CS 237: Probability in Computing

Wayne Snyder
Computer Science Department
Boston University

Lecture 17:

- Central Limit Theorem
- Sampling Theory (Application of the CLT)
- Point Estimates: warmup -- when the population parameters are known

The Central Limit Theorem

We will do the first part of the lecture from the notebook posted on the class web page.

Review: CLT and the Normal Distribution

Now suppose we consider the random variable $\overline{X}_n$ representing the mean of the X_i, i.e.,

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

The Central Limit Theorem

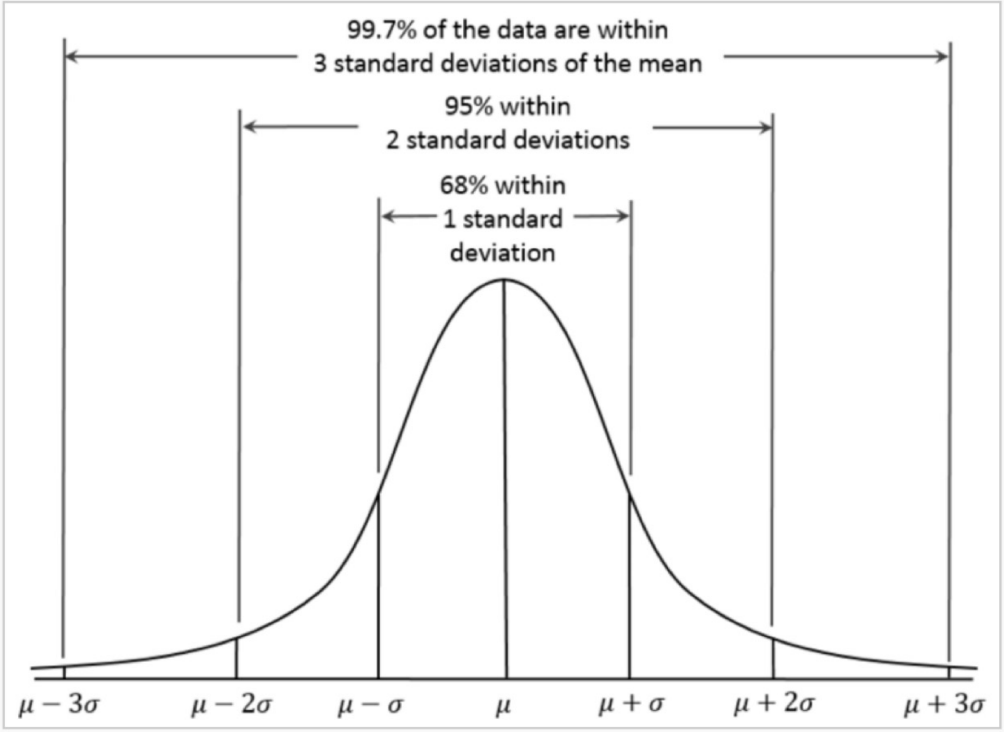
As n gets large, the random variable $\overline{X}_n$ converges to the distribution $N\left(\mu, \frac{\sigma^2}{n}\right)$.

There are several crucial things to remember about the CLT:

1. The mean μ of $\overline{X}_n$ is the same as the X_i.
2. But the standard deviation $\frac{\sigma}{\sqrt{n}}$ gets smaller as n gets larger, and approaches 0 as n approaches ∞.
3. The distributions of the X_i do NOT MATTER at all, and as long as they have a common mean and standard deviation, they can be completely different distributions. Typically, however, these are separate "pokes" of the same random variable.
4. We can use the strong properties of the normal distribution, such as the "68-95-99 rule," to quantify the randomness inherent in the sampling process. This will be the fundamental fact we will use in developing the various statistical procedures in elementary statistics.

Review: CLT and the Normal Distribution

The 68 – 95 – 99 Rule



For the normal distribution, the values less than one standard deviation away from the mean account for 68.27% of the set; while two standard deviations from the mean account for 95.45%; and three standard deviations account for 99.73%.

Actually, we can be more precise...

```
+/- 1 sigma = 0.682689492137
+/- 2 sigma = 0.954499736104
+/- 3 sigma = 0.997300203937
+/- 4 sigma = 0.999936657516
+/- 5 sigma = 0.999999426697
+/- 6 sigma = 0.999999998027
+/- 7 sigma = 0.999999999997
```

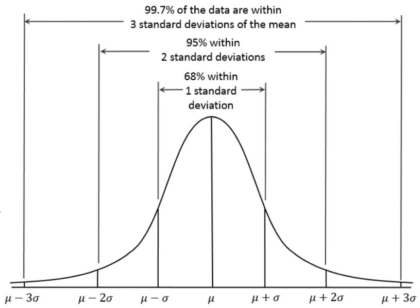
Review: CLT and the Normal Distribution

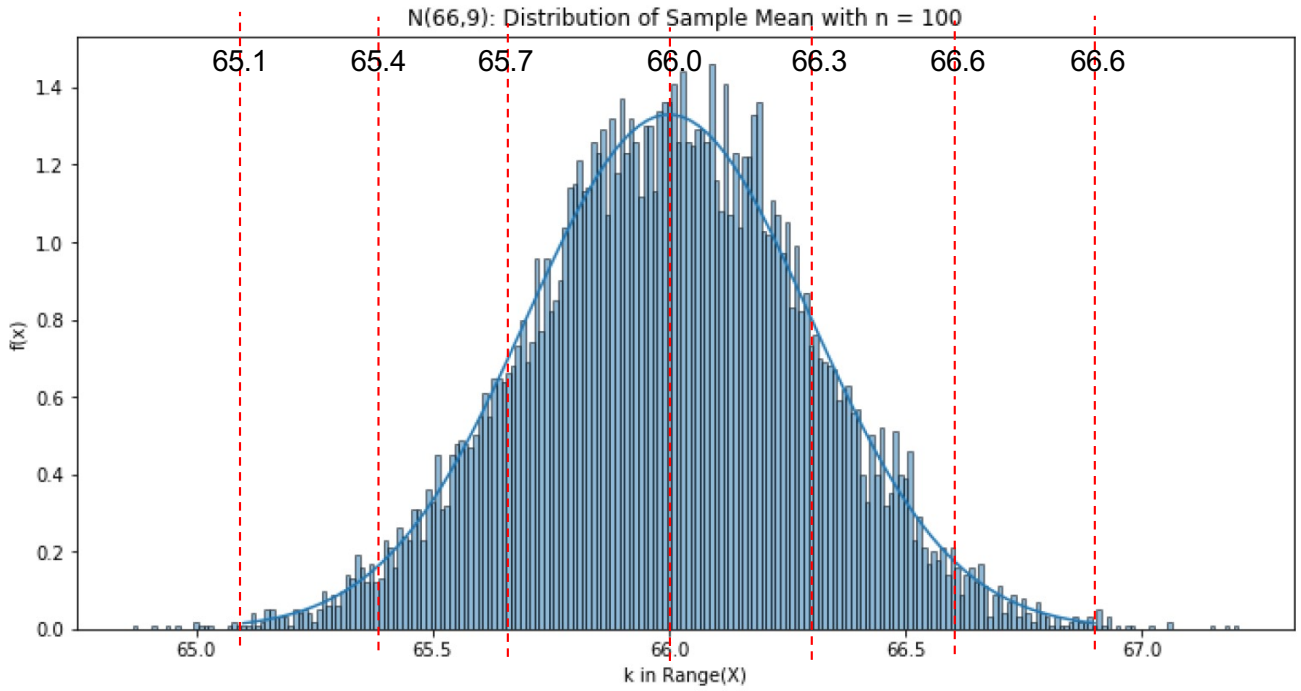
Example: Let X ~ N(66,3^2). We calculated the mean for n = 100, so we should get a standard deviation smaller by a factor of 10:

$$\frac{\sigma}{\sqrt{n}}$$

$$\overline{X}_{100} = \frac{X_1 + X_2 + \cdots + X_{100}}{100} \qquad \overline{X}_{100} \sim N(66, (0.3)^2) \qquad \sigma_{\overline{X}_{100}} = \frac{3}{\sqrt{100}} = \frac{3}{10} = 0.3$$

