دورهٔ آموزشی «علم داده»

Data Science Course

جلسهٔ دهم- بخش اول:

بازه‌ی اطمینان

Confidence Interval

مدرس: محمد فزونی
عضو هیات علمی دانشگاه گنبدکاووس
پائیز ۱۳۹۹

دورهٔ آموزشی «علم داده»

Data Science Course

جلسهٔ دهم – بخش دوم:

بازه‌ی اطمینان

Confidence Interval

مدرس: محمد فزونی
عضو هیات علمی دانشگاه گنبدکاووس
پائیز ۱۳۹۹

دورهٔ آموزشی «علم داده»

Data Science Course

جلسهٔ دهم- بخش سوم:

بازه‌ی اطمینان

Confidence Interval

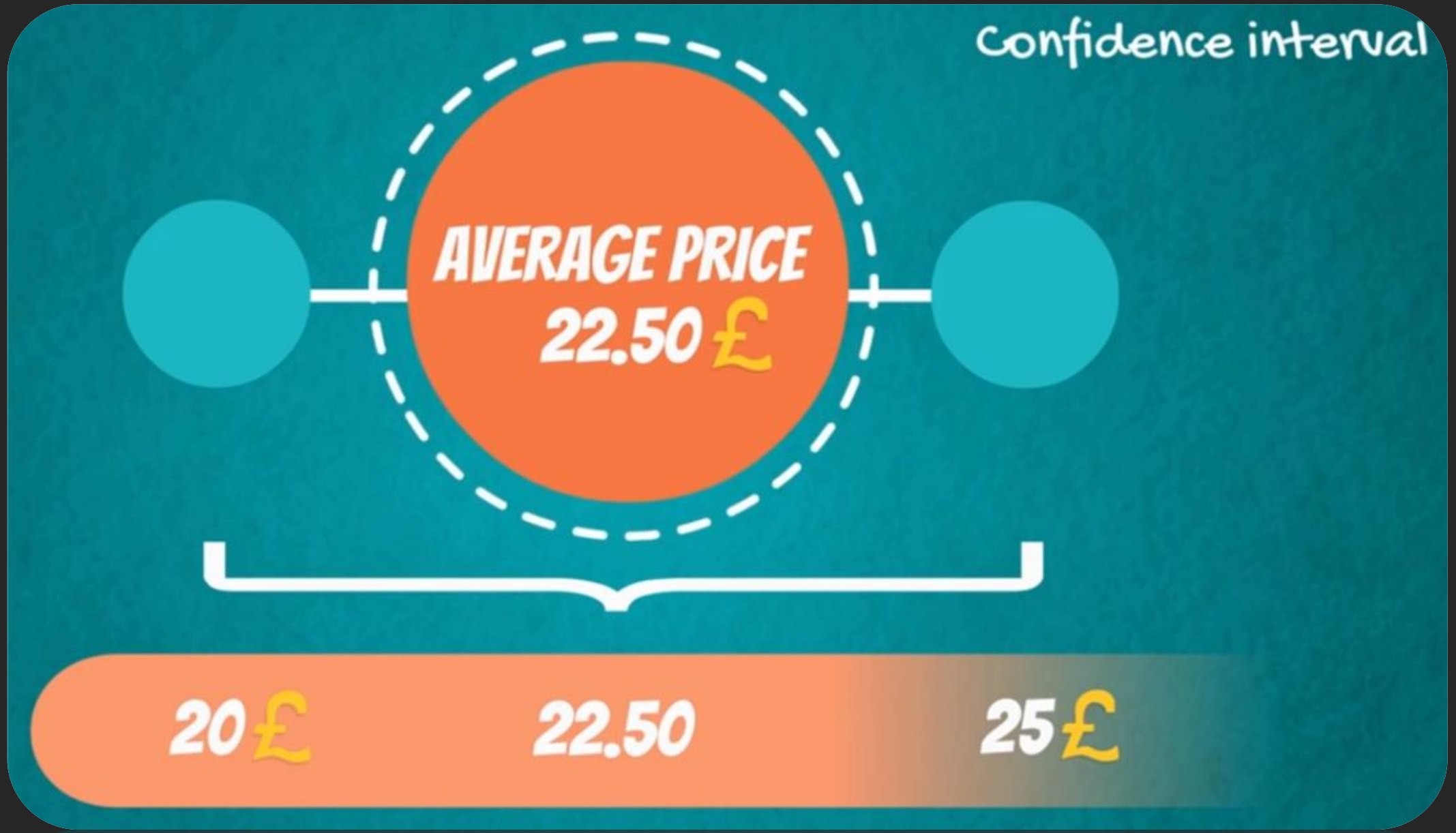
مدرس: محمد فزونی
عضو هیات علمی دانشگاه گنبدکاووس
پائیز ۱۳۹۹

