Contents

[Formula 1](#_Toc524396732)

[Standard error 1](#_Toc524396733)

[What is the difference between standard deviation and standard error 2](#_Toc524396734)

[Margin of error 3](#_Toc524396735)

[Point Estimate 4](#_Toc524396736)

[Confidence Interval 4](#_Toc524396737)

[Difference from point estimate 6](#_Toc524396738)

[Confidence interval in different distribution 8](#_Toc524396739)

[Interview Questions 10](#_Toc524396740)

[*Difference between confidence level and confidence interval* 12](#_Toc524396741)

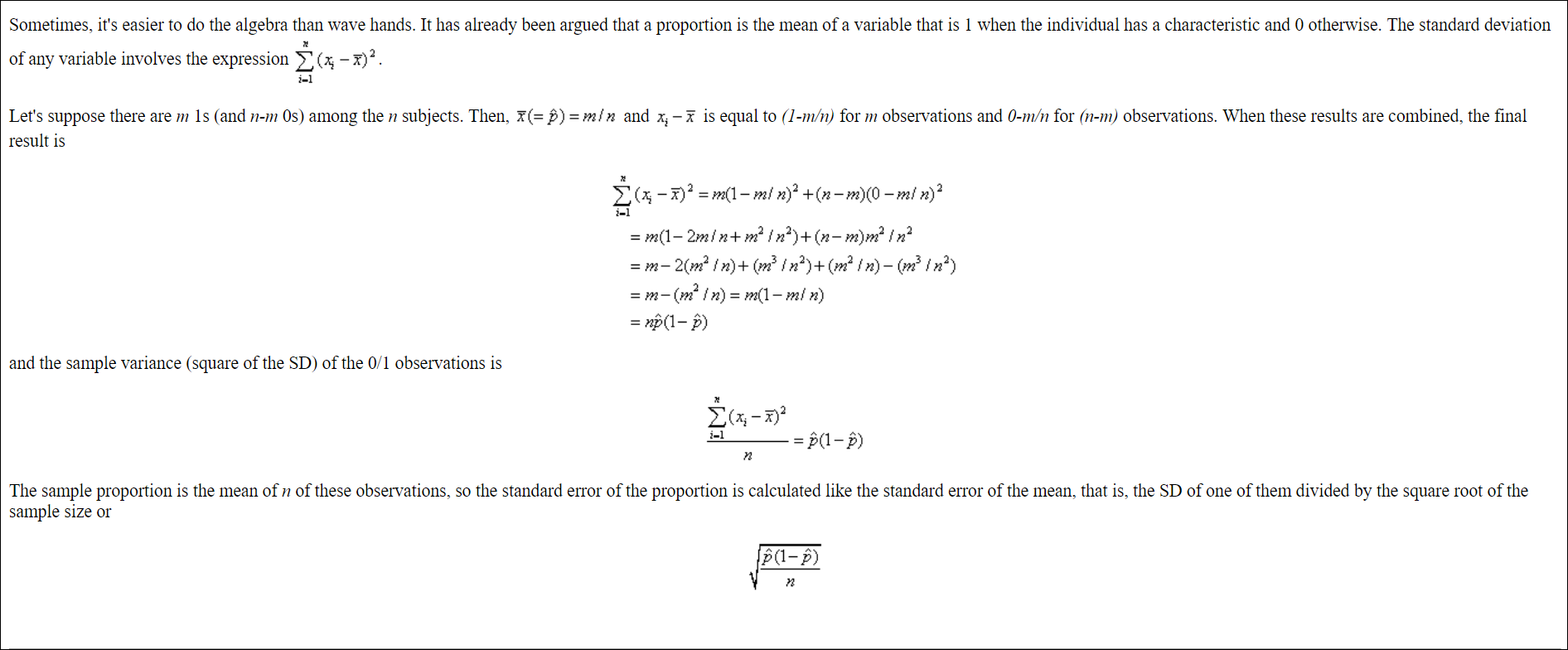
# Formula

standard error = standard deviation / sqrt n

confidence interval = X + Z s/sqrt(n)

# Standard error

standard error = standard deviation / sqrt n



# What is the difference between standard deviation and standard error

Well, **standard error** is to a **sample** what **standard deviation** is to a **population**

