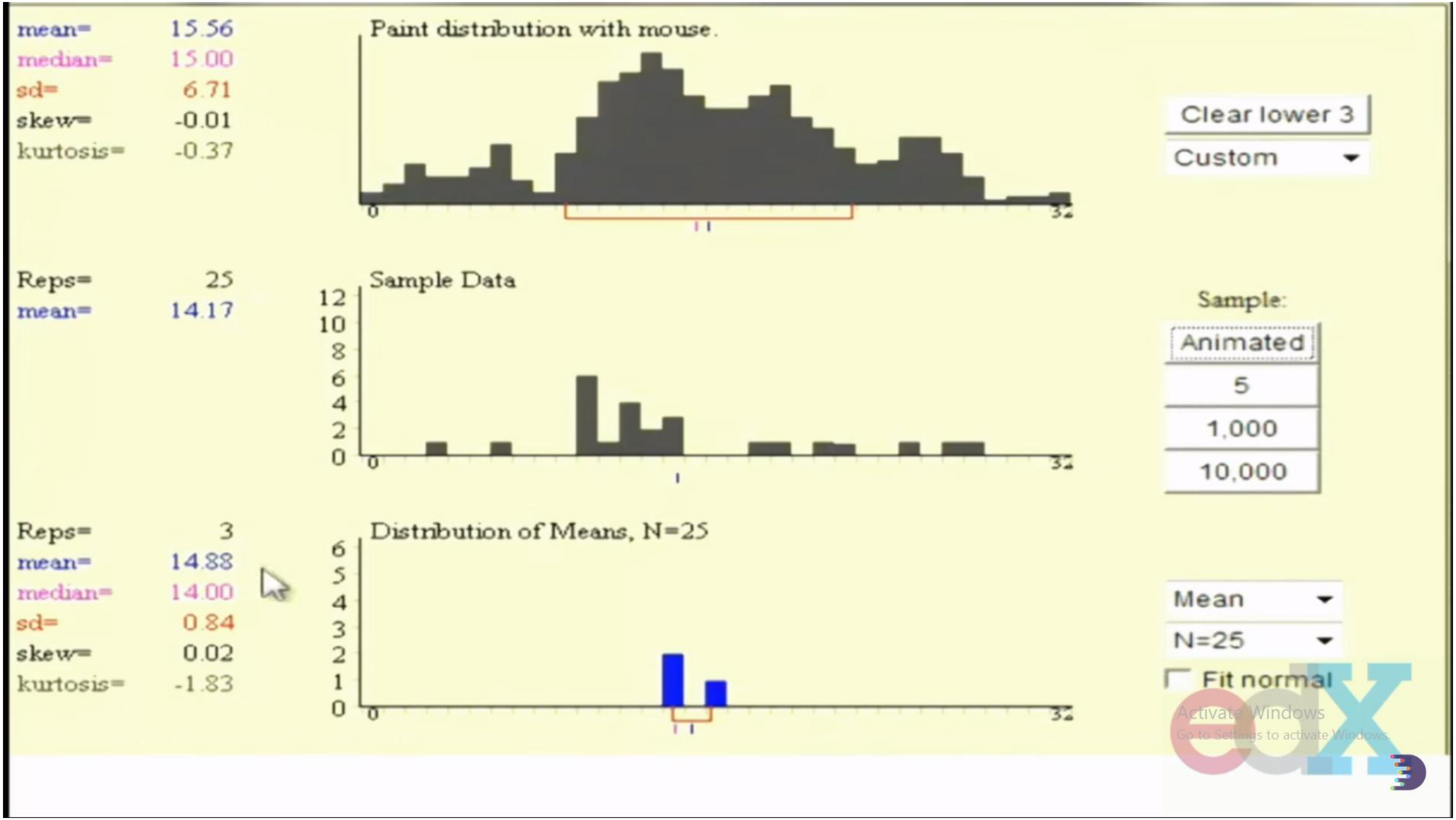
mean= 15.56
median= 15.00
sd= 6.71
skew= -0.01
kurtosis= -0.37
Paint distribution with mouse.
0
32
Clear lower 3
Custom
Sample Data
Sample:
Animated
5
1,000
10,000
Distribution of Means, N=25
Mean
N=25
Fit normal
Activate Windows
Go to Settings to activate Windows.



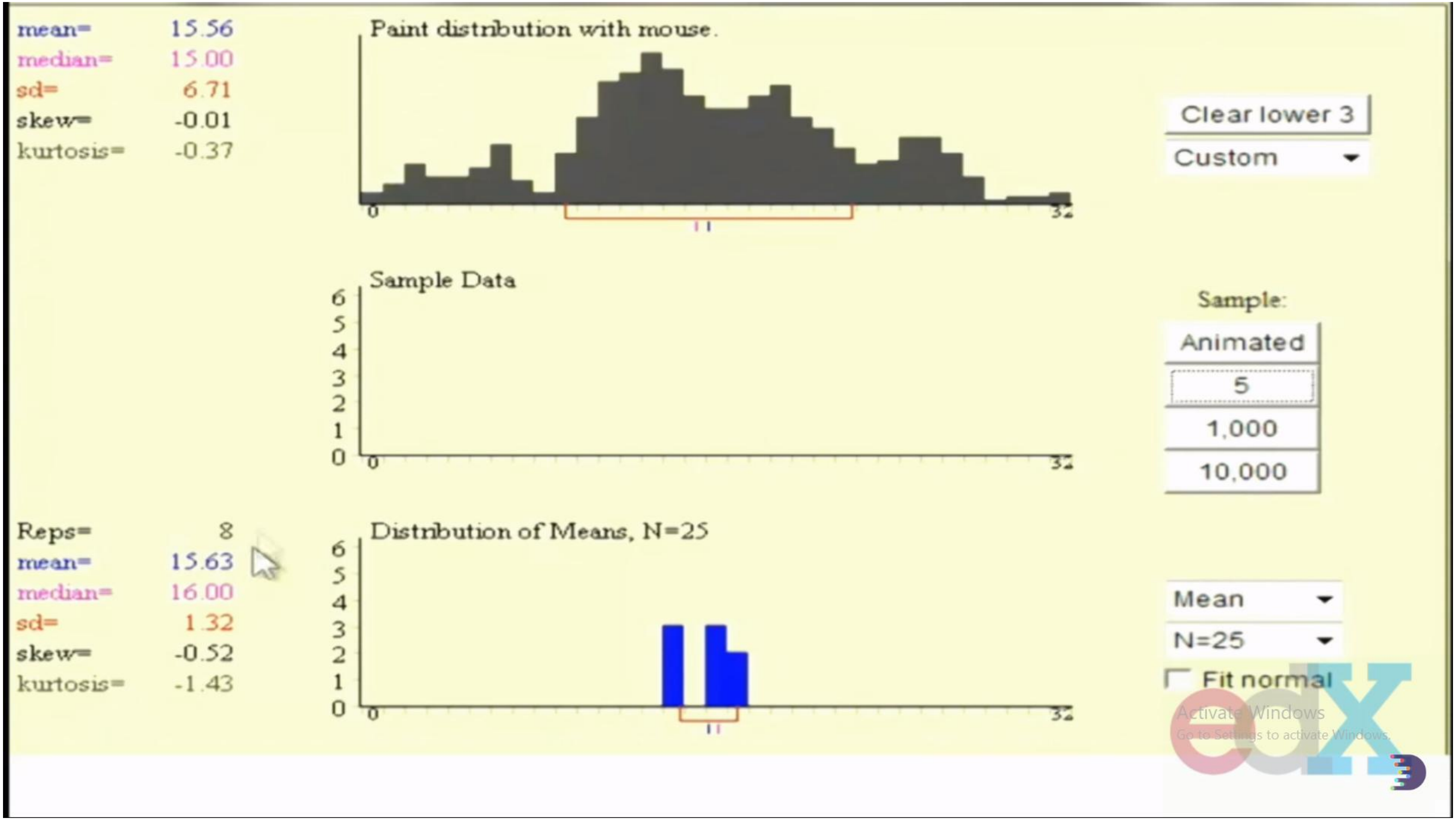
mean= 15.56
median= 15.00
sd= 6.71
skew= -0.01
kurtosis= -0.37
Paint distribution with mouse.
Clear lower 3
Custom
Reps= 25
mean= 14.40
Sample Data
Sample:
Animated
5
1,000
10,000
Reps= 1
mean= 14.40
median= 14.40
sd= 0.00
skew= 0.00
kurtosis= 0.00
Distribution of Means, N=25
Mean
N=25
Fit normal

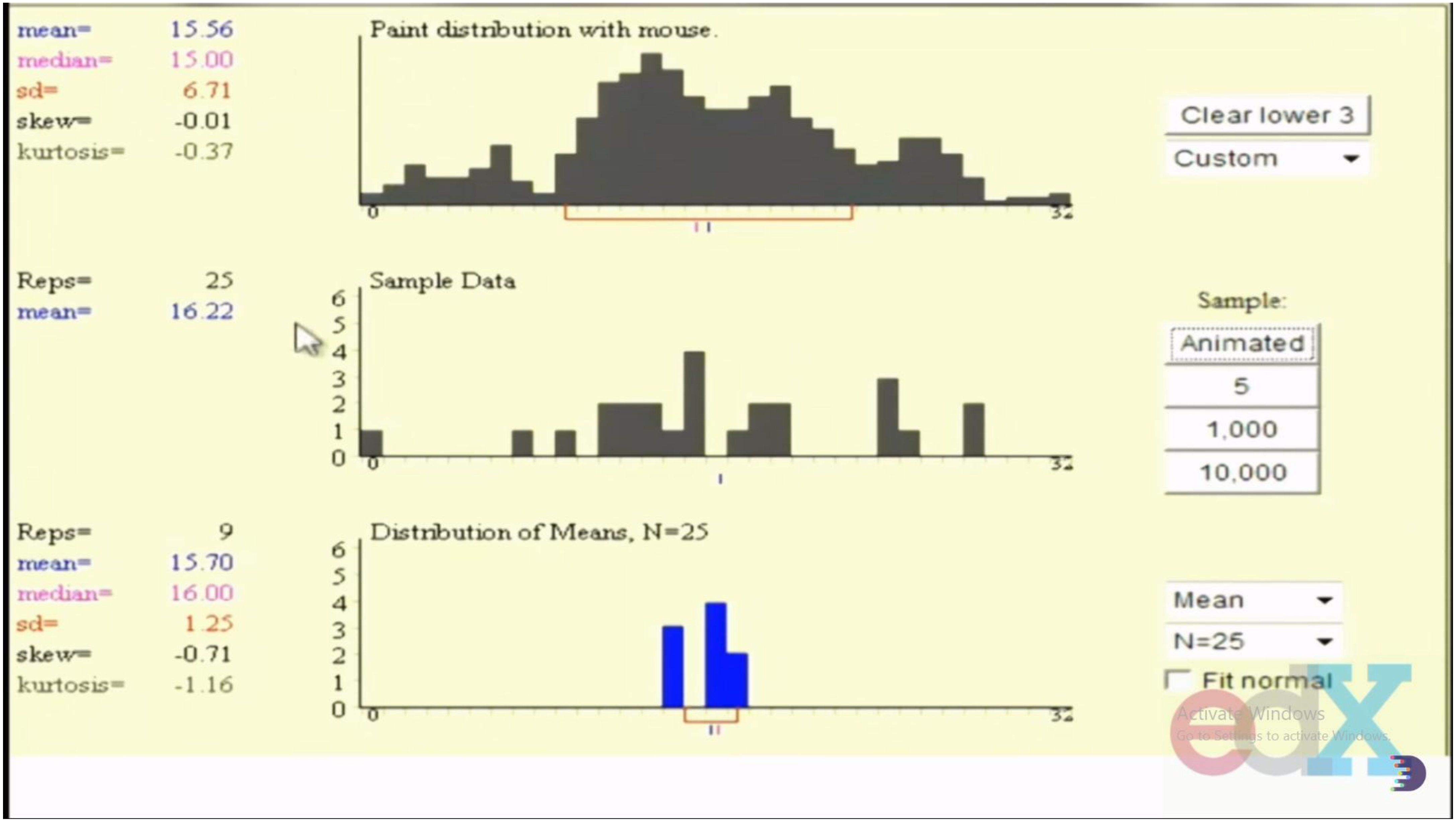


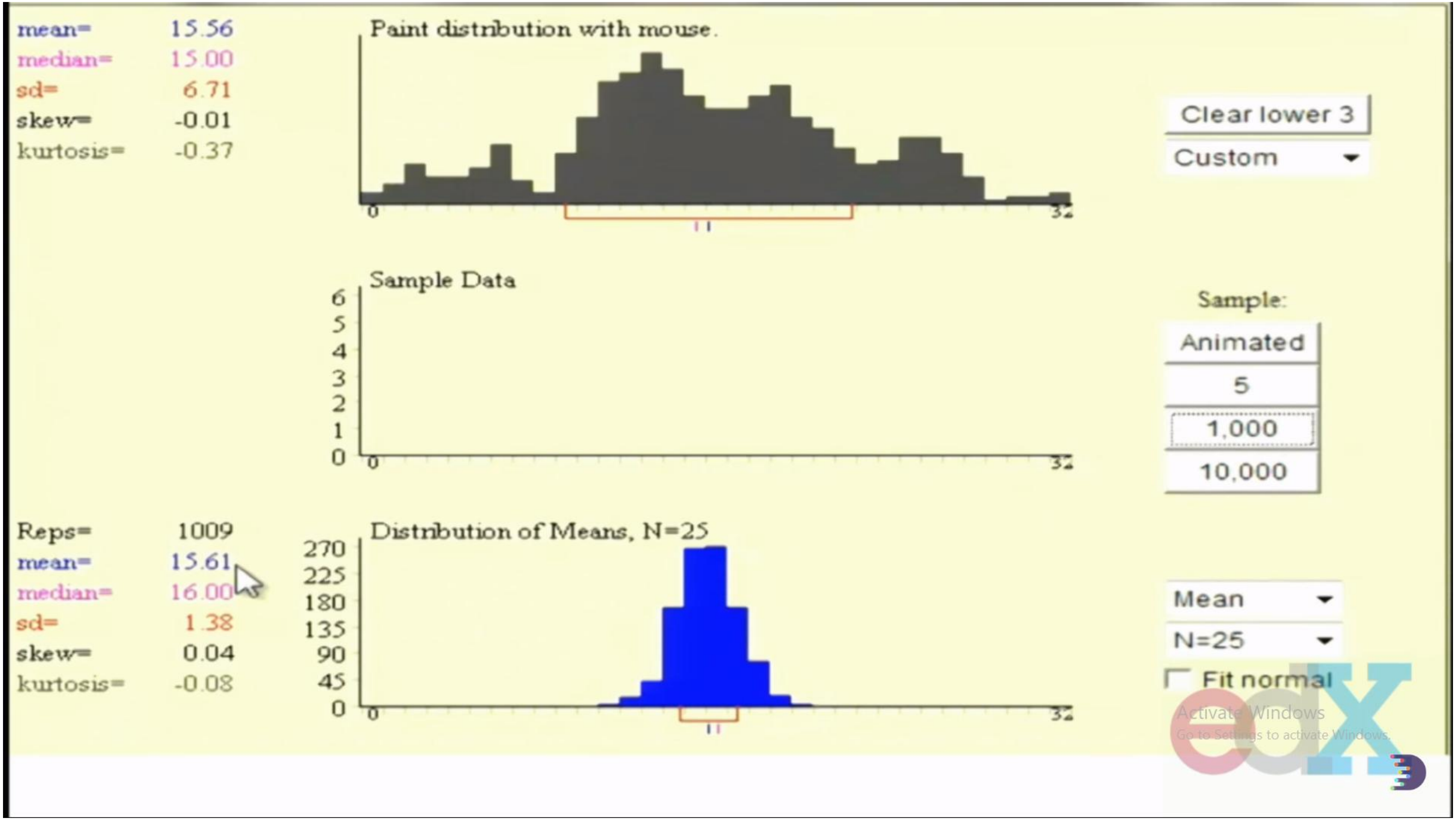
mean= 15.56
median= 15.00
sd= 6.71
skew= -0.01
kurtosis= -0.37
Paint distribution with mouse.
Clear lower 3
Custom
Reps= 25
mean= 16.06
Sample Data
Sample:
Animated
5
1,000
10,000
Reps= 2
mean= 15.23
median= 15.23
sd= 0.83
skew= -0.71
kurtosis= -1.67
Distribution of Means, N=25
Mean
N=25
Fit normal



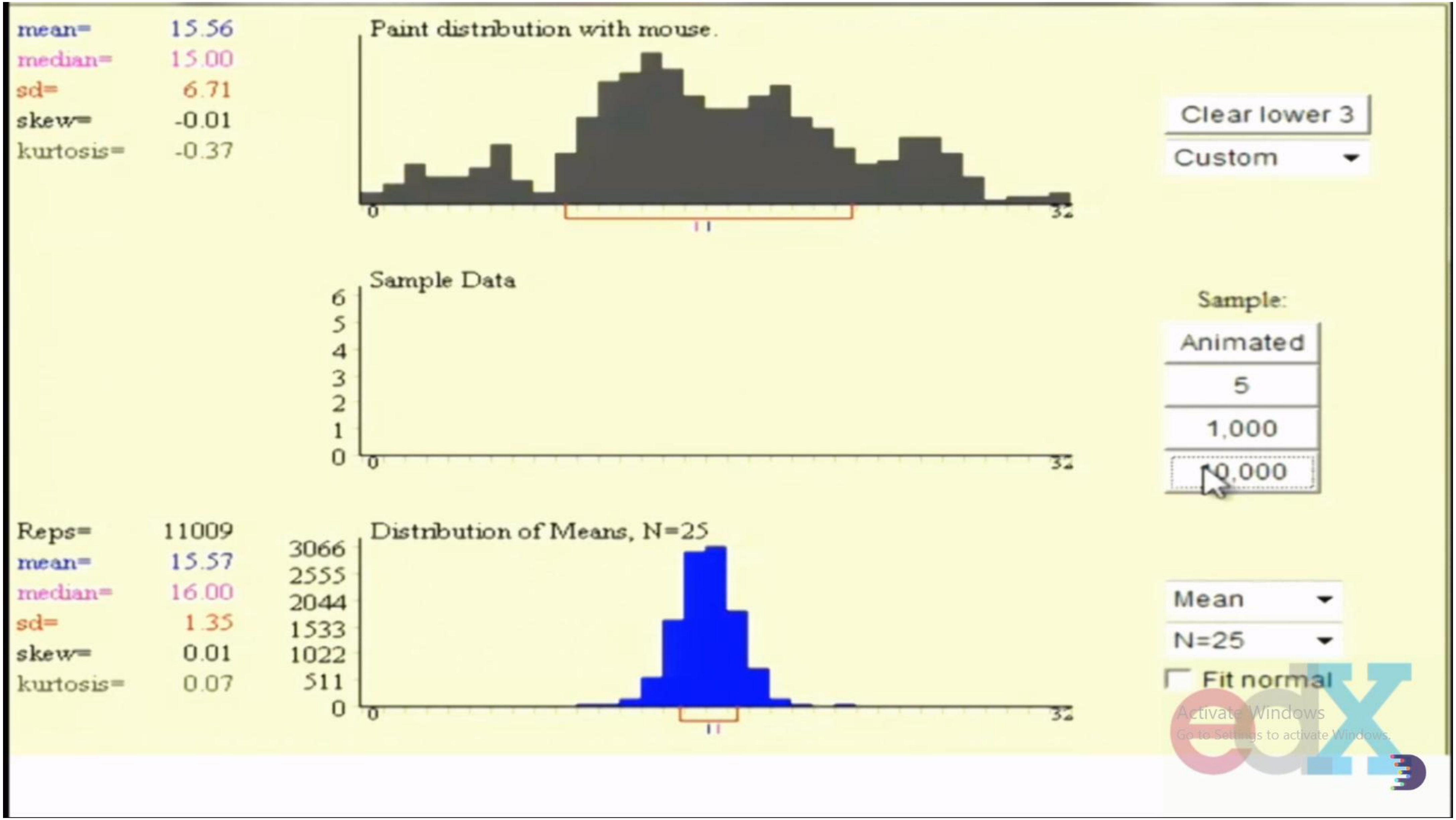
mean= 15.56
median= 15.00
sd= 6.71
skew= -0.01
kurtosis= -0.37
Paint distribution with mouse.
Clear lower 3
Custom
Reps= 25
mean= 14.17
Sample Data
Sample:
Animated
5
1,000
10,000
Reps= 3
mean= 14.88
median= 14.00
sd= 0.84
skew= 0.02
kurtosis= -1.83
Distribution of Means, N=25
Mean
N=25
Fit normal

