Course notes: inferential statistics

Probability theory

Distributions

INFERENTIAL STATISTICS

Distributions

Definition

In statistics, when we talk about distributions we usually mean probability distributions.

Definition (informal): A distribution is a function that shows the possible values for a variable and how often they occur.

Definition (Wikipedia): In probability theory and statistics, a probability distribution is a mathematical function that, stated in simple terms, can be thought of as providing the probabilities of occurrence of different possible outcomes in an experiment.

Examples: Normal distribution, Student's T distribution, Poisson distribution, Uniform distribution, Binomial distribution

Graphical representation

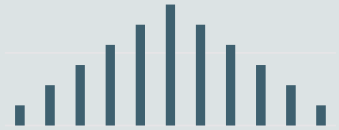
It is a common mistake to believe that the distribution is the graph. In fact the distribution is the 'rule' that determines how values are positioned in relation to each other.

Very often, we use a graph to visualize the data. Since different distributions have a particular graphical representation, statisticians like to plot them.

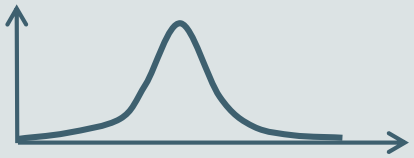
Examples:

Uniform distribution

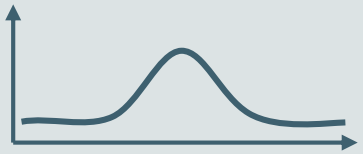
Binomial distribution



Normal distribution



Student's T distribution



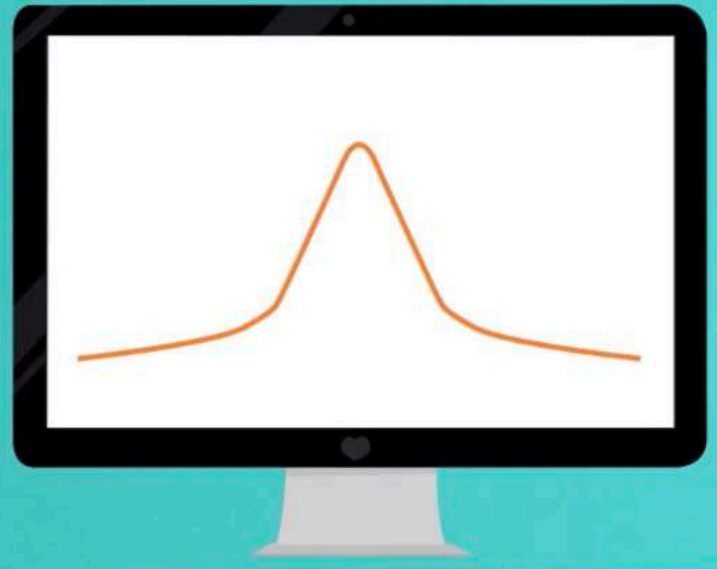
IN STATISTICS

DISTRIBUTION

=

PROBABILITY DISTRIBUTION

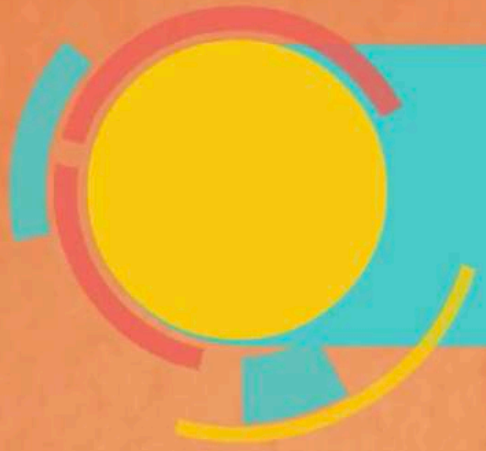
DISTRIBUTION = PROBABILITY DISTRIBUTION



NORMAL

BINOMIAL

UNIFORM



DEFINITION

A distribution is a function that shows the possible values for a variable and how often they occur.

ROLLING A DIE

DISCRETE UNIFORM DISTRIBUTION

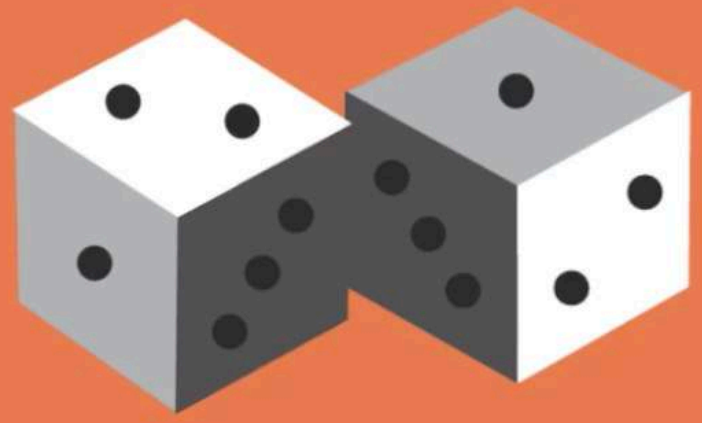
Fullscreen

OUTCOME	PROBABILITY
1; 2; 3; 4; 5; 6	0.17
All else	0

365 DataScience

ROLLING TWO DICE

OUTCOME	PROBABILITY
2	0.03
3	0.06
4	0.08
5	0.11
6	0.14
7	0.17
8	0.14
9	0.11
10	0.08
11	0.06
12	0.03
All else	0



SUM OF THE PROBABILITIES =

1 | 100%

=> we have exhausted all possibilities

NORMAL

STUDENT'S T

REASONS

- They approximate a wide variety of random variables
- Distributions of sample means with large enough sample sizes could be approximated to normal
- All computable statics are elegant
- Decisions based on normal distribution insights have a good track record

The Normal Distribution

The Normal distribution is also known as Gaussian distribution or the Bell curve. It is one of the most common distributions due to the following reasons:

- It approximates a wide variety of random variables
- Distributions of sample means with large enough samples sizes could be approximated to normal
- All computable statistics are elegant
- Heavily used in regression analysis
- Good track record

Examples:

- Biology. Most biological measures are normally distributed, such as: height; length of arms, legs, nails; blood pressure; thickness of tree barks, etc.
- IQ tests
- Stock market information

$$N \sim (\mu, \sigma^2)$$

N stands for normal;
$\sim$ stands for a distribution;
μ is the mean;
σ^2 is the variance.

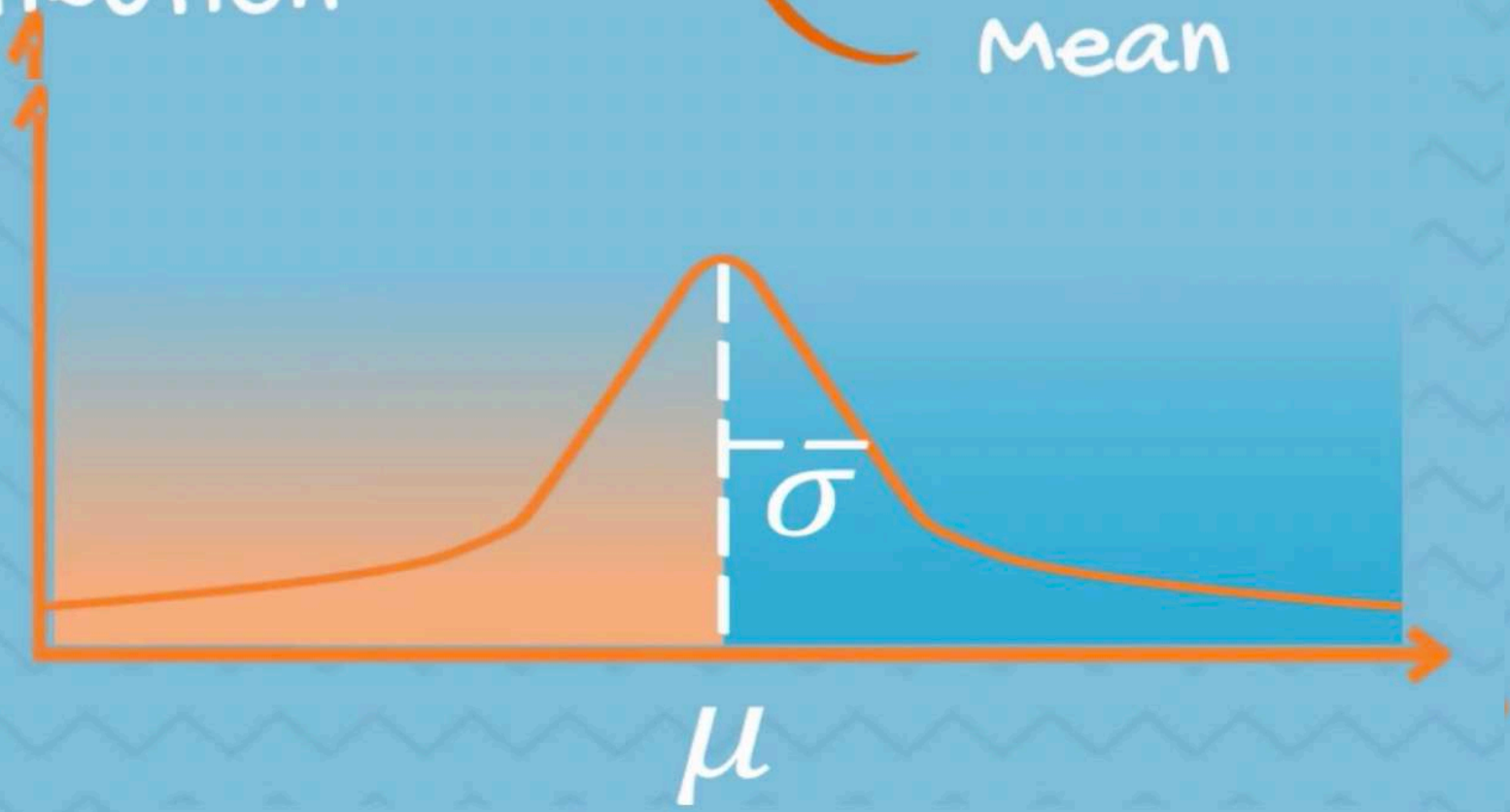
GAUSSIAN DISTRIBUTION

Normal → $N \sim (\mu, \sigma^2)$

Distribution

Mean

Variance



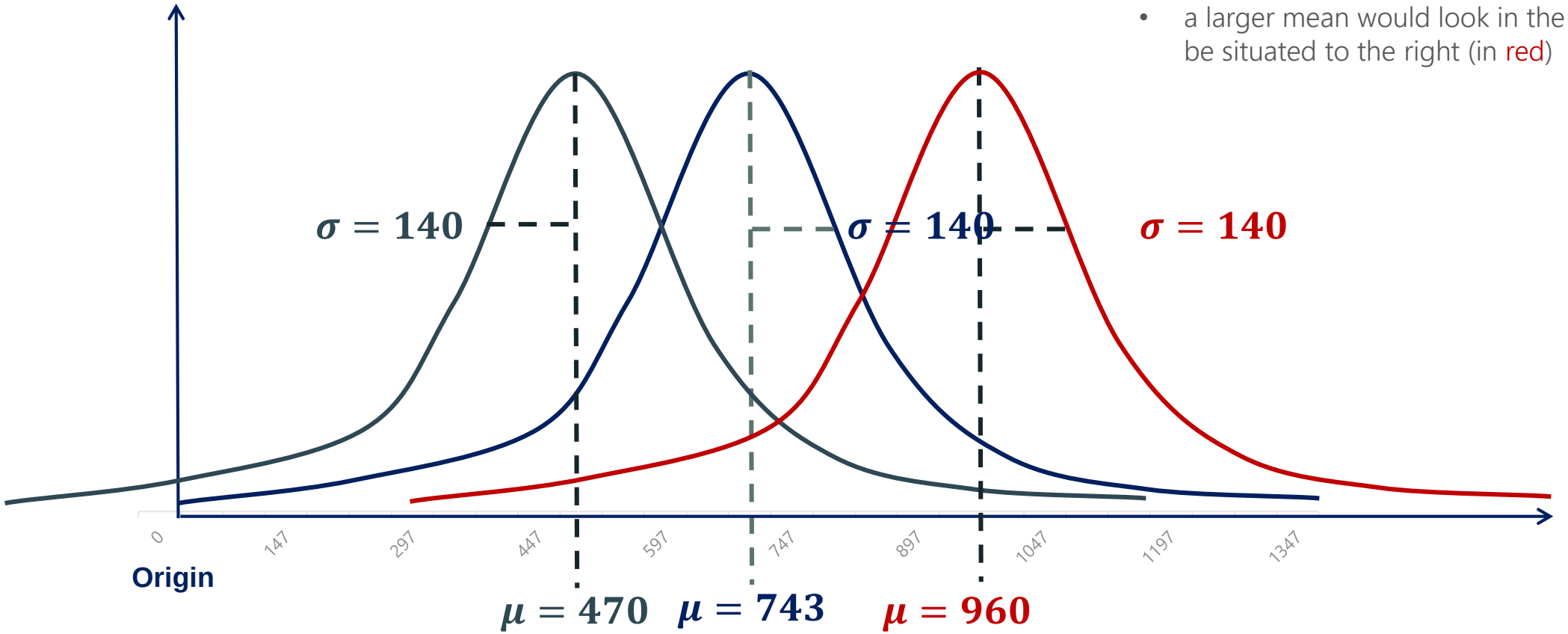
mean = median = mode

The Normal Distribution

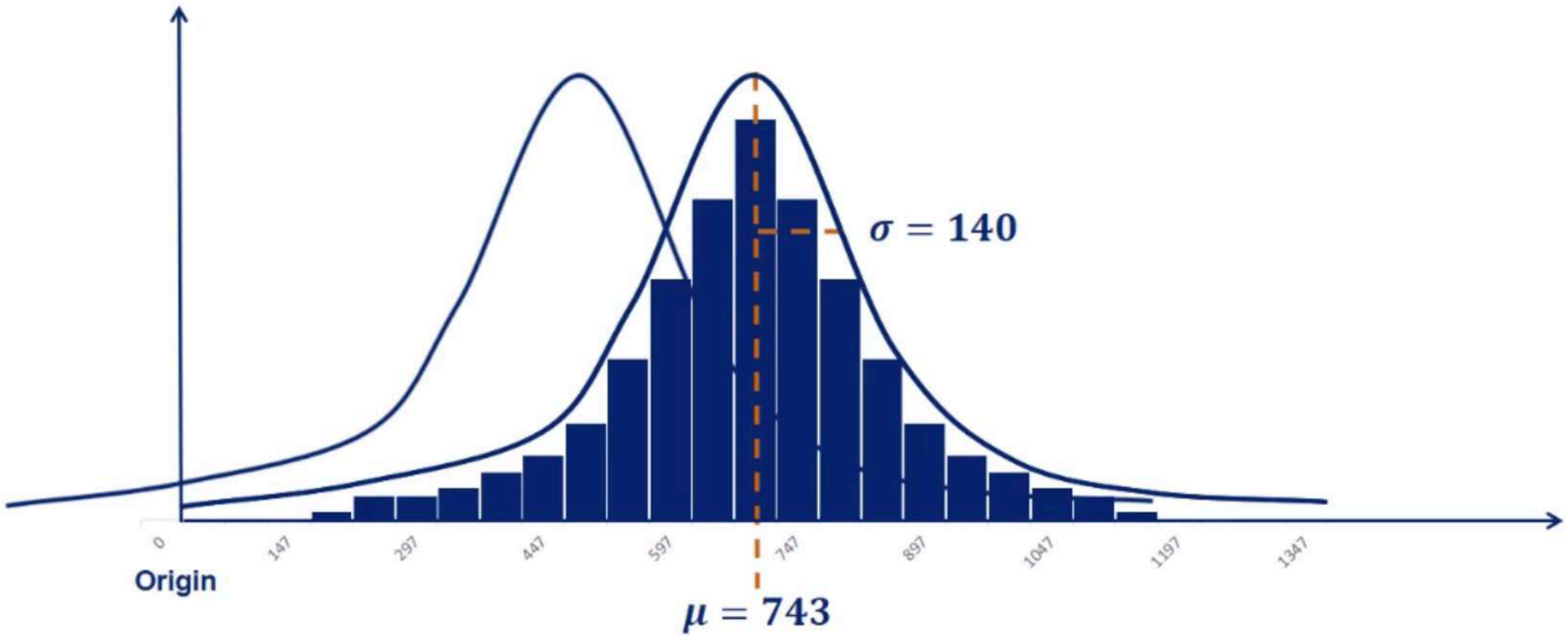
Controlling for the standard deviation

Keeping the standard deviation constant, the graph of a normal distribution with:

- a smaller mean would look in the same way, but be situated to the left (in gray)
- a larger mean would look in the same way, but be situated to the right (in red)



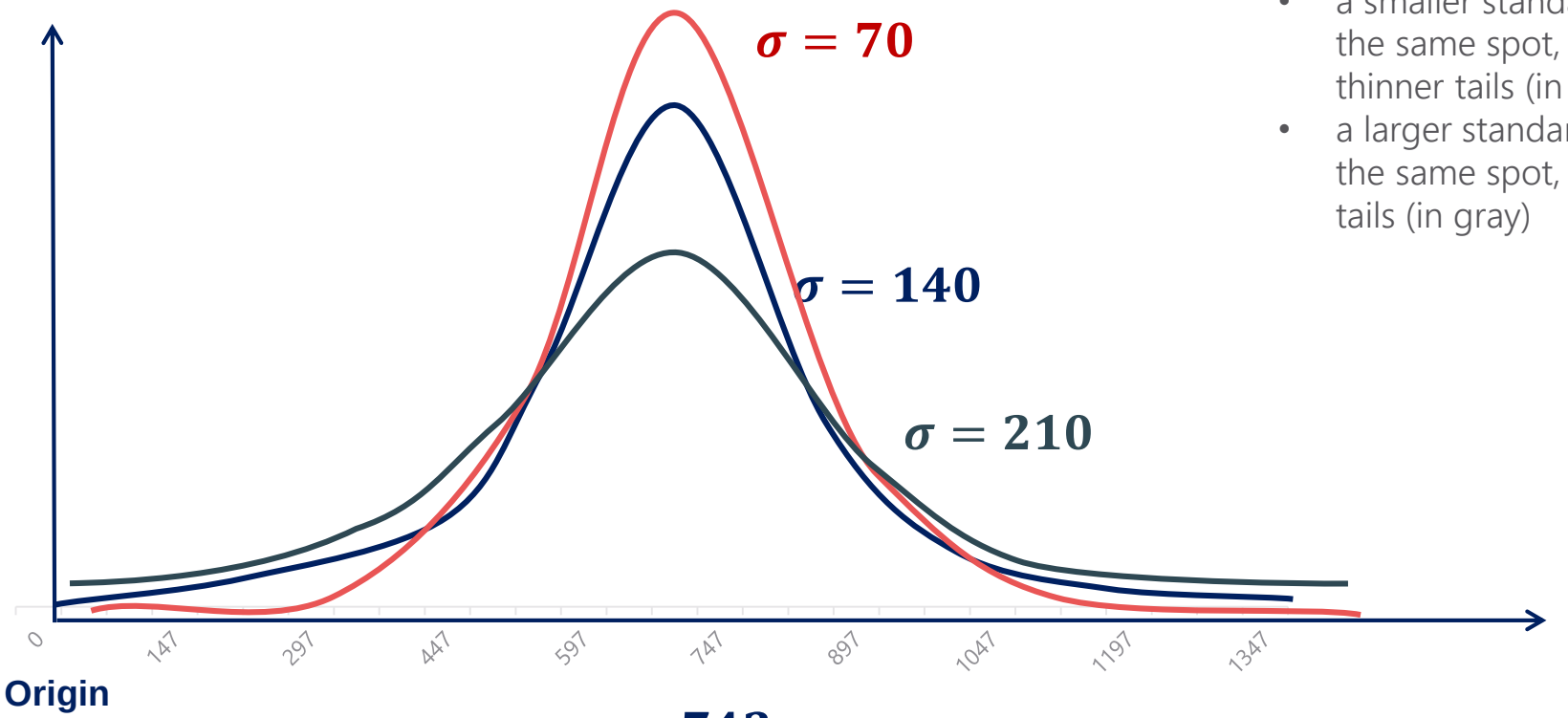
Normal distribution. Controlling for standard deviation



A lower mean would result in the same shape of the distribution, but on the left side of the plane

The Normal Distribution

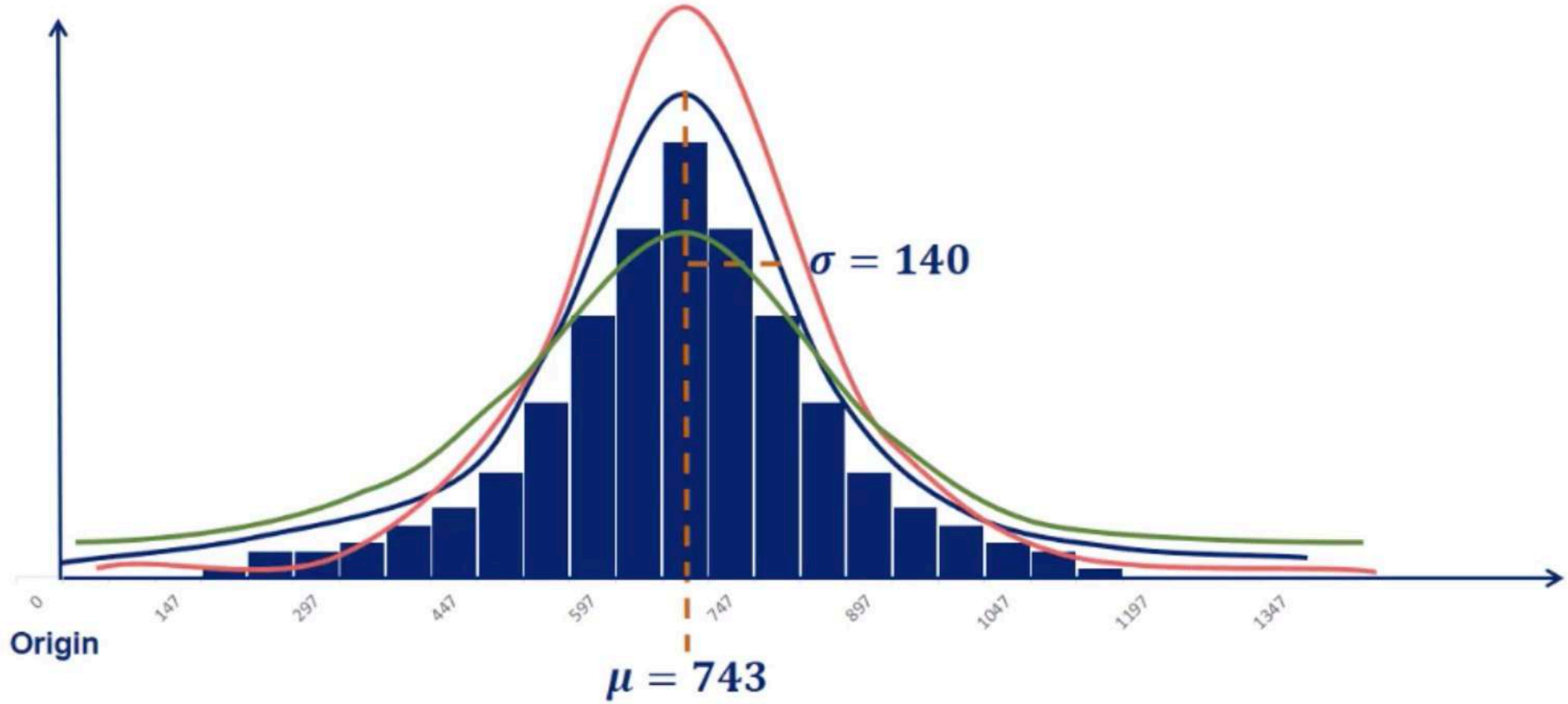
Controlling for the mean



Keeping the mean constant, a normal distribution with:

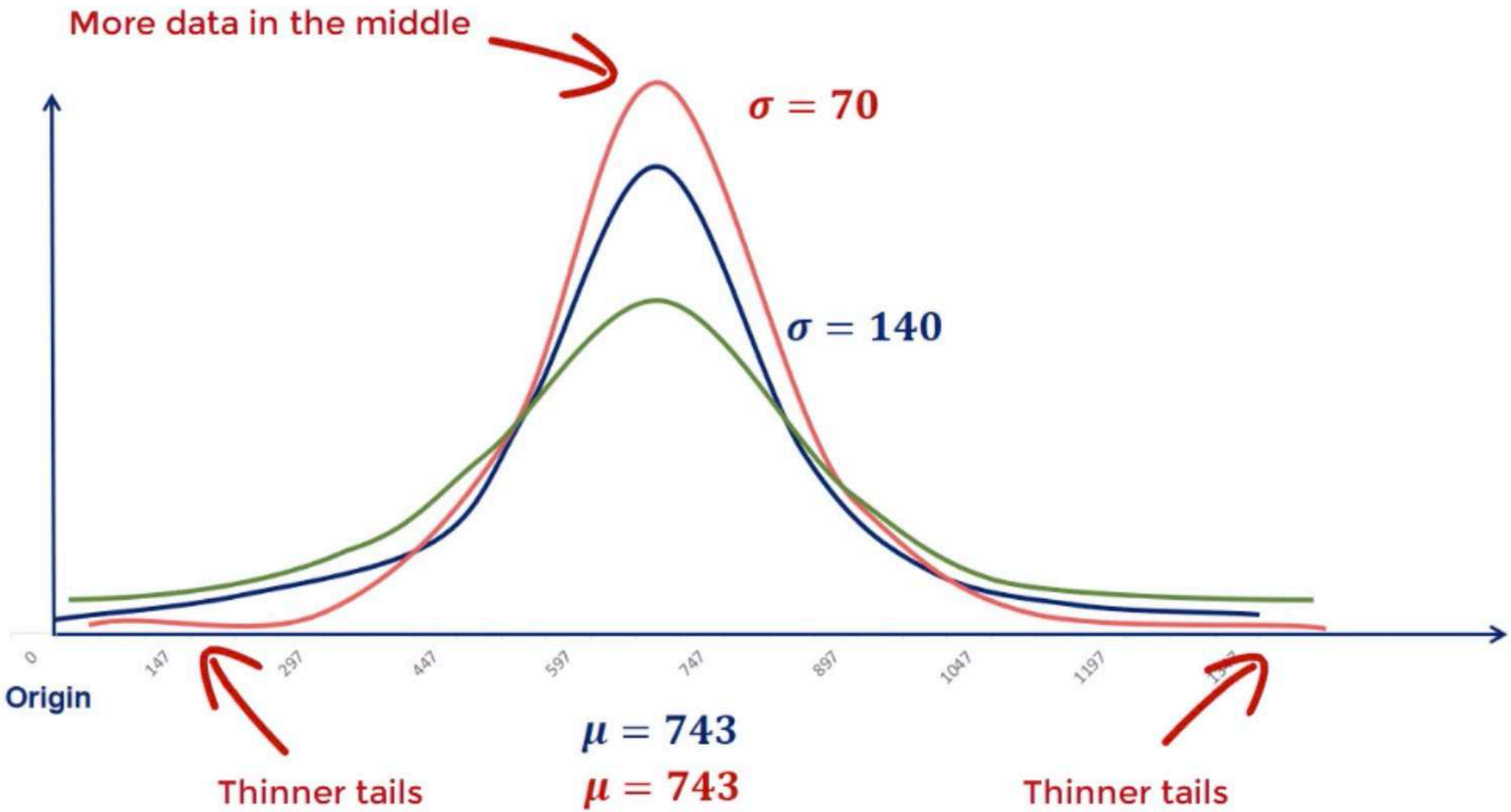
- a smaller standard deviation would be situated in the same spot, but have a higher peak and thinner tails (in red)
- a larger standard deviation would be situated in the same spot, but have a lower peak and fatter tails (in gray)

