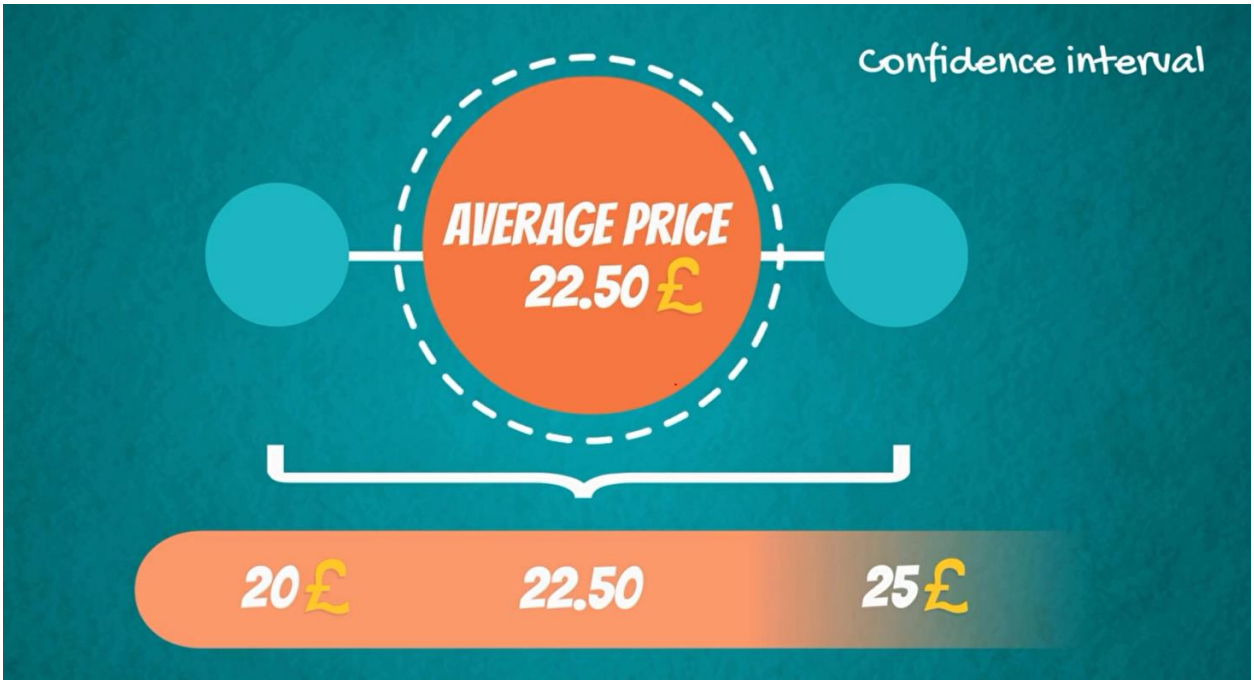
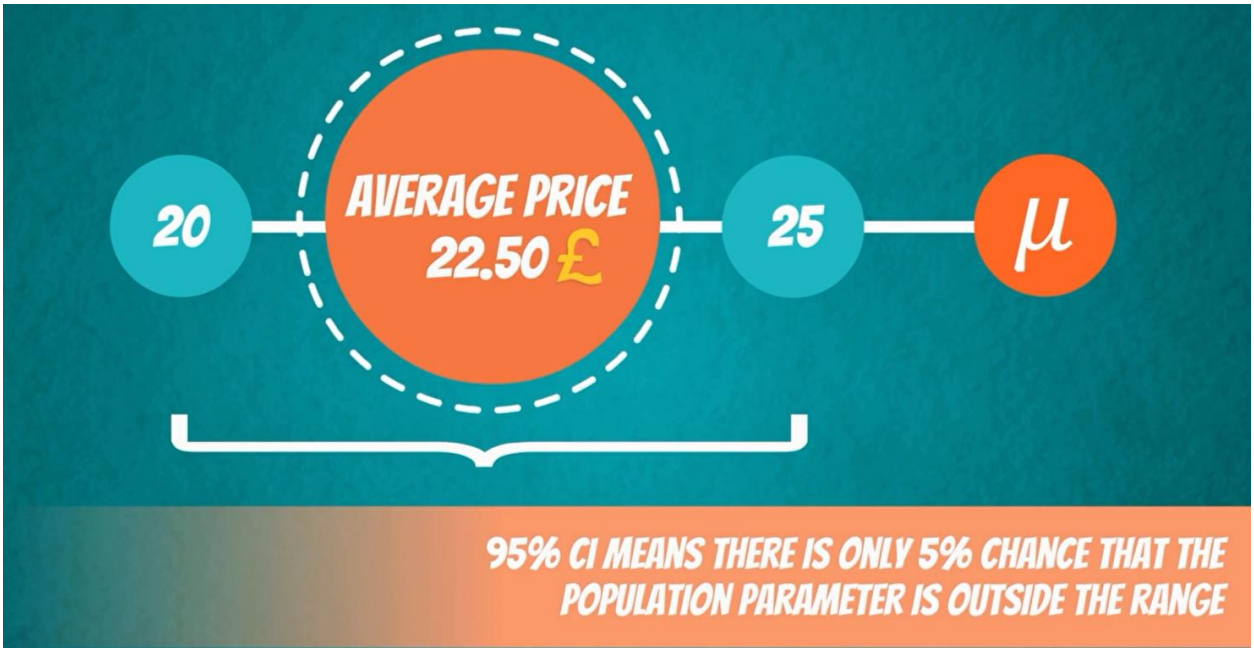
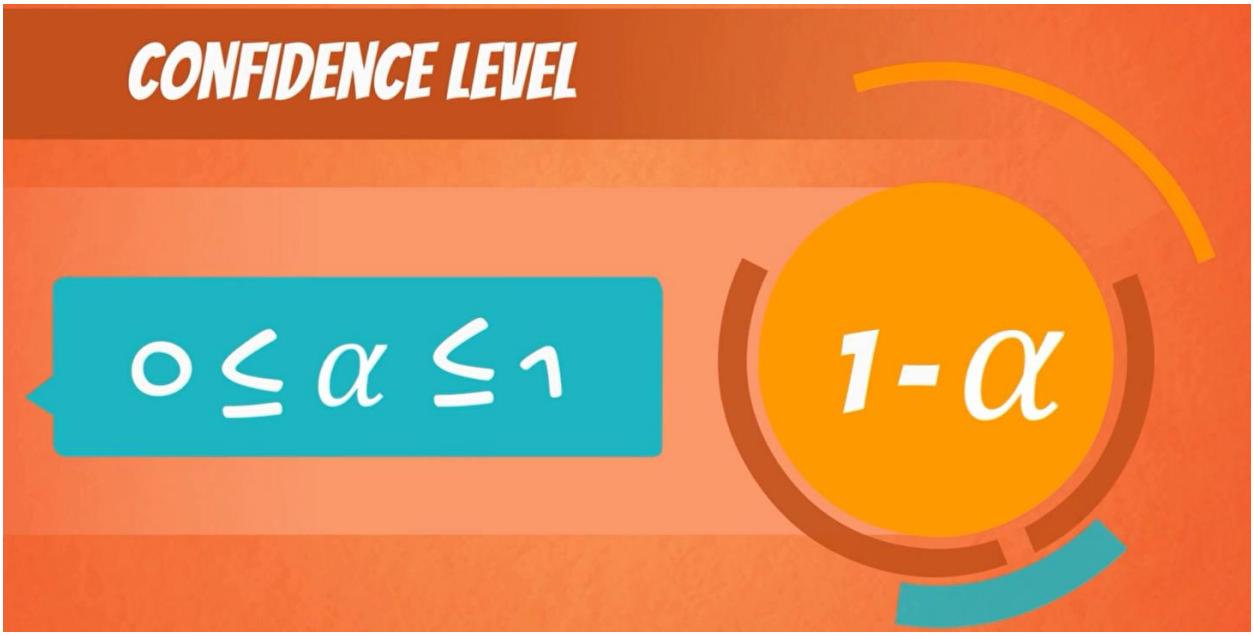
1. What are Confidence Intervals



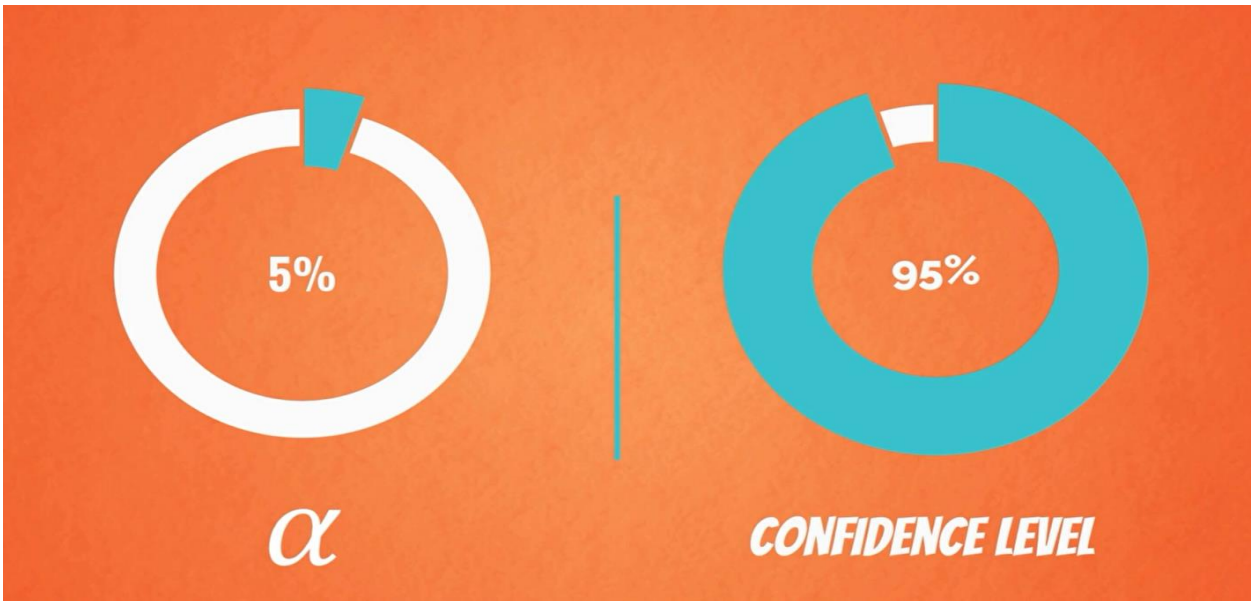
confidence interval is a much more accurate representation of reality

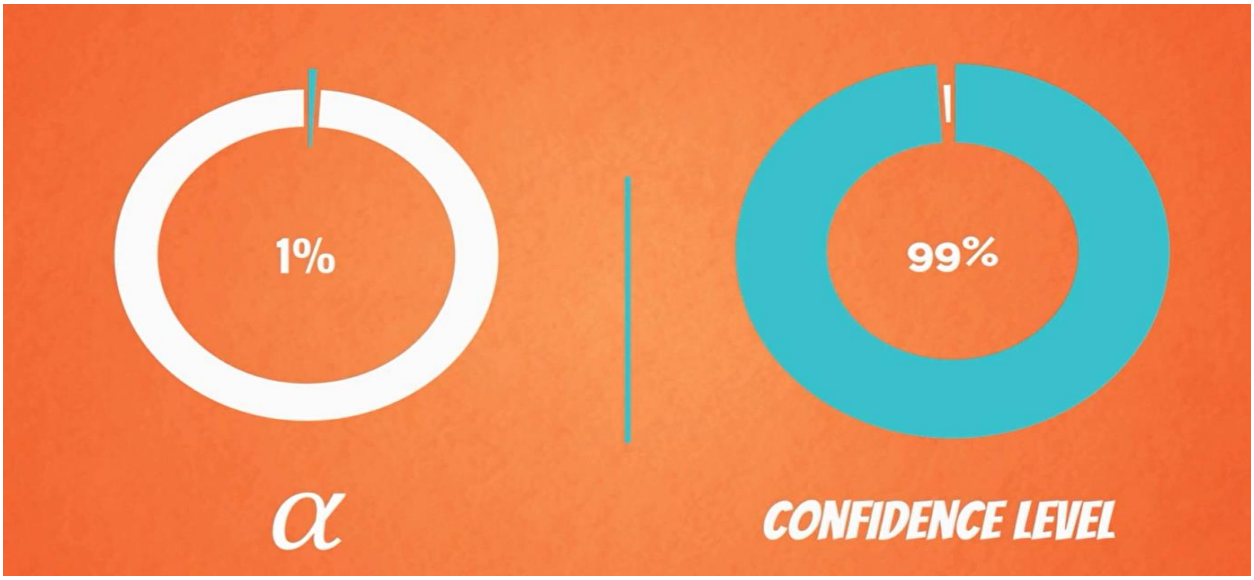


The level of confidence it is denoted by 1 minus Alpha and is called the confidence level of the interval. alpha is a value between 0 and 1

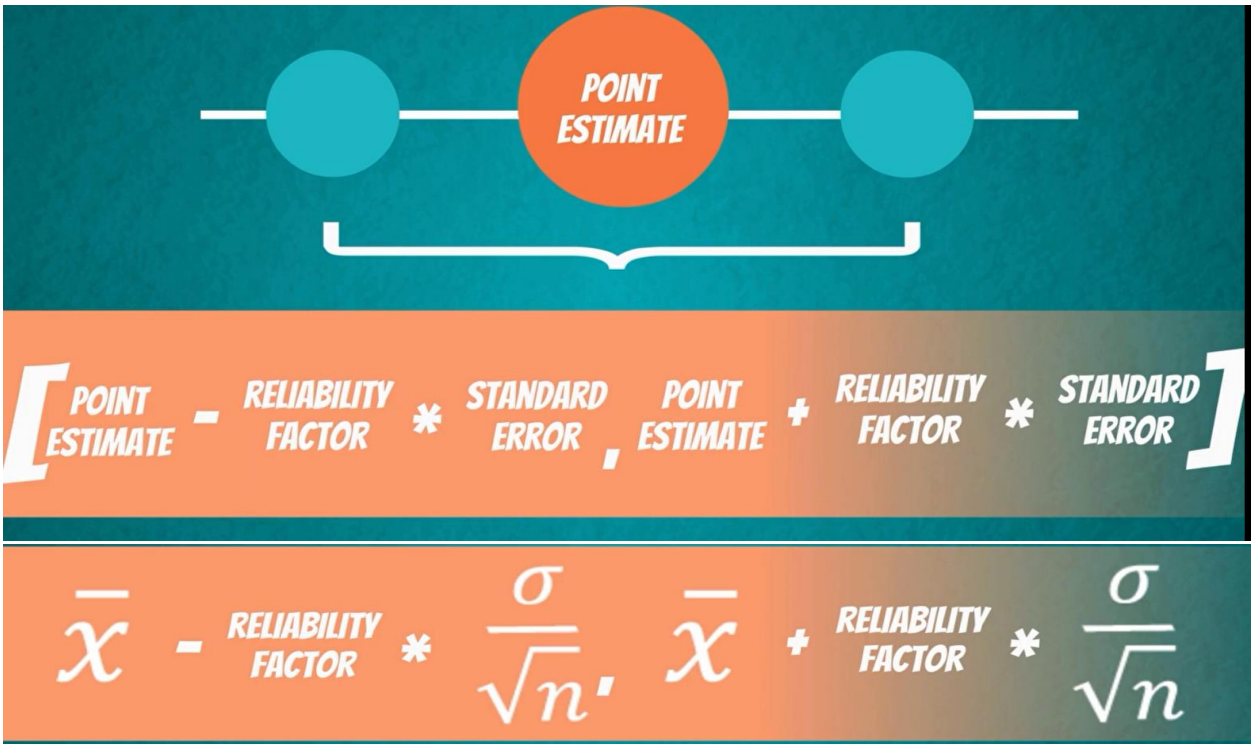


Example of confidence Interval

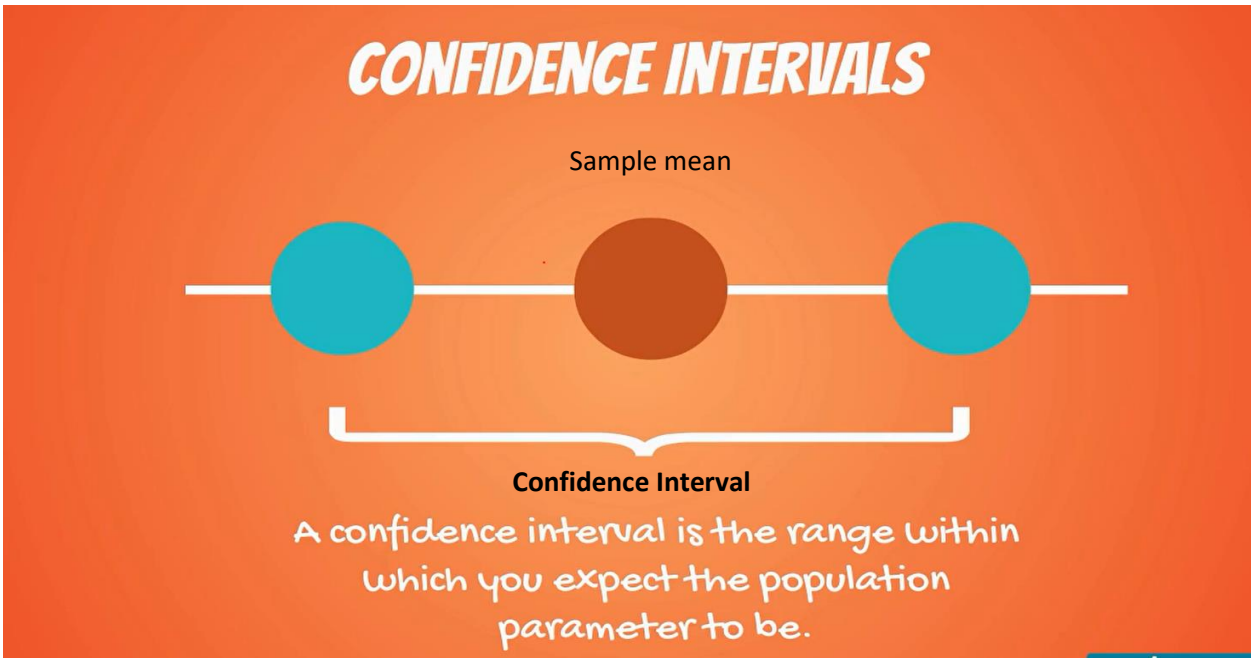




The formula for all confidence interval



2. Confidence Intervals; Population Variance Known; z-score



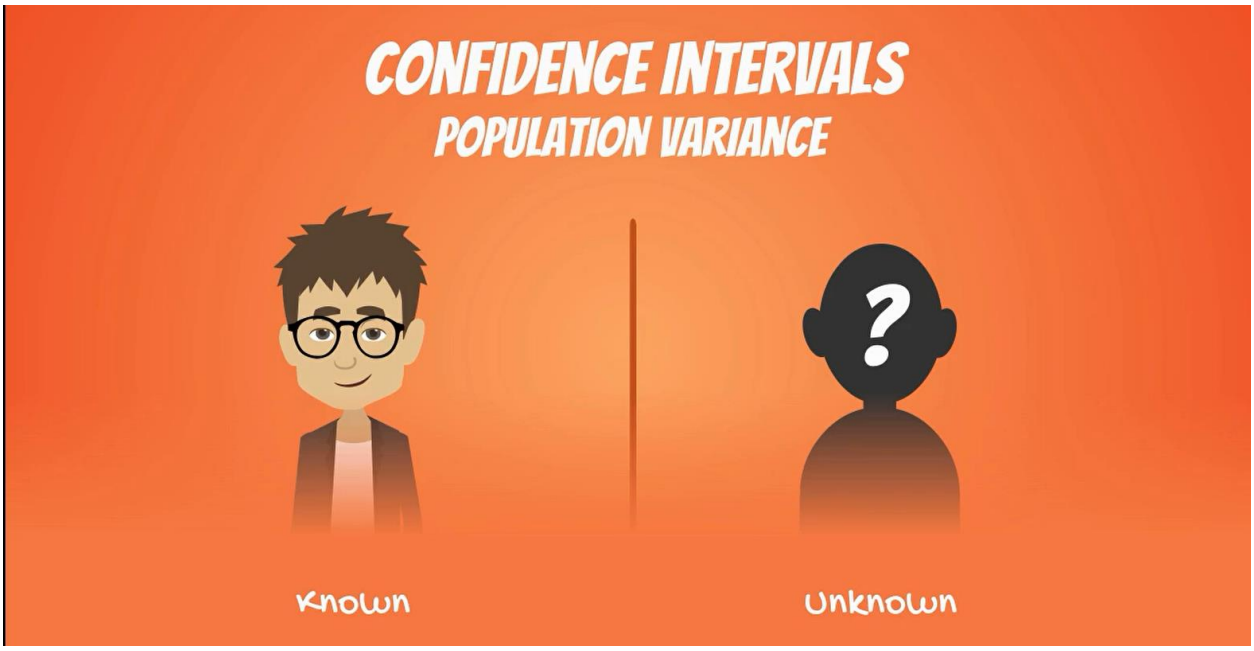
The sample mean is the point estimate.