Standard deviation is a measure of spread - how spread out are the heights of children in my sons class? It is the average deviation from the average. To say that the standard deviation is 1, is to say that, **on average, the difference between the mean of a population, and any randomly picked member of that population, is 1.**

Standard error is also a measure of spread. It is actually just another word for standard deviation, it is simply used in two very specific circumstances:

1. Imagine a population of size N. If you take every possible sample of size n from that population, and calculate the mean for each possible sample, the means will be normally distributed, with mean of means being the same as the population mean. The standard deviation of this hypothetical distribution of means, called the sampling distribution, is called standard error, because it is how much you expect any given sample to "err" from the mean.

2. The standard deviation of any sample can also be referred to as the standard error, to highlight the fact that it is an estimate, or statistic, and not the actual population parameter.

# Margin of error

 Margin of error is an interval estimate—a pair of percentages surrounding a guess about some attribute of the full population based on a random sample from that population.  “Margin of error allows us to feel confident a certain percentage of the time, within a range above or below the ideal guess, represented by a margin we believe is least in error” (Statistics Solutions, 2013a, para. 5).

**Why do we use margin of error?**

Whenever we use a representative sample to guess something about a full population, our guess will contain some uncertainty. Using our sample statistic, we have to infer the real statistic—and that inference will mean our guess will usually be somewhere near the actual figure (a bit too low or a bit too high, Statistics Solutions, 2013b).

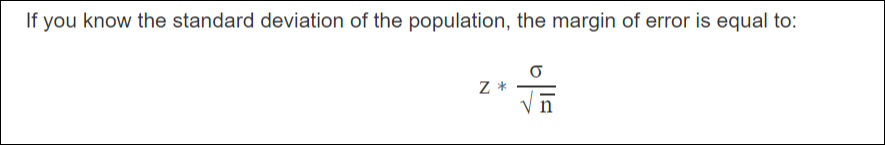
How does margin of error work?

Let’s say we conduct a survey of college students at four year institutions asking whether they prefer physical text books or electronic books (eBooks). According to the U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics (2012) there are about 13,494,131 students at four year institutions. We can’t realistically survey 13,494,131 students, so we gather a random sample from 2,500 that are representative of the full population.

Let’s say our data show that 1,875 out of 2,500 prefer eBooks (1875 / 2500 = .75 or 75%). Our margin of error, at a 95% confidence level, would be ±2% (M = 75, 95% CI [73.0, 77.0]). But how did we get the ±2% figure for margin of error (our confidence interval)? According to Sullivan (2006), our basic formula for a dichotomous outcome is:

*margin of error = critical value \* standard error*

**Margin of Error** - The resulting margin of error is what we will add or subtract from our guess to create our confidence interval.



**Critical value –** The critical value is a cut-off value that tells us how far from the sample mean we can vary and remain confident—usually one standard deviation from the mean.  We usually look it up in a z table or t table, although we can also compute it.  For our example, for a large sample size of 2,500 and a 95% level of confidence, our critical value would be ±1.96.

**What do we mean by standard error?**

Well, standard error is to a sample what standard deviation is to a population. To compute it, per Smith (2009), we estimate the population proportion (a number between 0 and 1). Our statistic is 75%, so as a proportion that would be 75 / 100 = .75. We might call this p but we don’t know p for sure, so we use p̂ (pronounced p-hat). Now we’re ready to calculate standard error using our statistic (.75). The formula is just:

SE = sqrt (p̂ (1 – p̂) / n)

Where: SE stands for standard error, p̂ is our estimated population proportion, and n is our sample size. Substituting our values we get:

SE = sqrt (.75 (1 – .75) / 2500) = .0087

Now let’s use our complete formula (margin of error = critical value \* standard error)

Margin of error (E) = 1.96 \* sqrt (p̂ (1 – p̂) / n) = (1.96 \* .0087) = .017

To get our interval, all we need to do is subtract from or add to our guess (while rounding to 2 decimal places):