Normal distribution. Controlling for the mean

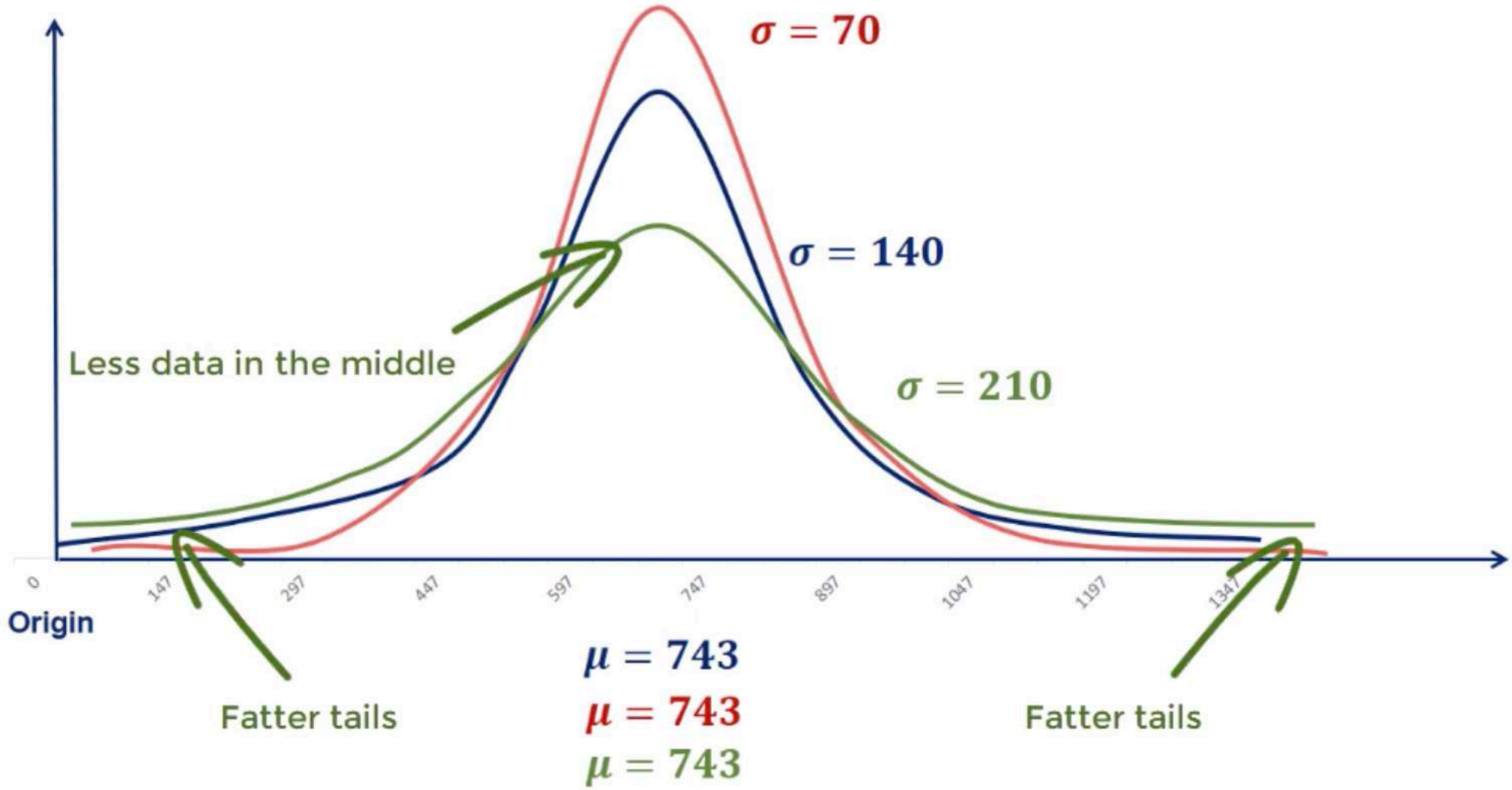


The graph is not moving, but rather - reshaping!

Normal distribution. Controlling for the mean



Normal distribution. Controlling for the mean



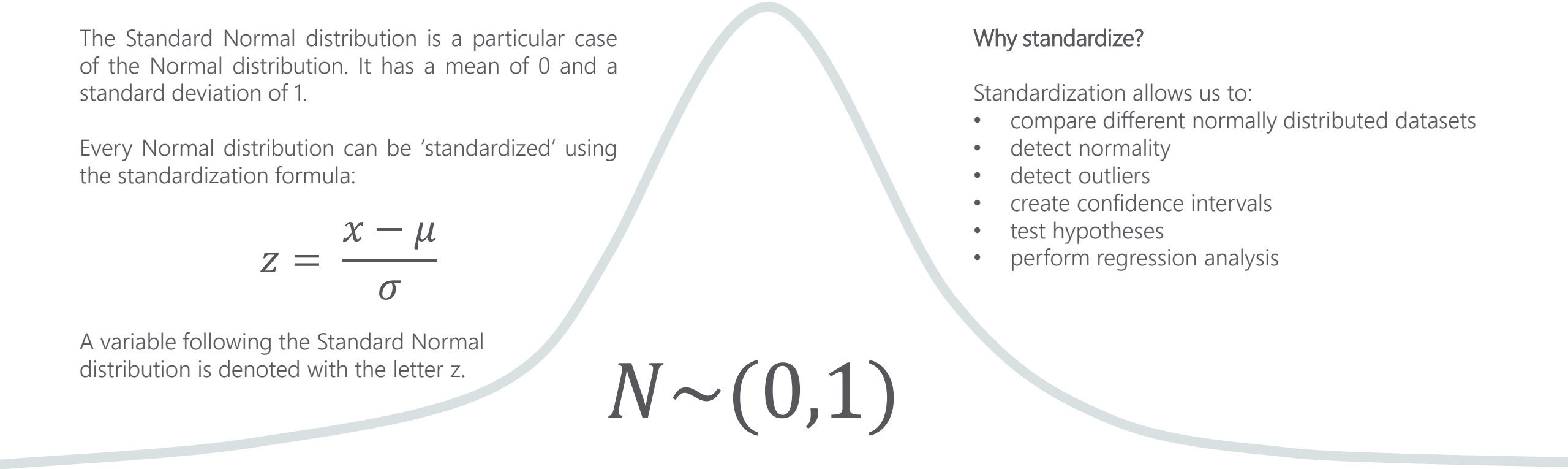
The Standard Normal Distribution

The Standard Normal distribution is a particular case of the Normal distribution. It has a mean of 0 and a standard deviation of 1.

Every Normal distribution can be 'standardized' using the standardization formula:

$$z = \frac{x - \mu}{\sigma}$$

A variable following the Standard Normal distribution is denoted with the letter z.



Why standardize?

Standardization allows us to:

- compare different normally distributed datasets
- detect normality
- detect outliers
- create confidence intervals
- test hypotheses
- perform regression analysis

Rationale of the formula for standardization:

We want to transform a random variable from $N \sim (\mu, \sigma^2)$ to $N \sim (0,1)$. Subtracting the mean from all observations would cause a transformation from $N \sim (\mu, \sigma^2)$ to $N \sim (0, \sigma^2)$, moving the graph to the origin. Subsequently, dividing all observations by the standard deviation would cause a transformation from $N \sim (0, \sigma^2)$ to $N \sim (0,1)$, standardizing the peak and the tails of the graph.

STANDARDIZATION

every distribution can be standardized

STANDARDIZATION

$$\sim (\mu, \sigma^2) \Rightarrow \sim (0,1)$$

$$\frac{x - \mu}{\sigma}$$

STANDARDIZATION

of a Normal distribution

$$\sim N(\mu, \sigma^2) \rightarrow \sim N(0,1)$$

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

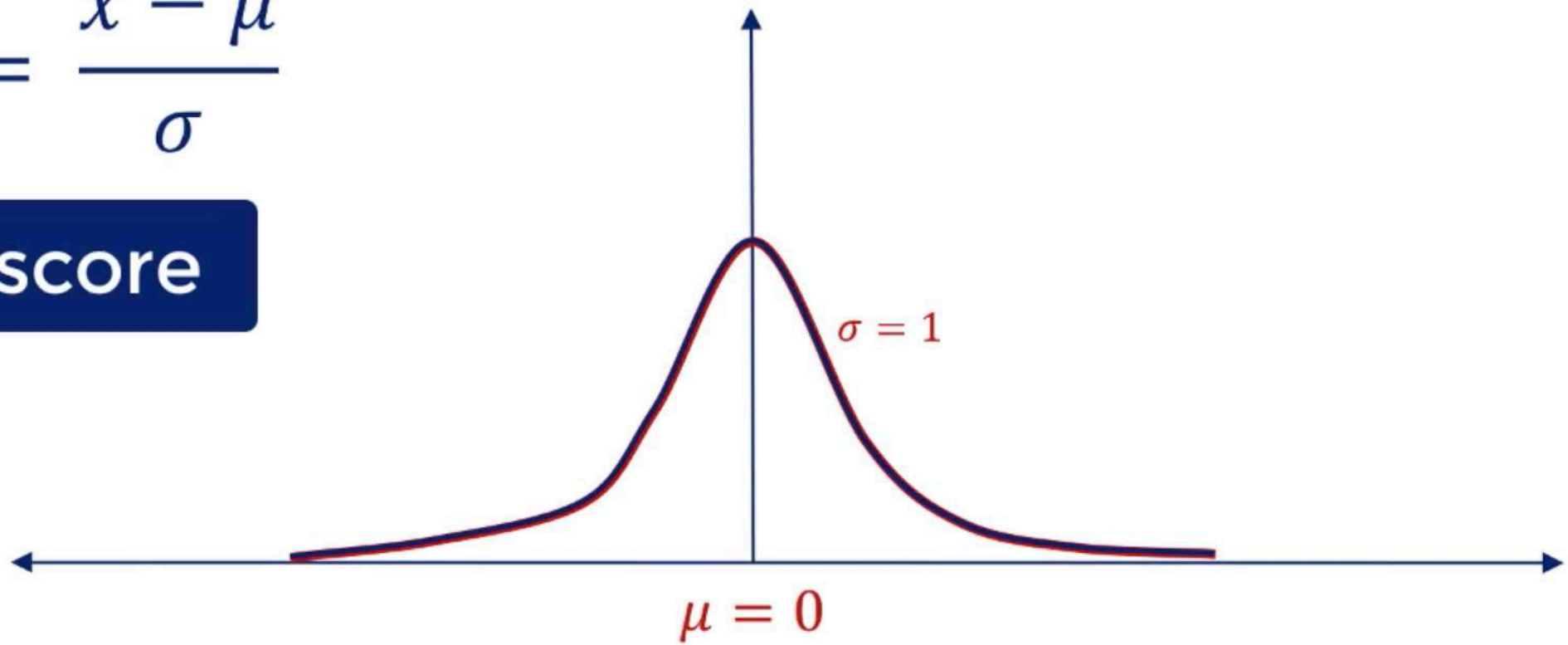
When we standardize a Normal distribution, the result is a Standard Normal distribution

1.50

STANDARDIZATION

$$z = \frac{x - \mu}{\sigma}$$

z-score



$$z \sim N(0,1)$$

Standard normal distribution

Standardization

Original dataset
1
2
2
3
3
3
4
4
5

Mean 3
St. dev 1.22

N~(3,1.49)

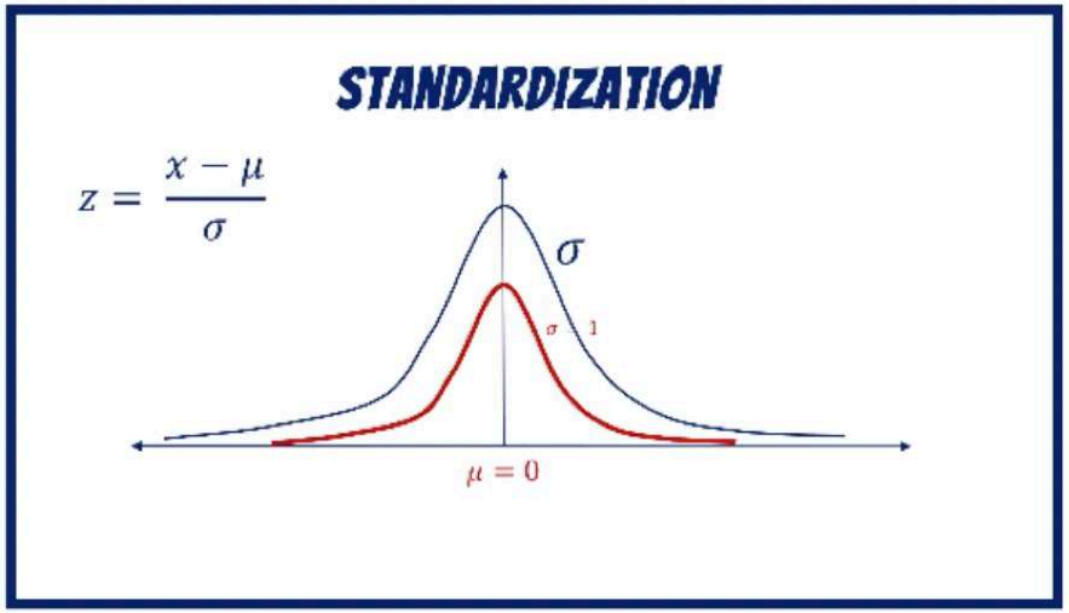
Subtract mean
-2
-1
-1
0
0
0
1
1
2

Mean 0
St. dev 1.22

N~(0,1.49)

x

$x - \mu$



Adding and subtracting values from all data points does not change the standard deviation

Standard normal distribution

Standardization

Original dataset
1
2
2
3
3
3
4
4
5

Mean 3
St. dev 1.22

N~(3,1.49)

x

Subtract mean
-2
-1
-1
0
0
0
1
1
2

Mean 0
St. dev 1.22

N~(0,1.49)

$x - \mu$

Divide by std
-1.63
-0.82
-0.82
0.00
0.00
0.00
0.82

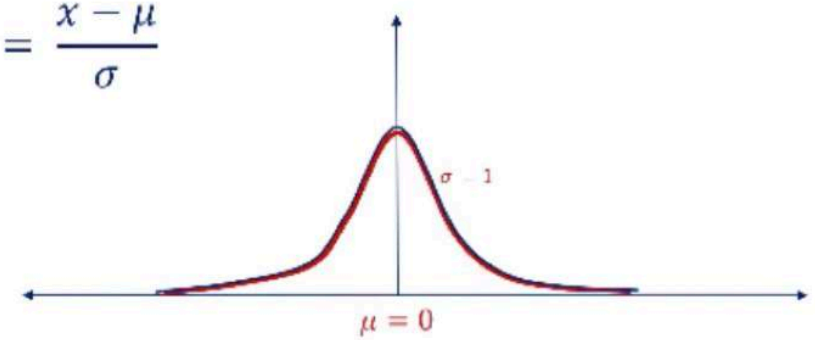
Mean 0.00
St. dev 1.00

N~(0,1)

$$\frac{x - \mu}{\sigma}$$

STANDARDIZATION

$$z = \frac{x - \mu}{\sigma}$$



We keep the curve at the same position but reshape it a bit.

Subtract mean | Subracted mean and std | Divide by std | Divided mean and std

Ready

365 DataScience

The Central Limit Theorem

The Central Limit Theorem (CLT) is one of the greatest statistical insights. It states that no matter the underlying distribution of the dataset, the sampling distribution of the means would approximate a normal distribution. Moreover, the mean of the sampling distribution would be equal to the mean of the original distribution and the variance would be n times smaller, where n is the size of the samples. The CLT applies whenever we have a sum or an average of many variables (e.g. sum of rolled numbers when rolling dice).

The theorem

- No matter the distribution
- The distribution of $\overline{x_1}$, $\overline{x_2}$, $\overline{x_3}$, $\overline{x_4}$, …, $\overline{x_k}$ would tend to $N \sim \left(\mu, \frac{\sigma^2}{n}\right)$
- The more samples, the closer to Normal (k -> ∞)
- The bigger the samples, the closer to Normal (n -> ∞)

Why is it useful?

The CLT allows us to assume normality for many different variables. That is very useful for confidence intervals, hypothesis testing, and regression analysis. In fact, the Normal distribution is so predominantly observed around us due to the fact that following the CLT, many variables converge to Normal.

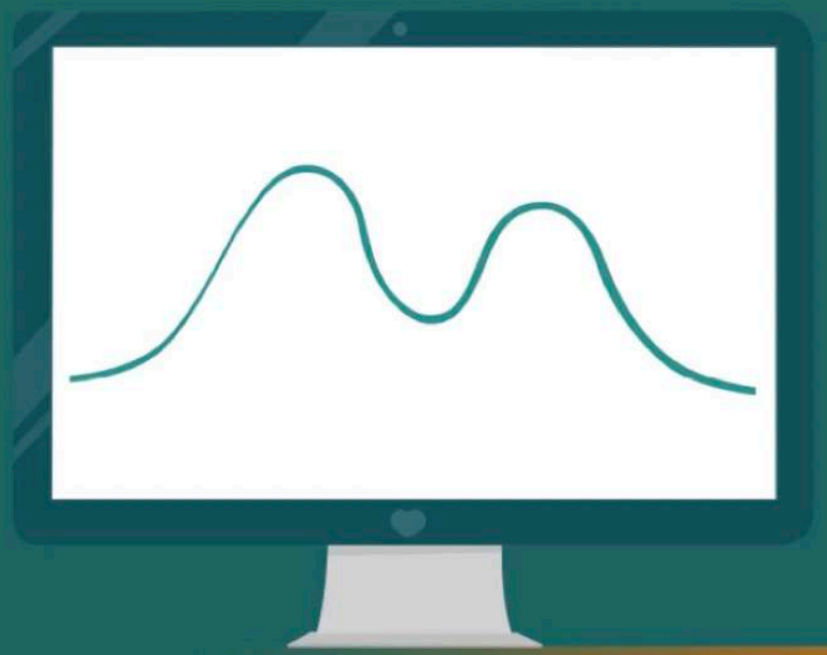
Click here for a CLT simulator.

Where can we see it?

Since many concepts and events are a sum or an average of different effects, CLT applies and we observe normality all the time. For example, in regression analysis, the dependent variable is explained through the sum of error terms.

SAMPLE OF USED CARS IN A CAR SHOP

$\mu, \sigma_x, \sigma_{xy}$



DISTRIBUTION OF CAR PRICES

Population

.70

Taking a single value, as we did in descriptive statistics is definitely suboptimal

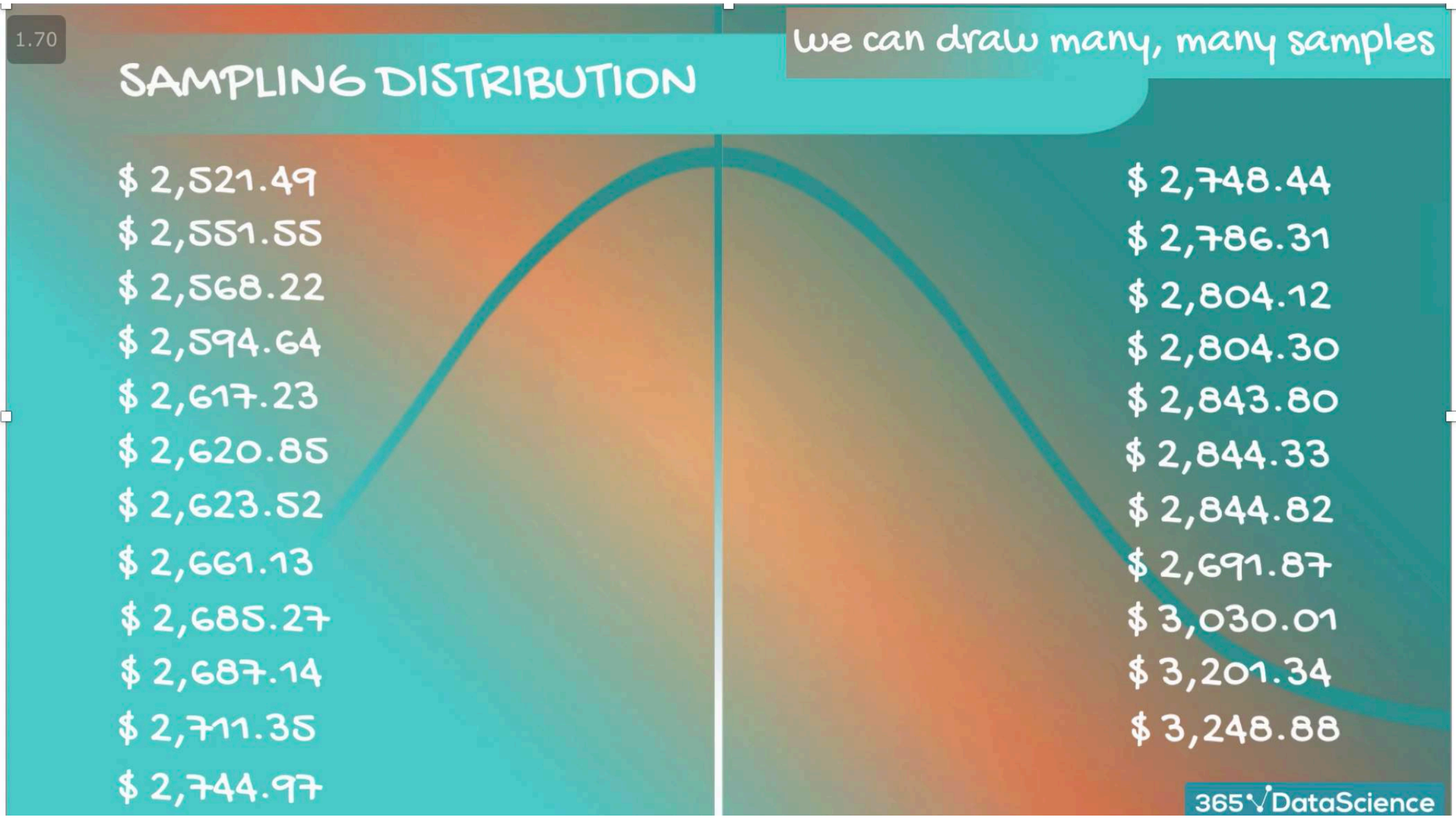
THE MEAN

$\bar{x}$

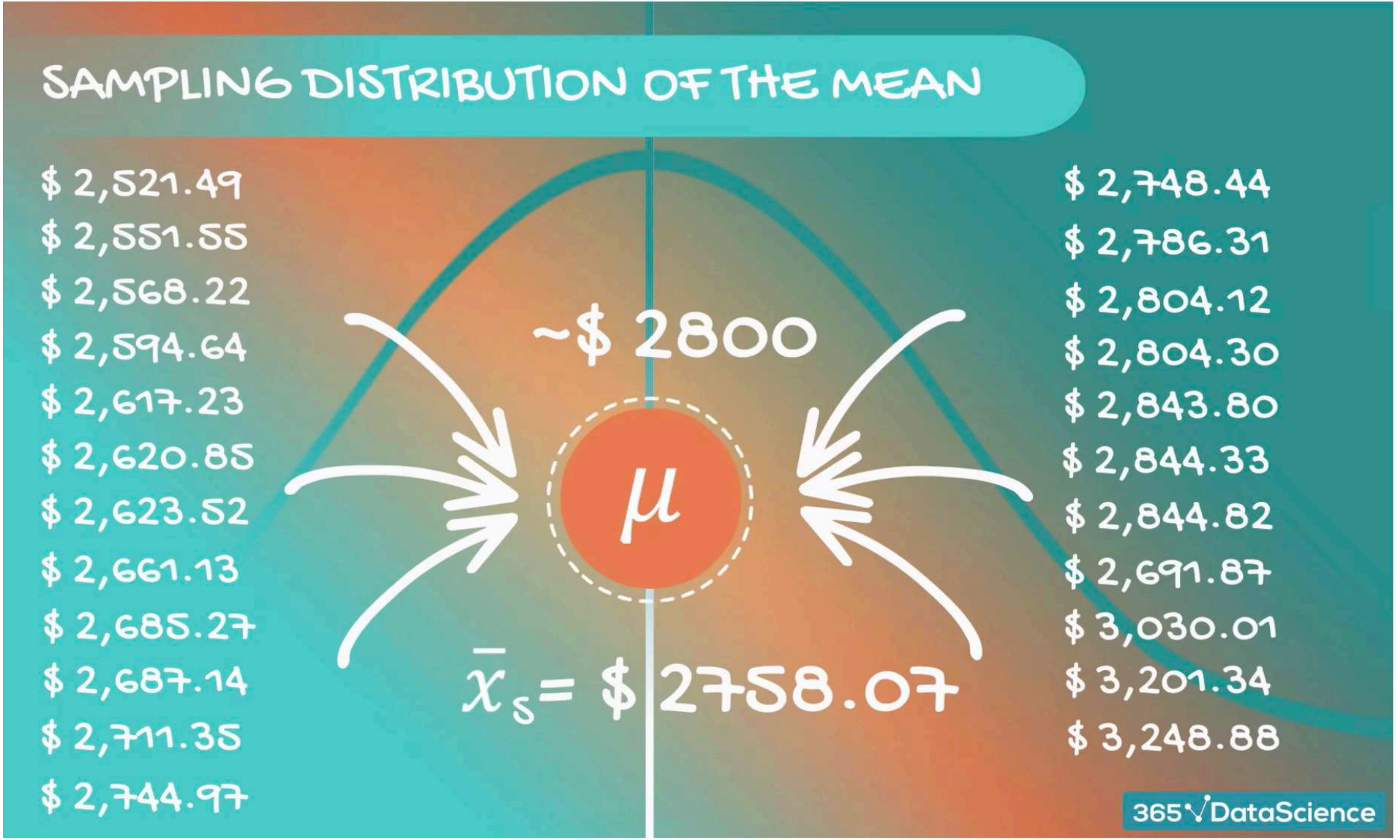
Sample #1:	$2,617.23
Sample #2:	$3,201.34
Sample #3:	$2,844.33