Its estimation is based on the data we have in our sample

There can be two main situations when we calculate the confidence intervals for a population where

1. the population barrier is known and
2. when it is unknown

depending on which situation we are in. we would use a different calculation method.



Confidence intervals. Population known, z-score

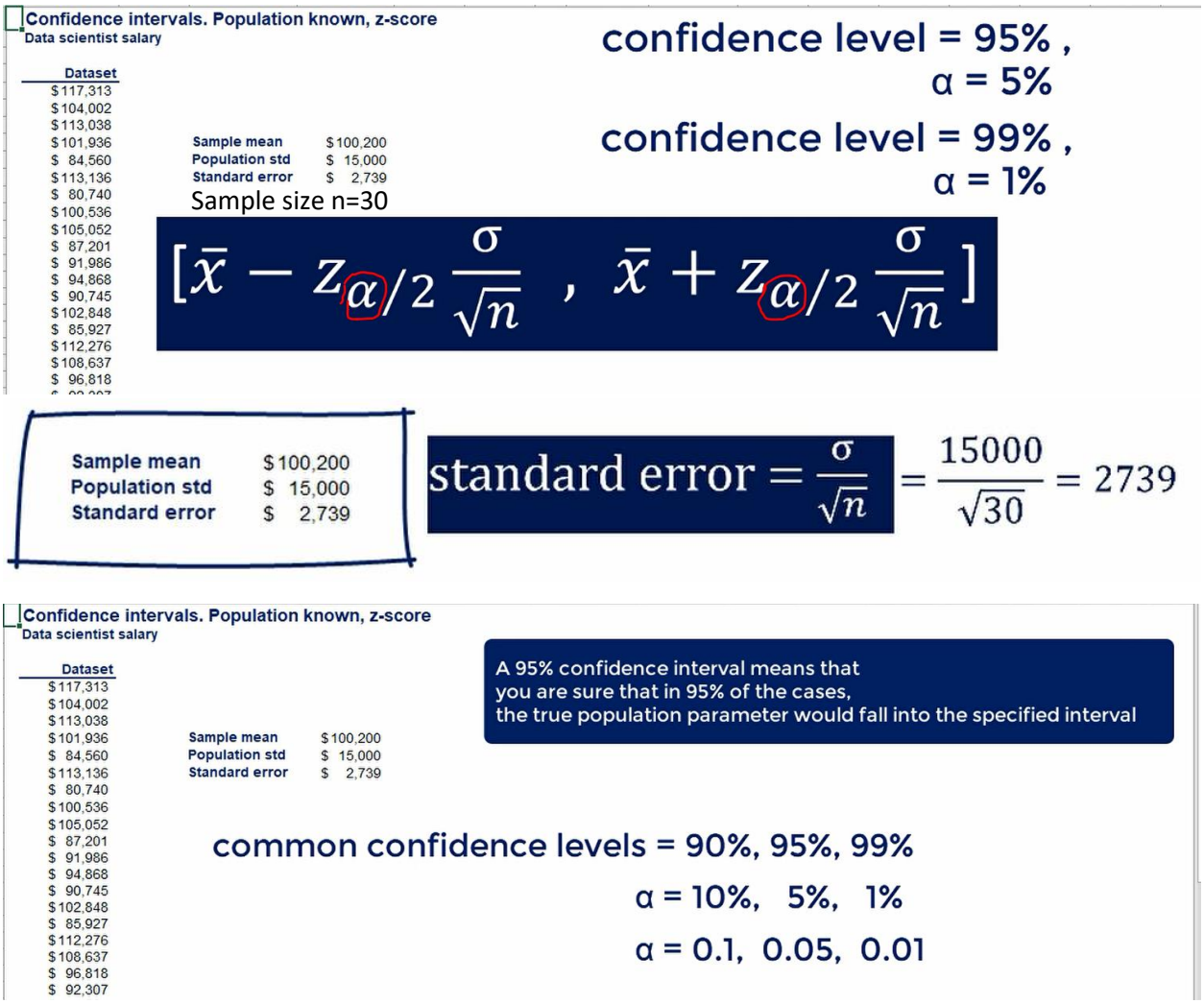
Data scientist salary

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

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

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



Z of Alpha

The z of Alpha comes from the so-called standard normal distribution table.

let's say that we want to find the values for the 95% confidence interval Alpha is zero 0.05.

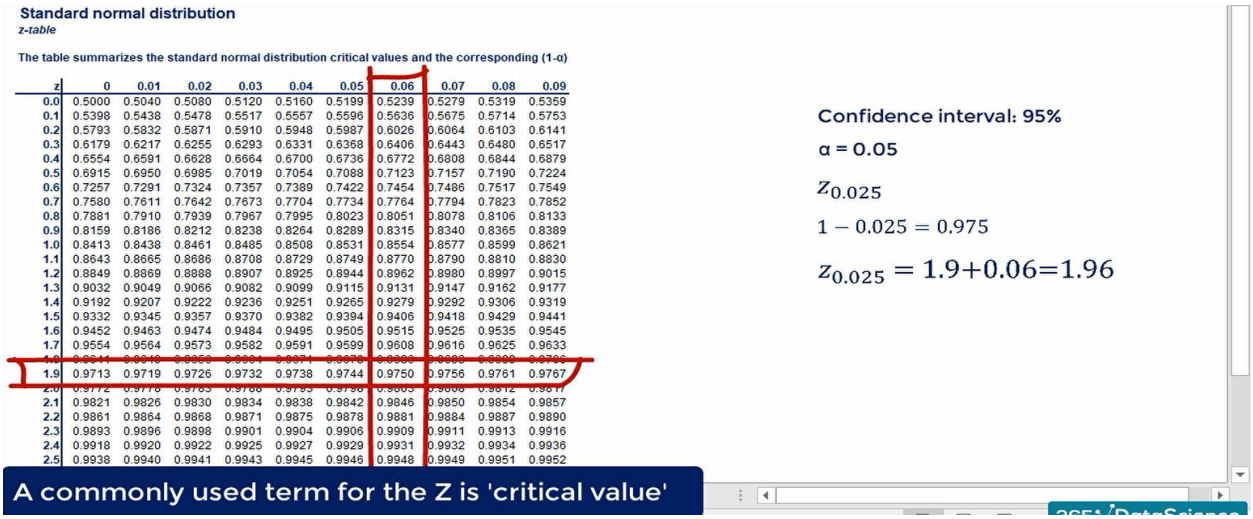
Standard normal distribution Z table

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
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990



z comes from the sum of the row and column table headers.

Confidence Interval

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

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

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

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

Figure: 95% Confidence 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
1.8	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

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

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

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

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

Figure: 99% Confidence Interval

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

Confidence intervals. Population known, z-score

Data scientist salary

Dataset
$ 117,313
$ 117,313
$ 117,313
$ 117,313
$ 117,313
$ 117,313
$ 117,313
$ 117,313
$ 117,313
$ 117,313
$ 117,313
$ 117,313
$ 117,313
$ 117,313
$ 117,313
$ 117,313
$ 117,313
$ 117,313
$ 117,313
$ 117,313
$ 117,313
$ 117,313
$ 117,313
$ 117,313
$ 117,313

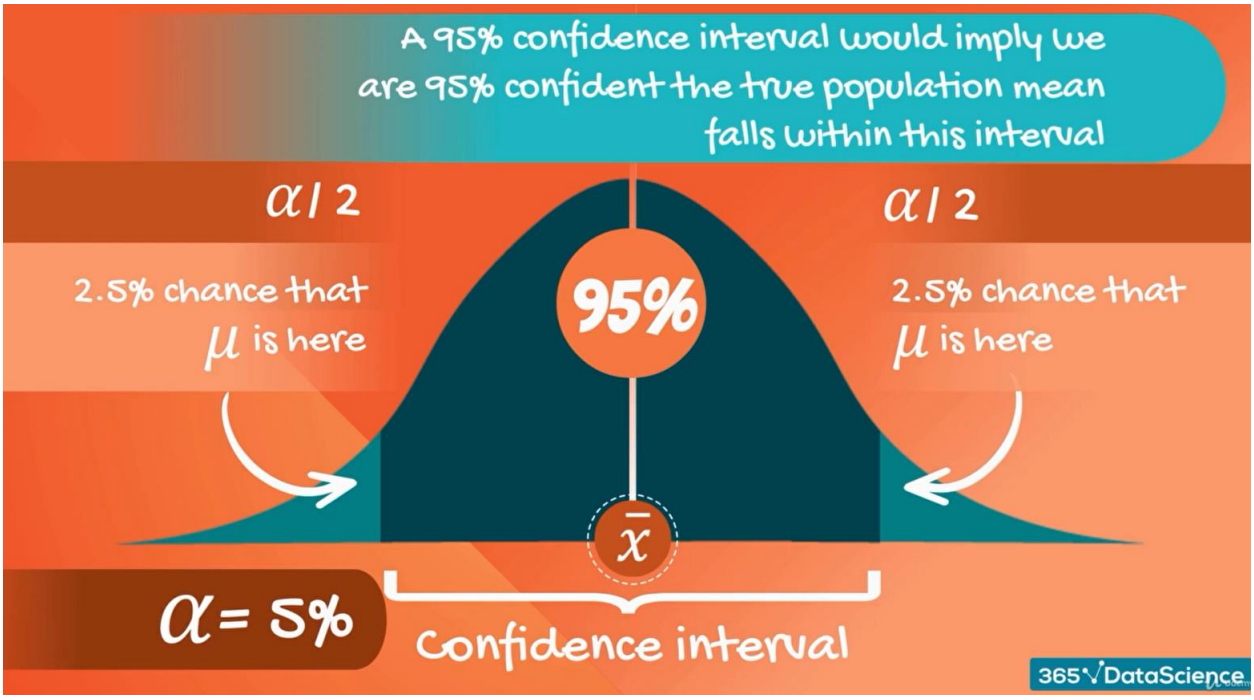
Sample mean	$ 117,313
Population std	$ 15,000
Standard error	$ 2,739
95% CI, $z_{0.025}$	1.96
99% CI, $z_{0.005}$	2.58

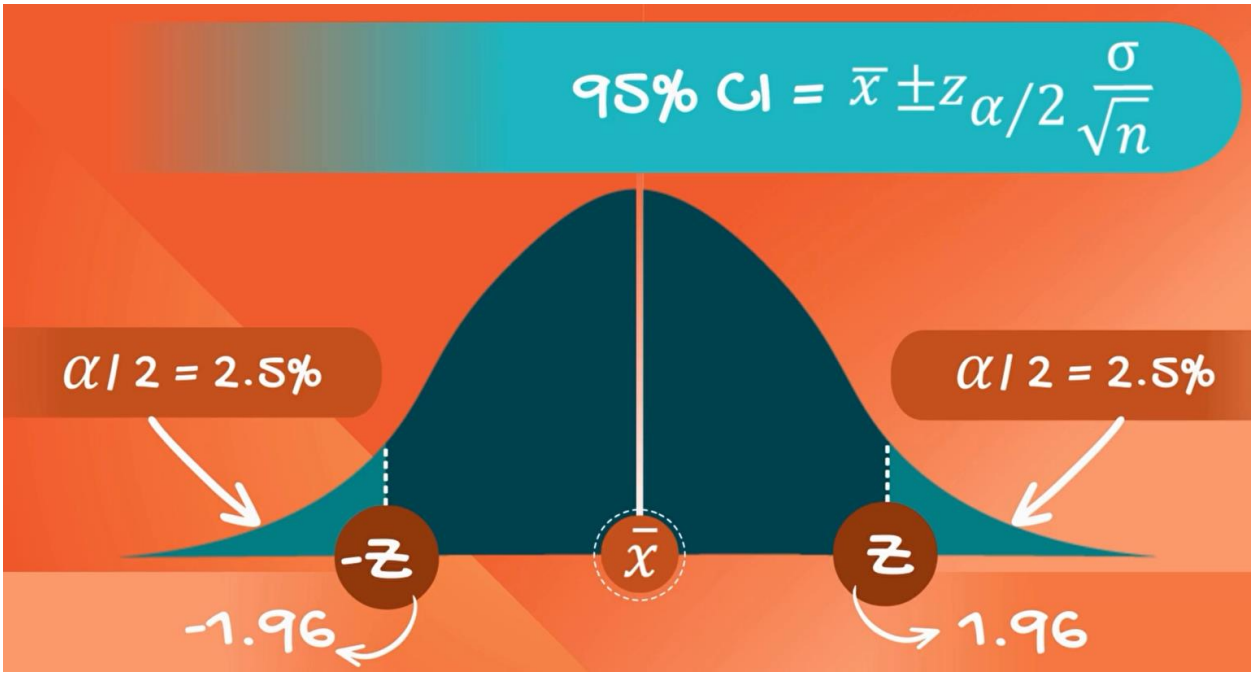
Confidence intervals

Z	CI low	CI high
###	######## #	$ 122,681
###	######## #	$ 124,379

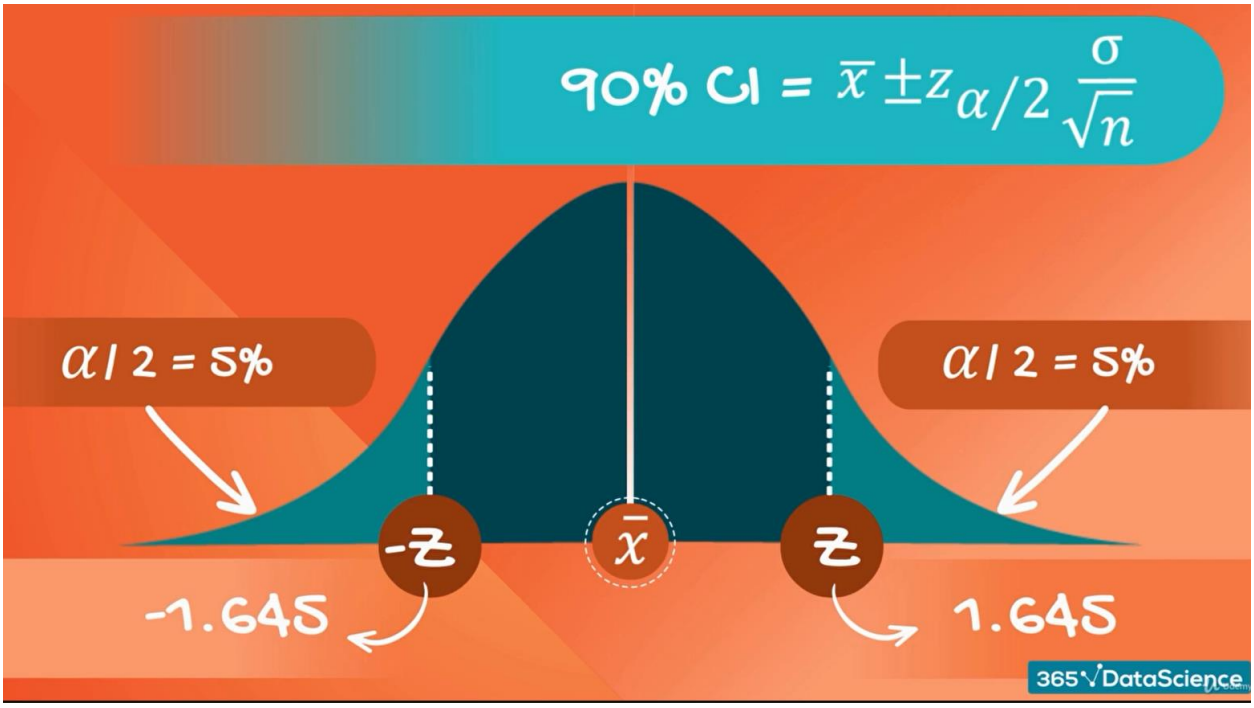
*CI low stands for lower bound of the confidence interval
CI high stands for higher bound of the confidence interval
The percentages indicate the confidence
Z stands for the fact that we are using the z-statistic
Note that this is not a common way to summarize the data. It is just a clear and useful one in Excel