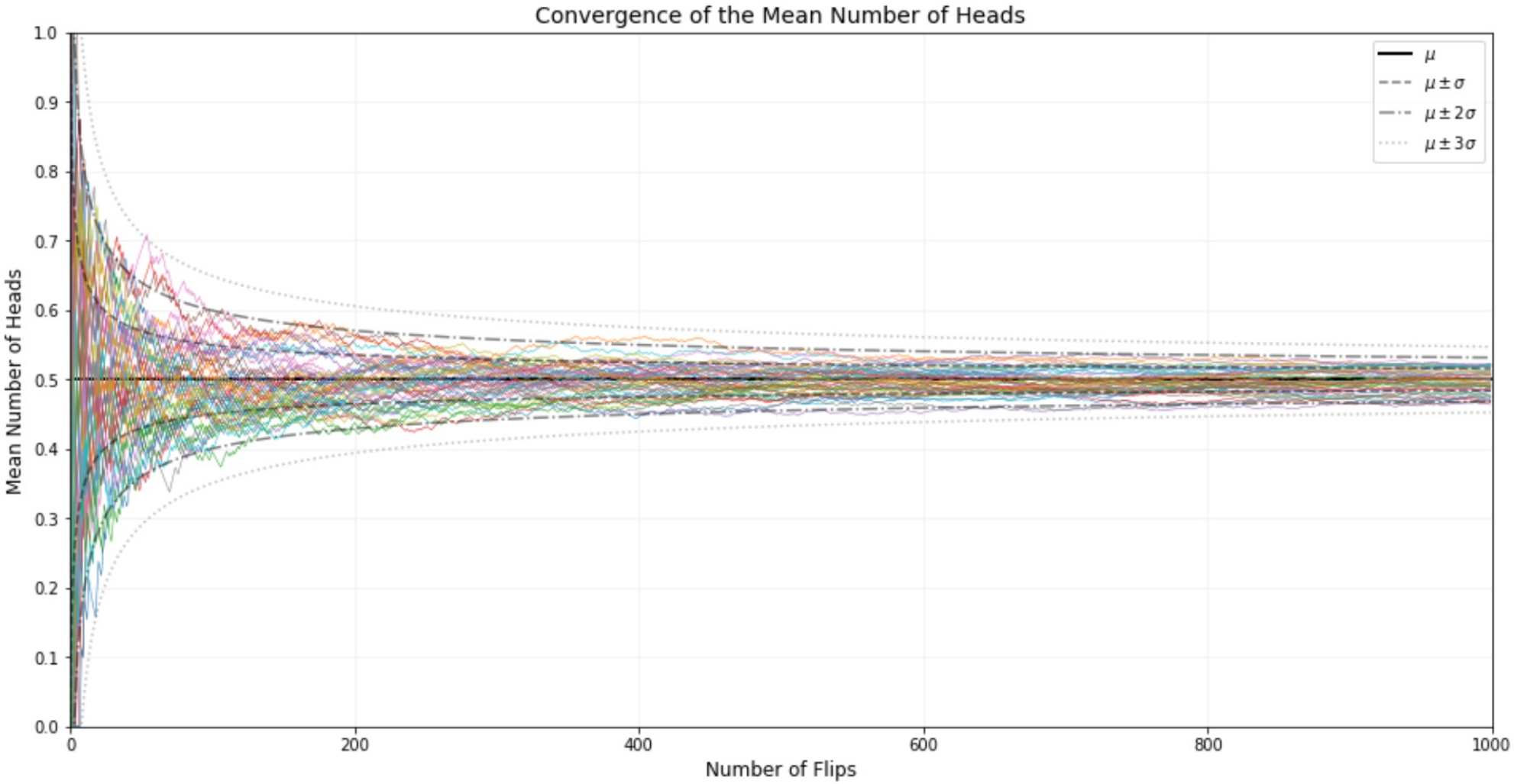
```
mu = 66
sigma = 3
n = 100                    # try for 1, 2, 5, 10, 30, 100
num_trials = 10000
display_sample_mean_normal(mu,sigma,n,num_trials,2)
```





Review: CLT and the Normal Distribution

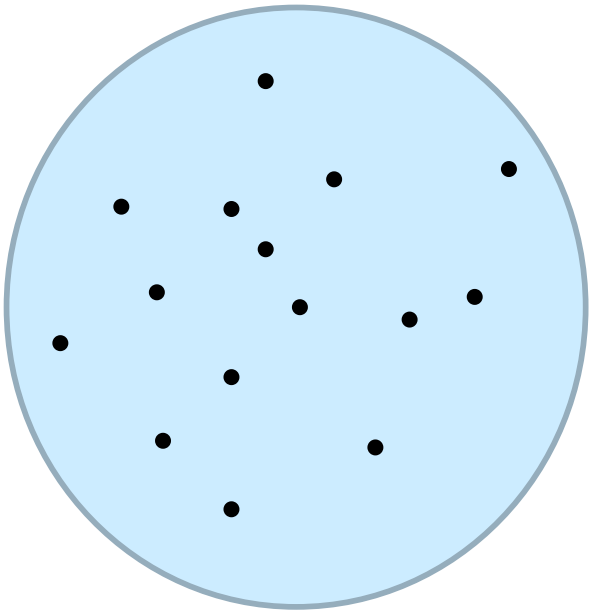
Graphically, you can see this in the experiment with flipping coins:



Sampling Theory

Recall: Sampling is the process of randomly selecting outcomes from a population, **which is really just a random variable**; the terminology for samples is slightly different for characteristics of the sample and population:

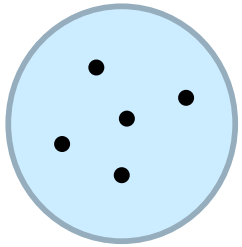
Population X



Randomly sample n outcomes

A "trial" is one such selection of n samples.

Sample of size n:



Sampling is generally done with replacement, but if the population is very large (perhaps infinite) it does not matter!

Population Parameters

mean	μ
variance	σ^2
standard deviation	σ

Sample Statistics

mean	$\bar{x}$
variance	s^2
standard deviation	s

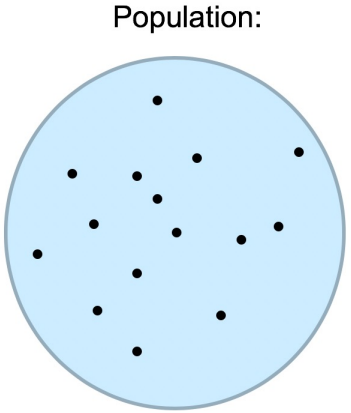
Sampling Theory

The sample statistics are estimators of the population parameters. They are also random variables (a function of the original random variable X). We will focus on the sample mean:

$$\bar{x} = \bar{x}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

In particular, we will use the CLT and focus on the sampling distribution of the sample mean, e.g.,

$$\bar{x} \sim N(66, (0.3)^2)$$



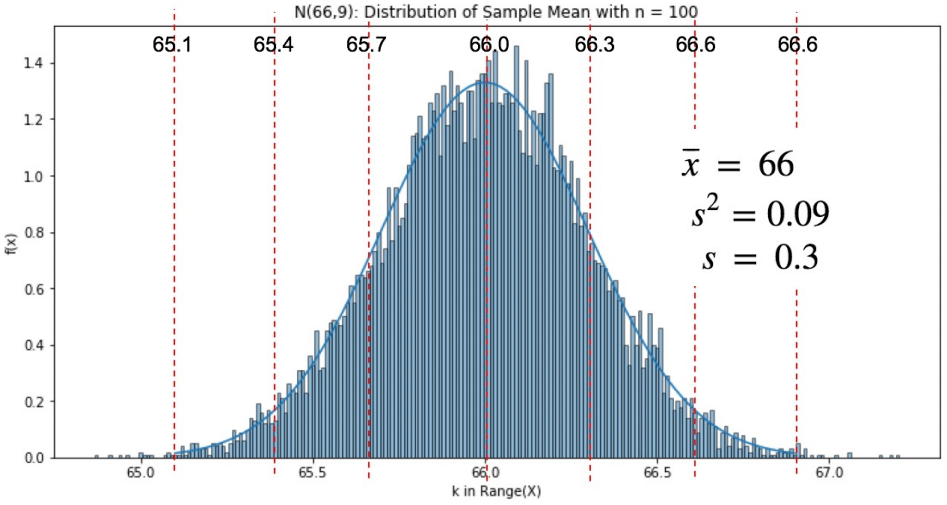
Population Parameters

mean	μ
variance	σ^2
standard deviation	σ

Sample Statistics

mean	$\bar{x}$
variance	s^2
standard deviation	s

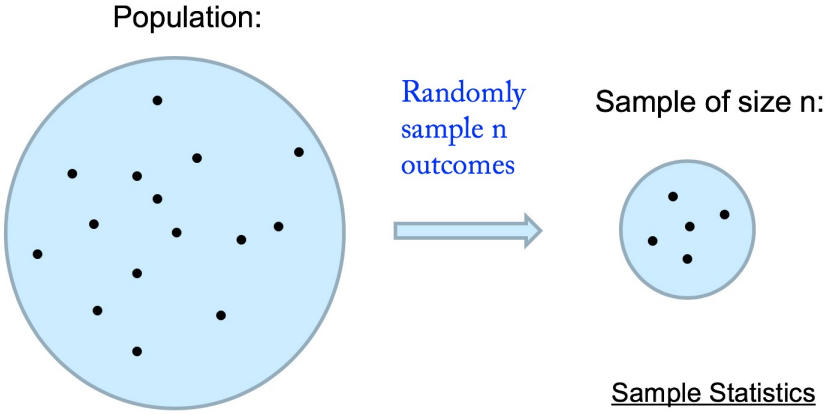
```
mu = 66
sigma = 3
n = 100                    # try for 1, 2, 5, 10, 30, 100
num_trials = 10000
display_sample_mean_normal(mu,sigma,n,num_trials,2)
```



Sampling Theory

Analogy: You want to know the height of BU students. Every day you select 100 students and measure them and take the mean. This is one trial (one "poke" of the sample mean random variable $\bar{x}$) and produces one number (a sample statistic). This sampling distribution of the sample mean is what results when you do 10,000 trials on 10,000 days, or 10,000 "pokes" of the sample mean random variable.