SAMPLING DISTRIBUTION

we can draw many, many samples



SAMPLING DISTRIBUTION OF THE MEAN



CENTRAL LIMIT THEOREM

Original distribution

$\mu \quad \sigma^2$

Sampling distribution

$N\left(\mu, \frac{\sigma^2}{n}\right)$

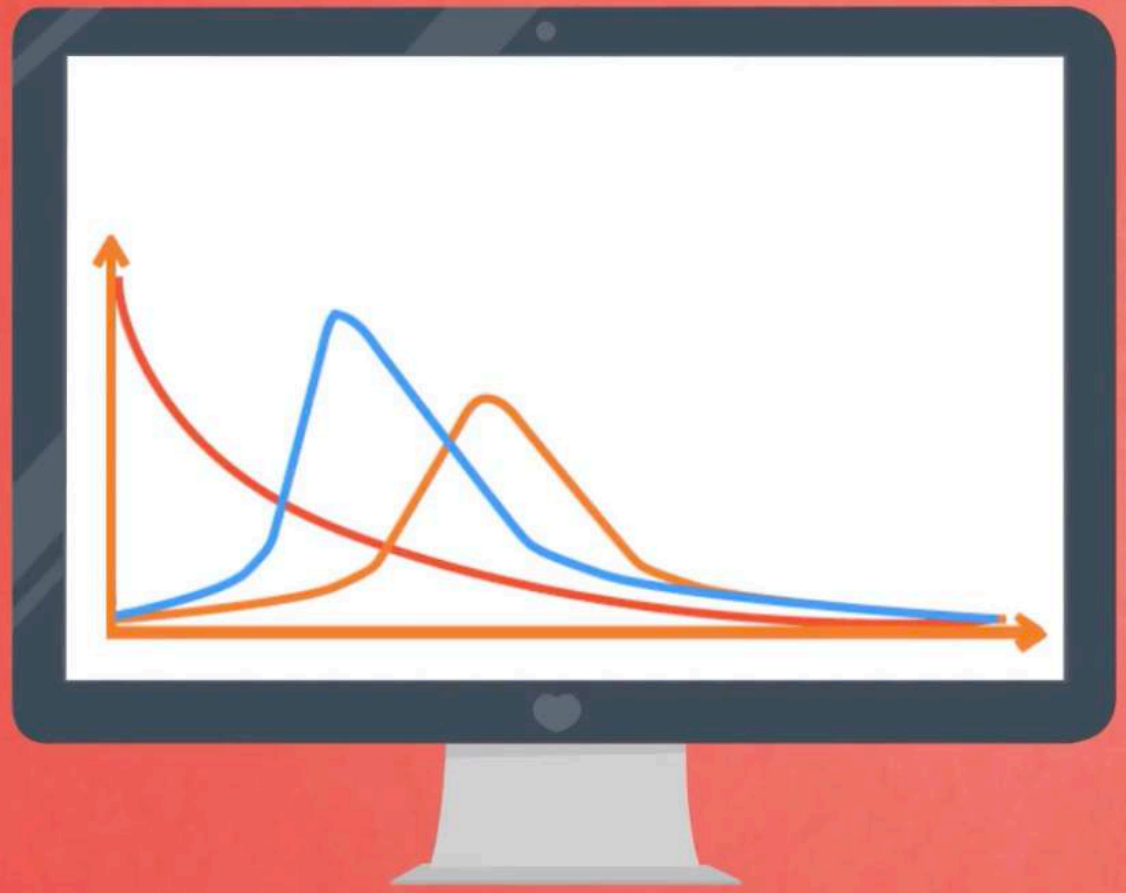
Same mean as in the population

No matter the underlying distribution,
the sampling distribution approximates a Normal

Sampling distribution ~ $N\left(\mu, \frac{\sigma^2}{n}\right)$, $n > 30$

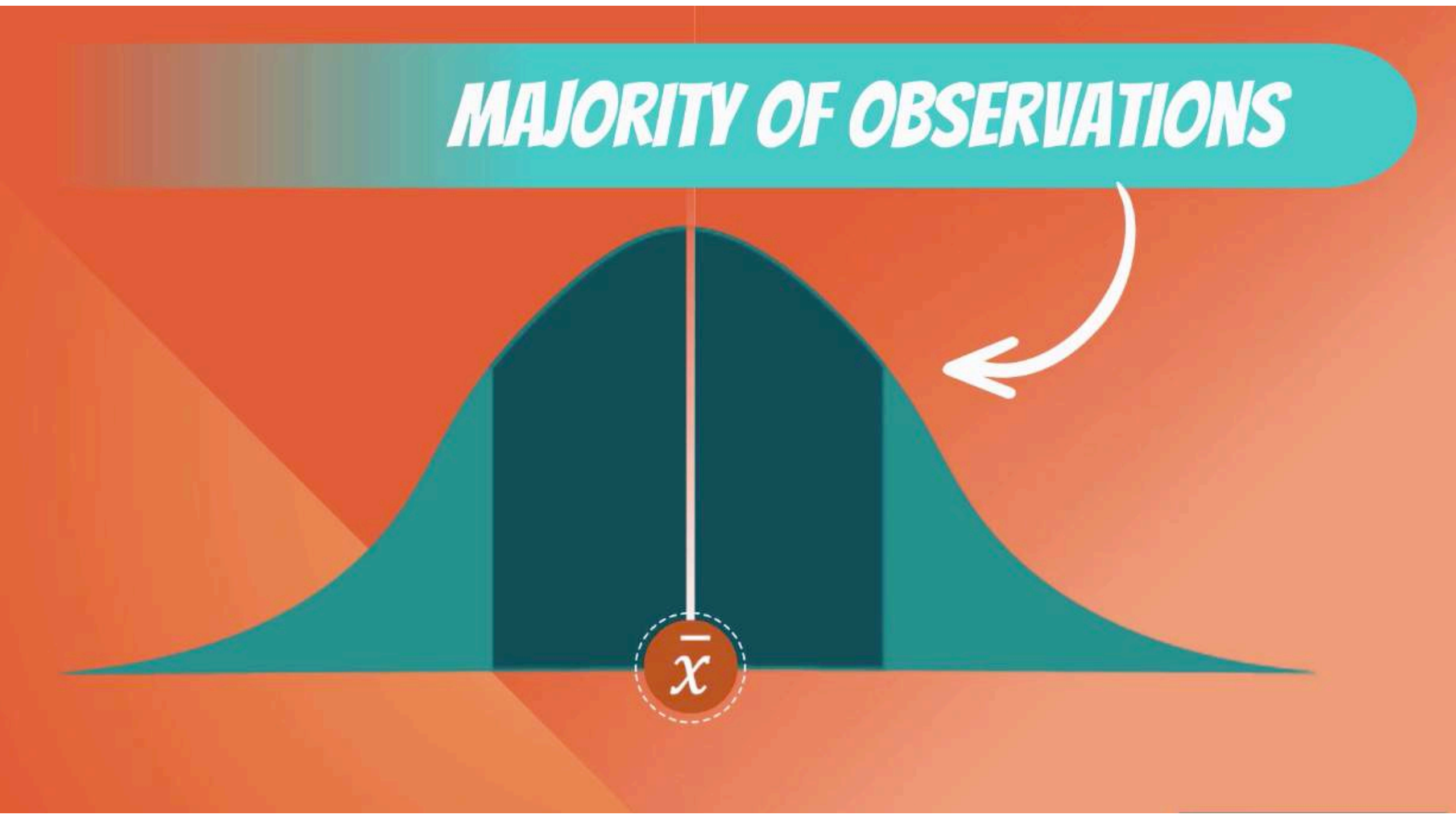
365 DataScience

NO MATTER THE DISRTIBUTION



$$\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots, \bar{x}_k$$

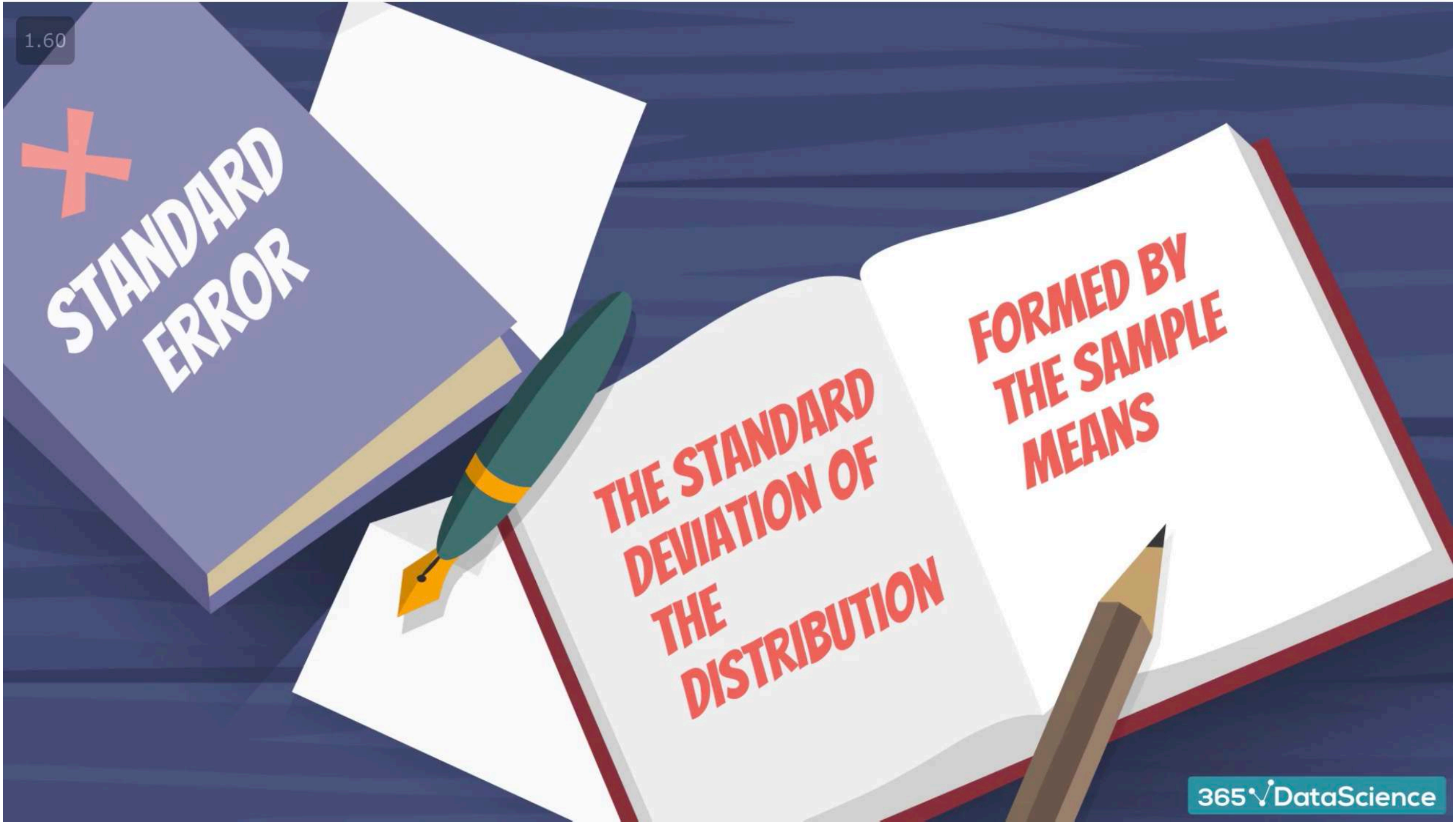
$$N \sim \left(\mu, \frac{\sigma^2}{n}\right)$$


MAJORITY OF OBSERVATIONS
$\bar{x}$

REASONS TO USE THE NORMAL DISTRIBUTION

CLT allows us to perform tests, solve problems and make inferences using the Normal distribution, even when the population is not normally distributed

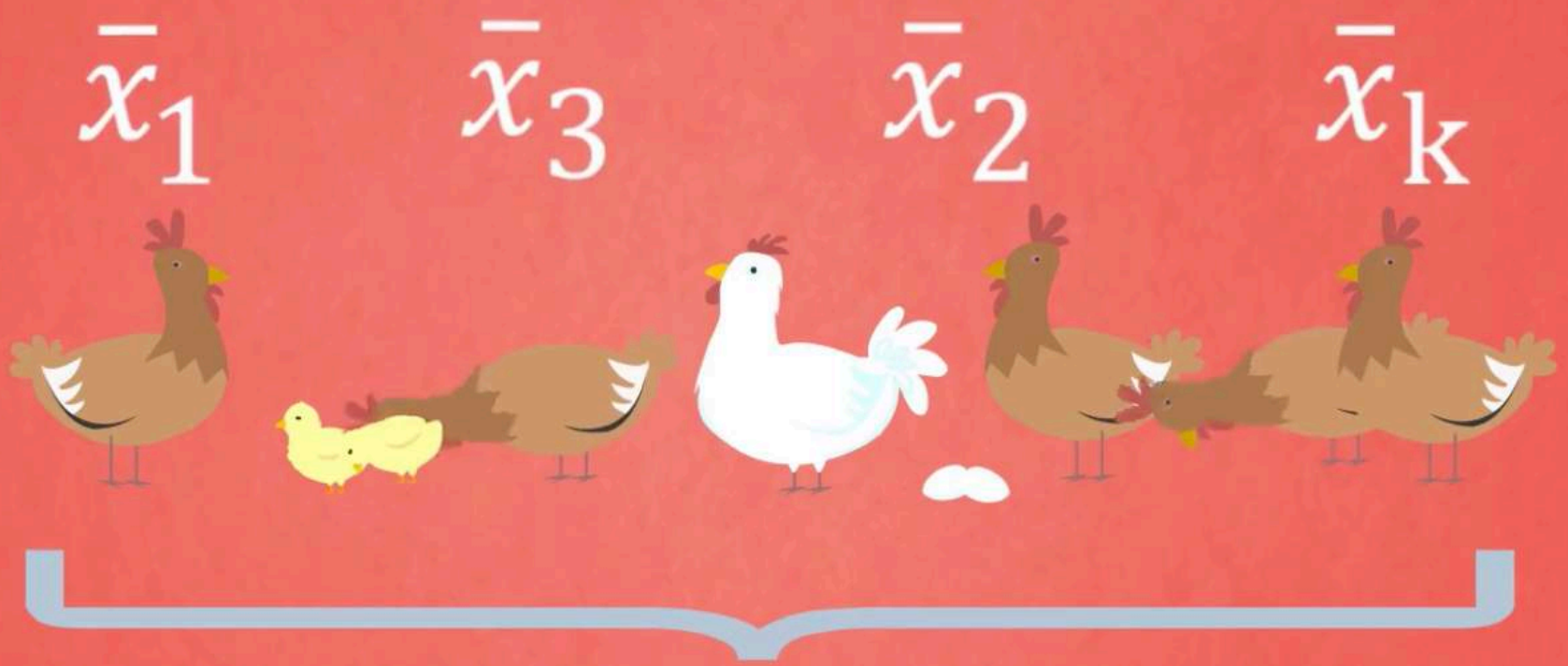
- They approximate a wide variety of random variables
- Distributions of sample means with large enough sample sizes could be approximated to normal
- All computable statics are elegant
- Decisions based on normal distribution insights have a good track record


1.60
STANDARD ERROR
THE STANDARD DEVIATION OF THE DISTRIBUTION
FORMED BY THE SAMPLE MEANS
365 DataScience

HOW DO WE FIND THE STANDARD ERROR?

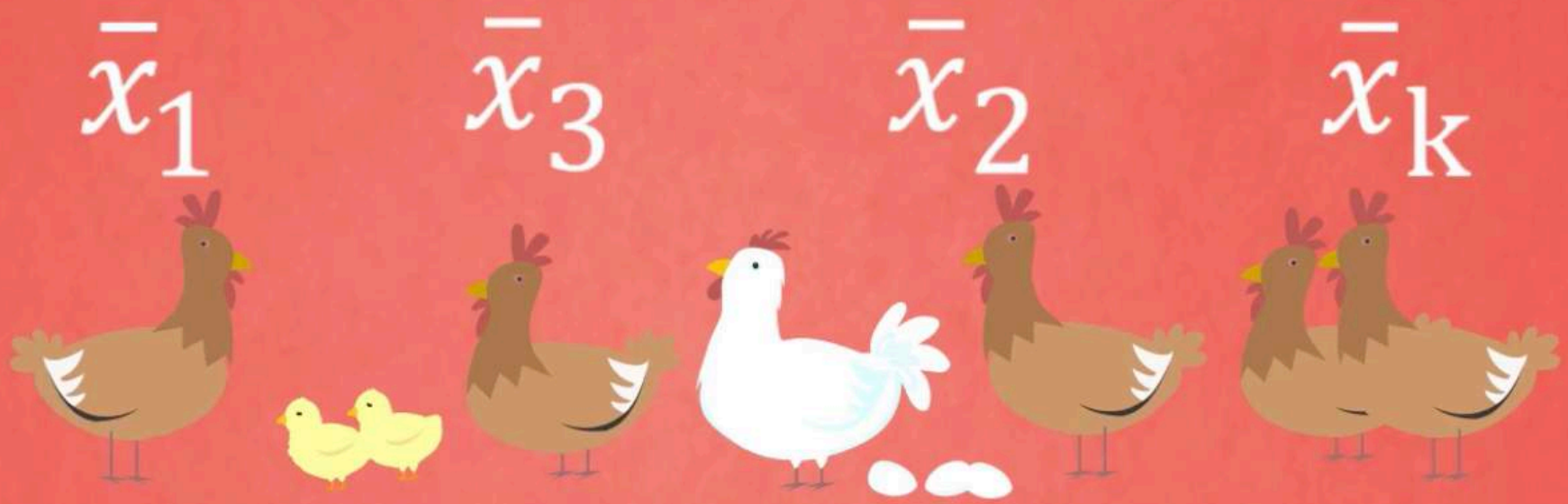
standard deviation (of the sampling distribution) $= \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$

MEANING OF THE STANDARD ERROR



Like any standard deviation, it shows variability

MEANING OF THE STANDARD ERROR

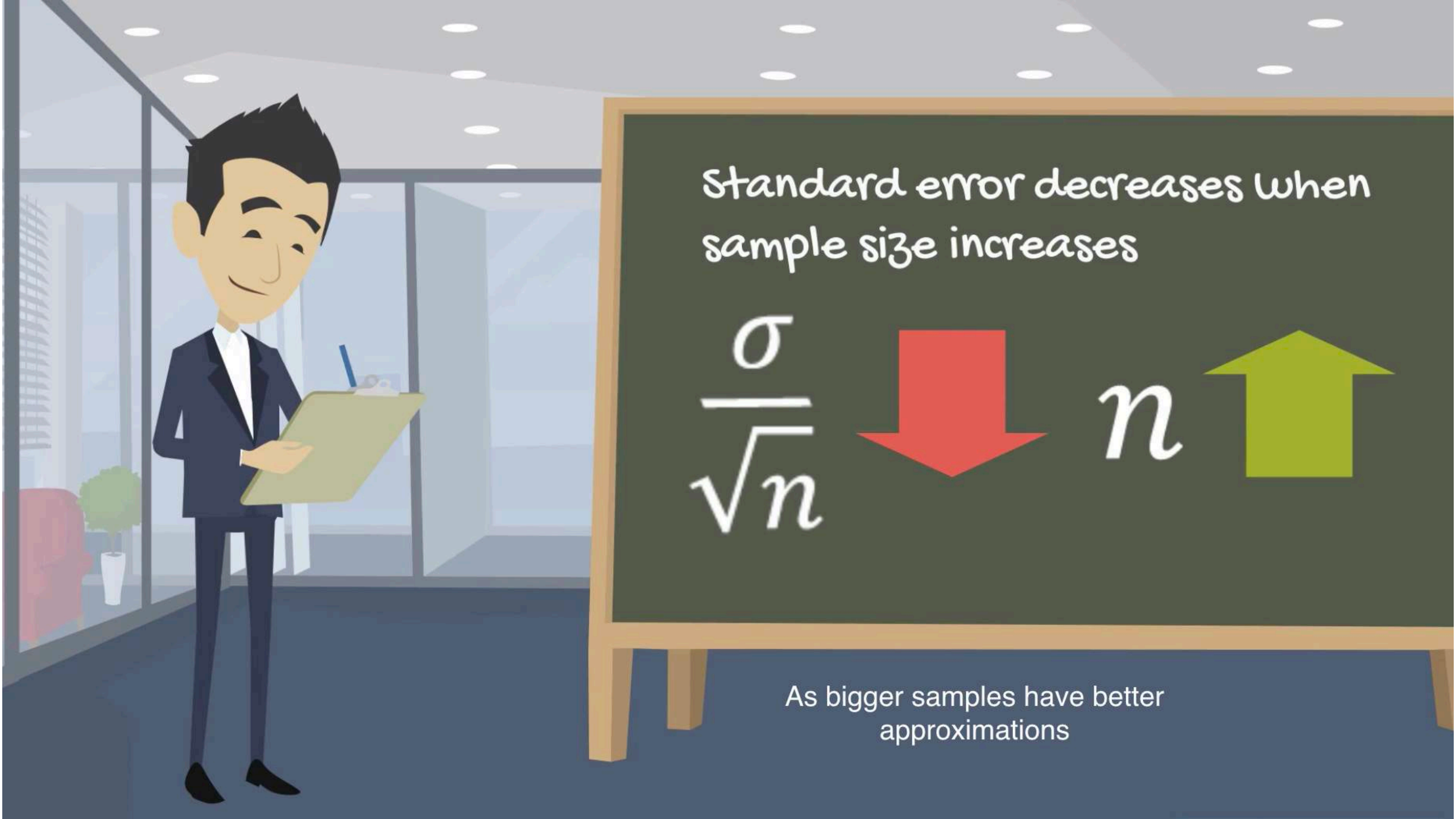


Variability.. of sample means

WHY IS IT IMPORTANT?

Used in most statistical tests

Because it shows how well you approximated the true mean


Standard error decreases when sample size increases
$\frac{\sigma}{\sqrt{n}}$
n
As bigger samples have better approximations

Estimators and Estimates

Estimators

Broadly, an estimator is a mathematical function that approximates a population parameter depending only on sample information.

Examples of estimators and the corresponding parameters:

Term	Estimator	Parameter
Mean	$\bar{x}$	μ
Variance	s^2	σ^2
Correlation	r	ρ

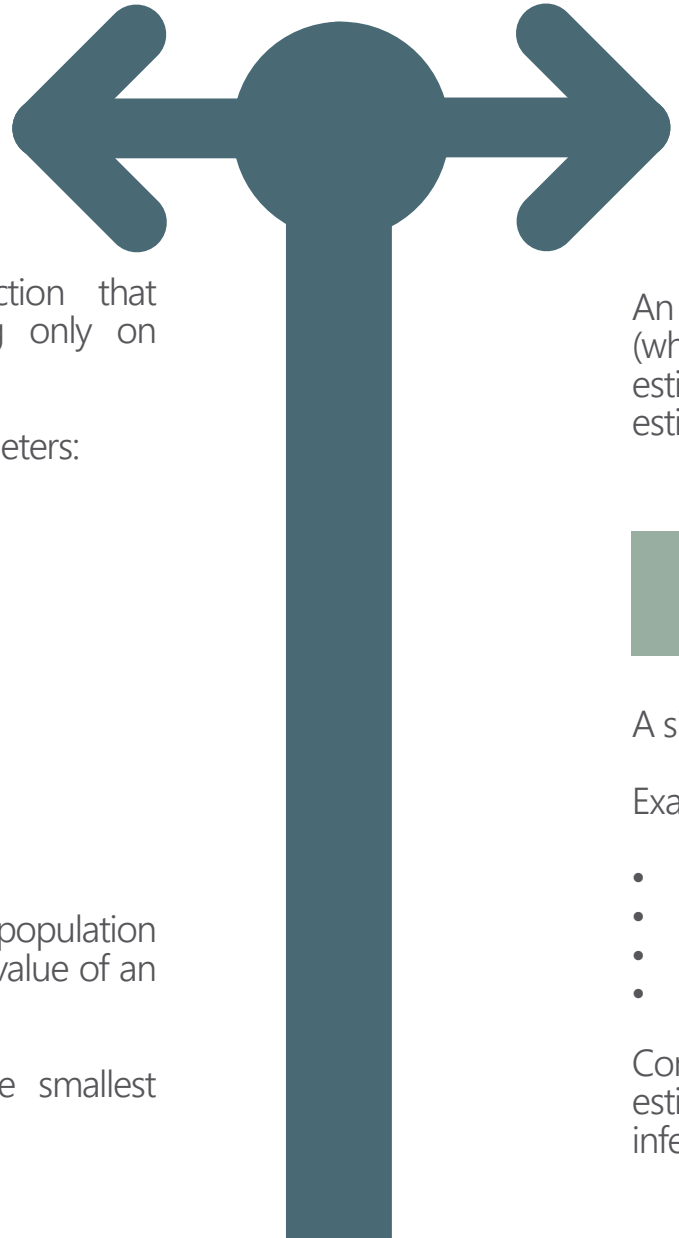
Estimators have two important properties:

- Bias

The expected value of an unbiased estimator is the population parameter. The bias in this case is 0. If the expected value of an estimator is (parameter + b), then the bias is b.

- Efficiency

The most efficient estimator is the one with the smallest variance.



Estimates

An estimate is the output that you get from the estimator (when you apply the formula). There are two types of estimates: point estimates and confidence interval estimates.

Point estimates

A single value.

Examples:

- 1
- 5
- 122.67
- 0.32

Confidence intervals

An interval.

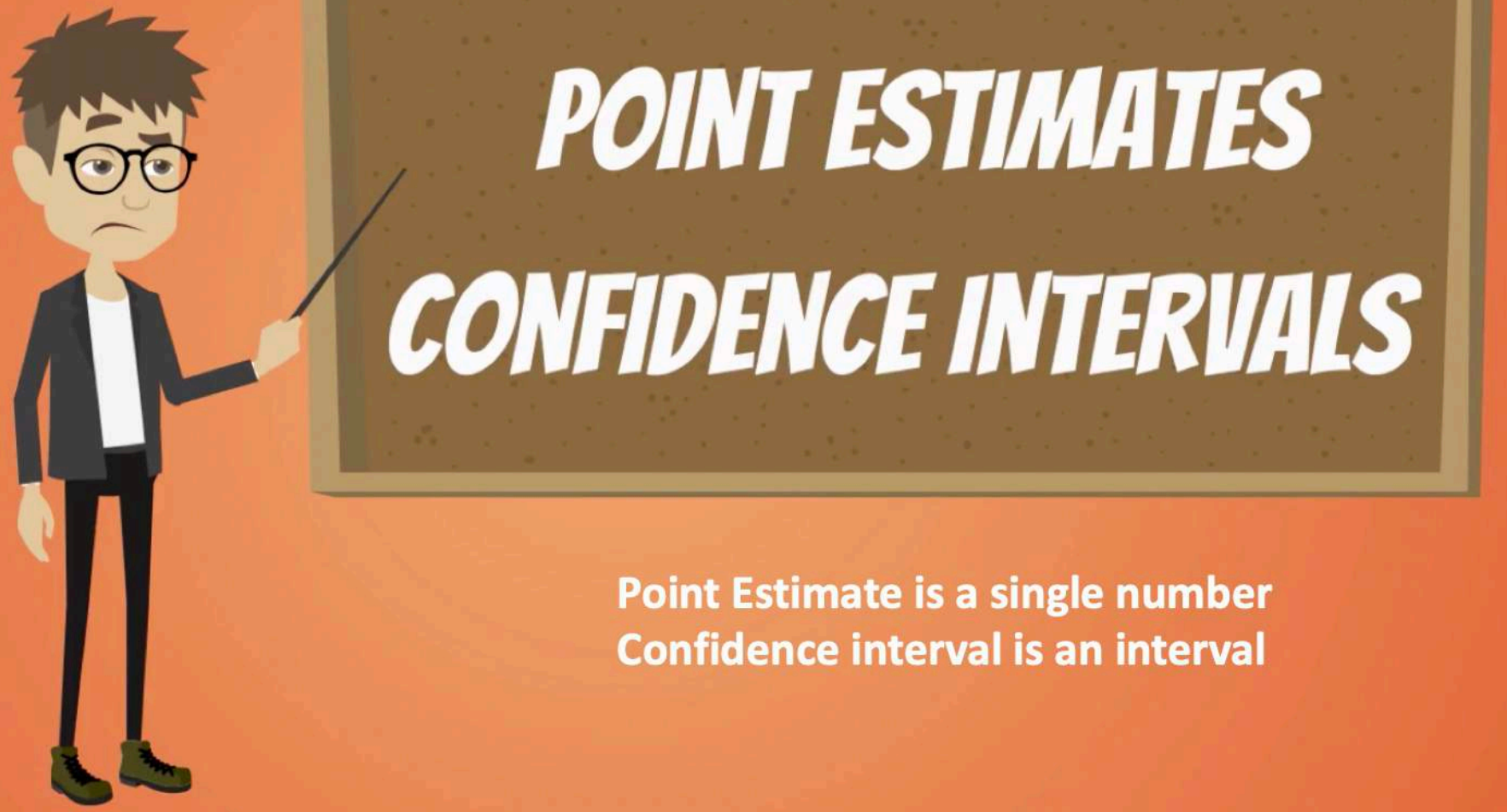
Examples:

- (1 , 5)
- (12 , 33)
- (221.78 , 745.66)
- (- 0.71 , 0.11)

Confidence intervals are much more precise than point estimates. That is why they are preferred when making inferences.