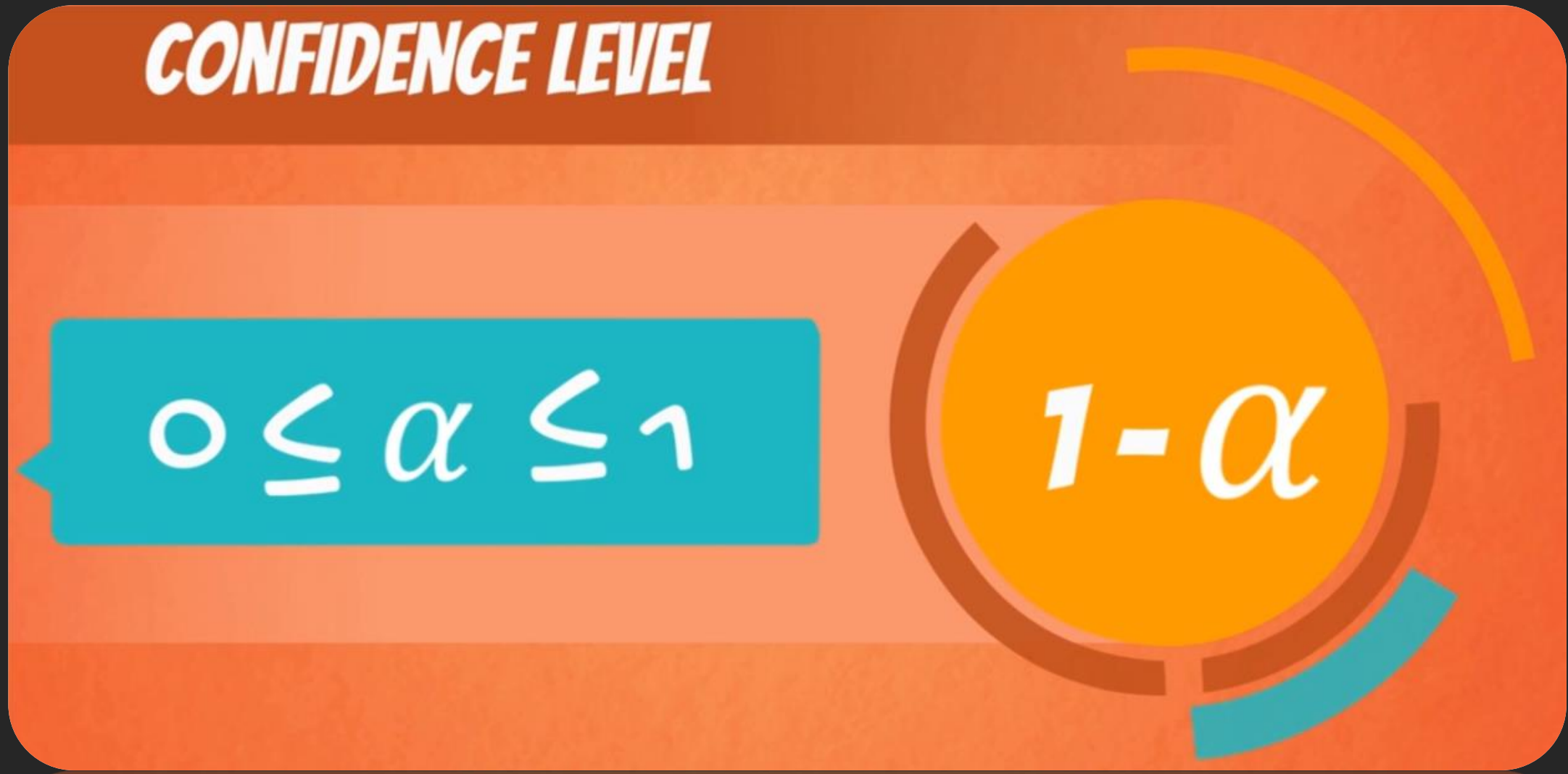
About me

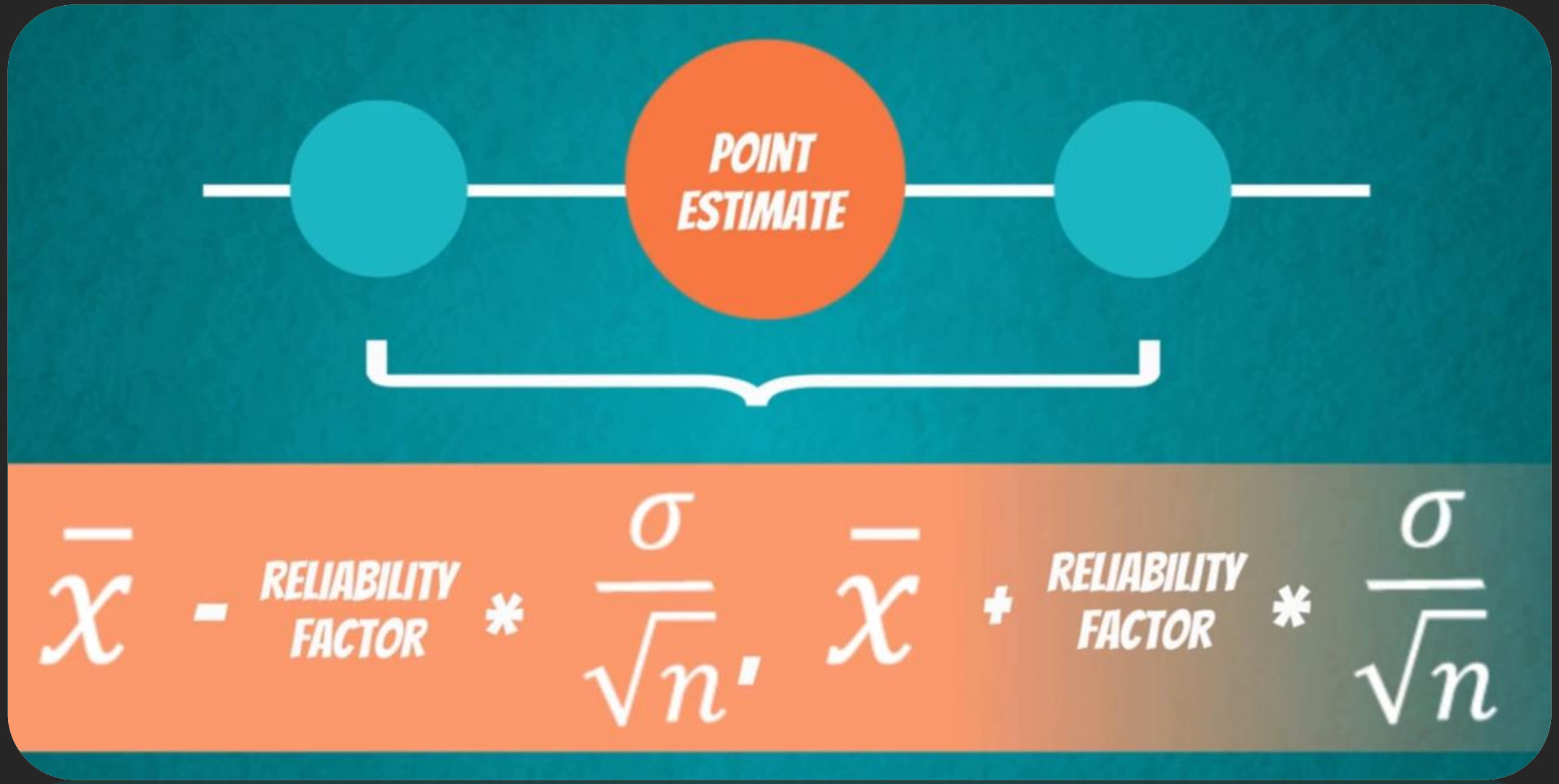
Mohammad Fozouni (Ph.D.)
Dep. of Math. & Stat.
Gonbad Kavous University

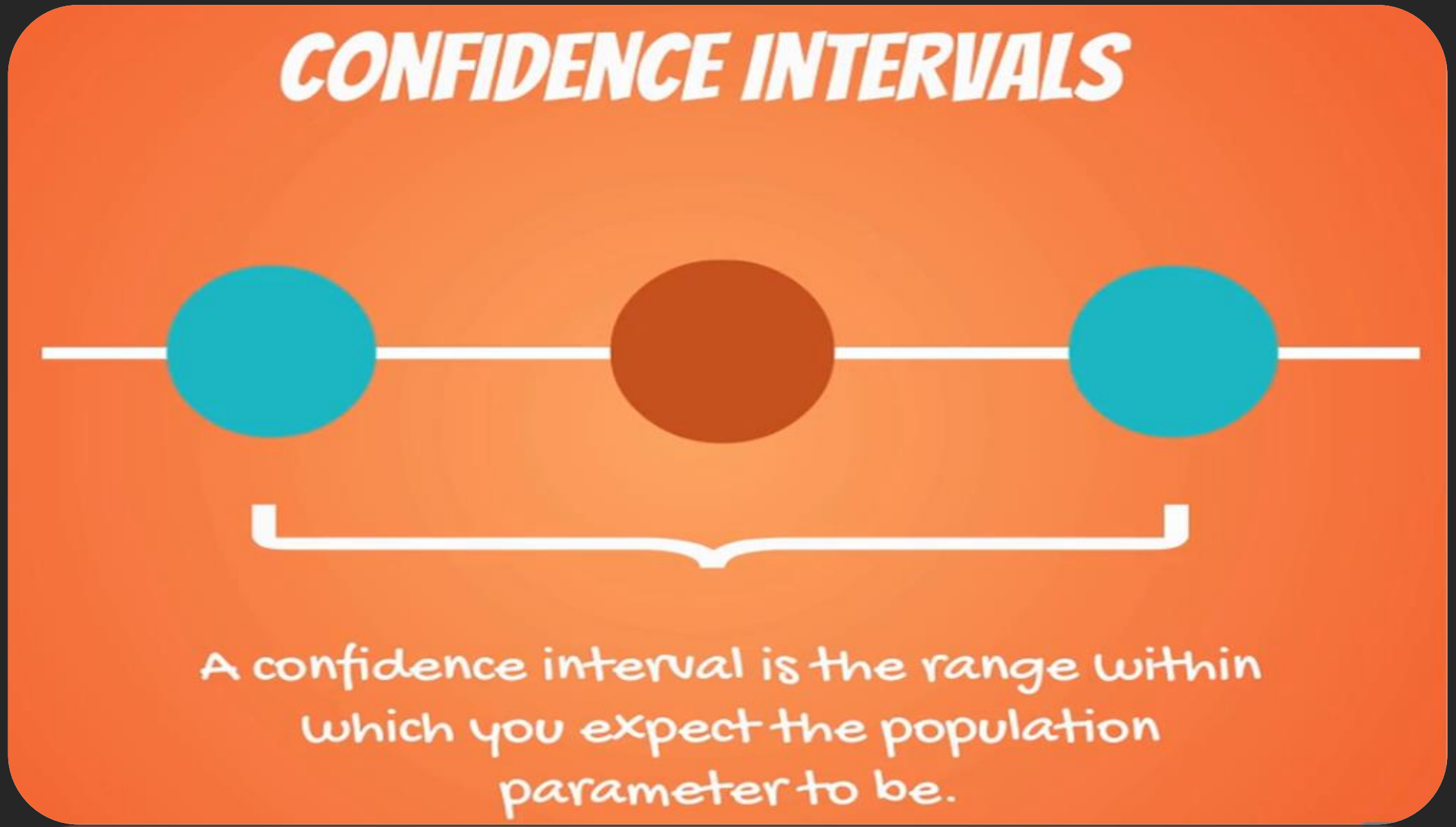
- fozouni@hotmail.com
- https://m-fozouni.ir
- http://profs.gonbad.ac.ir/fozouni/en