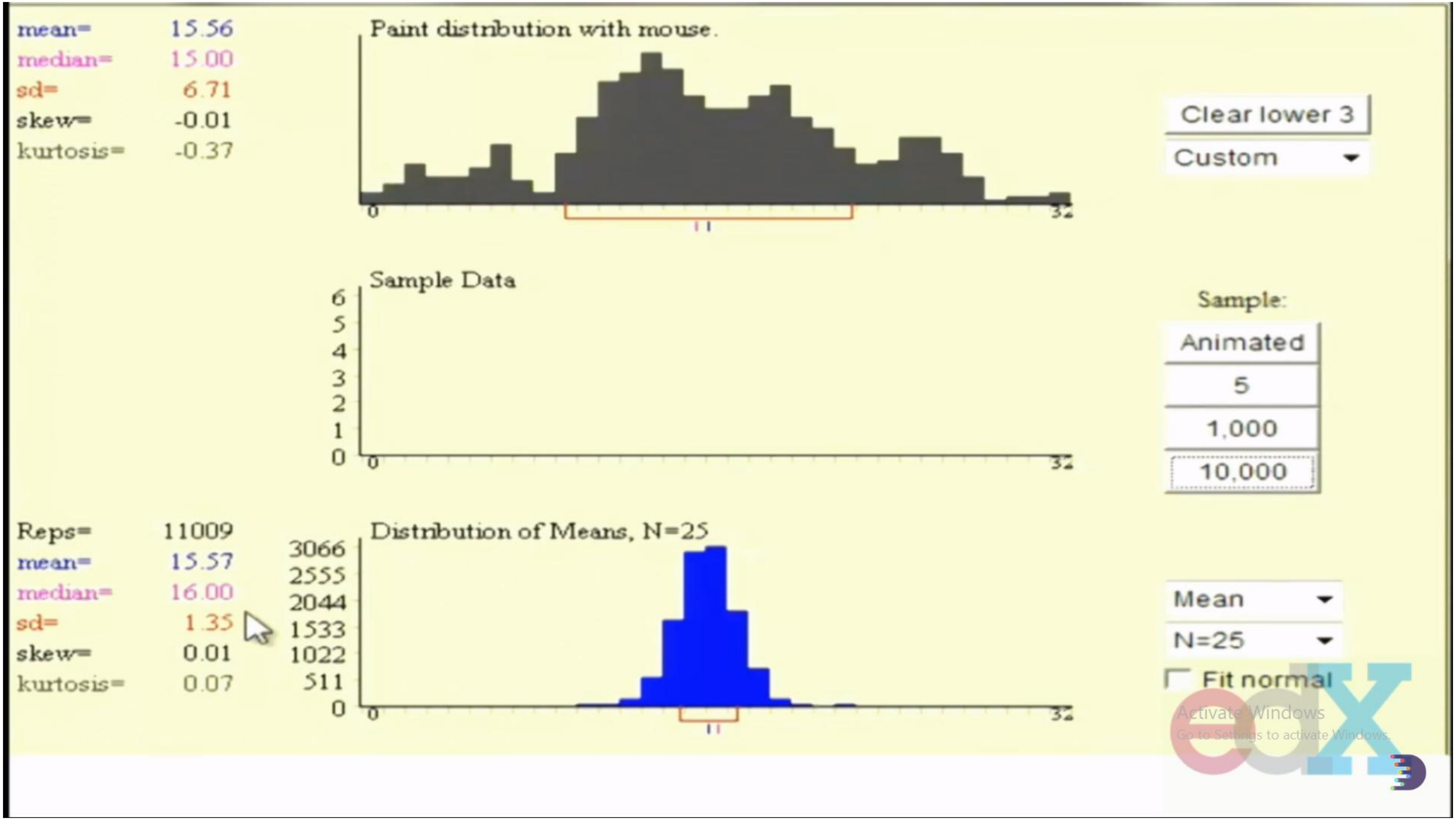

mean= 15.56
median= 15.00
sd= 6.71
skew= -0.01
kurtosis= -0.37
Paint distribution with mouse.
0
32
Clear lower 3
Custom
Sample Data
6
5
4
3
2
1
0
0
32
Sample:
Animated
5
1,000
10,000
Reps= 8
mean= 15.63
median= 16.00
sd= 1.32
skew= -0.52
kurtosis= -1.43
Distribution of Means, N=25
6
5
4
3
2
1
0
0
32
Mean
N=25
Fit normal
Activate Windows
Go to Settings to activate Windows.

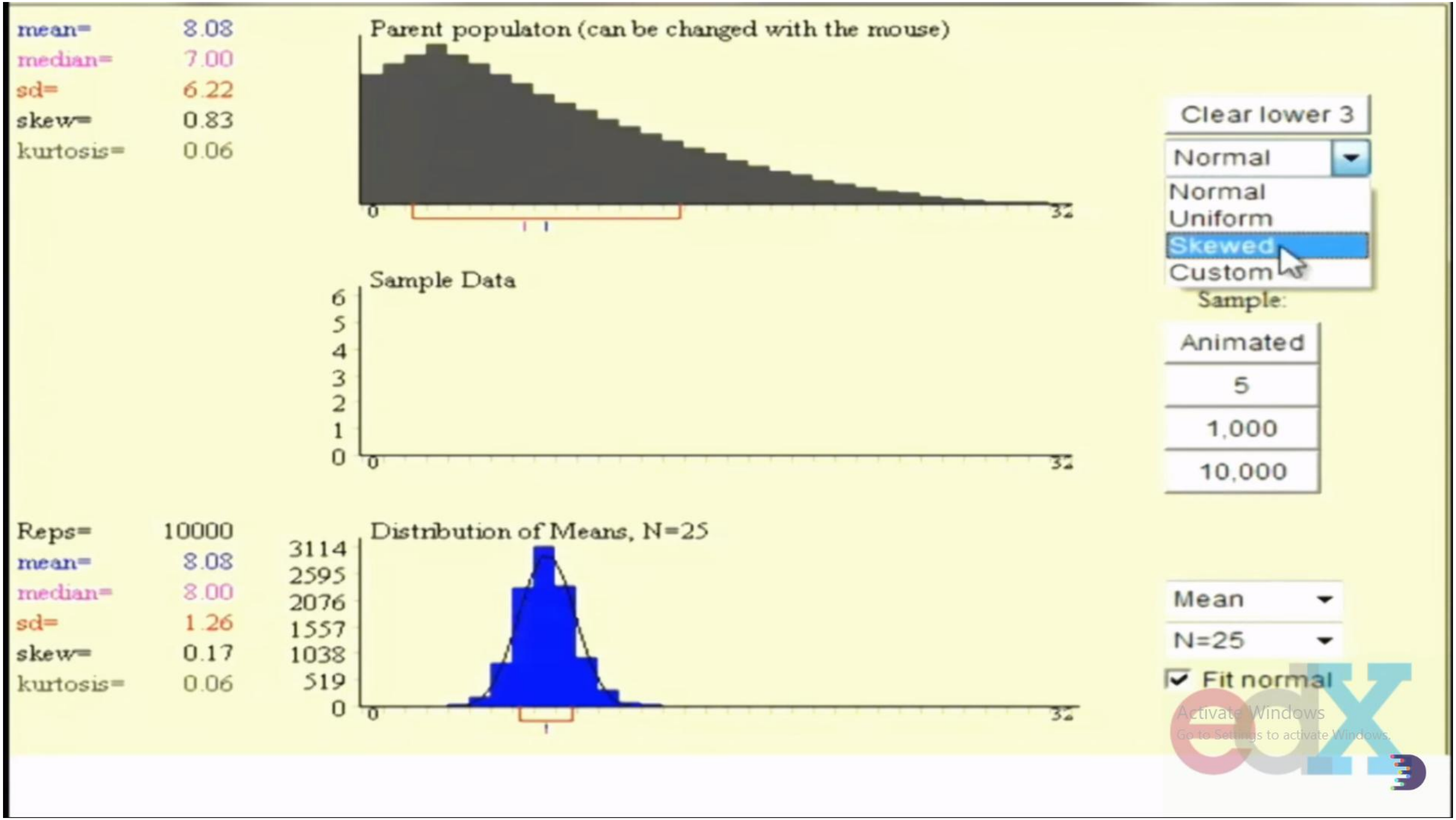

mean= 15.56
median= 15.00
sd= 6.71
skew= -0.01
kurtosis= -0.37
Paint distribution with mouse.
Clear lower 3
Custom
Reps= 25
mean= 16.22
Sample Data
Sample:
Animated
5
1,000
10,000
Reps= 9
mean= 15.70
median= 16.00
sd= 1.25
skew= -0.71
kurtosis= -1.16
Distribution of Means, N=25
Mean
N=25
Fit normal



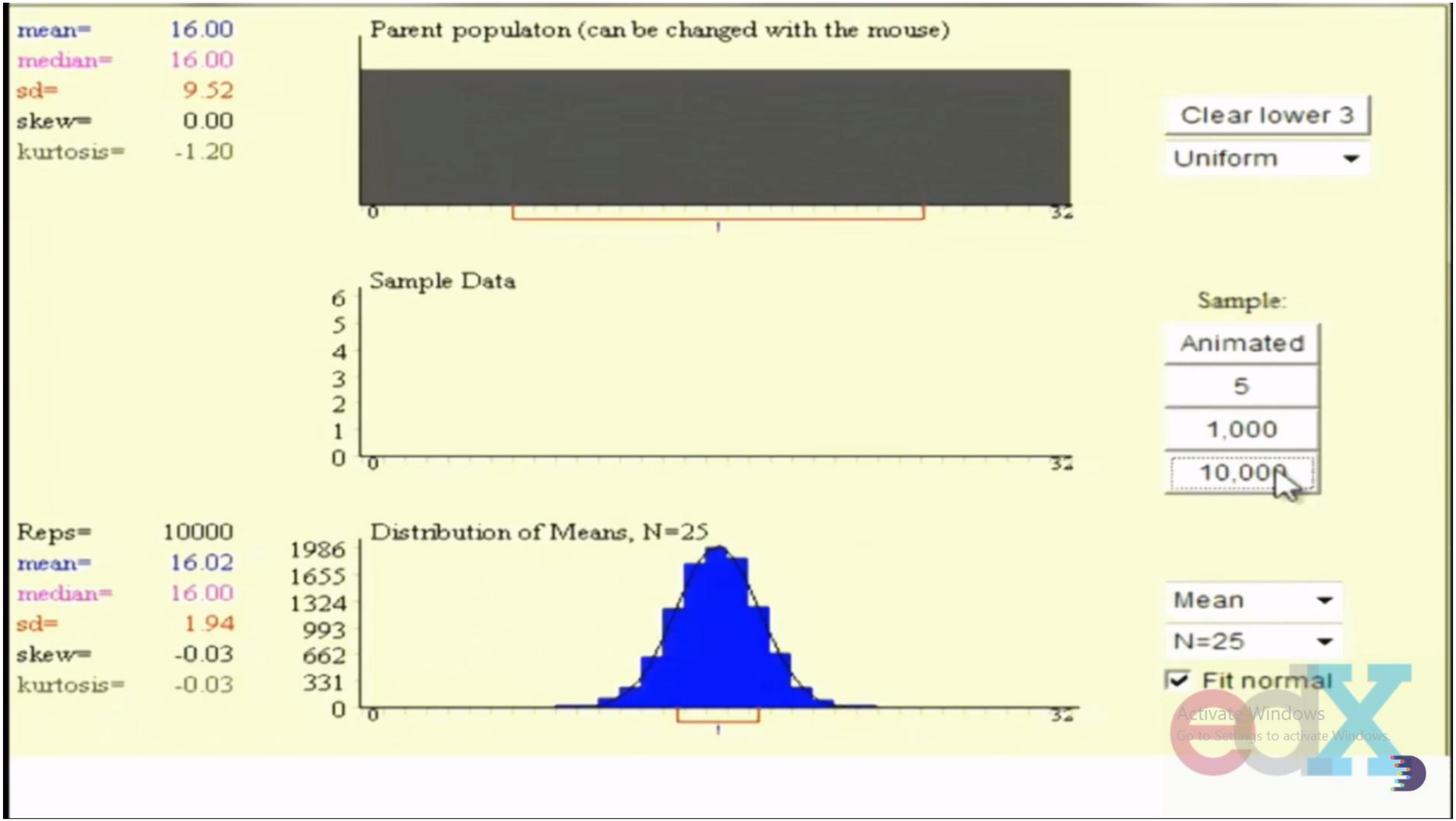
mean= 15.56
median= 15.00
sd= 6.71
skew= -0.01
kurtosis= -0.37
Paint distribution with mouse.
0
32
Clear lower 3
Custom
Sample Data
6
5
4
3
2
1
0
Sample:
Animated
5
1,000
10,000
Reps= 1009
mean= 15.61
median= 16.00
sd= 1.38
skew= 0.04
kurtosis= -0.08
Distribution of Means, N=25
270
225
180
135
90
45
0
Mean
N=25
Fit normal
Activate Windows
Go to Settings to activate Windows.


mean= 15.56
median= 15.00
sd= 6.71
skew= -0.01
kurtosis= -0.37
Paint distribution with mouse.
0
32
Clear lower 3
Custom
Sample Data
6
5
4
3
2
1
0
0
32
Sample:
Animated
5
1,000
10,000
Reps= 11009
mean= 15.57
median= 16.00
sd= 1.35
skew= 0.01
kurtosis= 0.07
Distribution of Means, N=25
3066
2555
2044
1533
1022
511
0
0
32
Mean
N=25
Fit normal
Activate Windows
Go to Settings to activate Windows.


mean= 15.56
median= 15.00
sd= 6.71
skew= -0.01
kurtosis= -0.37
Paint distribution with mouse.
0
32
Clear lower 3
Custom
Sample Data
6
5
4
3
2
1
0
0
32
Sample:
Animated
5
1,000
10,000
Reps= 11009
mean= 15.57
median= 16.00
sd= 1.35
skew= 0.01
kurtosis= 0.07
Distribution of Means, N=25
3066
2555
2044
1533
1022
511
0
0
32
Mean
N=25
Fit normal
Activate Windows
Go to Settings to activate Windows.



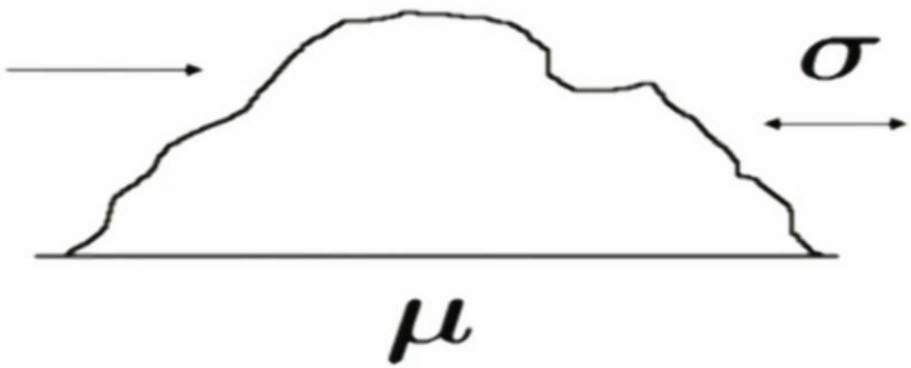
mean= 8.08
median= 7.00
sd= 6.22
skew= 0.83
kurtosis= 0.06
Parent populaton (can be changed with the mouse)
0
32
Clear lower 3
Normal
Normal
Uniform
Skewed
Custom
Sample:
Sample Data
6
5
4
3
2
1
0
Animated
5
1,000
10,000
Reps= 10000
mean= 8.08
median= 8.00
sd= 1.26
skew= 0.17
kurtosis= 0.06
Distribution of Means, N=25
3114
2595
2076
1557
1038
519
0
Mean
N=25
Fit normal
Activate Windows
Go to Settings to activate Windows.



mean= 16.00
median= 16.00
sd= 9.52
skew= 0.00
kurtosis= -1.20
Parent populaton (can be changed with the mouse)
0
32
Clear lower 3
Uniform
Sample Data
6
5
4
3
2
1
0
Sample:
Animated
5
1,000
10,000
Reps= 10000
mean= 16.02
median= 16.00
sd= 1.94
skew= -0.03
kurtosis= -0.03
Distribution of Means, N=25
1986
1655
1324
993
662
331
0
Mean
N=25
Fit normal
Activate Windows
Go to Settings to activate Windows.