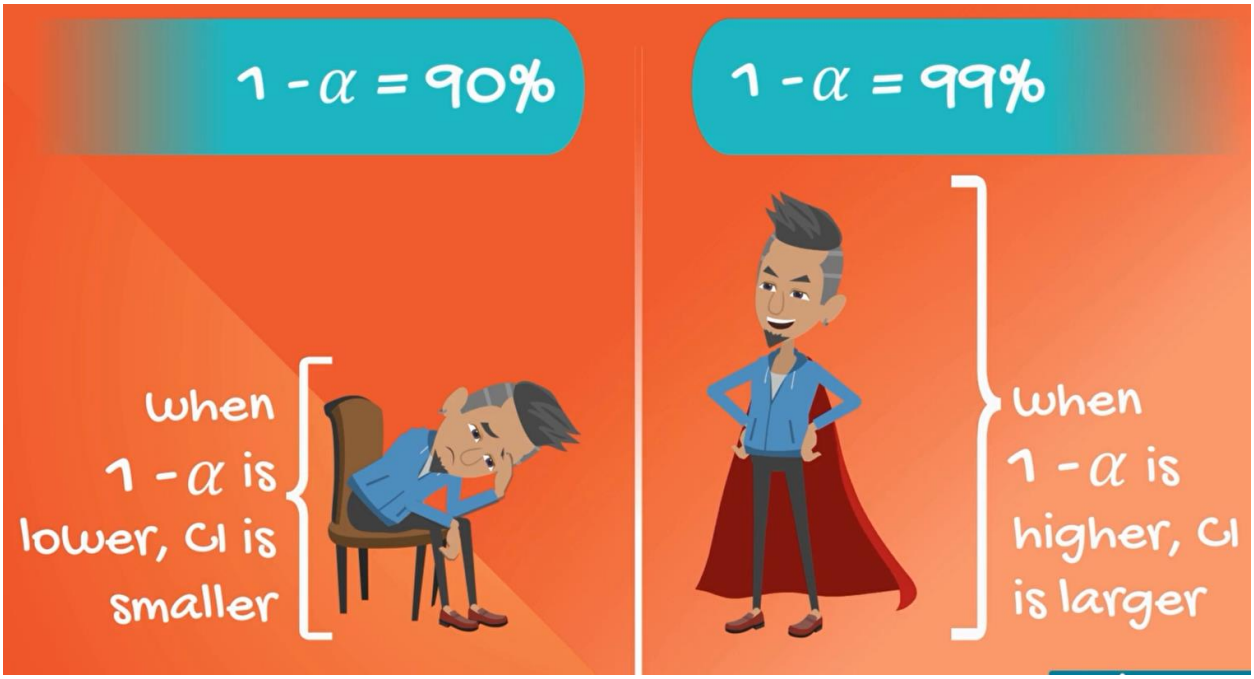
5. Confidence Interval Clarifications

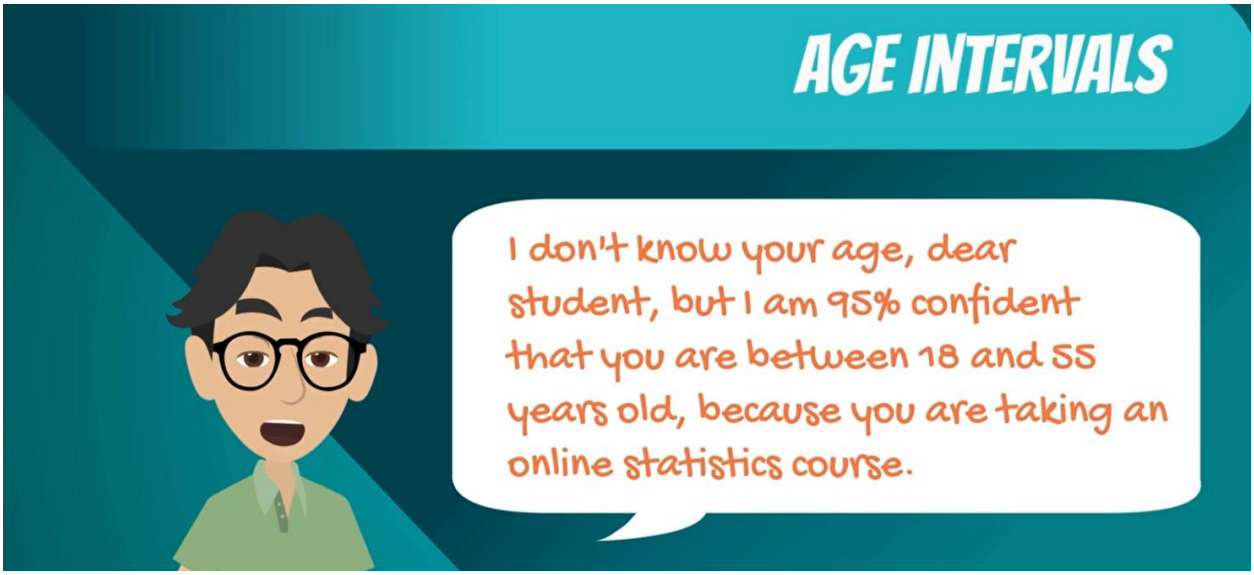


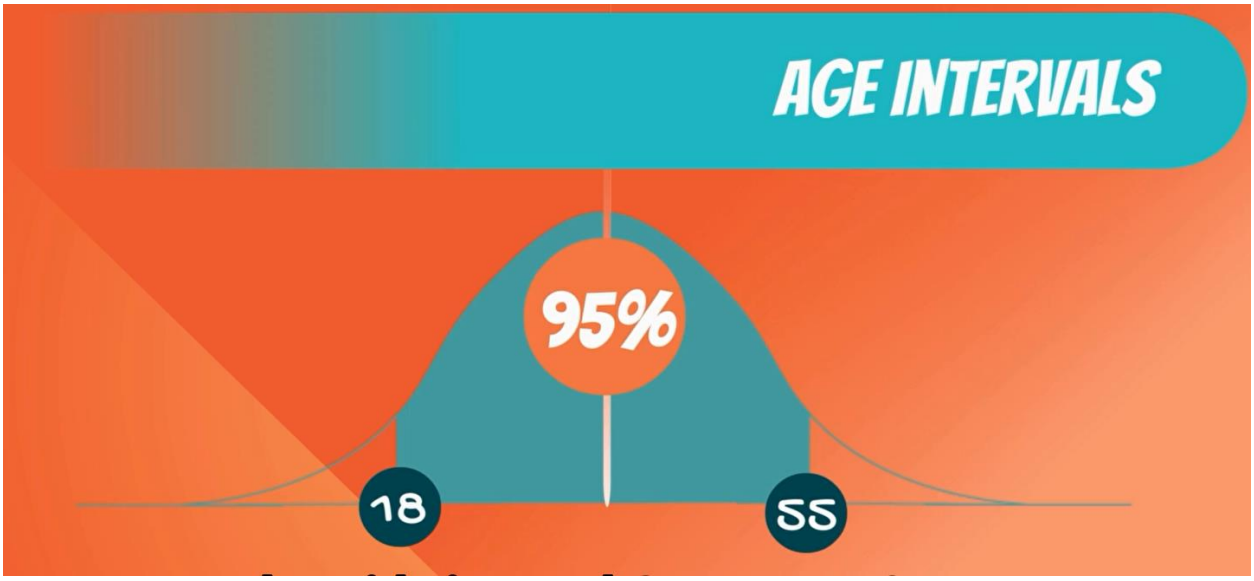


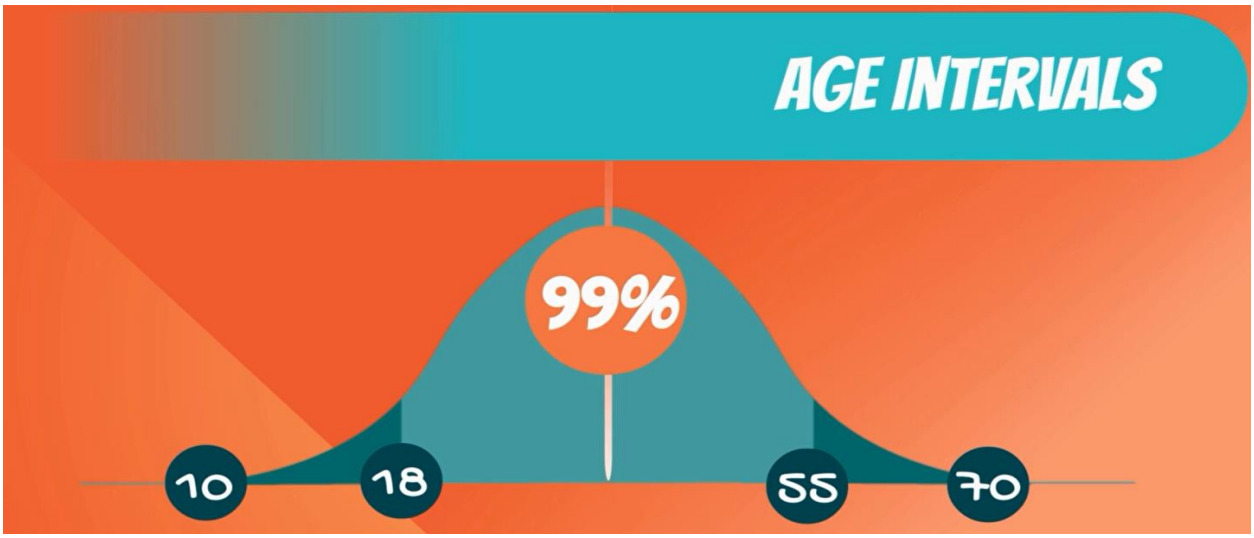
90% Confidence Interval

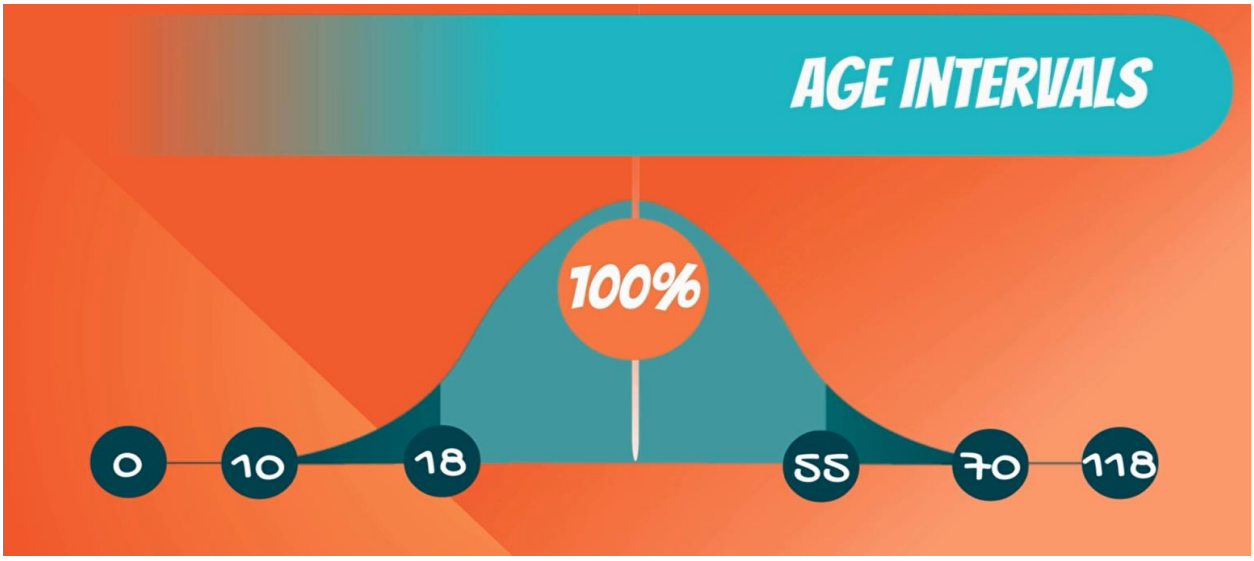


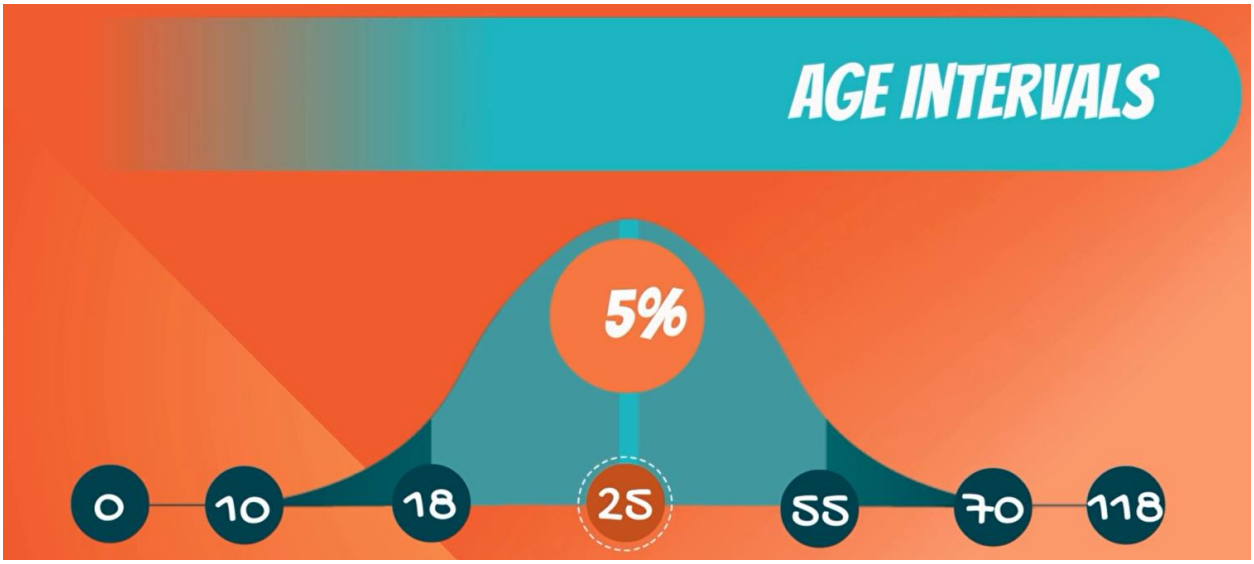


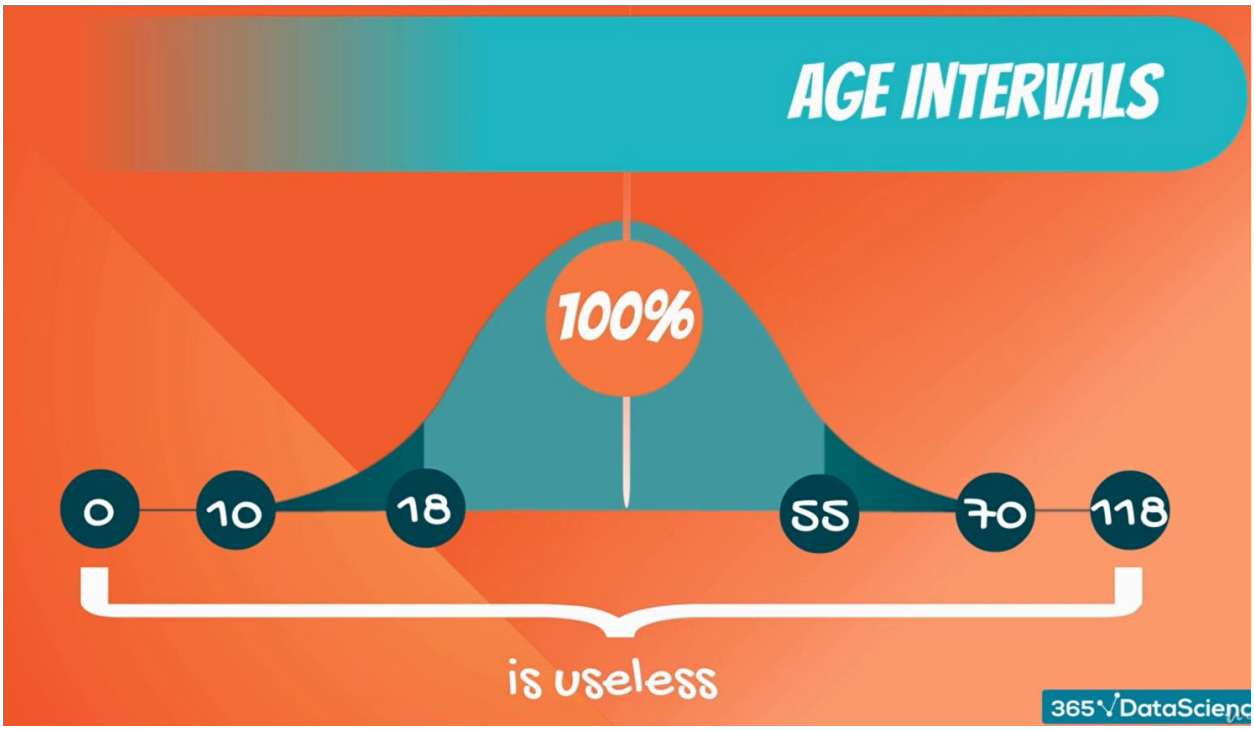

AGE INTERVALS
I don't know your age, dear student, but I am 95% confident that you are between 18 and 55 years old, because you are taking an online statistics course.