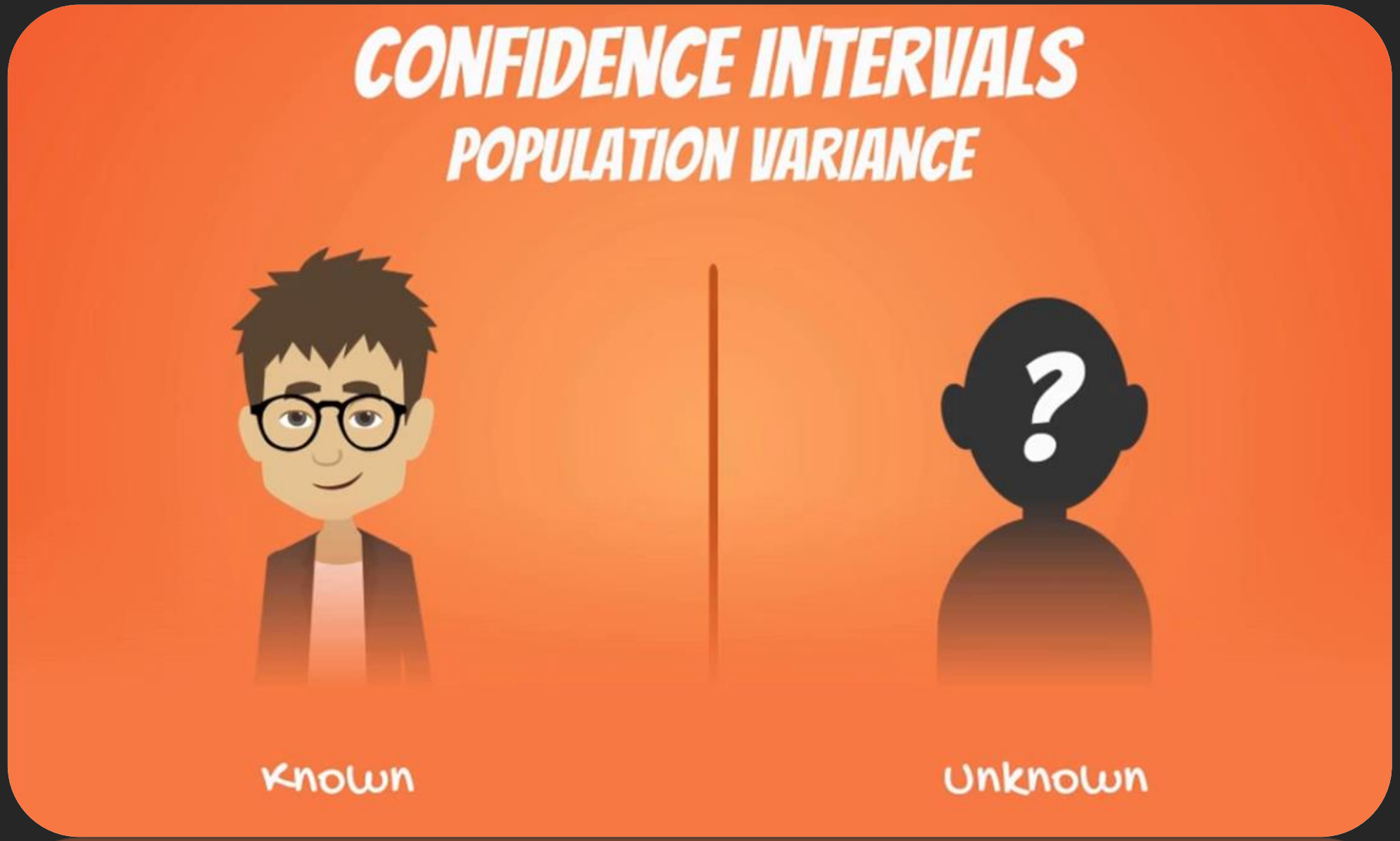
Instagram: @elmedade
#data_science_fozouni


Confidence interval
AVERAGE PRICE
22.50 £
20 £
22.50
25 £


CONFIDENCE LEVEL
$0 \leq \alpha \leq 1$
$1-\alpha$

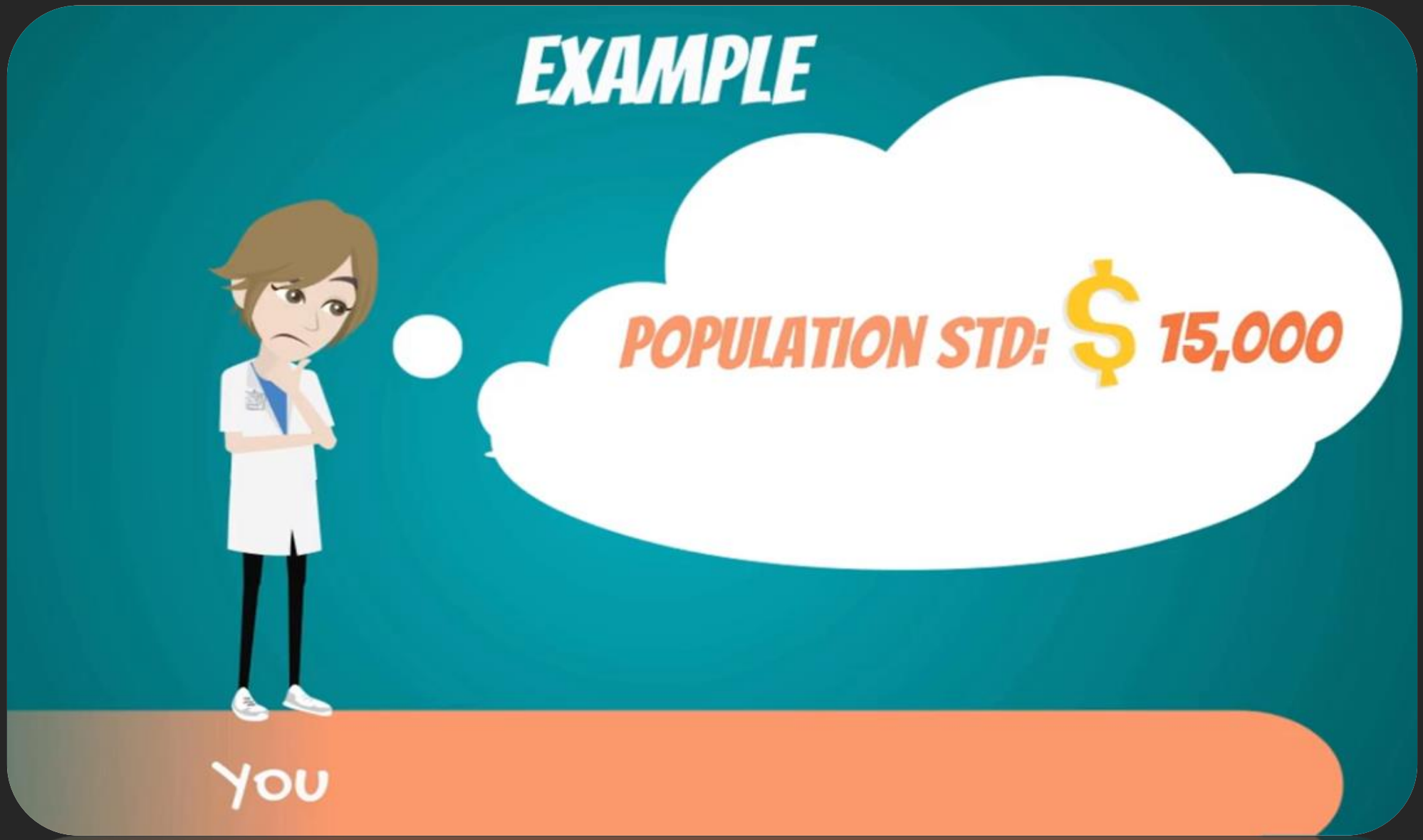

POINT ESTIMATE
$\bar{x}$ - RELIABILITY FACTOR * $\frac{\sigma}{\sqrt{n}}$, $\bar{x}$ + RELIABILITY FACTOR * $\frac{\sigma}{\sqrt{n}}$


