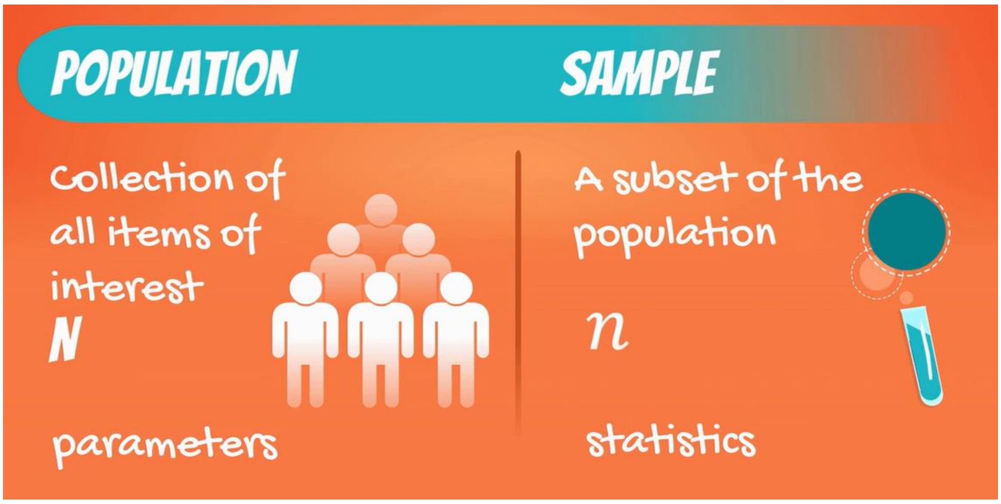
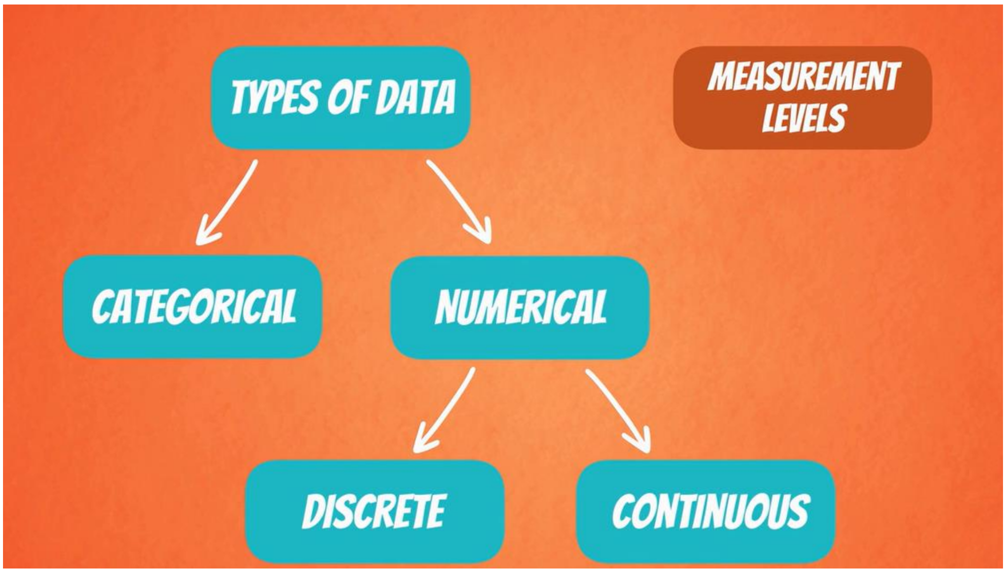
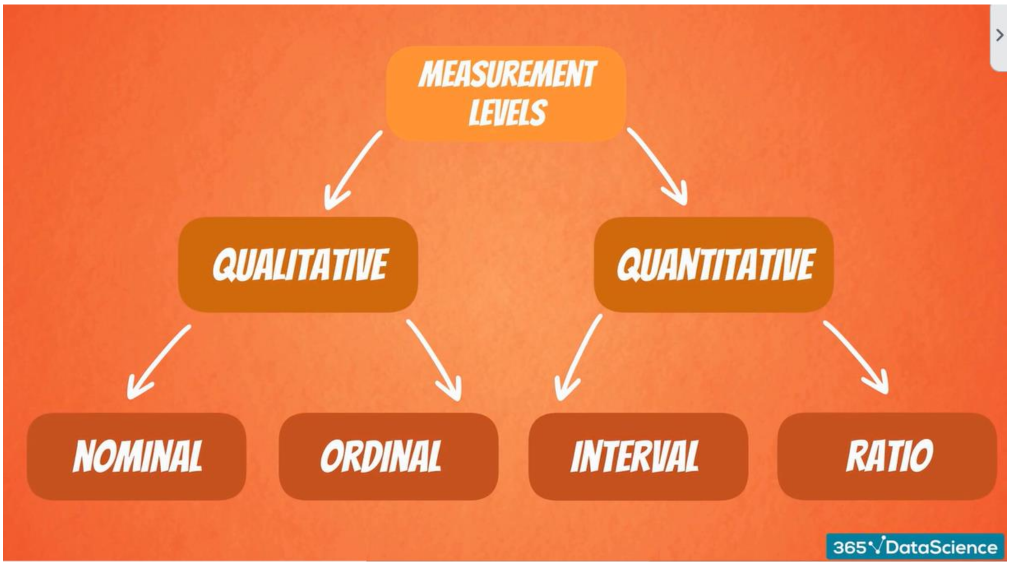
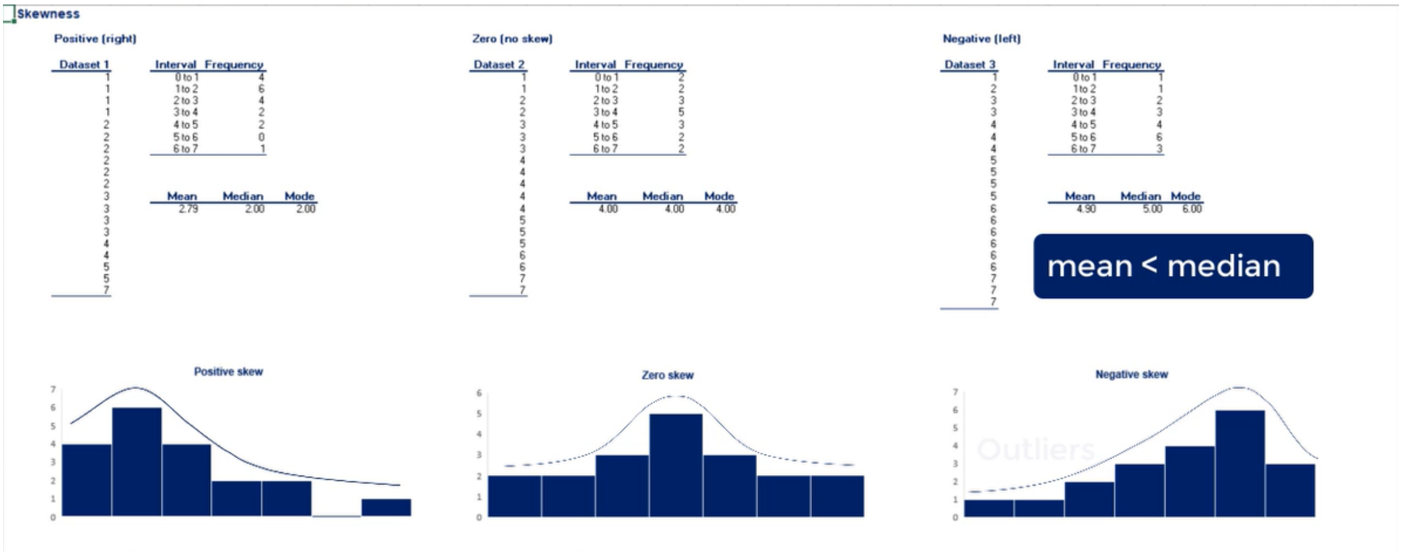

POPULATION
SAMPLE
Collection of all items of interest
N
parameters
A subset of the population
n
statistics


TYPES OF DATA
MEASUREMENT LEVELS
CATEGORICAL
NUMERICAL
DISCRETE
CONTINUOUS


MEASUREMENT LEVELS
QUALITATIVE
QUANTITATIVE
NOMINAL
ORDINAL
INTERVAL
RATIO
365√DataScience



Skewness

Positive (right)

Dataset 1
1
1
1
1
2
2
2
2
2
2
3
3
3
3
4
4
5
5
7

Interval	Frequency
0 to 1	4
1 to 2	6
2 to 3	4
3 to 4	2
4 to 5	2
5 to 6	0
6 to 7	1

Mean	Median	Mode
2.79	2.00	2.00

Zero (no skew)

Dataset 2
1
1
2
2
3
3
3
4
4
4
4
4
5
5
5
6
6
7
7

Interval	Frequency
0 to 1	2
1 to 2	2
2 to 3	3
3 to 4	5
4 to 5	3
5 to 6	2
6 to 7	2

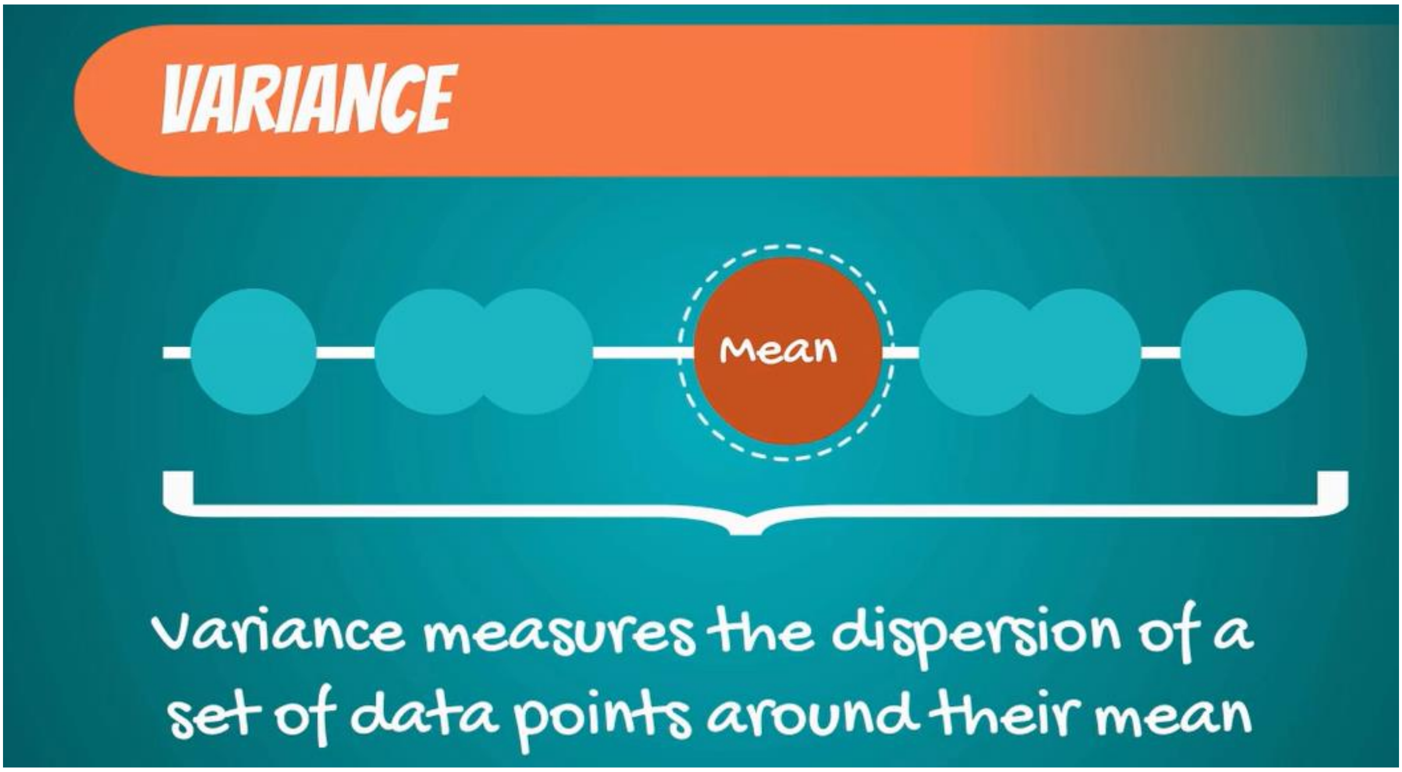
Mean	Median	Mode
4.00	4.00	4.00

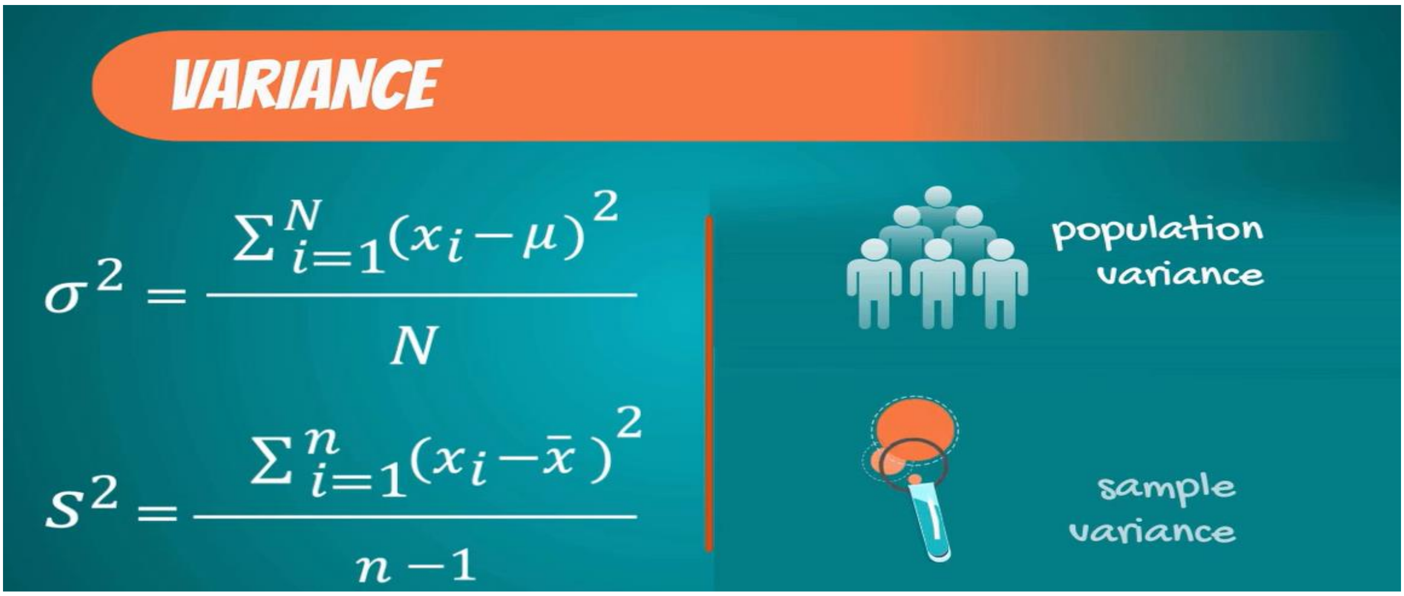
Negative (left)

Dataset 3
1
2
3
3
4
4
4
5
5
5
5
6
6
6
6
6
6
7
7
7

Interval	Frequency
0 to 1	1
1 to 2	1
2 to 3	2
3 to 4	3
4 to 5	4
5 to 6	6
6 to 7	3

Mean	Median	Mode
4.90	5.00	6.00





STANDARD DEVIATION FORMULAS

$$\sigma = \sqrt{\sigma^2}$$

population standard deviation

sample standard deviation

$$s = \sqrt{s^2}$$

COEFFICIENT OF VARIATION (CV)

/relative standard deviation/

$$\frac{\text{standard deviation}}{\text{mean}}$$

σ

Standard deviation is the most common measure of variability for a SINGLE DATASET

Comparing TWO OR MORE datasets

c_v

Standard deviation and coefficient of variation

Pizza price example

NY Dollars			Pesos
$	1.00	MXN	18.81
$	2.00	MXN	37.62
$	3.00	MXN	56.43
$	3.00	MXN	56.43
$	5.00	MXN	94.05
$	6.00	MXN	112.86
$	7.00	MXN	131.67
$	8.00	MXN	150.48
$	9.00	MXN	169.29
$	11.00	MXN	206.91

	Dollars			Pesos
Mean	$	5.50	MXN	103.46
Sample variance	$\$^2$	10.72	MXN^2	3793.69
Sample standard deviation	$	3.27	MXN	61.59