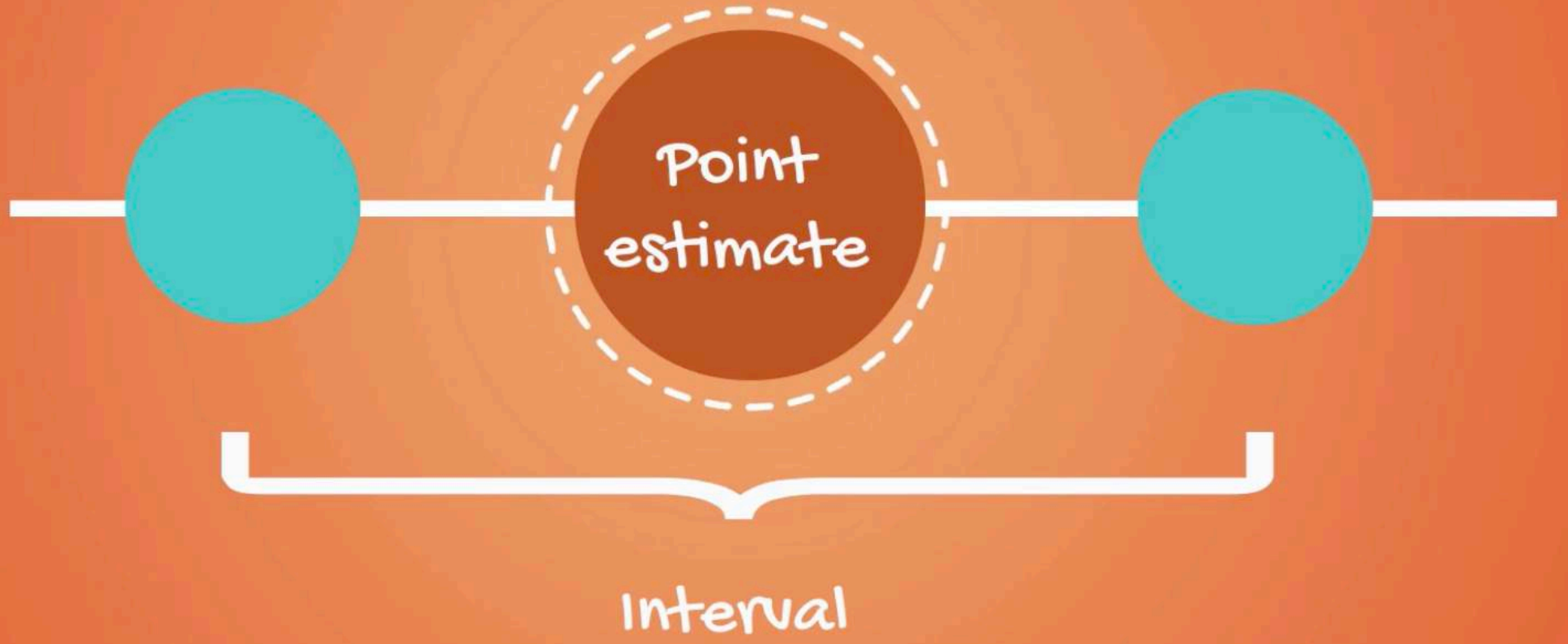
APPROXIMATION DEPENDING SOLELY ON SAMPLE INFORMATION

Types of Estimates



Point Estimate is a single number
Confidence interval is an interval

CONFIDENCE INTERVAL ESTIMATES



POINT ESTIMATORS AND ESTIMATES

Estimator /how to estimate/	Parameter /what to estimate/	Estimate /concrete result/
$\bar{x}$	of μ	52.22
s^2	of σ^2	1724.93

EFFICIENCY

MOST EFFICIENT

BIAS

UNBIASED

UNBIASED ESTIMATOR

expected value = population parameter

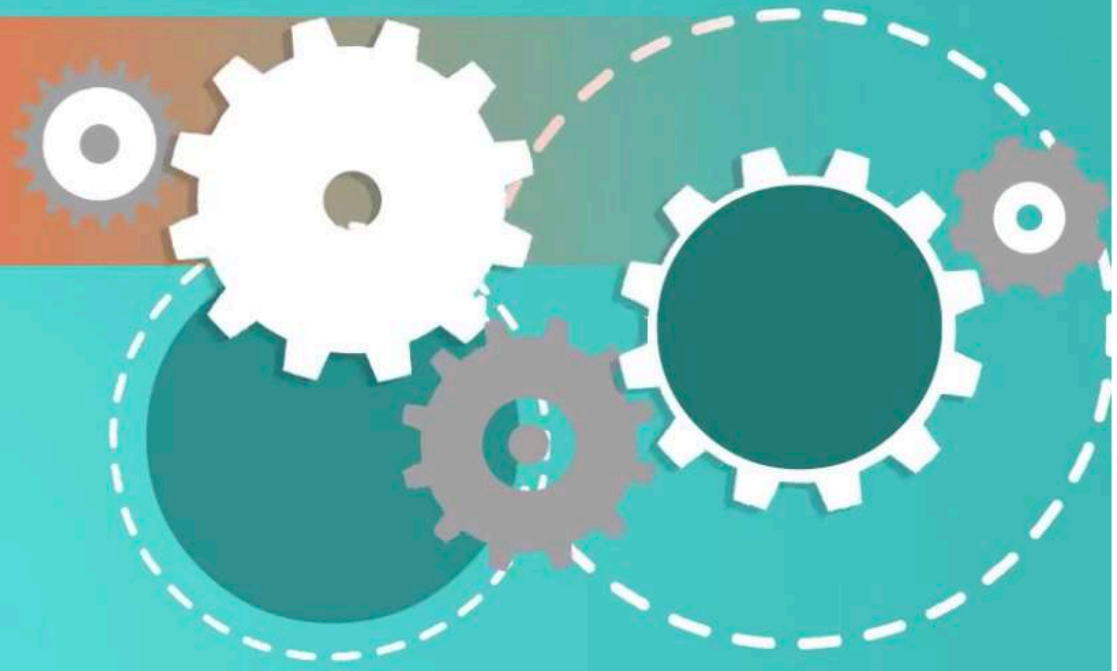
e.g. $\bar{x}$ has an expected value of μ

BIAS

$\bar{x}$ estimates μ with **NO BIAS**

$\bar{x}$ **+1FT** estimates μ with a bias of **+1FT**

EFFICIENCY



The most efficient estimator is the unbiased estimator with smallest variance

1.60

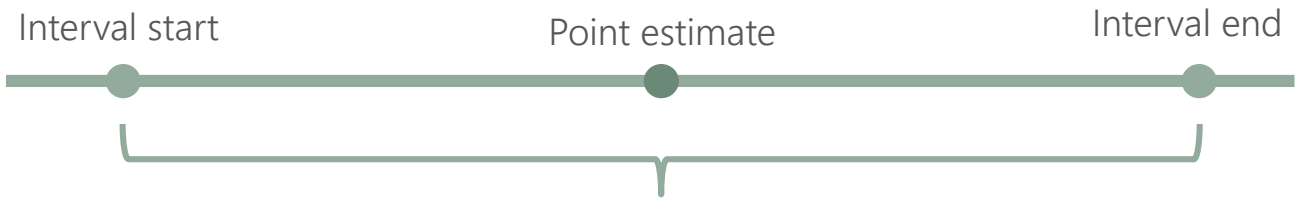
STATISTICS

broader term

ESTIMATORS

a type of statistic

Confidence Intervals and the Margin of Error



Definition: A confidence interval is an interval within which we are confident (with a certain percentage of confidence) the population parameter will fall.

We build the confidence interval **around** the point estimate.

(1-α) is the level of confidence. We are (1-α)*100% confident that the population parameter will fall in the specified interval. Common alphas are: 0.01, 0.05, 0.1.

General formula:

[$\bar{x}$ - ME, $\bar{x}$ + ME] , where ME is the margin of error.

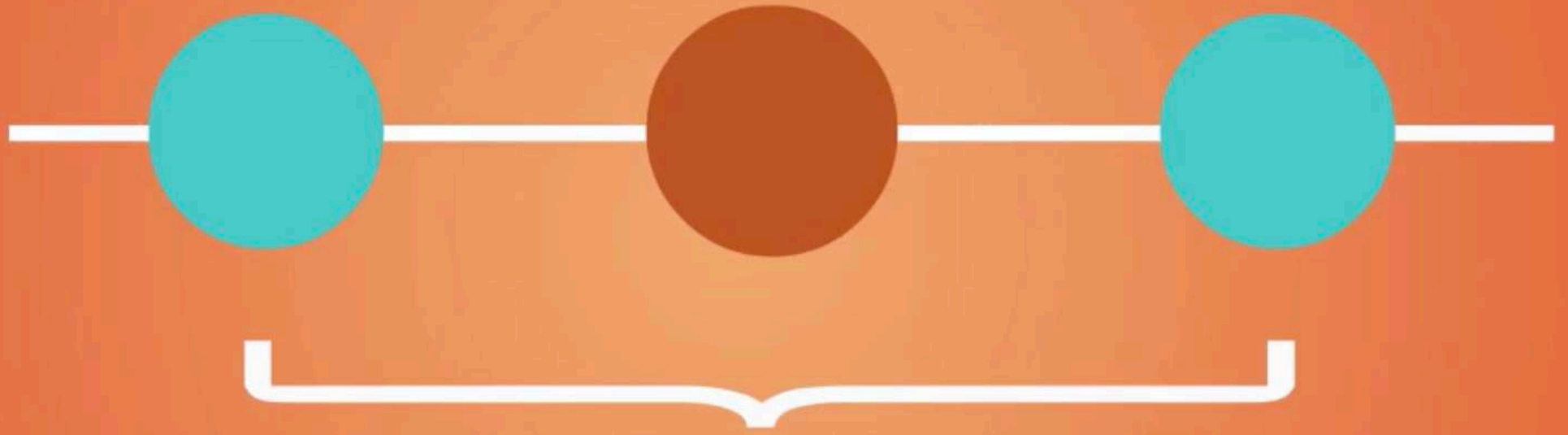
$$\textbf{ME} = reliability\ factor * \frac{standard\ deviation}{\sqrt{sample\ size}}$$

$$z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}$$

$$t_{\upsilon,\alpha/2} * \frac{s}{\sqrt{n}}$$

Term	Effect on width of CI
(1-α) ↑	↑
σ ↑	↑
n ↑	↓

CONFIDENCE INTERVALS



A confidence interval is the range within which you expect the population parameter to be.

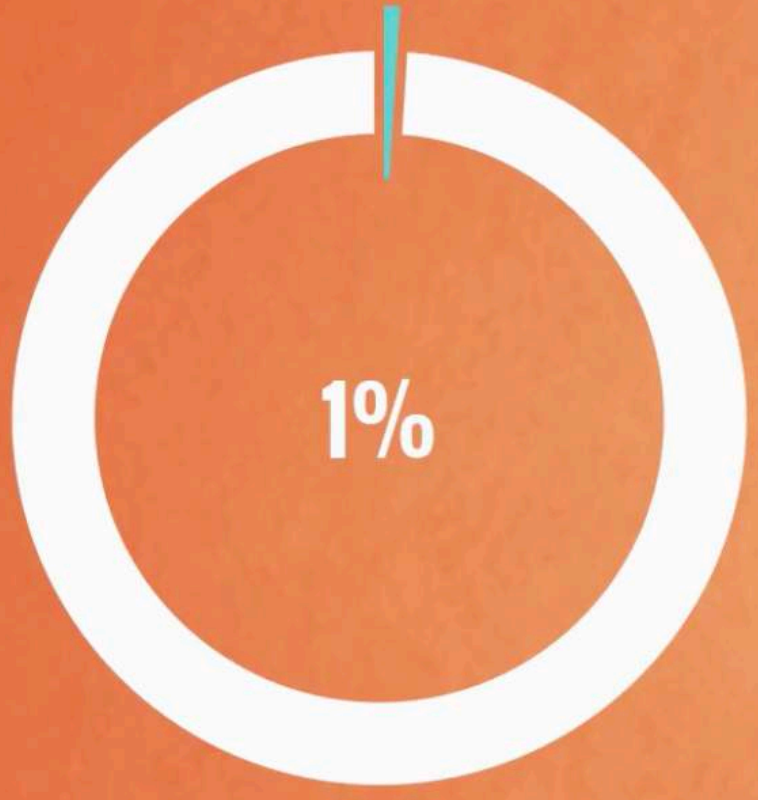
You cannot be 100% confident
unless you test the whole population

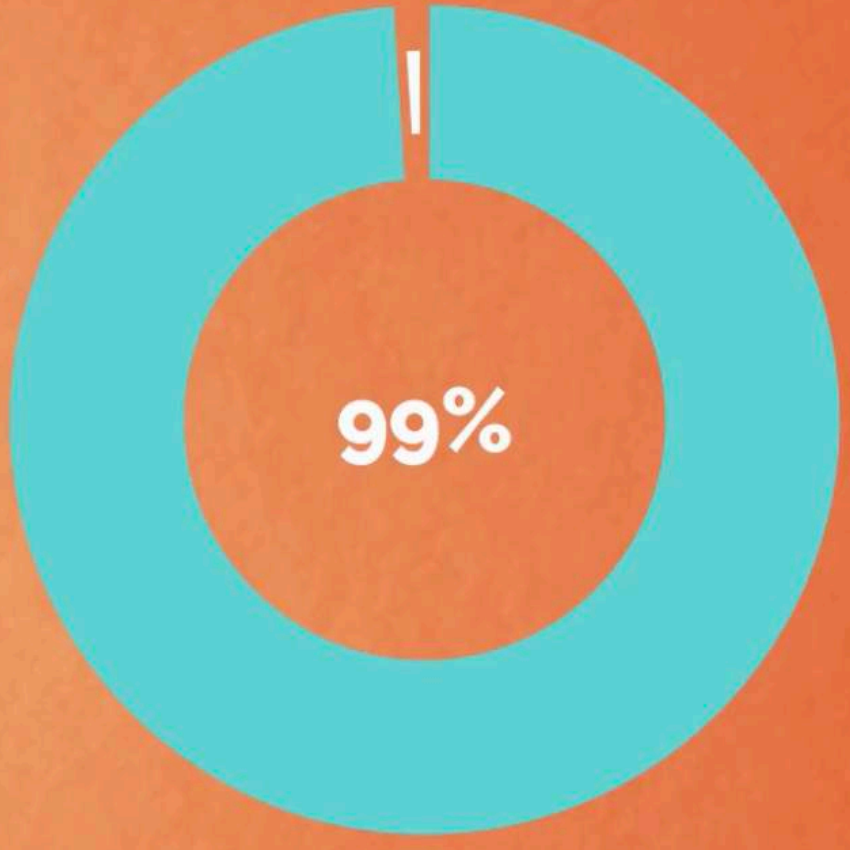
95% CI MEANS THERE IS ONLY 5% CHANCE THAT THE POPULATION PARAMETER IS OUTSIDE THE RANGE

CONFIDENCE LEVEL

$$0 \leq \alpha \leq 1$$

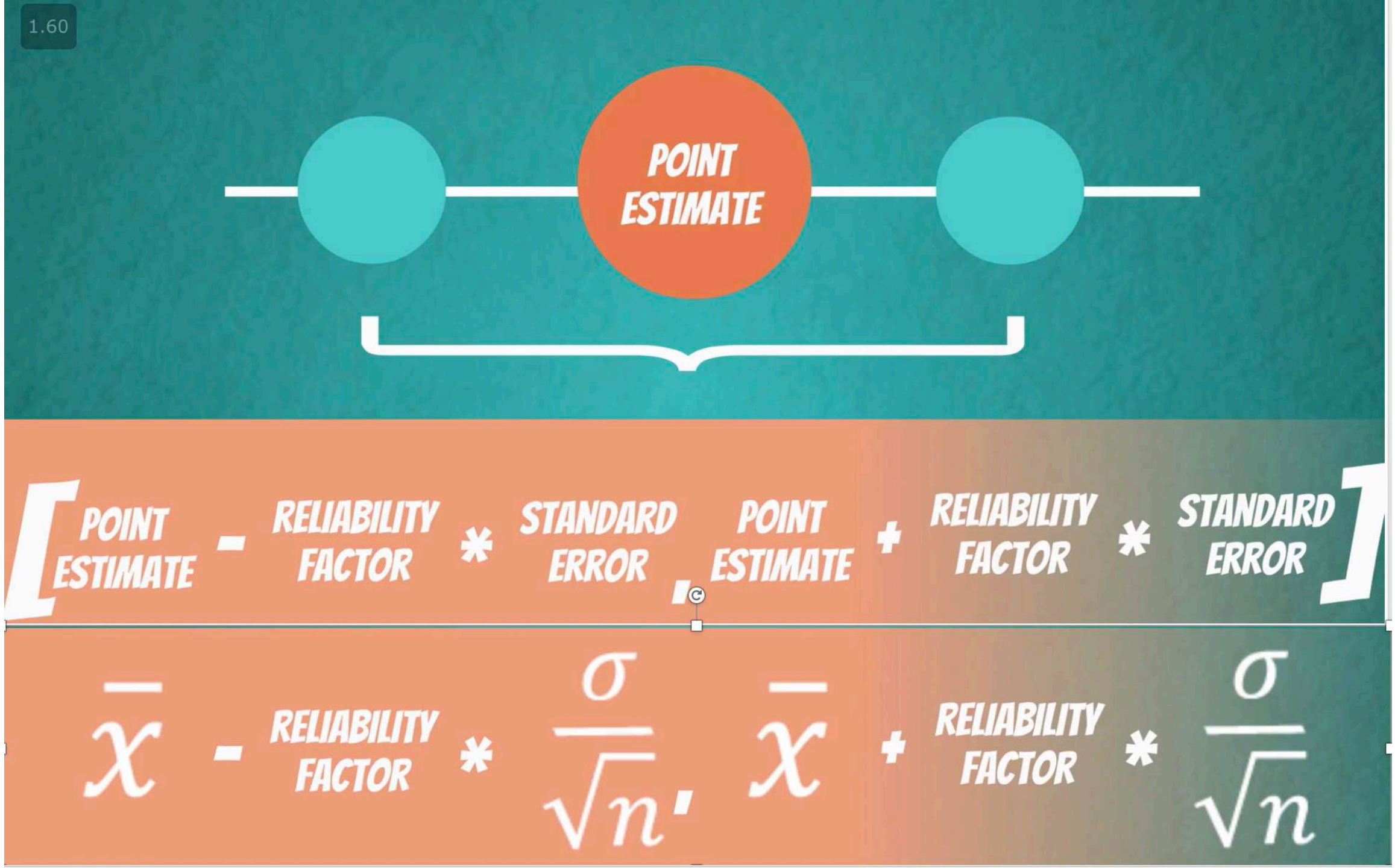
$$1-\alpha$$





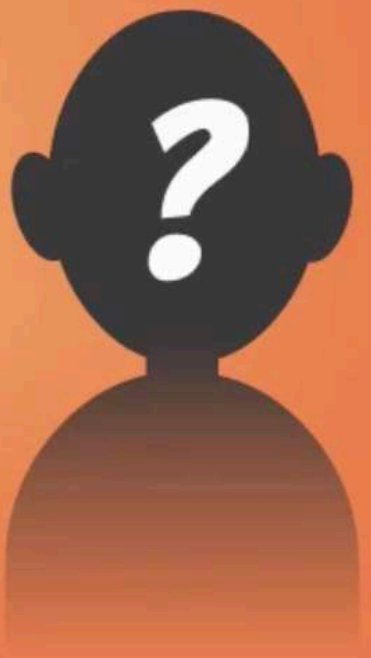
α

CONFIDENCE LEVEL


1.60
POINT ESTIMATE
[POINT ESTIMATE - RELIABILITY FACTOR * STANDARD ERROR , POINT ESTIMATE + RELIABILITY FACTOR * STANDARD ERROR]
$\bar{x}$ - RELIABILITY FACTOR * $\frac{\sigma}{\sqrt{n}}$, $\bar{x}$ + RELIABILITY FACTOR * $\frac{\sigma}{\sqrt{n}}$

CONFIDENCE INTERVALS

POPULATION VARIANCE



Known

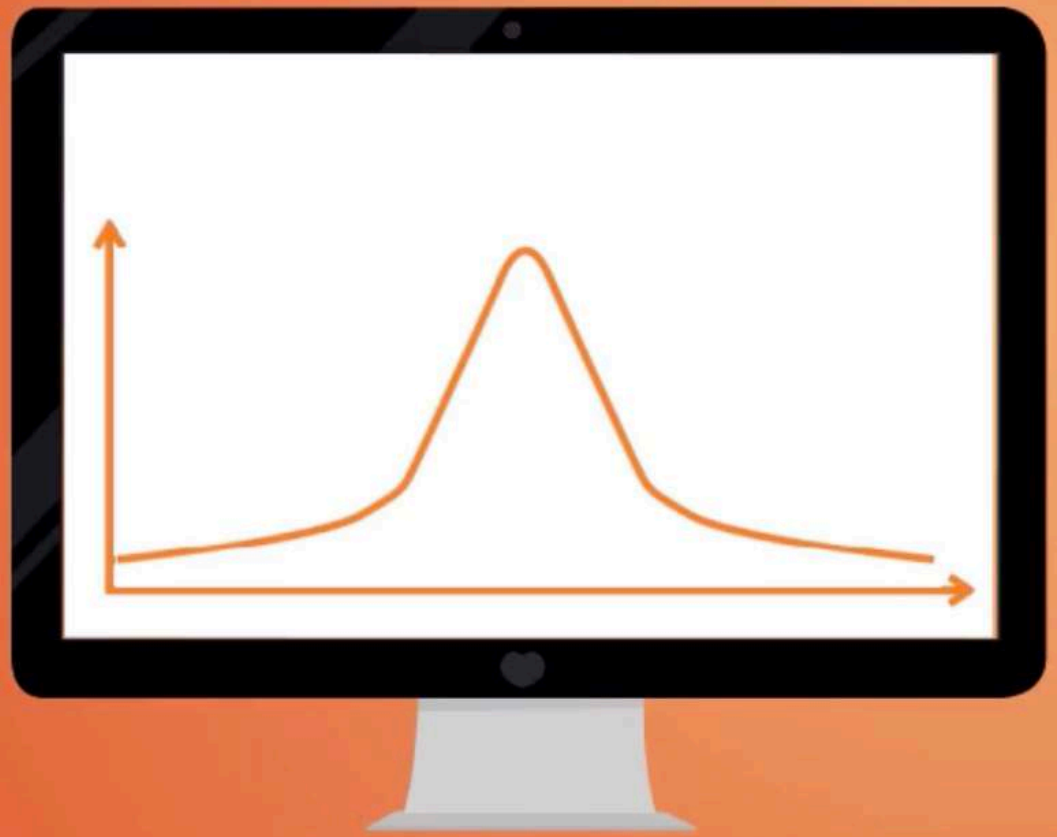
Unknown

CONFIDENCE INTERVALS

POPULATION VARIANCE

Known

CONFIDENCE INTERVALS
POPULATION VARIANCE



$$N \sim (\mu, \sigma^2)$$

CLT

CONFIDENCE INTERVALS
POPULATION VARIANCE

with a sample which is large enough you can assume sample means

Confidence Intervals. Population known, Z-score

[POINT ESTIMATE − RELIABILITY FACTOR * STANDARD ERROR , POINT ESTIMATE + RELIABILITY FACTOR * STANDARD ERROR]

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \;,\; \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

$$z \sim N(0,1)$$

$$\text{standard error} = \frac{\sigma}{\sqrt{n}}$$

common confidence levels

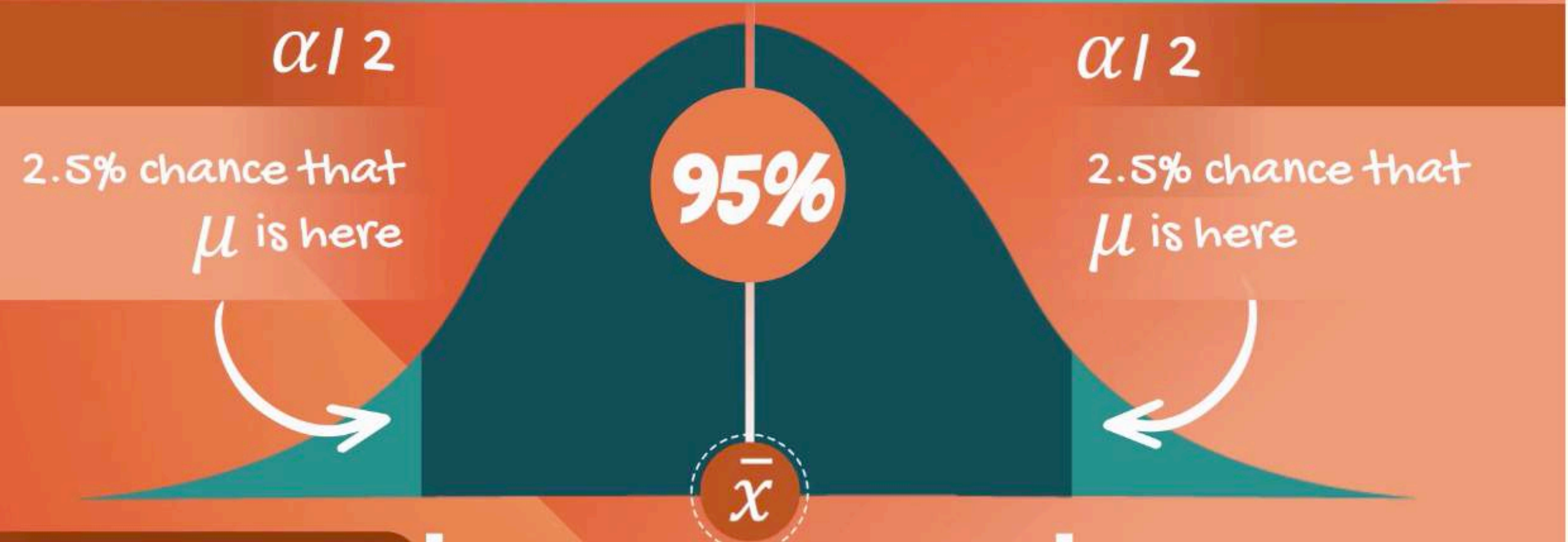
α = 90%, 95%, 99%

α = 10%, 5%, 1%

α = 0.1, 0.05, 0.01

A 95% confidence interval would imply we are 95% confident the true population mean falls within this interval