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


CONFIDENCE INTERVALS
POPULATION VARIANCE
?
Known
Unknown


even the company that has the most data...
doesn't necessarily have population data

Finding the CI whenever we know the population variance


EXAMPLE
POPULATION STD: $ 15,000
You

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

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

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
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964

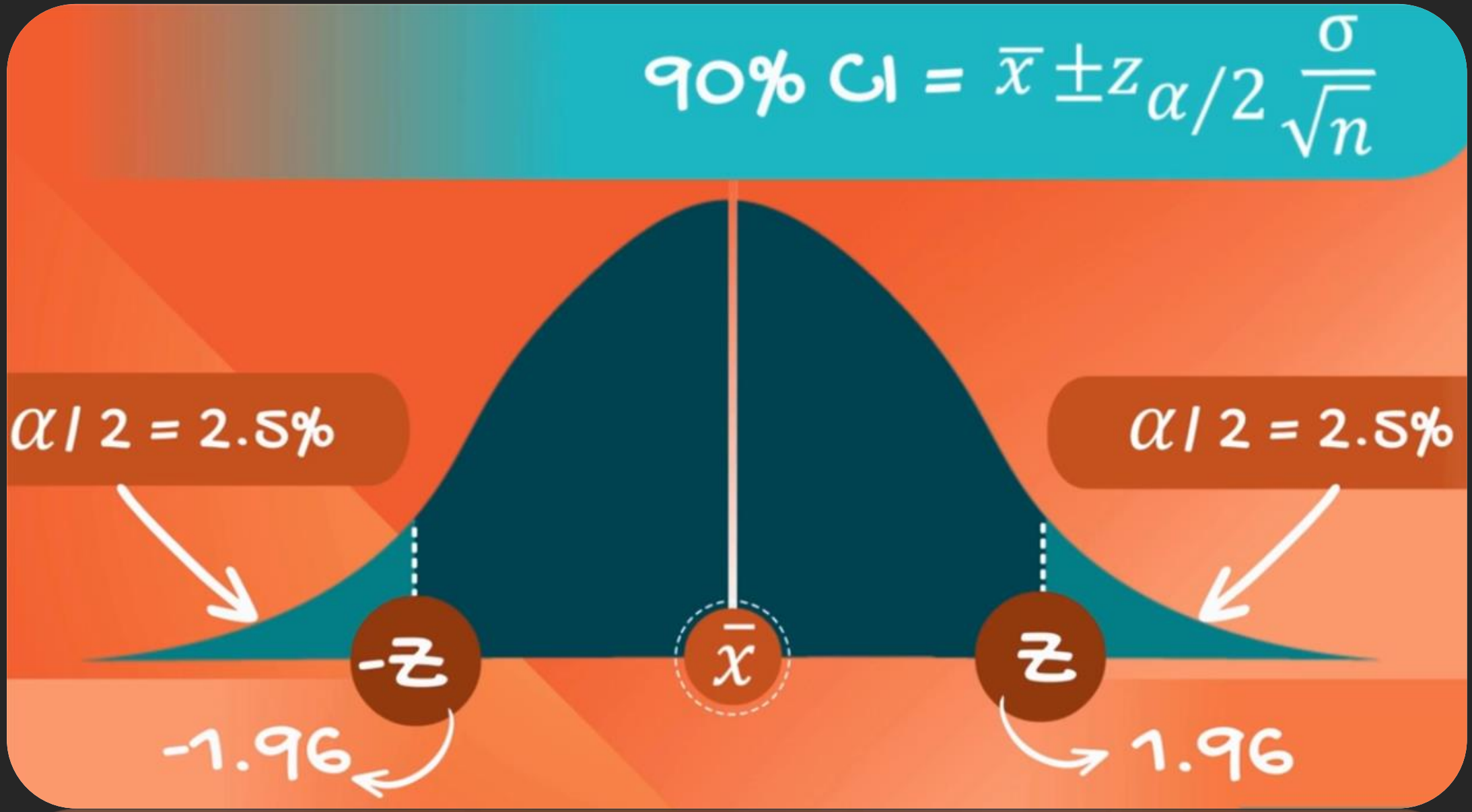
Confidence interval: 95%

$\alpha = 0.05$

$z_{0.025}$

$1 - 0.025 = 0.975$

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


90% CI = $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
α / 2 = 2.5%
α / 2 = 2.5%
-z
$\bar{x}$
z
-1.96
1.96

B C D E F G H I J

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

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

Confidence interval: 95% $= [94833 , 105568]$

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	[illegible]									0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	[illegible]	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	[illegible]	0.9964

Confidence interval: 99%

$\alpha = 0.01$

$1-0.005=0.995$

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

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]

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


A 95% confidence interval would imply we are 95% confident the true population mean falls within this interval
2.5% chance that μ is here
95%
2.5% chance that μ is here
$\bar{x}$
Confidence interval