Sampling Distribution of the Mean

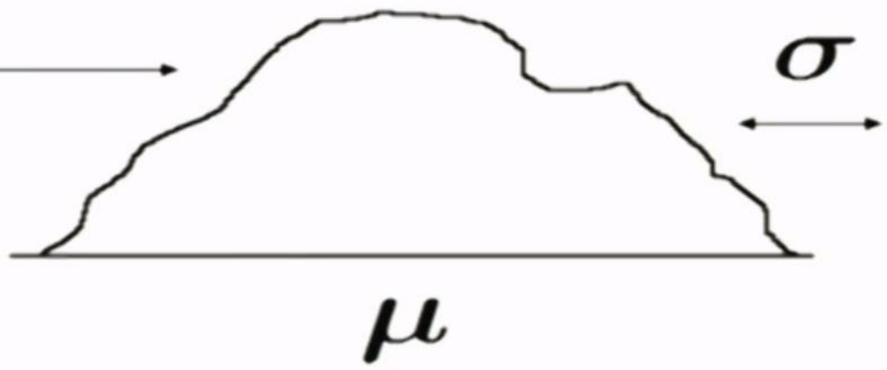
Population



$$\mu = ?$$

Sampling Distribution of the Mean

Population

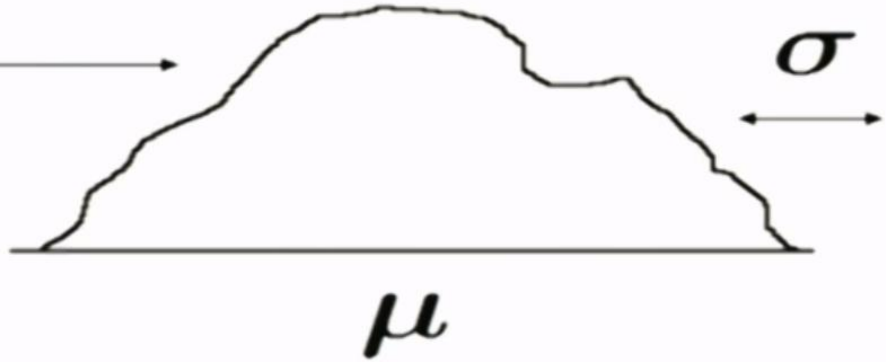


$\mu = ?$

Sample 1: $x_{11}, x_{12}, \ldots, x_{1n}$

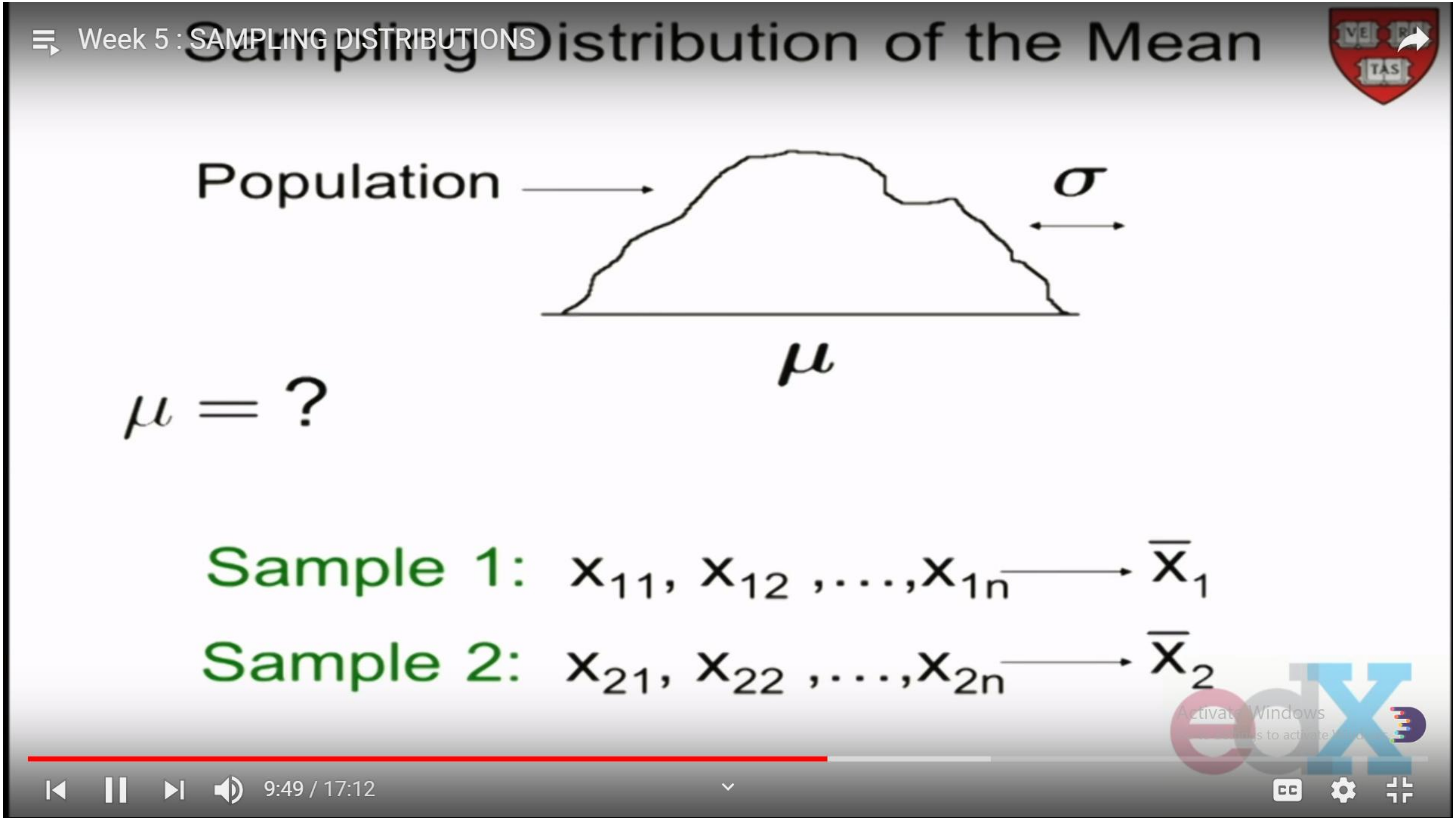
Sampling Distribution of the Mean

Population



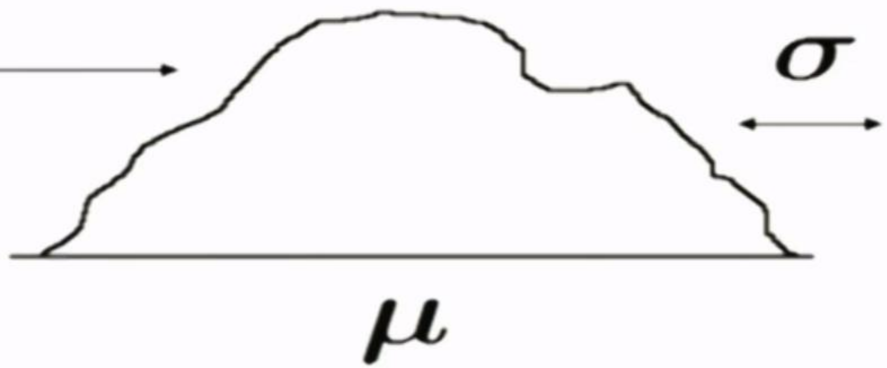
$\mu = ?$

Sample 1: $x_{11}, x_{12}, \ldots, x_{1n} \longrightarrow \bar{x}_1$


Week 5 : SAMPLING DISTRIBUTIONS
Distribution of the Mean
Population
σ
μ
$\mu = ?$
Sample 1: $x_{11}, x_{12}, \ldots, x_{1n} \rightarrow \bar{x}_1$
Sample 2: $x_{21}, x_{22}, \ldots, x_{2n} \rightarrow \bar{x}_2$
9:49 / 17:12

Sampling Distribution of the Mean

Population $\longrightarrow$



$$\mu = ?$$

Sample 1: $x_{11}, x_{12}, \ldots, x_{1n} \longrightarrow \bar{x}_1$

Sample 2: $x_{21}, x_{22}, \ldots, x_{2n} \longrightarrow \bar{x}_2$

Sample 3: $x_{31}, x_{32}, \ldots, x_{3n}$ ....

Population: μ, σ

Samples of size: n

Population: μ, σ

Samples of size: n

Sample means: $\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots$

Population: μ, σ

Samples of size: n

Sample means: $\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots$

Distribution of $\bar{X}$

Population: $\mu, \quad \sigma$

Samples of size: n

Sample means: $\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots$

Distribution of $\bar{X}$

1. Has mean μ
2. Has standard deviation $\dfrac{\sigma}{\sqrt{n}}$

Population: μ, σ

Samples of size: n

Sample means: $\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots$

Distribution of $\bar{X}$

1. Has mean μ
2. Has standard deviation $\dfrac{\sigma}{\sqrt{n}}$
3. Is Normal

Population: μ, σ

Samples of size: n

Sample means: $\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots$

Distribution of $\bar{X}$

1. Has mean μ
2. Has standard deviation $\dfrac{\sigma}{\sqrt{n}}$
3. Is Normal as $n \rightarrow \infty$

The Central Limit Theorem

Population: μ, σ

Samples of size: n

Sample means: $\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots$

Distribution of $\bar{X}$

1. Has mean μ
2. Has standard deviation $\dfrac{\sigma}{\sqrt{n}}$
3. Is Normal as $n \to \infty$

Example: In our population, the total cholesterol level at visit one had

$$\mu_X = 237 \quad mg/100ml$$

$$\sigma_X = 44.7 \quad mg/100ml$$

Example: In our population, the total cholesterol level at visit one had

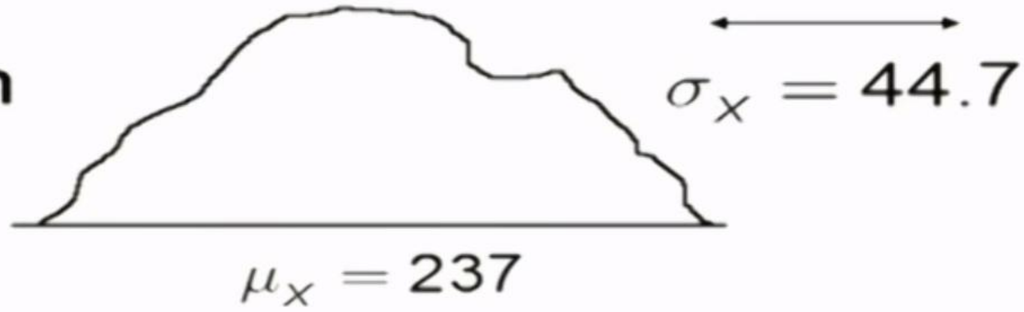
$$\mu_X = 237 \quad mg\,/\,100ml$$

$$\sigma_X = 44.7 \quad mg\,/\,100ml$$

Take *repeated* samples of size 25 from this population. What proportion of these samples will have means $\geqslant$ 260 mg/100ml ?

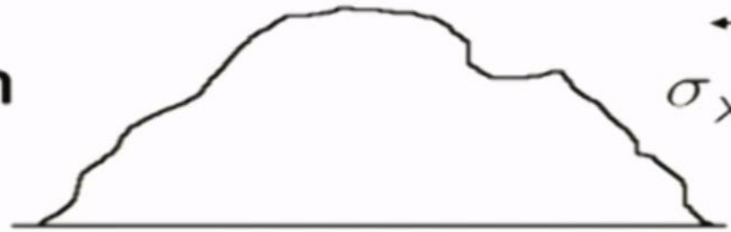
Systolic bp

Population



Systolic bp

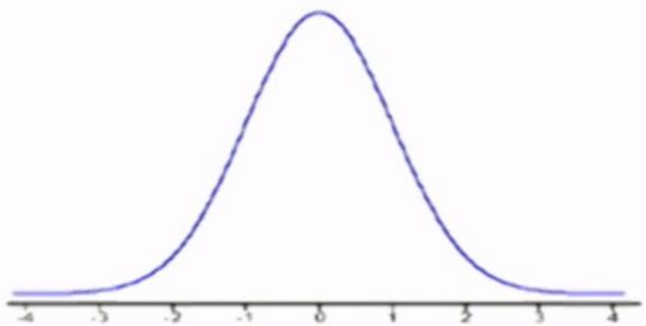
Population



$\sigma_X = 44.7$

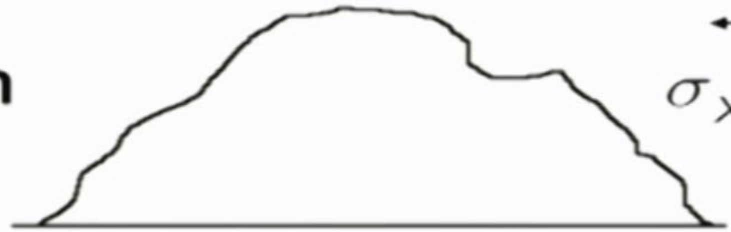
$\mu_X = 237$

Sample means



Systolic bp

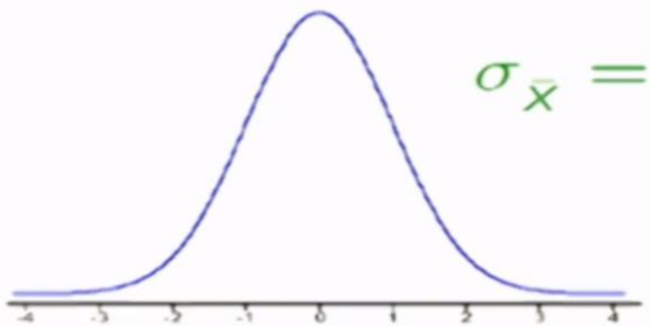
Population



$\sigma_X = 44.7$

$\mu_X = 237$

Sample means



$$\sigma_{\bar{X}} = \frac{44.7}{\sqrt{25}} = 8.94$$

$\mu_{\bar{X}} = 237$

What proportion of the $\bar{X} \geq 260$?

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

$$= \frac{260 - 237}{44.7 / \sqrt{25}} = 2.57$$

From Stata :

```
. di 1 - normal(2.57)
.00508493
```

So the probability of getting a sample mean of 260 or higher when taking a sample of 25 is about 0.005 or $\approx$ 0.5%

What proportion of the $\bar{X} \geq 260$?

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

$$= \frac{260 - 237}{44.7 / \sqrt{25}} = 2.57$$

From Stata : `. di 1 - normal(2.57)`
`.00508493`

So the probability of getting a sample mean of 260 or higher when taking a sample of 25 is about 0.005 or $\approx$ 0.5%

How big a sample do we need to be 95% sure that the *sample* mean for total cholesterol level is within $\pm$ 25 mg/100ml of the population mean?

How big a sample do we need to be 95% sure that the *sample* mean for total cholesterol level is within $\pm$ 25 mg/100ml of the population mean?

$$\Pr\left\{-25 \leq \bar{X} - \mu \leq 25\right\} = 0.95$$

How big a sample do we need to be 95% sure that the *sample* mean for total cholesterol level is within ± 25 mg/100ml of the population mean?

$$\Pr\left\{-25 \le \bar{X} - \mu \le 25\right\} = 0.95$$

$$\Pr\left\{\frac{-25}{44.7/\sqrt{n}} \le \frac{\bar{X} - \mu}{44.7/\sqrt{n}} \le \frac{25}{44.7/\sqrt{n}}\right\} = 0.95$$

How big a sample do we need to be 95% sure that the *sample* mean for total cholesterol level is within $\pm$ 25 mg/100ml of the population mean?

$$\Pr\{-25 \le \bar{X} - \mu \le 25\} = 0.95$$

$$\Pr\left\{\frac{-25}{44.7/\sqrt{n}} \le \boxed{\frac{\bar{X}-\mu}{44.7/\sqrt{n}}} \le \frac{25}{44.7/\sqrt{n}}\right\} = 0.95$$

$$\Pr\left\{\frac{-25}{44.7/\sqrt{n}} \le Z \le \frac{25}{44.7/\sqrt{n}}\right\} = 0.95$$

How big a sample do we need to be 95% sure that the *sample* mean for total cholesterol level is within $\pm$ 25 mg/100ml of the population mean?

$$\Pr\left\{-25 \le \bar{X} - \mu \le 25\right\} = 0.95$$

$$\Pr\left\{\frac{-25}{44.7/\sqrt{n}} \le \frac{\bar{X} - \mu}{44.7/\sqrt{n}} \le \frac{25}{44.7/\sqrt{n}}\right\} = 0.95$$

$$\Pr\left\{\frac{-25}{44.7/\sqrt{n}} \le Z \le \frac{25}{44.7/\sqrt{n}}\right\} = 0.95$$

$$\Rightarrow \quad \frac{25}{44.7/\sqrt{n}} = 1.96 \qquad \Rightarrow \quad n = 12.3 \quad \Rightarrow \quad n = 13$$

How big a sample do we need to be 95% sure that the *sample* mean for total cholesterol level is within $\pm$ 25 mg/100ml of the population mean?

$$\Pr\{-25 \le \bar{X} - \mu \le 25\} = 0.95$$

$$\Pr\left\{\frac{-25}{44.7/\sqrt{n}} \le \frac{\bar{X} - \mu}{44.7/\sqrt{n}} \le \frac{25}{44.7/\sqrt{n}}\right\} = 0.95$$

$$\Pr\left\{\frac{-25}{44.7/\sqrt{n}} \le Z \le \frac{25}{44.7/\sqrt{n}}\right\} = 0.95$$

$$\Rightarrow \frac{25}{44.7/\sqrt{n}} = 1.96 \qquad \Rightarrow n = 12.3 \Rightarrow n = 13$$

$$\Rightarrow \frac{12.5}{44.7/\sqrt{n}} = 1.96 \qquad \Rightarrow n = 49.1 \Rightarrow n = 50$$

Sample Size

So, in general if we want to be 95% sure that the sample mean will be within $\pm\,\Delta$ of the population mean, then we need a sample of size

$$\left(\frac{1.96\sigma}{\Delta}\right)^2$$

where σ is the population standard deviation.

Sample Size

So, in general if we want to be 95% sure that the sample mean will be within $\pm\,\Delta$ of the population mean, then we need a sample of size

$$n = \left(\frac{1.96\sigma}{\Delta}\right)^2$$

where σ is the population standard deviation.

Confidence Interval on μ (σ known)

$$\Pr\left\{-1.96 \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le 1.96\right\} = 0.95$$

Confidence Interval on μ (σ known)

$$\Pr\left\{-1.96 \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq 1.96\right\} = 0.95$$

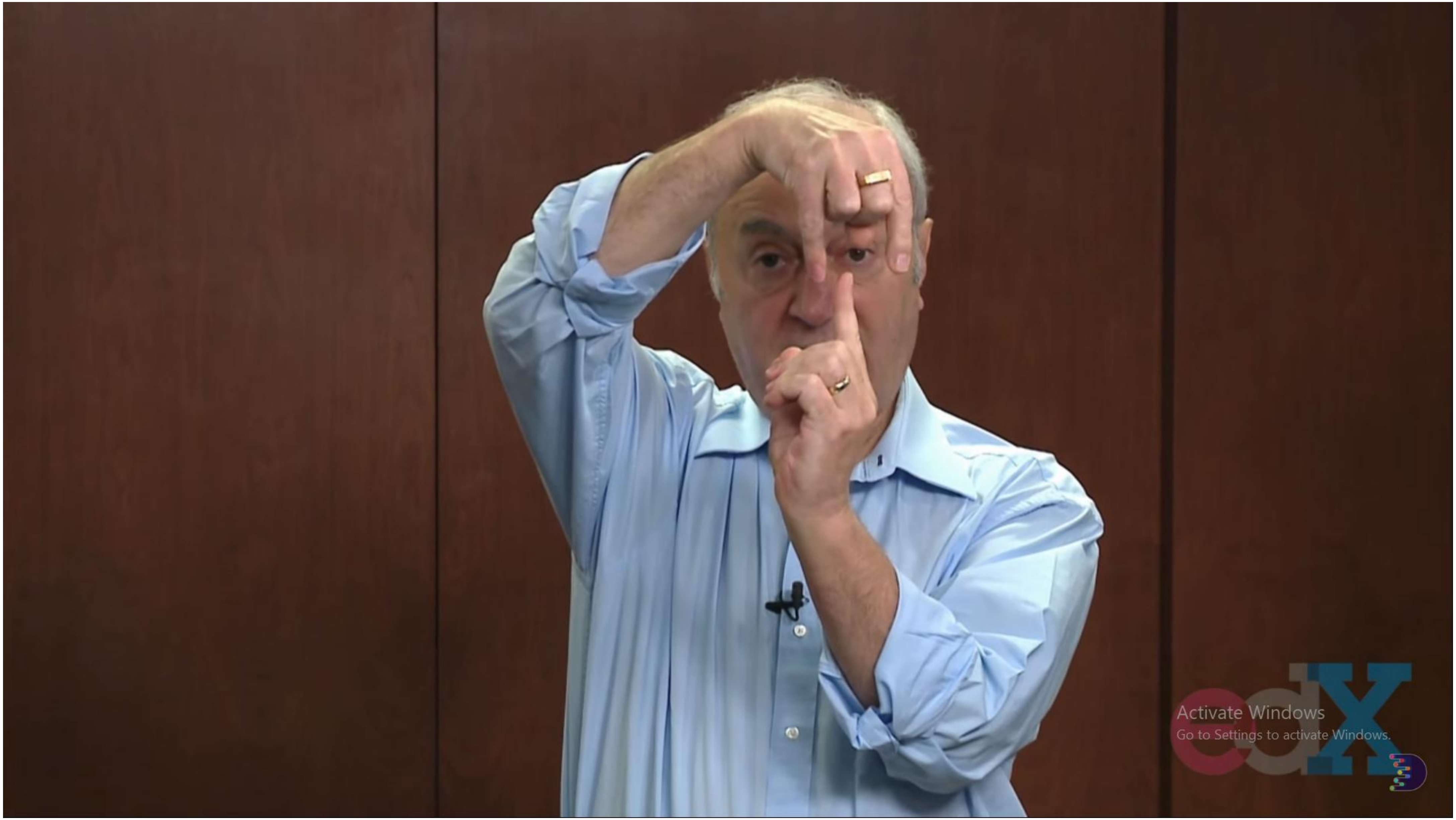
$$\Pr\left\{\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right\} = 0.95$$

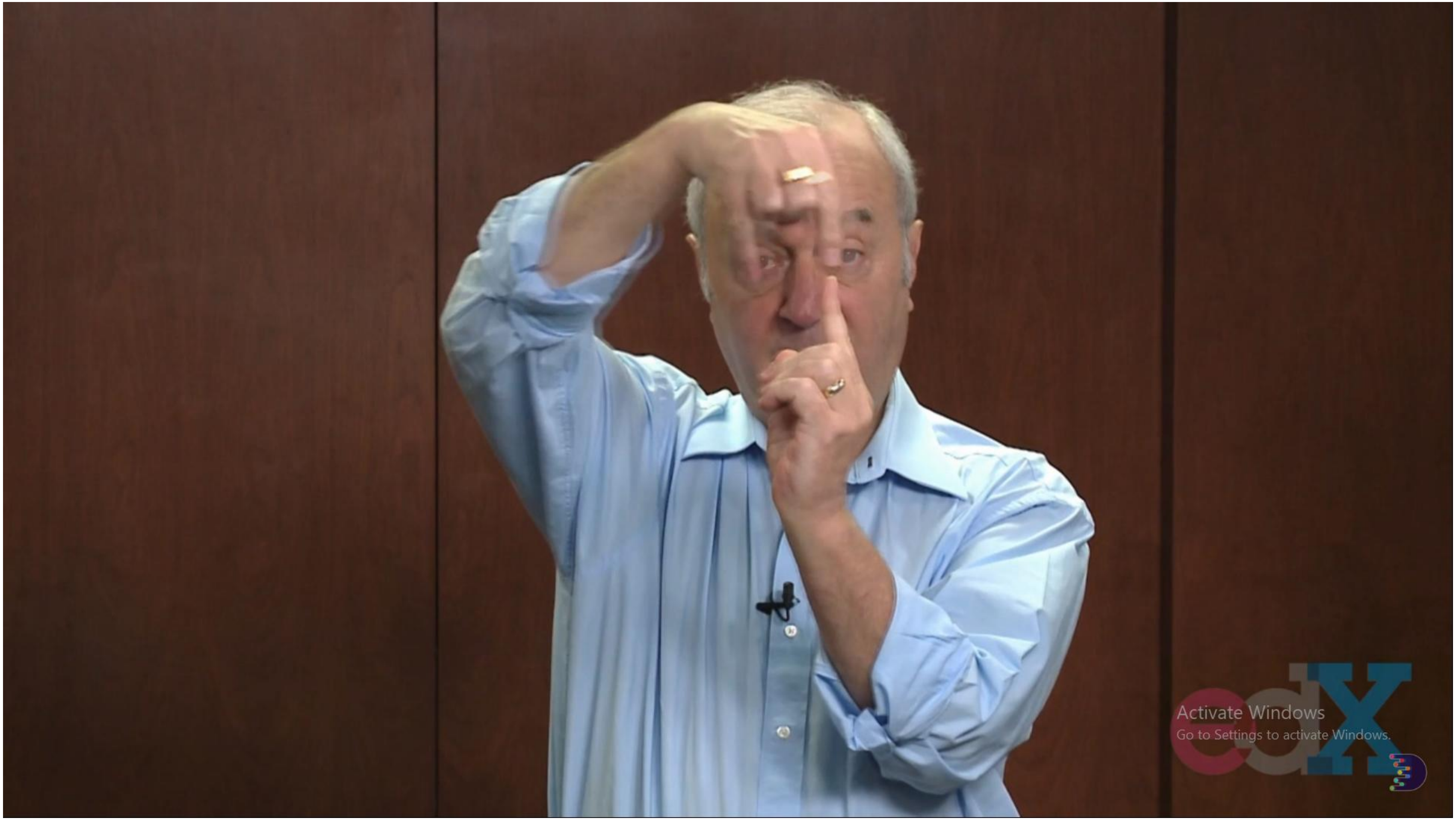
Confidence Interval on μ (σ known)