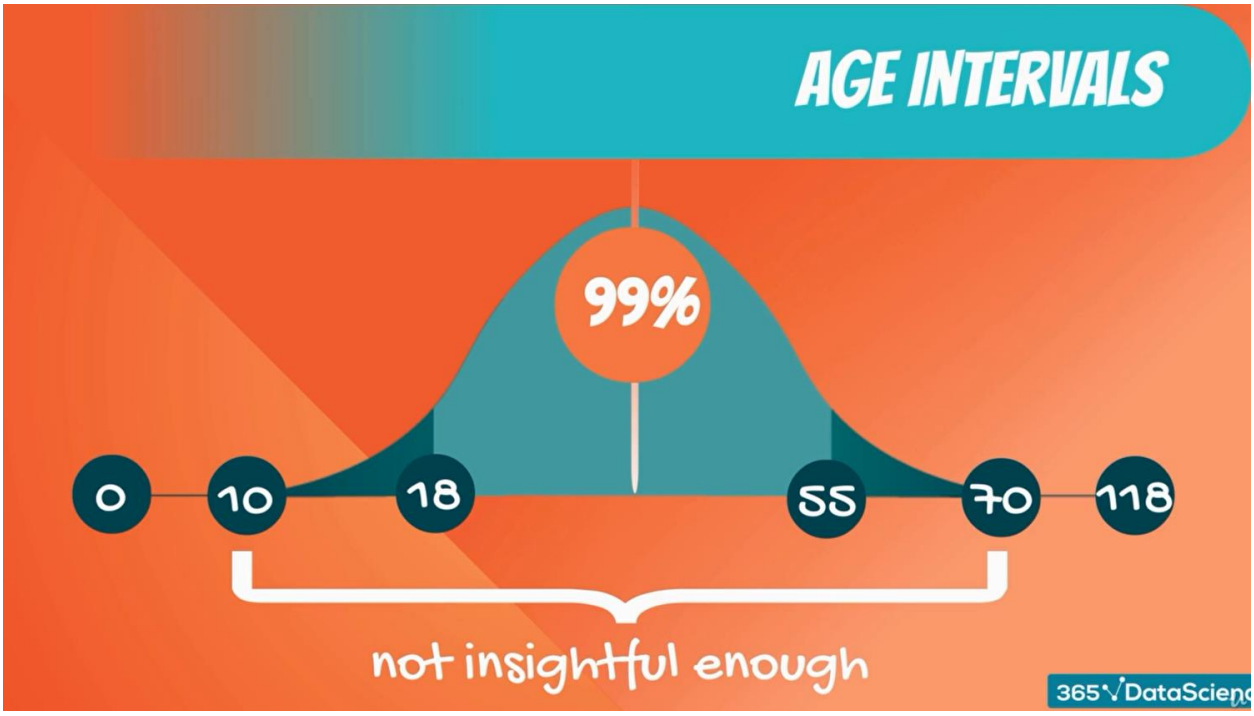

AGE INTERVALS
95%
18
55

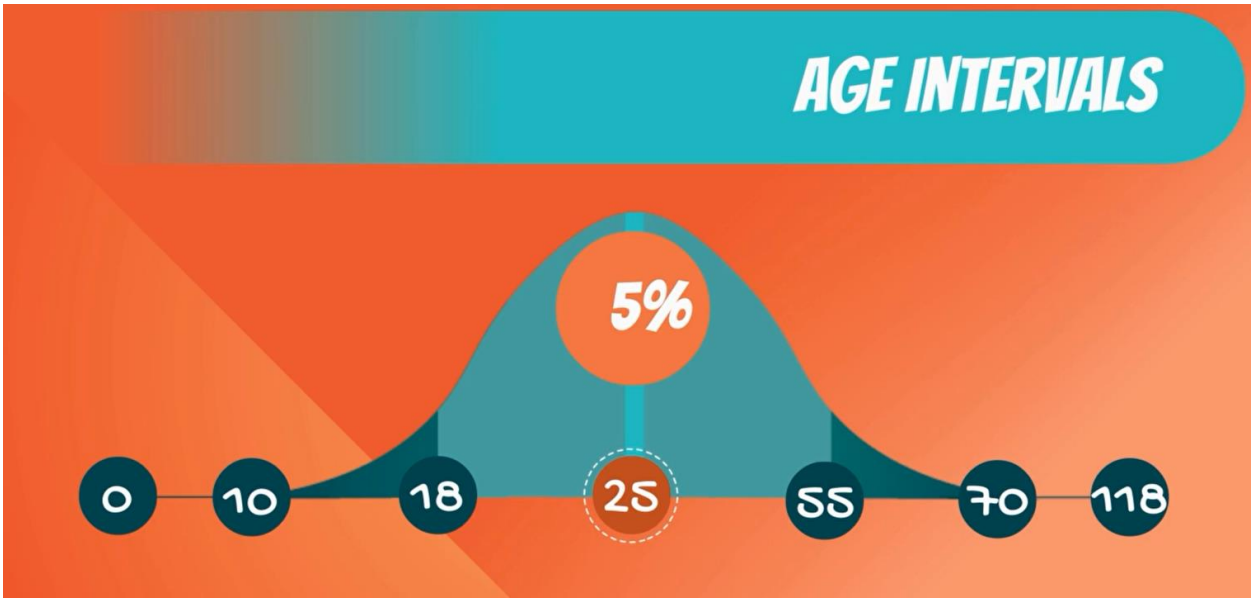

AGE INTERVALS
99%
10
18
55
70


AGE INTERVALS
100%
0
10
18
55
70
118

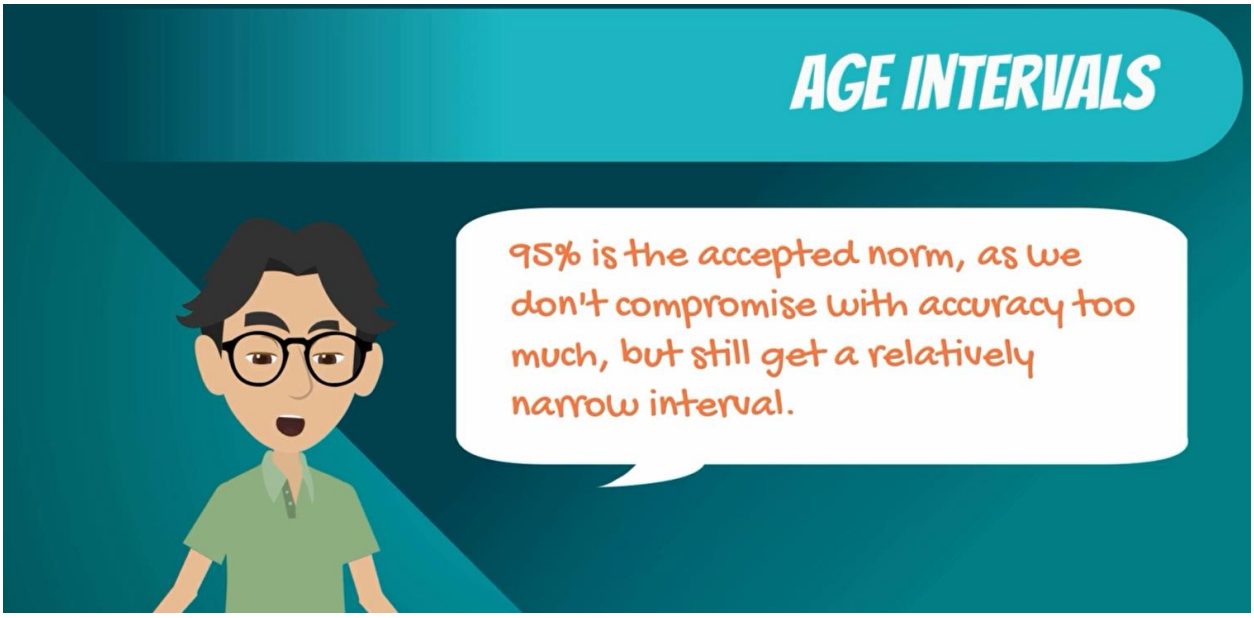

AGE INTERVALS
5%
0
10
18
25
55
70
118


AGE INTERVALS
100%
0
10
18
55
70
118
is useless
365√DataScienc

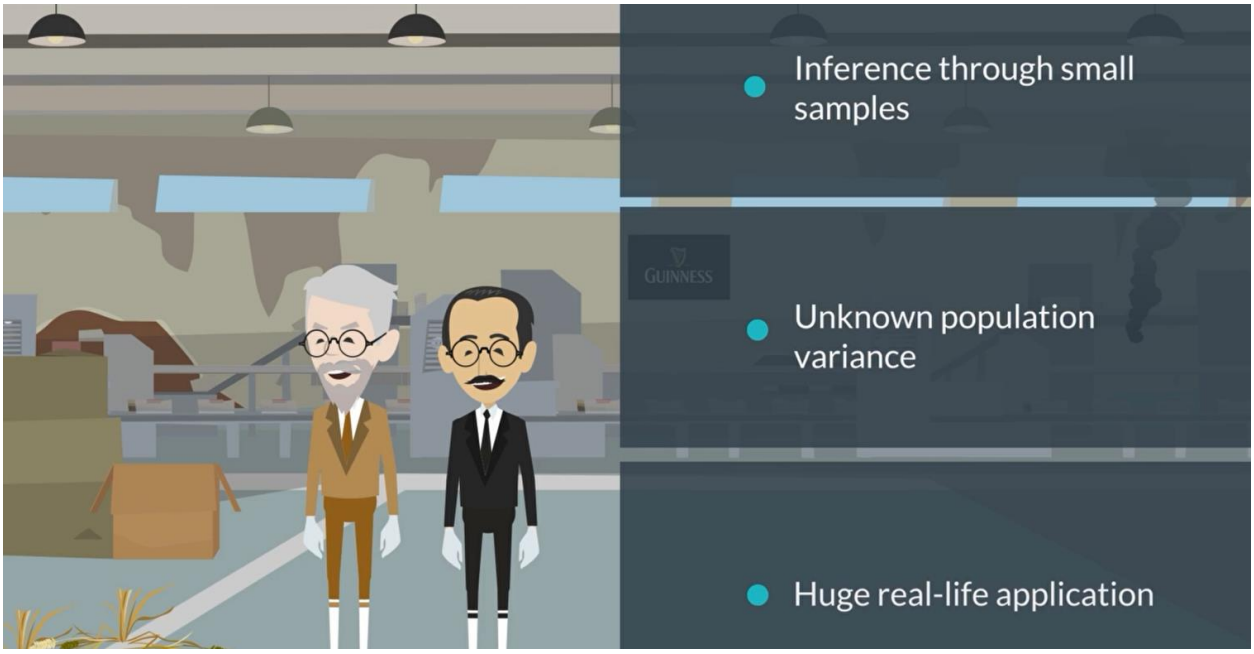

AGE INTERVALS
99%
0
10
18
55
70
118
not insightful enough
365√DataScienc



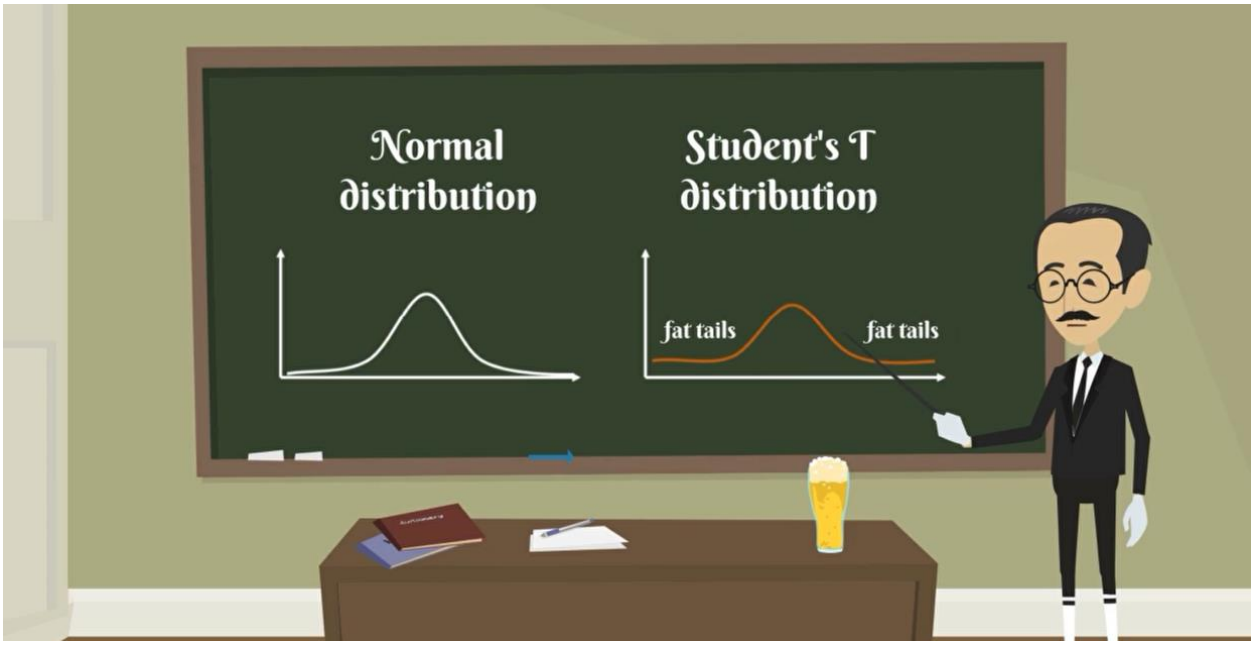
25 years old on the other hand is a pretty useful estimate as we have an exact number but the level of confidence of 5% is too small for us to make use of in any meaningful analysis.

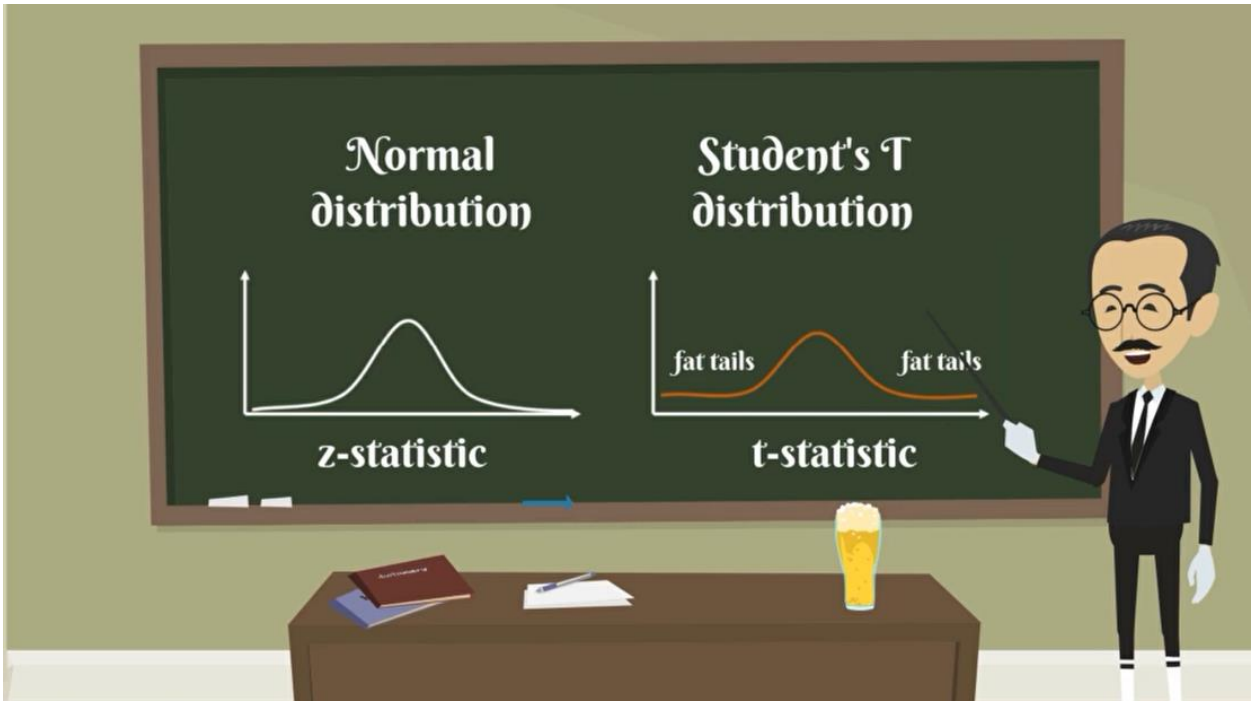


6. Student's T Distribution



It allowed inferences through small samples with an unknown population variance this setting can be applied big part of the statistical problems.

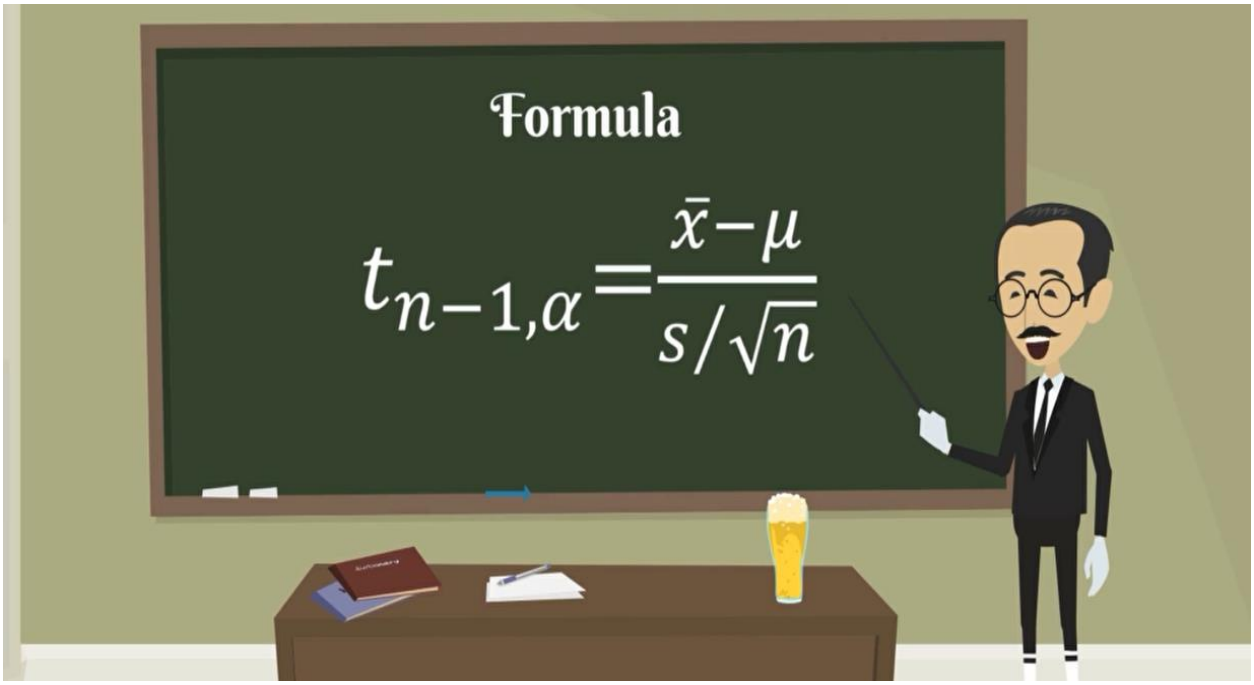




visually the student's t distribution looks much like a normal distribution but generally has fatter tails, fatter tails as you may remember allows for a higher Dispersion of variables and there is more uncertainty.

Z statistic is related to the standard normal distribution.

The t statistic is related to the student's t distribution.