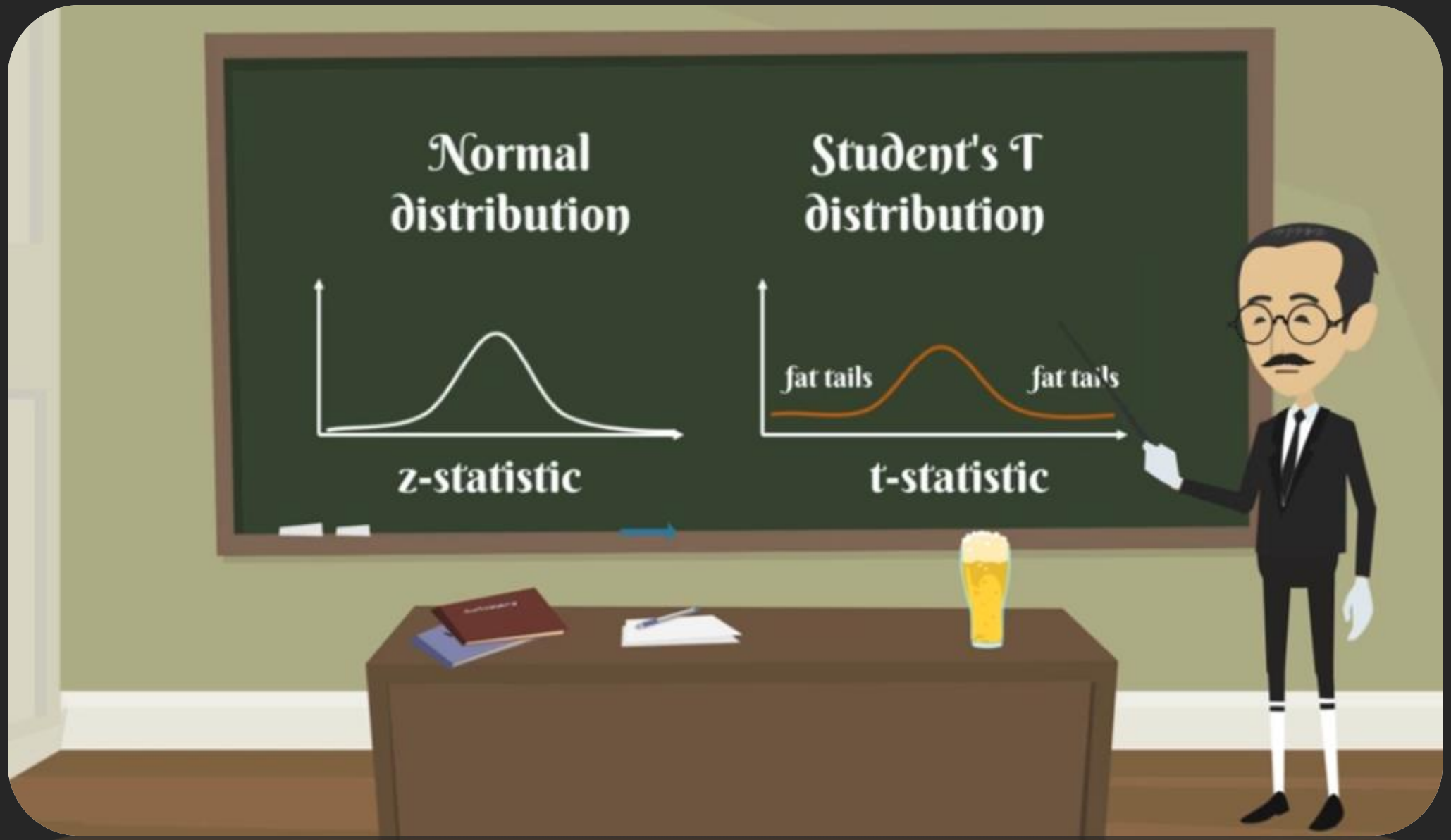
$1-\alpha = 90\%$

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

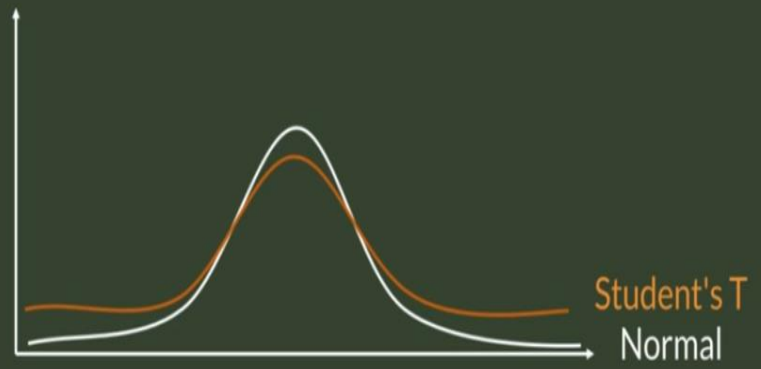
$1-\alpha = 99\%$

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

We do not know the variance of population


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

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

The 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
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

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

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

Again, we work on the "Data Scientists" salary problem. But this time, assume that we do not know the variance.

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

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%

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

$$t_{8,0.025}$$

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

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

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

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

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

When we know the variance and sample is bigger as in slide 17

CONFIDENCE INTERVALS FORMULAS

POPULATION VARIANCE

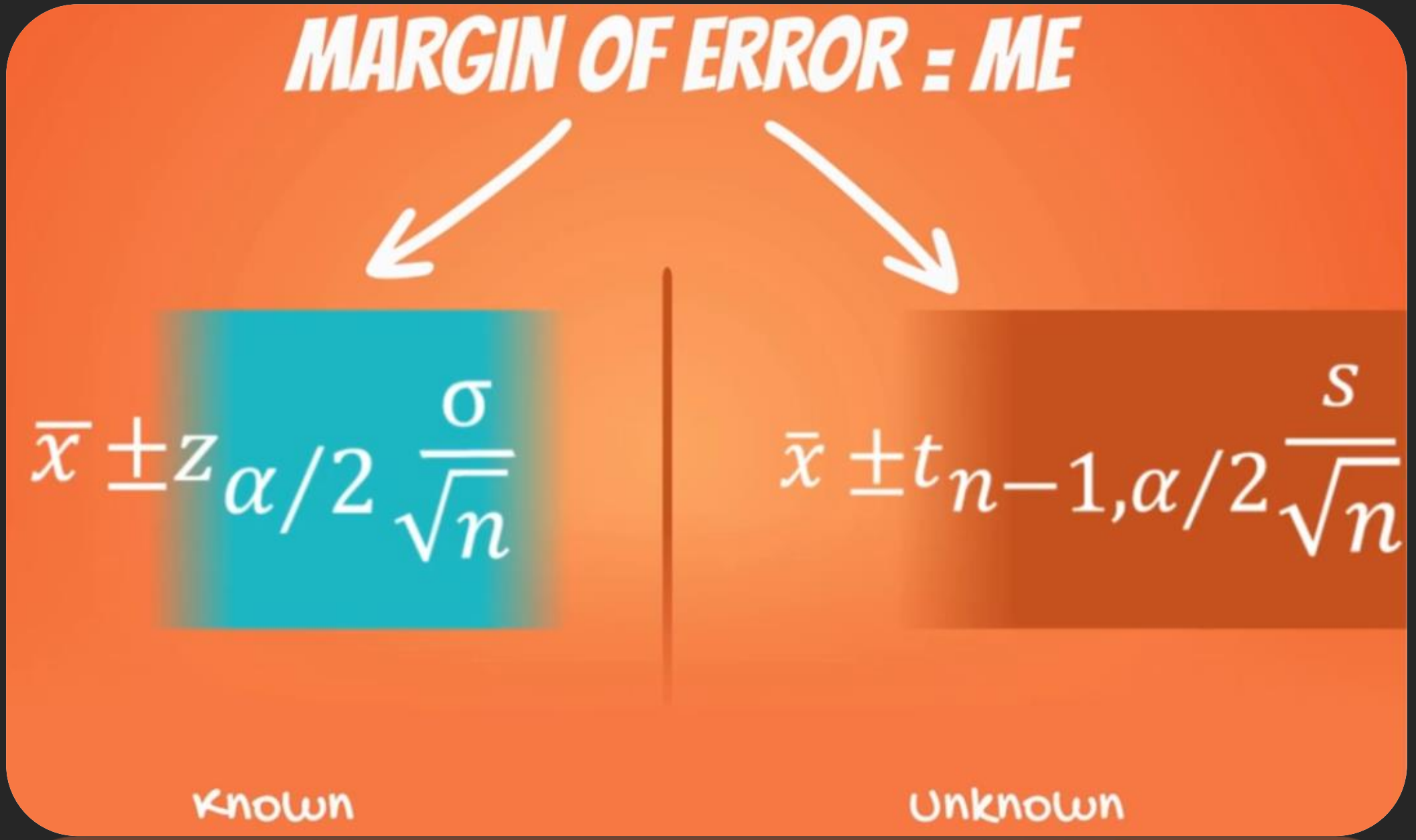
$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