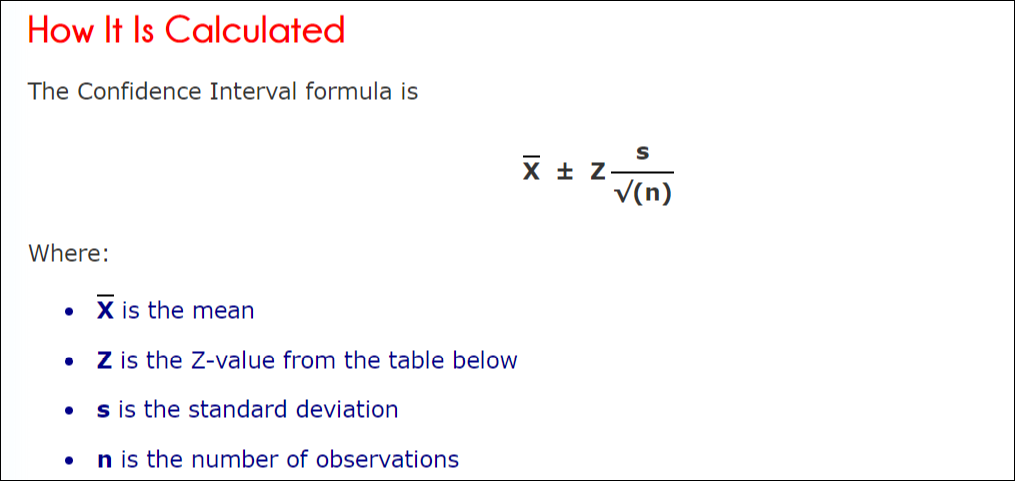
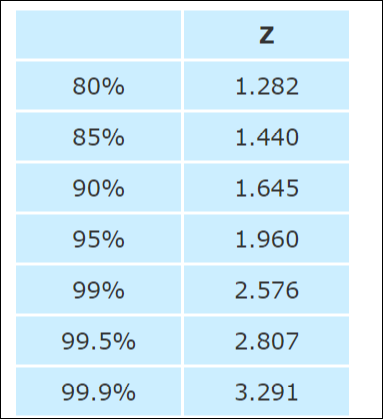
= (p̂ – (1.96 \* SE), p̂ + (1.96 \* SE) ) = (.75 – .017), (.75 + .017) = (.73, .77)

So, our margin of error, at a 95% confidence level, would be M = 75, 95% CI [73.0, 77.0]

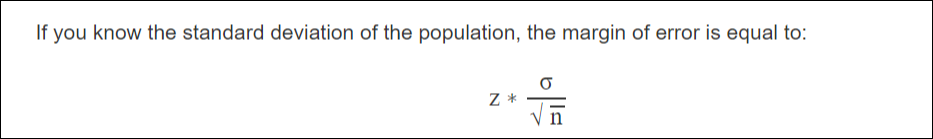
# Point Estimate

Suppose we are interested some characteristic of a population; for example, the average height *h* of all adult males in the U.S. We can estimate *h* by drawing a random sample of adult males in the U.S. and calculating the average height *H*in the sample. This is called a ***point******estimate*** of *h*.

# Confidence Interval

Calculate a confidence interval by taking a point estimate and then adding and subtracting a margin of error to create a range. Margin of error is based on your desired confidence level, the spread of the data and the size of your sample.



Where σ (sigma) is the population standard deviation, n is sample size, and z is a number known as the z-critical value. The z-critical value is the number of standard deviations you'd have to go from the mean of the normal distribution to capture the proportion of the data associated with the desired confidence level. For instance, we know that roughly 95% of the data in a normal distribution lies within 2 standard deviations of the mean, so we could use 2 as the z-critical value for a 95% confidence interval (although it is more exact to get z-critical values with stats.norm.ppf().).

|  |
| --- |
| np.random.seed(10)  sample\_size = 1000  sample = np.random.choice(a= population\_ages, size = sample\_size)  sample\_mean = sample.mean()  z\_critical = stats.norm.ppf(q = 0.975) *# Get the z-critical value\**  print("z-critical value:") *# Check the z-critical value*  print(z\_critical)  pop\_stdev = population\_ages.std() *# Get the population standard deviation*  margin\_of\_error = z\_critical \* (pop\_stdev/math.sqrt(sample\_size))  confidence\_interval = (sample\_mean - margin\_of\_error,  sample\_mean + margin\_of\_error)  print("Confidence interval:")  print(confidence\_interval) |

\*Note: We use stats.norm.ppf(q = 0.975) to get the desired z-critical value instead of q = 0.95 because the distribution has two tails.