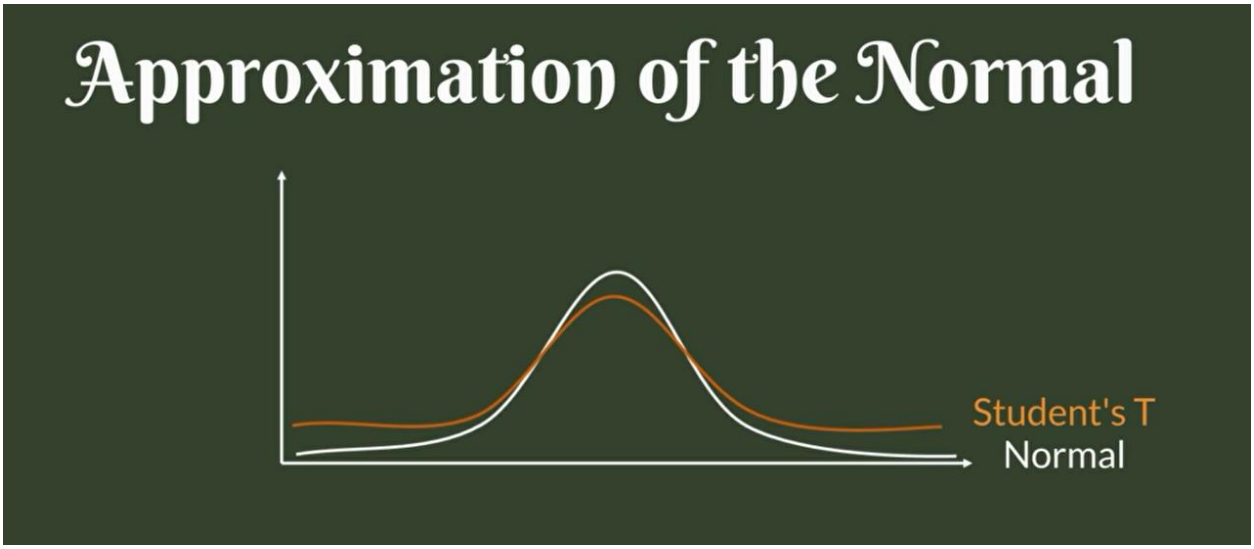
Degrees of freedom= t_{n-1} (t with n-1)

Sample mean = $\bar{x}$

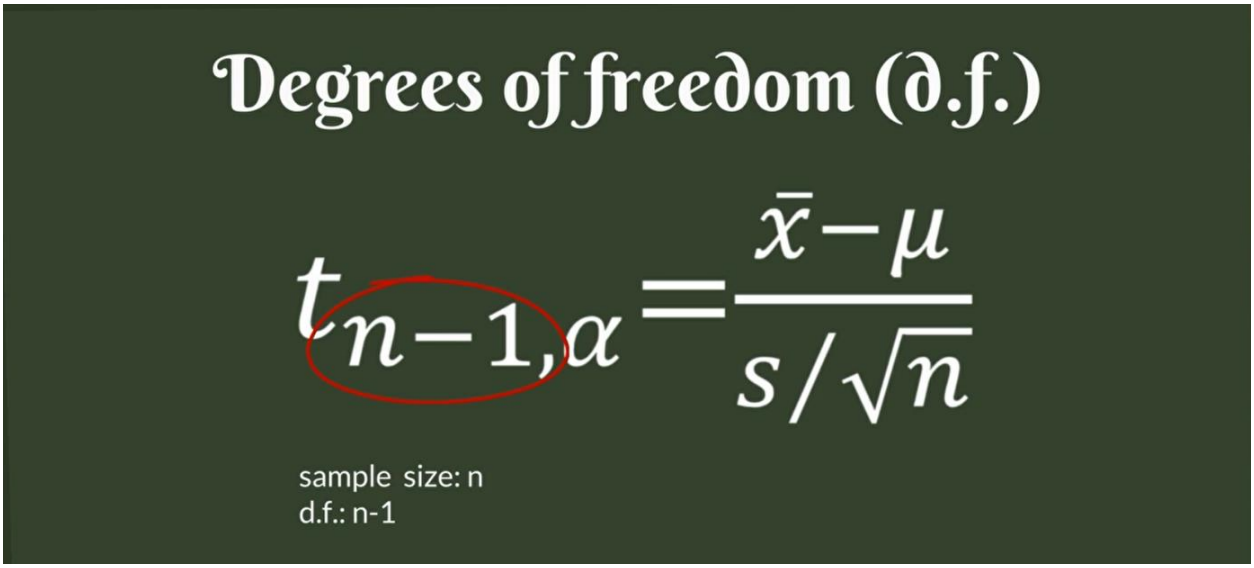
Population mean= μ (mu)

Significant of alpha = α

Standard error= $\sigma/\sqrt{n}$ (σ/s standard deviation)

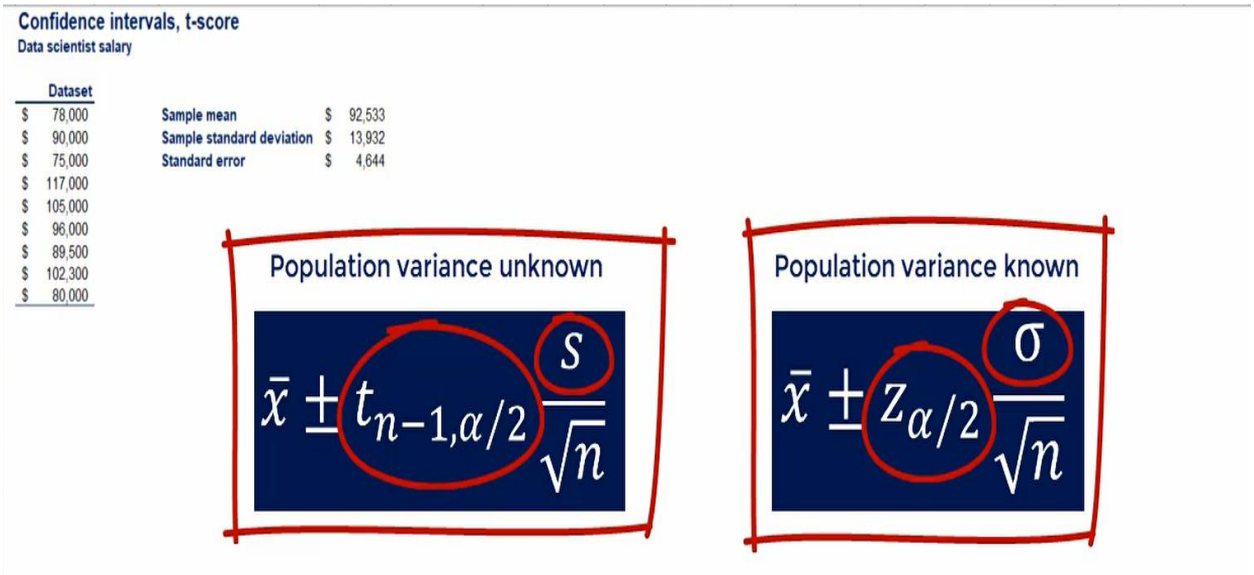


Tt is very similar to the Z, after all this is an approximation of the normal distribution.



The last characteristic of a student's t statistic is that there are degrees of freedom, usually for a Sample of N, we have n -1 of degrees of freedom.

8. Confidence Intervals; Population Variance Unknown; t-score



There are two key differences.

1. first instead of a z statistic we have a t statistic and
2. second instead of population standard deviation we have sample standard deviation.

otherwise, everything is the same. so, it should not be that difficult.

when population variance is unknown sample standard deviation goes with the t statistic.

when population variance is known population standard deviation goes with the z statistic.

Confidence Interval

in this example we are going to use a confidence interval of 95%, this means that Alpha is equal to 5%. therefore, half of Alpha would be 2.5%.

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 statistic

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

instead of finding Alpha we can just check the 95% confidence interval and get the same result.

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

365 √DataScience

When population variance is unknown => t-statistic

Confidence interval, t score

Confidence intervals, t-score

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

Mean	$	92,533
St. deviation	$	13,932
Standard error	$	4,644
95% CI, $t_{8,0.025}$		2.31

Confidence interval

T	CI low	CI high
95%	$81,806	$103,261

*CI low stands for lower bound of the confidence interval
CI high stands for higher bound of the confidence interval
The percentages indicate the confidence
T stands for the fact that we are using the t-statistic
Note that this is not a common way to summarize the data. It is just a clear and useful one in Excel

t - table

Student's T distribution
***t* –table**

The table summarizes the t–distribution critical values. The rows represent the degrees of freedom, while the columns – common alphas.

d.f. *l*	0.1	0.1	0.03	0.01	0.01
1	####	6.314	12.706	31.821	#####
2	1.886	####	4.303	6.965	9.925
3	1.638	####	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%

*CI stands for confidence intervals

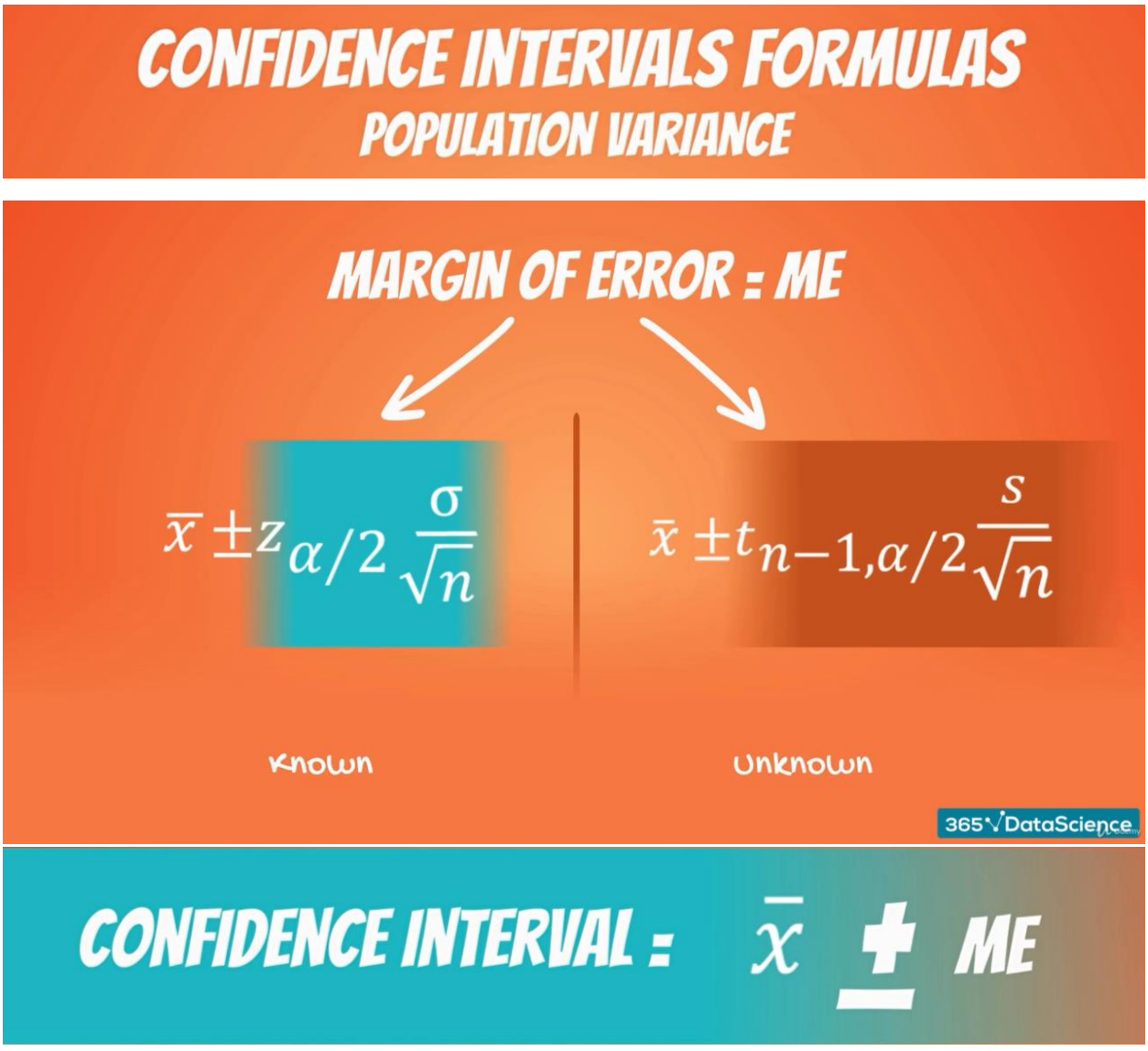
when we know the population variance, we get a narrow confidence interval.

when we do not know the population variance there is a higher uncertainty that is reflected by wider boundaries for our Interval.

We learned today

when we do not know the population variance, we can still make predictions but there would be less accurate

10. Margin of Error

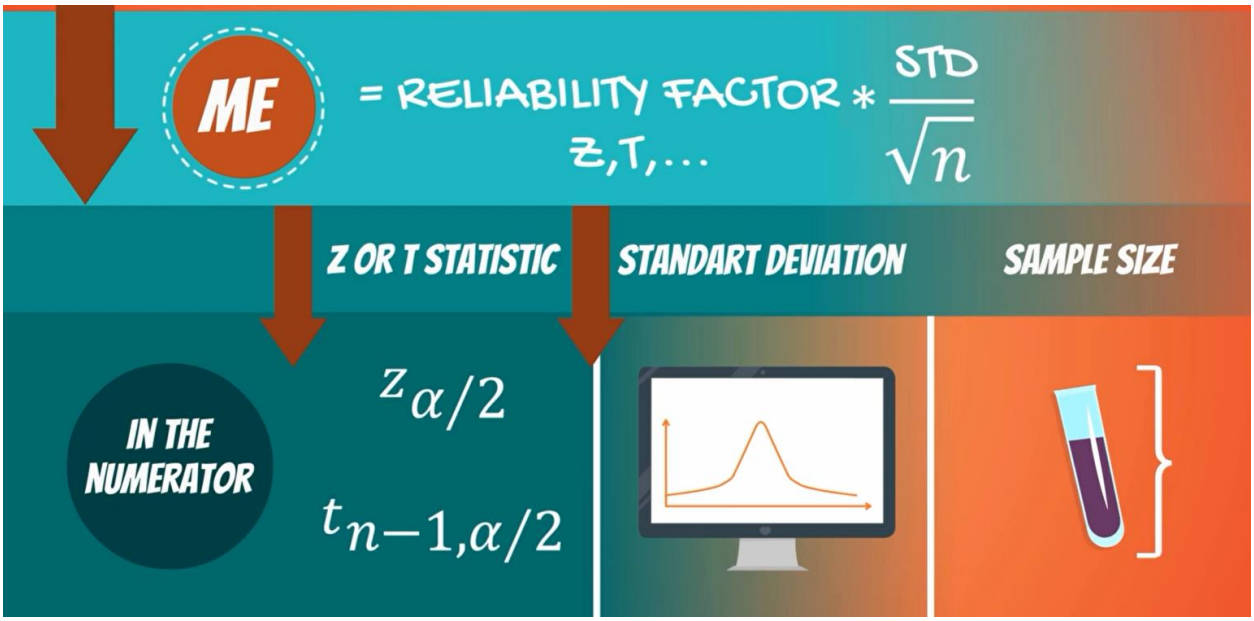


the margin of Error is equal to this expression and in the case of population variance unknown

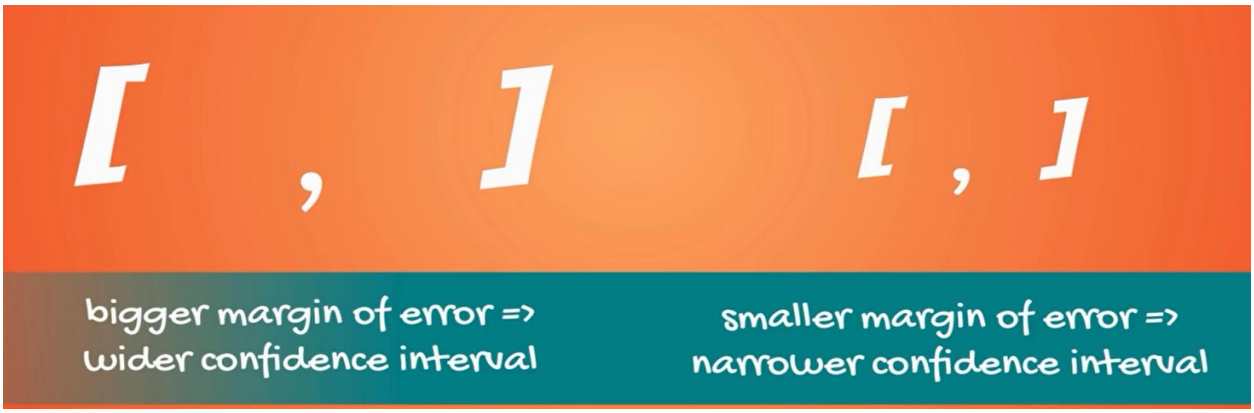
the margin of Error is equal to this expression and in the case of population variance known

getting a smaller margin of error means that the confidence interval would be narrower as we want a better prediction.

Control the margin of error



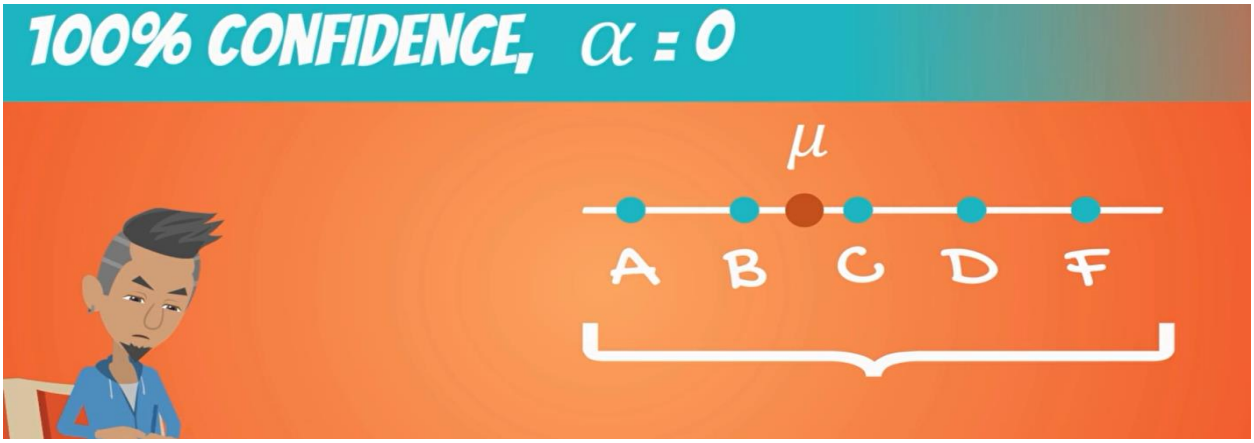
There is a statistic, a standard deviation and the sample size, statistic and the standard deviation are in the numerator. So, smaller statistics and smaller standard deviations will reduce the margin of error.