Sample standard deviation

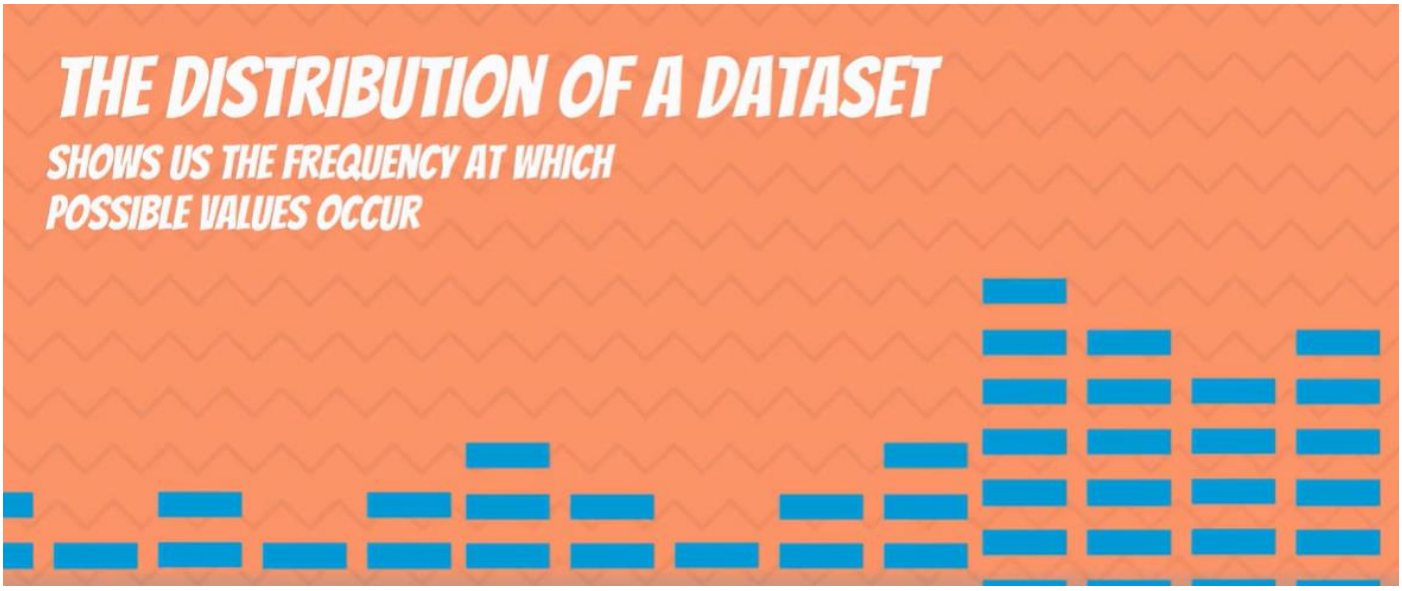
$$\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$$

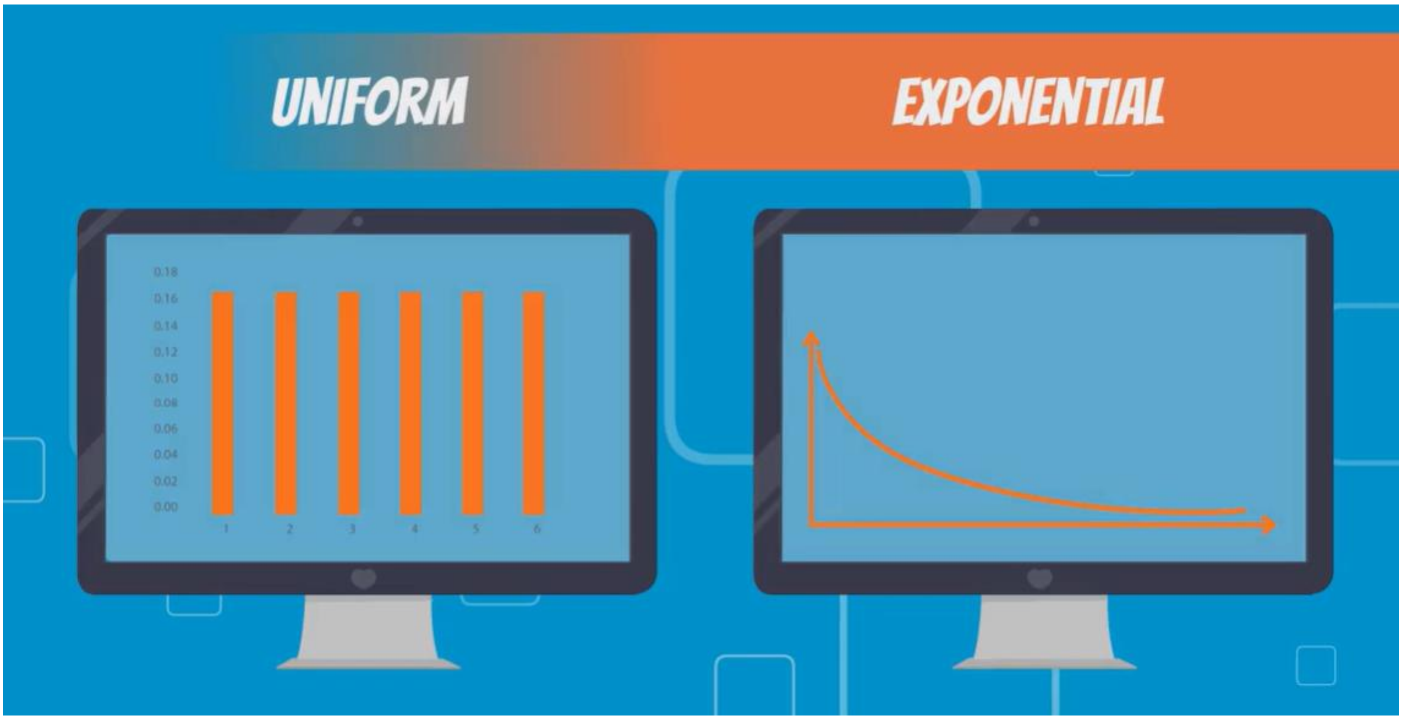
Step 1: Sample or population?

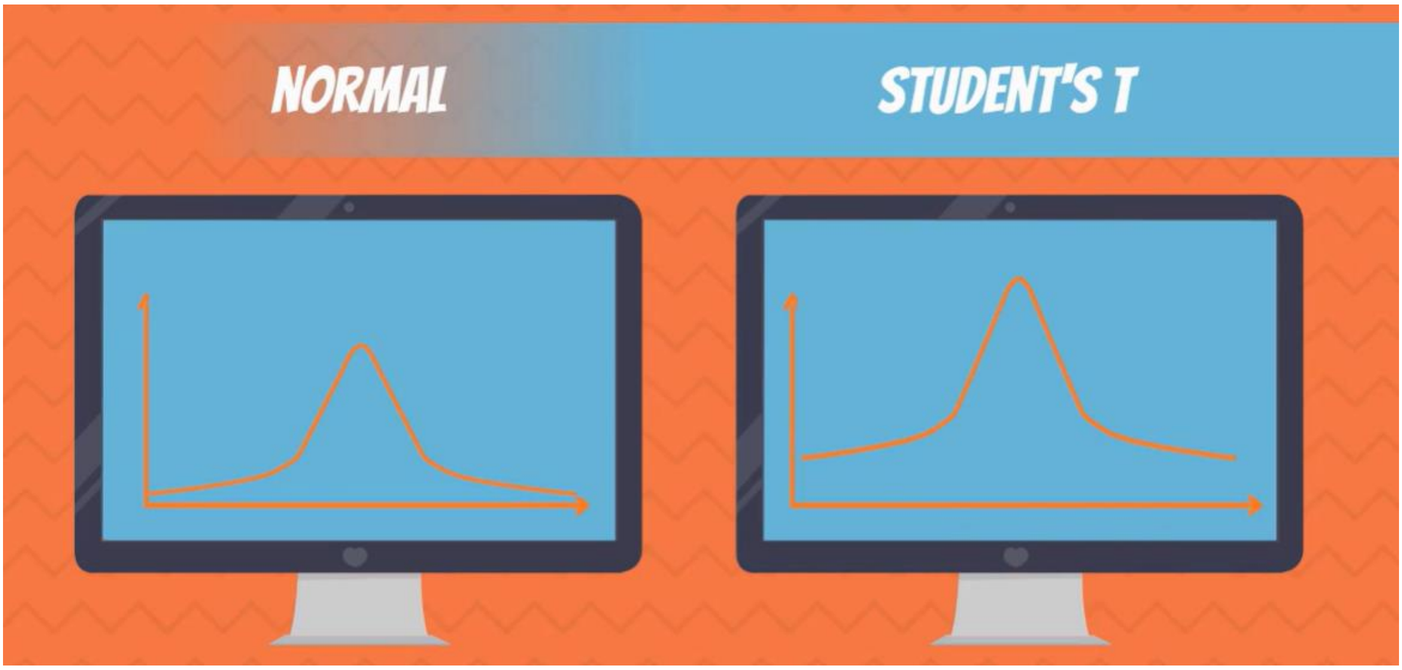
Step 2: Find the mean

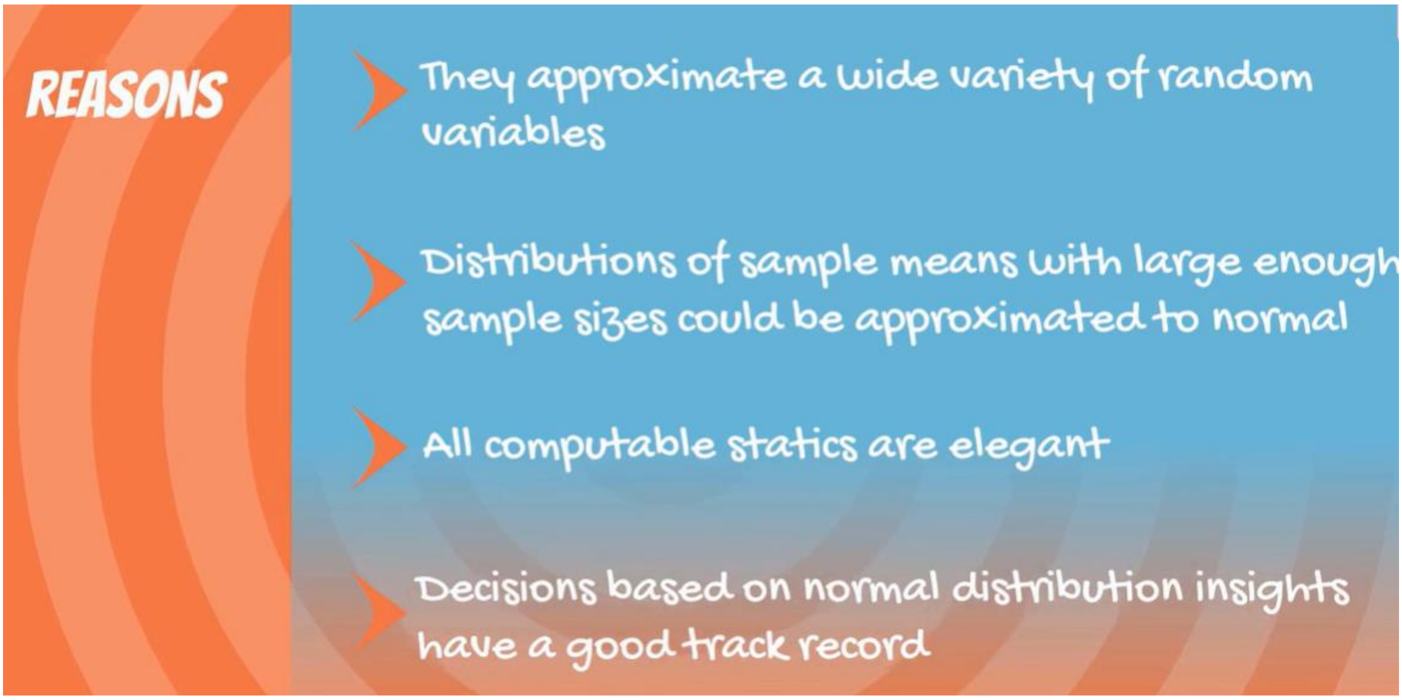
Step 3: Find the sample variance

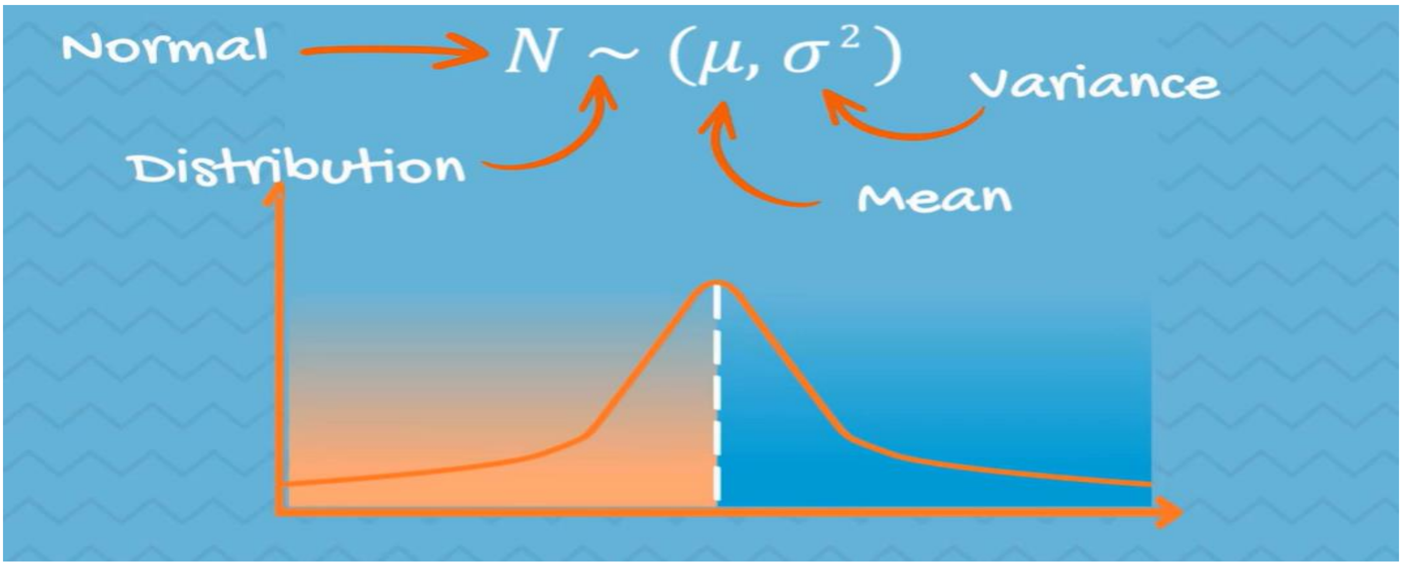
Step 4: Find the sample standard deviation