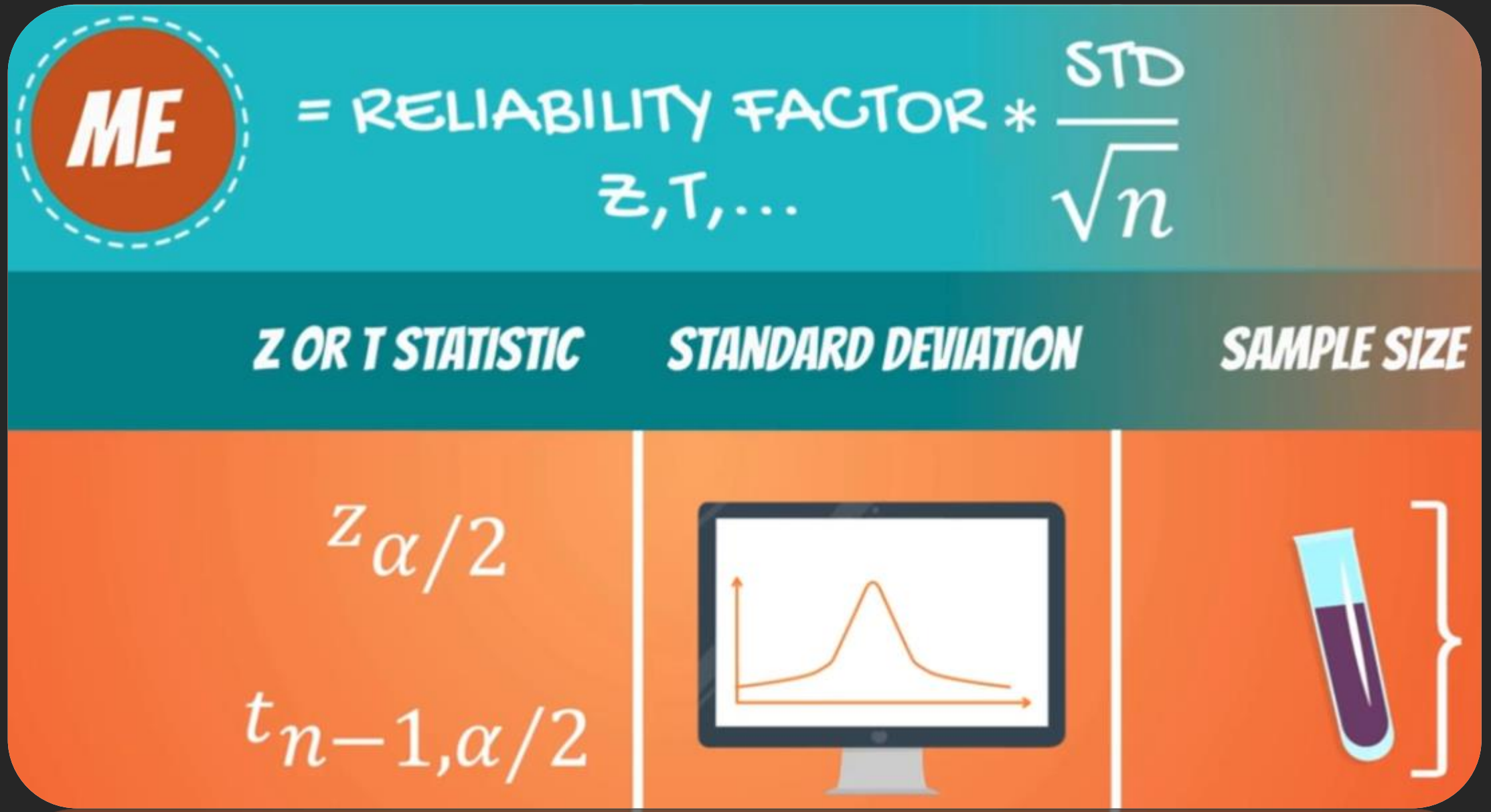
Known

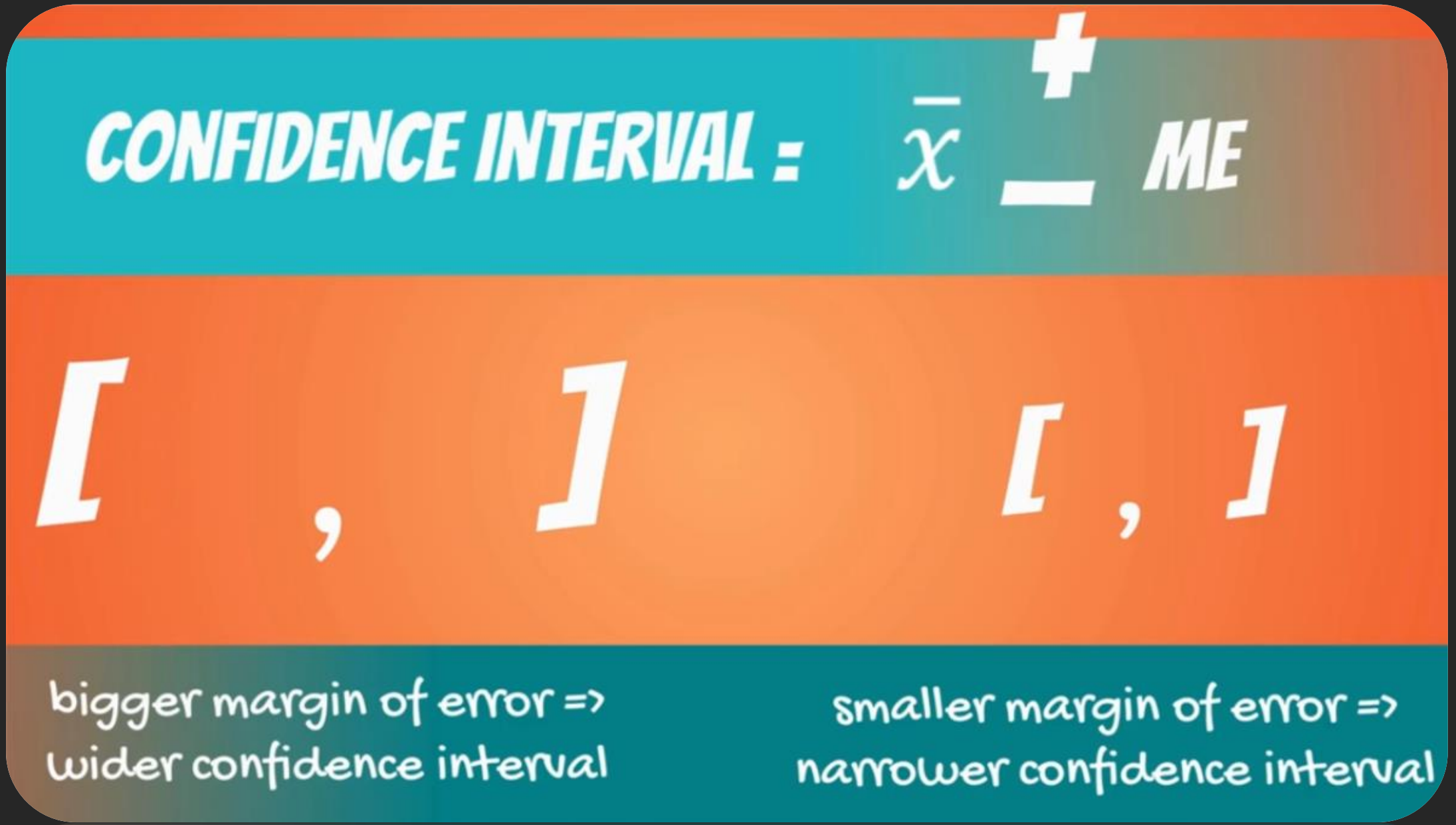
$$\bar{x} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$$