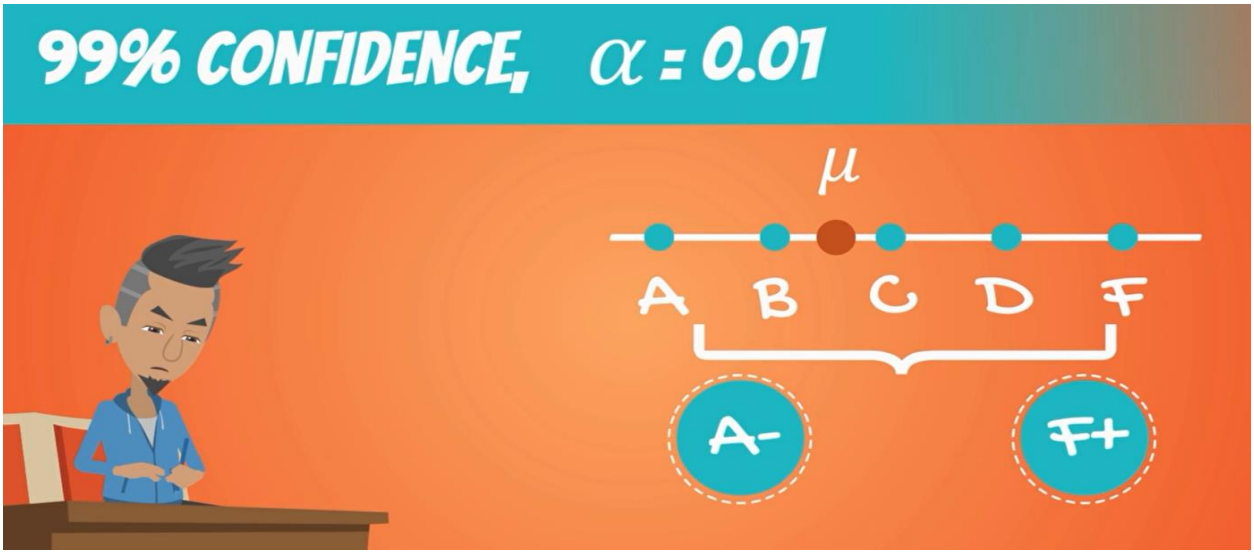


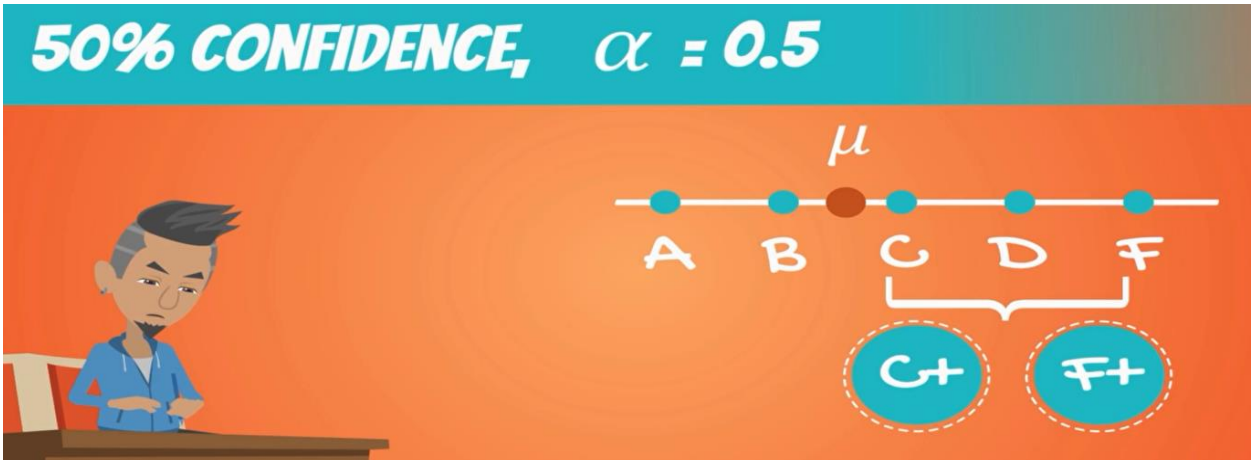
How do we do that a higher level of competence increases the statistics, a higher statistic means a higher margin of error.

This leads to a wider confidence interval.

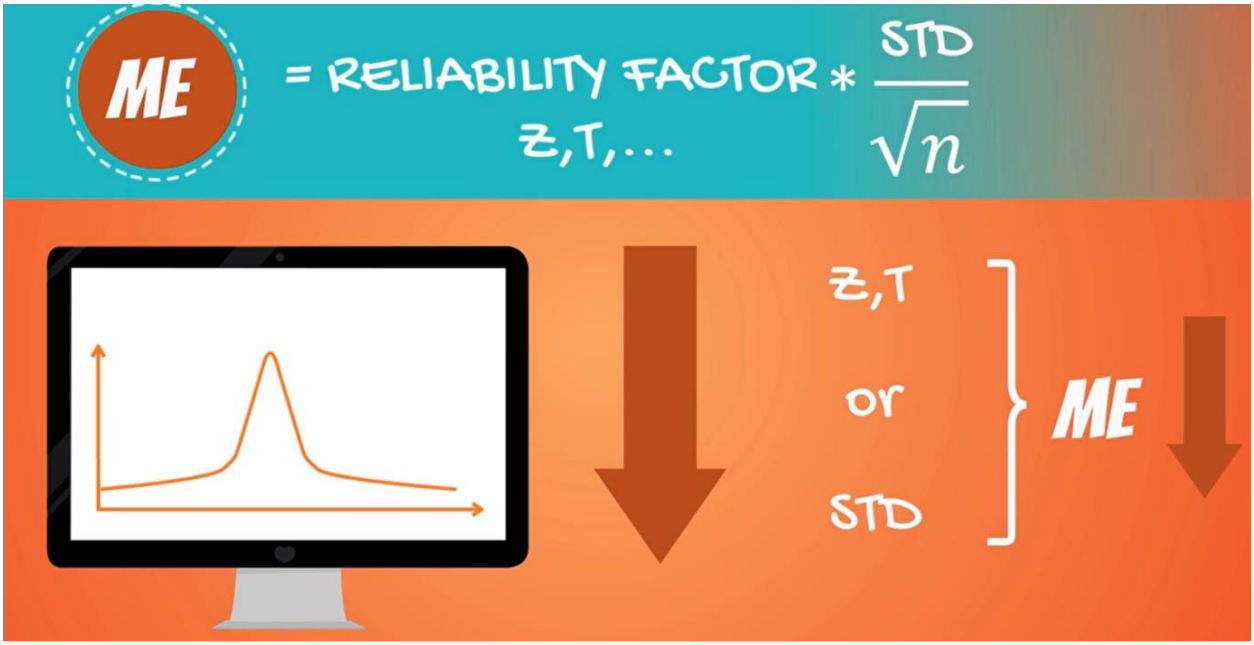
Example



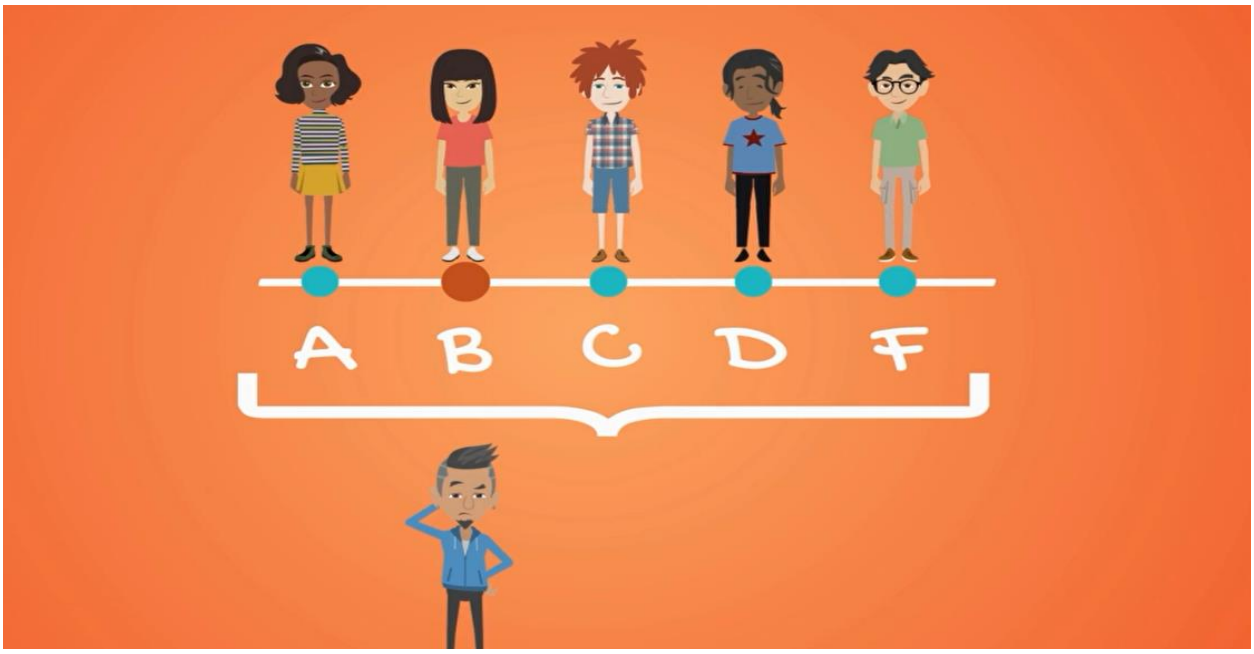




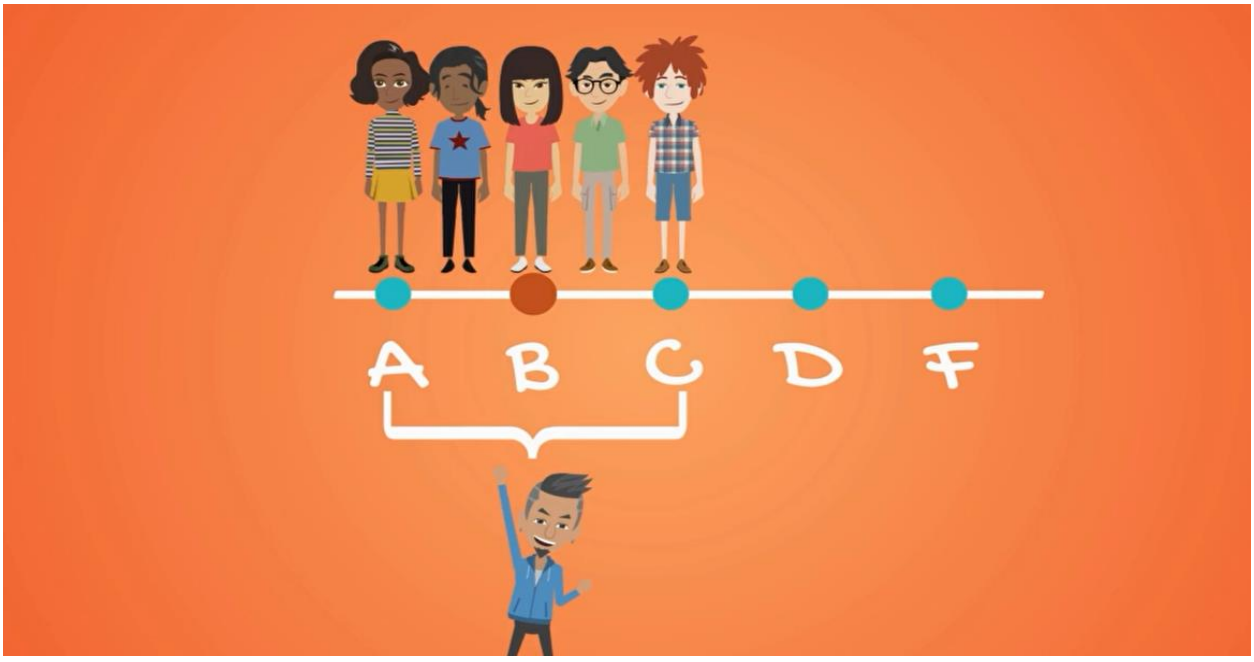
Lower confidence level result in a narrower interval.



a lower standard deviation means that the data set is more concentrated around the mean



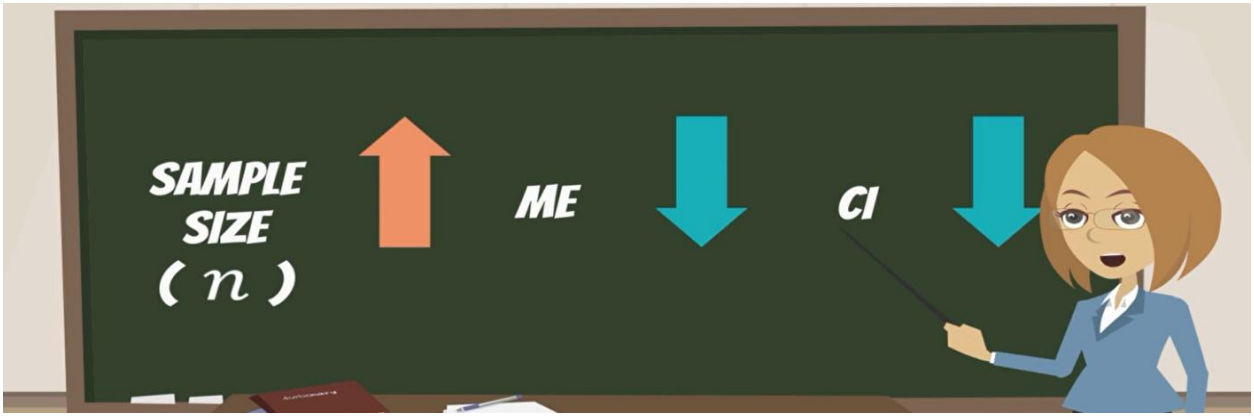
for instance, the mean grade in your class is B and you know that there are people with in A's B's C's D and a few F's. how likely is it that you got a B now?



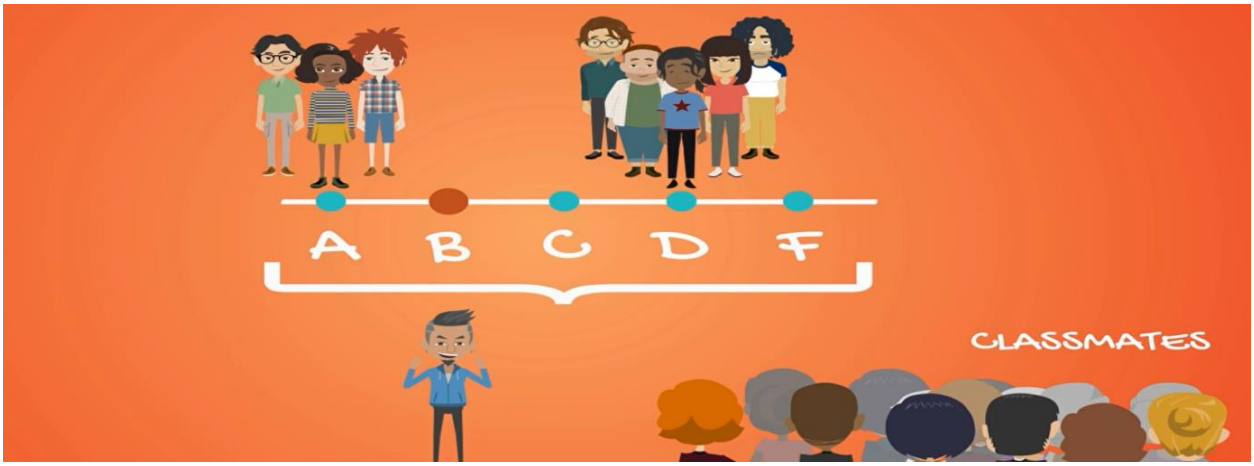
compare this to a situation when the teacher said the mean of the class is around B and the lowest grade is C.

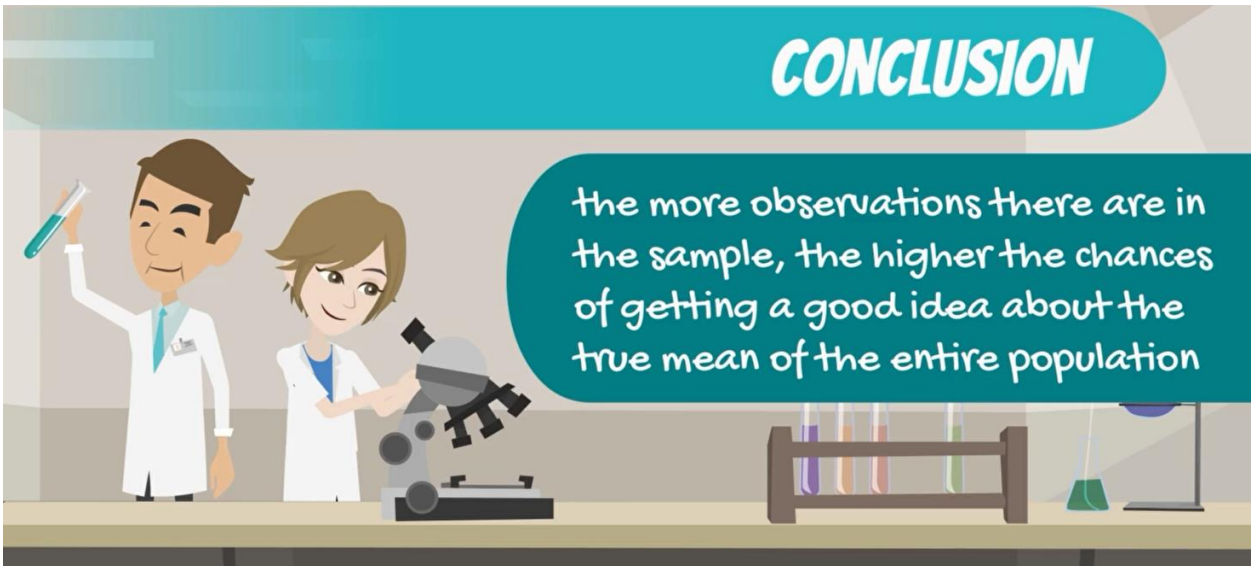
in this case you are much more likely to get a B right, in the first case the grades are dispersed while in the second they are concentrated.

lastly the higher sample size will decrease the margin of error.



Sample N= 30 students





12. Confidence intervals. Two means. Dependent samples.

We will explore confidence interval looking into two population.

In some cases, the samples that we have taken from that populations will be dependent on each other and in others they will be independent, dependent samples are easier.

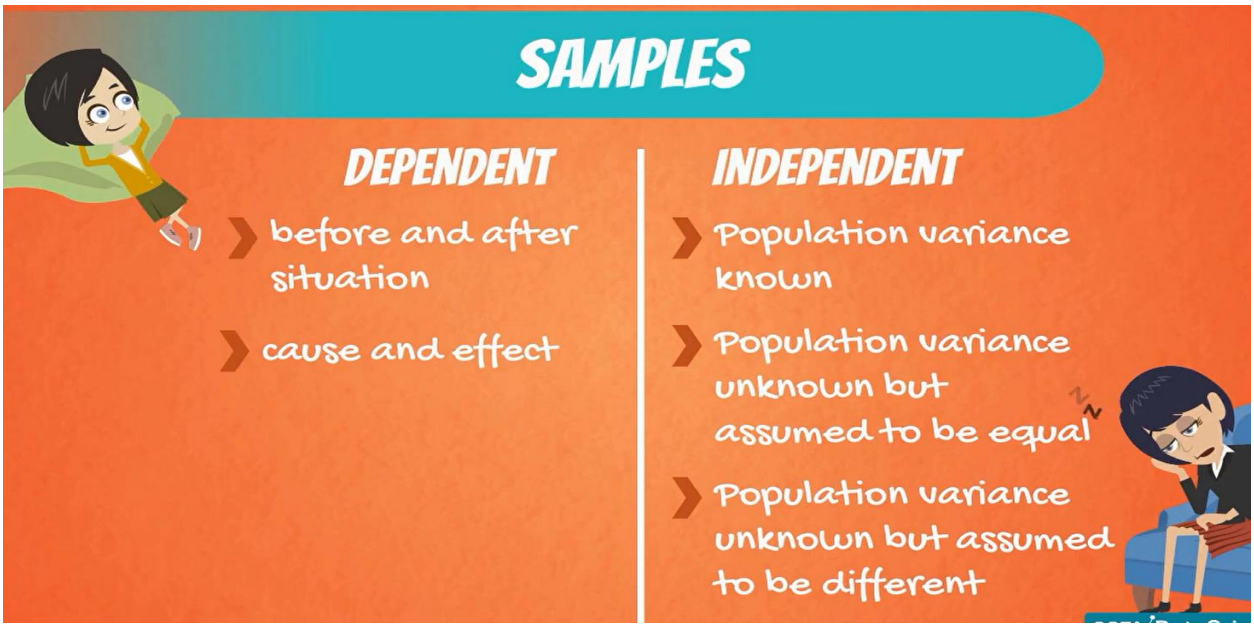
first when we are researching the same subject. examples are **weight loss** and **blood samples.**

essentially, we are looking at the same person before and after. these two examples will be explored and detailed in these lessons.