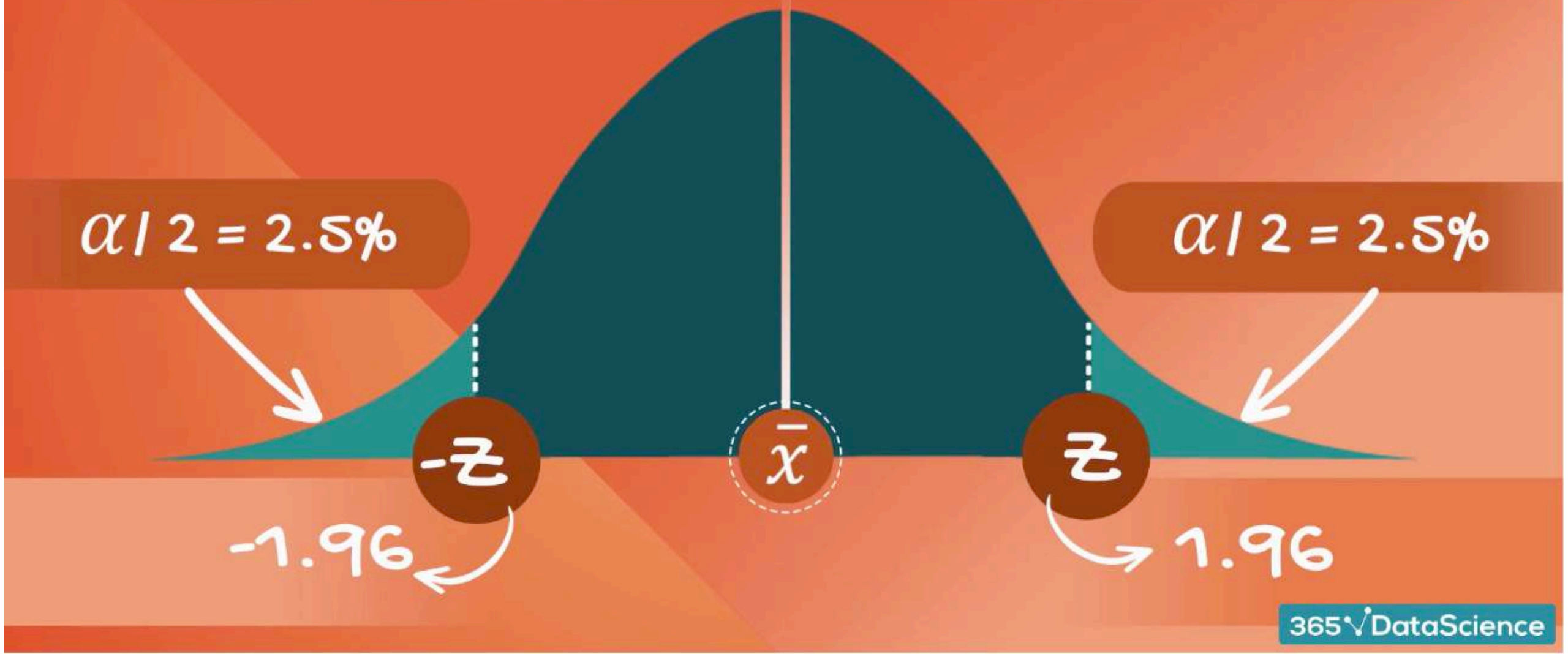
A 95% confidence interval would imply we are 95% confident the true population mean falls within this interval



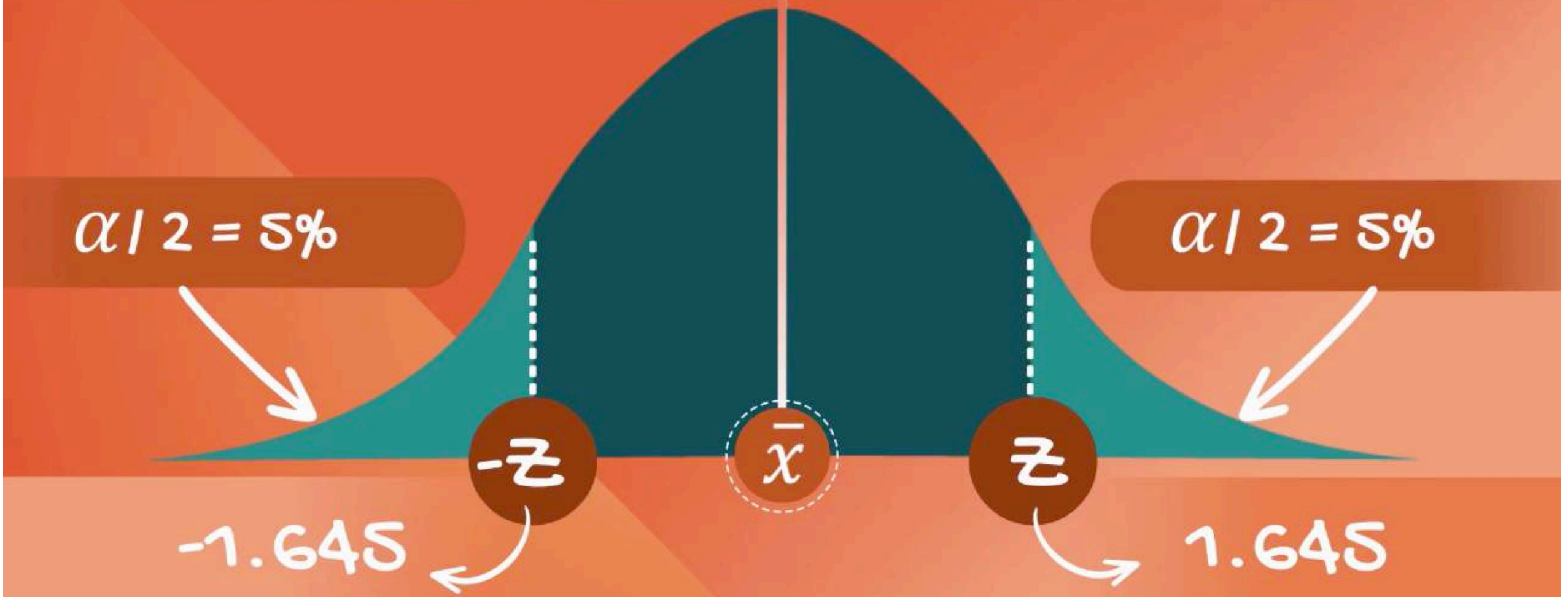
$\alpha = 5\%$

Confidence interval

$$95\%\ CI = \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$



$$90\%\ CI = \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$



1.80

$1-\alpha = 90\%$

When $1-\alpha$ is lower, CI is smaller

$1-\alpha = 99\%$

When $1-\alpha$ is higher, CI is larger

365 DataScience

File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do Share

A1

A 95% confidence interval means that
you are sure that in 95% of the cases,
the true population parameter would fall into the specified interval

Standard normal distribution

z-table

The table summarizes the standard normal distribution critical values and the corresponding (1-α)

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952

Confidence interval: 95%

$\alpha = 0.05$

$Z_{\alpha/2}$ $z_{0.025}$

$1 - 0.025 = 0.975$

$z_{0.025} = 1.9+0.06=1.96$

A commonly used term for the Z is 'critical value'

Ready

365 DataScience

Confidence intervals. Population known, z-score

Data scientist salary

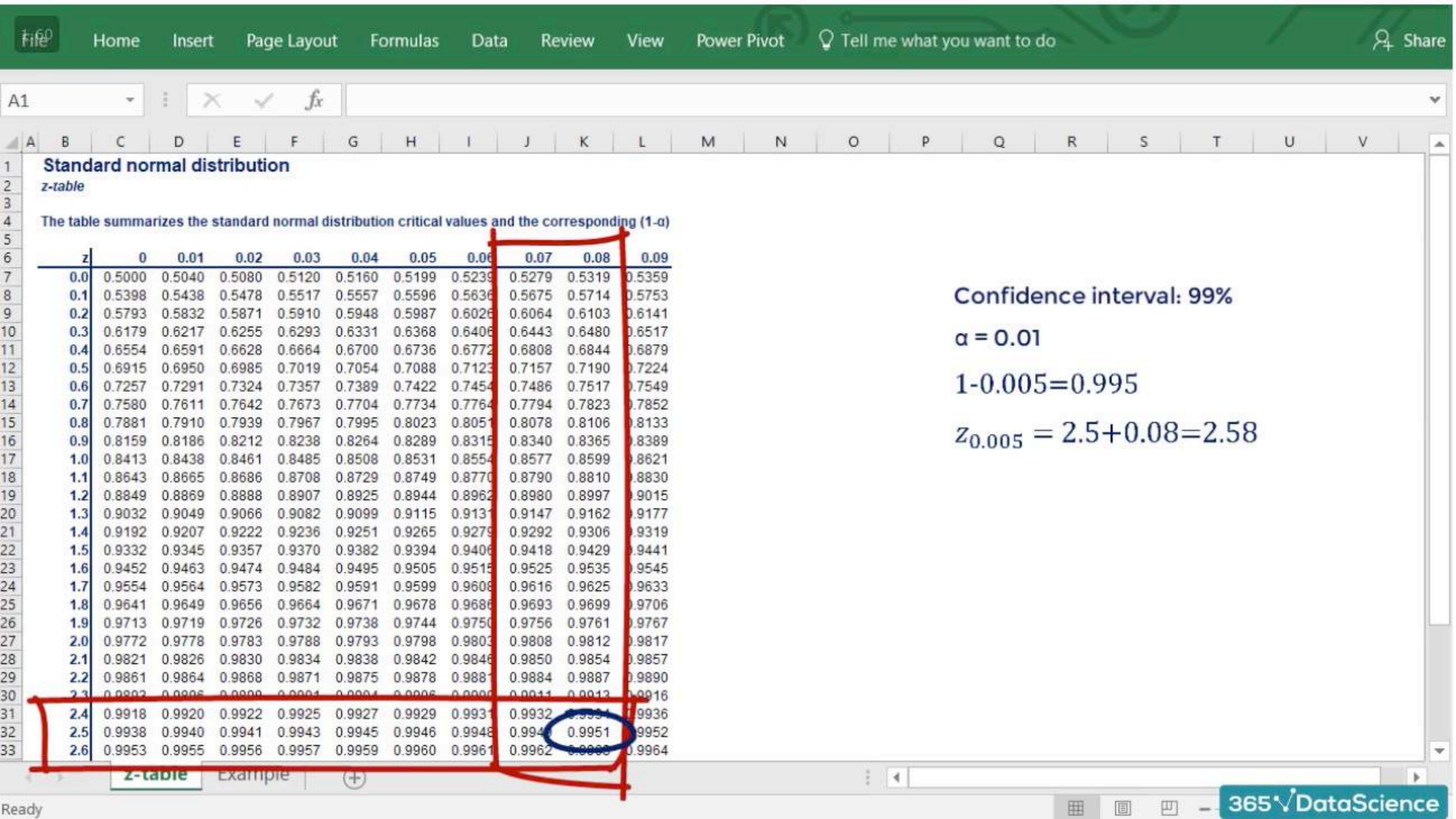
Dataset
$117,313
$104,002
$113,038
$101,936
$ 84,560
$113,136
$ 80,740
$100,536
$105,052
$ 87,201
$ 91,986
$ 94,868
$ 90,745
$102,848
$ 85,927
$112,276
$108,637
$ 96,818
$ 92,307
$114,564

Sample mean	$100,200
Population std	$ 15,000
Standard error	$ 2,739

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \;,\; \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

$$\left[100200 - 1.96\frac{15000}{\sqrt{30}} \;,\; 100200 + 1.96\frac{15000}{\sqrt{30}}\right] = [94833\,,\,105568]$$

We are 95% confident that the average data scientist salary will be in the interval [$94833 , $105568]



Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do Share

A1

Standard normal distribution

z-table

The table summarizes the standard normal distribution critical values and the corresponding (1-α)

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964

Confidence interval: 99%

α = 0.01

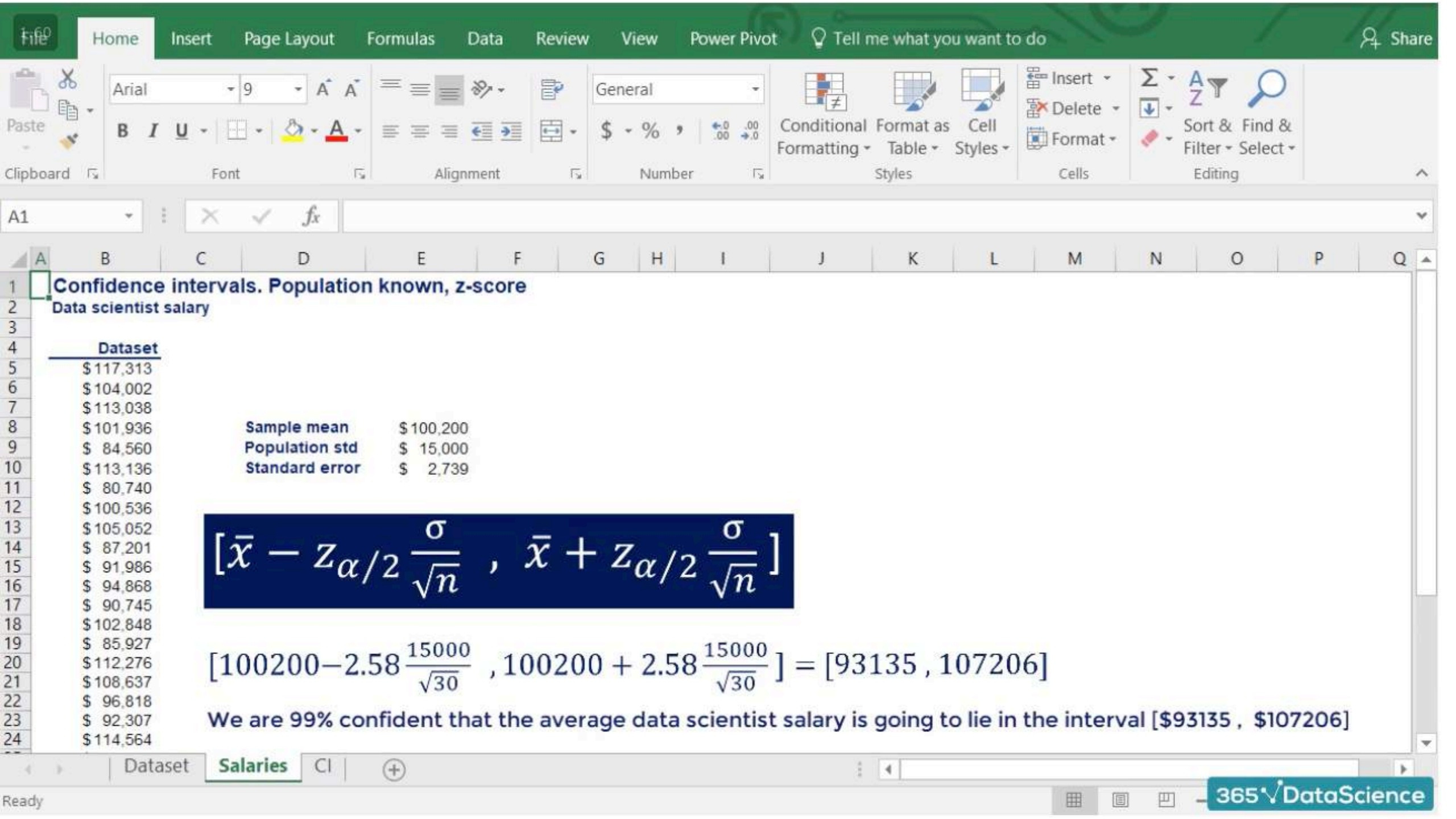
1-0.005=0.995

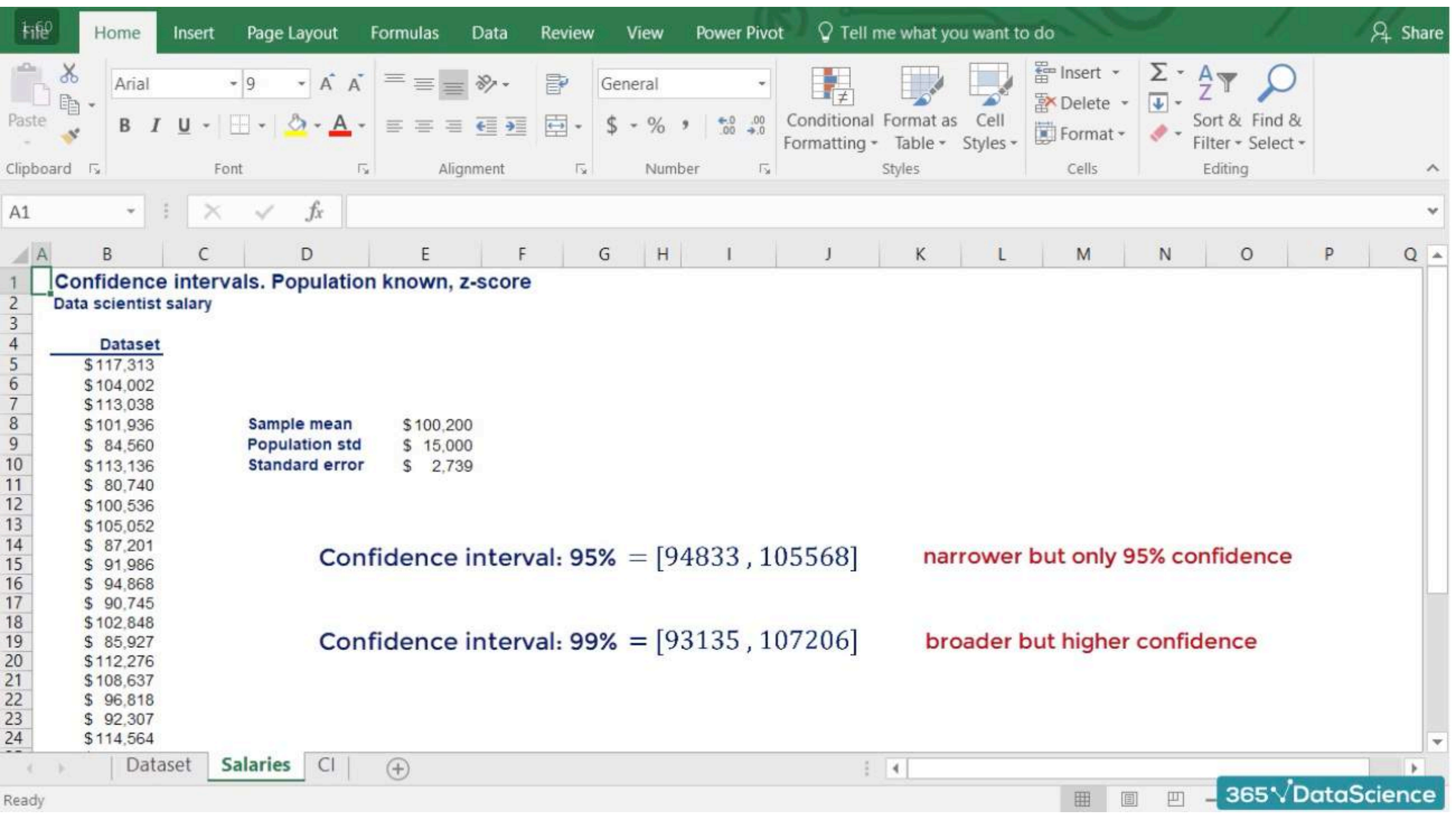
$z_{0.005} = 2.5+0.08=2.58$

z-table Example

Ready

365 DataScience


File
Home
Insert
Page Layout
Formulas
Data
Review
View
Power Pivot
Tell me what you want to do
Share
Paste
Arial
9
General
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
Sort & Filter
Find & Select
Clipboard
Font
Alignment
Number
Styles
Cells
Editing
A1
Confidence intervals. Population known, z-score
Data scientist salary
Dataset
$117,313
$104,002
$113,038
$101,936
$ 84,560
$113,136
$ 80,740
$100,536
$105,052
$ 87,201
$ 91,986
$ 94,868
$ 90,745
$102,848
$ 85,927
$112,276
$108,637
$ 96,818
$ 92,307
$114,564
Sample mean $100,200
Population std $ 15,000
Standard error $ 2,739
$[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} , \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$
$[100200-2.58\frac{15000}{\sqrt{30}} , 100200 + 2.58\frac{15000}{\sqrt{30}}] = [93135 , 107206]$
We are 99% confident that the average data scientist salary is going to lie in the interval [$93135 , $107206]
Dataset
Salaries
CI
Ready
365 DataScience

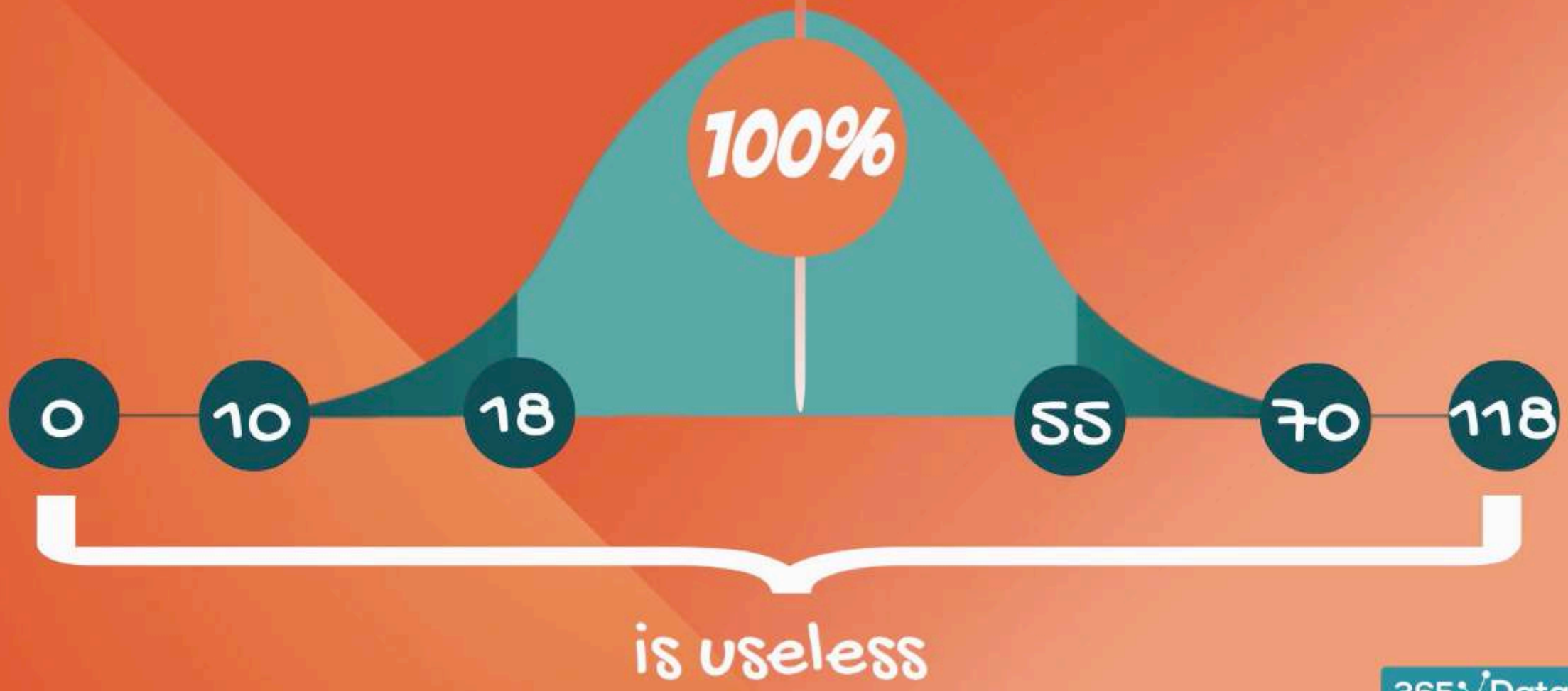

File
Home
Insert
Page Layout
Formulas
Data
Review
View
Power Pivot
Tell me what you want to do
Share
Paste
Arial
9
General
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
Sort & Filter
Find & Select
Clipboard
Font
Alignment
Number
Styles
Cells
Editing
A1
Confidence intervals. Population known, z-score
Data scientist salary
Dataset
$117,313
$104,002
$113,038
$101,936
$ 84,560
$113,136
$ 80,740
$100,536
$105,052
$ 87,201
$ 91,986
$ 94,868
$ 90,745
$102,848
$ 85,927
$112,276
$108,637
$ 96,818
$ 92,307
$114,564
Sample mean $100,200
Population std $ 15,000
Standard error $ 2,739
Confidence interval: 95% = [94833 , 105568]
narrower but only 95% confidence
Confidence interval: 99% = [93135 , 107206]
broader but higher confidence
Dataset
Salaries
CI
Ready
365 DataScience

1.80

AGE INTERVALS

I don't know your age, dear student, but I am 95% confident that you are between 18 and 55 years old, because you are taking an online statistics course.

AGE INTERVALS



AGE INTERVALS

AGE INTERVALS

Range of 5% confidence is too small to be usefull for analysis, although it is an exact number.

AGE INTERVALS

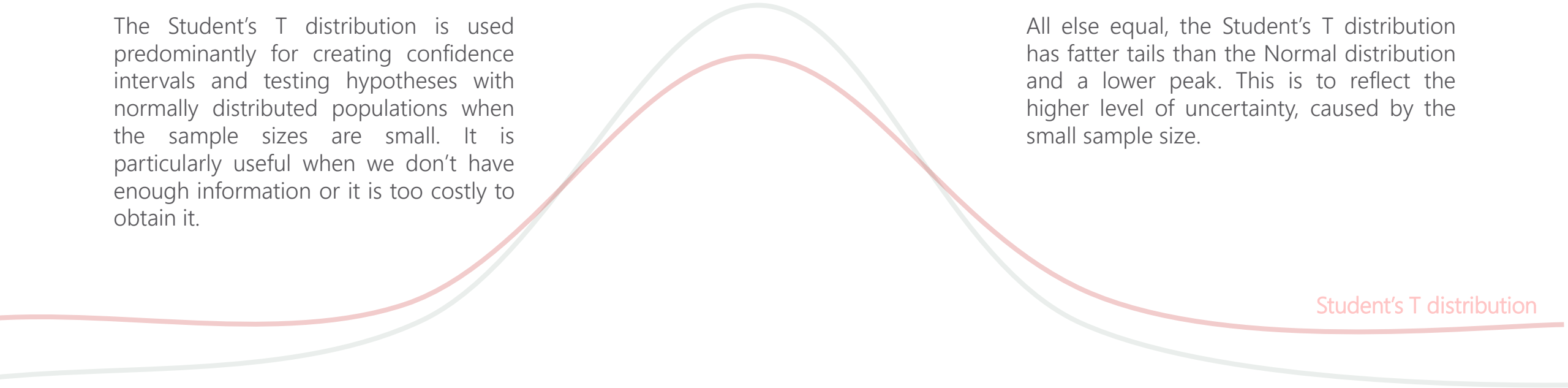
AGE INTERVALS

95% is the accepted norm, as we don't compromise with accuracy too much, but still get a relatively narrow interval.