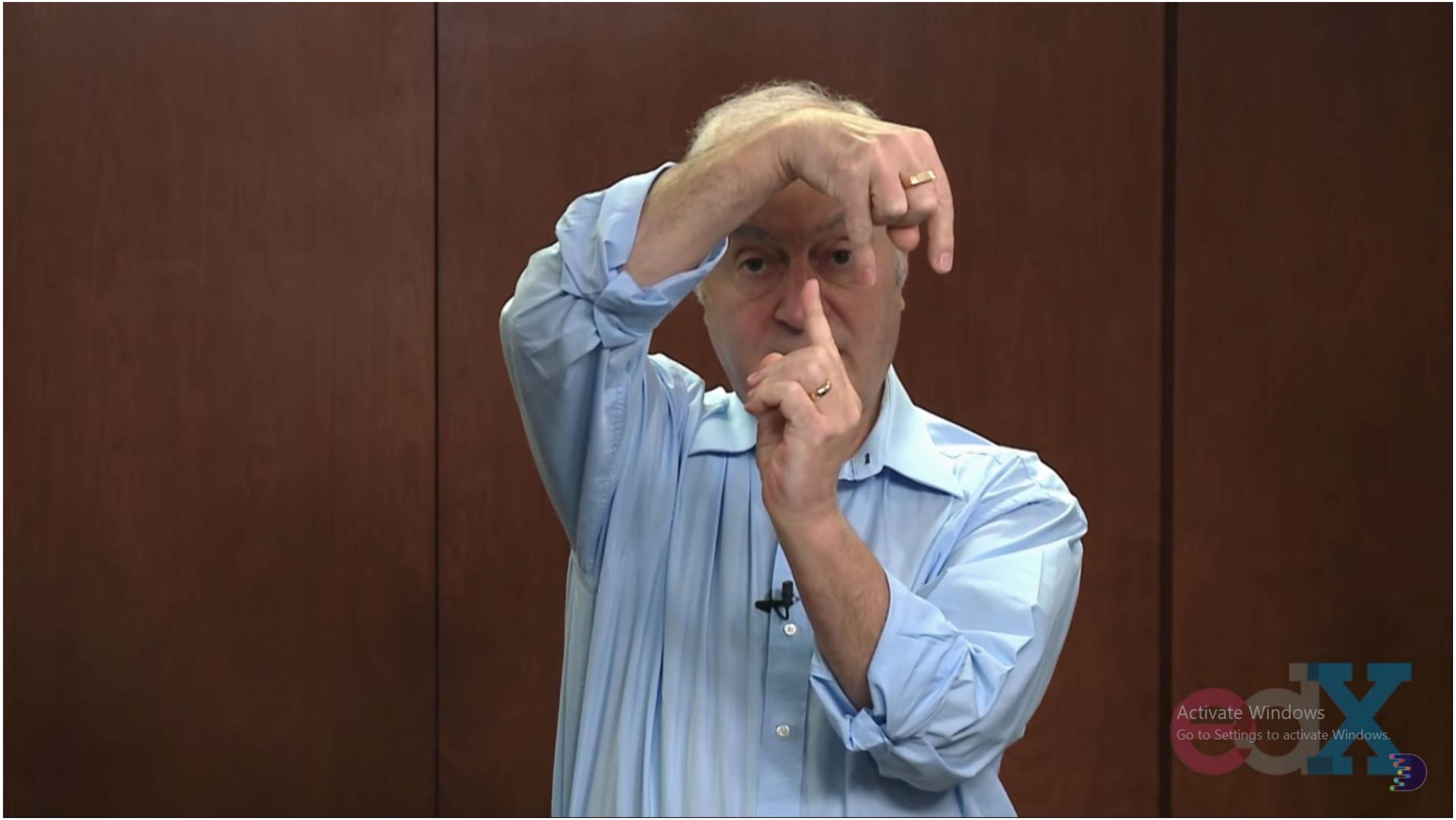
$$\Pr\left\{-1.96 \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq 1.96\right\} = 0.95$$

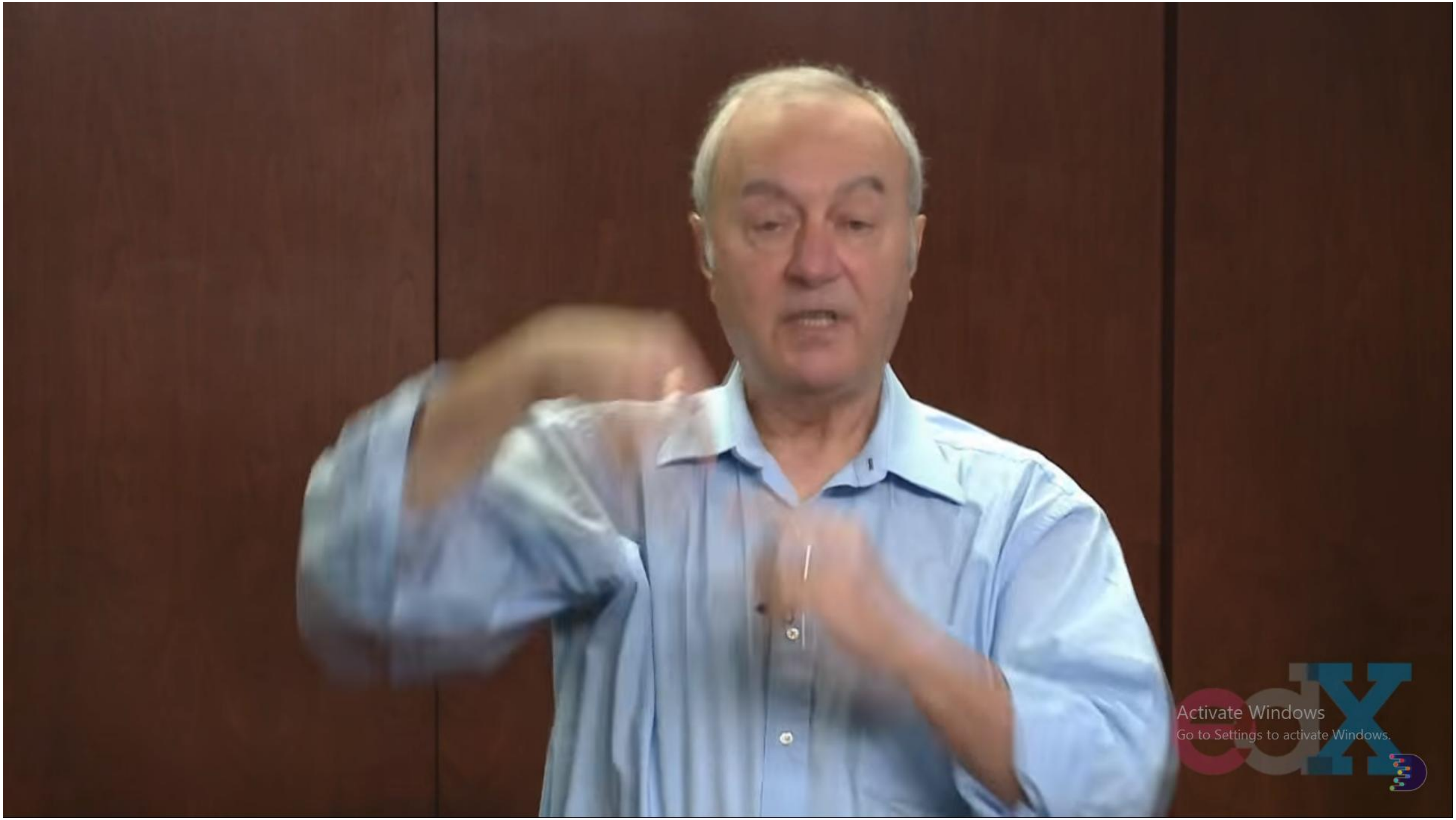
$$\Pr\left\{\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right\} = 0.95$$

$$\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$


Activate Windows
Go to Settings to activate Windows.


Activate Windows
Go to Settings to activate Windows.


Activate Windows
Go to Settings to activate Windows.


Activate Windows
Go to Settings to activate Windows.

Confidence Interval on μ (σ known)

$$\Pr\left\{-1.96 \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le 1.96\right\} = 0.95$$

$$\Pr\left\{\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right\} = 0.95$$

$$\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}},\ \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

is a 95% confidence interval for μ .

Confidence Interval on μ (σ known)

Before taking the sample:

$$\Pr\left\{\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right\} = 0.95$$

the interval:

$$\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

has a 95% chance of covering μ .

95% confidence interval for total cholesterol mean

e.g. If we take a sample of size $n = 49$ from our Framingham population where $\sigma = 47.7$ mg/100ml for the total cholesterol level, then the interval

$$\left(\bar{X} - 1.96\frac{47.7}{\sqrt{49}}, \bar{X} + 1.96\frac{47.7}{\sqrt{49}}\right)$$

$$\left(\bar{X} - 13.4, \bar{X} + 13.4\right)$$

has a 95% chance of covering μ (which we know is 237.0 mg/100ml).

95% confidence interval for total cholesterol mean

e.g. If we take a sample of size $n = 49$ from our Framingham population where $\sigma = 47.7$ mg/100ml for the total cholesterol level, then the interval

$$\left(\bar{X} - 1.96\frac{47.7}{\sqrt{49}}, \bar{X} + 1.96\frac{47.7}{\sqrt{49}}\right)$$

$$(\bar{X} - 13.4\ ,\ \bar{X} + 13.4)$$

has a 95% chance of covering μ (which we know is 237.0 mg/100ml).

95% confidence interval for total cholesterol mean

We observed $\bar{x} = 233.7$ so the 95% confidence interval is $\boxed{(220.3\ ,\ 247.1)}$.

Here we know the answer because we know the population, but in general, this interval may or may *not* contain the mean of the population, we do not know. But we followed the rules that give us a 95% chance of being correct– confident.

Predictive Interval

If we have a Normal distribution with mean μ and standard deviation σ, then for a single observation X,

$$Z = \frac{X - \mu}{\sigma}$$

$$\Pr\left\{-1.96 \le \frac{X - \mu}{\sigma} \le 1.96\right\} = 0.95$$

$$\Pr\{\mu - 1.96\sigma \le X \le \mu + 1.96\sigma\} = 0.95$$

Predictive Interval

If we have a Normal distribution with mean μ and standard deviation σ, then for a single observation X,

$$Z = \frac{X - \mu}{\sigma}$$

$$\Pr\left\{-1.96 \leq \frac{X - \mu}{\sigma} \leq 1.96\right\} = 0.95$$

$$\Pr\{\mu - 1.96\sigma \leq X \leq \mu + 1.96\sigma\} = 0.95$$

So, $(\mu - 1.96\sigma, \mu + 1.96\sigma)$

is a predictive interval for X, just as

$$\left(\mu - 1.96\frac{\sigma}{\sqrt{n}}, \mu + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

is a predictive interval for $\overline{X}$, So, $\left(\mu - 1.96\sigma, \mu + 1.96\sigma\right)$

is a predictive interval for X, just as

$$\left(\mu - 1.96\frac{\sigma}{\sqrt{n}}, \mu + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

is a predictive interval for $\overline{\mathbf{X}}$, So, $(\mu - 1.96\sigma, \mu + 1.96\sigma)$

is a predictive interval for X, just as

$$\left(\mu - 1.96\boxed{\frac{\sigma}{\sqrt{n}}}, \mu + 1.96\boxed{\frac{\sigma}{\sqrt{n}}}\right)$$

is a predictive interval for $\bar{X}$, and

$$\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

is a confidence interval for μ.

e.g. So for total cholesterol at visit 1, $\mu = 237$ and $\sigma = 44.7$, so

$$(\mu - 1.96\sigma, \mu + 1.96\sigma) = (149.4, 324.6)$$

e.g. So for total cholesterol at visit 1, $\mu = 237$ and $\sigma = 44.7$, so

$$(\mu - 1.96\sigma, \mu + 1.96\sigma) = (149.4, 324.6)$$

is a 95% predictive interval for X, just as

$$\left(\mu - 1.96\frac{\sigma}{\sqrt{n}}, \mu + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

e.g. So for total cholesterol at visit 1, μ = 237 and σ = 44.7, so

$$\left(\mu - 1.96\sigma, \mu + 1.96\sigma\right) = (149.4, 324.6)$$

is a 95% predictive interval for X, just as

$$\left(\mu - 1.96\frac{\sigma}{\sqrt{n}}, \mu + 1.96\frac{\sigma}{\sqrt{n}}\right) = (224.5, 249.5)$$

e.g. So for total cholesterol at visit 1, μ = 237 and σ = 44.7, so

$$\left(\mu - 1.96\sigma\ ,\ \mu + 1.96\sigma\right) = (149.4,\ 324.6)$$

is a 95% predictive interval for X, just as

$$\left(\mu - 1.96\frac{\sigma}{\sqrt{n}}\ ,\ \mu + 1.96\frac{\sigma}{\sqrt{n}}\right) = (224.5,\ 249.5)$$

is a 95% predictive interval for $\bar{X}$, and

$$\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}},\ \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = (\bar{X} - 12.5,\ \bar{X} + 12.5)$$

e.g. So for total cholesterol at visit 1, $\mu = 237$ and $\sigma = 44.7$, so

$$\left(\mu - 1.96\sigma\,,\ \mu + 1.96\sigma\right) = (149.4,\ 324.6)$$

is a 95% predictive interval for X, just as

$$\left(\mu - 1.96\frac{\sigma}{\sqrt{n}}\,,\ \mu + 1.96\frac{\sigma}{\sqrt{n}}\right) = (224.5,\ 249.5)$$

is a 95% predictive interval for $\bar{X}$, and

$$\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}},\ \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = (\bar{X} - 12.5,\ \bar{X} + 12.5)$$

$$= (221.2,\ 246.2)$$

is a 95% confidence interval for μ.

Width of Confidence Interval

		width
95%	$\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$	$3.92 \frac{\sigma}{\sqrt{n}}$

Width of Confidence Interval

		width
95%	$\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$	$\boxed{3.92 \frac{\sigma}{\sqrt{n}}}$
99%	$\bar{X} \pm 2.58 \frac{\sigma}{\sqrt{n}}$	$5.16 \frac{\sigma}{\sqrt{n}}$

Width of Confidence Interval

		width
95%	$\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$	$3.92 \frac{\sigma}{\sqrt{n}}$
99%	$\bar{X} \pm 2.58 \frac{\sigma}{\sqrt{n}}$	$5.16 \frac{\sigma}{\sqrt{n}}$

- As confidence increases (95% to 99%) the width of the interval increases.

Width of Confidence Interval

		width
95%	$\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$	$3.92 \frac{\sigma}{\sqrt{n}}$
99%	$\bar{X} \pm 2.58 \frac{\sigma}{\sqrt{n}}$	$5.16 \frac{\sigma}{\sqrt{n}}$

- As confidence increases (95% to 99%) the width of the interval increases.
- As the sample size increases, the width decreases.

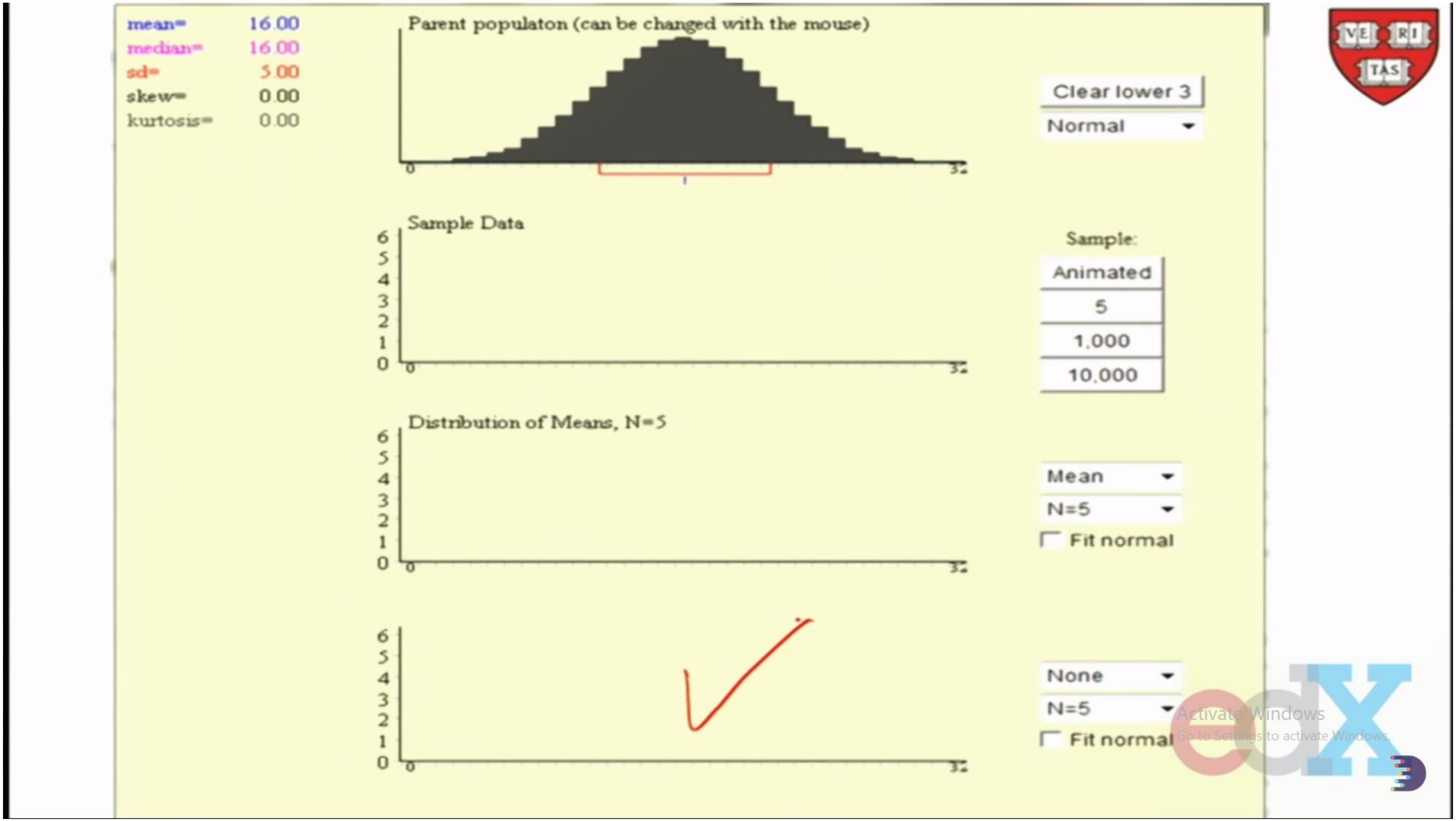
n	95% CI for μ	Interval width
10	$\bar{X} \pm 0.620\sigma$	1.240σ
100	$\bar{X} \pm 0.196\sigma$	0.392σ
1000	$\bar{X} \pm 0.062\sigma$	0.124σ