Another case in which we have dependent samples is when investigating couples or families. for instance, habits of husbands and wives.

They are obviously dependent on each other at the time, these people spend together at home often coincides watching TV, eating dinner, often sharing the same household income.

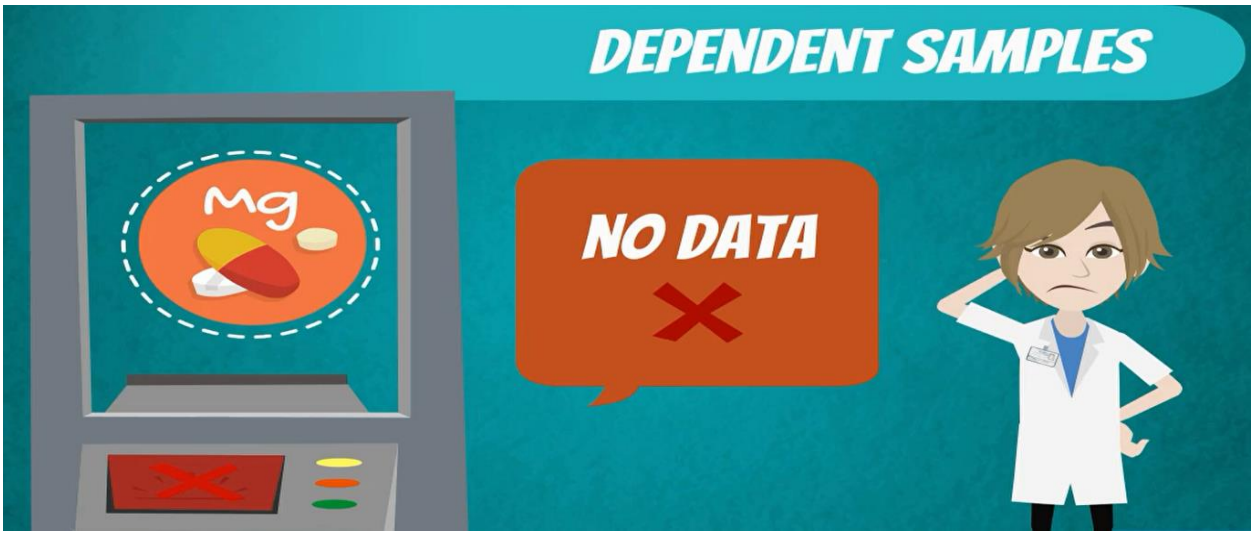
Types of samples



Dependent Samples

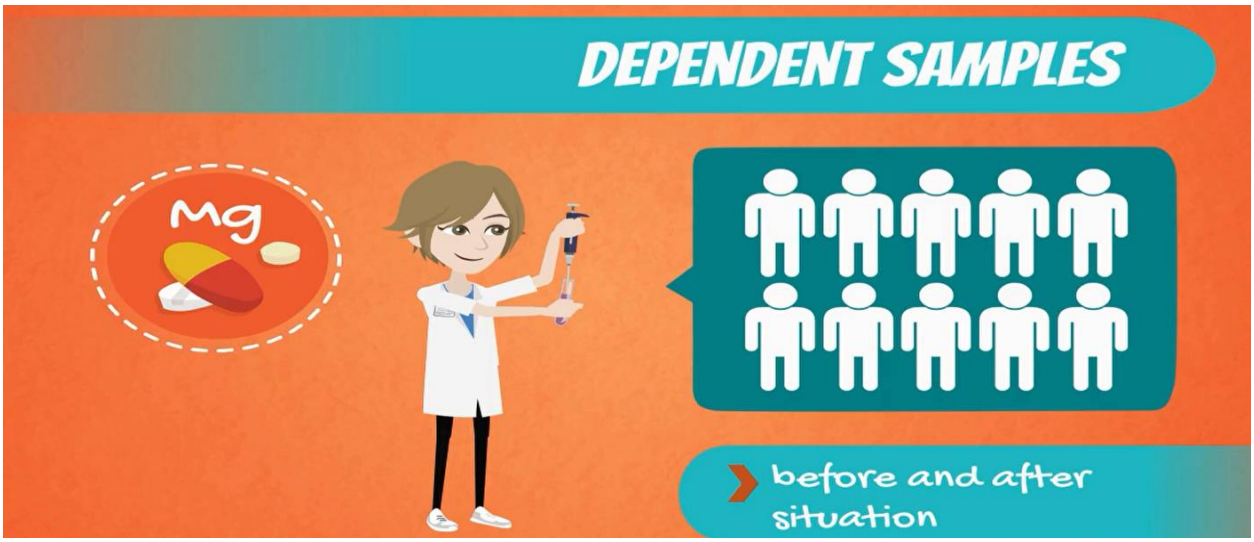
This statistical test is often used when developing medicine.

let's say you have developed a pill that increase the concentration of magnesium and blood.

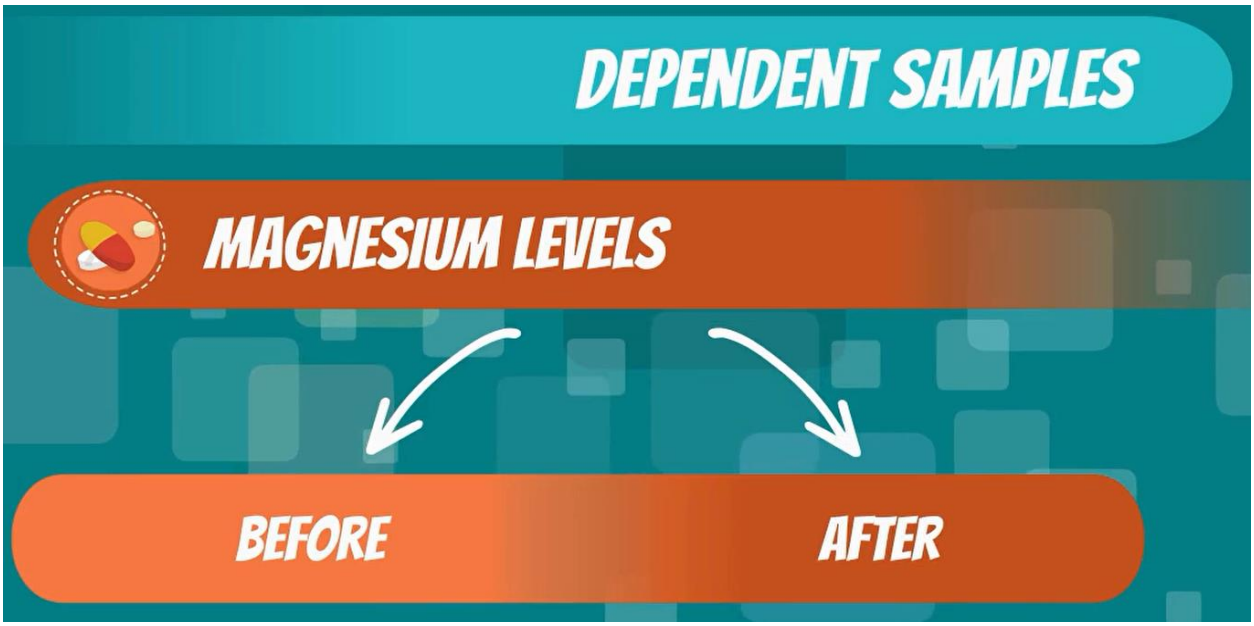


It is very promising but there is no data to support your claim.

after testing the drug in the laboratory, it is time to see its actual effect on people.



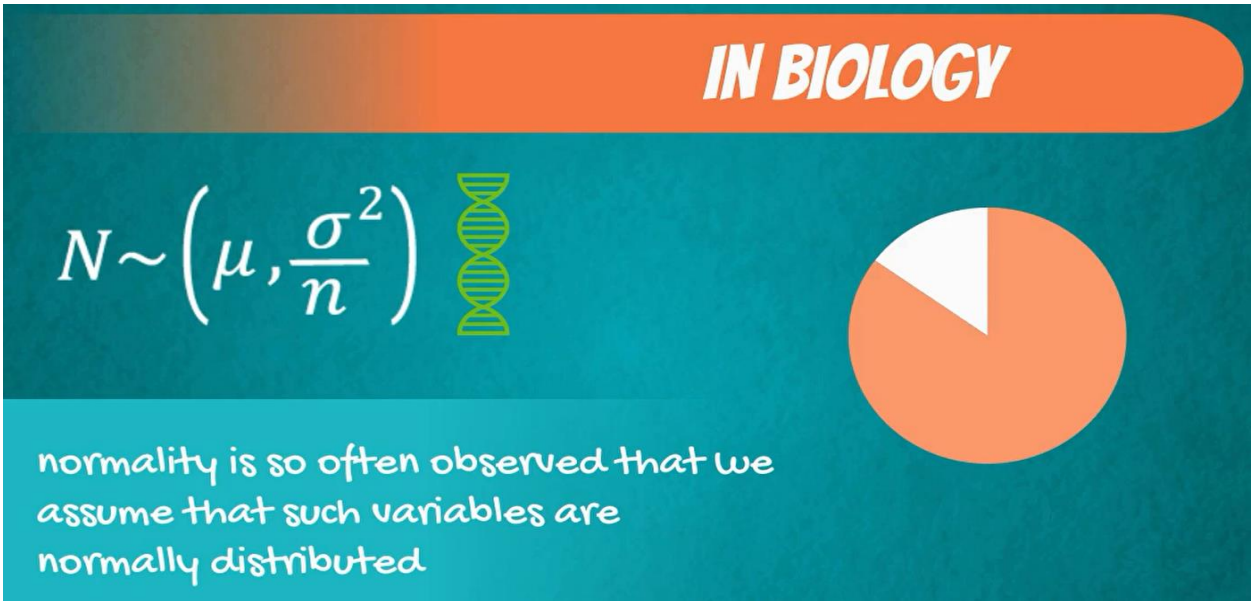
what you would typically do is take a sample of 10 people and tests or magnesium level before and after taking the pill.



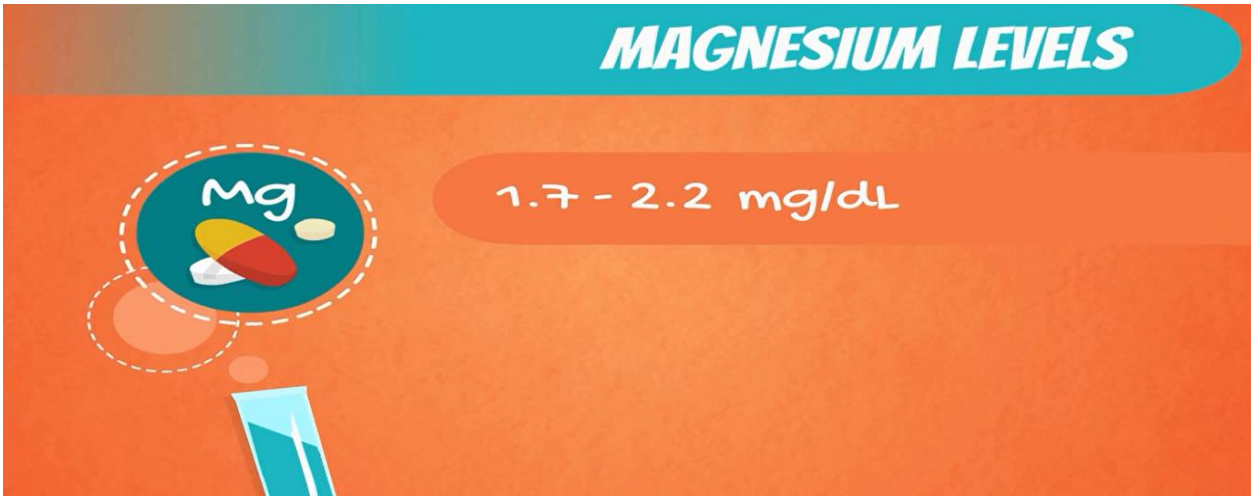
The two dependent samples are the magnesium levels before and the magnesium levels after. it is clear that it is the same people, we are testing because the samples are dependent.

Note

an important note is that the populations are normally distributed Actually when telling with biology.



whenever you take a blood test the magnesium levels are stated in mg per deciliter and a healthy person would usually have somewhere between 1.7 and 2.2 mg of magnesium mg per deciliter.



Confidence interval for difference of two means, dependent samples

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

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

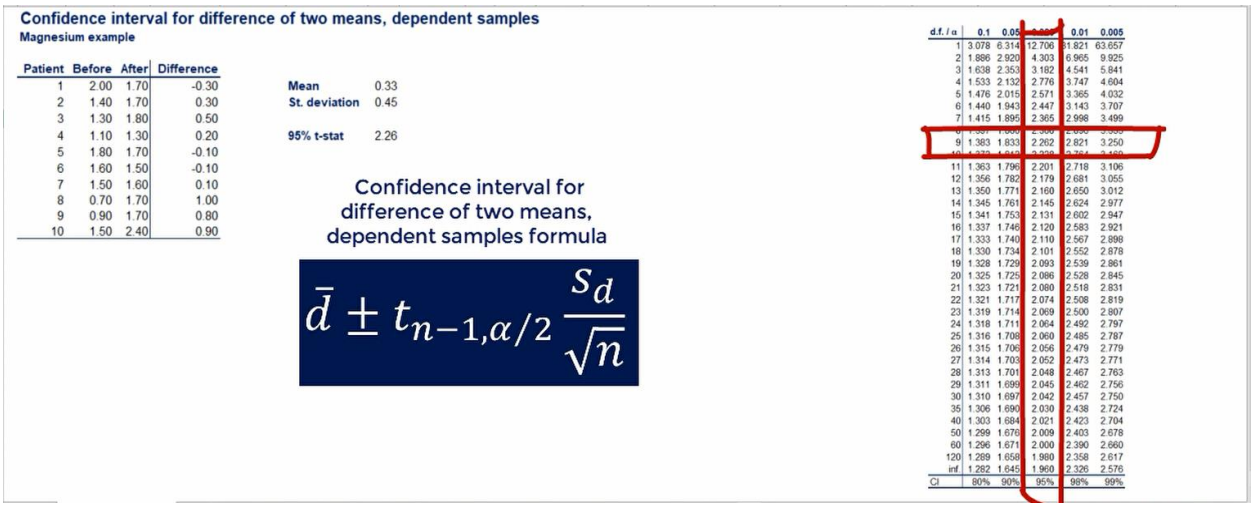
Confidence interval for a single population. Population variance unknown formula

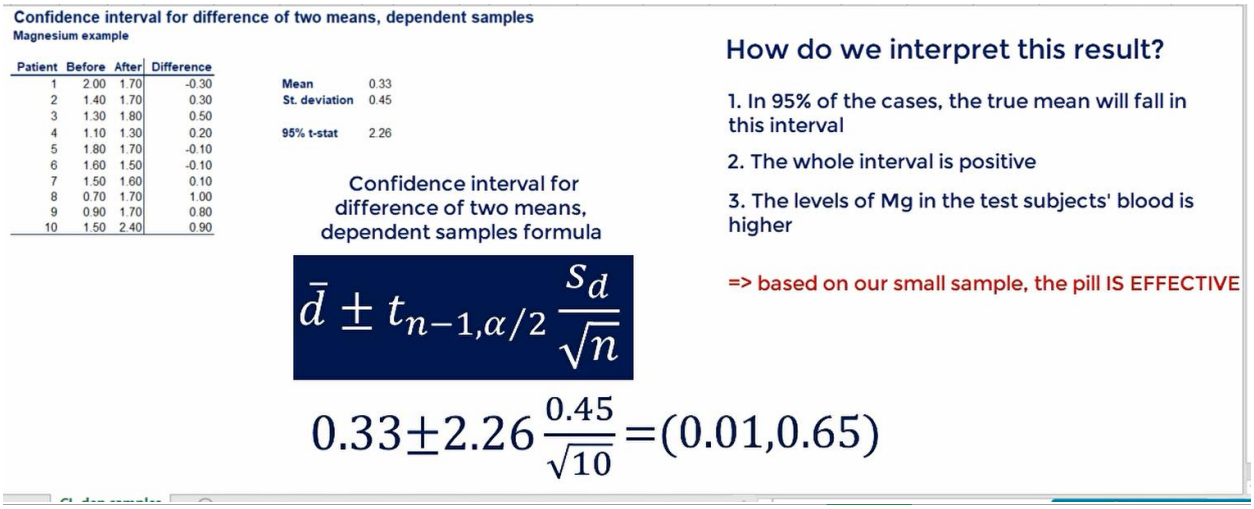
$$\bar{x} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$$

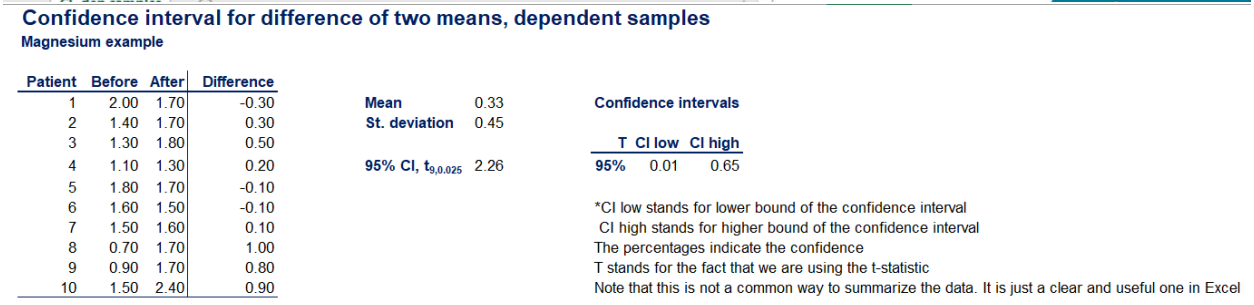
Difference = After - Before

In this way, the data looks as a single population

95% confidence is one of the most common ones







this shows that the true meaning of the difference is definitely positive. therefore with 95% certainty we can say that the levels of magnesium is the test subject blood is higher.

the purpose of this test was to determine whether the drug is effective based on our small sample.

this shows you some of the practical applications of inference

14. Confidence intervals. Two means. Independent samples (Part 1)

Known population variances

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

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	58	65
Population std	10	5

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

Thus, the variance is known

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. 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	