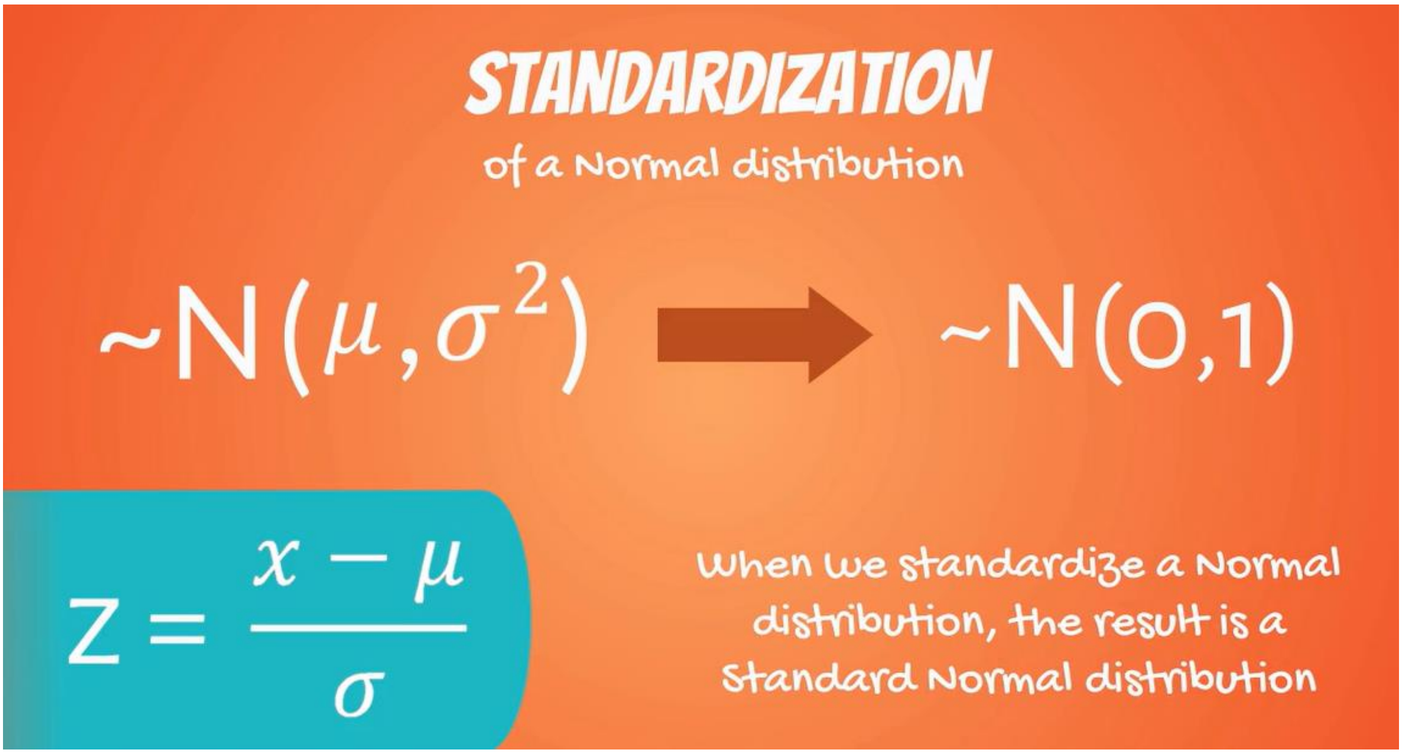


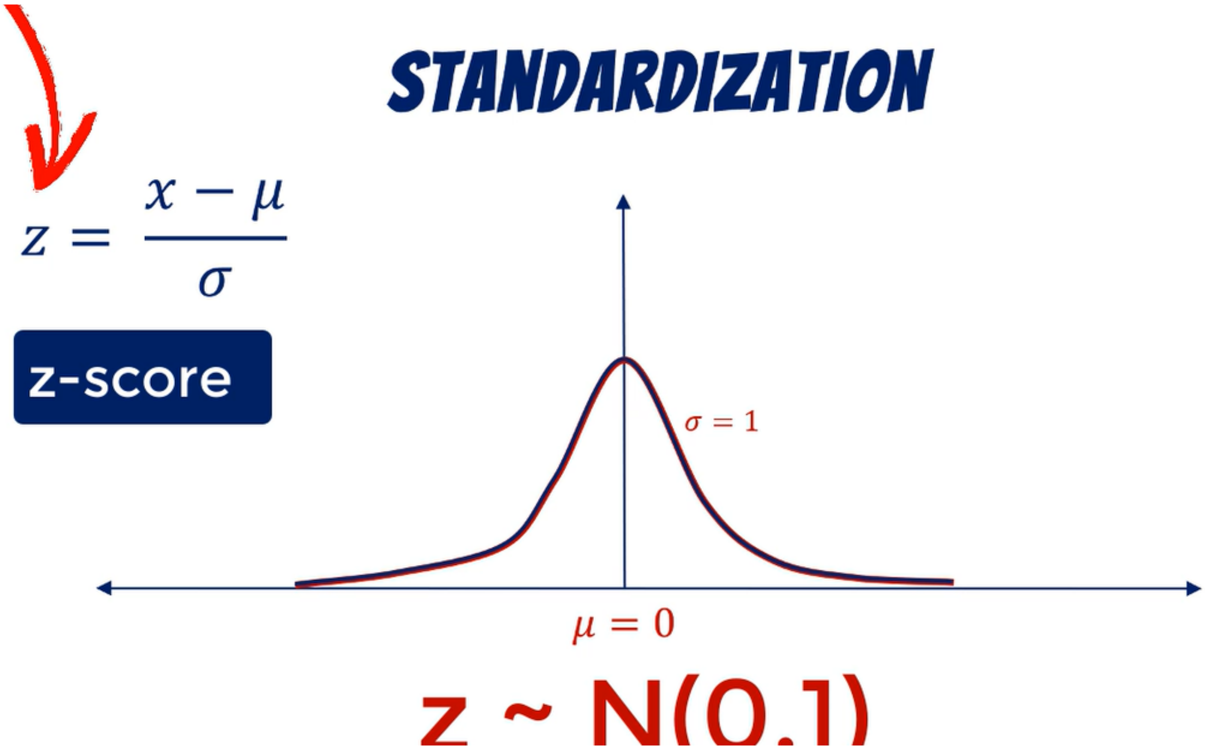



NORMAL
STUDENT'S T


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


Normal
$N \sim (\mu, \sigma^2)$
Variance
Distribution
Mean

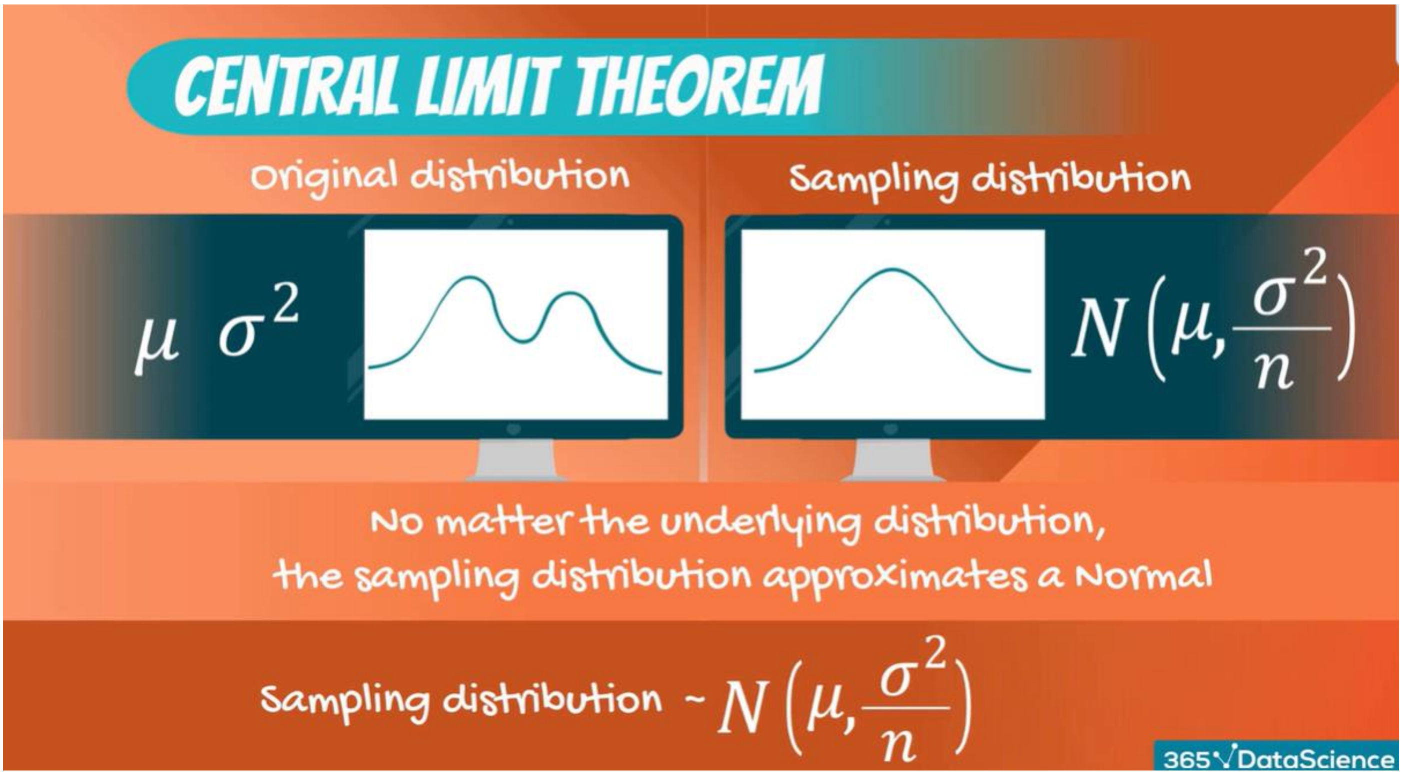

STANDARDIZATION
of a Normal distribution
~N(μ,σ²) ~N(0,1)
Z = (x − μ)/σ
When we standardize a Normal distribution, the result is a Standard Normal distribution

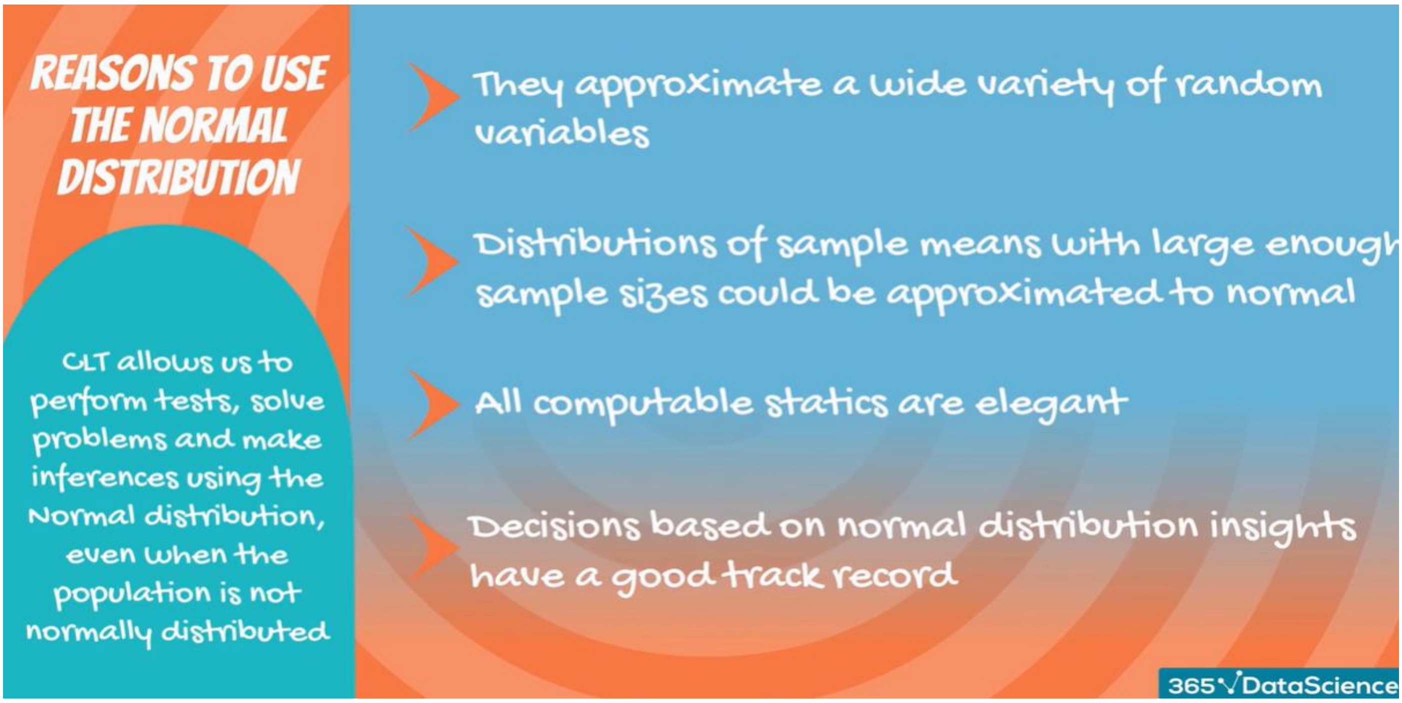

STANDARDIZATION
z = (x − μ)/σ
z-score
σ = 1
μ = 0
z ~ N(0,1)

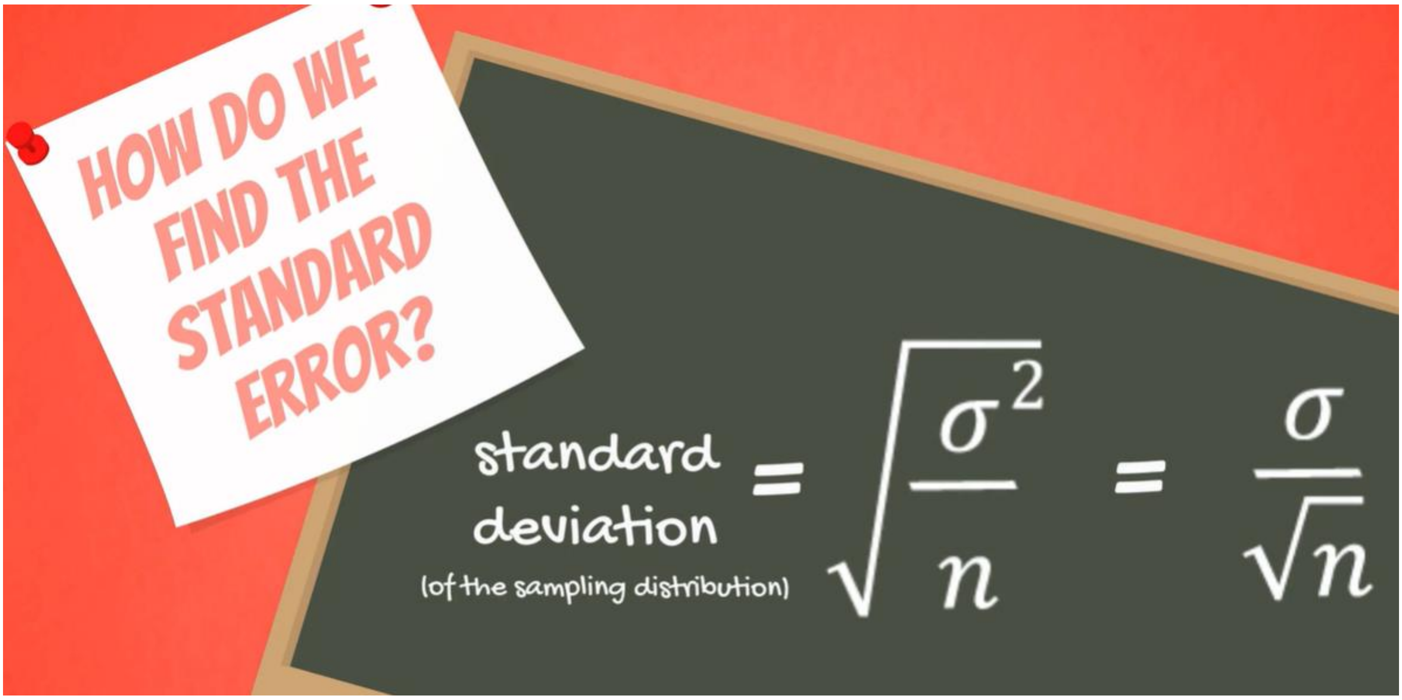
Standard normal distribution
Standardization

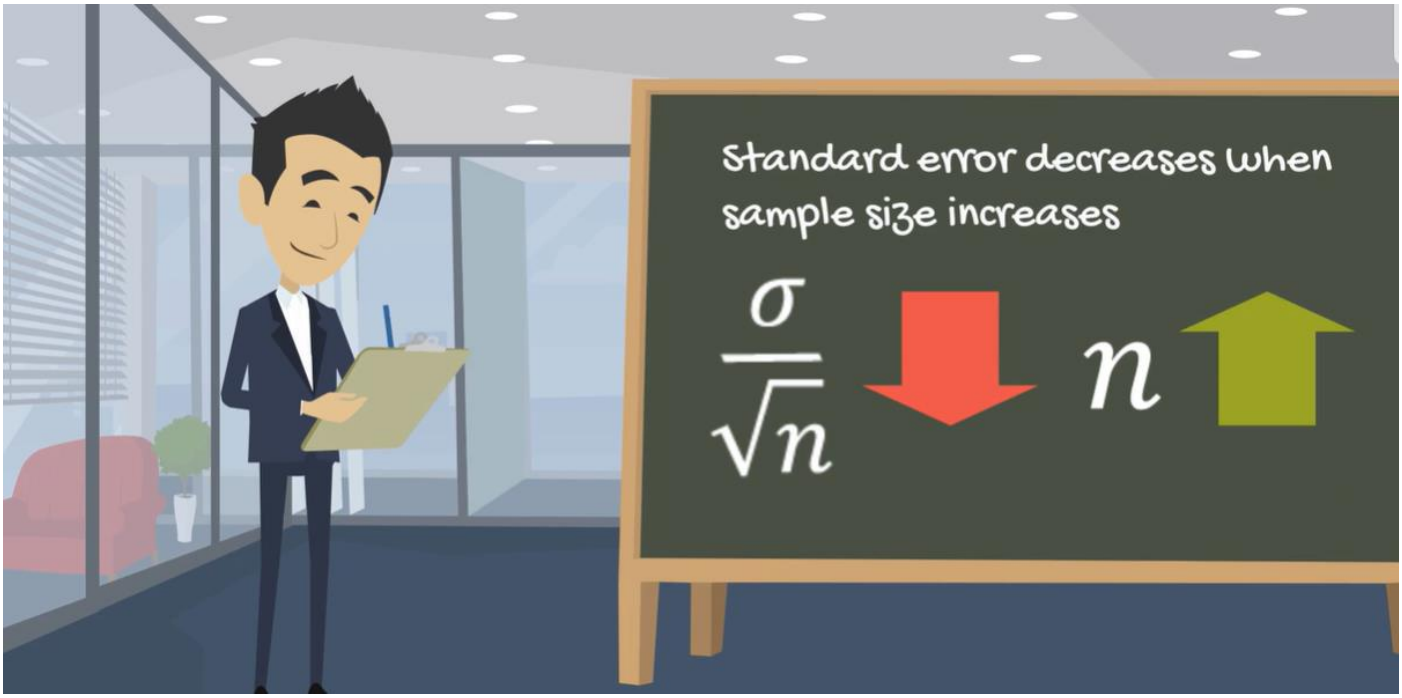
Original dataset			Subtract mean			Divide by std
1			-2			-1.63
2	Mean	3	-1	Mean	0	-0.82
2	St. dev	1.22	-1	St. dev	1.22	-0.82
3			0			0.00
3			0			0.00
3	N~(3,1.49)		0	N~(0,1.49)		0.00
4			1			0.82
4			1			0.82
5			2			1.63