Student's T Distribution

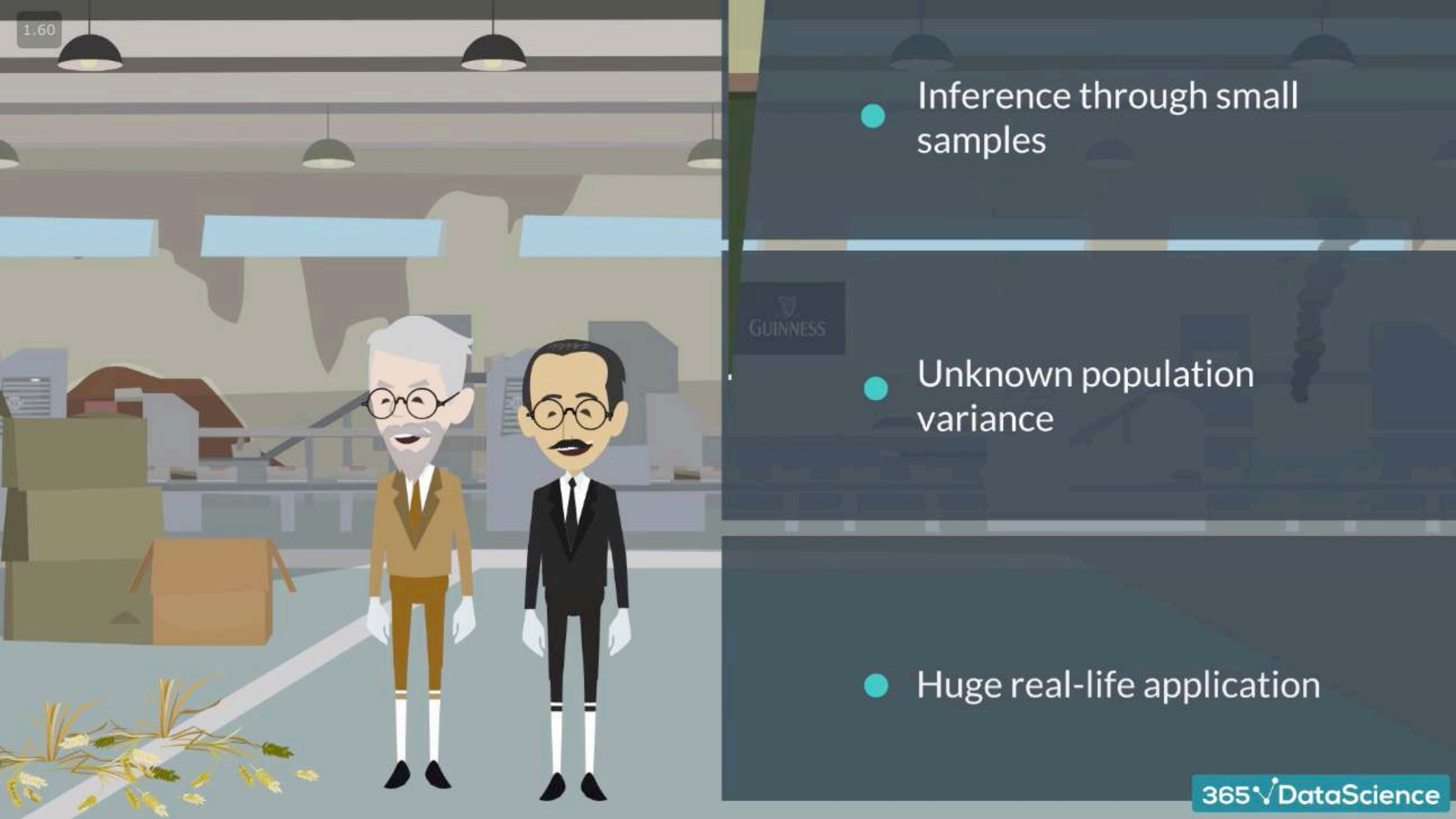
The Student's T distribution is used predominantly for creating confidence intervals and testing hypotheses with normally distributed populations when the sample sizes are small. It is particularly useful when we don't have enough information or it is too costly to obtain it.

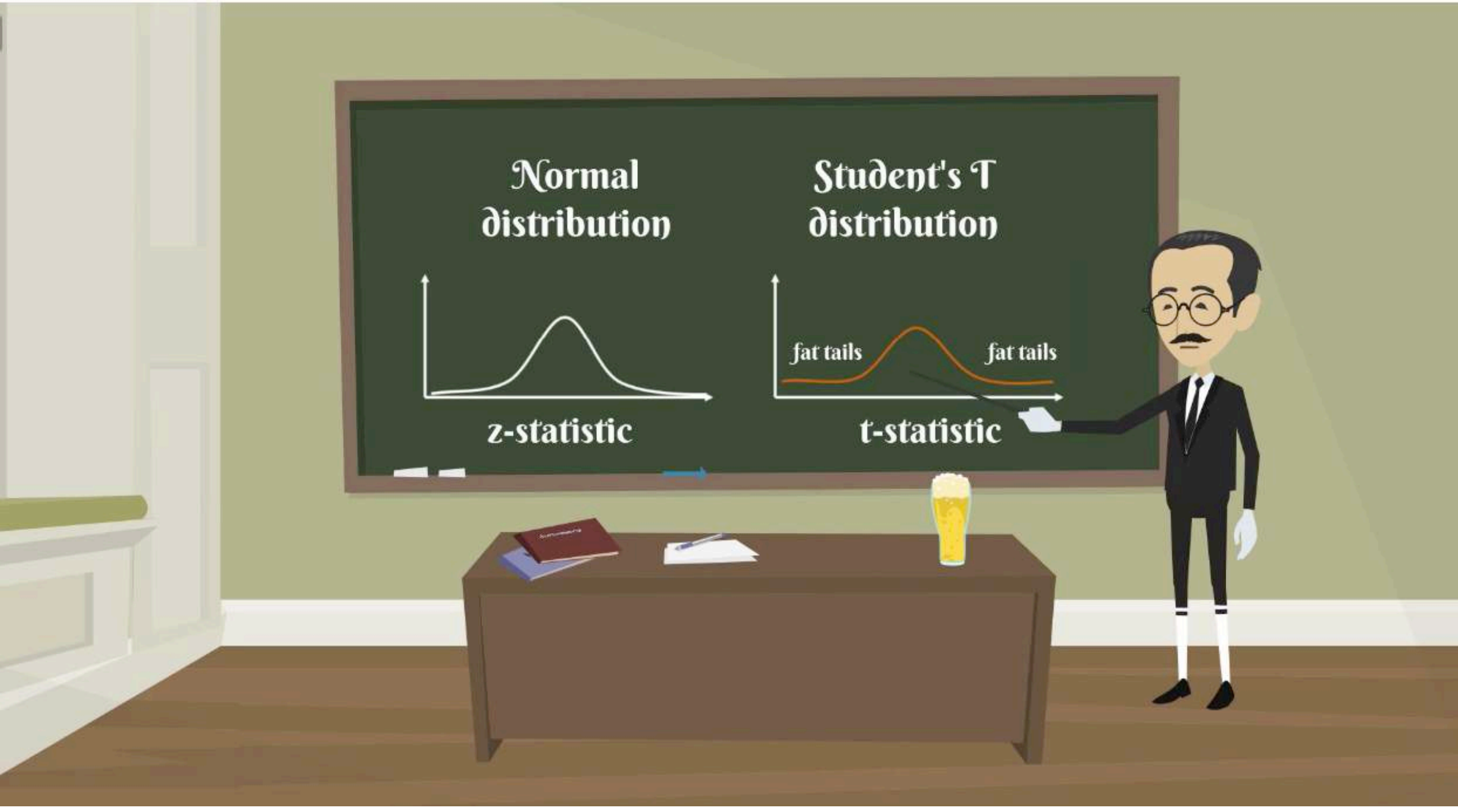
All else equal, the Student's T distribution has fatter tails than the Normal distribution and a lower peak. This is to reflect the higher level of uncertainty, caused by the small sample size.

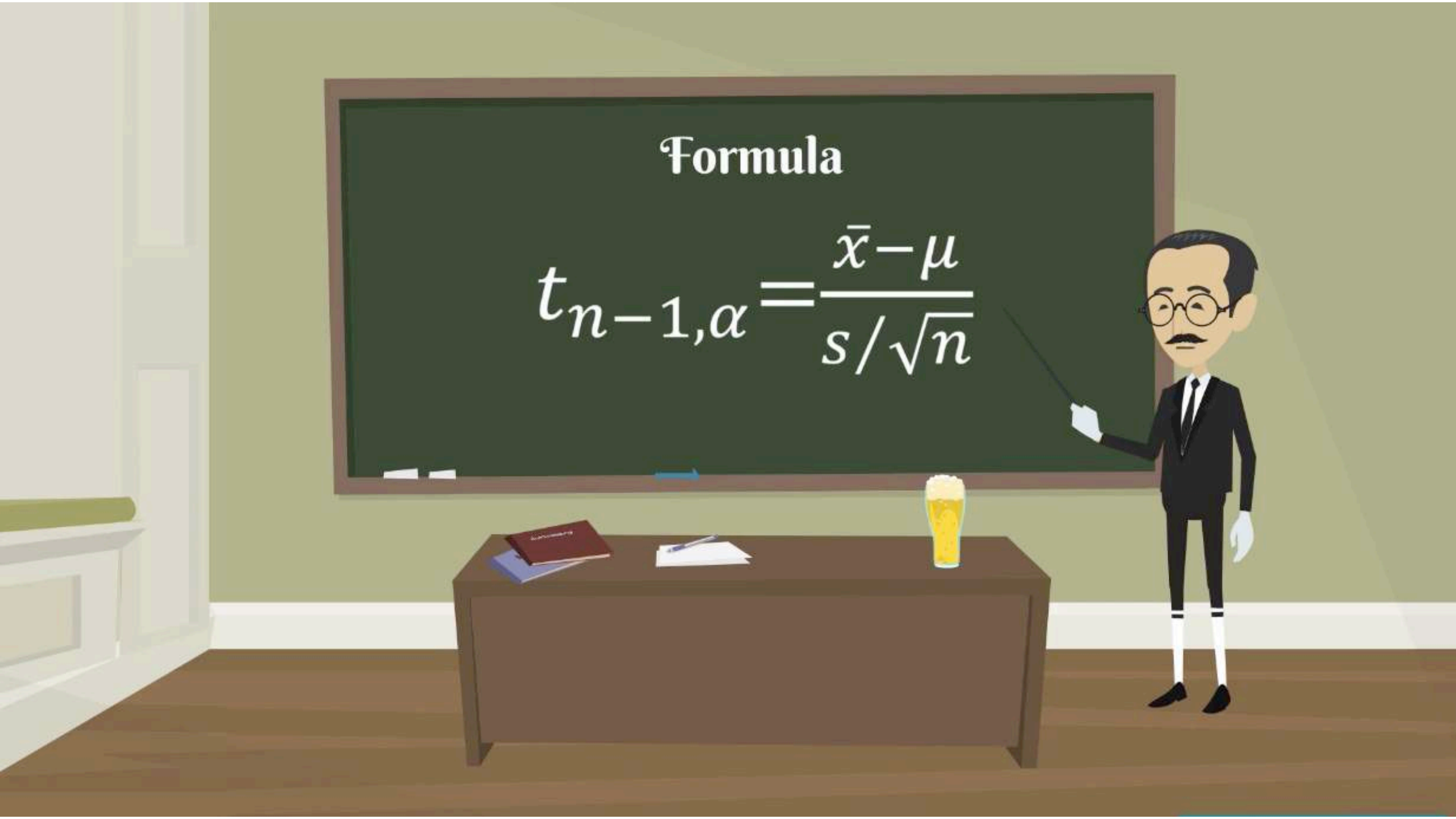


A random variable following the t-distribution is denoted $t_{\upsilon,\alpha}$, where υ are the degrees of freedom.

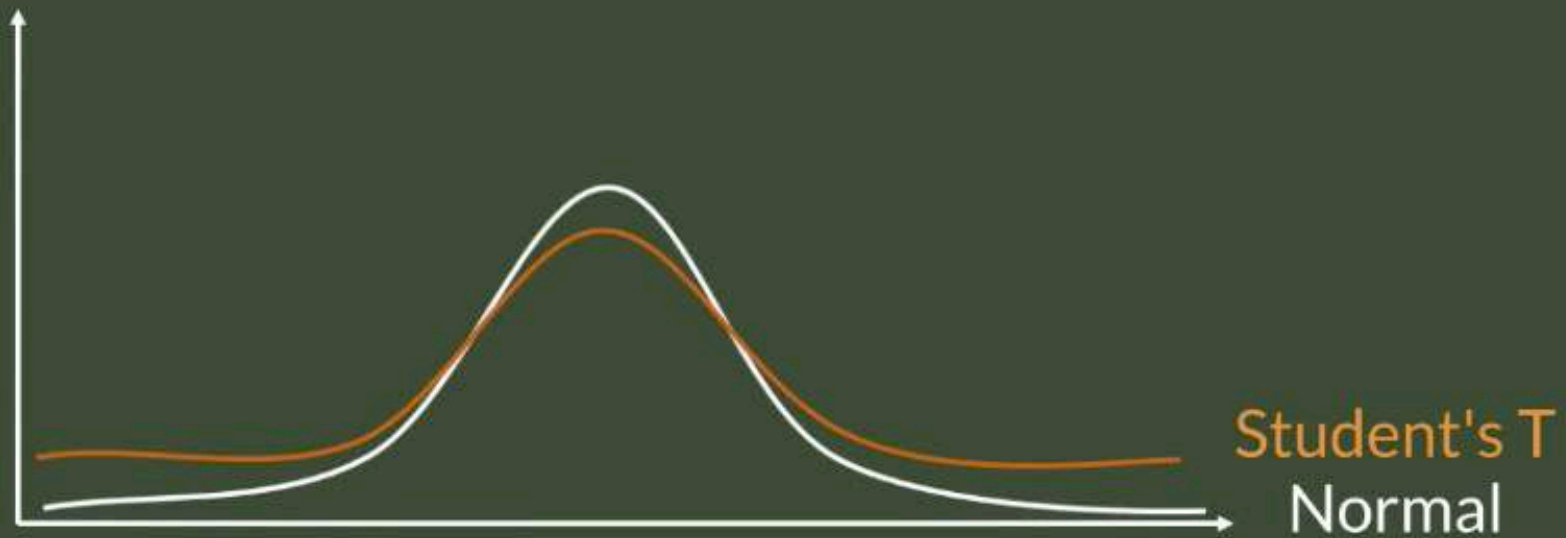
We can obtain the student's T distribution for a variable with a Normally distributed population using the formula: $t_{\upsilon,\alpha} = \frac{\bar{x}-\mu}{s/\sqrt{n}}$


1.60
GUINNESS
Inference through small samples
Unknown population variance
Huge real-life application
365 DataScience


Normal distribution
Student's T distribution
fat tails
fat tails
z-statistic
t-statistic


Formula
$t_{n-1,\alpha} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$

Approximation of the Normal



Degrees of freedom (d.f.)

$$t_{n-1,\alpha} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

sample size: n
d.f.: n-1

Population variance unknown, t-score

Population variance unknown

$$\bar{x} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$$

Population variance known

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

standard error = $\frac{s}{\sqrt{n}}$

Population variance is unknown. The sample size is small => Student's T distribution

d.f. / α	0.1	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
25	1.316	1.708	2.060	2.485	2.787
30	1.310	1.697	2.042	2.457	2.750
35	1.306	1.690	2.030	2.438	2.724
40	1.303	1.684	2.021	2.423	2.704
50	1.299	1.676	2.009	2.403	2.678
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
inf.	1.282	1.645	1.960	2.326	2.576

coincides with the z-statistic

Rule of thumb:
When the sample is bigger than 50, we use the z-statistic

t-table

d.f. / α	0.1	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
35	1.306	1.690	2.030	2.438	2.724
40	1.303	1.684	2.021	2.423	2.704
50	1.299	1.676	2.009	2.403	2.678
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
inf	1.282	1.645	1.960	2.326	2.576
CI	80%	90%	95%	98%	99%

$$t_{n-1,\alpha/2}$$

$$t_{8,0.025} = 2.31$$

95% CI => alpha = 5%

Question: How much the mean data scientist salary is?

Confidence intervals, t-score

Data scientist salary

Data set
$ 78,000
$ 90,000
$ 75,000
$ 117,000
$ 105,000
$ 96,000
$ 89,500
$ 102,300
$ 80,000

Sample mean	$ 92,533
Sample standard deviation	$ 13,932
Standard error	$ 4,644
t-stat 95%	2.31

n = 9

$$\bar{x} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$$

$$92.533 \pm 2.31 * 4{,}644$$

95% CI => alpha = 5%

$$CI_{95\%,unknown} = (\$\ 81806\ ,\ \$\ 103261)$$

When population variance is unknown => t-statistic

$$CI_{95\%,unknown} = (\$\ 81806\ ,\ \$\ 103261) \quad \text{width} = \$21{,}455$$

$$CI_{95\%,known} = (\$\ 94833\ ,\ \$\ 105568) \quad \text{width} = \$10{,}735$$

*Here we've got two effects: 1) smaller sample size and 2) unknown population variance
Both contribute to the width of the interval

Less accurate when variance is unknown

Formulas for Confidence Intervals

# populations	Population variance	Samples	Statistic	Variance	Formula
One	known	-	z	σ^2	$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
One	unknown	-	t	s^2	$\bar{x} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$
Two	-	dependent	t	$s^2_{difference}$	$\bar{d} \pm t_{n-1,\alpha/2} \frac{s_d}{\sqrt{n}}$
Two	Known	independent	z	σ_x^2 , σ_y^2	$(\bar{x} - \bar{y}) \pm z_{\alpha/2} \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$
Two	unknown, assumed equal	independent	t	$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$	$(\bar{x} - \bar{y}) \pm t_{n_x+n_y-2,\alpha/2} \sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}$
Two	unknown, assumed different	independent	t	s_x^2 , s_y^2	$(\bar{x} - \bar{y}) \pm t_{\upsilon,\alpha/2} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$

1 POPULATION

Known

Unknown

CONFIDENCE INTERVALS FORMULAS

POPULATION VARIANCE

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Known

$$\bar{x} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$$

Unknown

MARGIN OF ERROR = ME

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Known

$$\bar{x} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$$

Unknown

CONFIDENCE INTERVAL = $\bar{x} \pm ME$

$[\ ,\ ]$

smaller margin of error =>
narrower confidence interval

WE CAN CONTROL THE MARGIN OF ERROR

ME = RELIABILITY FACTOR $\times$ $\frac{\text{STD}}{\sqrt{n}}$

z, T, ...

1.60

ME = RELIABILITY FACTOR $*$ $\frac{\text{STD}}{\sqrt{n}}$

z, t, ...

IN THE NUMERATOR

Z OR T STATISTIC

$z_{\alpha/2}$

$t_{n-1,\alpha/2}$

STANDARD DEVIATION

SAMPLE SIZE

CONFIDENCE INTERVAL = $\bar{x} \pm ME$

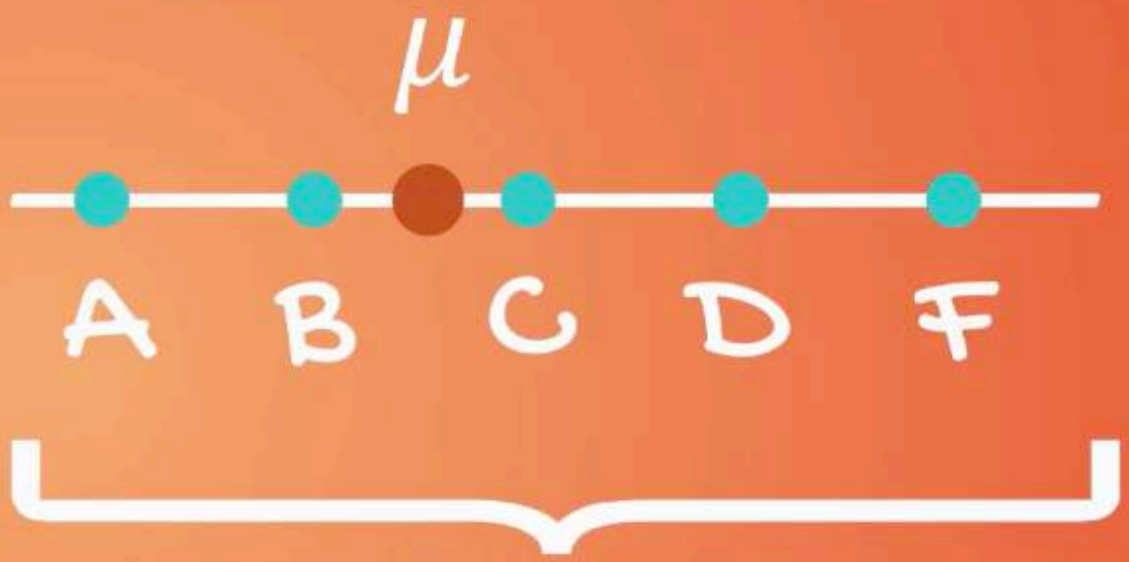
[,]

[,]

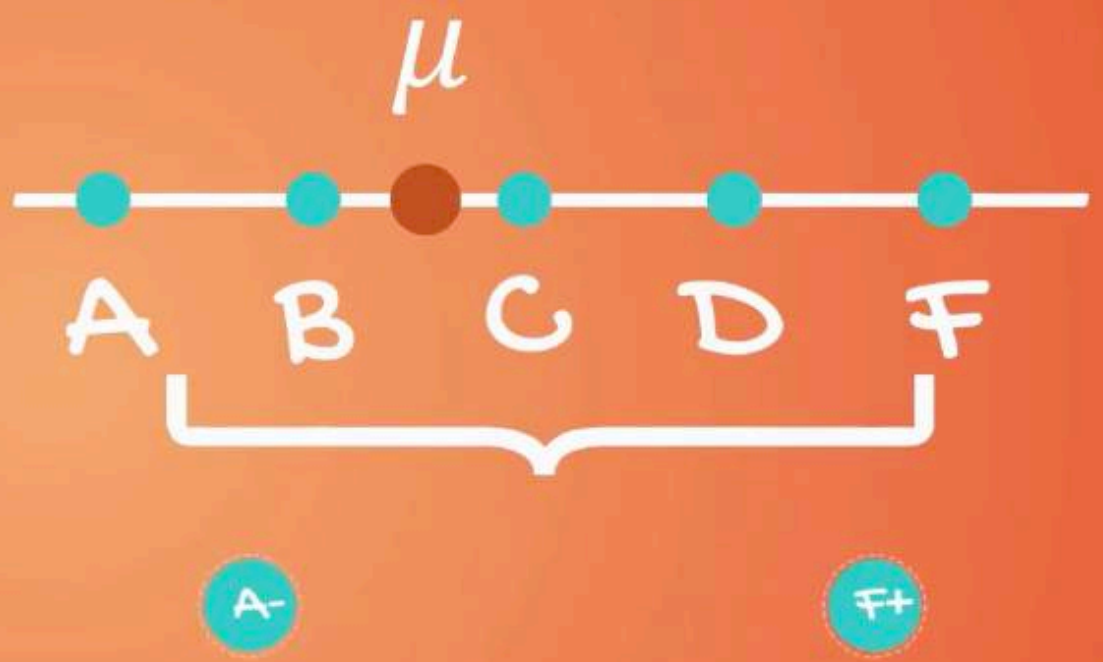
bigger margin of error =>
wider confidence interval

smaller margin of error =>
narrower confidence interval

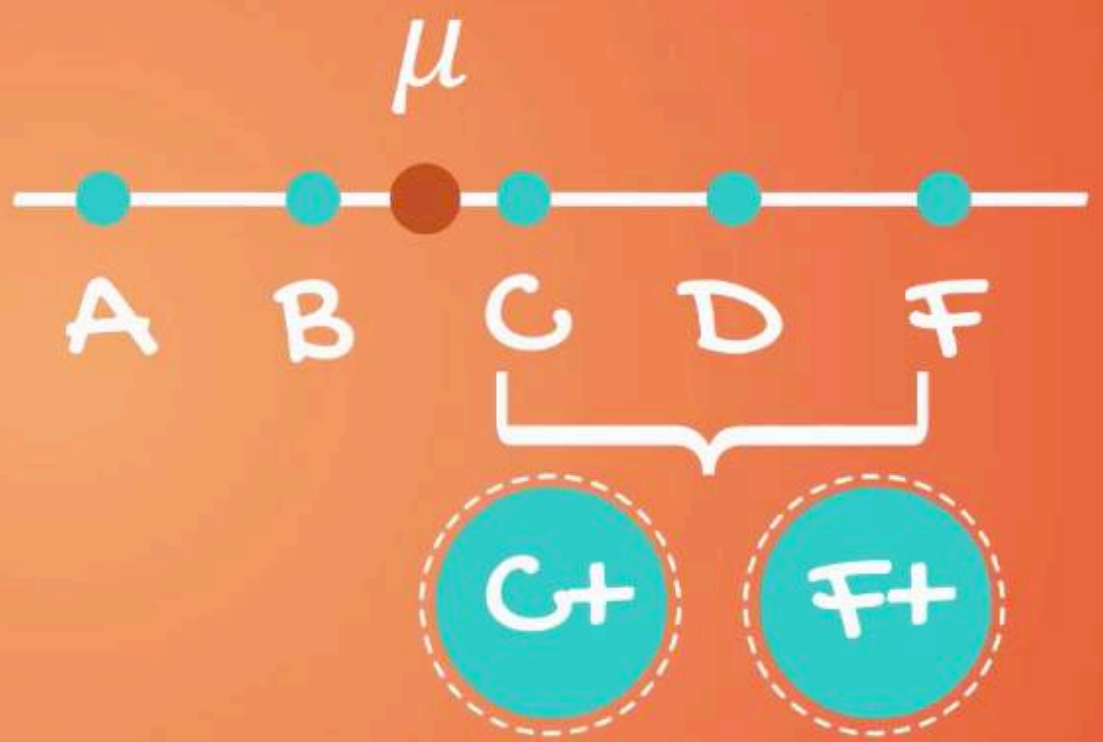
100% CONFIDENCE, $\alpha = 0$



99% CONFIDENCE, $\alpha = 0.01$

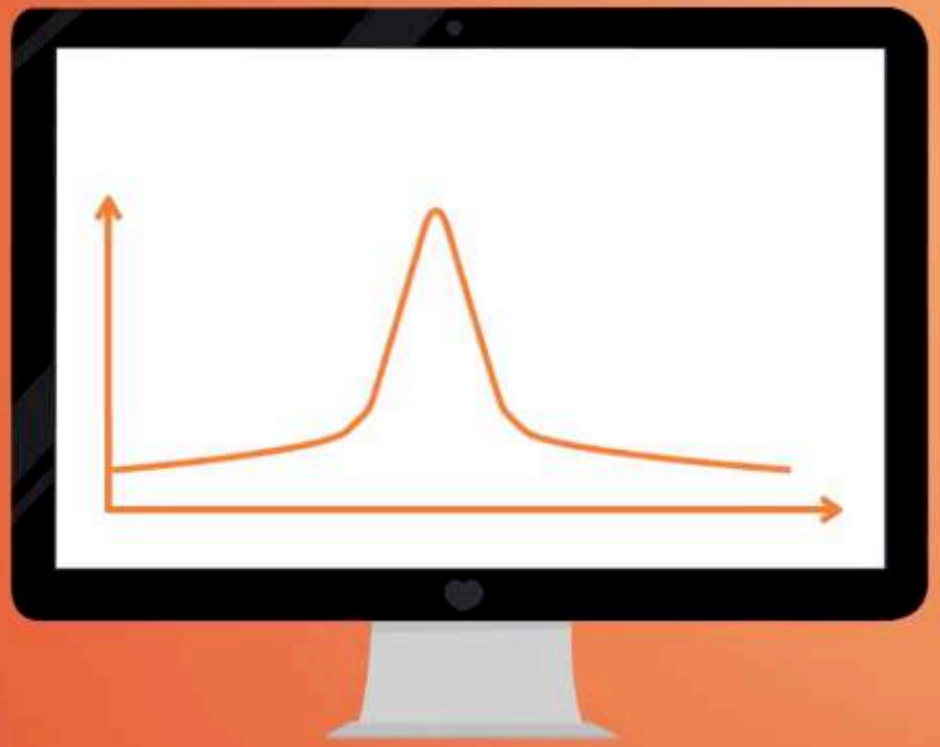


50% CONFIDENCE, α : 0.5



ME = RELIABILITY FACTOR $*$ $\frac{\text{STD}}{\sqrt{n}}$

z, t, ...