It's random variables, functions of random variables, and distributions all over again!



Population Parameters

mean	μ
variance	σ^2
standard deviation	σ

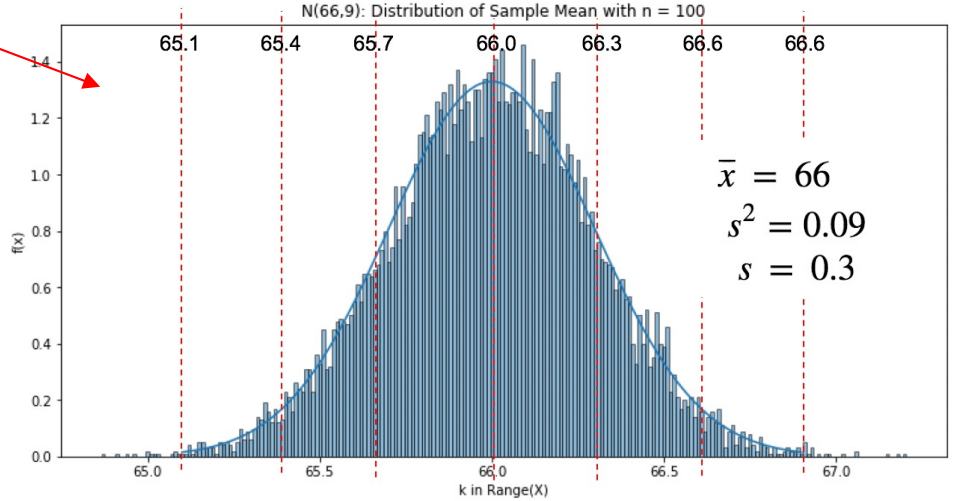
Sample Statistics

mean	$\bar{x}$
variance	s^2
standard deviation	s

$$\bar{x} = \bar{x}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

```
mu = 66
sigma = 3
n = 100
num_trials = 10000
display_sample_mean_normal(mu,sigma,n,num_trials,2)
```

$$\bar{x} \sim N(66, (0.3)^2)$$

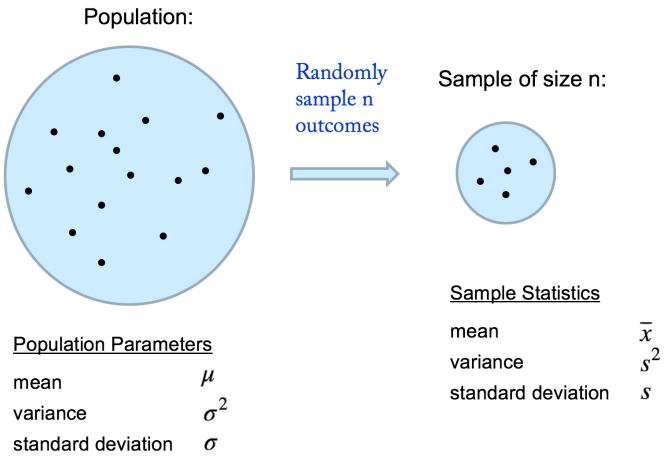


Sampling Theory When Population Parameters are Known

This is a warm-up to the real situation.....

Suppose (humor me!) that you have
the actual height data about all BU students,
including the mean and standard deviation,
but then you LOSE all the data, but somehow
you remember that the standard deviation is

σ = 3 inches.



$$\bar{x} = \bar{x}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

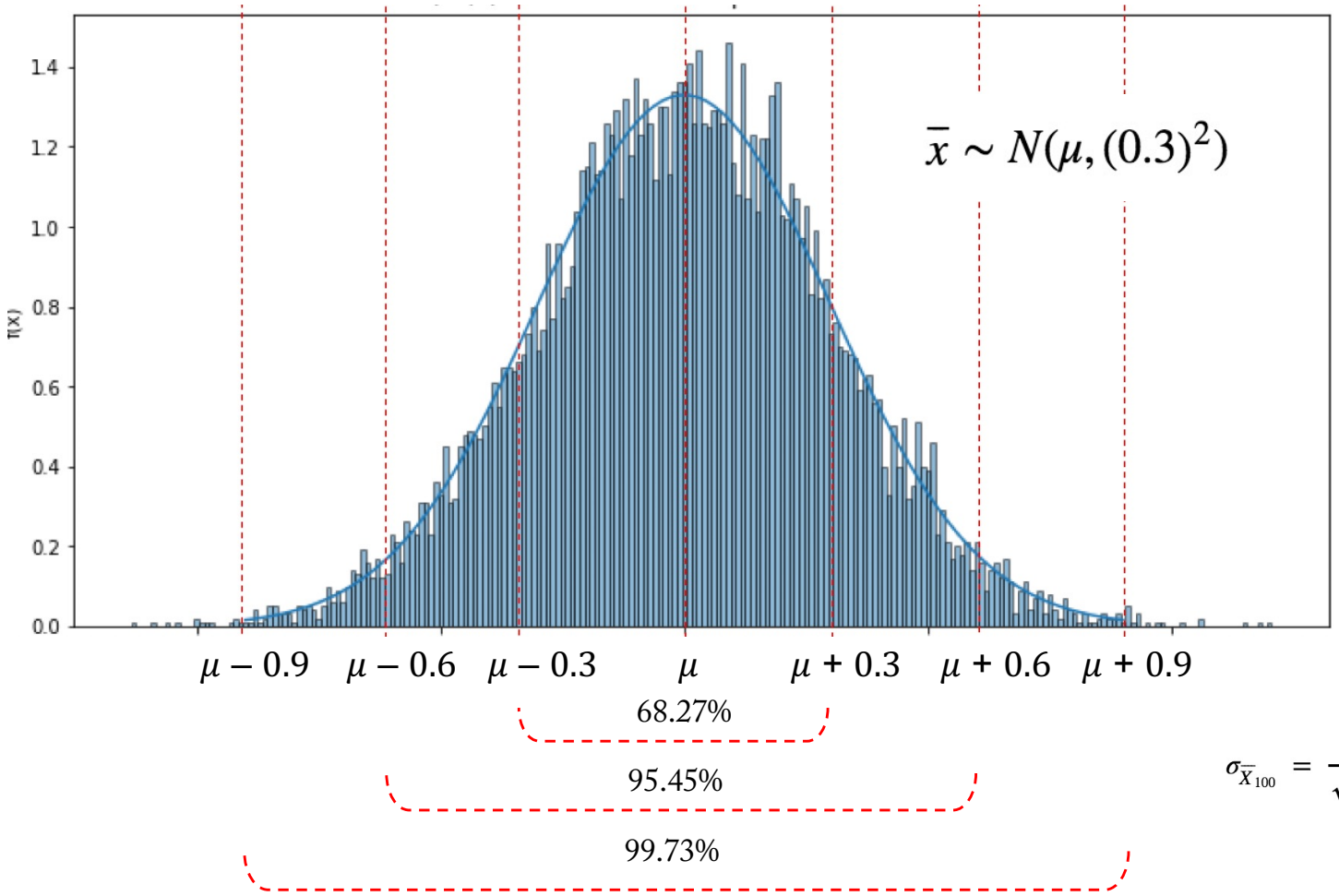
Furthermore, you need to the know the mean height, but you don't have a lot of time, and in any case you only need an approximation (an estimate) of the true mean μ.

What to do? Sample 100 randomly-selected students (**one trial**) and use the sample mean as your estimate! (Think polling: you ask 100 random people who they voted for.)

When you report your result, you have an estimate, and you can use the CLT to give precise information about how accurate your estimate is. This is called a Confidence Interval...