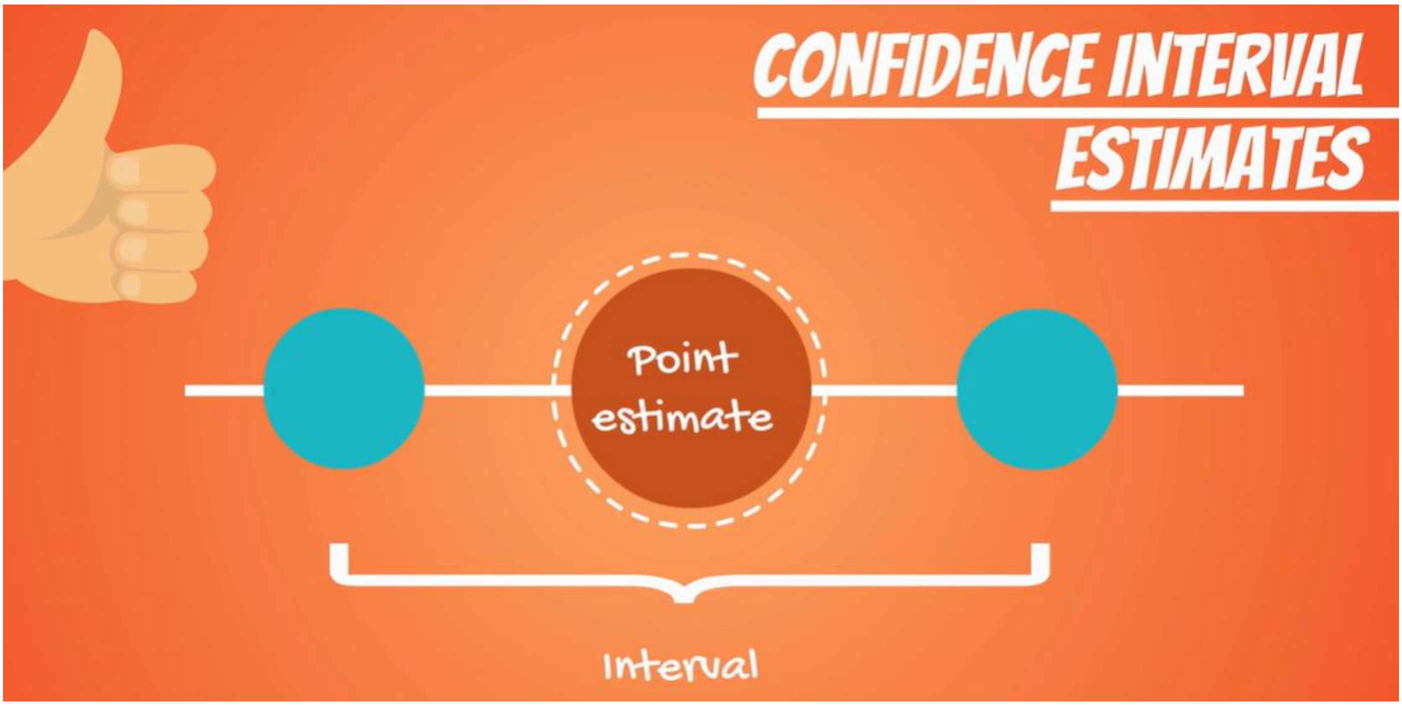
$$x \qquad\qquad x-\mu \qquad\qquad \frac{x-\mu}{\sigma}$$

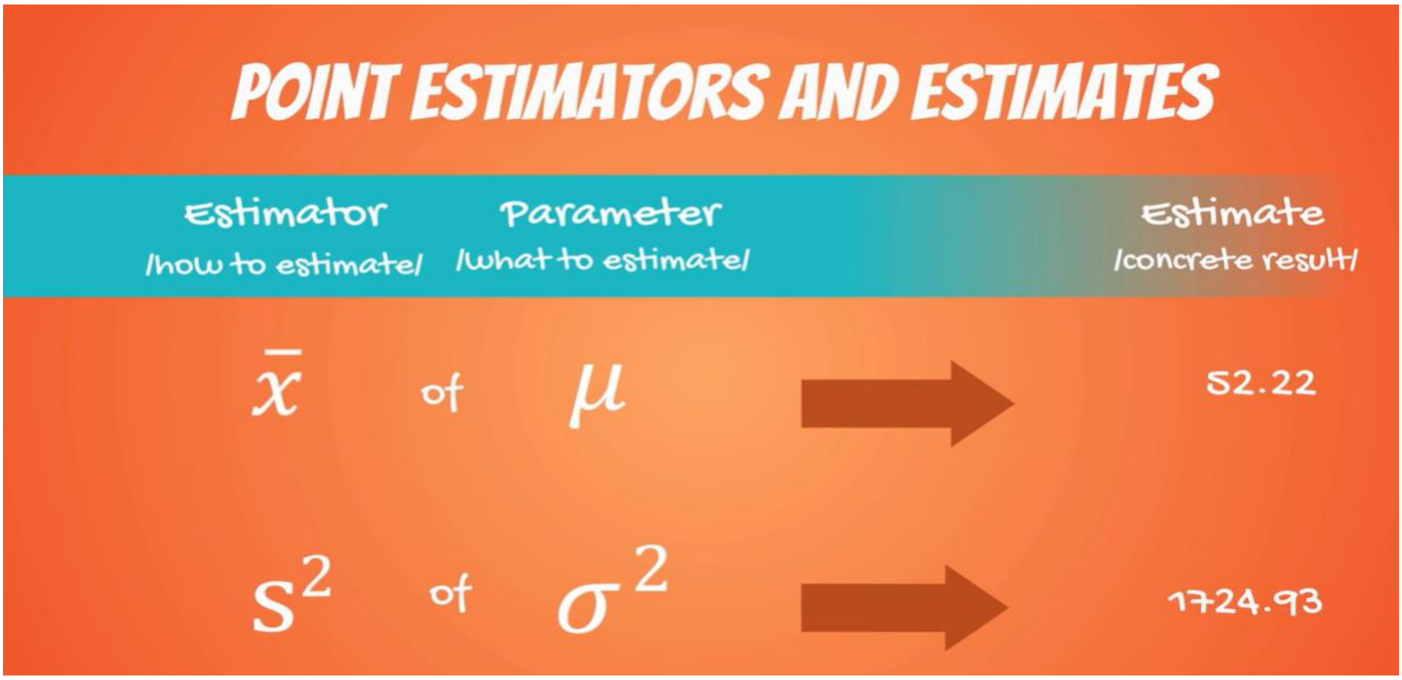




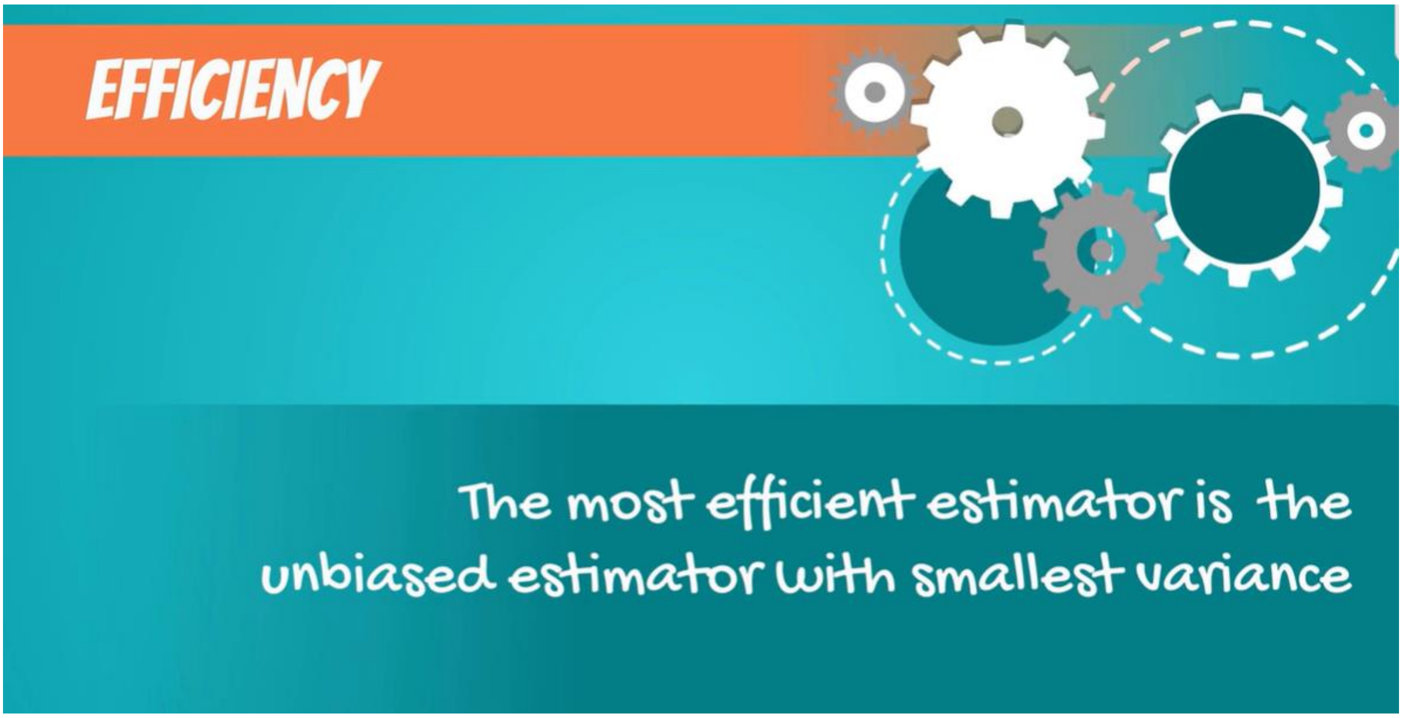

HOW DO WE FIND THE STANDARD ERROR?
standard deviation (of the sampling distribution) $= \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$

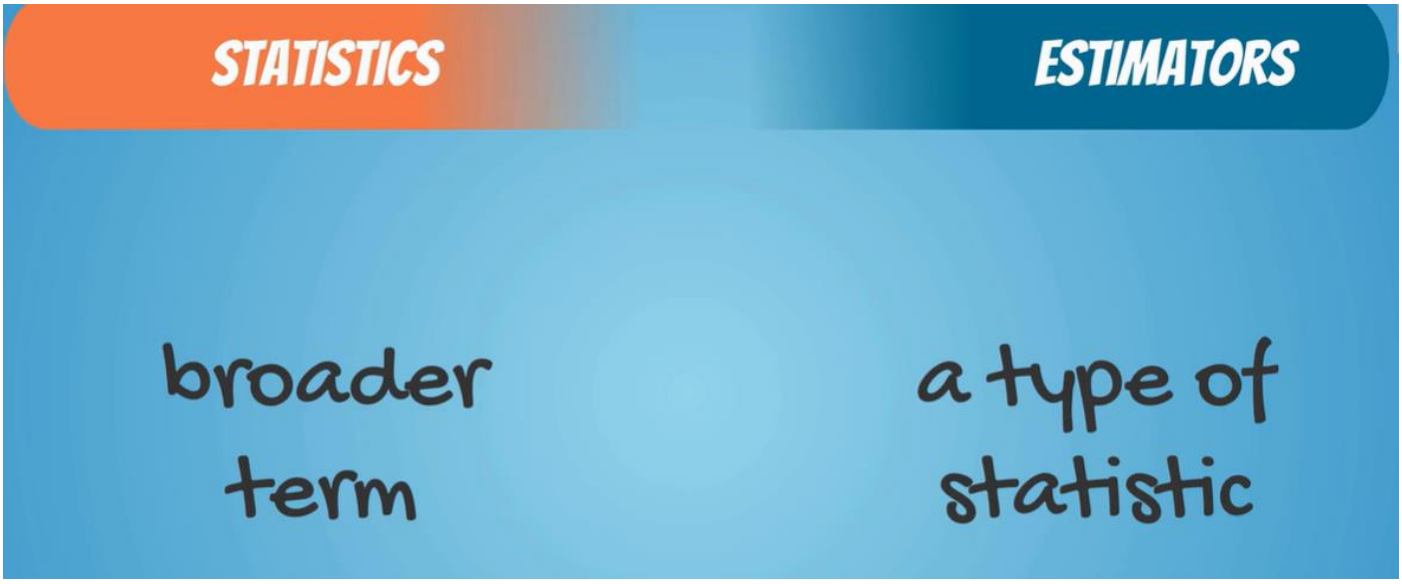

Standard error decreases when sample size increases
$\frac{\sigma}{\sqrt{n}}$
n


CONFIDENCE INTERVAL ESTIMATES
Point estimate
Interval

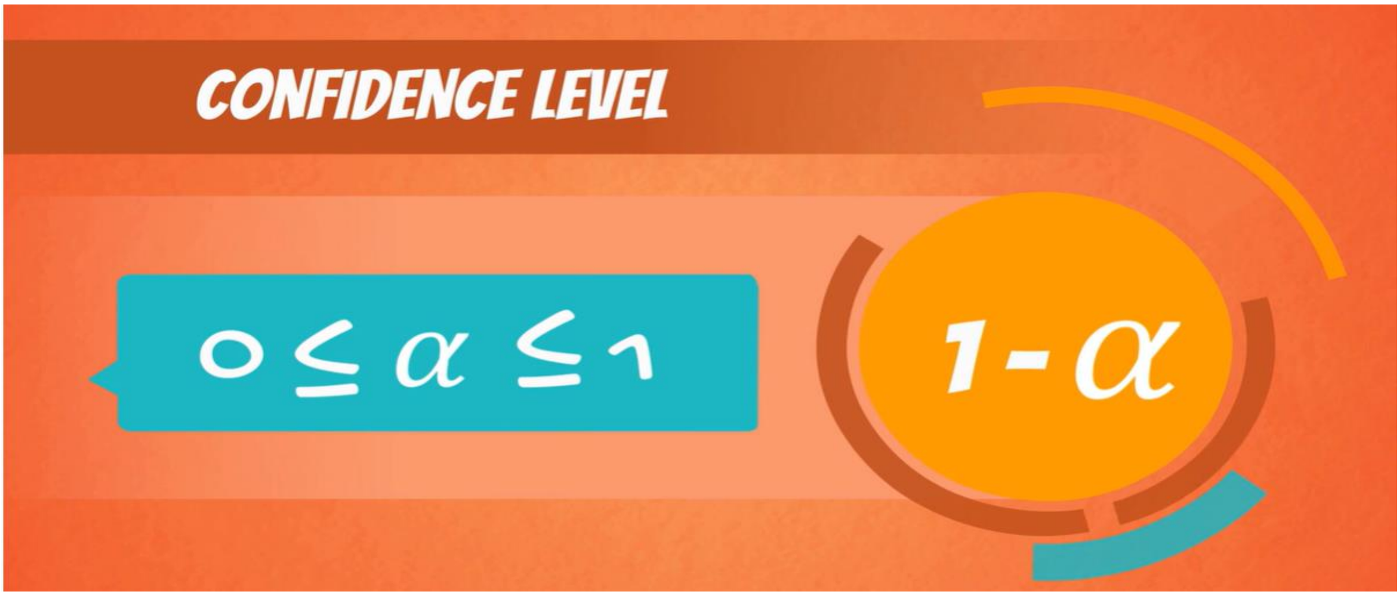

POINT ESTIMATORS AND ESTIMATES
Estimator /how to estimate/
Parameter /what to estimate/
Estimate /concrete result/
$\bar{x}$ of μ
52.22
s^2 of σ^2
1724.93

