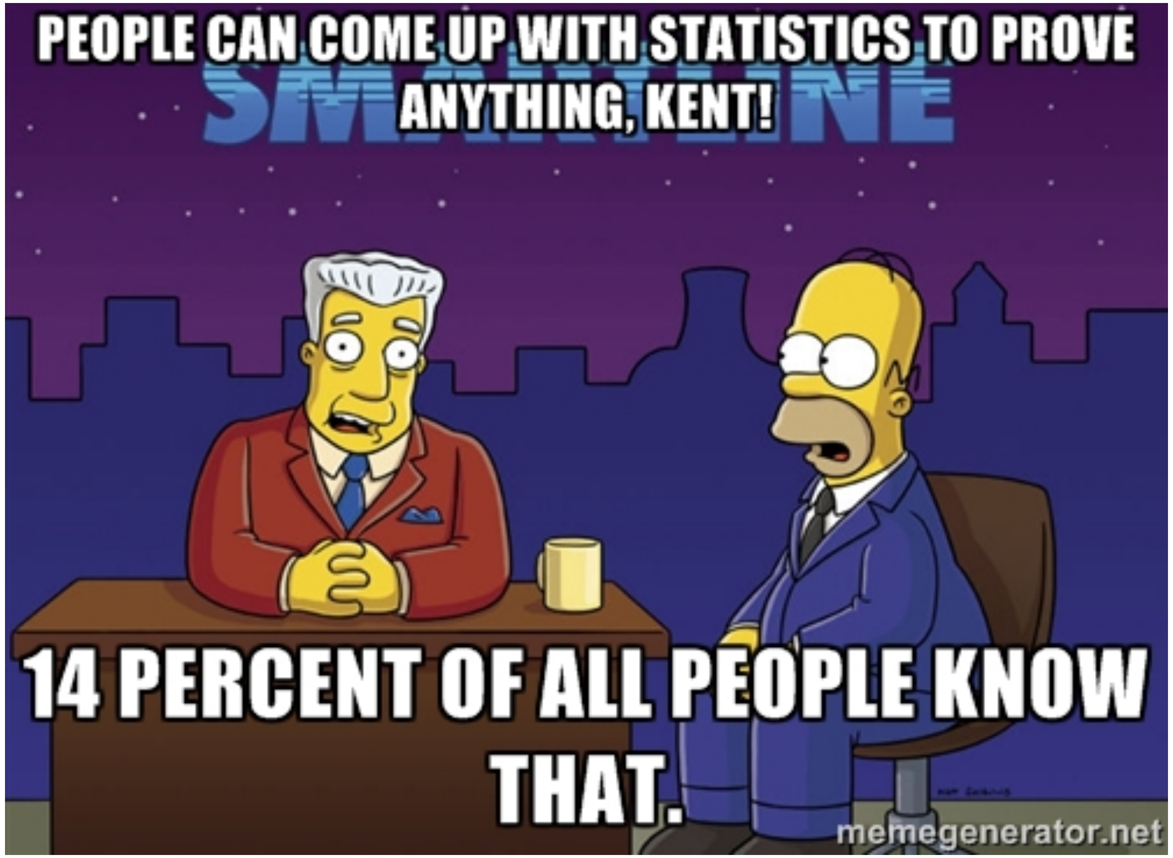
Hypothesis Testing

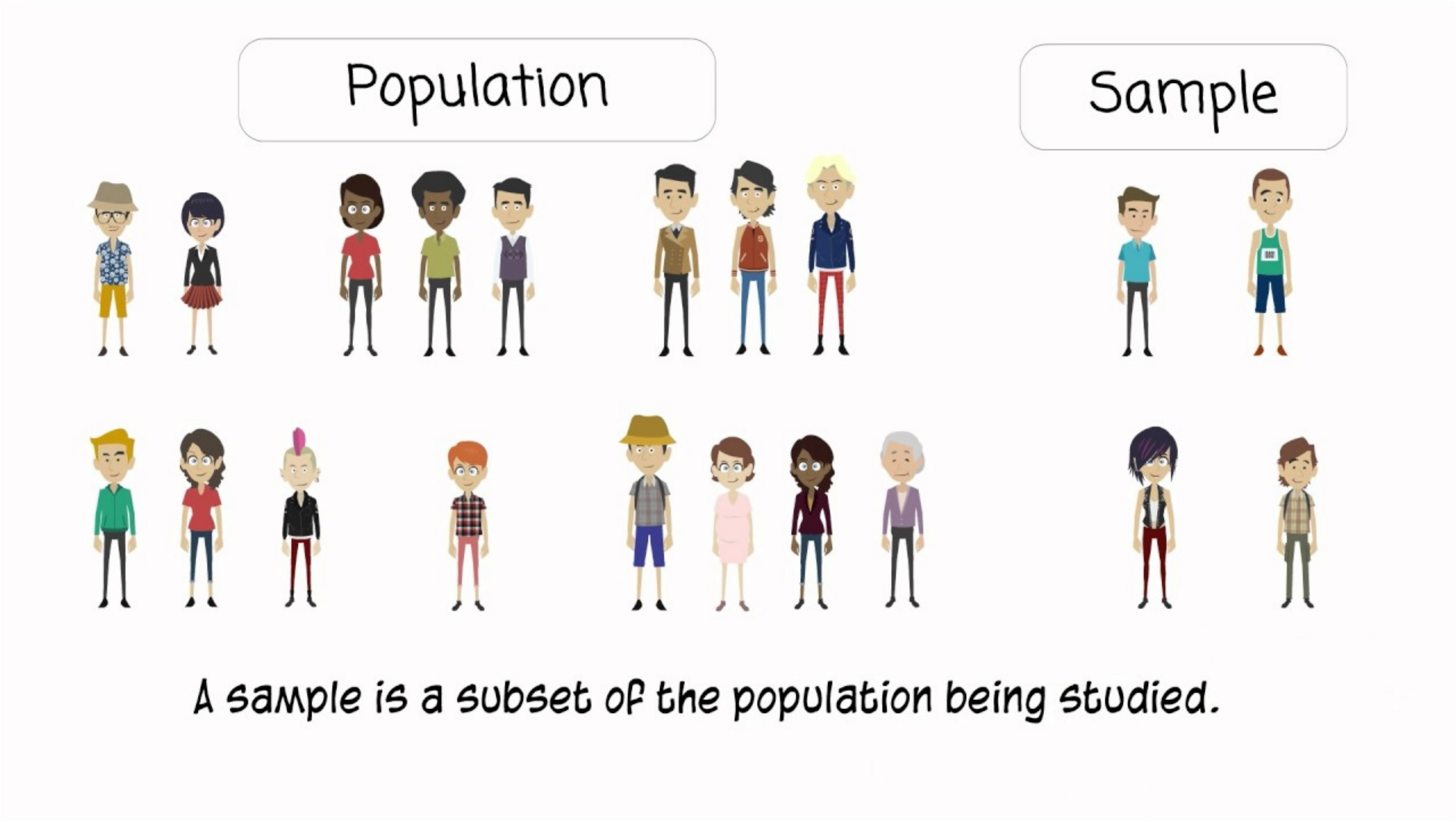

PEOPLE CAN COME UP WITH STATISTICS TO PROVE ANYTHING, KENT!
SMARTLINE
14 PERCENT OF ALL PEOPLE KNOW THAT.
memegenerator.net

Statistics = Use random samples to make confident statements about entire populations. Intelligent guesses/speculation.

Sample, shape, location, and spread

- Sample = make sure it's random, handle missing data (mcar, mar, nmar), imputation methods. NMAR!
- Shape = Is the data skewed, normal, or flat? If normal then we can use statistical analysis for normal distributions
- Location = Where does the data accumulate? Is it skewed, if so the median tells a better story.
- Spread = How much does the data differ? Standard deviations!

Samples give statistics, and populations (which we may never know give parameters)