Notice that the confidence interval we calculated captures the true population mean of 43.0023.

If you don't know the standard deviation of the population, you have to use the standard deviation of your sample as a stand in when creating confidence intervals. Since the sample standard deviation may not match the population parameter the interval will have more error when you don't know the population standard deviation. To account for this error, we use what's known as a t-critical value instead of the z-critical value. The t-critical value is drawn from what's known as a t-distribution--a distribution that closely resembles the normal distribution but that gets wider and wider as the sample size falls. The t-distribution is available in scipy.stats with the nickname "t" so we can get t-critical values with stats.t.ppf().

|  |
| --- |
| np.random.seed(10)  sample\_size = 25  sample = np.random.choice(a= population\_ages, size = sample\_size)  sample\_mean = sample.mean()  t\_critical = stats.t.ppf(q = 0.975, df=24) *# Get the t-critical value\**  print("t-critical value:") *# Check the t-critical value*  print(t\_critical)  sample\_stdev = sample.std() *# Get the sample standard deviation*  sigma = sample\_stdev/math.sqrt(sample\_size) *# Standard deviation estimate*  margin\_of\_error = t\_critical \* sigma  confidence\_interval = (sample\_mean - margin\_of\_error,  sample\_mean + margin\_of\_error)  print("Confidence interval:")  print(confidence\_interval) |

## Difference from point estimate

Suppose we are interested some characteristic of a population; for example, the average height *h* of all adult males in the U.S. We can estimate *h* by drawing a random sample of adult males in the U.S. and calculating the average height *H*in the sample. This is called a *point estimate* of *h*. If the sample is large, *H*will be a good estimate of *h,*but by itself it does not tell you how good it is.  
  
A 95% confidence interval is a different kind of estimate. It consists of *two*numbers *L*(lower) and *U*(upper), which are derived from the sample in some way without knowledge of the unknown *h*(or any other unknown parameters)*.*The interval (*L*,*U*) is supposed to contain the unknown *h.*A procedure for finding (*L*,*U*) which does in fact contain *h* for 95% of the possible samples is called a 95% confidence interval. If the interval is short, it gives us a small range of "likely" values for *h.*

That is the definition. Now, a few comments. Why the strange word "confidence," which is never used by itself in probability or statistics? Why the scare quotes around the word "likely" in the previous paragraph?   
    
Confidence intervals are a tool ofthe *frequentist* school of statistics, which holds that we should use the concepts of probability and randomness only to describe the mechanics of certain kinds of sampling from populations, and **not**to describe our certainty or degree of belief. Frequentists aim to use probability in an objective way.    
    
For a frequentist, a statement like "the probability that the average height *h*of the all males in the US lies between 70 and 74 inches is 95%" is meaningless: *h*is just a number we don't know. It either lies in the interval (70, 74) or it doesn't.