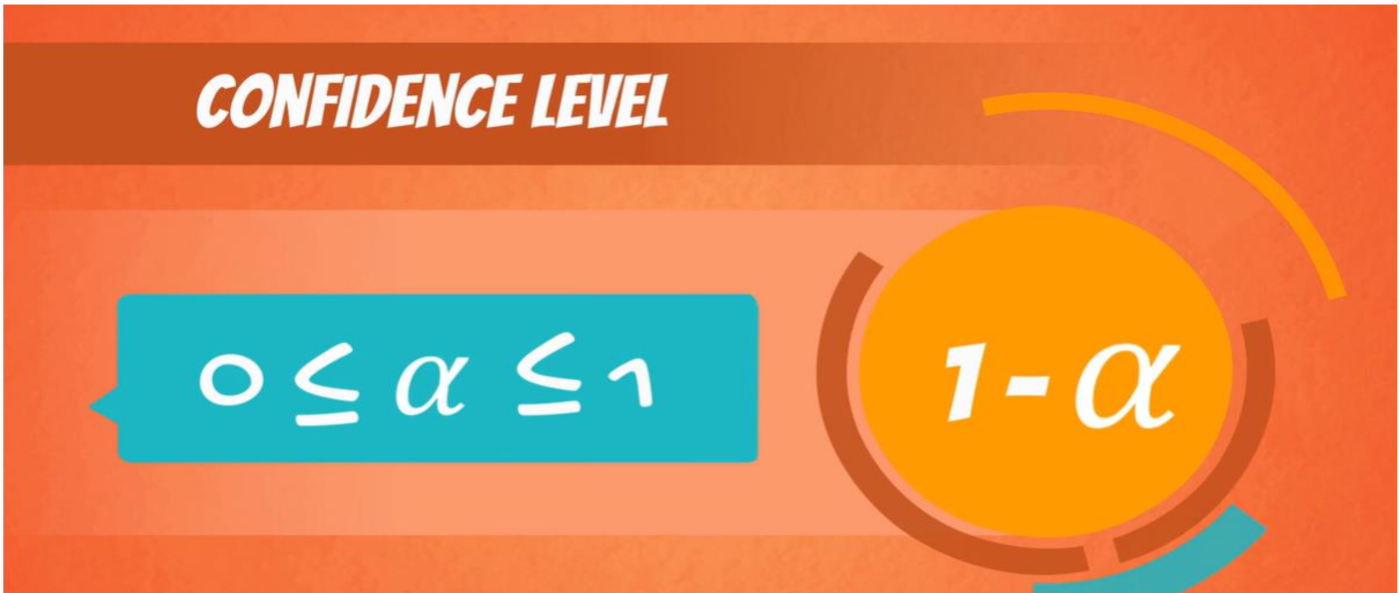

UNBIASED ESTIMATOR
expected value = population parameter
e.g. $\bar{x}$ has an expected value of μ

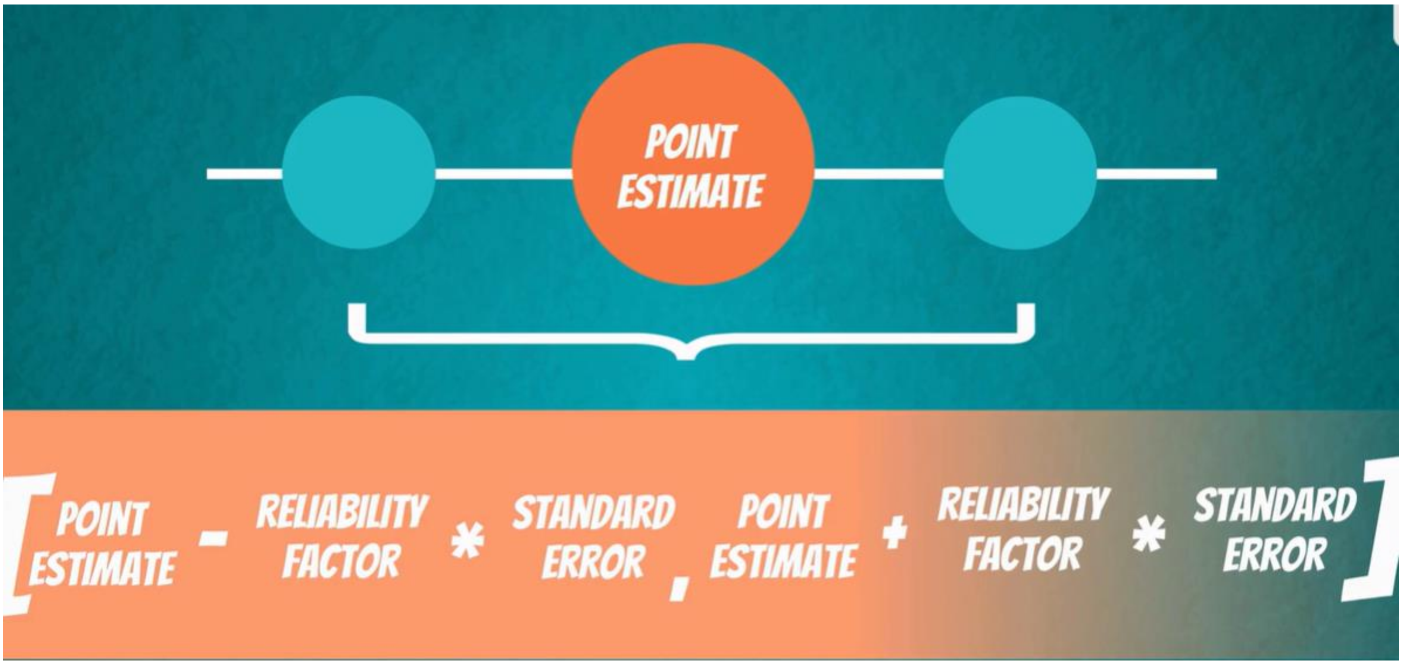

EFFICIENCY
The most efficient estimator is the unbiased estimator with smallest variance

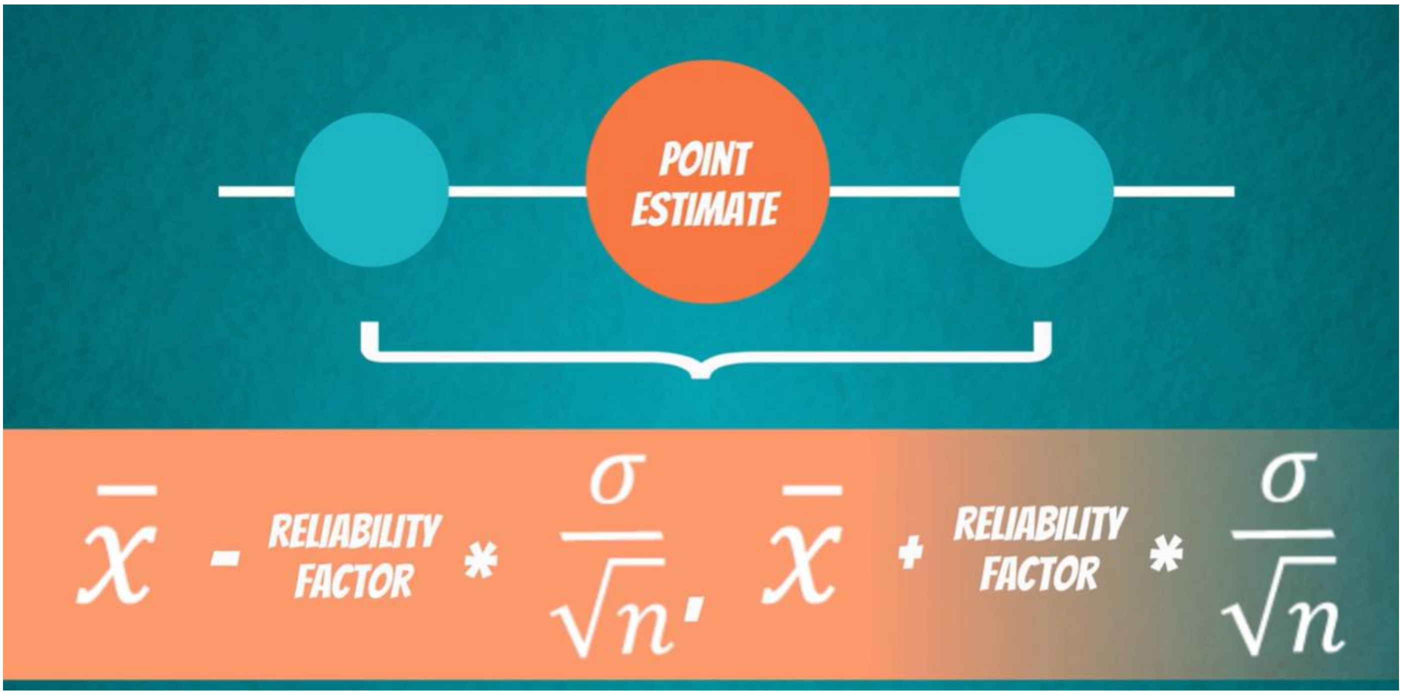

STATISTICS
ESTIMATORS
broader term
a type of statistic


20
AVERAGE PRICE
22.50 £
25
μ
95% CI MEANS THERE IS ONLY 5% CHANCE THAT THE
POPULATION PARAMETER IS OUTSIDE THE RANGE


CONFIDENCE LEVEL
0 ≤ α ≤ 1
1 - α


CONFIDENCE LEVEL
0 ≤ α ≤ 1
1 - α





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

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

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

Confidence intervals. Population known, z-score

Data scientist salary

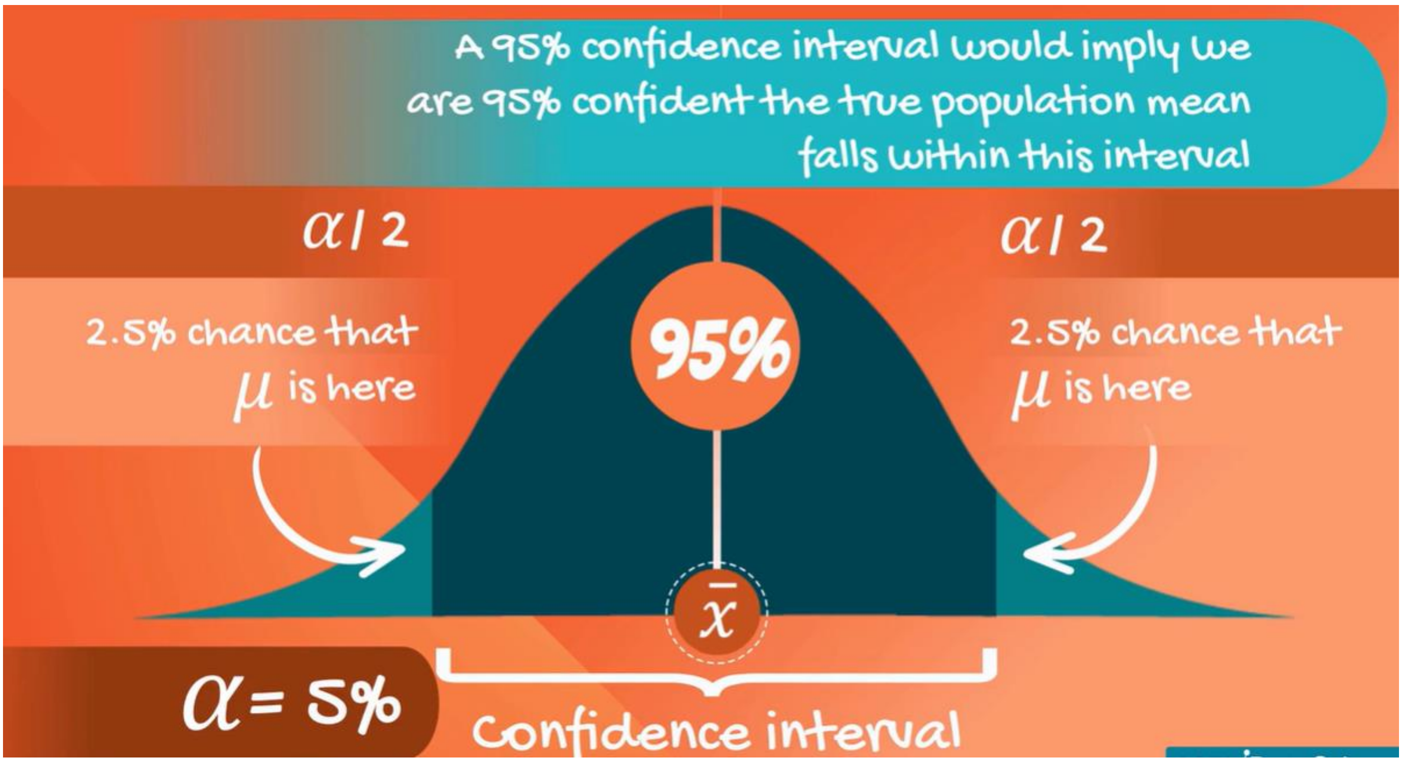
Dataset
$117,313
$104,002
$113,038
$101,936
$ 84,560
$113,136
$ 80,740
$100,536
$105,052
$ 87,201
$ 91,986
$ 94,868
$ 90,745
$102,848
$ 85,927
$112,276
$108,637
$ 96,818
$ 92,307

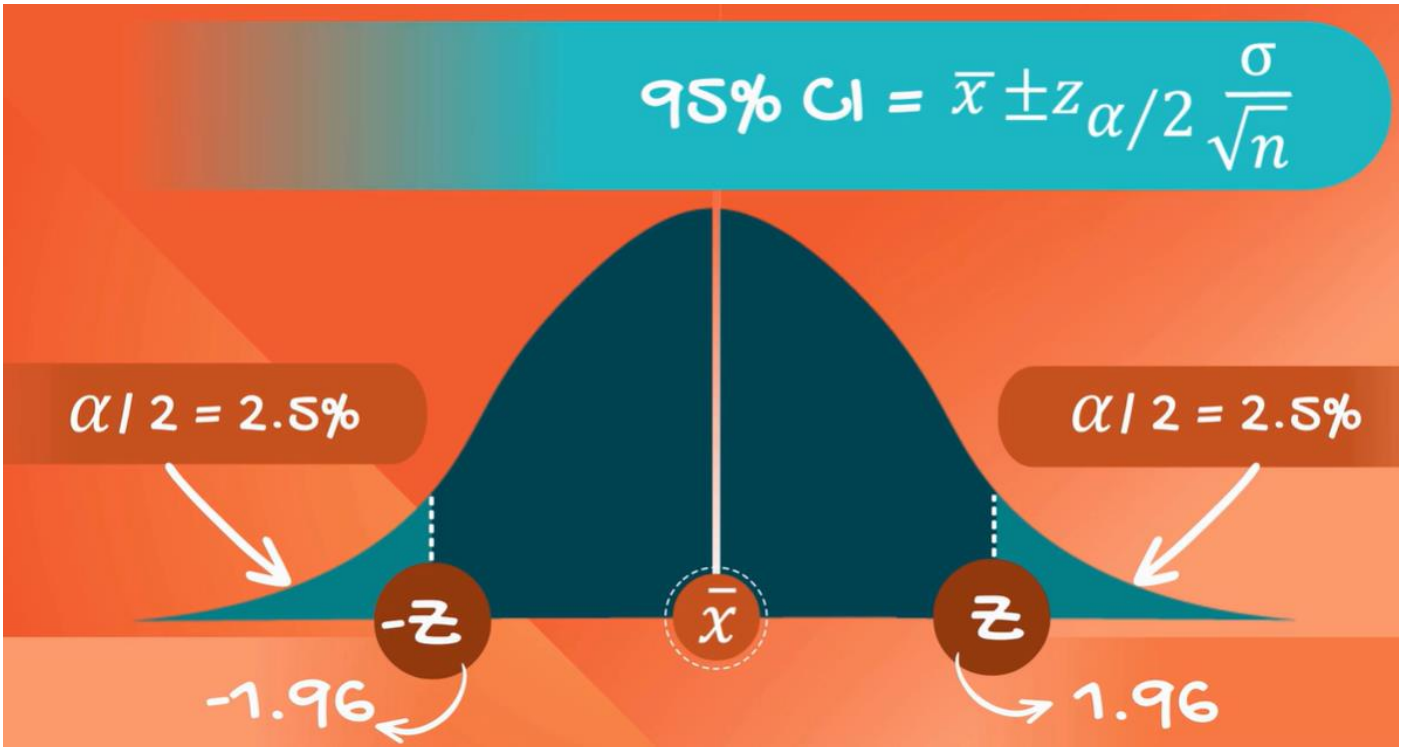
Sample mean	$100,200
Population std	$ 15,000
Standard error	$ 2,739