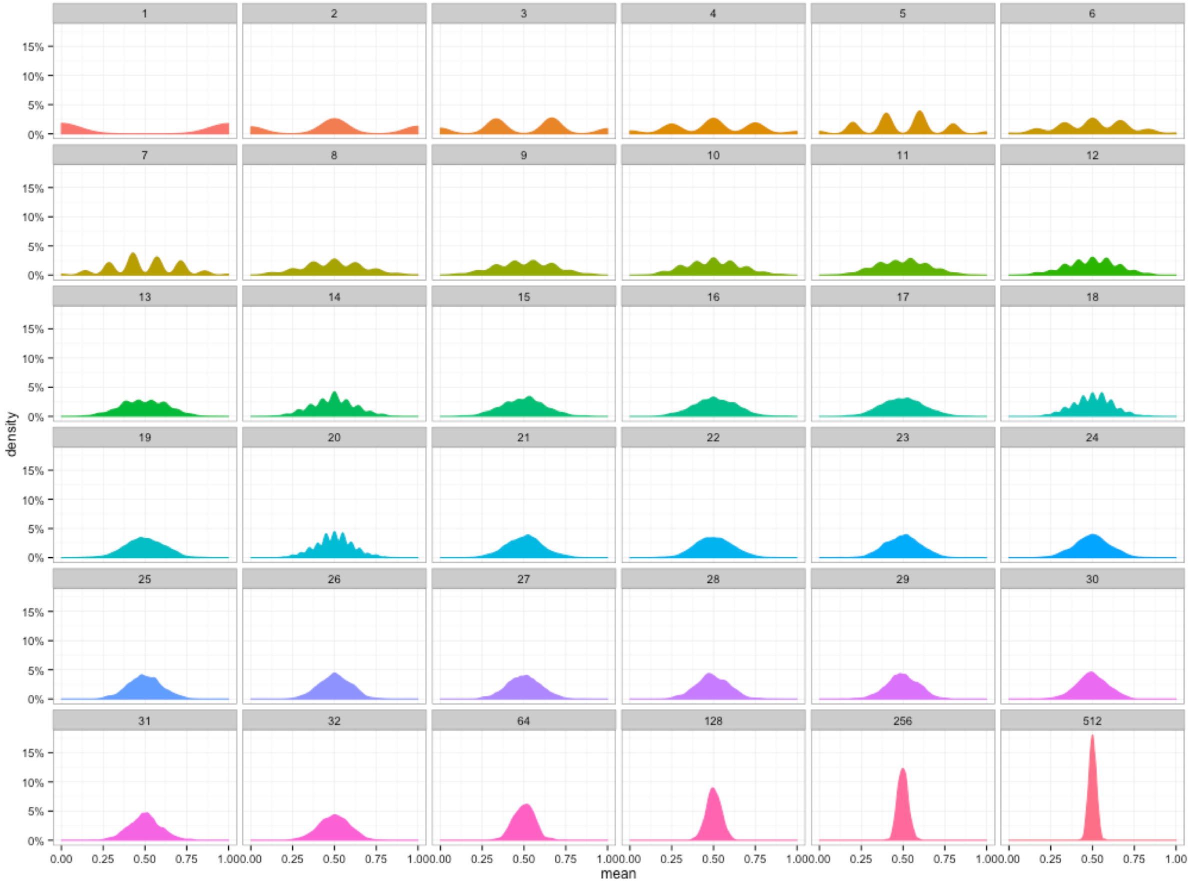


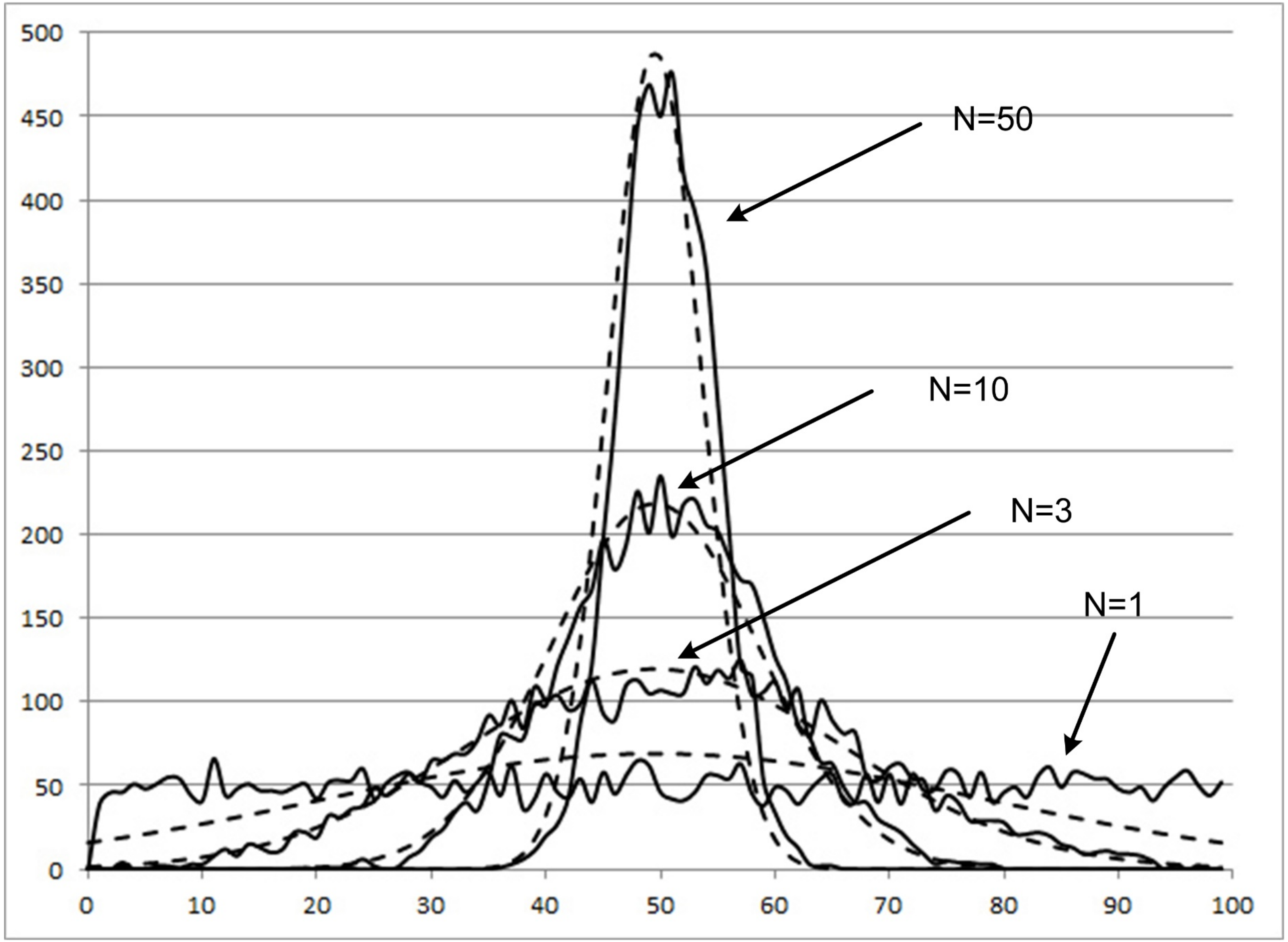
The Central Limit Theorem

If we:

1. Sample randomly
2. And use the averages of the sampled data

In the long term an n goes ballistic, the distribution will be normal and narrow. Why?



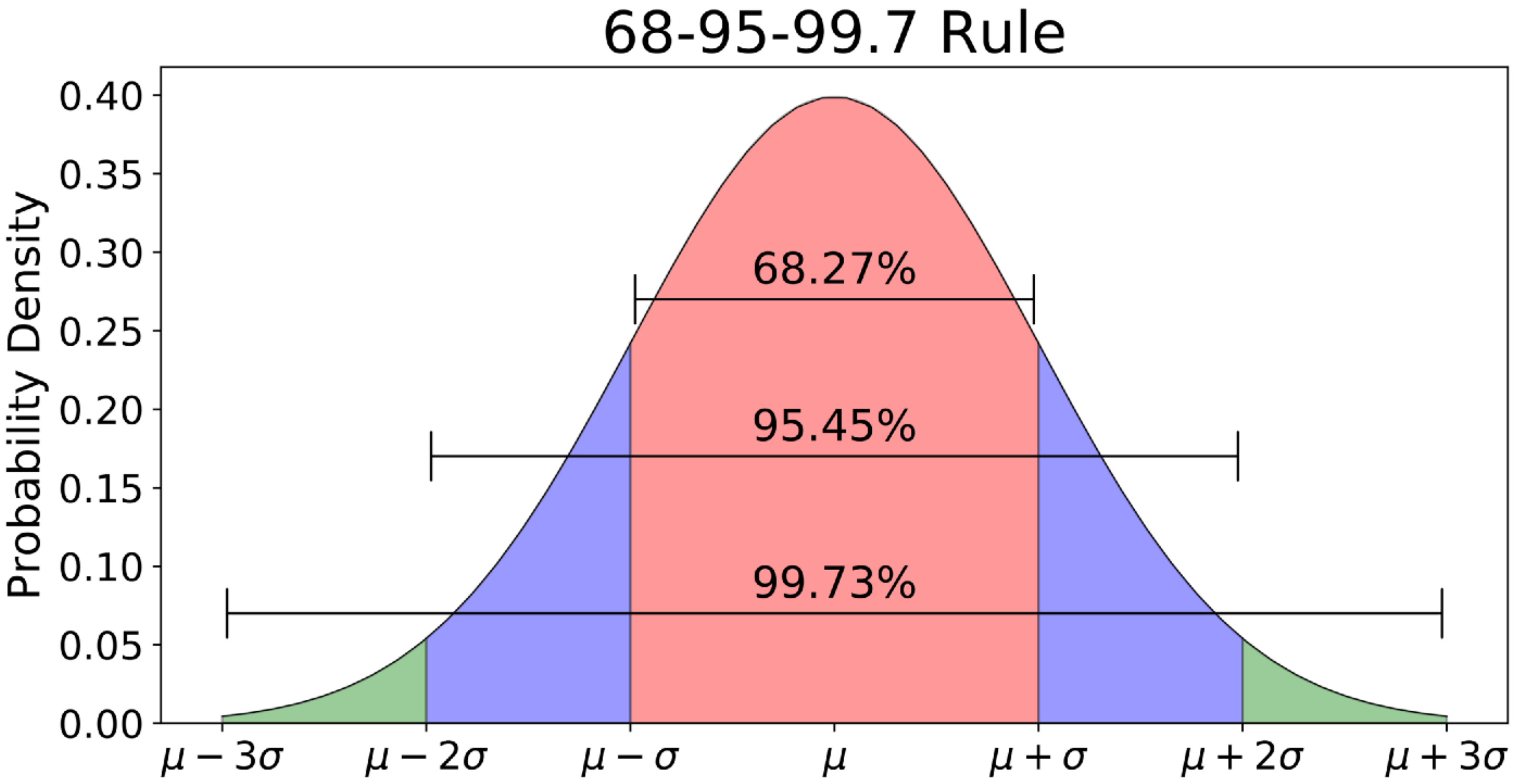

500
450
400
350
300
250
200
150
100
50
0
0
10
20
30
40
50
60
70
80
90
100
N=50
N=10
N=3
N=1

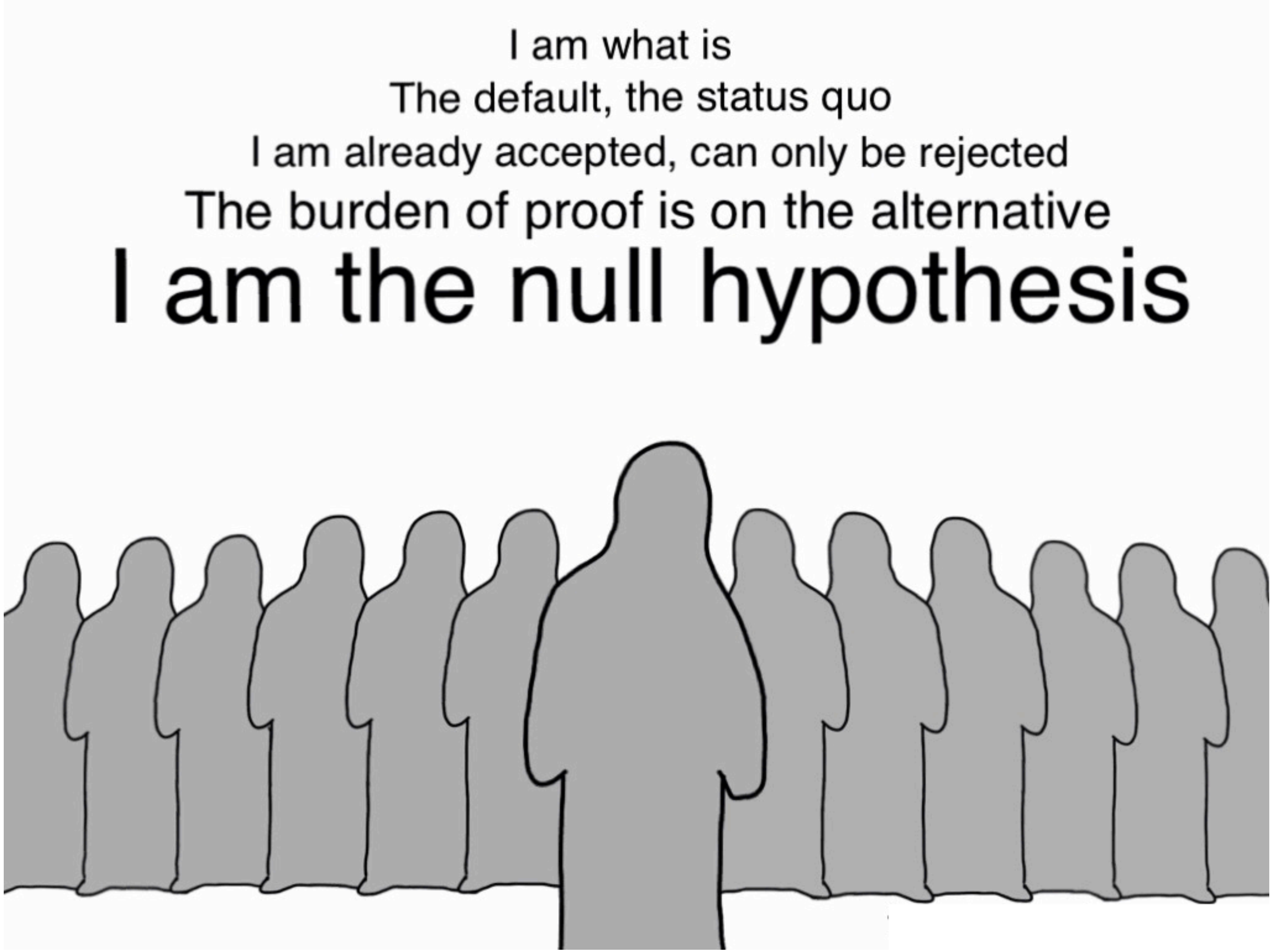
We're all
[redacted] until
proven [redacted].

How do we prove our statistical claims?

- Hypothesis Testing! The point of hypothesis testing is to make sure we don't jump to bad conclusions. Conclusions can be confusing (xylitol vs. fluoride). We are inherently speculating albeit rigorously. So we try to control the guessing by being conservative and use innocent until proven guilty.