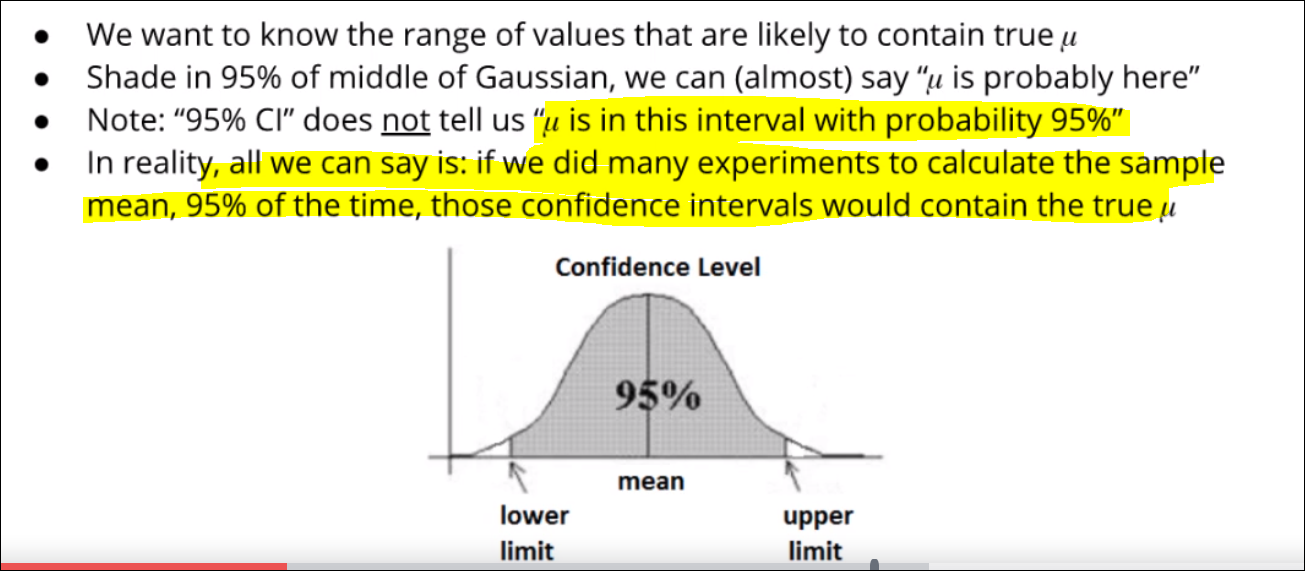
Confidence intervals are a trick frequentists use to make statements resembling the one above without violating their rules about how probability should be used. According to the definition given above, it is legitimate to write:   
P(L≤h≤U)=95%P(L≤h≤U)=95%   
if (*L, U)*is derived according to a rule so that it does contain *h*for 95% of samples. This resembles a subjective statement about our certainty that *h*lies in the range (*L, U*). But it isn't: it's an objective statement about how often, in the long run, our random interval will contain the fixed but unknown *h*, according to the randomness in our sampling.  The "subject" of the probability statement above looks like it is *h* (and that is the trick)but actually it is the interval.

**Formula**

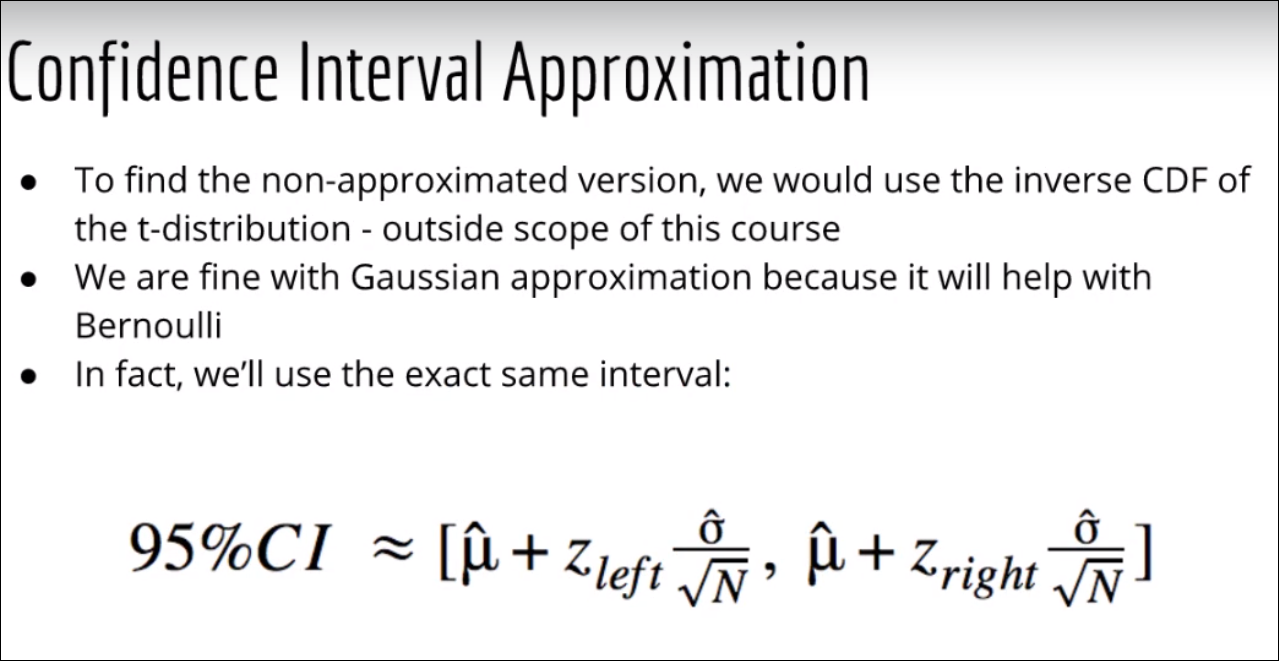
margin of error = critical value \* standard error