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

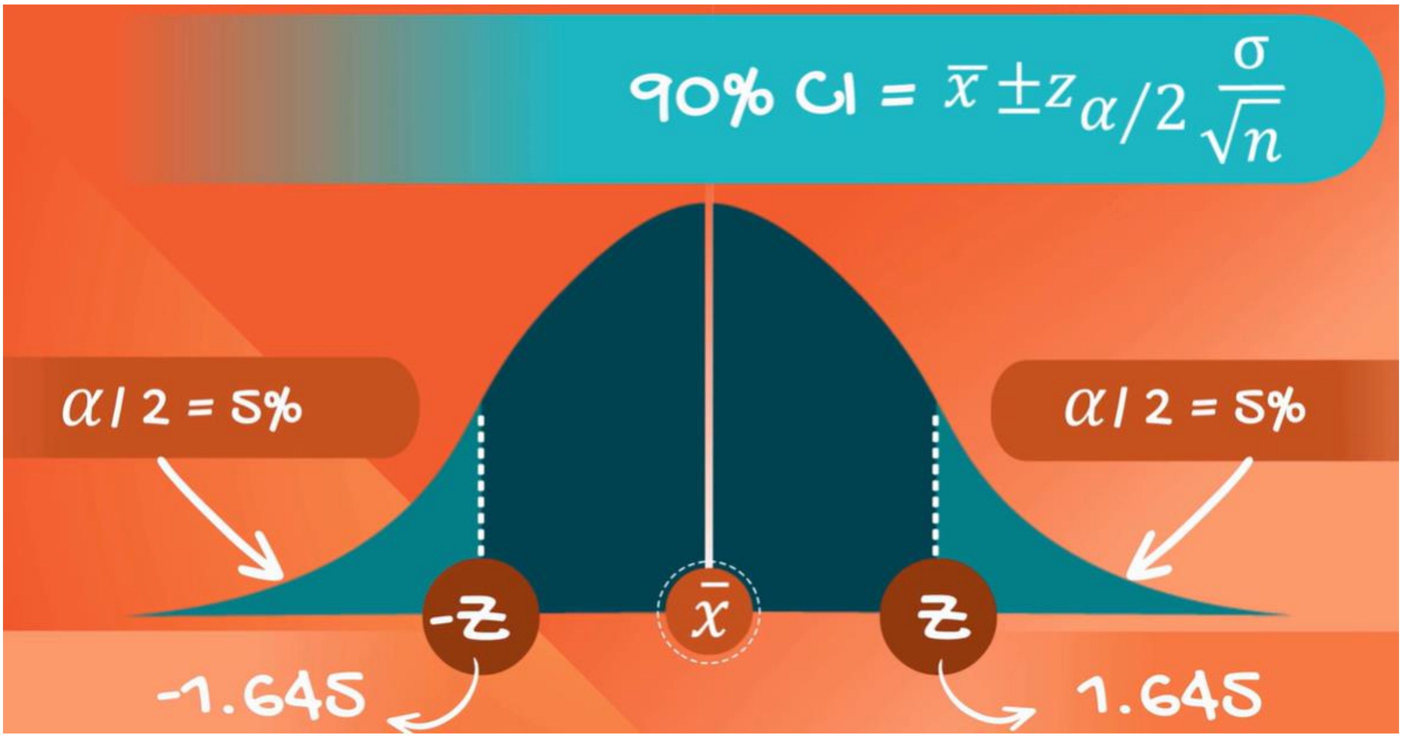
$$[100200-2.58\frac{15000}{\sqrt{30}} \ , 100200 + 2.58\frac{15000}{\sqrt{30}}] = [93135 \ , 107206]$$

We are 99% confident that the average data scientist salary is going to lie in the interval [$93135 , $107206]

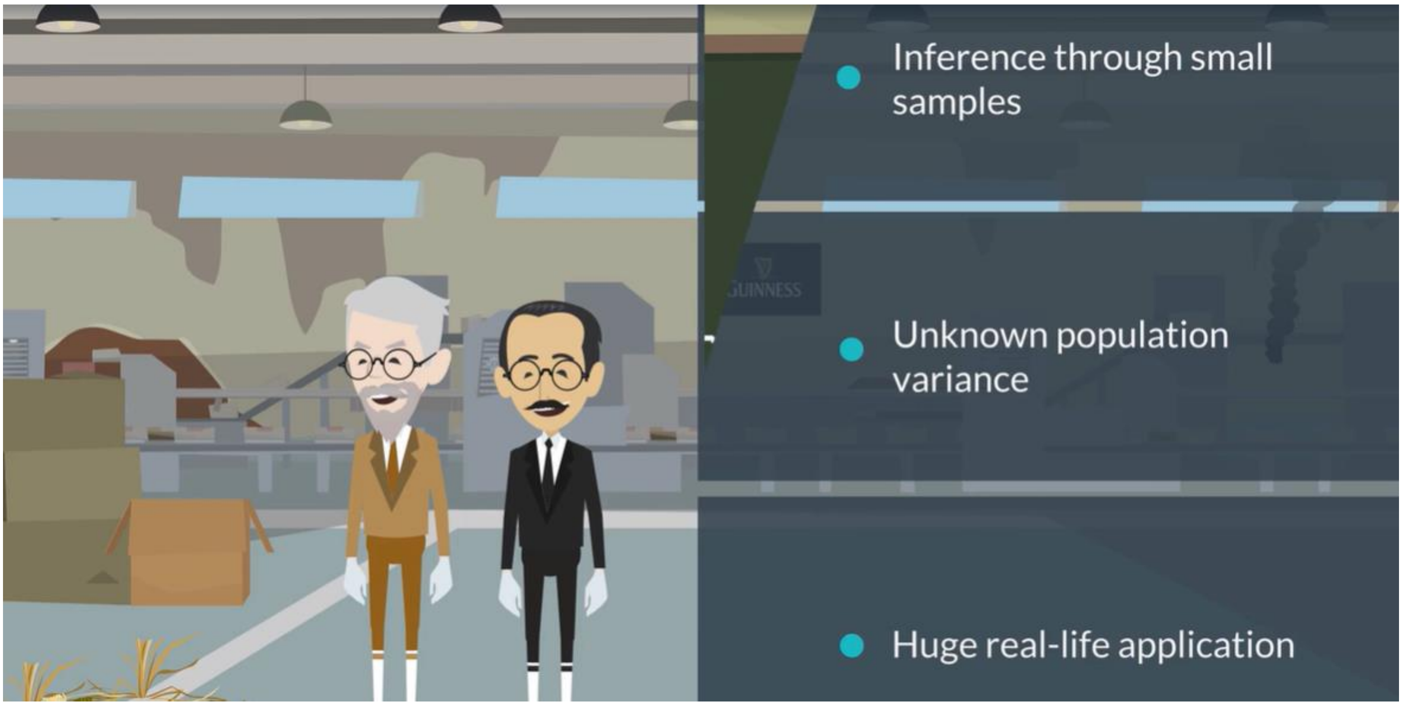


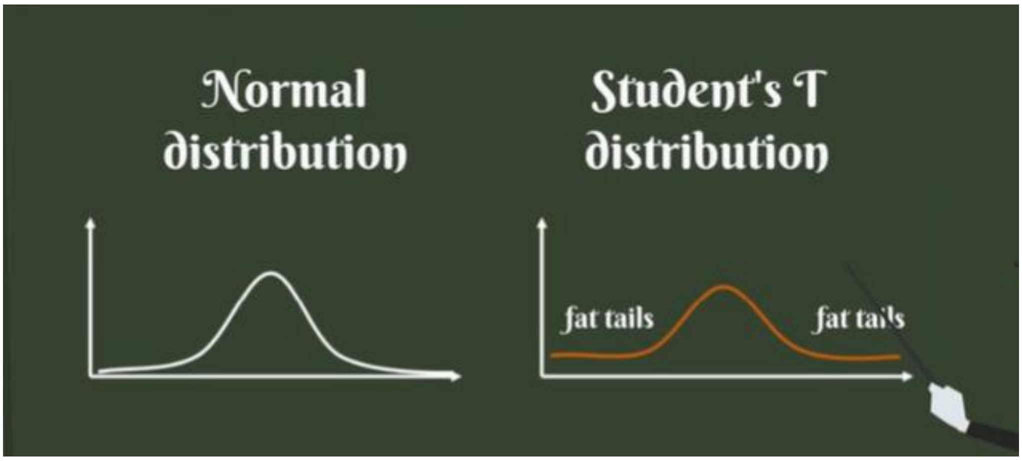


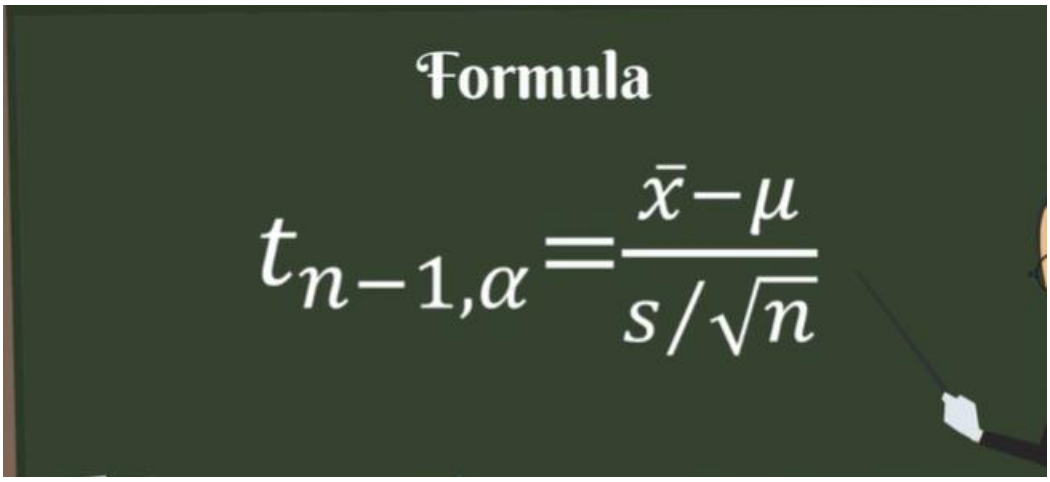
$95\%\ CI = \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
$\alpha / 2 = 2.5\%$
$\alpha / 2 = 2.5\%$
-z
$\bar{x}$
z
-1.96
1.96



$90\%\ CI = \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
$\alpha / 2 = 5\%$
$\alpha / 2 = 5\%$
-z
$\bar{x}$
z
-1.645
1.645


Inference through small samples
Unknown population variance
Huge real-life application


Normal distribution
Student's T distribution
fat tails
fat tails


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

Degrees of freedom (∂.f.)

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

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

Confidence intervals, t-score

Data scientist salary

	Dataset
$	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

Population variance unknown

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

Population variance known

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

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%

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

$$t_{8,}$$

95% CI => alpha = 5%

Confidence intervals, t-score

Data scientist salary

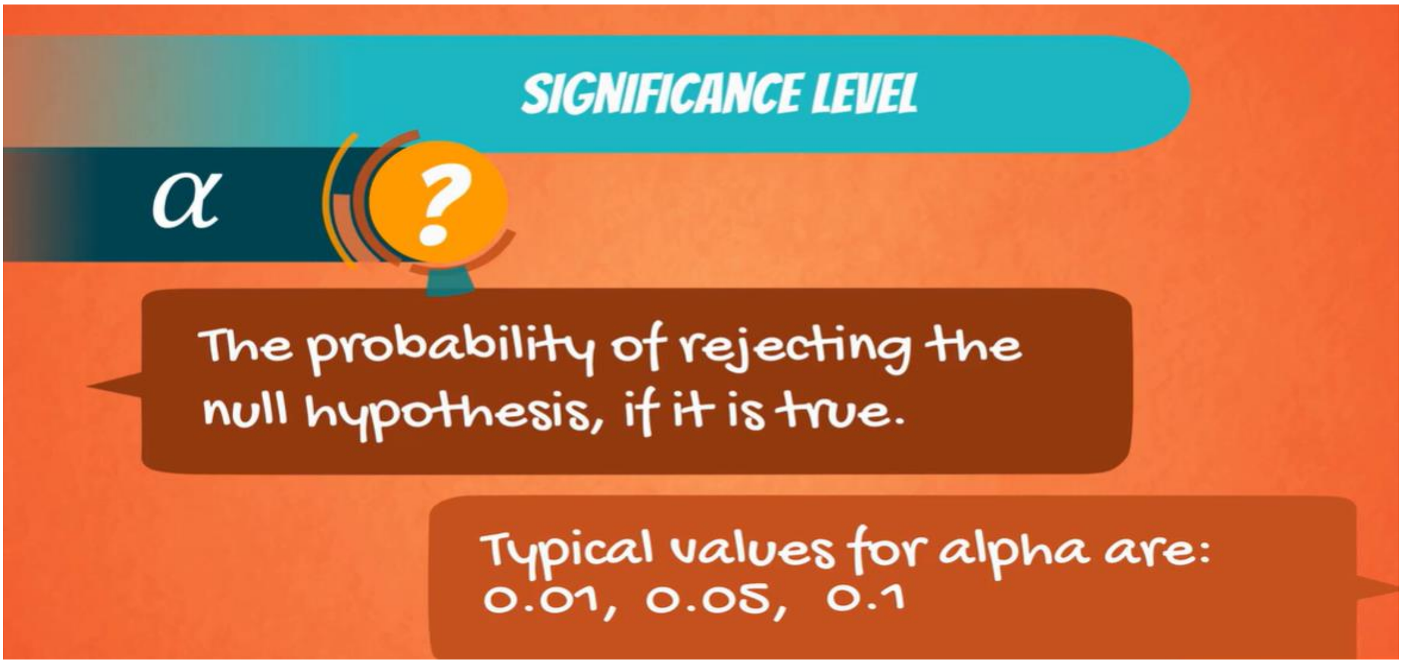
Data set
$ 78,000
$ 90,000
$ 75,000
$ 117,000
$ 105,000
$ 96,000
$ 89,500
$ 102,300
$ 80,000

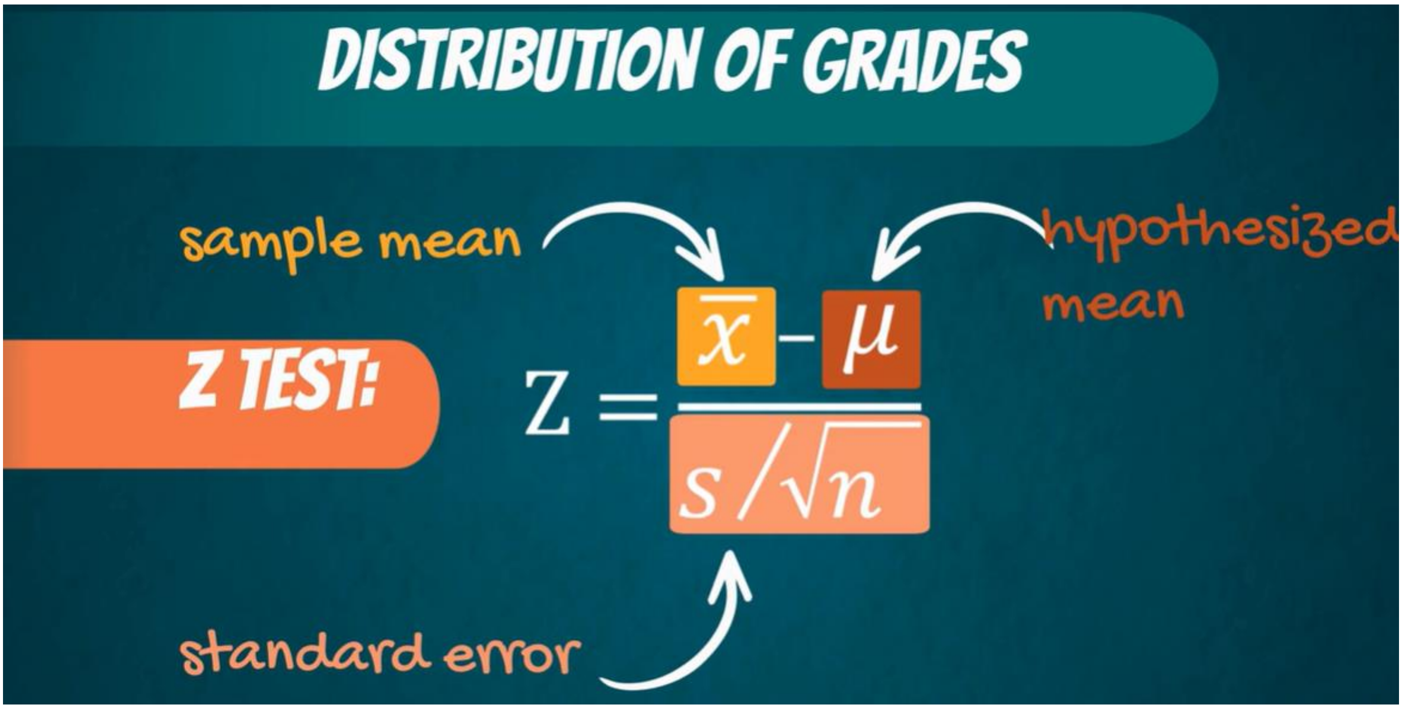
Sample mean	$ 92,533
Sample standard deviation	$ 13,932
Standard error	$ 4,644
t-stat 95%	2.31