Confidence Intervals When Population Parameters are Known

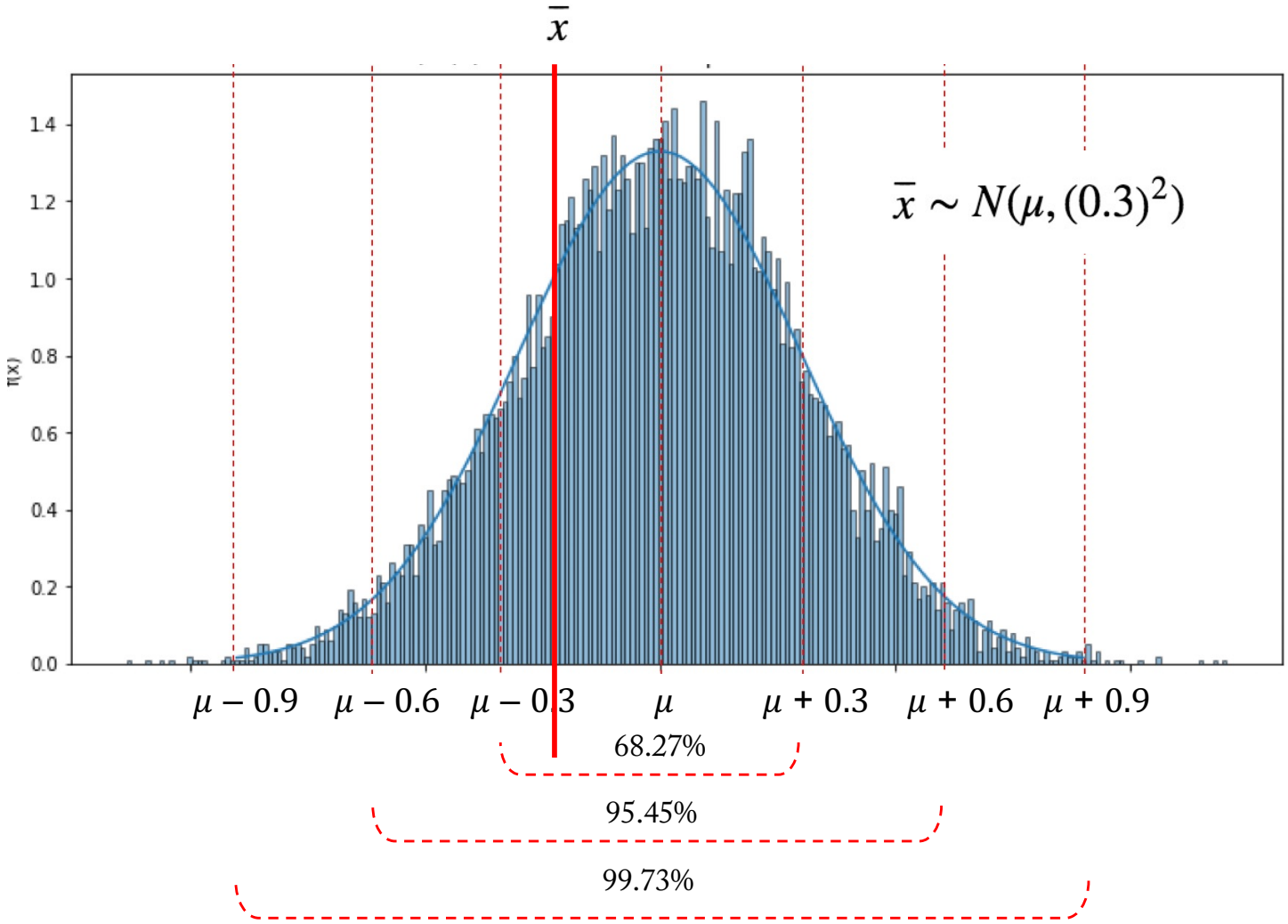
So you know that the actual standard deviation is σ = 3 inches and you want to estimate the unknown actual mean height μ by using one trial, one "poke" of the sample mean estimator $\bar{x}$, and you know by the CLT what the sampling distribution looks like.
You just don't know where the centerpoint μ is:



$$\sigma_{\bar{X}_{100}} = \frac{3}{\sqrt{100}} = \frac{3}{10} = 0.3$$

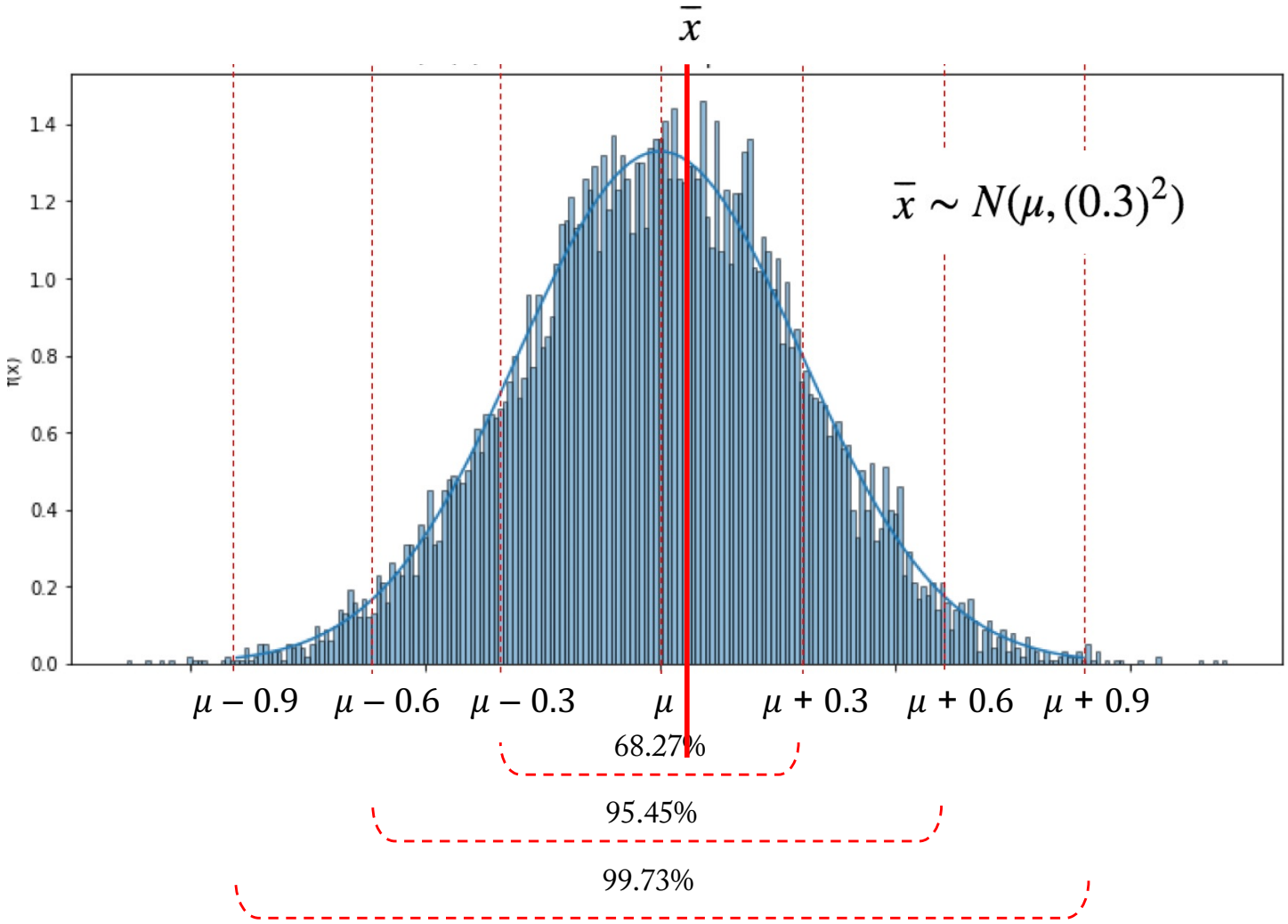
Confidence Intervals When Population Parameters are Known

But what you DO know is that whatever number you get for $\bar{x}$ from one trial of measuring 100 students, you have 68.27% chance of being within 0.3 inches of the true mean, 95.45% chance of being within 0.6 inches, and 99.73% of being within 0.9 inches:



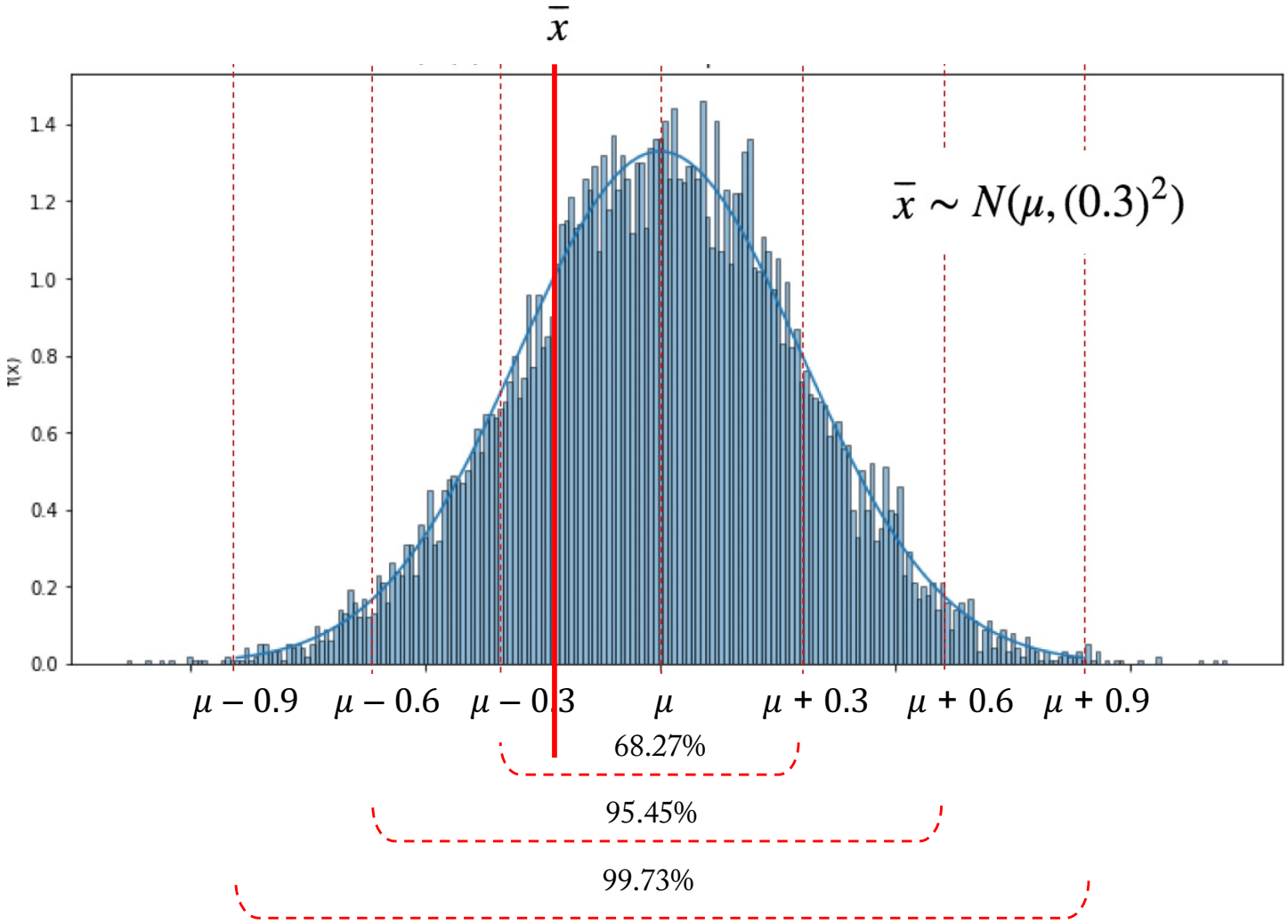
Confidence Intervals When Population Parameters are Known

But what you DO know is that whatever number you get for $\bar{x}$ from one trial of measuring 100 students, you have 68.27% chance of being within 0.3 inches of the true mean, 95.45% chance of being within 0.6 inches, and 99.73% of being within 0.9 inches:



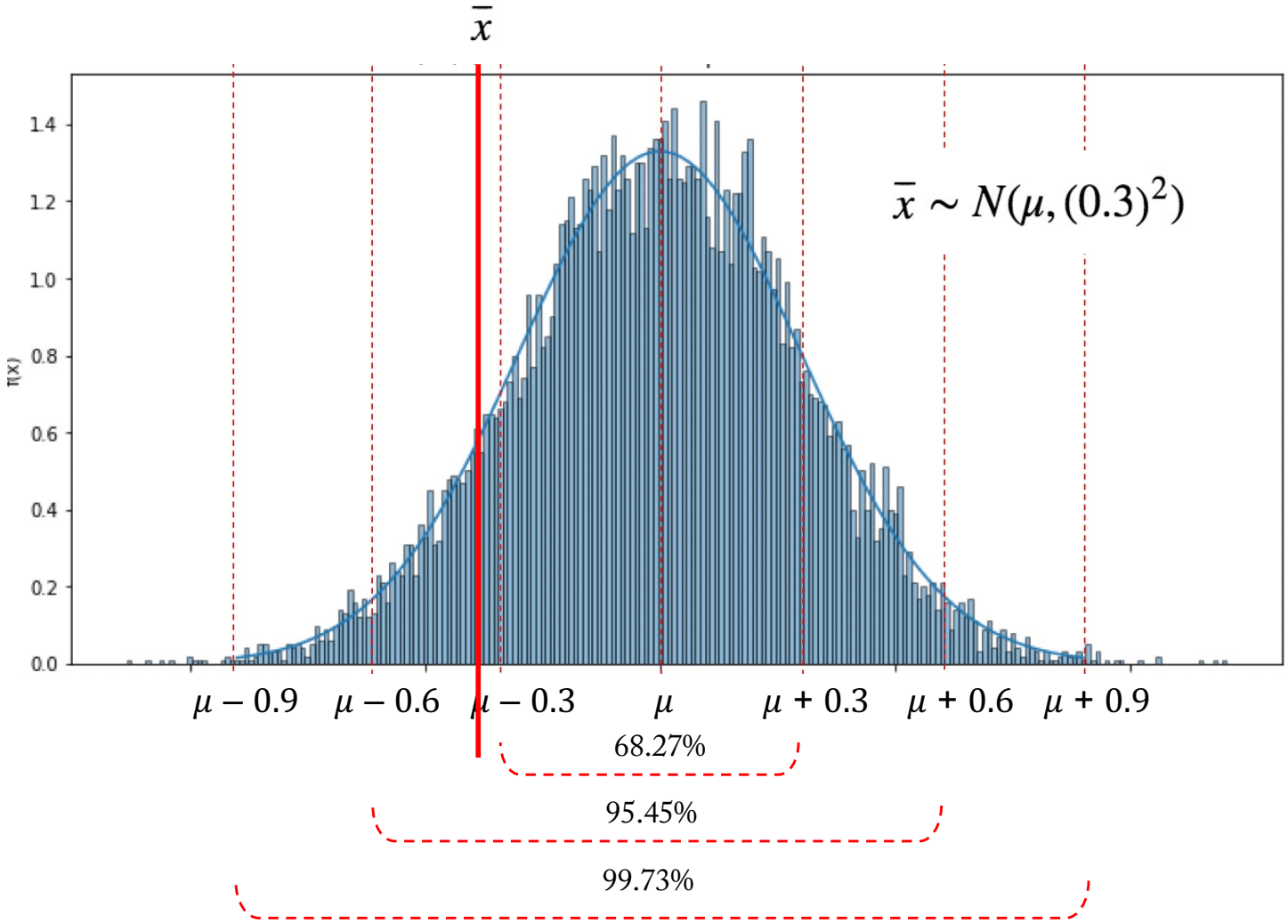
Confidence Intervals When Population Parameters are Known

But what you DO know is that whatever number you get for $\bar{x}$ from one trial of measuring 100 students, you have 68.27% chance of being within 0.3 inches of the true mean, 95.45% chance of being within 0.6 inches, and 99.73% of being within 0.9 inches:



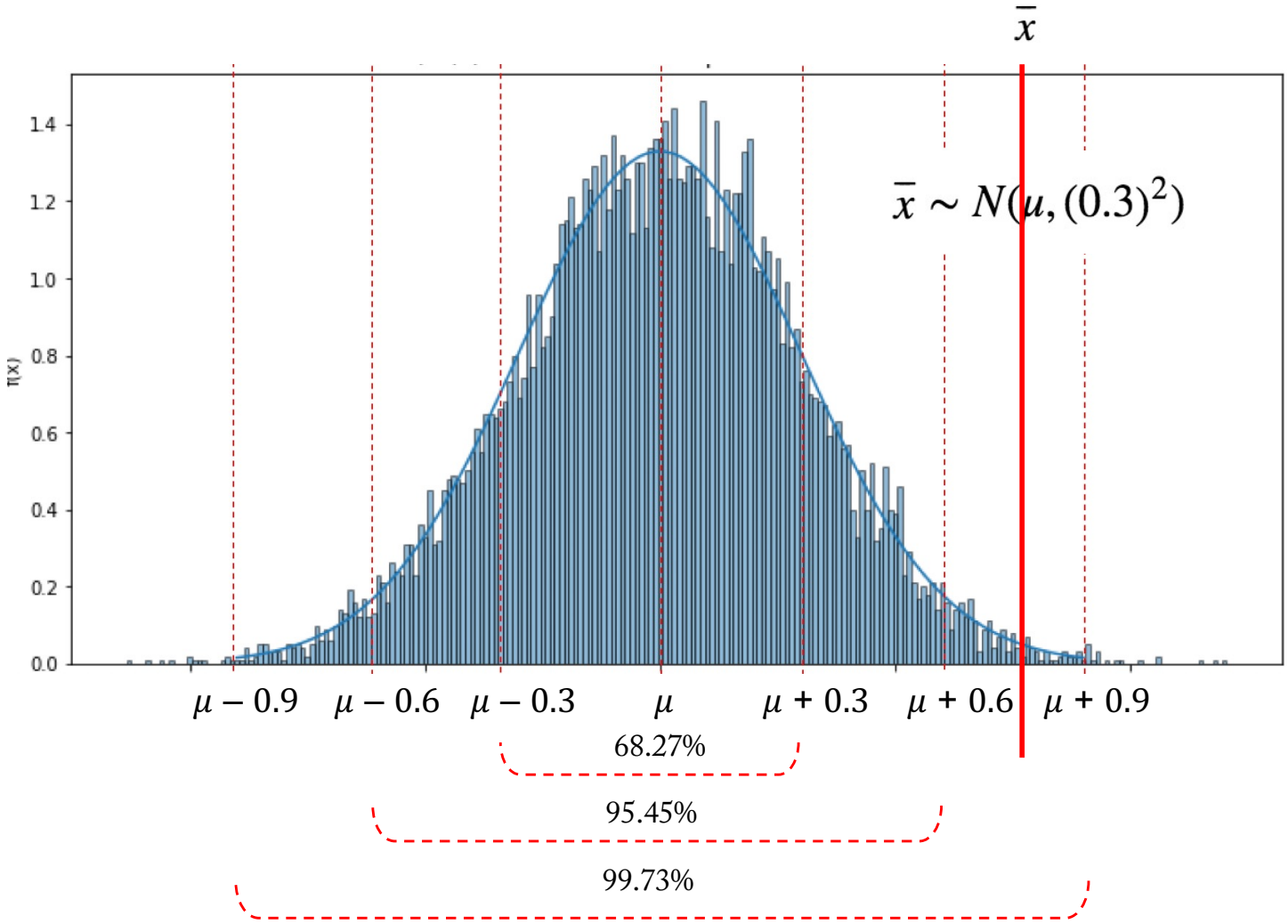
Confidence Intervals When Population Parameters are Known