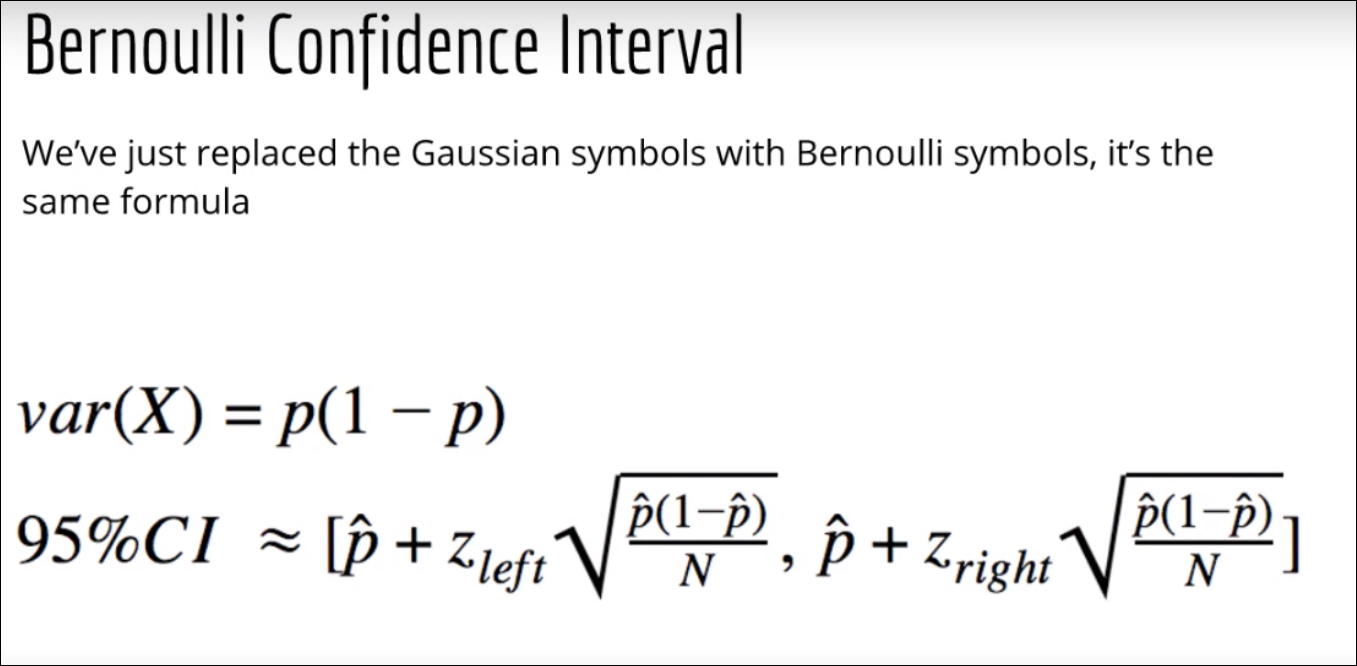
If we perform the experiemnt 1000 times 95% of the times to

Calculate sample mean , 95% if the time those confidence interval would contain true mean