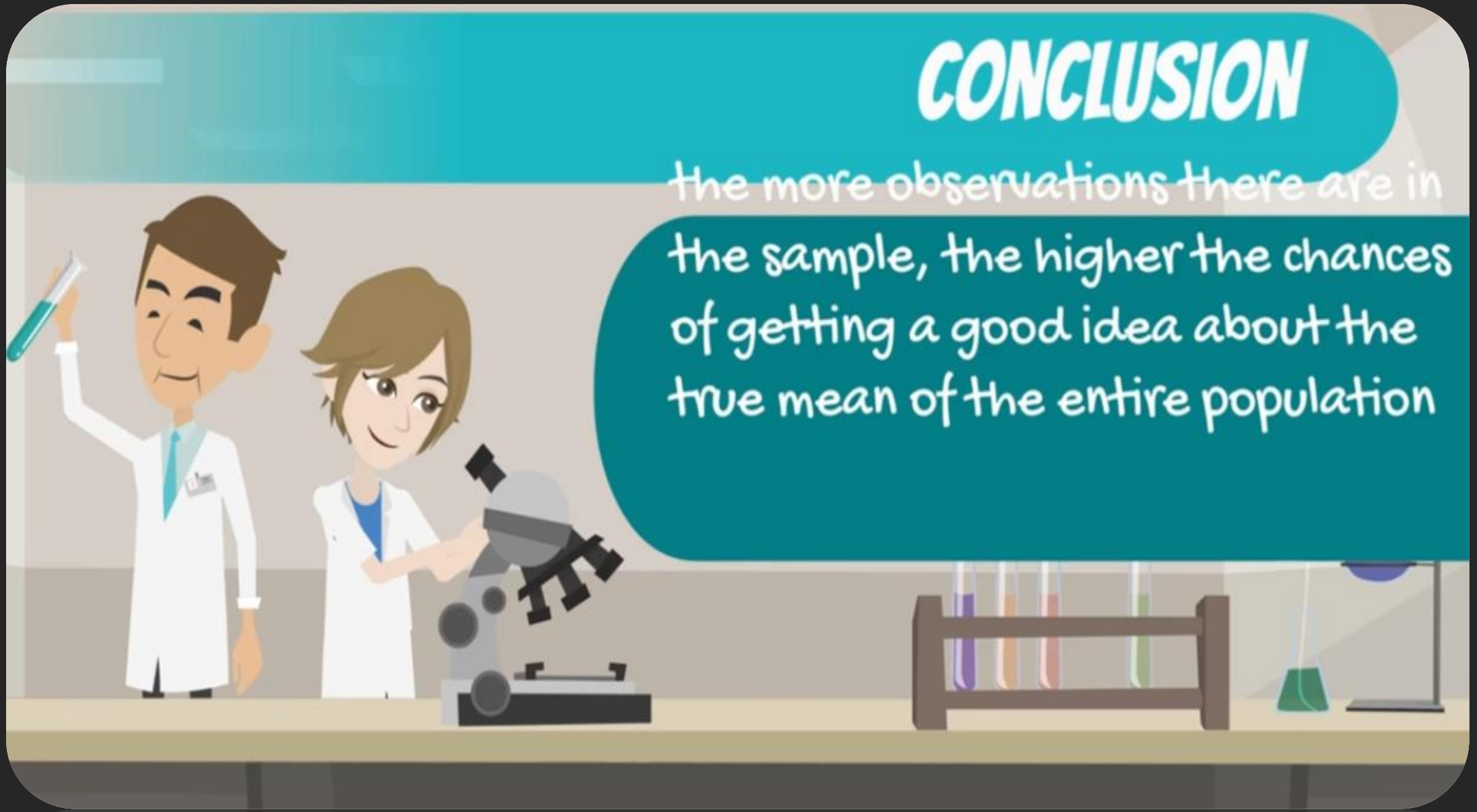
Unknown

MARGIN OF ERROR = ME

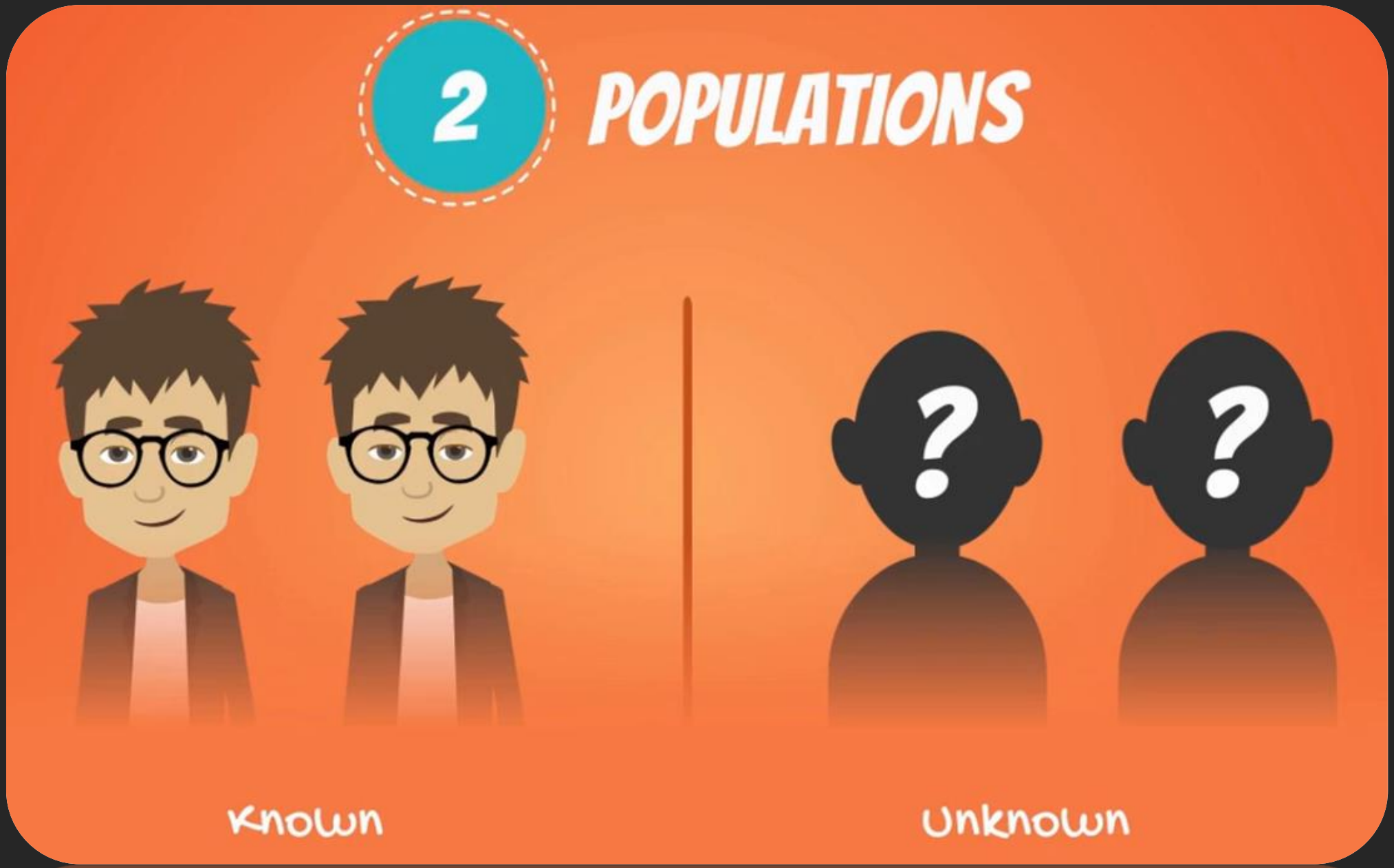


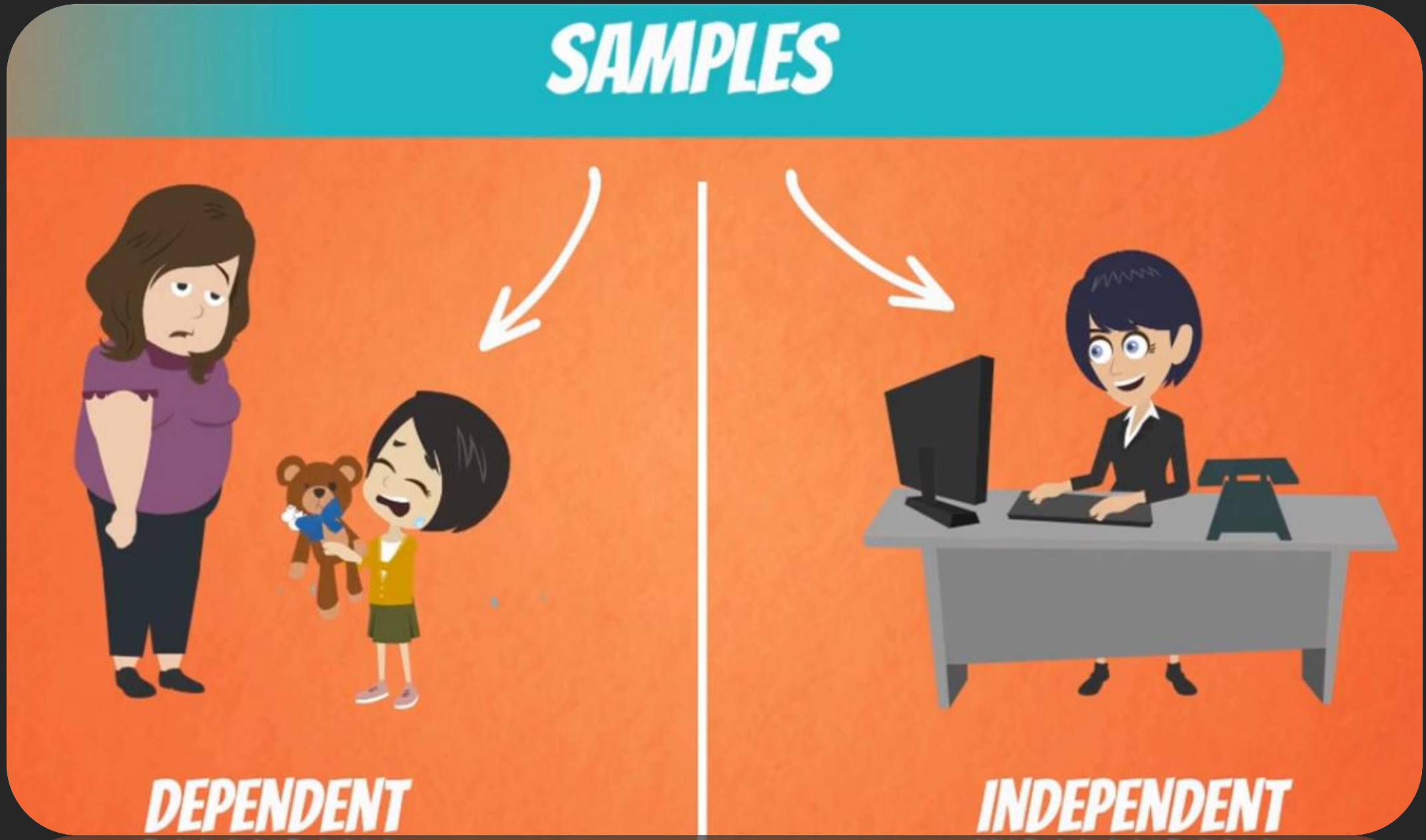

ME
= RELIABILITY FACTOR * STD / √n
z,T,...
Z OR T STATISTIC
STANDARD DEVIATION
SAMPLE SIZE
$z_{\alpha/2}$
$t_{n-1,\alpha/2}$