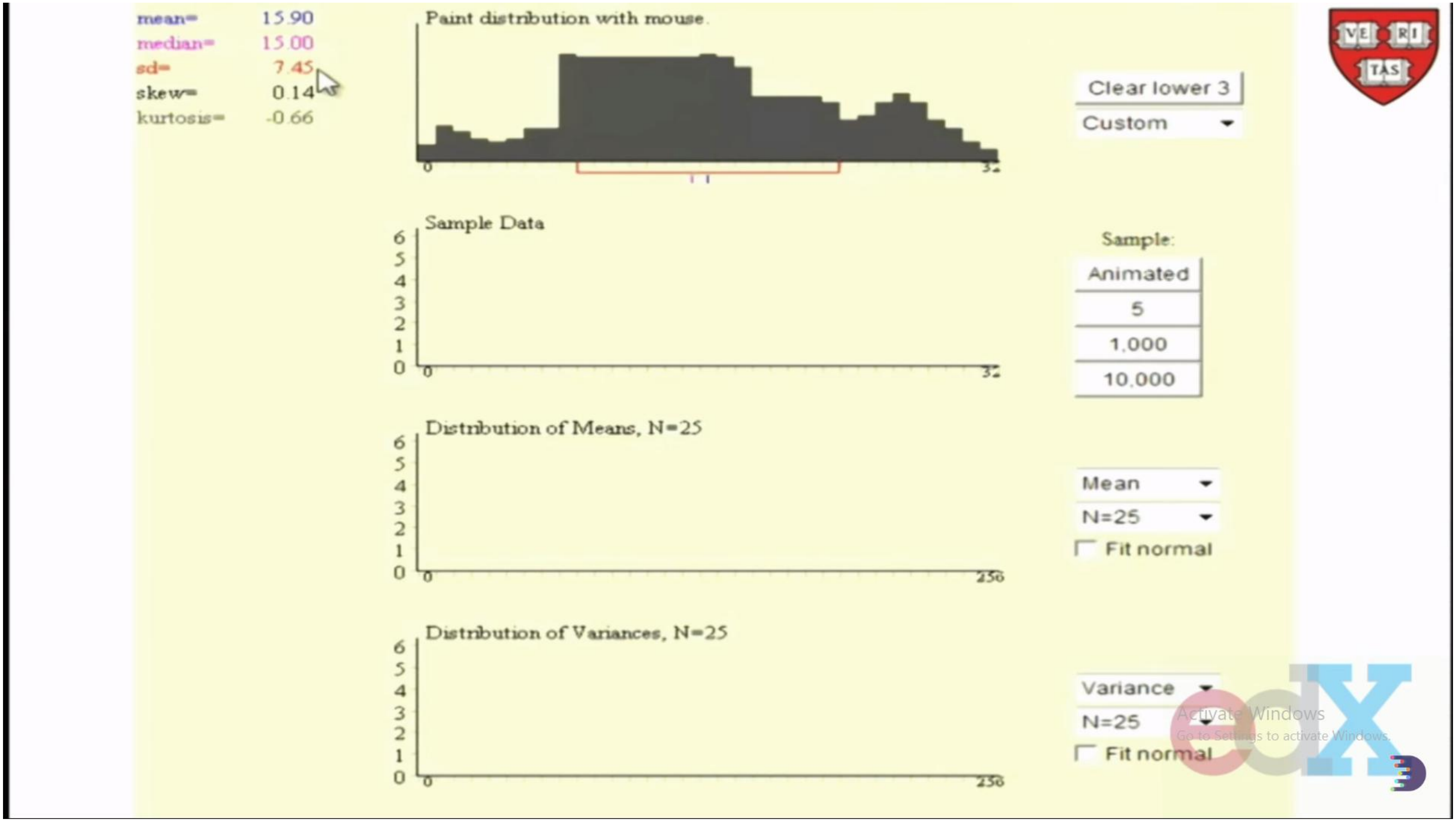
Smaller is σ, the tighter are the bounds – more homogeneous.

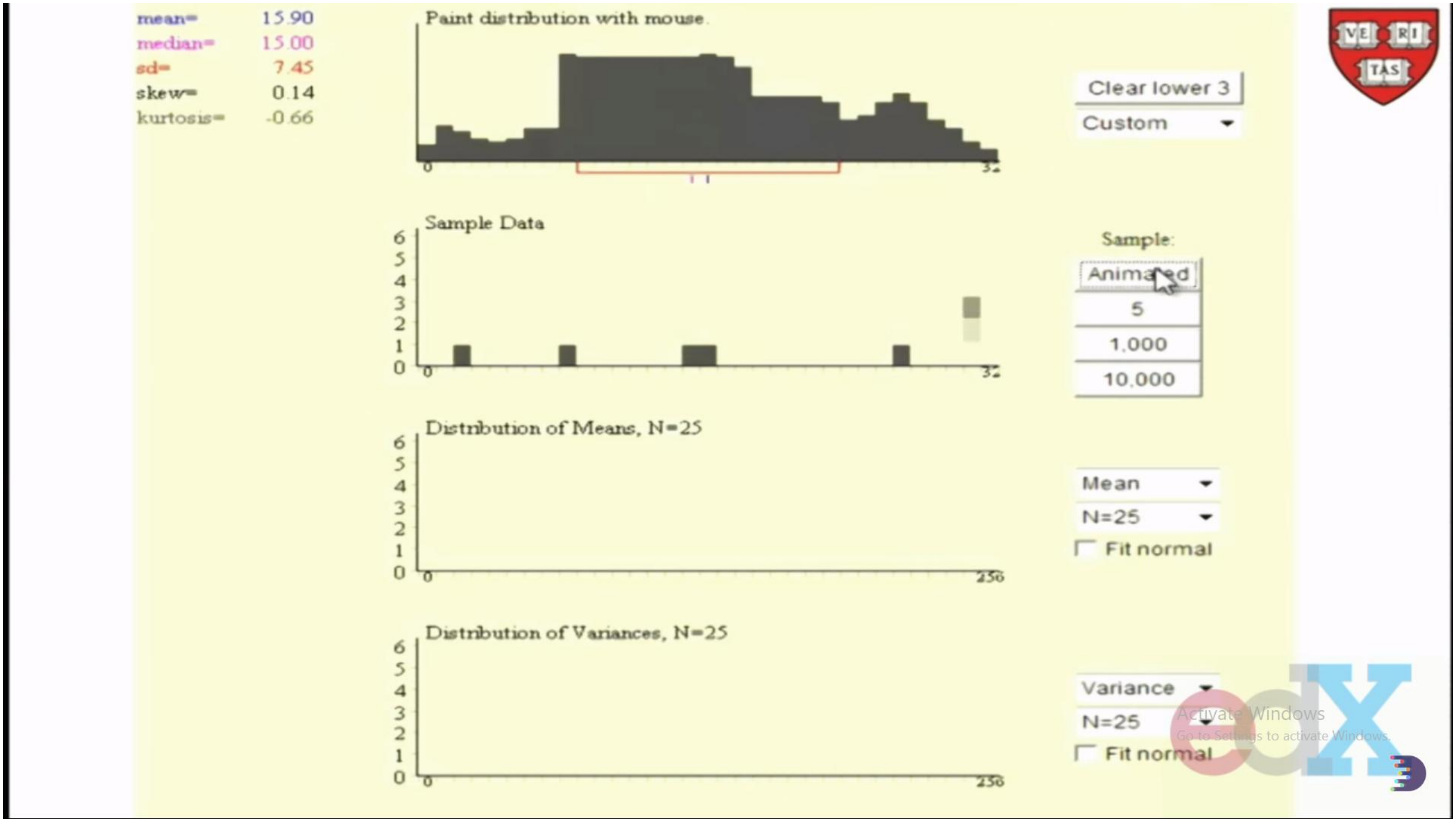
n	95% CI for μ	Interval width
10	$\bar{X} \pm 0.620\sigma$	1.240σ
100	$\bar{X} \pm 0.196\sigma$	0.392σ
1000	$\bar{X} \pm 0.062\sigma$	0.124σ

Smaller is σ, the tighter are the bounds – more homogeneous.

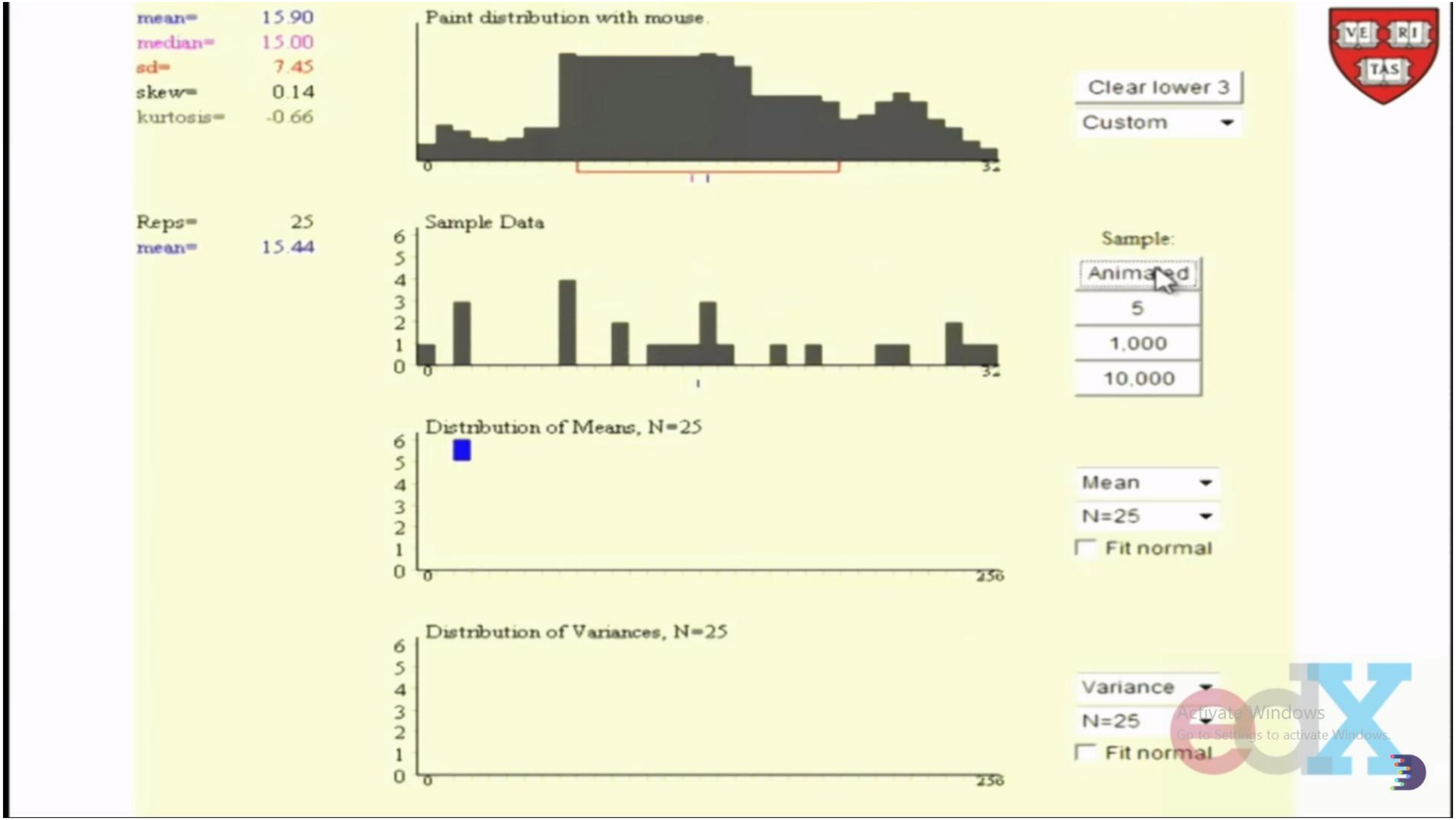
What if σ is unknown?


mean= 16.00
median= 16.00
sd= 5.00
skew= 0.00
kurtosis= 0.00
Parent populaton (can be changed with the mouse)
Clear lower 3
Normal
Sample Data
Sample:
Animated
5
1,000
10,000
Distribution of Means, N=5
Mean
N=5
Fit normal
None
N=5
Fit normal

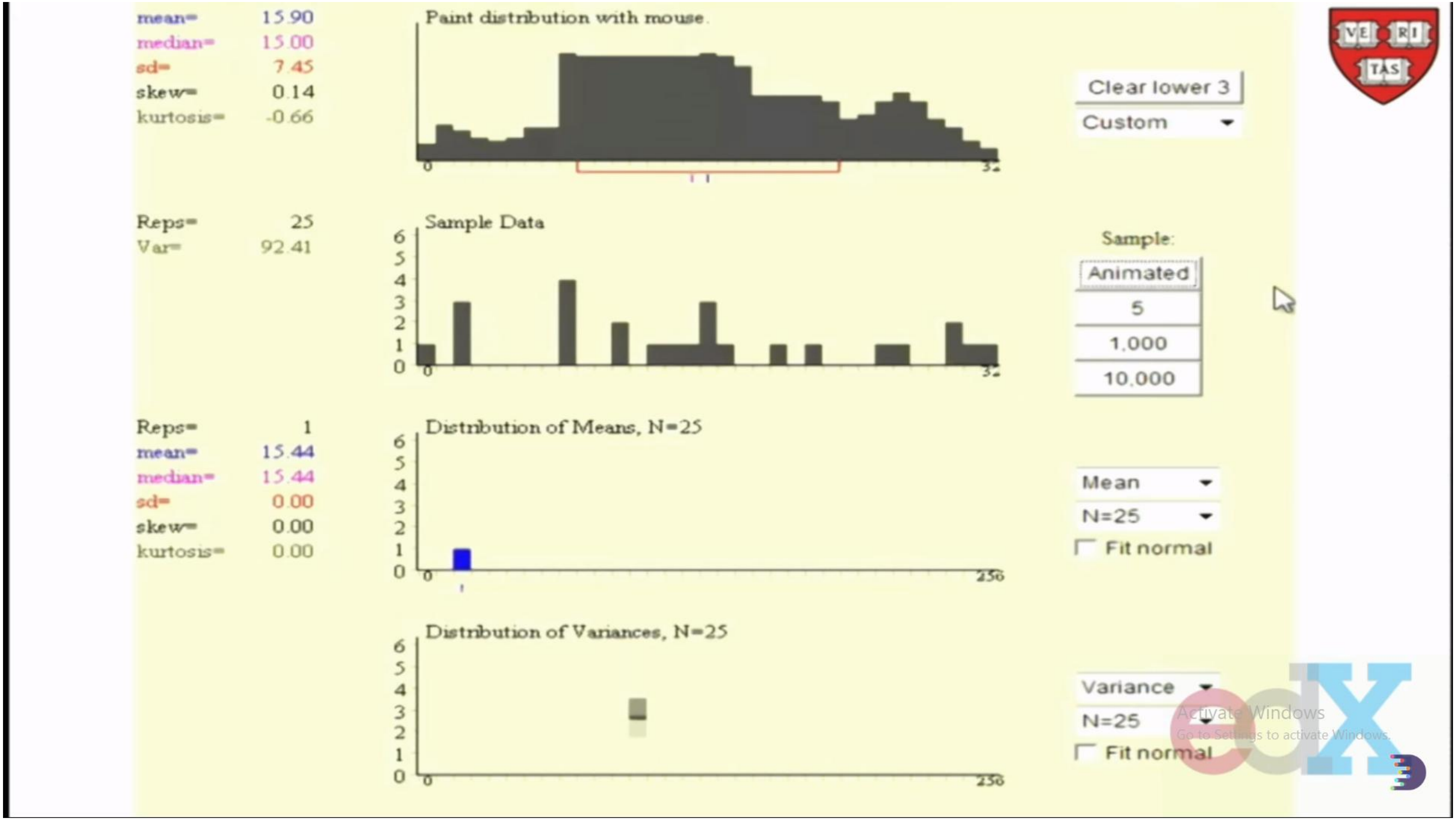

mean= 15.90
median= 15.00
sd= 7.45
skew= 0.14
kurtosis= -0.66
Paint distribution with mouse.
0
32
Clear lower 3
Custom
Sample Data
6
5
4
3
2
1
0
0
32
Sample:
Animated
5
1,000
10,000
Distribution of Means, N=25
6
5
4
3
2
1
0
0
256
Mean
N=25
Fit normal
Distribution of Variances, N=25
6
5
4
3
2
1
0
0
256
Variance
N=25
Fit normal
VE
RI
TAS
Activate Windows



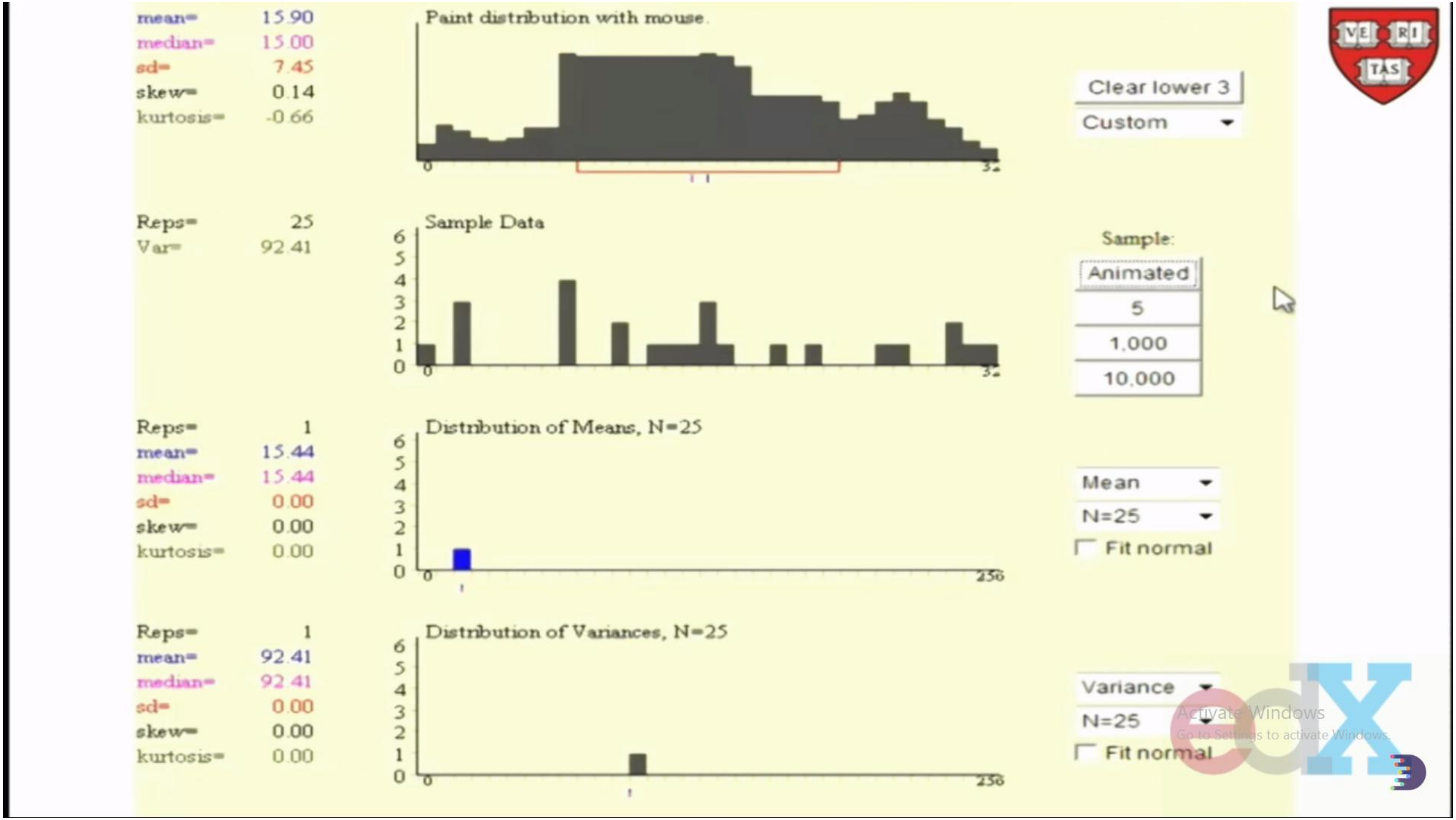
mean= 15.90
median= 15.00
sd= 7.45
skew= 0.14
kurtosis= -0.66
Paint distribution with mouse.
Clear lower 3
Custom
Sample Data
Sample:
Animated
5
1,000
10,000
Distribution of Means, N=25
Mean
N=25
Fit normal
Distribution of Variances, N=25
Variance
N=25
Fit normal
VE RI TAS
Activate Windows
Go to Settings to activate Windows.