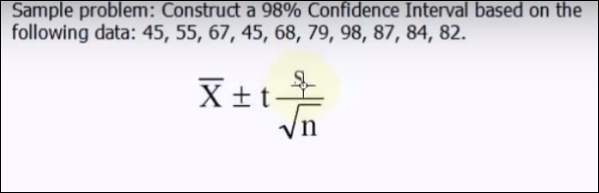


## Confidence interval in different distribution





**Exmaple**



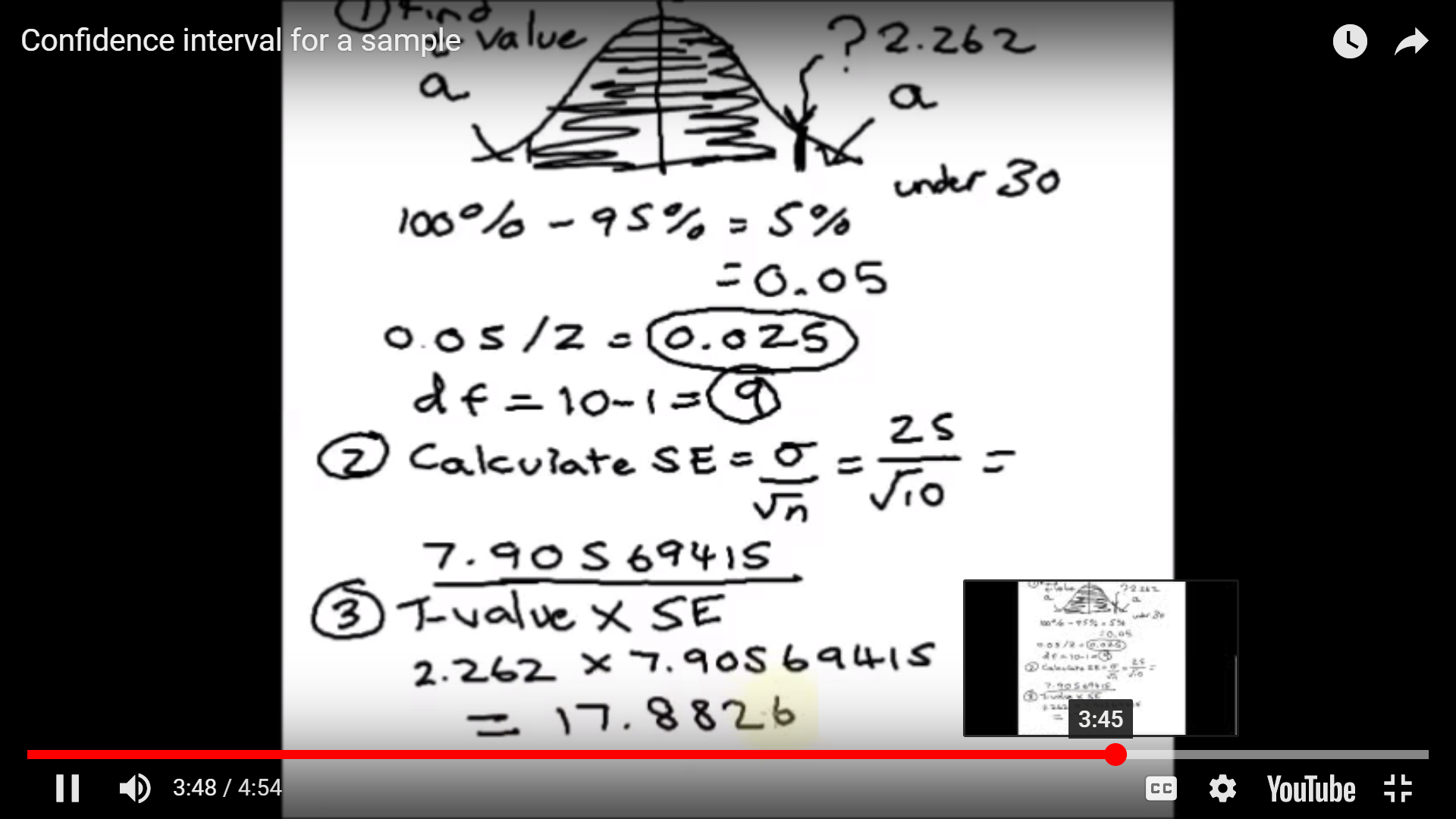
T = t value

S = standard deviation

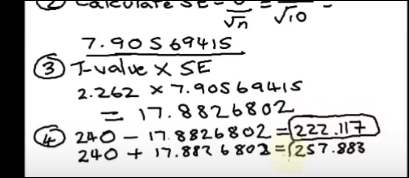
Step 1 find the t value

Calculate the alpha

Step 2 calculate the standard error = standard deviation / sqrt n

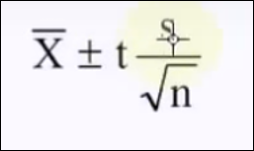


Step 3 – compute margin of error = standard error \* critical value(T value)