Standard deviations and probability of the population mean





I am what is

The default, the status quo

I am already accepted, can only be rejected

The burden of proof is on the alternative

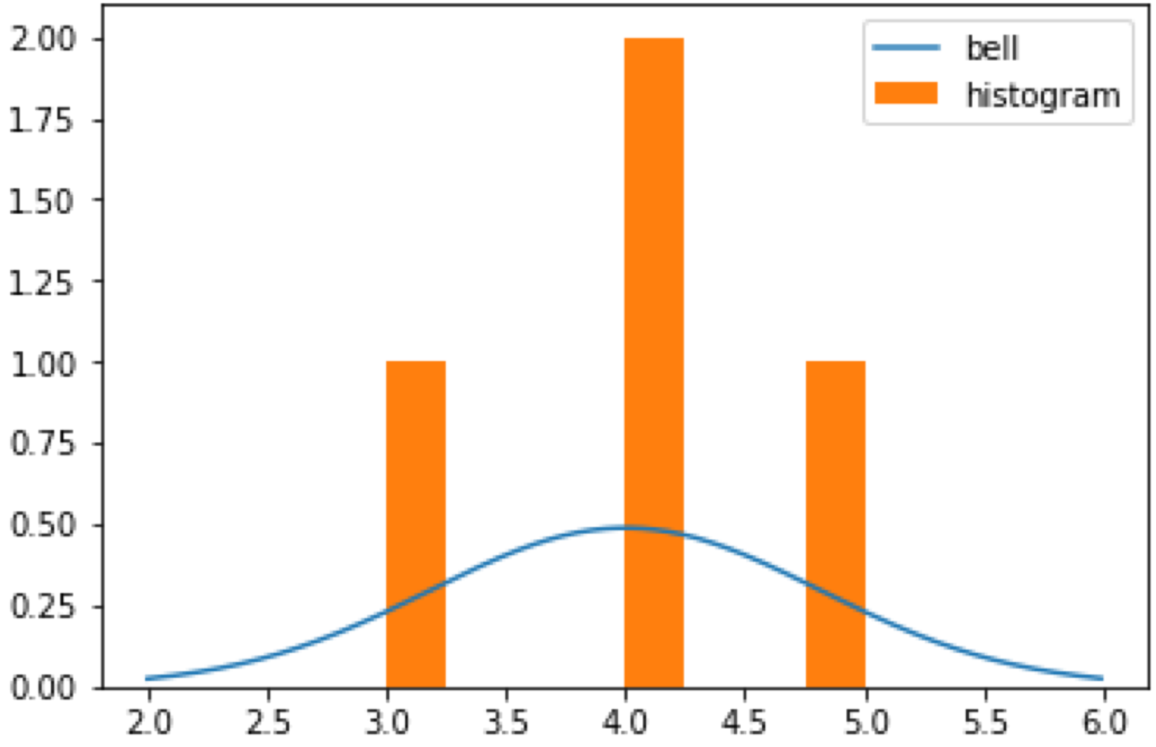
I am the null hypothesis

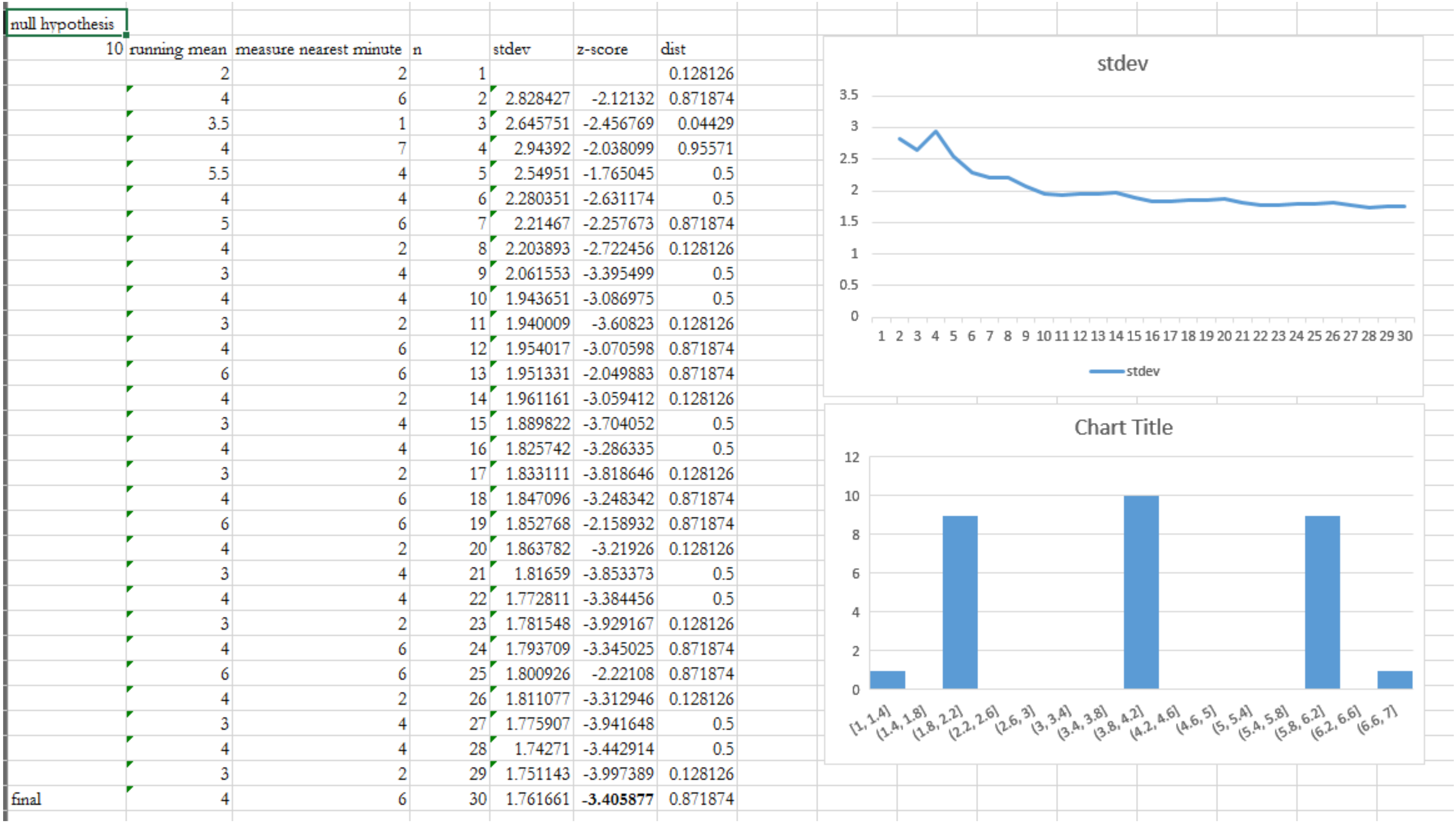
The alternative hypothesis tries to nullify the ghosts!