But what you DO know is that whatever number you get for $\bar{x}$ from one trial of measuring 100 students, you have 68.27% chance of being within 0.3 inches of the true mean, 95.45% chance of being within 0.6 inches, and 99.73% of being within 0.9 inches:



Confidence Intervals When Population Parameters are Known

But what you DO know is that whatever number you get for $\bar{x}$ from one trial of measuring 100 students, you have 68.27% chance of being within 0.3 inches of the true mean, 95.45% chance of being within 0.6 inches, and 99.73% of being within 0.9 inches:



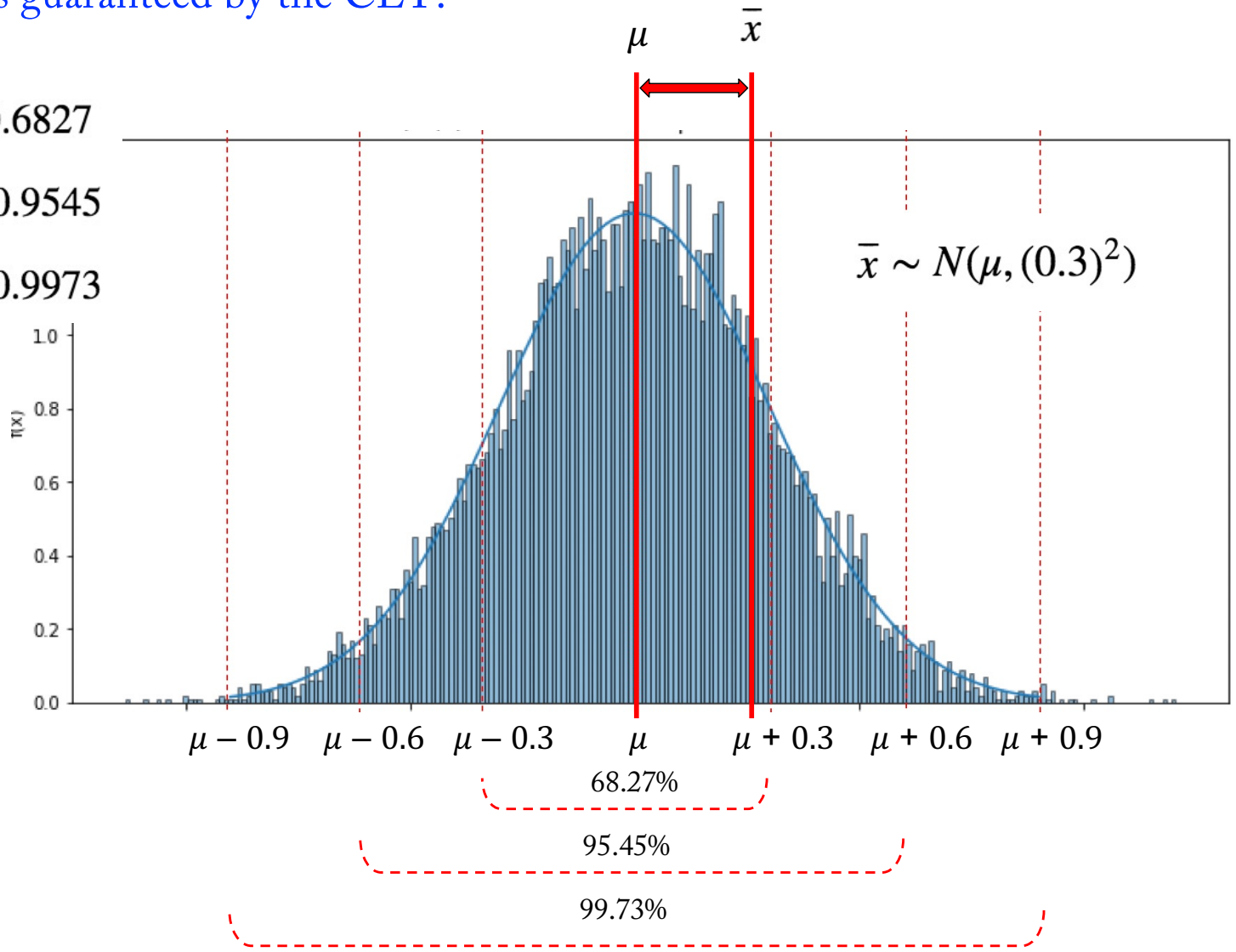
Confidence Intervals When Population Parameters are Known

But notice that what we are really talking about is the probability of the distance $|\bar{x} - \mu|$ being within bounds guaranteed by the CLT:

$$P(|\bar{x} - \mu| \leq \sigma) = 0.6827$$

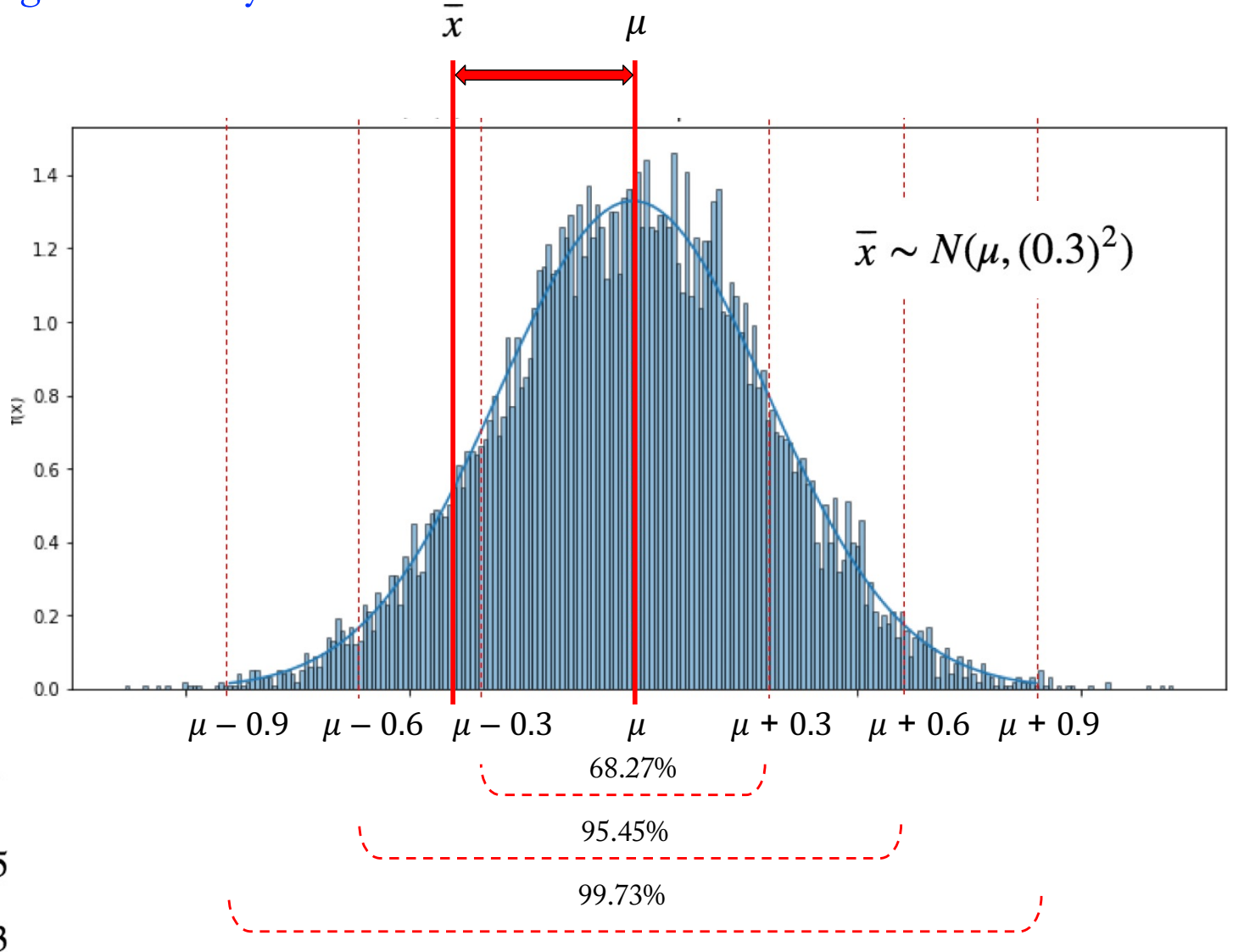
$$P(|\bar{x} - \mu| \leq 2\sigma) = 0.9545$$

$$P(|\bar{x} - \mu| \leq 3\sigma) = 0.9973$$



Confidence Intervals When Population Parameters are Known

But notice that what we are really talking about is the probability of the distance $|\bar{x} - \mu|$ being within bounds guaranteed by the CLT:



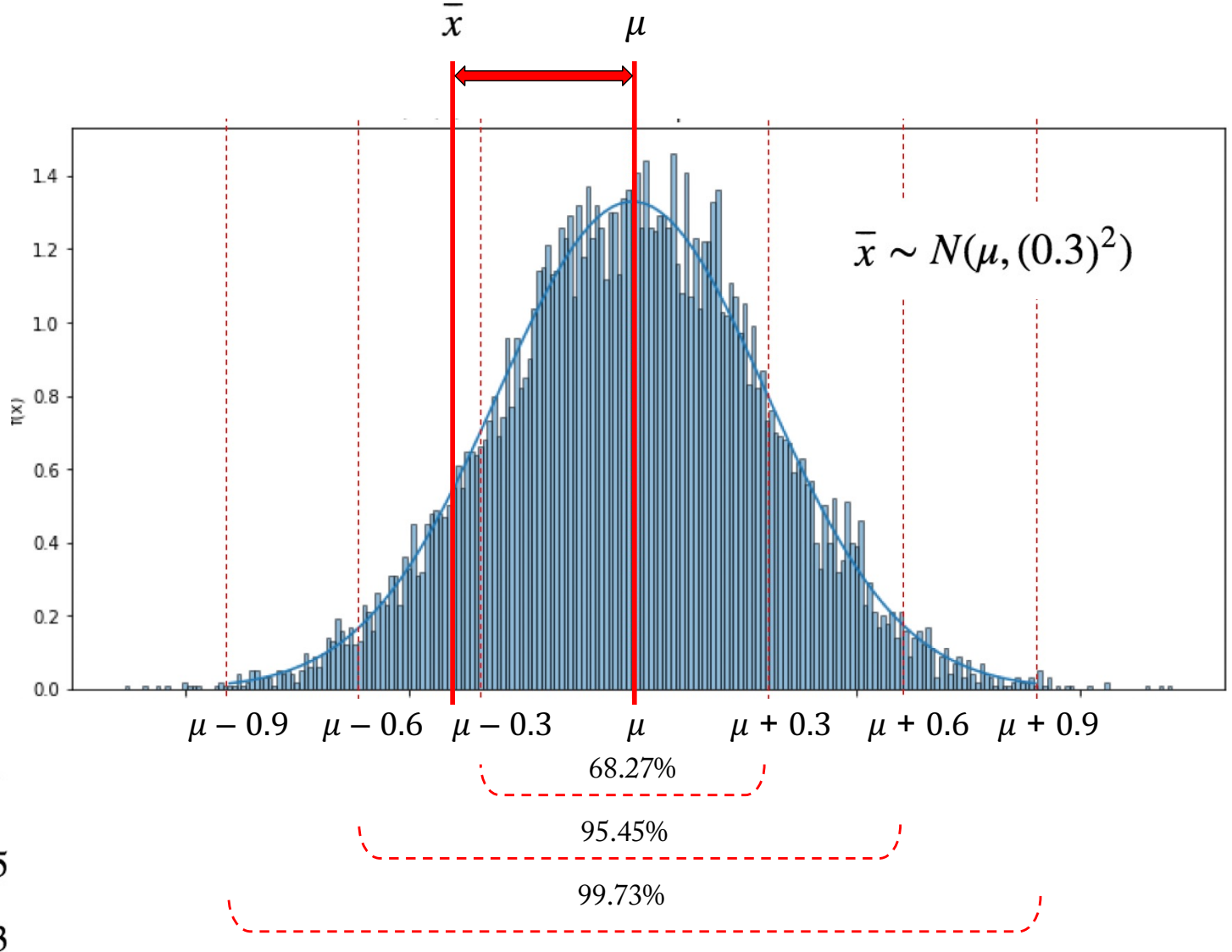
$$P(|\bar{x} - \mu| \leq \sigma) = 0.6827$$

$$P(|\bar{x} - \mu| \leq 2\sigma) = 0.9545$$

$$P(|\bar{x} - \mu| \leq 3\sigma) = 0.9973$$

Confidence Intervals When Population Parameters are Known

But then because the normal is symmetric, it does not matter if we change our perspective to use a sampling distribution centered on μ or on $\bar{x}$:



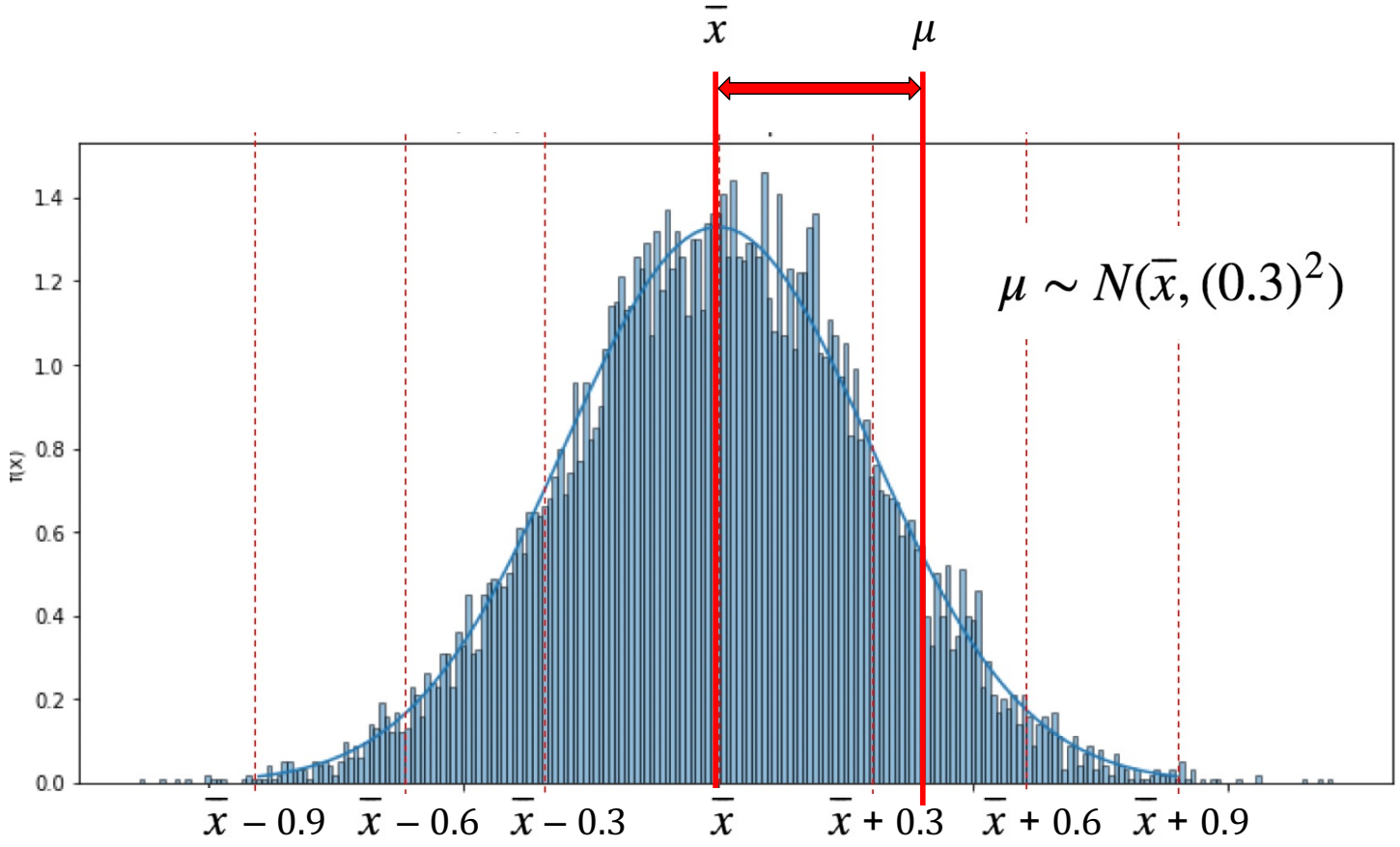
$$P(|\bar{x} - \mu| \leq \sigma) = 0.6827$$