Step 4 confidence interval = mean – margin of error

**Construct a 95% confidence interval for sample 101,102,103,104,105**



t = 1.96

103+ 1.96\*s/sqrt(5)

S= sqrt((x1-xmean)^2 + (x2-xmean)^2 + (x3-xmean)^2)

=Sqrt(4+1+0+1+4/5)

=sqrt(10/5)

Sqrt(2)

1.5

1.5

# Prediction Interval

Confidence interval is about the mean value

Prediction interval is about the individual value of y

# Interview Questions

#### **How would you explain a confidence interval to an engineer with no statistics background? What does 95% confidence mean?**

https://www.quora.com/What-is-a-confidence-interval-in-laymans-terms

# Suppose we are interested some characteristic of a population; for example, the average height h of all adult males in the U.S. We can estimate h by drawing a random sample of adult males in the U.S. and calculating the average height H in the sample. This is called a point estimate of h. If the sample is large, H will be a good estimate of h, but by itself it does not tell you how good it is.

# A 95% confidence interval is a different kind of estimate. It consists of two numbers L (lower) and U (upper), which are derived from the sample in some way without knowledge of the unknown h (or any other unknown parameters). The interval (L,U) is supposed to contain the unknown h. A procedure for finding (L,U) which does in fact contain h for 95% of the possible samples is called a 95% confidence interval. If the interval is short, it gives us a small range of "likely" values for h.

***What is the relation between Margin of error and sample size***

In statistics, the two most important ideas regarding sample size and margin of error are, first, sample size and margin of error have an inverse relationship; and second, after a point, increasing the sample size beyond what you already have gives you a diminished return because the increased accuracy will be negligible.

The relationship between margin of error and sample size is simple: As the sample size increases, the margin of error decreases. This relationship is called an inverse because the two move in opposite directions. If you think about it, it makes sense that the more information you have, the more accurate your results are going to be (in other words, the smaller your margin of error will get). (That assumes, of course, that the data were collected and handled properly.)

**For example**

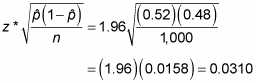
|  |  |  |
| --- | --- | --- |
| Suppose that the Gallup Organization’s latest poll sampled 1,000 people from the United States, and the results show that 520 people (52%) think the president is doing a good job, compared to 48% who don’t think so. First, assume you want a 95% level of confidence, so you find*z\**using the following table*.* |  |  |

From the table, you find that *z\**= 1.96.

The number of Americans in the sample who said they approve of the president was found to be 520. This means that the sample proportion,

image0.png

is 520 / 1,000 = 0.52. (The sample size,*n,*was 1,000.) The margin of error for this polling question is calculated in the following way:



According to this data, you conclude with 95% confidence that 52% of all Americans approve of the president, plus or minus 3.1%.

Using the same formula, you can look at how the margin of error changes dramatically for samples of different sizes. Suppose in the presidential approval poll that*n*was 500 instead of 1,000. Now the margin of error for 95% confidence is

image2.png

which is equivalent to 4.38%. If*n*is increased to 1,500, the margin of error (with the same level of confidence) becomes

image3.png

or 2.53%. Finally, when*n*= 2,000, the margin of error is

image4.png

or 2.19%.

Looking at these different results, you can see that larger sample sizes decrease the margin of error, but after a certain point, you have a diminished return. Each time you survey one more person, the cost of your survey increases, and going from a sample size of, say, 1,500 to a sample size of 2,000 decreases your margin of error by only 0.34% (one third of one percent!) — from 0.0253 to 0.0219. The extra cost and trouble to get that small decrease in the margin of error may not be worthwhile. Bigger isn’t always that much better!

# *Difference between confidence level and confidence interval*

A **confidence interval** is a range of values that is likely to contain an unknown population parameter. If you draw a random sample many times, a certain percentage of the**confidence intervals** will contain the population mean. This percentage is the **confidence level**.