An example

- Ghosts say the best server config for a query we run is X config and the best results are returned in 10 minutes.
- Alternative hypothesis: A different server config is better and returns results in 4 minutes.
- Can we reject the null hypothesis and kill the ghosts?

Collect data, randomly and take averages





null hypothesis						
10	running mean	measure nearest minute	n	stdev	z-score	dist
	2	2	1			0.128126
	4	6	2	2.828427	-2.12132	0.871874
	3.5	1	3	2.645751	-2.456769	0.04429
	4	7	4	2.94392	-2.038099	0.95571
	5.5	4	5	2.54951	-1.765045	0.5
	4	4	6	2.280351	-2.631174	0.5
	5	6	7	2.21467	-2.257673	0.871874
	4	2	8	2.203893	-2.722456	0.128126
	3	4	9	2.061553	-3.395499	0.5
	4	4	10	1.943651	-3.086975	0.5
	3	2	11	1.940009	-3.60823	0.128126
	4	6	12	1.954017	-3.070598	0.871874
	6	6	13	1.951331	-2.049883	0.871874
	4	2	14	1.961161	-3.059412	0.128126
	3	4	15	1.889822	-3.704052	0.5
	4	4	16	1.825742	-3.286335	0.5
	3	2	17	1.833111	-3.818646	0.128126
	4	6	18	1.847096	-3.248342	0.871874
	6	6	19	1.852768	-2.158932	0.871874
	4	2	20	1.863782	-3.21926	0.128126
	3	4	21	1.81659	-3.853373	0.5
	4	4	22	1.772811	-3.384456	0.5
	3	2	23	1.781548	-3.929167	0.128126
	4	6	24	1.793709	-3.345025	0.871874
	6	6	25	1.800926	-2.22108	0.871874
	4	2	26	1.811077	-3.312946	0.128126
	3	4	27	1.775907	-3.941648	0.5
	4	4	28	1.74271	-3.442914	0.5
	3	2	29	1.751143	-3.997389	0.128126
final	4	6	30	1.761661	**-3.405877**	0.871874

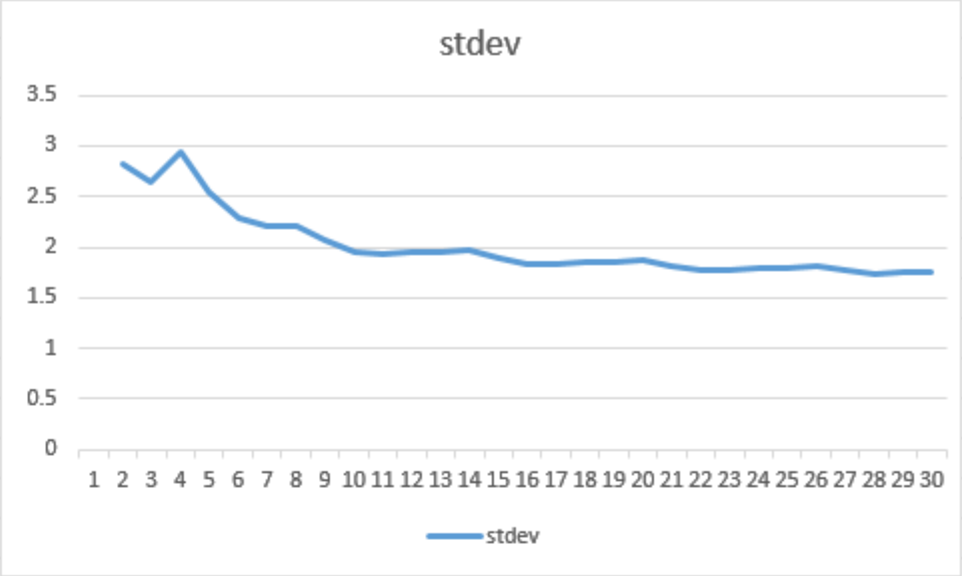
Warning

- It's important to note that statistics and inferences about "populations" are always approximations. We aren't using probabilities when we do have the entire population e.g., the entire Data Science class is our population. Here I can get the population mean etc... No guess work is necessary. Now what if we said all data science students in the world currently? That’s much harder if not impossible. Inference time!