CONFIDENCE INTERVAL = $\bar{x} \pm ME$
[,]
[,]
bigger margin of error =>
wider confidence interval
smaller margin of error =>
narrower confidence interval


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

Toward the real world problems


2
POPULATIONS
Known
Unknown


SAMPLES
DEPENDENT
INDEPENDENT


WHEN WE ARE RESEARCHING THE SAME SUBJECT OVER TIME
weight loss
before and after
blood samples

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

DEPENDENT SAMPLES

- before and after situation

MAGNESIUM LEVELS

1.7 - 2.2 mg/dL

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

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 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
95% t-stat	2.26

Confidence interval for difference of two means, dependent samples formula

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

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

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

Confidence intervals. Independent samples:

1. Known population variances
2. Unknown population variances but assumed to be equal
3. Unknown population variances but assumed to be different

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

University example

	Engineering	Management
Size	100	70
Sample mean	59	65
Population std	10	5

Two Faculties in UK. Find the difference between grades

Considerations:

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

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

University example

	Engineering	Management
Size	100	70
Sample mean	58	65
Population std	10	5

Considerations:

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

The two samples are truly independent

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

University example

	Engineering	Management	Difference
Size	100	70	?
Sample mean	58	65	-7.00
Population std	10	5	

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

	Engineering	Management	Difference
Size	100	70	?
Sample mean	58	65	-7.00
Population std	10	5	

Considerations:

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

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

University example

	Engineering	Management	Difference
Size	100	70	?
Sample mean	58	65	-7.00
Population std	10	5	1.16

Variance of the difference

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

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

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

University example

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

$$(\bar{x} - \bar{y}) \pm z_{\alpha/2} \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} = (-9.28, -4.72)$$

95% confidence interval

difference point estimator

test statistic

standard error

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

University example

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

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

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

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 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	

iPhone “price difference” problem

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

We know that the population variance are equal, but unknown. So, we estimate it. Its estimator is called “Pooled Variance”.

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

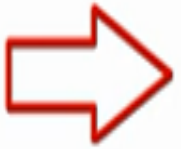
Apples example

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

1. Population variance unknown
2. Small samples



T

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	

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

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

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

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%

$$n_x + n_y - 2 = 10 + 8 - 2 = 16$$

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	
95% t-stat	2.12	

Takeaway:

Apples in NY are much more expensive than in LA

$$(\bar{x} - \bar{y}) \pm t_{n_x+n_y-2,\alpha/2} \sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}} = (3.94 - 3.25) \pm 2.12 \sqrt{\frac{0.05}{10} + \frac{0.05}{8}}$$

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

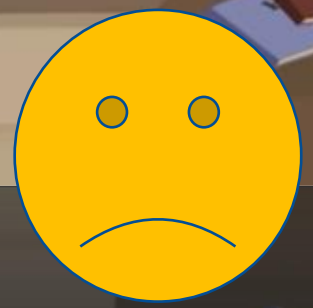
WE WANT TO FIND THE CONFIDENCE INTERVAL FOR TWO SAMPLE MEANS

Independent

Variance unknown

Assumed to be different

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



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

JUST take it easy and go forward…

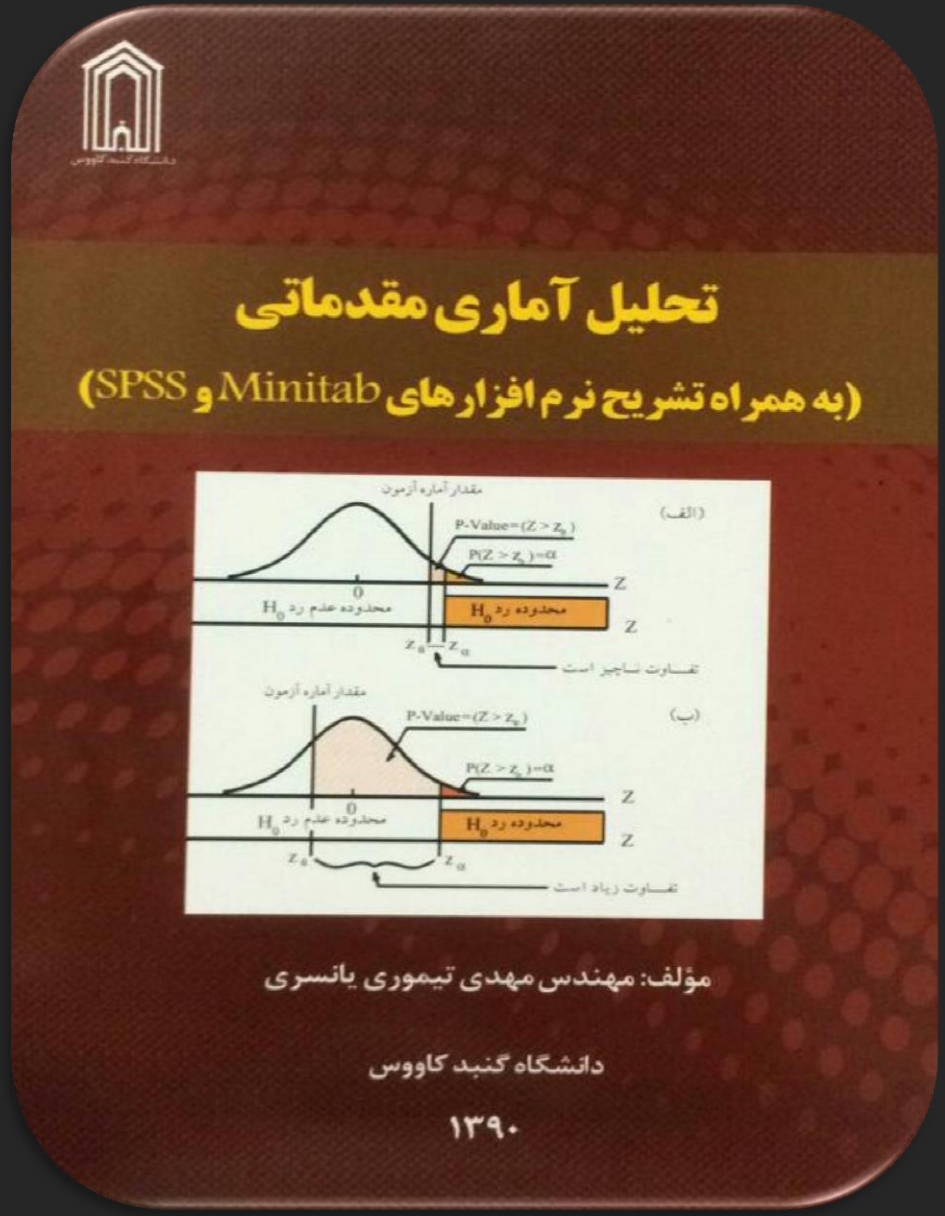
Times for review

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

NEXT LESSON: HYPOTHESIS TESTING

A good reference for studying more about these topics. Give it a try.



Thanks for watching
AMIGOS, see you soon

Also thanks to Data Science Course 2020 by Udemy and Data Science 365 team. Almost all the slides have been duplicated from this wonderful course.