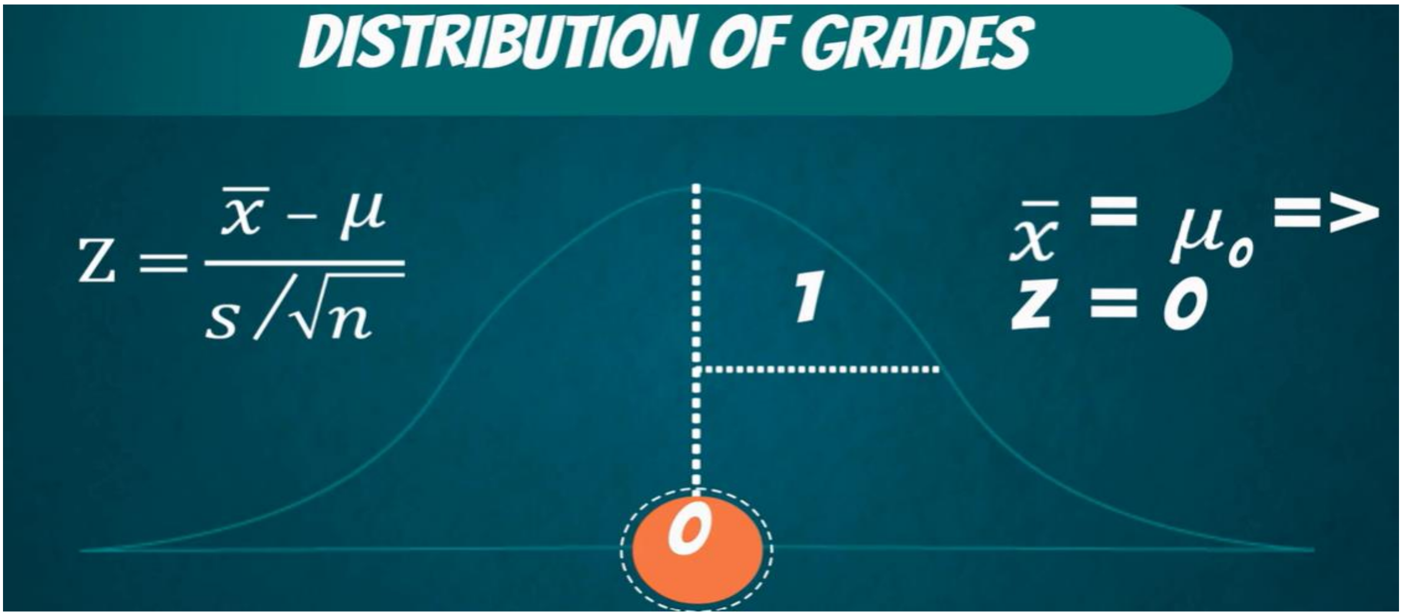
$$CI_{95\%,unknown} = (\$\,81806\ ,\ \$\,103261)$$ width = $21,455

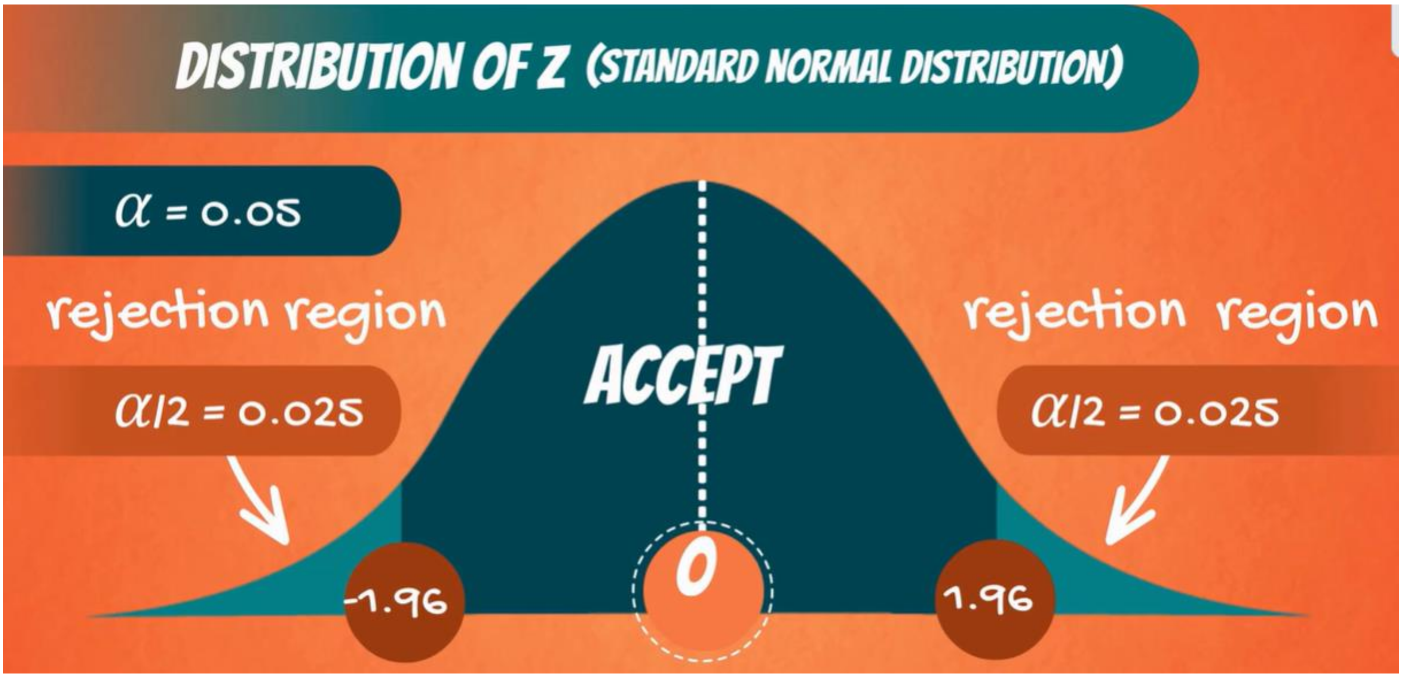
$$CI_{95\%,known} = (\$\,94833\ ,\ \$\,105568)$$ width = $10,735

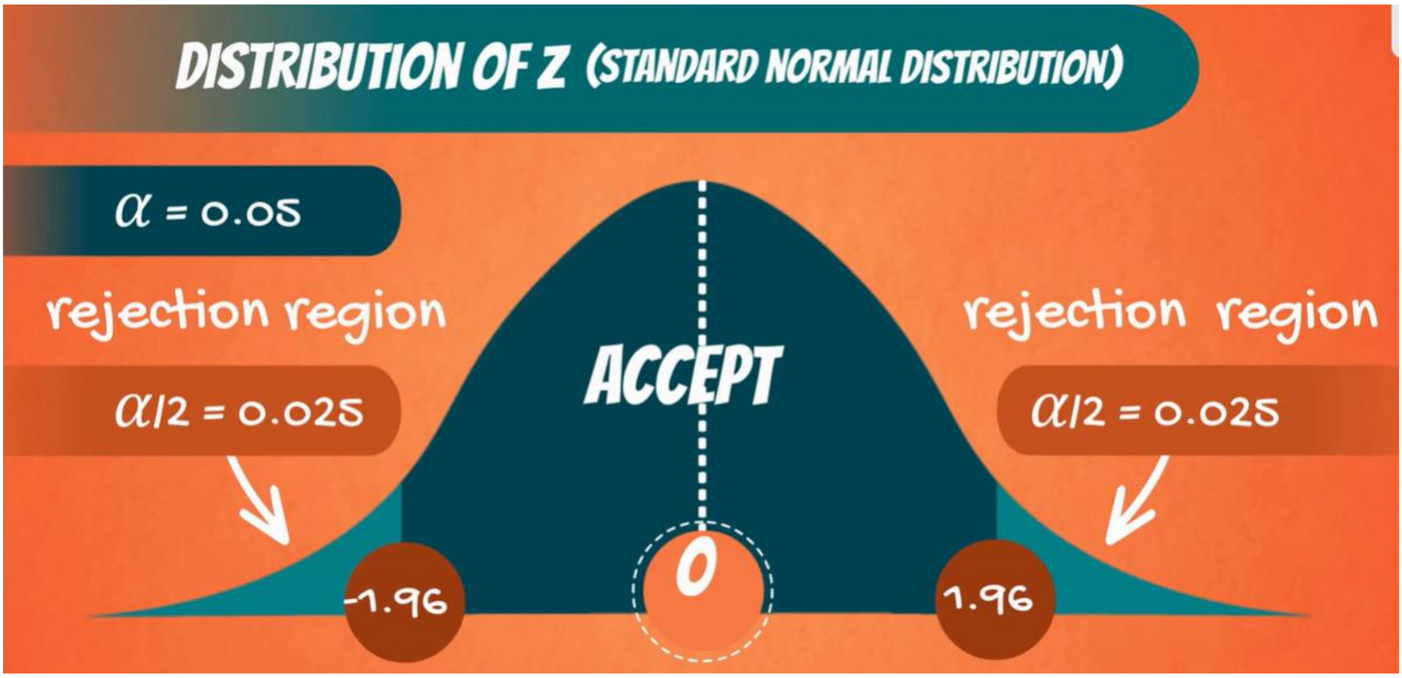
*Here we've got two effects: 1) smaller sample size and 2) unknown population variance
Both contribute to the width of the interval

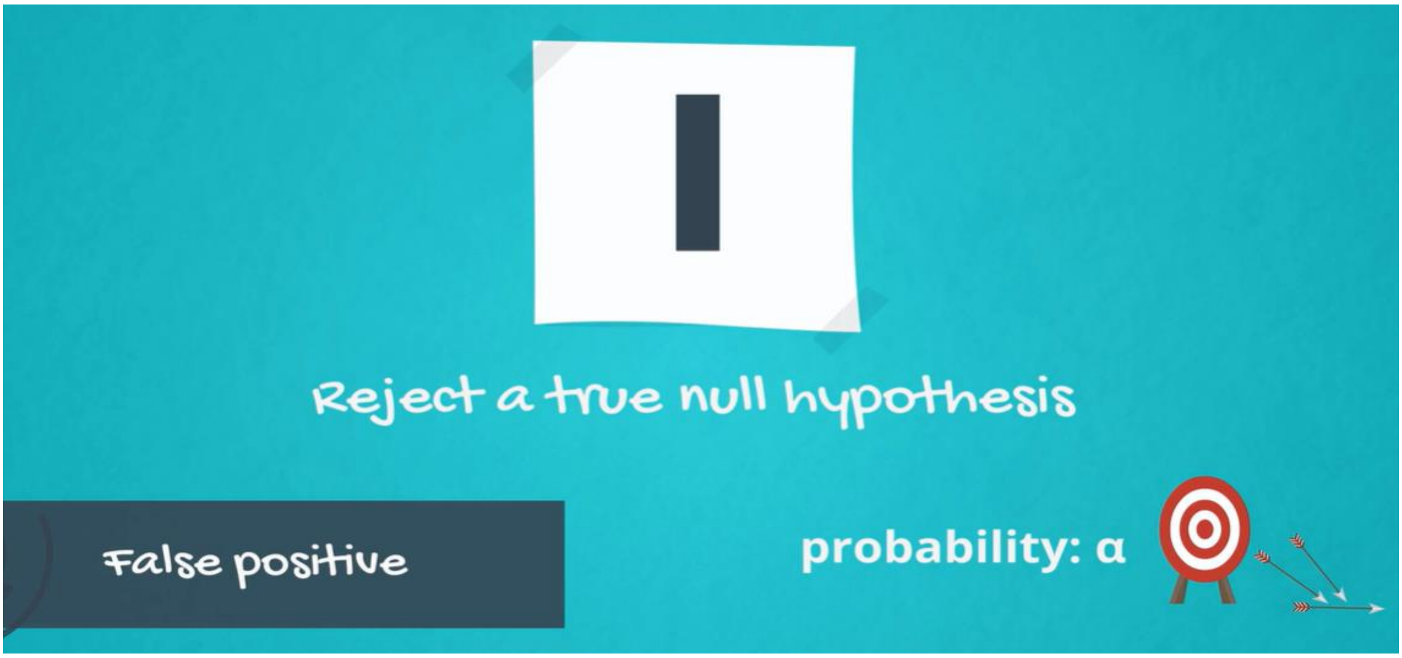


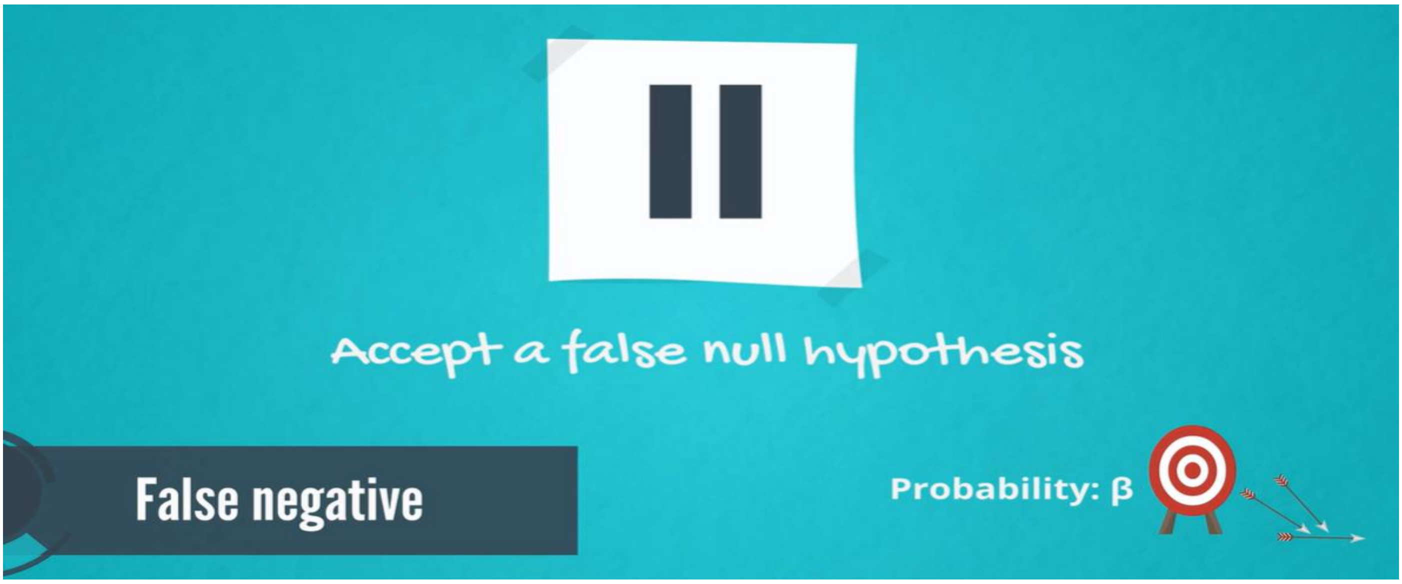

DISTRIBUTION OF GRADES
sample mean
hypothesized mean
Z TEST:
$Z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$
standard error


DISTRIBUTION OF GRADES
$Z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$
1
$\bar{x} = \mu_0 \Rightarrow$
$Z = 0$
0

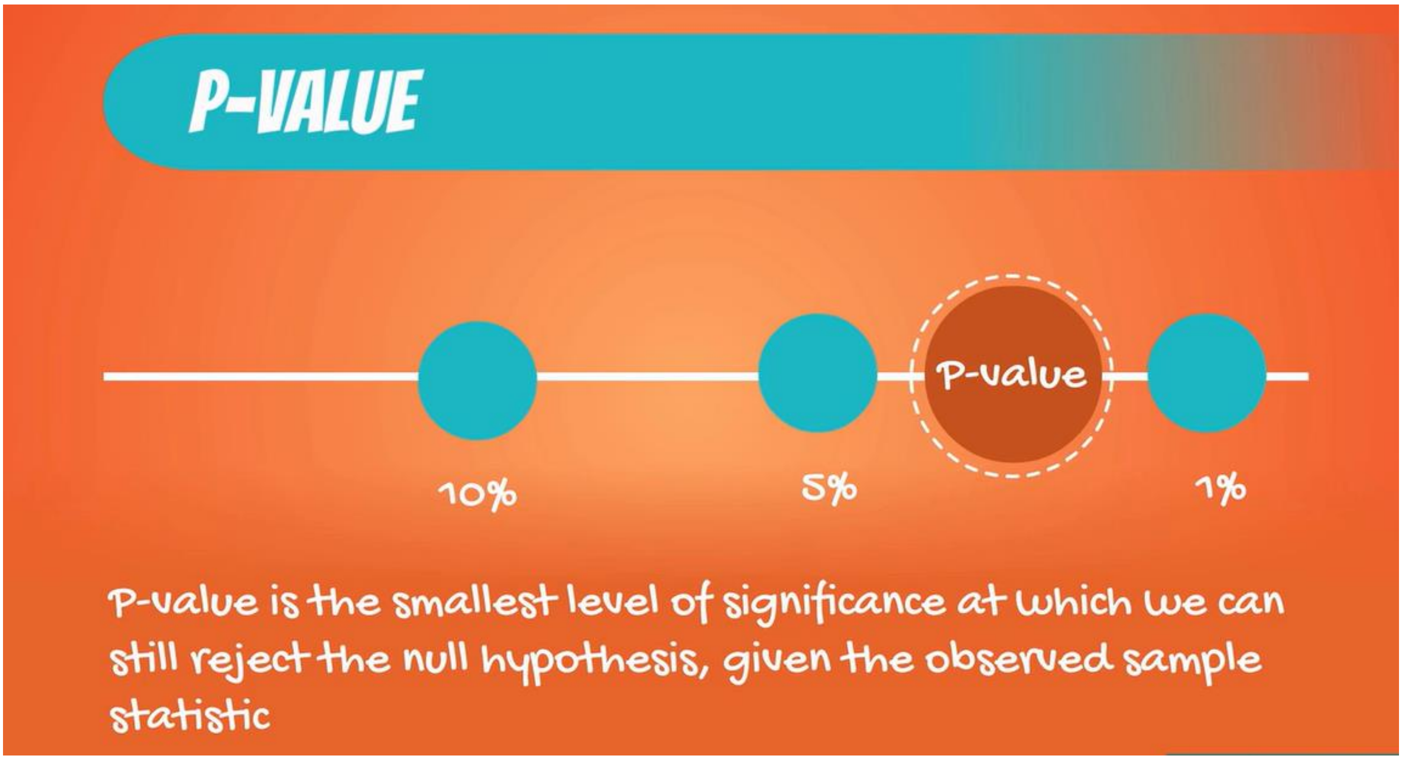

DISTRIBUTION OF Z (STANDARD NORMAL DISTRIBUTION)
$\alpha = 0.05$
rejection region
$\alpha/2 = 0.025$
ACCEPT
rejection region
$\alpha/2 = 0.025$
-1.96
0
1.96