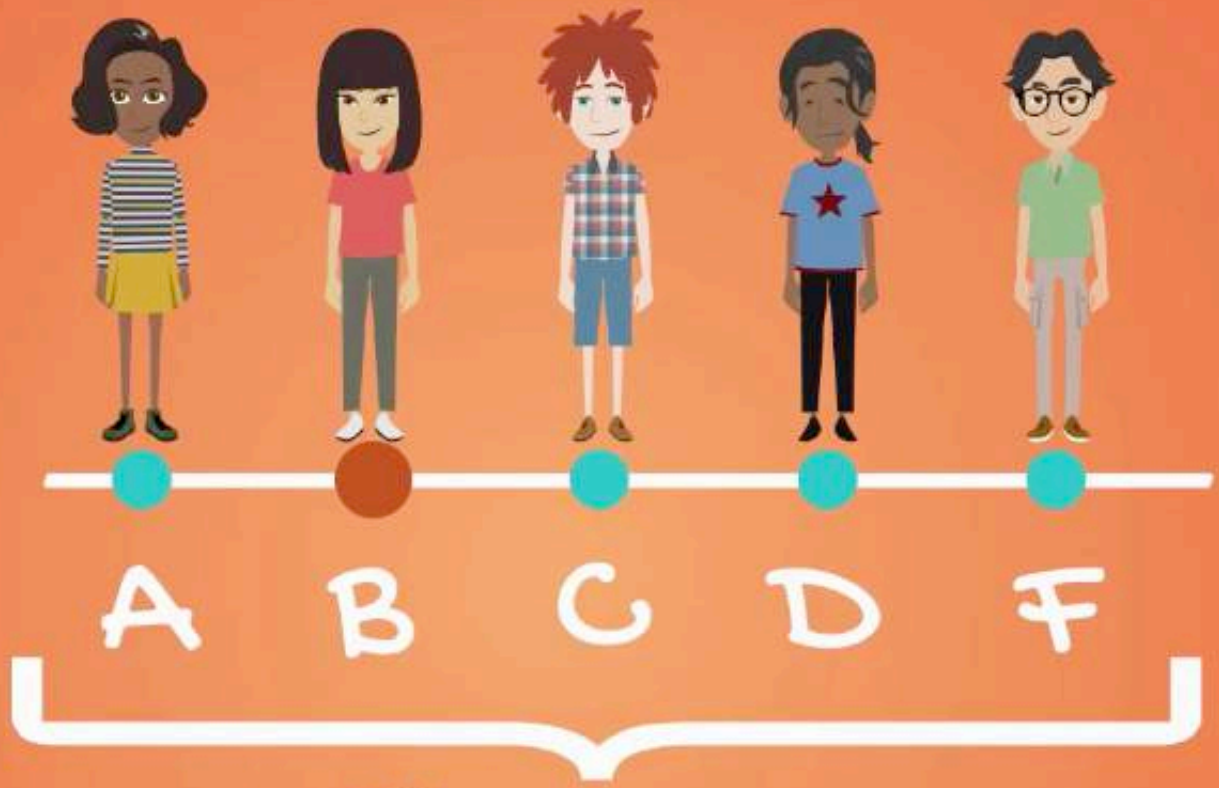


z, t

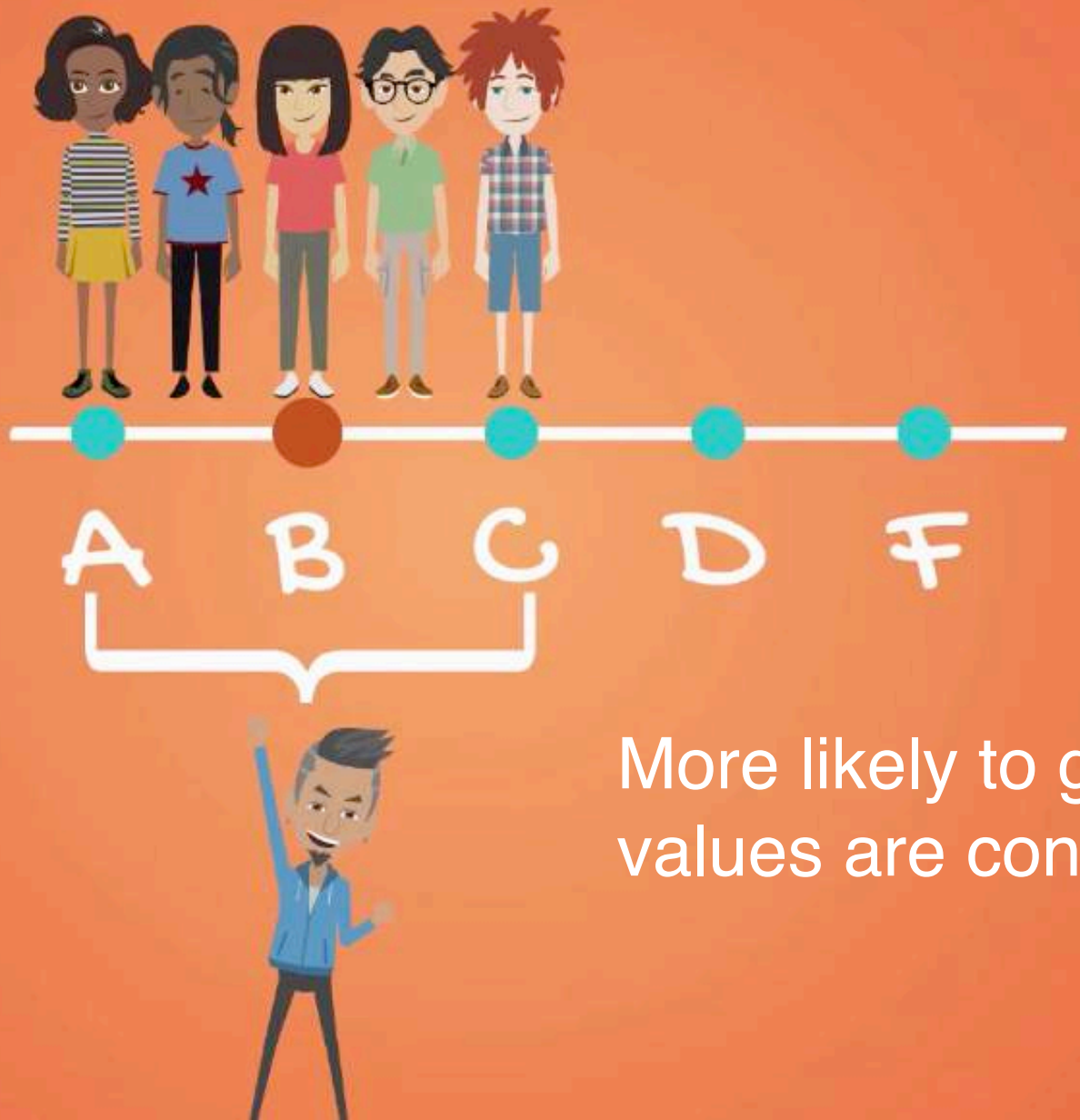
or

STD

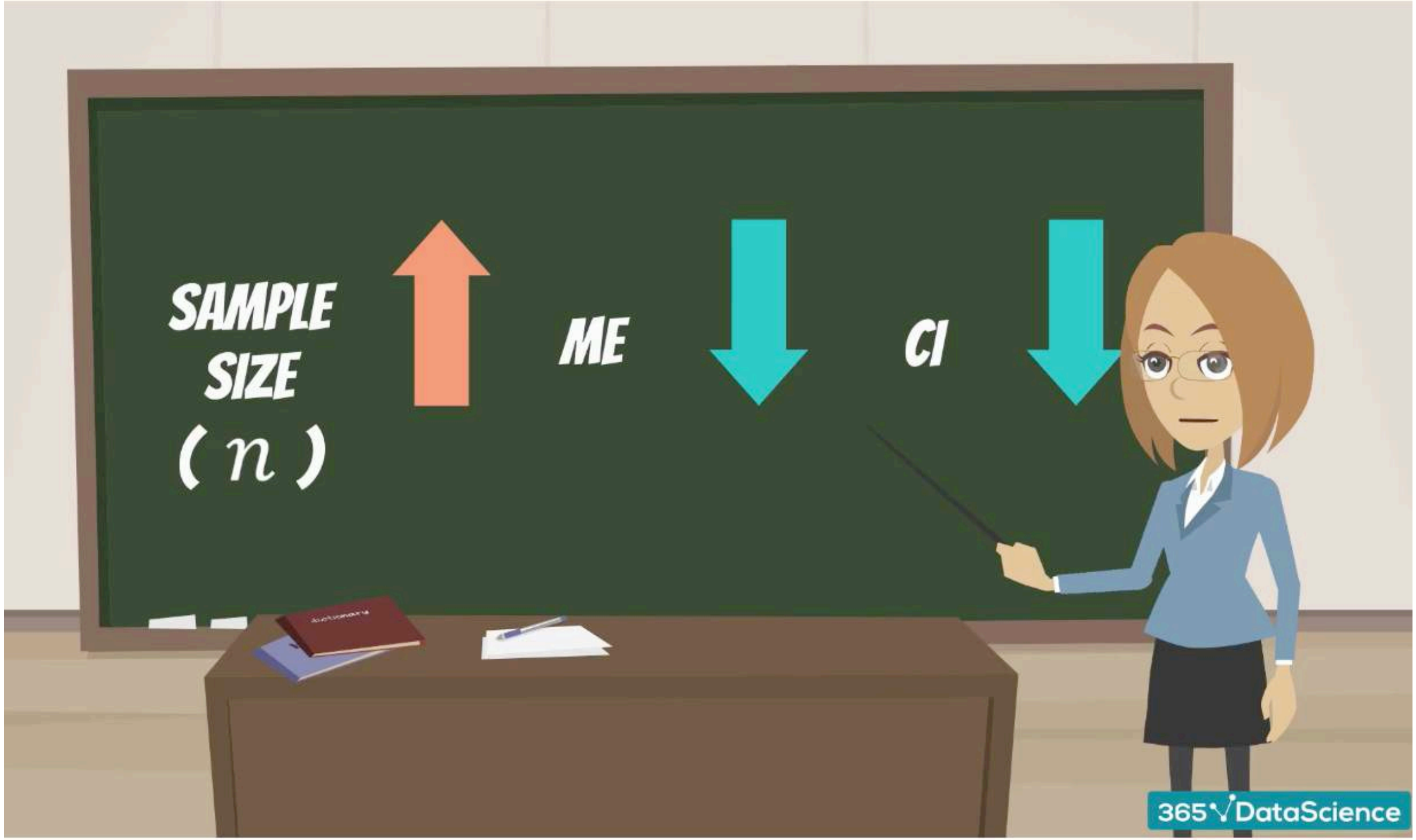
} **ME**



Likelihood to get a B is lower
having a big dispersion



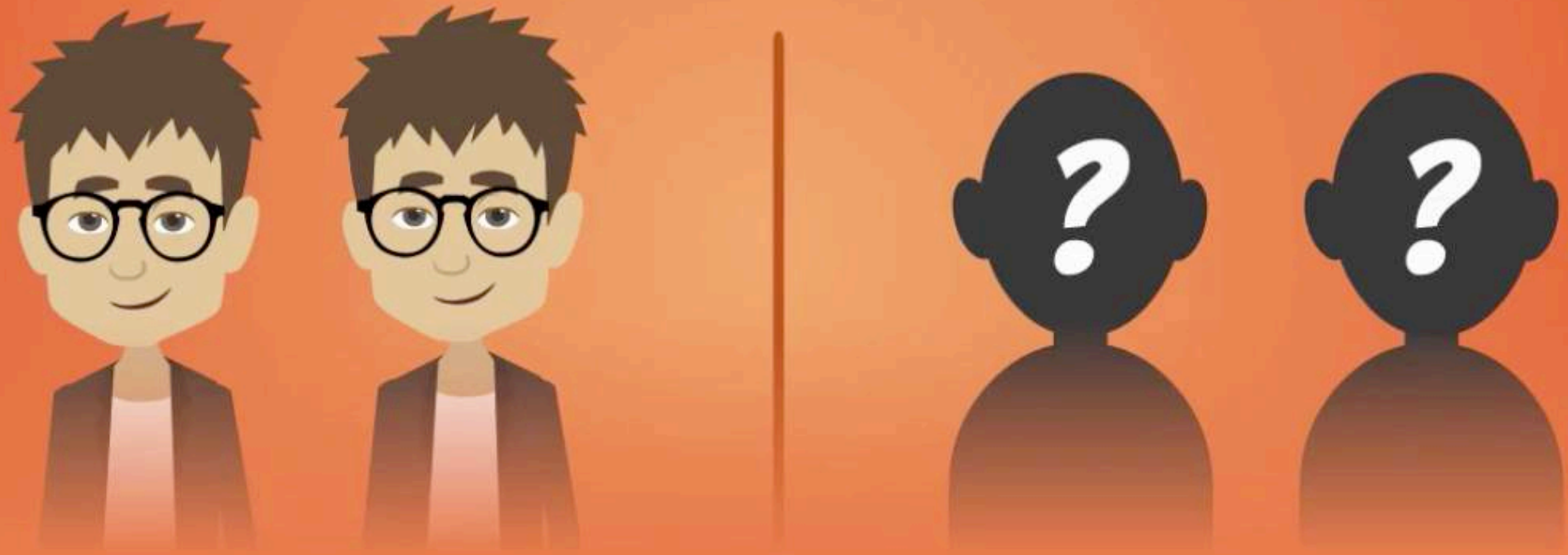
More likely to get a B as values are concentrated


SAMPLE SIZE (n)
ME
CI
365 DataScience

CONCLUSION

the more observations there are in the sample, the higher the chances of getting a good idea about the true mean of the entire population

2 POPULATIONS



Known

Unknown

SAMPLES

DEPENDENT

INDEPENDENT

SAMPLES

DEPENDENT

- before and after situation

INDEPENDENT

Dependent examples

WHEN WE ARE RESEARCHING THE SAME SUBJECT OVER TIME

weight loss

before and after

blood samples

1.40

SAMPLES

DEPENDENT

- before and after situation
- cause and effect

INDEPENDENT


1.40
A+
FAIL
EXAMPLE
One relates to the SAT
the other to the admittance outcome

TESTING APPROACHES

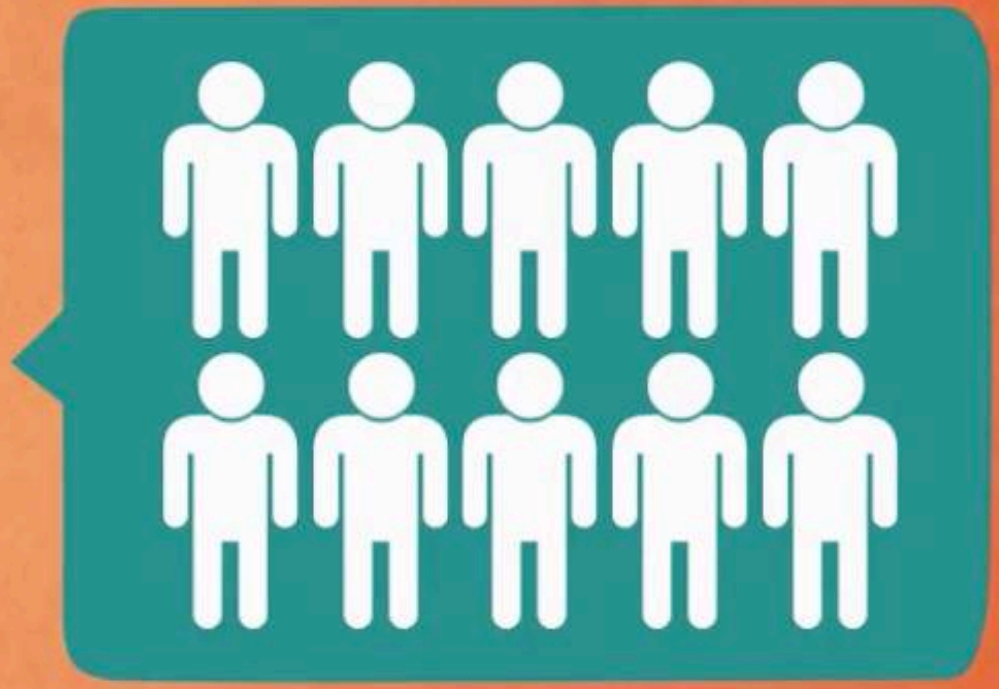
- confidence intervals for dependent samples
- statistical methods like regressions

DEPENDENT SAMPLES

You are developing a drug that increases Mg on the blood.

After testing it in the lab, you want to see the effect on people (data)

DEPENDENT SAMPLES



- before and after situation

IN BIOLOGY

$$N \sim \left(\mu, \frac{\sigma^2}{n}\right)$$

normality is so often observed that we assume that such variables are normally distributed

MAGNESIUM LEVELS

1.7 - 2.2 mg/dL

Healthy person

Confidence interval for difference of two means, dependent samples

Magnesium example

Patient	Before	After	Difference
1	2.00	1.70	-0.30
2	1.40	1.70	0.30
3	1.30	1.80	0.50
4	1.10	1.30	0.20
5	1.80	1.70	-0.10
6	1.60	1.50	-0.10
7	1.50	1.60	0.10
8	0.70	1.70	1.00
9	0.90	1.70	0.80
10	1.50	2.40	0.90

Mean	0.33
St. deviation	0.45

n = 10

Difference = After - Before

Confidence interval for difference of two means, dependent samples formula

$$\bar{d} \pm t_{n-1,\alpha/2} \frac{s_d}{\sqrt{n}}$$

=

Confidence interval for a single population. Population variance unknown formula

$$\bar{x} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$$

In this way, the data looks as a single population

95% confidence is one of the most common ones

How do we interpret this result?

1. In 95% of the cases, the true mean will fall in this interval
2. The whole interval is positive
3. The levels of Mg in the test subjects' blood is higher

=> based on our small sample, the pill IS EFFECTIVE

$$0.33 \pm 2.26 \frac{0.45}{\sqrt{10}} = (0.01, 0.65)$$

SAMPLES

DEPENDENT

- before and after situation
- cause and effect

INDEPENDENT

- Population variance known
- Population variance unknown but assumed to be equal
- Population variance unknown but assumed to be different

Confidence interval for the difference of two means. Independent samples. Variance known

Considerations:

1. The populations are normally distributed
2. The population variances are known
3. The sample sizes are different

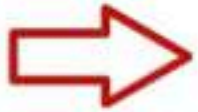
Considerations:

1. Different departments
2. Different teachers
3. Different grades
4. Different exams

The two samples are truly independent

Considerations:

1. Samples are big
2. Population variances are known
3. Populations are assumed to follow the Normal distribution



Z

Problem: We want to find a 95% confidence interval for the difference between the grades of the students from engineering and management

Confidence interval for the difference of two means. Independent samples, variance known

University example

	x Engineering	y Management	x-y Difference
Size	100	70	?
Sample mean	58	65	-7.00
Population std	10	5	1.16
95% z-stat	1.96		

← standard error

'From past years, we know that the population standard deviation is 10 percentage points.'

Thus, the variance is known

$$(\bar{x} - \bar{y}) \pm z_{\alpha/2} \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$$

difference point estimator

test statistic

standard error

Variance of the difference

$$\sigma_{diff}^2 = \frac{\sigma_e^2}{n_e} + \frac{\sigma_m^2}{n_m}$$

$$\sigma_{diff}^2 = \frac{10^2}{100} + \frac{5^2}{70} = 1.36$$

$$= (-9.28, -4.72)$$

95% confidence interval

Takeaways:

1. We are 95% confident that the true mean difference between engineering and management grades falls into this interval
2. The whole interval is negative => engineers were consistently getting lower grades
3. Had we calculated difference as: 'management - engineering', we would get a confidence interval: **(4.72,9.28)**

Confidence interval for the difference of two means. Independent samples. Variance known but assumed to be equal.

Pooled variance formula

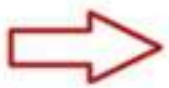
$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$$

$$(\bar{x} - \bar{y}) \pm t_{n_x+n_y-2,\alpha/2}\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}$$

The degrees of freedom are equal to the total sample size minus the number of variables.

$$n_x + n_y - 2$$

1. Population variance unknown
2. Small samples



T

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$$

Problem: Estimate the difference of price of apples in NY and LA

Confidence interval for difference of two means; independent samples, variances unknown but assumed to be equal

Apples example

NY apples	LA apples
$ 3.80	$ 3.02
$ 3.76	$ 3.22
$ 3.87	$ 3.24
$ 3.99	$ 3.02
$ 4.02	$ 3.06
$ 4.25	$ 3.15
$ 4.13	$ 3.81
$ 3.98	$ 3.44
$ 3.99	
$ 3.62	

	NY	LA
Mean	$3.94	$3.25
Std. deviation	$0.18	$0.27
Sample size	10	8
Pooled variance	0.05	
Pooled std	0.22	

You don't know what the population variance of apple prices in NY or LA is, but you assume it should be the same

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2} = \frac{(10 - 1)0.18^2 + (8 - 1)0.27^2}{10 + 8 - 2} = 0.05$$

$n_x + n_y - 2 = 10 + 8 - 2 = 16$ = Degrees of Freedom

$$(\bar{x} - \bar{y}) \pm t_{n_x+n_y-2,\alpha/2} \sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}$$

$$= (3.94 - 3.25) \pm 2.12 \sqrt{\frac{0.05}{10} + \frac{0.05}{8}}$$

$CI_{95\%} = (0.47, 0.92)$

Takeaway:

Apples in NY are much more expensive than in LA

We are 95% confident that the actual difference between the two populations, price the apple in NY & LA, is somewhere between 0.47 and 0.92

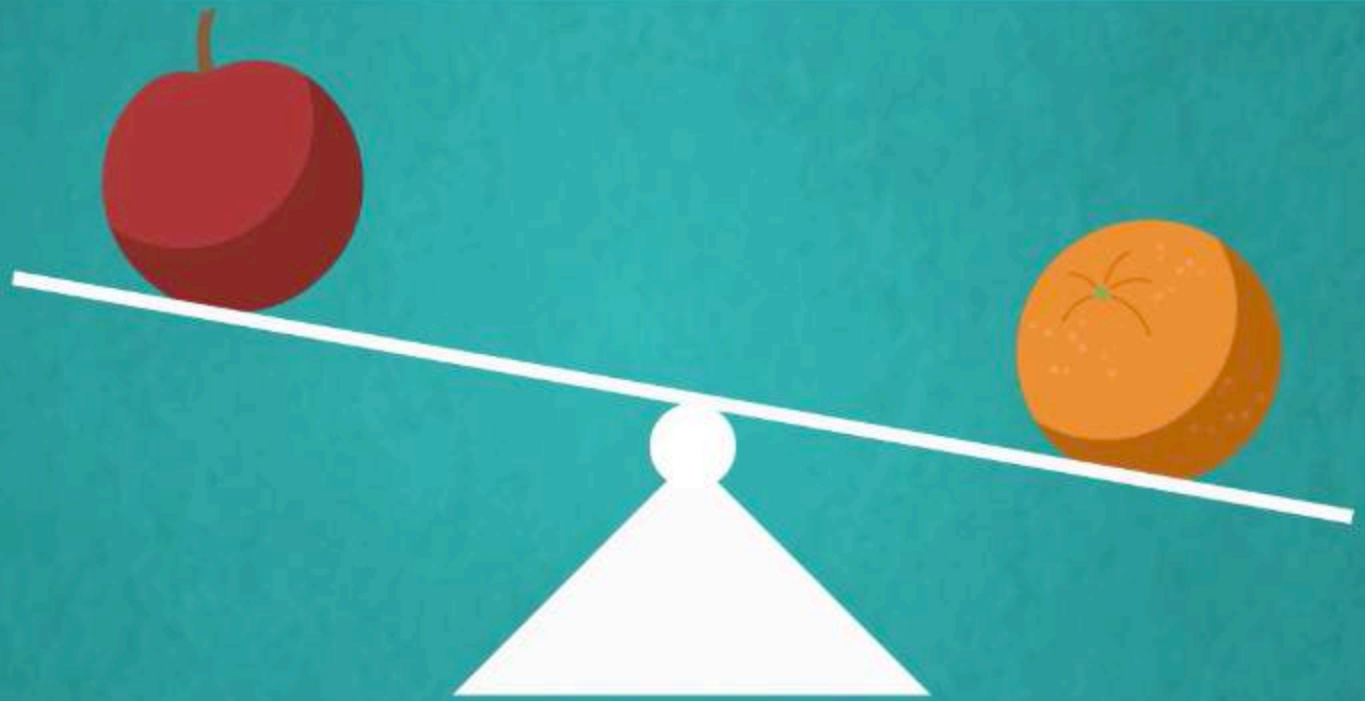
WE WANT TO FIND THE CONFIDENCE INTERVAL FOR TWO SAMPLE MEANS