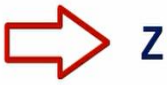
Problem: We want to find a 95% confidence interval for the difference between the grades of the students from engineering and management

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	

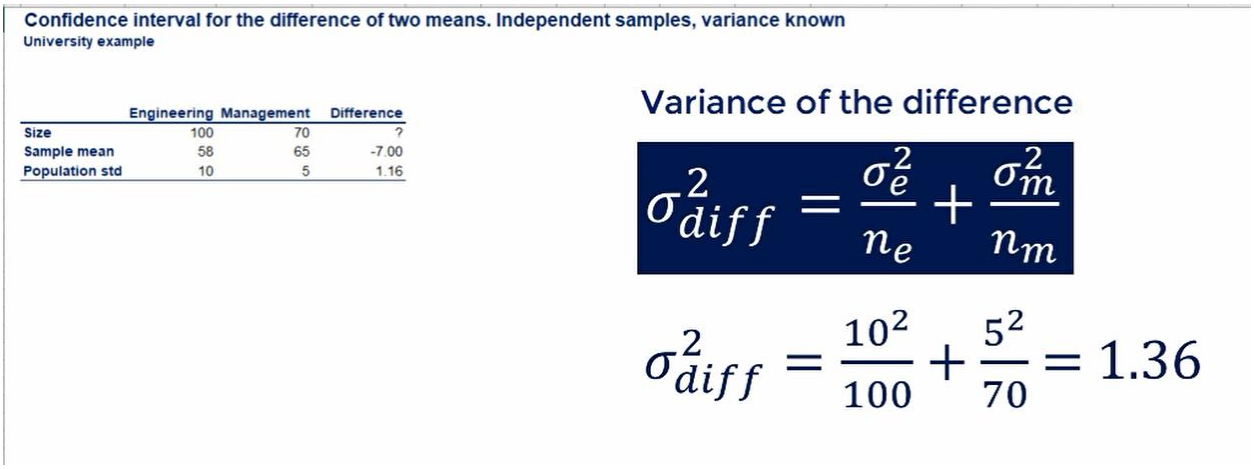
Considerations:

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



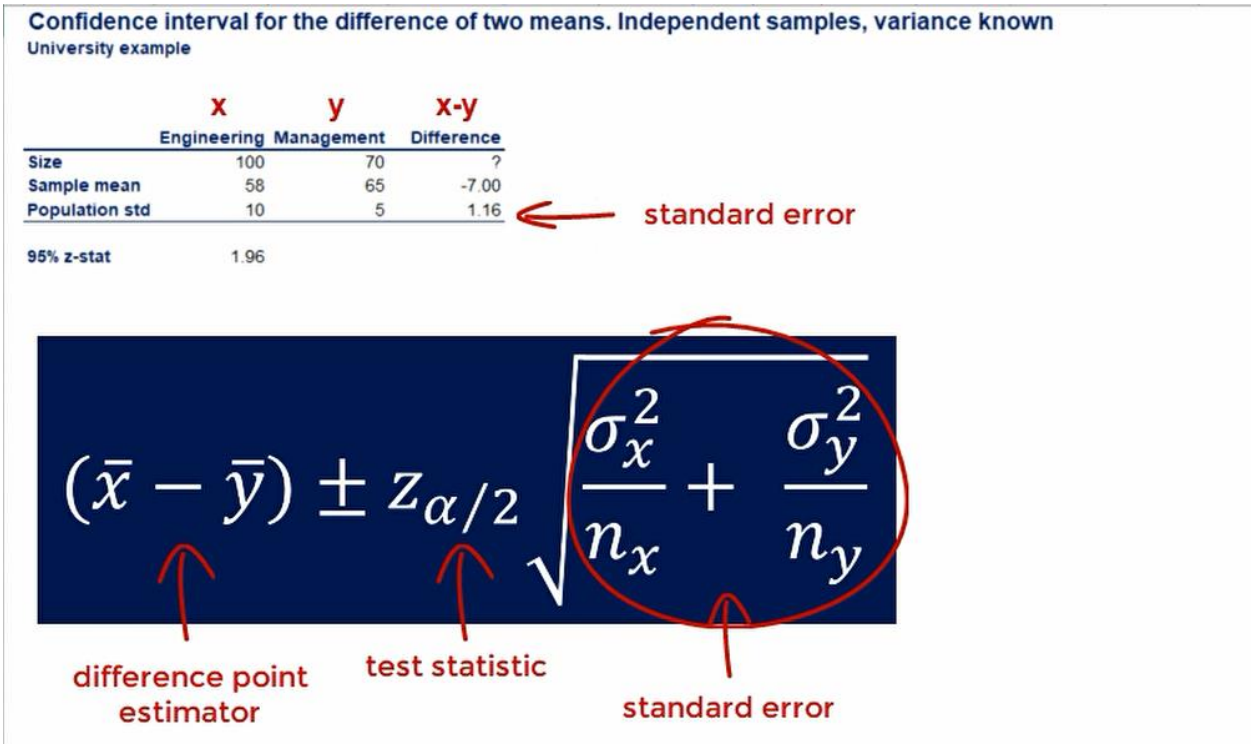
all this information points as to the z statistics instead of the t statistics

the variance of the difference between the two means

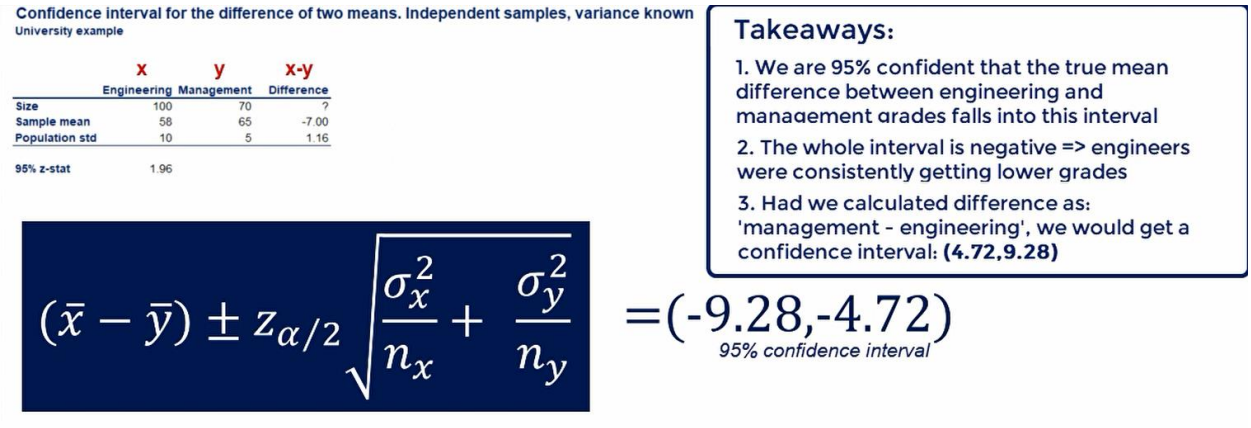


- The variance of the grades received by engineering students = σ_e^2
- Sample size of the engineering students = n_e
- The variance of the grades received by engineering students = σ_m^2
- Sample size of the engineering students = n_m

Confidence intervals



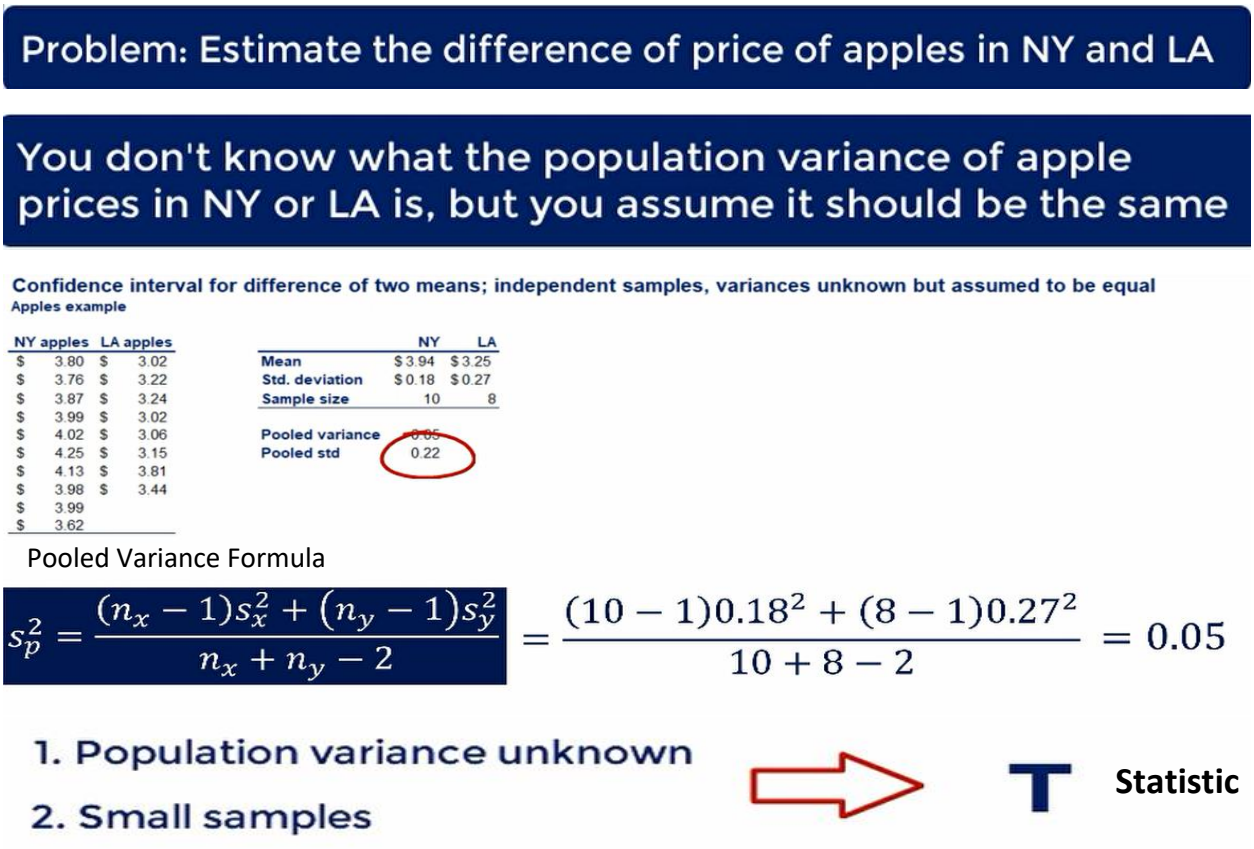
interpretation



note that this time the whole interval is negative because we are calculating engineering grade minus management grade, as the engineers why were constantly getting lower grades. So, the difference is negative.

16. Confidence intervals. Two means. Independent samples (Part 2)

population variances Unknown but assumed to equal.



we assume that the population variances are equal

so, we have to estimate them, the unbiased estimate or in this case is called the **pooled sample variance**

- Sample Standard Deviation of NL apples = s_X^2
- Sample size of the NL apples = s_X
- Sample Standard Deviation of LA apples = s_Y^2
- Sample size of the LA apples = S_Y

Confidence Interval for unknown variance

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

Confidence Interval for known variance

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

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	$ 0.15	$ 0.27
Sample size	10	8
Pooled variance	0.05	
Pooled std	0.22	

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

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

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

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 Calculation

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:

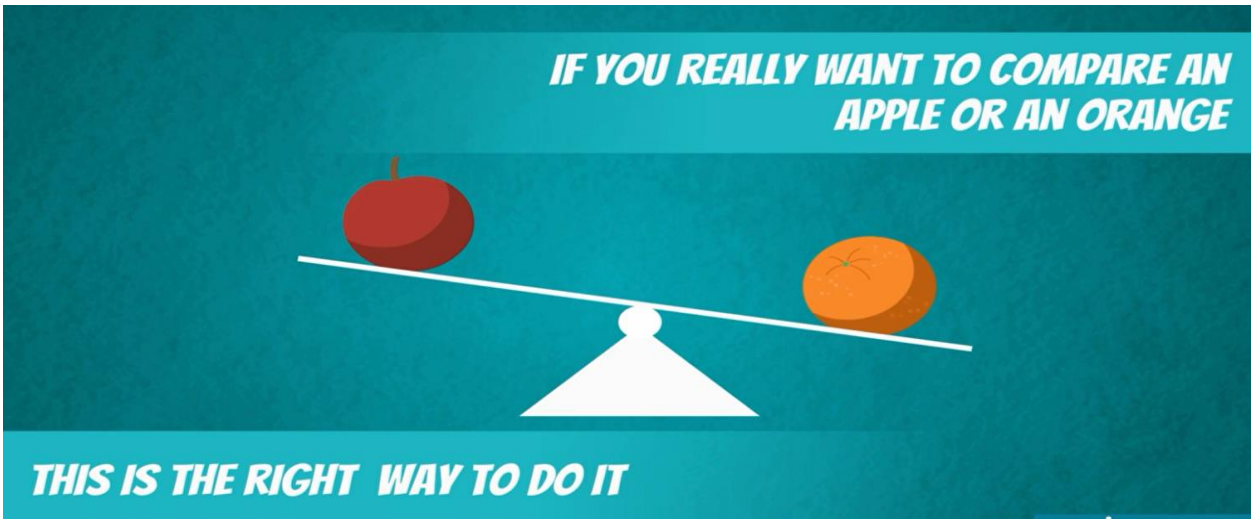
Apples in NY are much more expensive than in LA

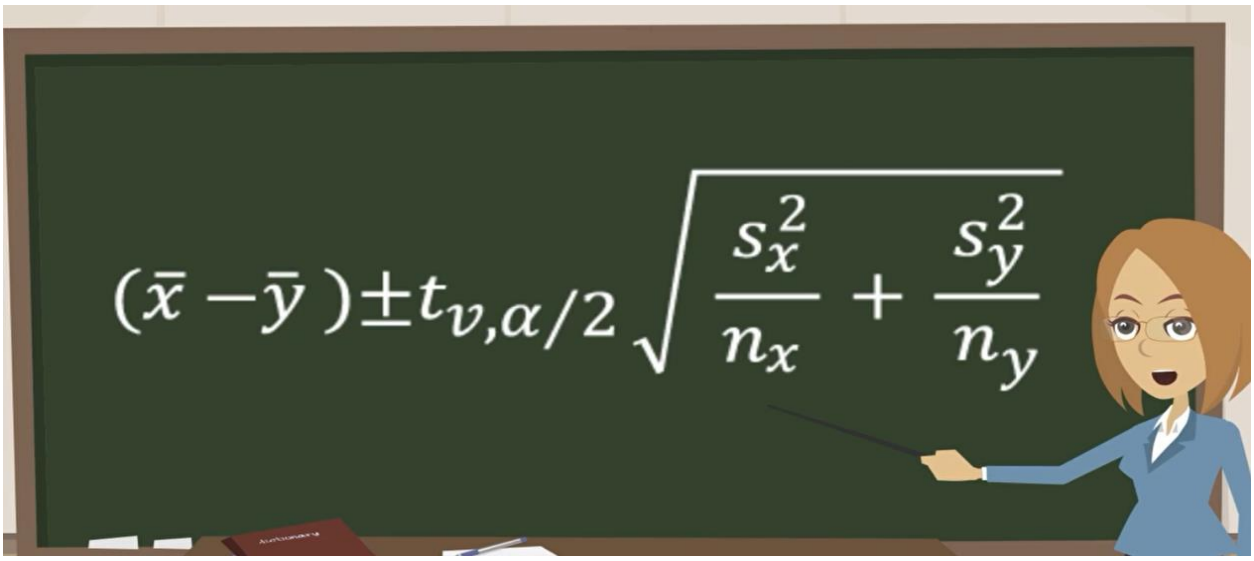
$$(\bar{x} - \bar{y}) \pm t_{n_x+n_y-2,\alpha/2}\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}} = (3.94 - 3.25) \pm 2.12\sqrt{\frac{0.05}{10} + \frac{0.05}{8}}$$

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

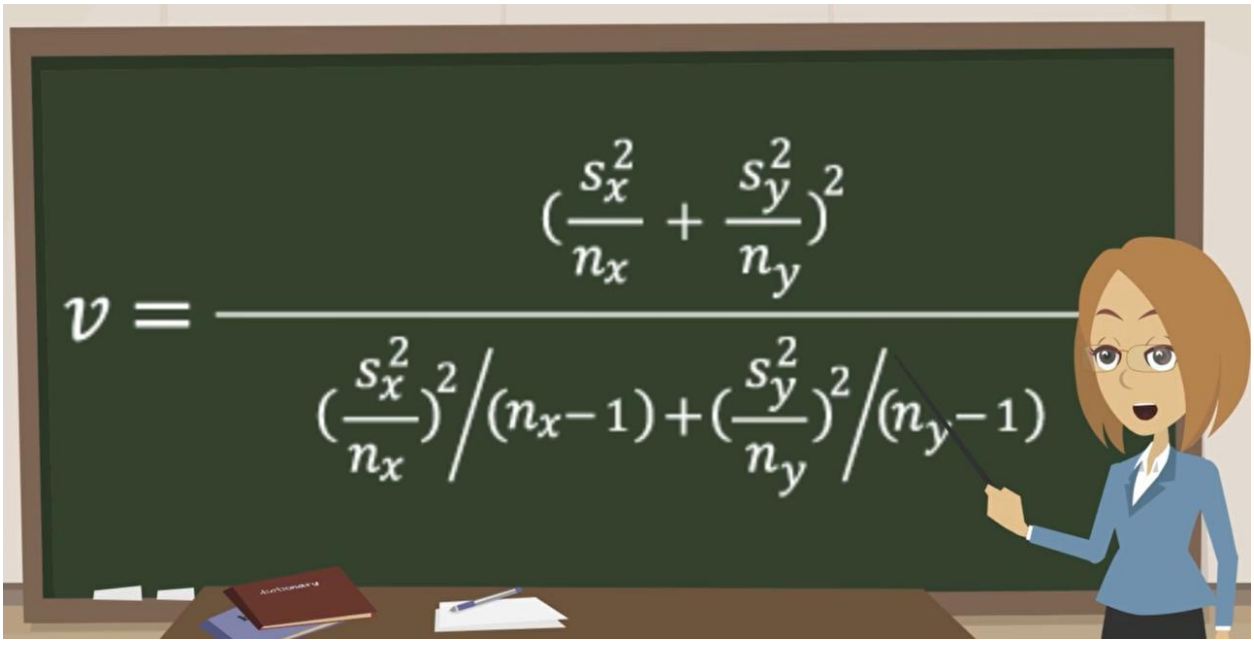
18. Confidence intervals. Two means. Independent samples (Part 3)

population variances Unknown but assumed to different.





we have the difference of the means of the two samples the variance or the sample variance of each of the two variables and here are the respective sample sizes.



the tough things about this is to estimate the degrees of freedom. statisticians have come up with a formula that allows us to do just that