Sample Size?

- THE MATH IN STATISTICS IS BUILT AROUND YOU NOT KNOWING THE POPULATION SIZE. WHICH IS WHY WE CAN PICK N WITHOUT KNOWING IT AND N IS REALLY A FUNCTION OF TIME+COST.

null hypothesis						
10	running mean	measure nearest minute	n	stdev	z-score	dist
	2	2	1			0.128126
	4	6	2	2.828427	-2.12132	0.871874
	3.5	1	3	2.645751	-2.456769	0.04429
	4	7	4	2.94392	-2.038099	0.95571
	5.5	4	5	2.54951	-1.765045	0.5
	4	4	6	2.280351	-2.631174	0.5
	5	6	7	2.21467	-2.257673	0.871874
	4	2	8	2.203893	-2.722456	0.128126
	3	4	9	2.061553	-3.395499	0.5
	4	4	10	1.943651	-3.086975	0.5
	3	2	11	1.940009	-3.60823	0.128126
	4	6	12	1.954017	-3.070598	0.871874
	6	6	13	1.951331	-2.049883	0.871874
	4	2	14	1.961161	-3.059412	0.128126
	3	4	15	1.889822	-3.704052	0.5
	4	4	16	1.825742	-3.286335	0.5
	3	2	17	1.833111	-3.818646	0.128126
	4	6	18	1.847096	-3.248342	0.871874
	6	6	19	1.852768	-2.158932	0.871874
	4	2	20	1.863782	-3.21926	0.128126
	3	4	21	1.81659	-3.853373	0.5
	4	4	22	1.772811	-3.384456	0.5
	3	2	23	1.781548	-3.929167	0.128126
	4	6	24	1.793709	-3.345025	0.871874
	6	6	25	1.800926	-2.22108	0.871874
	4	2	26	1.811077	-3.312946	0.128126
	3	4	27	1.775907	-3.941648	0.5
	4	4	28	1.74271	-3.442914	0.5
	3	2	29	1.751143	-3.997389	0.128126
final	4	6	30	1.761661	**-3.405877**	0.871874


stdev
stdev

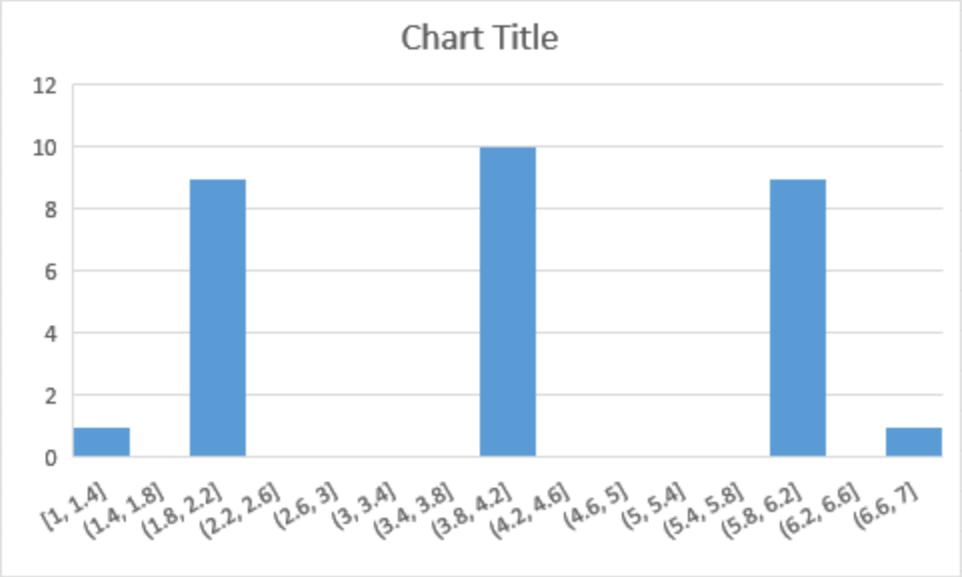


Chart Title
[1, 1.4]
(1.4, 1.8]
(1.8, 2.2]
(2.2, 2.6]
(2.6, 3]
(3, 3.4]
(3.4, 3.8]
(3.8, 4.2]
(4.2, 4.6]
(4.6, 5]
(5, 5.4]
(5.4, 5.8]
(5.8, 6.2]
(6.2, 6.6]
(6.6, 7]