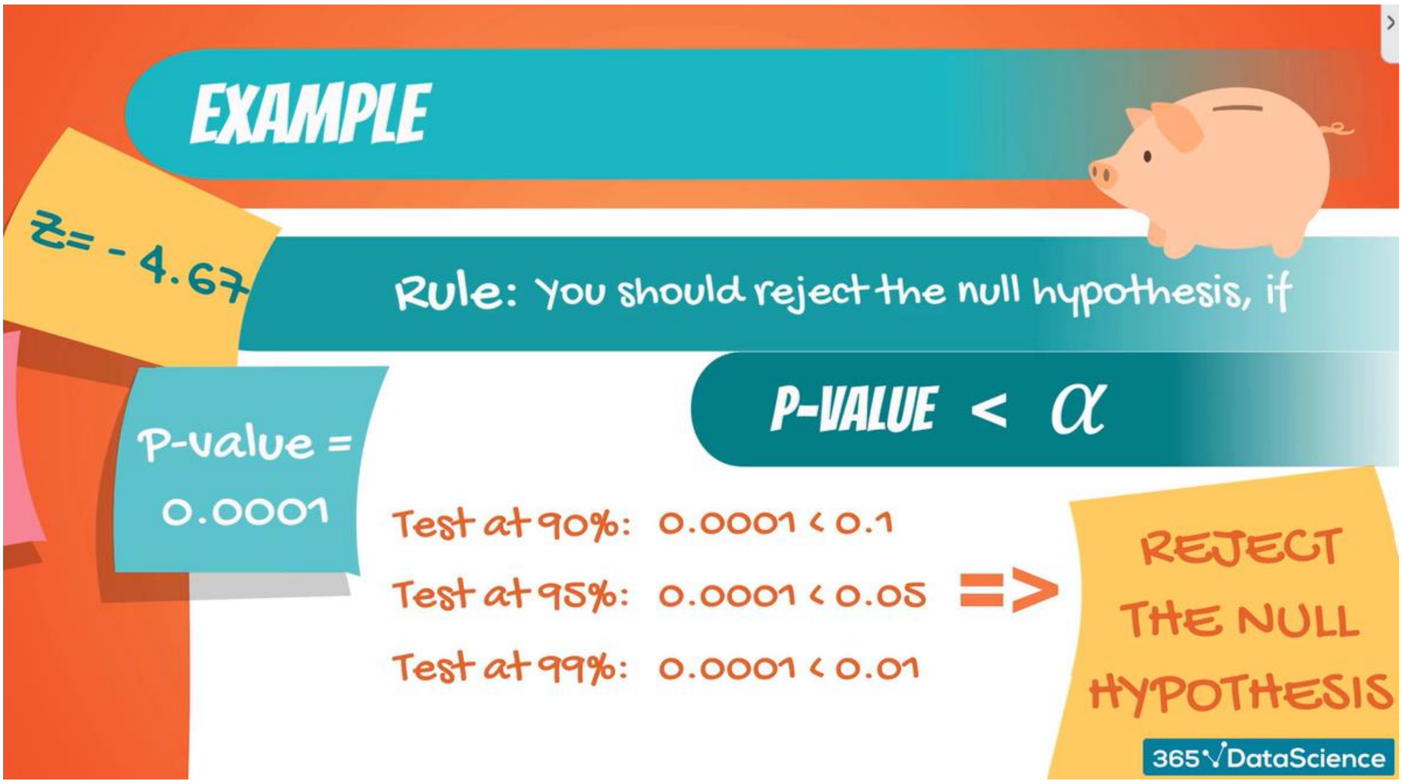

DISTRIBUTION OF Z (STANDARD NORMAL DISTRIBUTION)
α = 0.05
rejection region
α/2 = 0.025
ACCEPT
rejection region
α/2 = 0.025
-1.96
0
1.96

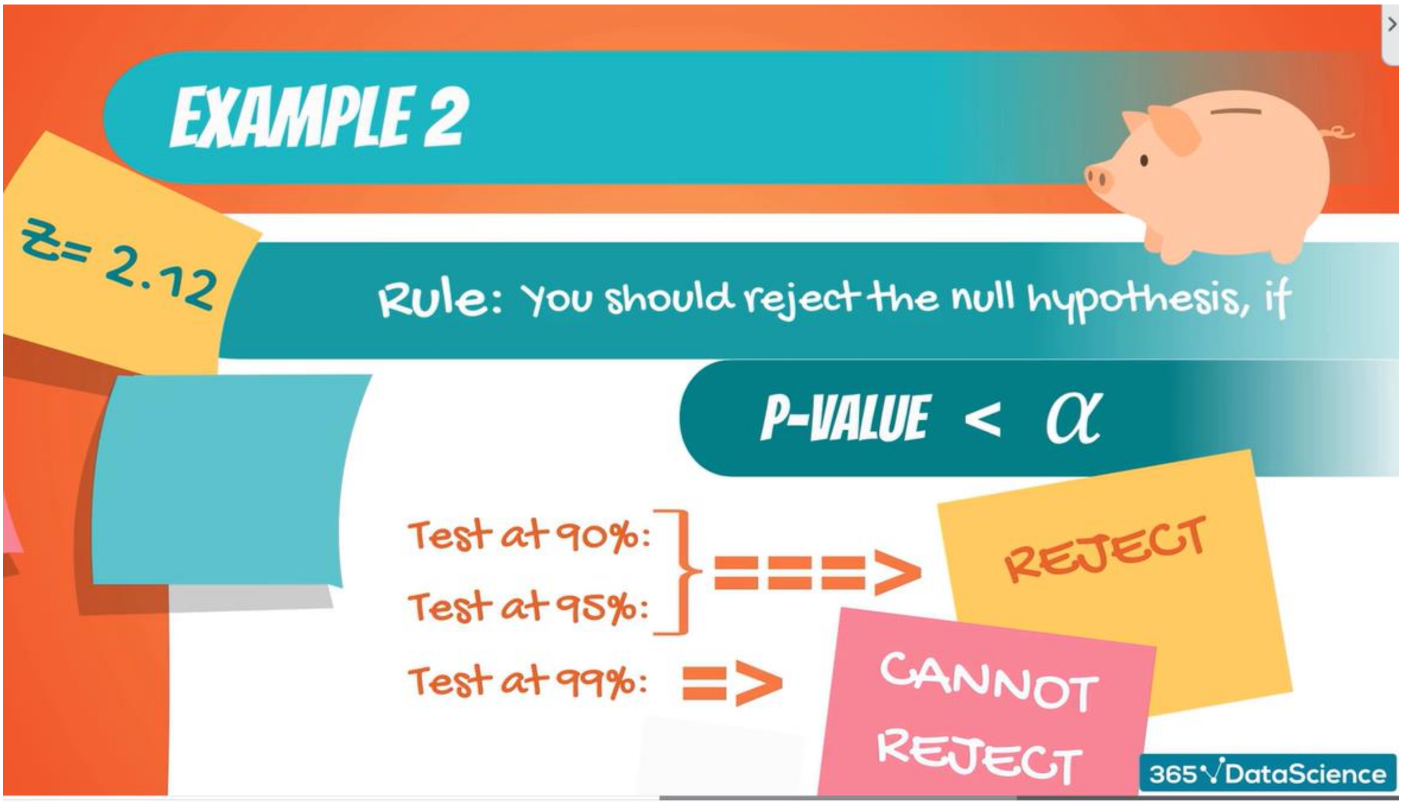

I
Reject a true null hypothesis
False positive
probability: α

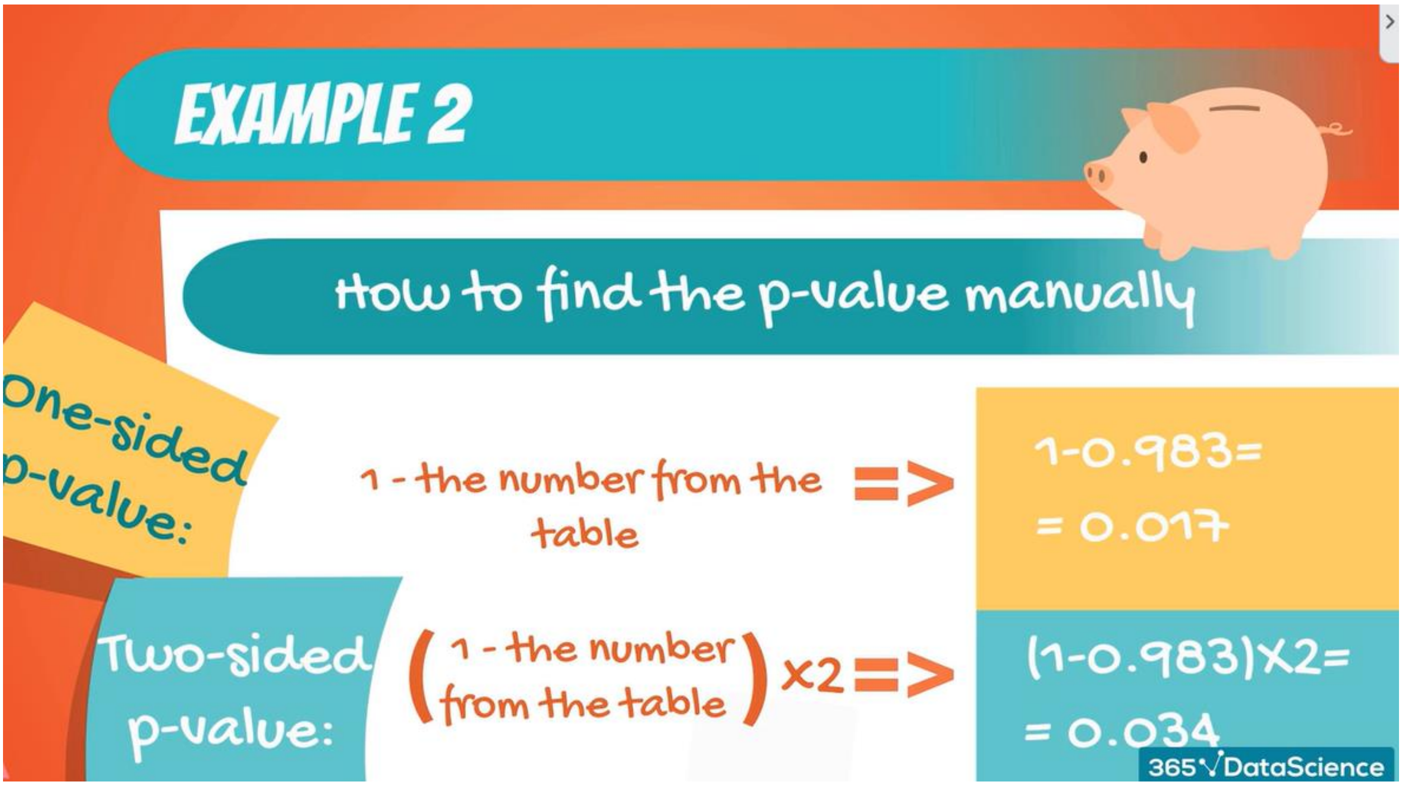

II
Accept a false null hypothesis
False negative
Probability: β

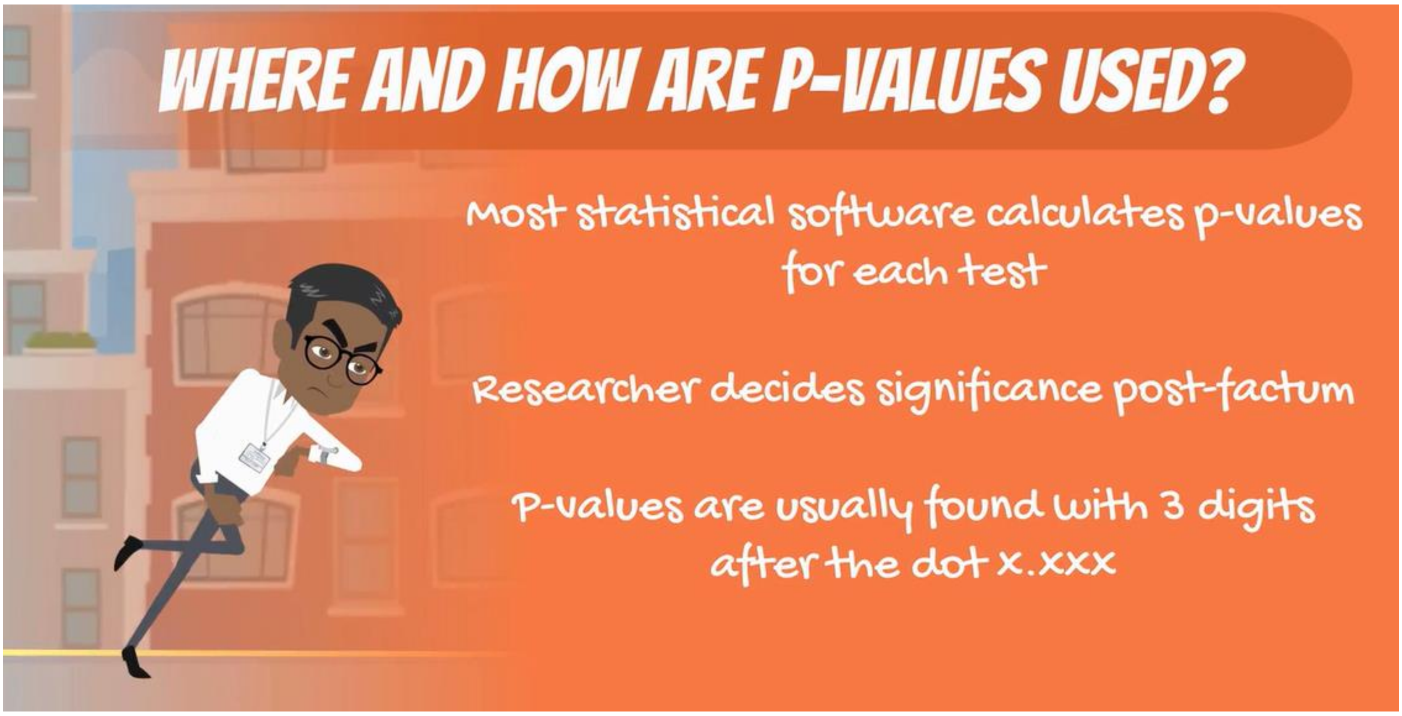

Goal of hypothesis testing
Rejecting a false null hypothesis
Probability:1-β
a.k.a. power of the test


P-VALUE
P-value
10%
5%
1%
P-value is the smallest level of significance at which we can still reject the null hypothesis, given the observed sample statistic


EXAMPLE
Z= - 4.67
Rule: You should reject the null hypothesis, if
P-VALUE < α
P-value =
0.0001
Test at 90%: 0.0001 < 0.1
Test at 95%: 0.0001 < 0.05
Test at 99%: 0.0001 < 0.01
=>
REJECT
THE NULL
HYPOTHESIS
365√DataScience


EXAMPLE 2
Z= 2.12
Rule: You should reject the null hypothesis, if
P-VALUE < α
Test at 90%:
Test at 95%:
====>
REJECT
Test at 99%:
=>
CANNOT
REJECT
365√DataScience


EXAMPLE 2
How to find the p-value manually
One-sided p-value:
1 - the number from the table =>
1-0.983=
= 0.017
Two-sided p-value:
(1 - the number from the table) x2 =>
(1-0.983)x2=
= 0.034
365√DataScience


WHERE AND HOW ARE P-VALUES USED?
Most statistical software calculates p-values for each test
Researcher decides significance post-factum
P-values are usually found with 3 digits after the dot x.xxx


IF THE P-VALUE IS LOWER THAN THE LEVEL OF SIGNIFICANCE
YOU REJECT THE NULL HYPOTHESIS