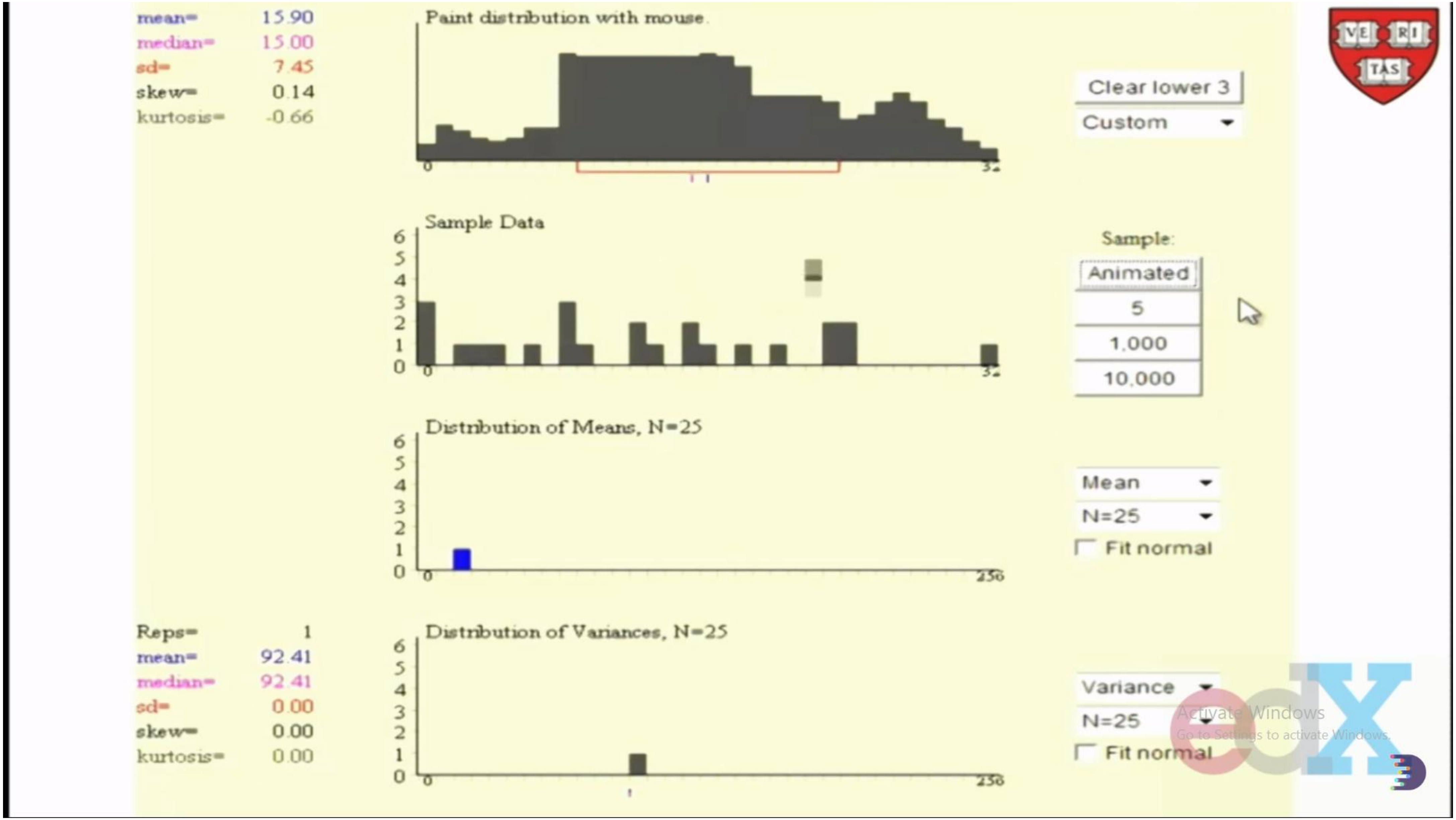
mean= 15.90
median= 15.00
sd= 7.45
skew= 0.14
kurtosis= -0.66
Paint distribution with mouse.
Clear lower 3
Custom
Reps= 25
mean= 15.44
Sample Data
Sample:
Animated
5
1,000
10,000
Distribution of Means, N=25
Mean
N=25
Fit normal
Distribution of Variances, N=25
Variance
N=25
Fit normal



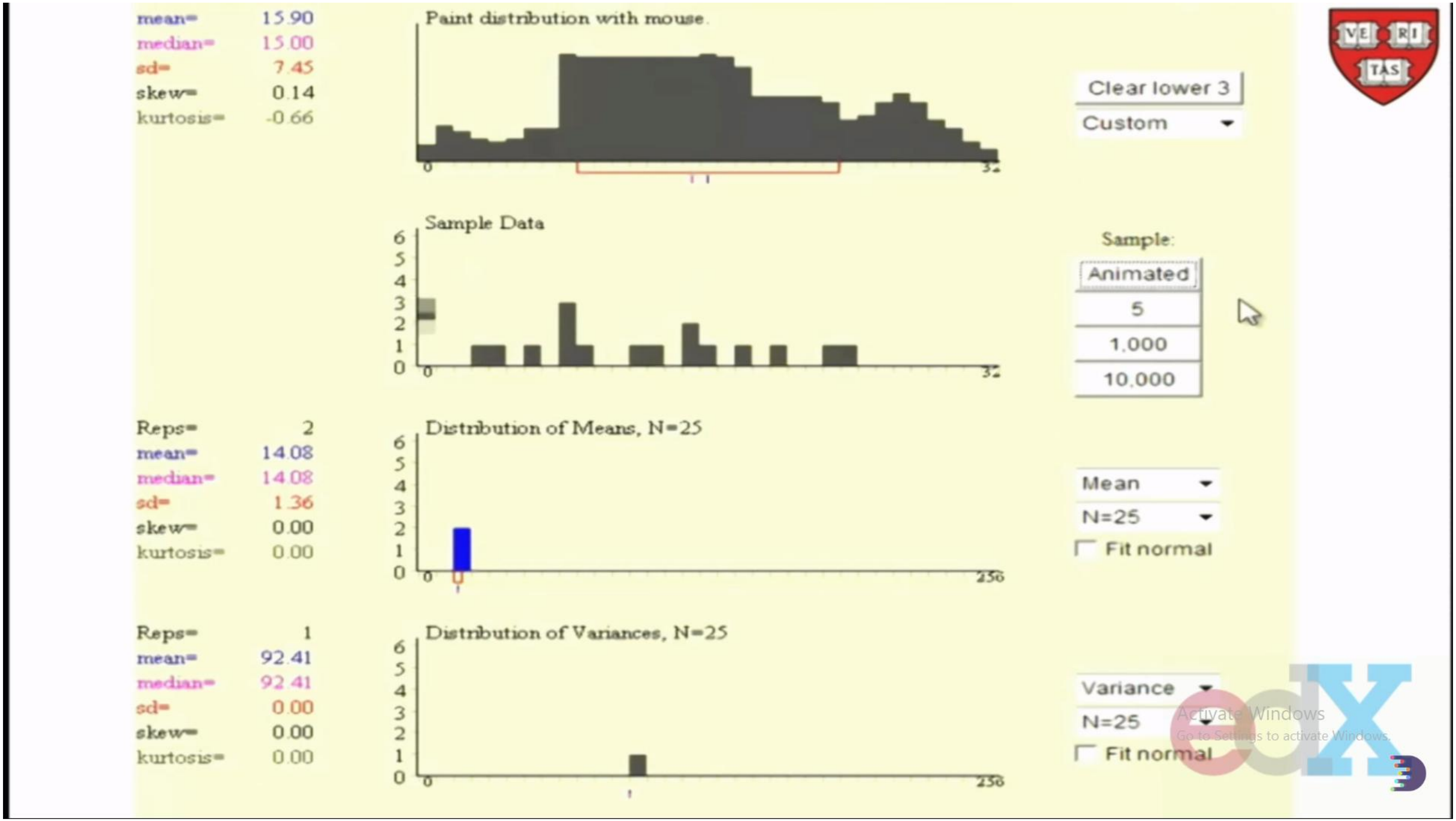
mean= 15.90
median= 15.00
sd= 7.45
skew= 0.14
kurtosis= -0.66
Paint distribution with mouse.
Clear lower 3
Custom
Reps= 25
Var= 92.41
Sample Data
Sample:
Animated
5
1,000
10,000
Reps= 1
mean= 15.44
median= 15.44
sd= 0.00
skew= 0.00
kurtosis= 0.00
Distribution of Means, N=25
Mean
N=25
Fit normal
Distribution of Variances, N=25
Variance
N=25
Fit normal

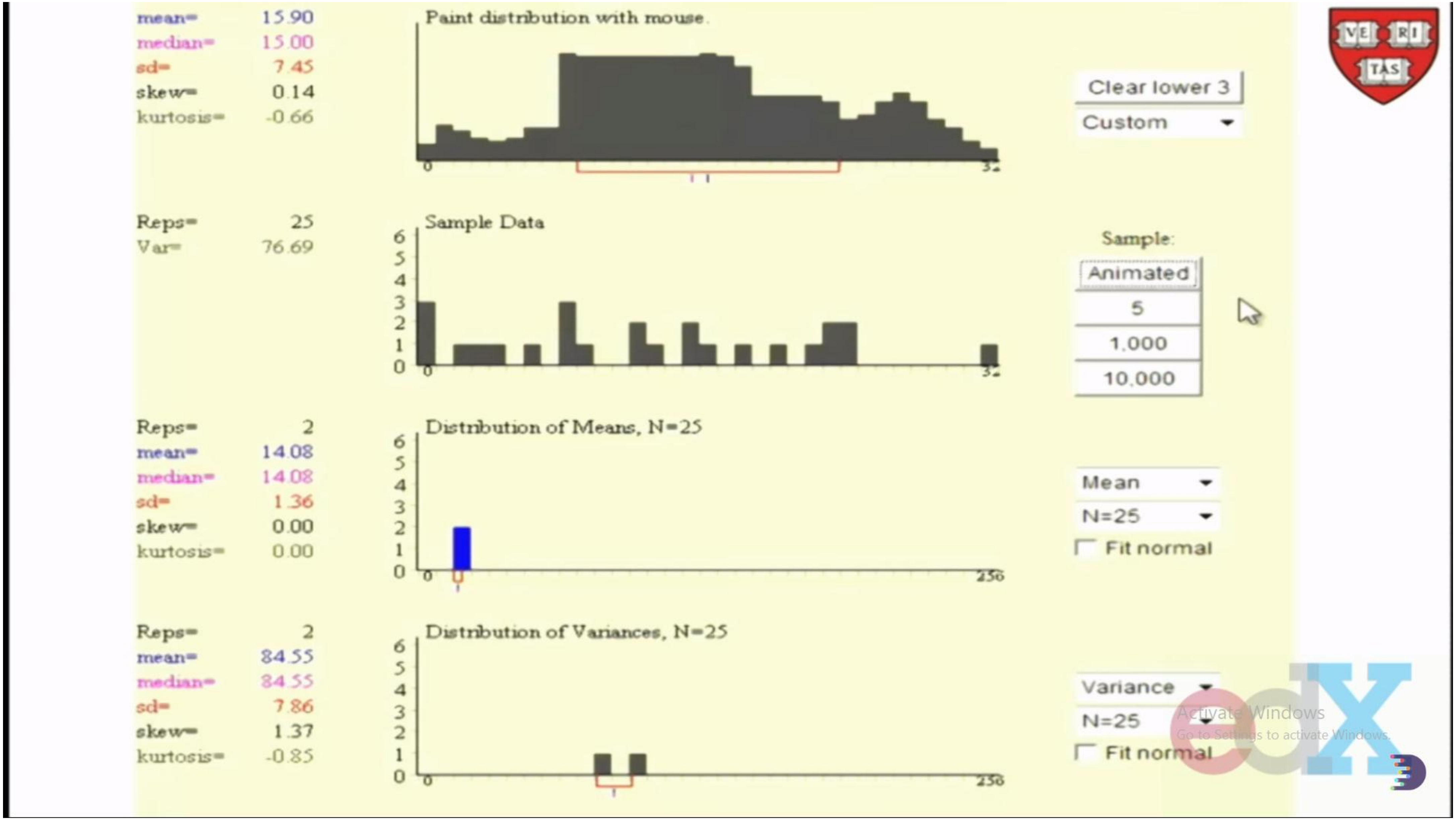


mean= 15.90
median= 15.00
sd= 7.45
skew= 0.14
kurtosis= -0.66
Paint distribution with mouse.
Clear lower 3
Custom
Reps= 25
Var= 92.41
Sample Data
Sample:
Animated
5
1,000
10,000
Reps= 1
mean= 15.44
median= 15.44
sd= 0.00
skew= 0.00
kurtosis= 0.00
Distribution of Means, N=25
Mean
N=25
Fit normal
Reps= 1
mean= 92.41
median= 92.41
sd= 0.00
skew= 0.00
kurtosis= 0.00
Distribution of Variances, N=25
Variance
N=25
Fit normal

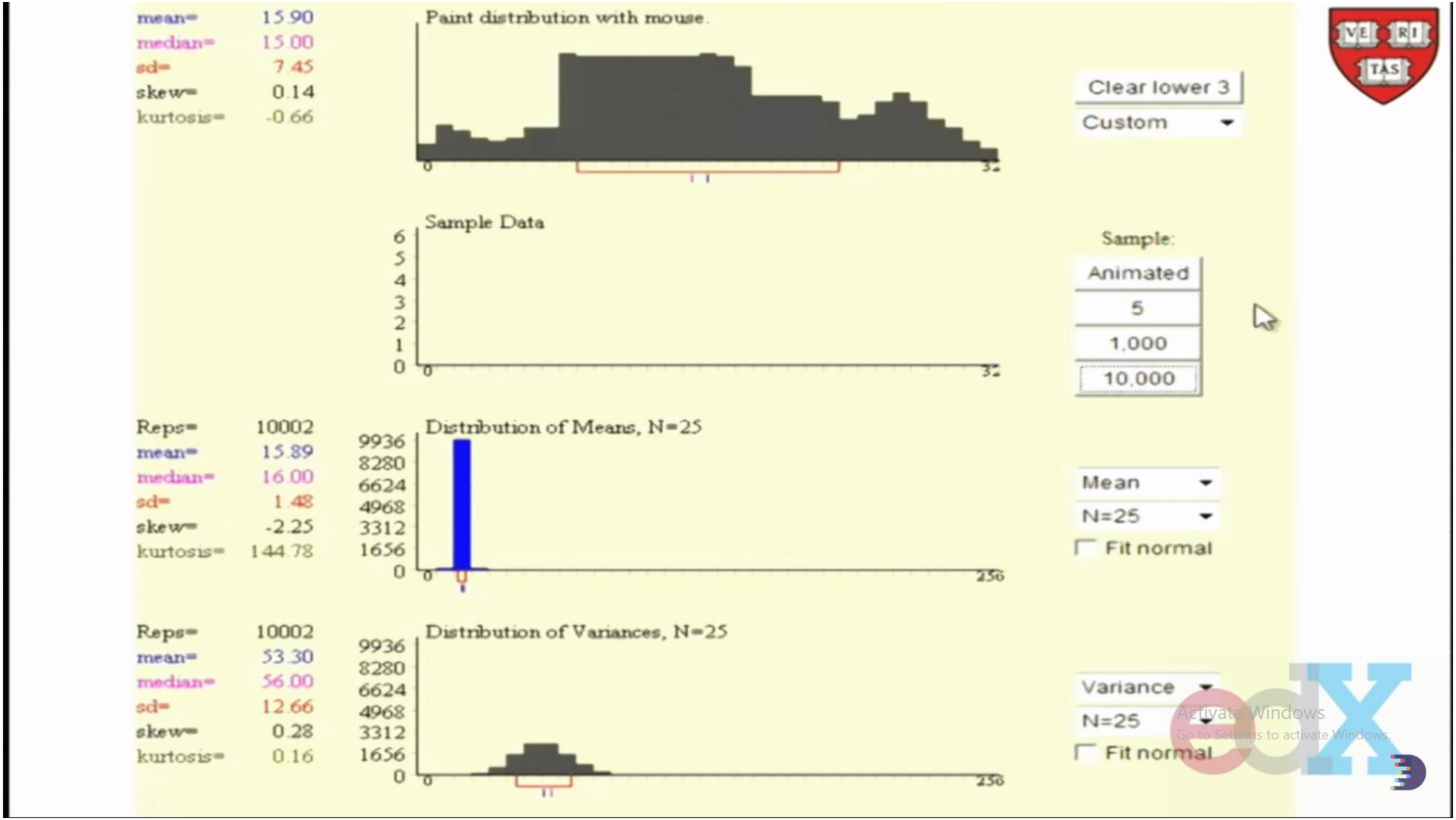


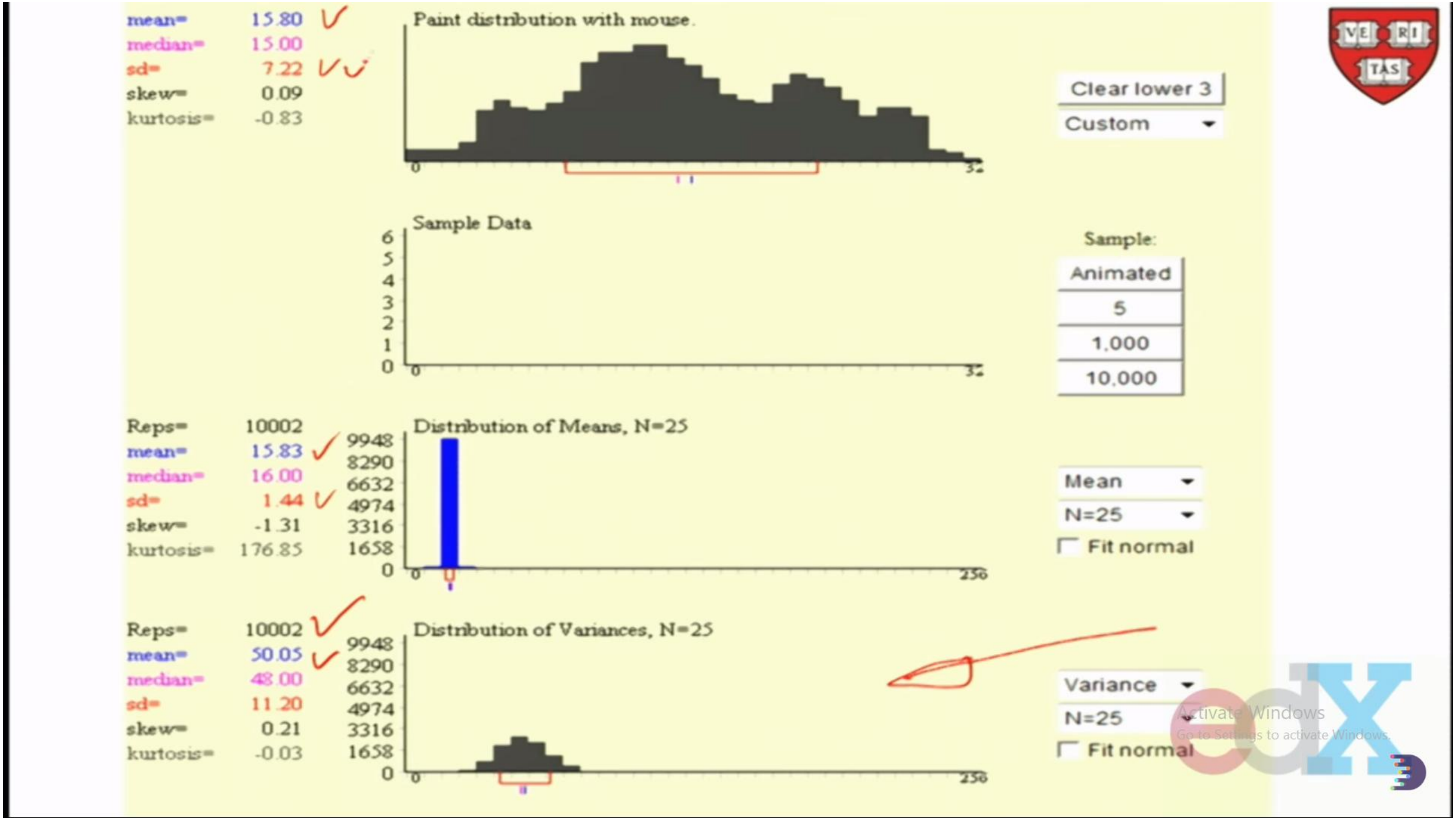
mean= 15.90
median= 15.00
sd= 7.45
skew= 0.14
kurtosis= -0.66
Paint distribution with mouse.
Clear lower 3
Custom
Sample Data
Sample:
Animated
5
1,000
10,000
Distribution of Means, N=25
Mean
N=25
Fit normal
Reps= 1
mean= 92.41
median= 92.41
sd= 0.00
skew= 0.00
kurtosis= 0.00
Distribution of Variances, N=25
Variance
N=25
Fit normal


mean= 15.90
median= 15.00
sd= 7.45
skew= 0.14
kurtosis= -0.66
Paint distribution with mouse.
Clear lower 3
Custom
Sample Data
Sample:
Animated
5
1,000
10,000
Reps= 2
mean= 14.08
median= 14.08
sd= 1.36
skew= 0.00
kurtosis= 0.00
Distribution of Means, N=25
Mean
N=25
Fit normal
Reps= 1
mean= 92.41
median= 92.41
sd= 0.00
skew= 0.00
kurtosis= 0.00
Distribution of Variances, N=25
Variance
N=25
Fit normal