Independent

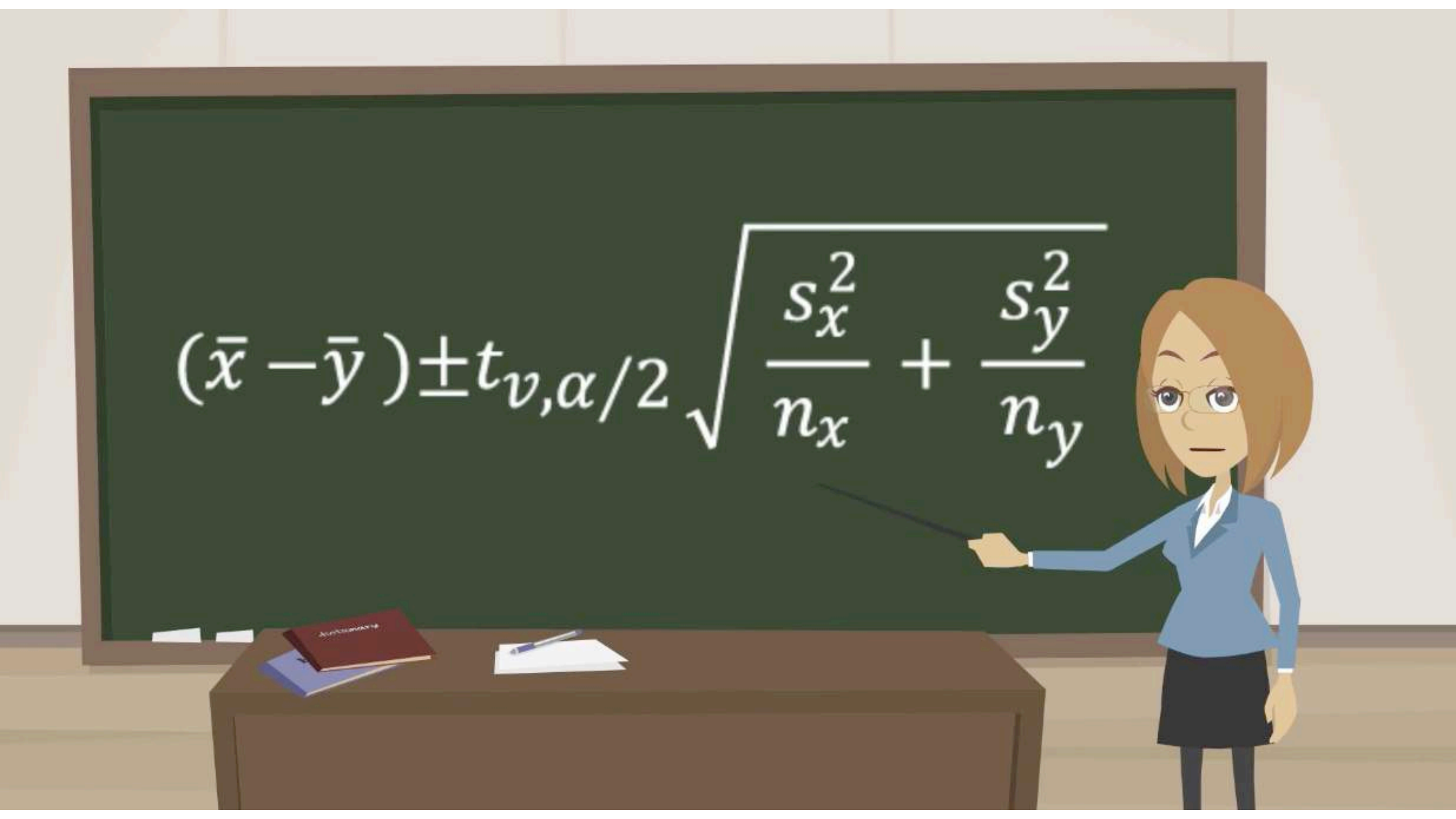
Variance unknown

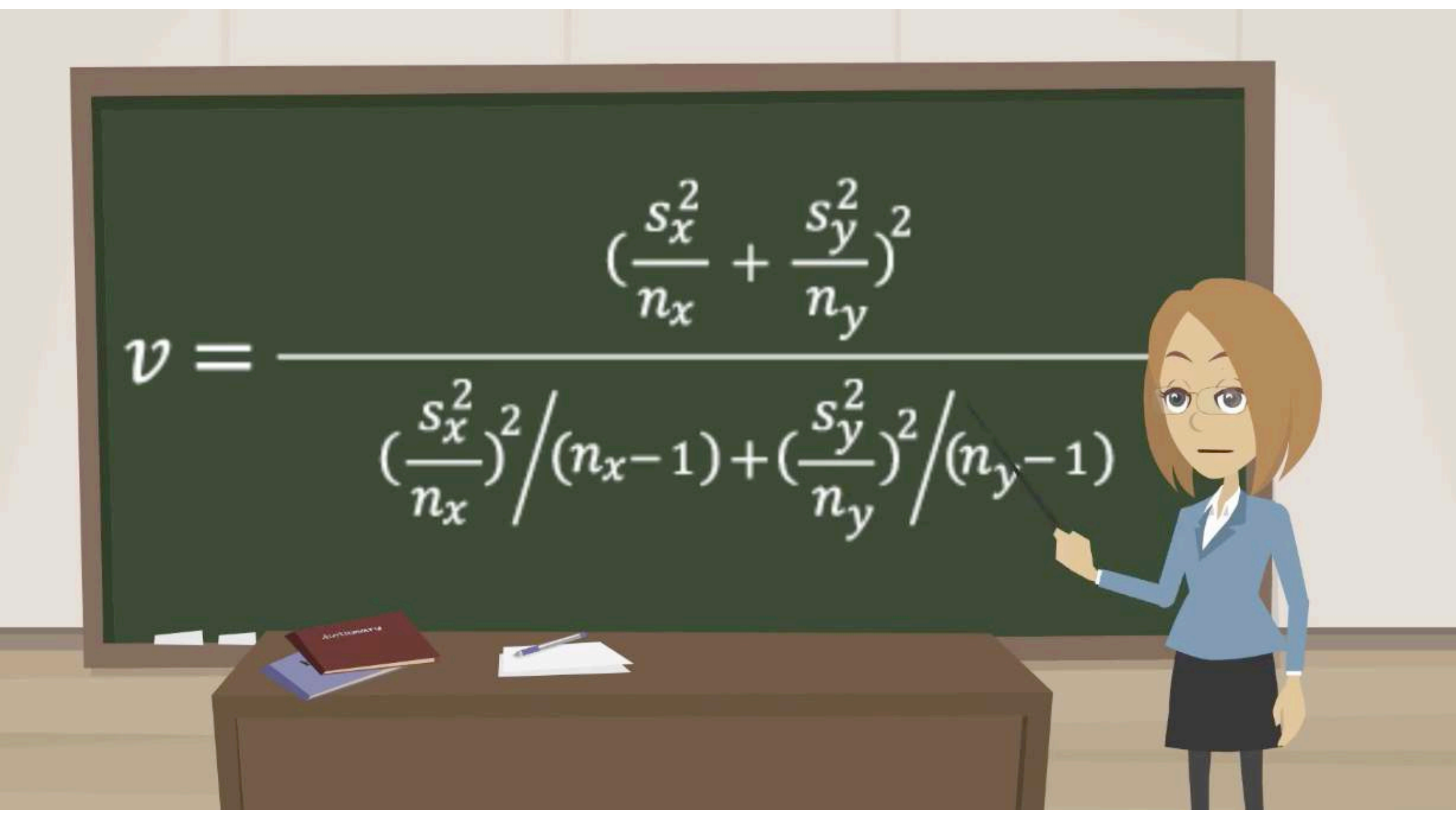
Assumed to be different

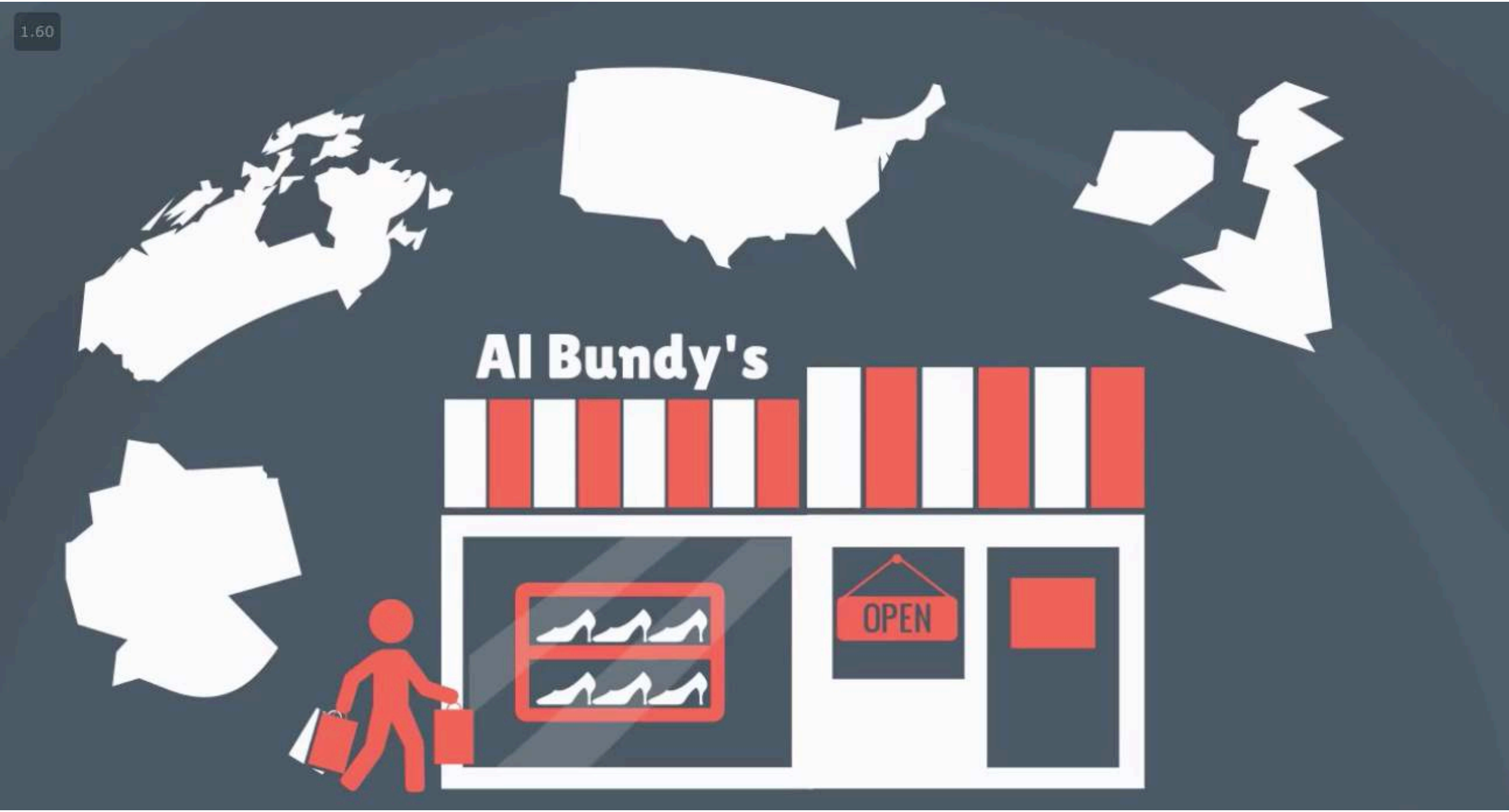
IF YOU REALLY WANT TO COMPARE AN APPLE OR AN ORANGE

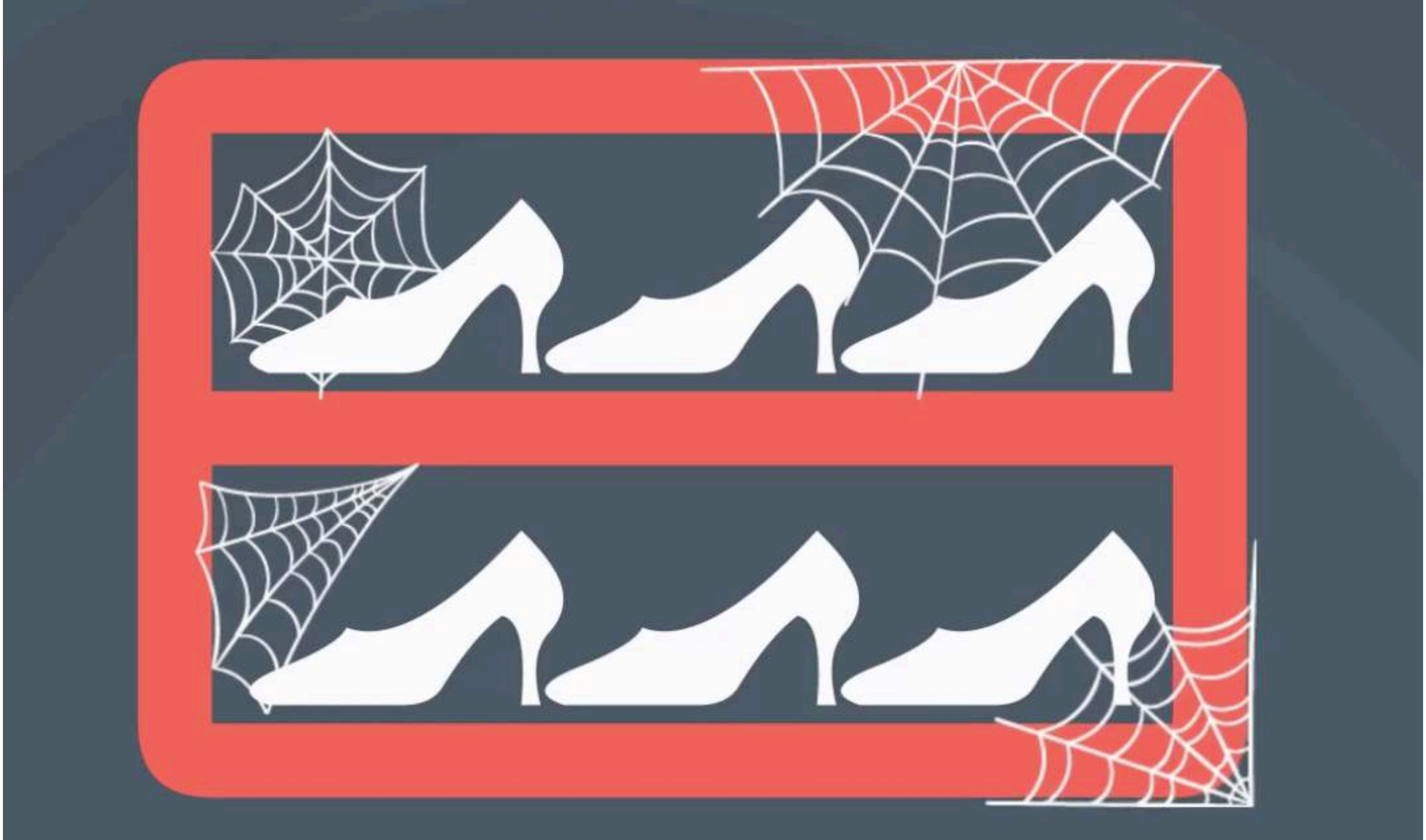


THIS IS THE RIGHT WAY TO DO IT


$(\bar{x} - \bar{y}) \pm t_{v,\alpha/2} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$


$$v = \frac{(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y})^2}{(\frac{s_x^2}{n_x})^2 / (n_x - 1) + (\frac{s_y^2}{n_y})^2 / (n_y - 1)}$$


1.60
Al Bundy's
OPEN



Inventory Management Problem

Inferential statistics. Confidence intervals

Al Bundy's shoe shop

InvoiceNo	Date	Country	ProductID	Shop	Gender	Size (US)	Size (Europe)	Size (UK)	UnitPrice	Discount	SalePrice
52389	1/1/2014	United Kingdom	2152	UK2	Male	11	44	10.5	$ 159.00	0%	$ 159.00
52390	1/1/2014	United States	2230	US15	Male	11.5	44-45	11	$ 199.00	20%	$ 159.20
52391	1/1/2014	Canada	2160	CAN7	Male	9.5	42-43	9	$ 149.00	20%	$ 119.20
52392	1/1/2014	United States	2234	US6	Female	9.5	40	7.5	$ 159.00	0%	$ 159.00
52393	1/1/2014	United Kingdom	2222	UK4	Female	9	39-40	7	$ 159.00	0%	$ 159.00
52394	1/1/2014	United States	2173	US15	Male	10.5	43-44	10	$ 159.00	0%	$ 159.00
52395	1/2/2014	Germany	2200	GER2	Female	9	39-40	7	$ 179.00	0%	$ 179.00
52396	1/2/2014	Canada	2238	CAN5	Male	10	43	9.5	$ 169.00	0%	$ 169.00
52397	1/2/2014	United States	2191	US13	Male	10.5	43-44	10	$ 139.00	0%	$ 139.00
52398	1/2/2014	United Kingdom	2237	UK1	Female	9	39-40	7	$ 149.00	0%	$ 149.00
52399	1/2/2014	United States	2197	US1	Male	10	43	9.5	$ 129.00	0%	$ 129.00
52399	1/2/2014	United States	2213	US11	Female	9.5	40	7.5	$ 169.00	10%	$ 152.10
52399	1/2/2014	United States	2206	US2	Female	9.5	40	7.5	$ 139.00	0%	$ 139.00
52400	1/2/2014	United States	2152	US15	Male	8	41	7.5	$ 139.00	0%	$ 139.00
52401	1/3/2014	Germany	2235	GER1	Male	10.5	43-44	10	$ 169.00	50%	$ 84.50
52401	1/3/2014	Germany	2197	GER1	Female	8.5	39	6.5	$ 179.00	20%	$ 143.20
52402	1/3/2014	Canada	2240	CAN6	Male	9.5	42-43	9	$ 199.00	30%	$ 139.30
52403	1/3/2014	United States	2221	US7	Male	11	44	10.5	$ 149.00	50%	$ 74.50
52404	1/3/2014	United States	2234	US6	Female	9.5	40	7.5	$ 159.00	0%	$ 159.00
52404	1/3/2014	United States	2197	US1	Male	10	43	9.5	$ 129.00	0%	$ 129.00
52404	1/3/2014	United States	2213	US11	Female	9.5	40	7.5	$ 169.00	10%	$ 152.10
52405	1/3/2014	United States	2213	US7	Female	8	38-39	6	$ 139.00	0%	$ 139.00
52406	1/3/2014	United States	2147	US15	Male	9.5	42-43	9	$ 139.00	0%	$ 139.00
52407	1/4/2014	United States	2224	US13	Male	9	42	8.5	$ 149.00	10%	$ 134.10
52408	1/4/2014	Germany	2206	GER2	Male	8.5	41-42	8	$ 149.00	20%	$ 119.20
52409	1/4/2014	Germany	2157	GER2	Male	12	45	11.5	$ 149.00	20%	$ 119.20
						8	38-39	6	$ 129.00	0%	$ 129.00
						10.5	43-44	10	$ 139.00	0%	$ 139.00

distribution tables | United States Men | G ...

Segment the data by:
1. Shoe size
2. Country
3. Gender

Sample or population data

- Sample: Data from 2014 to 2016

Men and women shoes are different. Bundling them together will yield deceiving results

Frequency distribution tables

Problem: What is the number of shoes that are likely to be sold, based on historical data?

Men sizes

	Country				
US	Canada	United States	United Kingdom	Germany	Total
6	15	54	6	30	**105**
6.5	15	45	12	18	**90**
7	24	39	21	30	**114**
7.5	45	66	12	48	**171**
8	51	141	45	117	**354**
8.5	192	225	87	174	**678**
9	324	492	183	348	**1347**
9.5	375	741	225	549	**1890**
10	237	543	156	411	**1347**
10.5	243	462	150	453	**1308**
11	114	213	69	156	**552**
11.5	75	156	39	129	**399**
12	51	87	24	78	**240**
13	12	39	3	33	**87**
14	21	60	15	30	**126**
15	27	24	12	48	**111**
16	0	0	0	0	**0**

Women sizes

	Country				
US	Canada	United States	United Kingdom	Germany	Total
4	0	0	0	0	**0**
4.5	6	21	15	9	**51**
5	6	9	9	12	**36**
5.5	6	42	6	9	**63**
6	21	33	12	15	**81**
6.5	51	93	24	84	**252**
7	93	147	27	156	**423**
7.5	153	318	87	222	**780**
8	192	618	168	324	**1302**
8.5	171	399	129	339	**1038**
9	213	384	93	264	**954**
9.5	84	189	57	126	**456**
10	48	75	21	87	**231**
10.5	36	87	18	57	**198**
11.5	12	30	3	15	**60**
Total	**1092**	**2445**	**669**	**1719**	**5925**

Game plan: Find the 95% confidence interval using
1. Last 12 months of sales
2. Only for men shoes (as the problem is identical)
3. Only for the USA (as the problem is identical)

Inventory Management Problem

First, we need to calculate the means

Second, population variance is unknown

One population, population variance unknown -> t-statistic

$t_{11,0.025} = 2.20$

Third, compute the standard errors

$$\frac{s}{\sqrt{n}}$$

Finally, calculate the confidence intervals

$$\bar{x} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$$

Frequency distribution tables

By size and month

Problem: What is the number of shoes that are likely to be sold, based on historical data?

Men shoes sales

	United States, 2016												Mean	Standard error	ME	95% CI	
US	1	2	3	4	5	6	7	8	9	10	11	12	2016	2016	2016	2016	
6	4	1	3	1	3	3	3	4	3	7	3	0	2.92	0.51	1.12	1.80	4.04
6.5	3	2	0	1	0	0	1	7	2	1	2	1	1.67	0.56	1.21	0.46	2.88
7	0	0	1	0	6	4	4	2	3	0	0	0	1.67	0.61	1.32	0.34	2.99
7.5	3	2	3	1	7	0	7	3	4	6	1	1	3.17	0.69	1.51	1.65	4.68
8	7	9	7	3	12	2	9	4	7	5	2	6	6.08	0.88	1.92	4.16	8.01
8.5	12	12	8	8	15	9	17	17	6	9	10	6	10.75	1.12	2.45	8.30	13.20
9	17	13	13	11	21	22	25	30	26	25	13	10	18.83	1.97	4.29	14.54	23.12
9.5	19	25	27	24	26	33	25	47	31	44	37	26	30.33	2.45	5.33	25.00	35.67
10	17	26	26	19	16	31	25	24	23	31	15	20	22.75	1.57	3.42	19.33	26.17
10.5	13	16	22	14	28	19	18	15	19	21	16	10	17.58	1.37	2.98	14.60	20.56
11	5	16	13	10	10	11	15	8	9	7	6	7	9.75	1.01	2.20	7.55	11.95
11.5	4	3	6	3	3	5	6	4	5	12	13	5	5.75	0.96	2.10	3.65	7.85
12	3	0	0	4	4	4	3	12	4	9	2	1	3.83	1.01	2.21	1.62	6.04
13	1	1	2	0	3	2	1	0	0	4	3	2	1.58	0.38	0.82	0.76	2.41
14	2	6	3	3	5	3	2	1	0	1	2	1	2.42	0.50	1.09	1.33	3.50
15	0	0	0	1	1	0	4	0	0	0	0	2	0.67	0.36	0.77	-0.11	1.44
16	0	0	0	0	0	0	0	0	0	0	0	0	0.00	0.00	0.00	0.00	0.00
Total	110	132	134	103	160	148	165	178	142	182	125	98					

	2016
n	12
t_{95}	2.18

In 95% of the cases, the true population mean of the sales for each shoe size will fall into the respective interval

Frequency distribution tables

By size and month

Problem: What is the number of shoes that are likely to be sold, based on historical data?

Men shoes sales

	United States, 2016												Mean	Standard error	ME	95% CI		Number of pairs
US	1	2	3	4	5	6	7	8	9	10	11	12	2016	2016	2016	2016		
6	4	1	3	1	3	3	3	4	3	7	3	0	2.92	0.51	1.12	1.80	4.04	4
6.5	3	2	0	1	0	0	1	7	2	1	2	1	1.67	0.56	1.21	0.46	2.88	3
7	0	0	1	0	6	4	4	2	3	0	0	0	1.67	0.61	1.32	0.34	2.99	3
7.5	3	2	3	1	7	0	7	3	4	6	1	1	3.17	0.69	1.51	1.65	4.68	5
8	7	9	7	3	12	2	9	4	7	5	2	6	6.08	0.88	1.92	4.16	8.01	8
8.5	12	12	8	8	15	9	17	17	6	9	10	6	10.75	1.12	2.45	8.30	13.20	13
9	17	13	13	11	21	22	25	30	26	25	13	10	18.83	1.97	4.29	14.54	23.12	23
9.5	19	25	27	24	26	33	25	47	31	44	37	26	30.33	2.45	5.33	25.00	35.67	36
10	17	26	26	19	16	31	25	24	23	31	15	20	22.75	1.57	3.42	19.33	26.17	26
10.5	13	16	22	14	28	19	18	15	19	21	16	10	17.58	1.37	2.98	14.60	20.56	21
11	5	16	13	10	10	11	15	8	9	7	6	7	9.75	1.01	2.20	7.55	11.95	12
11.5	4	3	6	3	3	5	6	4	5	12	13	5	5.75	0.96	2.10	3.65	7.85	8
12	3	0	0	4	4	4	3	12	4	9	2	1	3.83	1.01	2.21	1.62	6.04	6
13	1	1	2	0	3	2	1	0	0	4	3	2	1.58	0.38	0.82	0.76	2.41	2
14	2	6	3	3	5	3	2	1	0	1	2	1	2.42	0.50	1.09	1.33	3.50	4
15	0	0	0	1	1	0	4	0	0	0	0	2	0.67	0.36	0.77	-0.11	1.44	1
16	0	0	0	0	0	0	0	0	0	0	0	0	0.00	0.00	0.00	0.00	0.00	0
Total	110	132	134	103	160	148	165	178	142	182	125	98						

	2016
n	12
$t_{\alpha;}$	2.18

The upper bound of the CI shows us the maximum number of pairs needed.
(we have rounded them mathematically, you can instead round them all up if you see fit

Frequency distribution tables

By size and month

Problem: By how much one shop outperforms the other in terms of sales?

Women shoe sales

$$\mathrm{ME} = t_{v,\alpha/2}\frac{s_p}{\sqrt{n}}$$

	Germany, GER1												Germany, GER2												Mean		Sample variance		Pooled variance	Margin of error	95% CI	
US	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	GER1	GER2	GER1	GER2				
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4.5	0	0	0	0	1	3	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0.42	0.08	0.81	0.08	0.45	0.57	-0.23	0.90
5	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0.17	0.17	0.33	0.33	0.33	0.49	-0.49	0.49
5.5	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	2	0	1	0.08	0.33	0.08	0.42	0.25	0.43	-0.68	0.18
6	0	2	0	0	0	0	0	0	0	0	0	0	0	1	3	1	2	0	0	0	0	0	0	0	0.17	0.58	0.33	0.99	0.66	0.69	-1.11	0.27
6.5	3	3	1	2	1	0	2	0	2	1	3	4	2	0	2	1	1	2	0	1	2	1	3	0	1.83	1.25	1.61	0.93	1.27	0.95	-0.37	1.54
7	0	3	3	4	1	0	1	0	2	0	0	1	0	0	0	4	1	3	1	1	1	3	1	4	1.25	1.58	2.02	2.27	2.14	1.24	-1.57	0.91
7.5	1	2	4	1	2	6	4	3	5	8	2	1	2	1	1	3	2	7	9	8	14	8	6	3	3.25	5.33	4.93	16.06	10.50	2.74	-4.83	0.66
8	6	10	3	9	1	3	6	8	3	12	3	9	13	6	5	13	5	3	11	6	6	9	8	3	6.08	7.33	12.27	12.24	12.25	2.96	-4.21	1.71
8.5	10	10	10	7	14	4	7	7	4	8	7	9	8	5	10	4	5	5	9	7	3	7	9	8	8.08	6.67	7.72	4.97	6.34	2.13	-0.72	3.55
9	1	3	8	6	3	1	4	4	0	2	4	2	5	2	2	9	3	1	1	7	2	1	4	2	3.17	3.25	5.06	6.57	5.81	2.04	-2.13	1.96
9.5	4	1	2	1	2	2	2	4	5	2	3	2	0	1	1	0	1	2	2	1	7	2	4	2	2.50	1.92	1.55	3.72	2.63	1.37	-0.79	1.96
10	0	1	1	1	1	1	3	1	0	0	0	1	0	1	1	0	0	0	2	3	0	2	0	0	0.83	0.75	0.70	1.11	0.91	0.81	-0.72	0.89
10.5	1	0	0	0	2	2	4	1	0	3	1	1	0	2	0	0	0	1	0	0	0	0	2	1	1.25	0.50	1.66	0.64	1.15	0.91	-0.16	1.66
11.5	0	0	0	2	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	1	0	0	0	0	0.17	0.50	0.33	2.09	1.21	0.93	-1.27	0.60
Total	26	35	32	33	28	22	35	28	21	36	25	30	30	19	30	35	20	24	35	38	35	36	37	24								

Assumption: same people don't buy shoes from different shops in the same year

The two samples are independent

The two samples are independent, population variance unknown, but assumed equal

All confidence intervals start in the negatives and finish in the positives

$t_{12+12-2,0.025}$

	GER1	GER2
n	12	12
$t_{95\%,22}$	2.07	

We cannot conclude one shop sells more shoes than the other for any size

For some sizes, GER1 is likely to sell more, while for others - vice versa

Insight: these two shops are so balanced in terms of sales, they may be bundled together