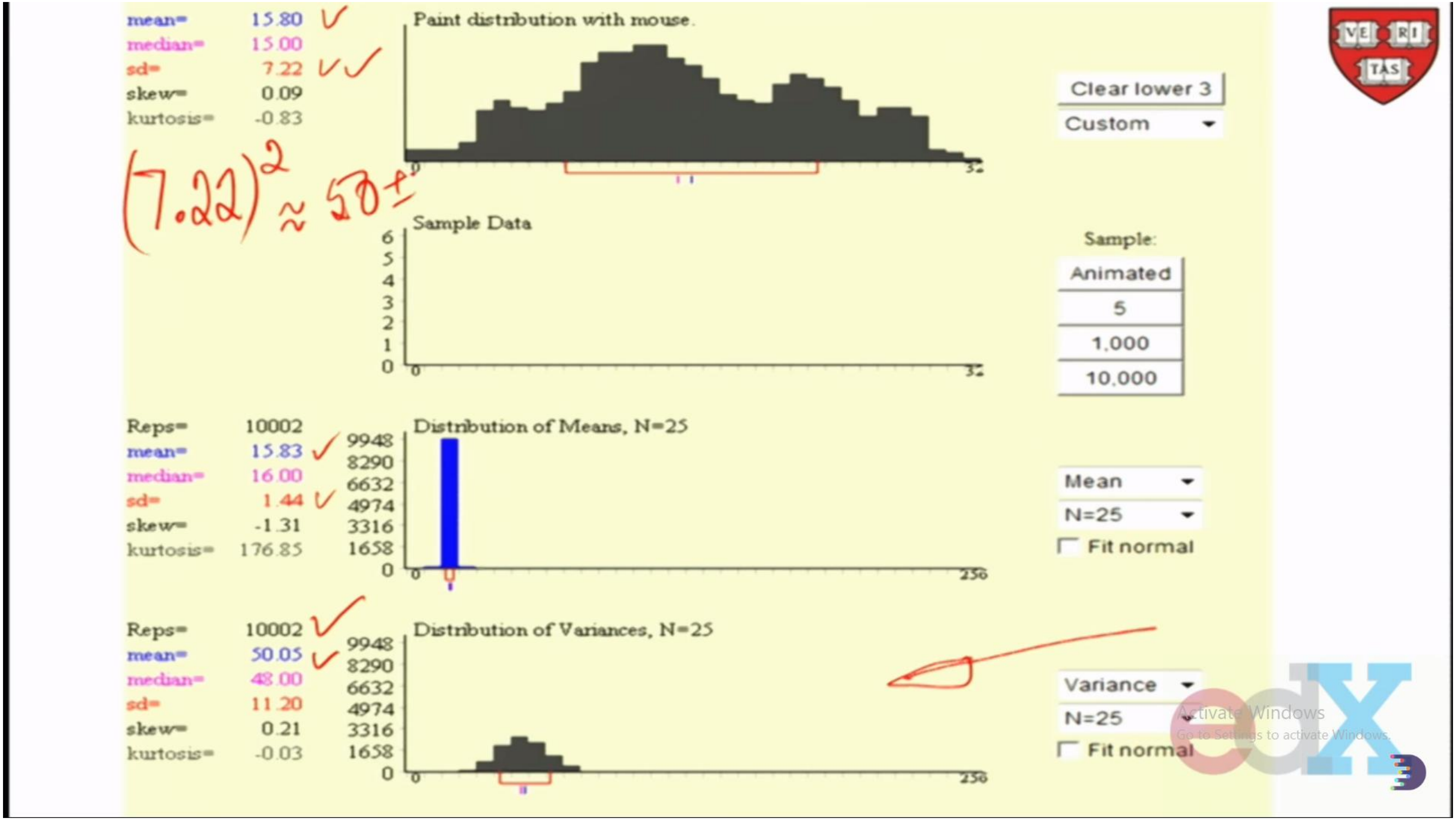


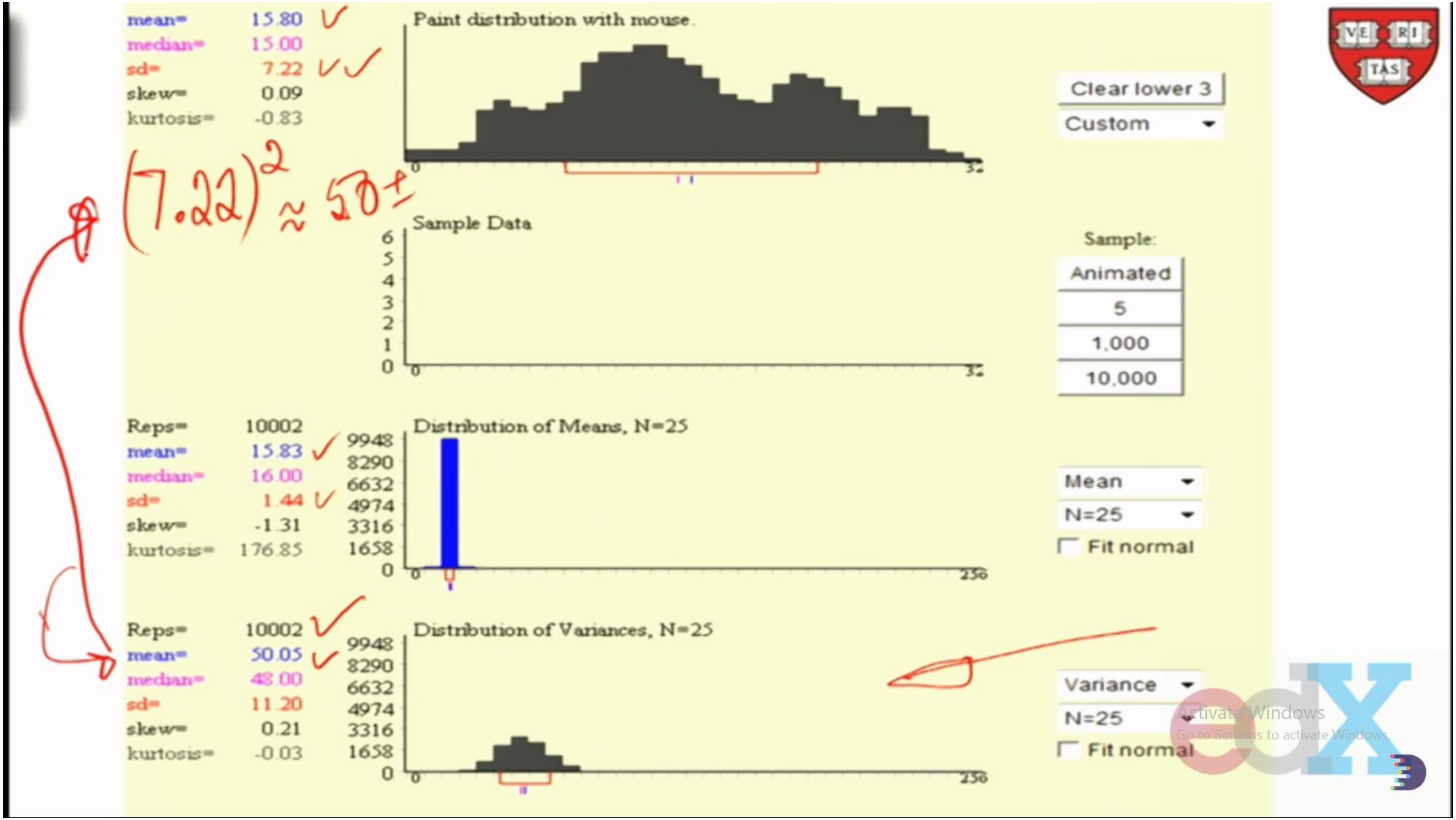
mean= 15.90
median= 15.00
sd= 7.45
skew= 0.14
kurtosis= -0.66
Paint distribution with mouse.
Clear lower 3
Custom
Reps= 25
Var= 76.69
Sample Data
Sample:
Animated
5
1,000
10,000
Reps= 2
mean= 14.08
median= 14.08
sd= 1.36
skew= 0.00
kurtosis= 0.00
Distribution of Means, N=25
Mean
N=25
Fit normal
Reps= 2
mean= 84.55
median= 84.55
sd= 7.86
skew= 1.37
kurtosis= -0.85
Distribution of Variances, N=25
Variance
N=25
Fit normal
VE RI TAS
Activate Windows
Go to Settings to activate Windows.
edX

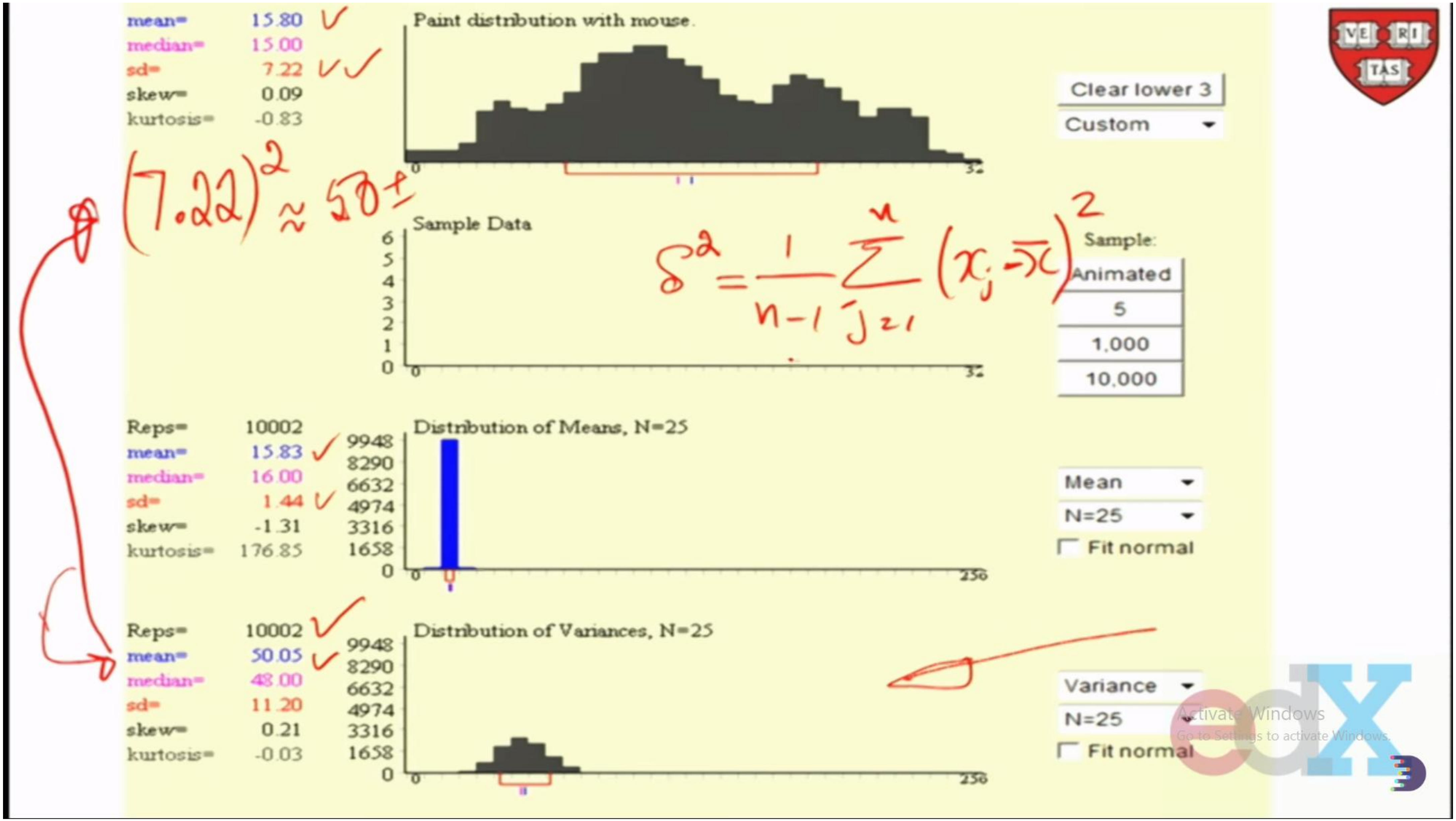

mean= 15.90
median= 15.00
sd= 7.45
skew= 0.14
kurtosis= -0.66
Paint distribution with mouse.
Clear lower 3
Custom
Sample Data
Sample:
Animated
5
1,000
10,000
Reps= 10002
mean= 15.89
median= 16.00
sd= 1.48
skew= -2.25
kurtosis= 144.78
Distribution of Means, N=25
Mean
N=25
Fit normal
Reps= 10002
mean= 53.30
median= 56.00
sd= 12.66
skew= 0.28
kurtosis= 0.16
Distribution of Variances, N=25
Variance
N=25
Fit normal
Activate Windows
Go to Settings to activate Windows.



mean= 15.80
median= 15.00
sd= 7.22
skew= 0.09
kurtosis= -0.83
Paint distribution with mouse.
Clear lower 3
Custom
Sample Data
Sample:
Animated
5
1,000
10,000
Reps= 10002
mean= 15.83
median= 16.00
sd= 1.44
skew= -1.31
kurtosis= 176.85
Distribution of Means, N=25
9948
8290
6632
4974
3316
1658
0
Mean
N=25
Fit normal
Reps= 10002
mean= 50.05
median= 48.00
sd= 11.20
skew= 0.21
kurtosis= -0.03
Distribution of Variances, N=25
Variance
N=25
Fit normal


mean= 15.80
median= 15.00
sd= 7.22
skew= 0.09
kurtosis= -0.83
Paint distribution with mouse.
(7.22)^2 ≈ 52±
Clear lower 3
Custom
Sample Data
Sample:
Animated
5
1,000
10,000
Reps= 10002
mean= 15.83
median= 16.00
sd= 1.44
skew= -1.31
kurtosis= 176.85
Distribution of Means, N=25
9948
8290
6632
4974
3316
1658
0
Mean
N=25
Fit normal
Reps= 10002
mean= 50.05
median= 48.00
sd= 11.20
skew= 0.21
kurtosis= -0.03
Distribution of Variances, N=25
Variance
N=25
Fit normal
Activate Windows
Go to Settings to activate Windows.


mean= 15.80
median= 15.00
sd= 7.22
skew= 0.09
kurtosis= -0.83
Paint distribution with mouse.
0
32
Clear lower 3
Custom
(7.22)² ≈ 52±
Sample Data
6
5
4
3
2
1
0
Sample:
Animated
5
1,000
10,000
Reps= 10002
mean= 15.83
median= 16.00
sd= 1.44
skew= -1.31
kurtosis= 176.85
Distribution of Means, N=25
9948
8290
6632
4974
3316
1658
0
256
Mean
N=25
Fit normal
Reps= 10002
mean= 50.05
median= 48.00
sd= 11.20
skew= 0.21
kurtosis= -0.03
Distribution of Variances, N=25
Variance
N=25
Fit normal
VE RI TAS
edX



mean= 15.80
median= 15.00
sd= 7.22
skew= 0.09
kurtosis= -0.83
Paint distribution with mouse.
(7.22)^2 ≈ 50 ±
Sample Data
s^2 = 1/(n-1) Σ_{j=1}^{n} (x_j - x̄)^2
Clear lower 3
Custom
Sample:
Animated
5
1,000
10,000
Reps= 10002
mean= 15.83
median= 16.00
sd= 1.44
skew= -1.31
kurtosis= 176.85
Distribution of Means, N=25
Mean
N=25
Fit normal
Reps= 10002
mean= 50.05
median= 48.00
sd= 11.20
skew= 0.21
kurtosis= -0.03
Distribution of Variances, N=25
Variance
N=25
Fit normal

What if σ is unknown?

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

What if σ is unknown?

$$t = \frac{\bar{X} - \mu}{s / \sqrt{n}}$$

Sample: size n

sample mean $\bar{X}$

sample standard deviation s

What if σ is unknown?

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$ has n-1 degrees of freedom

Sample: size n

sample mean $\bar{X}$

sample standard deviation s

What if σ is unknown?

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$ has n-1 degrees of freedom

Sample: size n

sample mean $\bar{X}$

sample standard deviation s

Population: X is approx. normal

mean μ

standard deviation σ

What if σ is unknown?

$$t = \frac{\bar{X} - \mu}{s / \sqrt{n}}$$ has n-1 degrees of freedom

Sample: size n

sample mean $\bar{X}$

sample standard deviation s

Population: X is approx. normal

mean μ

standard deviation σ

What if σ is unknown?

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$ has n-1 degrees of freedom

Sample: size n
sample mean $\bar{X}$
sample standard deviation s

William Sealy Gosset
1876 – 1937

Population: X is approx. normal
mean μ
standard deviation σ

What if σ is unknown?

Student's t

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$ has n-1 degrees of freedom

Sample: size n

sample mean $\bar{X}$

sample standard deviation s

William Sealy Gosset
1876 – 1937

Population: X is approx. normal

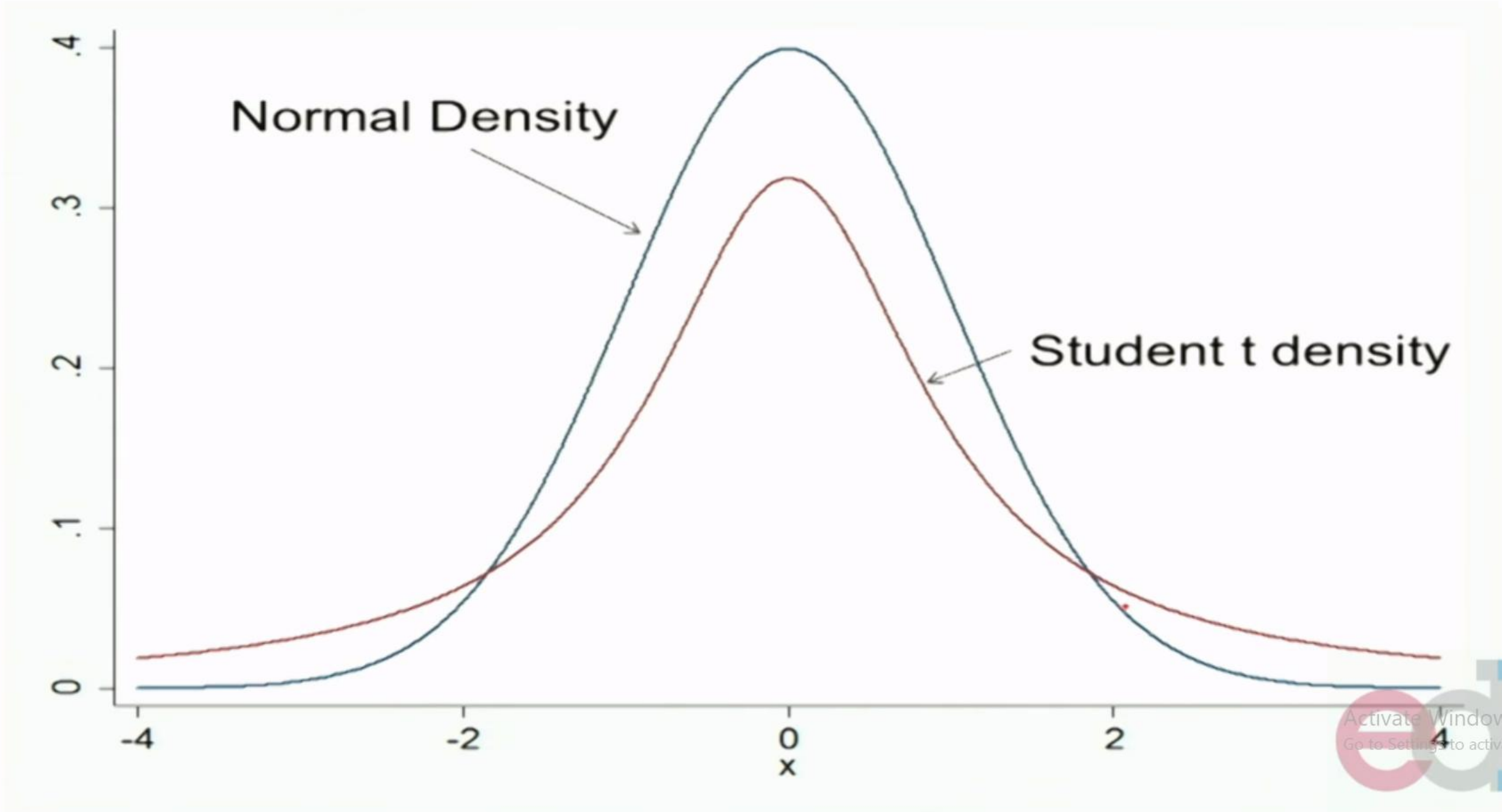
mean μ

standard deviation σ

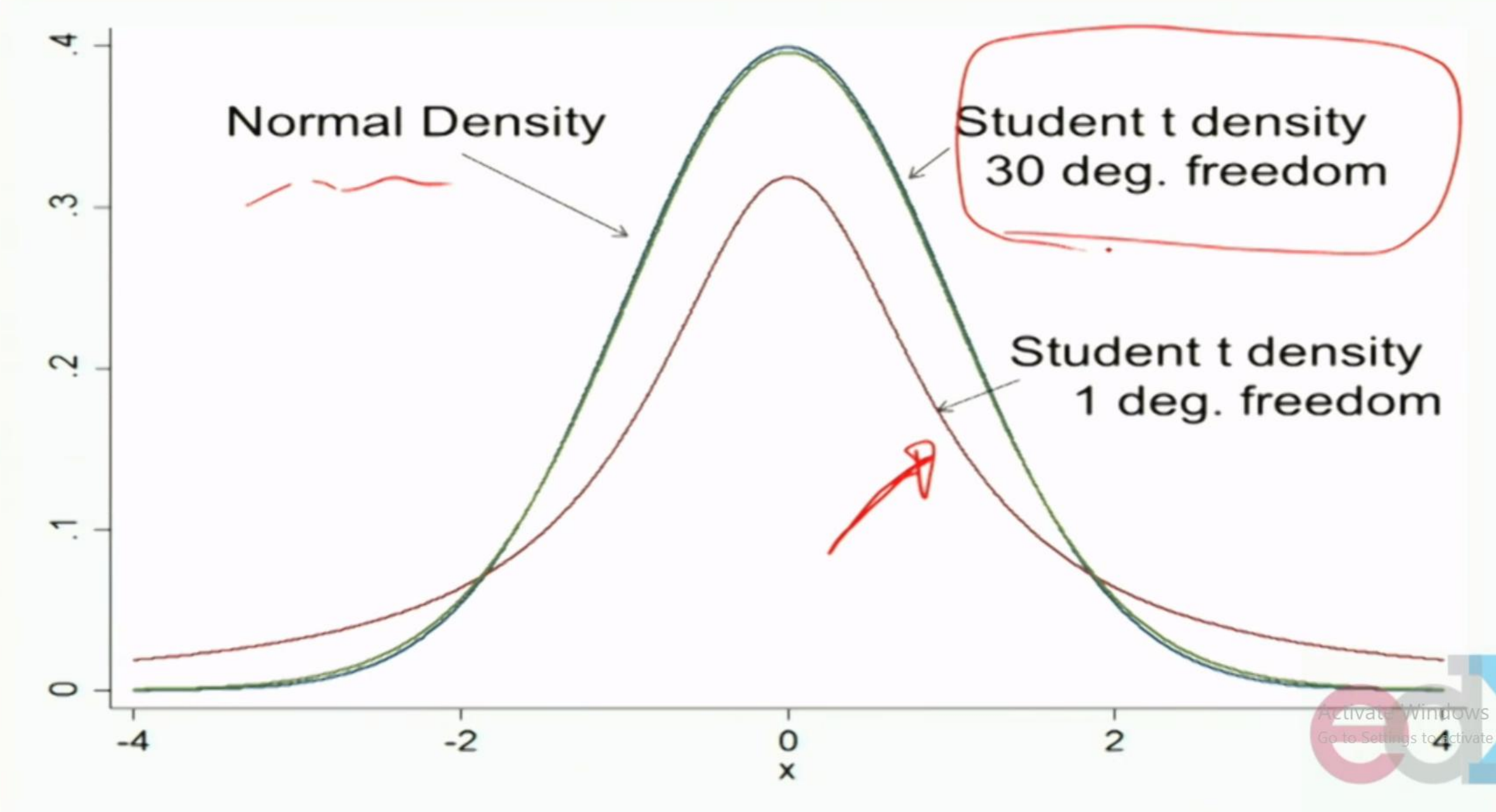
n fixed Std.dev.

$\bar{x}_1$	s_1	$\dfrac{\bar{x}_1 - \mu}{s_1/\sqrt{n}}$
$\bar{x}_2$	s_2	$\dfrac{\bar{x}_2 - \mu}{s_2/\sqrt{n}}$
$\bar{x}_3$	s_3	$\dfrac{\bar{x}_3 - \mu}{s_3/\sqrt{n}}$
$\vdots$	$\vdots$	$\vdots$
$\dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$		$\dfrac{\bar{X} - \mu}{S/\sqrt{n}}$

Normal and t densities



Normal and t densities



Hypotheticodeductive Method

Karl Popper's essence of the scientific method:

1. Set up falsifiable hypotheses
2. Test them

Conjectures and refutations: the growth of scientific knowledge.
NY Routledge & Kegan Paul, 1963

We know that total cholesterol levels in *our* Framingham population are distributed with mean $\mu = 237$ mg/100 ml and standard deviation $\sigma = 44.7$ mg/100ml.

We know that total cholesterol levels in *our* Framingham population are distributed with mean $\mu = 237$ mg/100 ml and standard deviation $\sigma = 44.7$ mg/100ml.

We have a sample of 49 total cholesterol levels and their average is:

$$\bar{x} = 230 \text{ mg} / 100 \text{ ml?}$$

We know that total cholesterol levels in *our* Framingham population are distributed with mean $\mu = 237$ mg/100 ml and standard deviation $\sigma = 44.7$ mg/100ml.

We have a sample of 49 total cholesterol levels and their average is:

$$\bar{x} = 230 \text{ mg} / 100 \text{ ml?}$$

Is it reasonable to assume that this is a sample from our population?

We know that total cholesterol levels in *our* Framingham population are distributed with mean μ=237 mg/100 ml and standard deviation σ = 44.7 mg/100ml.

We have a sample of 49 total cholesterol levels and their average is:

$$\bar{x} = 223 \text{ mg} / 100 \text{ ml?}$$

Is it reasonable to assume that this is a sample from our population?

We know that total cholesterol levels in *our* Framingham population are distributed with mean $\mu = 237$ mg/100 ml and standard deviation $\sigma = 44.7$ mg/100ml.

We have a sample of 49 total cholesterol levels and their average is:

$$\bar{x} = 215 \text{ mg}/100 \text{ ml?}$$

Is it reasonable to assume that this is a sample from our population?

Use of 95% confidence interval to infer value of μ ($\mu = 237?$)

$$\left(\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}\right) \rightarrow \left(\bar{X} \pm 1.96\frac{47.7}{\sqrt{49}}\right) \rightarrow (\bar{X} \pm 13.4)$$

has a 95% chance of including μ.

Use of 95% confidence interval to infer value of μ (μ = 237?)

$$\left(\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}\right) \rightarrow \left(\bar{X} \pm 1.96 \frac{47.7}{\sqrt{49}}\right) \rightarrow \boxed{(\bar{X} \pm 13.4)}$$

has a 95% chance of including μ.

If $\bar{X}$	95% confidence interval
230	(216.6, 243.4)

Use of 95% confidence interval to infer value of μ ($\mu = 237?$)

$$\left(\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}\right) \rightarrow \left(\bar{X} \pm 1.96 \frac{47.7}{\sqrt{49}}\right) \rightarrow \boxed{(\bar{X} \pm 13.4)}$$

has a 95% chance of including μ.

If $\bar{X}$	95% confidence interval
230	(216.6, 243.4)
223	(209.6, 236.4)

Use of 95% confidence interval to infer value of μ ($\mu = 237?$)

$$\left(\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}\right) \rightarrow \left(\bar{X} \pm 1.96\frac{47.7}{\sqrt{49}}\right) \rightarrow \boxed{(\bar{X} \pm 13.4)}$$

has a 95% chance of including μ.

If $\bar{X}$	95% confidence interval
230	(216.6, 243.4)
223	(209.6, 236.4)

Use of 95% confidence interval to infer value of μ (μ = 237?)

$$\left(\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}\right) \rightarrow \left(\bar{X} \pm 1.96\frac{47.7}{\sqrt{49}}\right) \rightarrow \boxed{(\bar{X} \pm 13.4)}$$

has a 95% chance of including μ.

If $\bar{X}$	95% confidence interval
230	(216.6, 243.4) ✓
223	(209.6, 236.4) ✗
215	(201.6, 228.4)

Alternatively,

IF

μ= 237 and σ = 47.7 and we take a sample of size n=49 from this population, then the Central Limit Theorem tells us that the sample mean is approximately normally distributed with mean μ = 237 and std. dev. 47.7/7;

i.e.

Alternatively,

IF

$\mu = 237$ and $\sigma = 47.7$ and we take a sample of size $n=49$ from this population, then the Central Limit Theorem tells us that the sample mean is approximately normally distributed with mean $\mu = 237$ and std. dev. $47.7/7$;

i.e.

$$\Pr\left\{-1.96 \le \boxed{\frac{\bar{X} - 237}{47.7/7}} \le 1.96\right\} = 0.95$$

$$\Pr\left\{223.6 \le \bar{X} \le 250.4\right\} = 0.95$$

Alternatively,

IF

μ= 237 and σ = 47.7 and we take a sample of size n=49 from this population, then the Central Limit Theorem tells us that the sample mean is approximately normally distributed with mean μ = 237 and std. dev. 47.7/7;

i.e.

$$\Pr\left\{-1.96 \le \frac{\bar{X} - 237}{47.7/7} \le 1.96\right\} = 0.95$$

$$\Pr\left\{223.6 \le \bar{X} \le 250.4\right\} = 0.95$$

The two approaches are consistent.

Individual on trial.
Did he commit the crime?

Individual on trial.

Did he commit the crime?

Evidence

Trial

Person

Innocent

Guilty

Individual on trial.

Did he commit the crime?

Evidence — Trial

Jury	Person	
	Innocent	Guilty
Not Guilty		
Guilty		

Individual on trial.
Did he commit the crime?

Evidence　　　　Trial

Jury	Person	
	Innocent	Guilty
Not Guilty	✓	
Guilty		✓

Individual on trial.
Did he commit the crime?

Evidence Trial

Jury	Person	
	Innocent	Guilty
Not Guilty	✓	x
Guilty	x	✓

Test of Hypothesis that $\mu = \mu_0$?

Evidence Trial

Jury	Person	
	Innocent	Guilty
Not Guilty	✓	x
Guilty	x	✓

Test of Hypothesis that $\mu = \mu_0$?

Sample **Trial**

Jury	Person	
	Innocent	Guilty
Not Guilty	√	x
Guilty	x	√

Test of Hypothesis that $\mu = \mu_0$?

Sample Analysis

Jury	Person	
	Innocent	Guilty
Not Guilty	✓	x
Guilty	x	✓

Test of Hypothesis that $\mu = \mu_0$?

Sample **Analysis**

Jury	Population	
	Innocent	Guilty
Not Guilty	√	x
Guilty	x	√

Test of Hypothesis that $\mu = \mu_0$?

Sample **Analysis**

Jury	Population	
	$\mu = \mu_0$	Guilty
Not Guilty	✓	x
Guilty	x	✓

Test of Hypothesis that $\mu = \mu_0$?

Sample Analysis

	Population	
	$\mu = \mu_0$	$\mu \neq \mu_0$
Not Guilty	✓	x
Guilty	x	✓

Test of Hypothesis that $\mu = \mu_0$?

Sample Analysis

Us	Population	
	$\mu = \mu_0$	$\mu \neq \mu_0$
Not reject	✓	x
Guilty	x	✓

Test of Hypothesis that $\mu = \mu_0$?

Sample Analysis

Us	Population	
	$\mu = \mu_0$	$\mu \neq \mu_0$
Not reject	$\checkmark$	x
Reject	x	$\checkmark$

Test of Hypothesis that $\mu = \mu_0$?

Sample **Analysis**

Us	Population	
	$\mu = \mu_0$	$\mu \neq \mu_0$
Not reject	✓	Type II
Reject	Type I	✓

Probability of Type I error is α
i.e. the probability of rejecting the null hypothesis when it is true.

Probability of Type II error is β
i.e the probability of not rejecting the null hypothesis when it is false.

$1-\beta$ is the *power* of the test.

Recap: Hypothesis testing about μ :

1° Hypothesize a value (μ_0)

2° Take a random sample (n).

3° Is it **likely** that the sample came from a population with mean μ_0 ($\alpha = 0.05$) ?

Recap: Hypothesis testing about μ :

1° Hypothesize a value (μ_0)

2° Take a random sample (n).

3° Is it likely that the sample came from a population with mean μ_0 ($\alpha = 0.05$) ?

Recap: Hypothesis testing about μ :

1° Hypothesize a value (μ_0)

2° Take a random sample (n).

3° Is it likely that the sample came from a population with mean μ_0 ($\alpha = 0.05$) ?

Decide on statistic: $\bar{X}$

Determine which values of $\bar{X}$ are consonant with the hypothesis that $\mu = \mu_0$ and which ones are not.

Look at $\dfrac{\bar{X} - \mu_0}{\sigma}$ and decide.

One sided or two?

Decide on statistic: $\bar{X}$

Determine which values of $\bar{X}$ are consonant with the hypothesis that $\mu = \mu_0$ and which ones are not.

Look at $\dfrac{\bar{X} - \mu_0}{\sigma}$ and decide.

One sided or two?

Need to set up 2 hypotheses to cover *all* possibilities for μ.

Choose one of three possibilities:

Two-sided	$H_0 : \mu = \mu_0$
	$H_A : \mu \neq \mu_0$

Need to set up 2 hypotheses to cover *all* possibilities for μ.

Choose one of three possibilities:

Two-sided	$H_0 : \mu = \mu_0$
	$H_A : \mu \neq \mu_0$
One-sided	$H_0 : \mu \geq \mu_0$
	$H_A : \mu < \mu_0$

Need to set up 2 hypotheses to cover *all* possibilities for μ.

Choose one of three possibilities:

Two-sided	$H_0 : \mu = \mu_0$
	$H_A : \mu \neq \mu_0$
One-sided	$H_0 : \mu \geq \mu_0$
	$H_A : \mu < \mu_0$
One-sided	$H_0 : \mu \leq \mu_0$
	$H_A : \mu > \mu_0$

$$H_0 : \mu = \mu_0$$
$$H_A : \mu \neq \mu_0$$

Look at

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

and reject H_0 if Z is too large, + or –, e.g.

Reject if Z is >1.96 or <-1.96, then

Pr(reject H_0 when true) = 0.05

$H_0 : \mu = \mu_0$
$H_A : \mu \neq \mu_0$

Look at

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

and reject H_0 if Z is too large, + or –, e.g.

Reject if Z is $\boxed{>1.96}$ or $\boxed{<-1.96}$, then

Pr(reject H_0 when true) = 0.05

$H_0 : \mu = 237$
$H_A : \mu \neq 237$

$\sigma = 47.7$ mg/100ml

Sample of 49 non-hypertensives have:

$\bar{x} = 221.9$ mg/100ml

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{221.9 - 237}{47.7 / \sqrt{49}} = -2.37$$

$H_0 : \mu = 237$
$H_A : \mu \neq 237$

$\sigma = 47.7$ mg/100ml

Sample of 49 non-hypertensives have:

$\overline{x} = 221.9$ mg/100ml

$$z = \frac{\overline{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{221.9 - 237}{47.7 / \sqrt{49}} = -2.37$$

$H_0 : \mu = 237$
$H_A : \mu \neq 237$

$\sigma = 47.7$ mg/100ml

Sample of 49 non-hypertensives have:

$\bar{x} = 221.9$ mg/100ml

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{221.9 - 237}{47.7 / \sqrt{49}} = -2.37$$

$H_0 : \mu = 237$
$H_A : \mu \neq 237$

$\sigma = 47.7$ mg/100ml

Sample of 49 non-hypertensives have:

$\bar{x} = 221.9$ mg/100ml

$$\boxed{z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}} = \frac{221.9 - 237}{47.7 / \sqrt{49}} = -2.37$$

So reject the null hypothesis.

```
. set seed 725764662

. drop if hyperten==1
(3252 observations deleted)

. sample 49 , count
(1133 observations deleted)

. mean totchol1

Mean estimation                    Number of obs   =       49

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
    totchol1 |   221.8776   4.614348      212.5998    231.1553
--------------------------------------------------------------

. di (221.8776-237)/(44.7/7)
-2.3681611
```

```
. set seed 725764662

. drop if hyperten==1
(3252 observations deleted)

. sample 49 , count
(1133 observations deleted)

. mean totchol1

Mean estimation                    Number of obs     =        49
```

t instead of Z

	Mean	Std. Err.	[95% Conf.	Interval]
totchol1	221.8776	4.614348	212.5998	231.1553

```
. di (221.8776-237)/(44.7/7)
-2.3681611
```

P-value

Some prefer to quote the p-value. The p-value answers the question, "What is the probability of getting as large, or larger, a discrepancy?" $(\mu - \bar{X})$

$$\begin{aligned} \Pr(Z > 2.37 \text{ or } Z < -2.37) &= 2\Pr(Z > 2.37) \\ &= 2 \times 0.0222 \\ &= 0.044 \end{aligned}$$

Stata:

```
. di normal(-2.0106348)
.02218202
```

P-value $= 0.044 < 0.05$

Some prefer to quote the p-value. The p-value answers the question, "What is the probability of getting as large, or larger, a discrepancy?" $(\mu - \bar{X})$

$$
\begin{aligned}
\Pr(Z > 2.37 \text{ or } Z < -2.37) &= 2\Pr(Z > 2.37) \\
&= 2 \times 0.0222 \\
&= \boxed{0.044}
\end{aligned}
$$

Stata:

```
. di normal(-2.0106348)
.02218202
```

$H_0 : \mu \leq 9.7$
$H_A : \mu > 9.7$

Blood glucose level of healthy persons has μ = 9.7 mmol/L and σ = 2.0 mmol/L

Sample of 64 diabetics yields

$\bar{x} = 13.1$ mmol/L

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = 13.60$$

p-value << 0.001

Blood glucose level of healthy persons has $\mu = 9.7$ mmol/L and $\sigma = 2.0$ mmol/L

$H_0 : \mu \leq 9.7$
$H_A : \mu > 9.7$

Sample of 64 diabetics yields

$\bar{x} = 13.1$ mmol/L

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = 13.60$$

p-value << 0.001

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

$$\alpha = 0.05$$

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

$\alpha = 0.05$

$H_0 : \mu = \mu_0$ $H_A : \mu \neq \mu_0$	Reject if $\lvert z \rvert > 1.96$
$H_0 : \mu \geq \mu_0$ $H_A : \mu < \mu_0$	Reject if $z < -1.645$
$H_0 : \mu \leq \mu_0$ $H_A : \mu > \mu_0$	Reject if $z > 1.645$