$$P(|\bar{x} - \mu| \leq 2\sigma) = 0.9545$$

$$P(|\bar{x} - \mu| \leq 3\sigma) = 0.9973$$

Confidence Intervals When Population Parameters are Known

But then because the normal is symmetric, it does not matter if we change our perspective to use a sampling distribution centered on μ or on $\bar{x}$:



$$P(|\bar{x} - \mu| \leq \sigma) = 0.6827$$

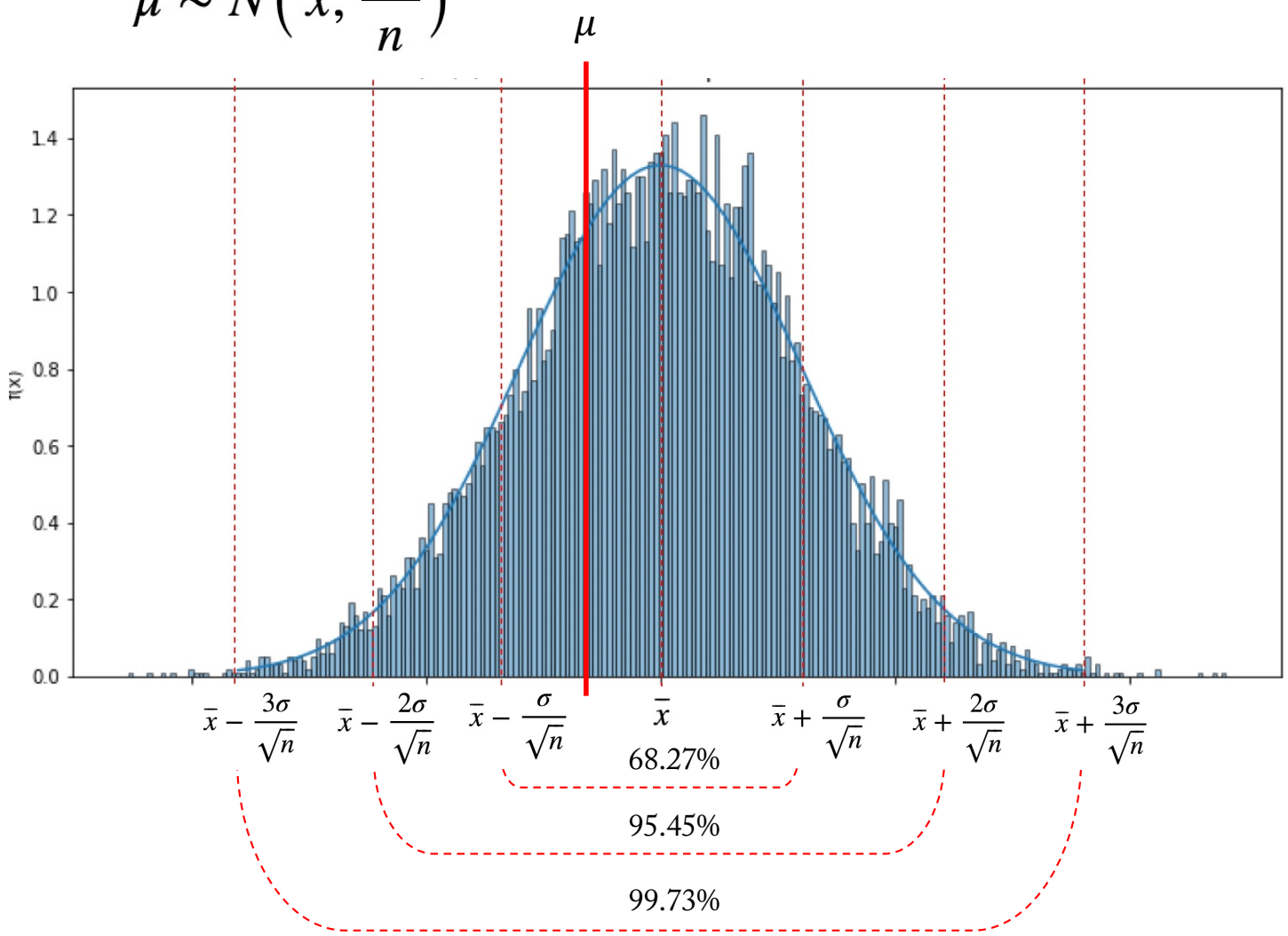
$$P(|\bar{x} - \mu| \leq 2\sigma) = 0.9545$$

$$P(|\bar{x} - \mu| \leq 3\sigma) = 0.9973$$

Confidence Intervals When Population Parameters are Known

So we can **pretend** that the population mean is normally distributed around the sample mean (not true in general, but for one sample, it is effectively the same thing).

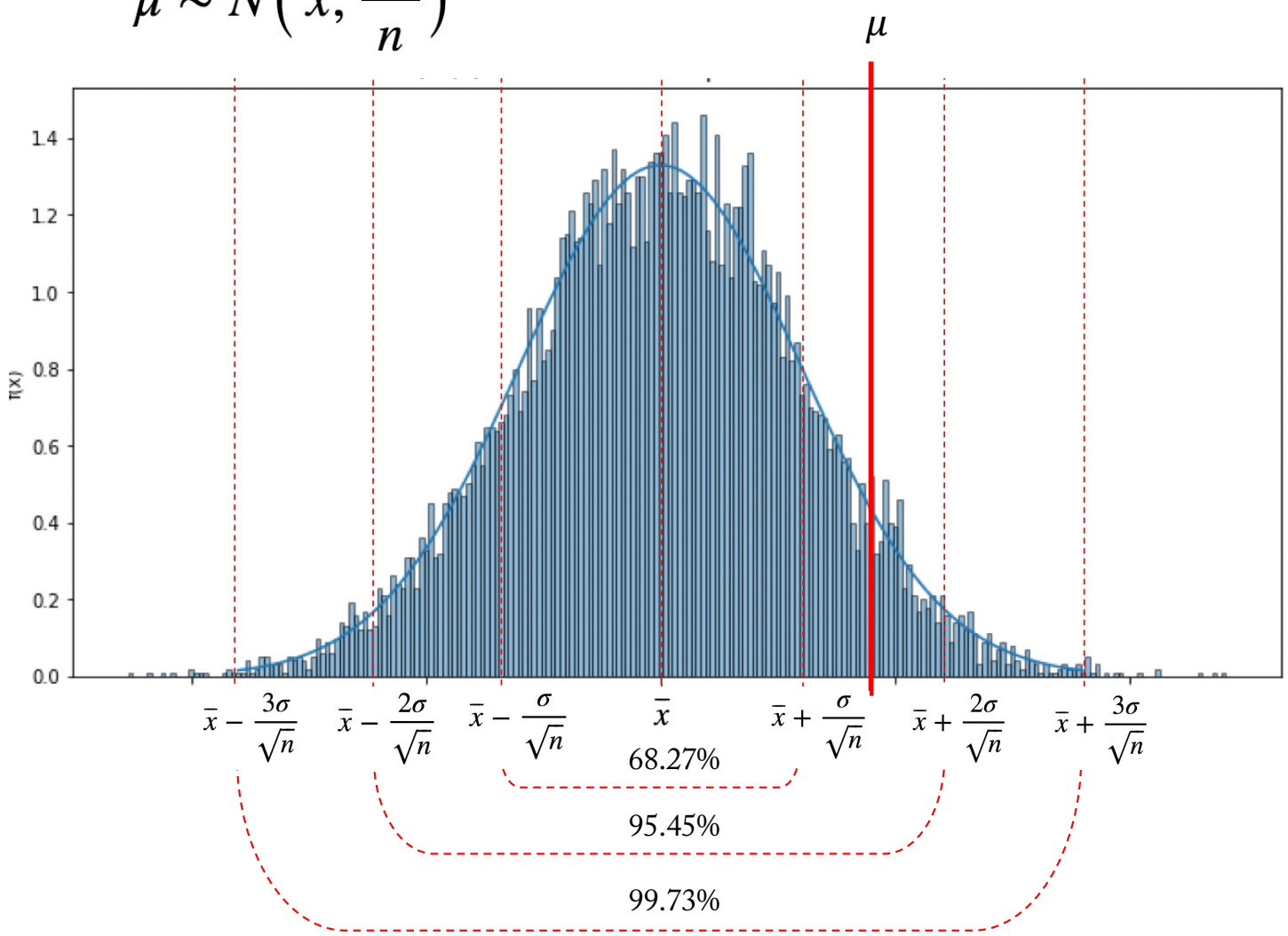
$$\mu \sim N\left(\bar{x}, \frac{\sigma^2}{n}\right)$$



Confidence Intervals When Population Parameters are Known

So we can **pretend** that the population mean is normally distributed around the sample mean (not true in general, but for one sample, it is effectively the same thing).

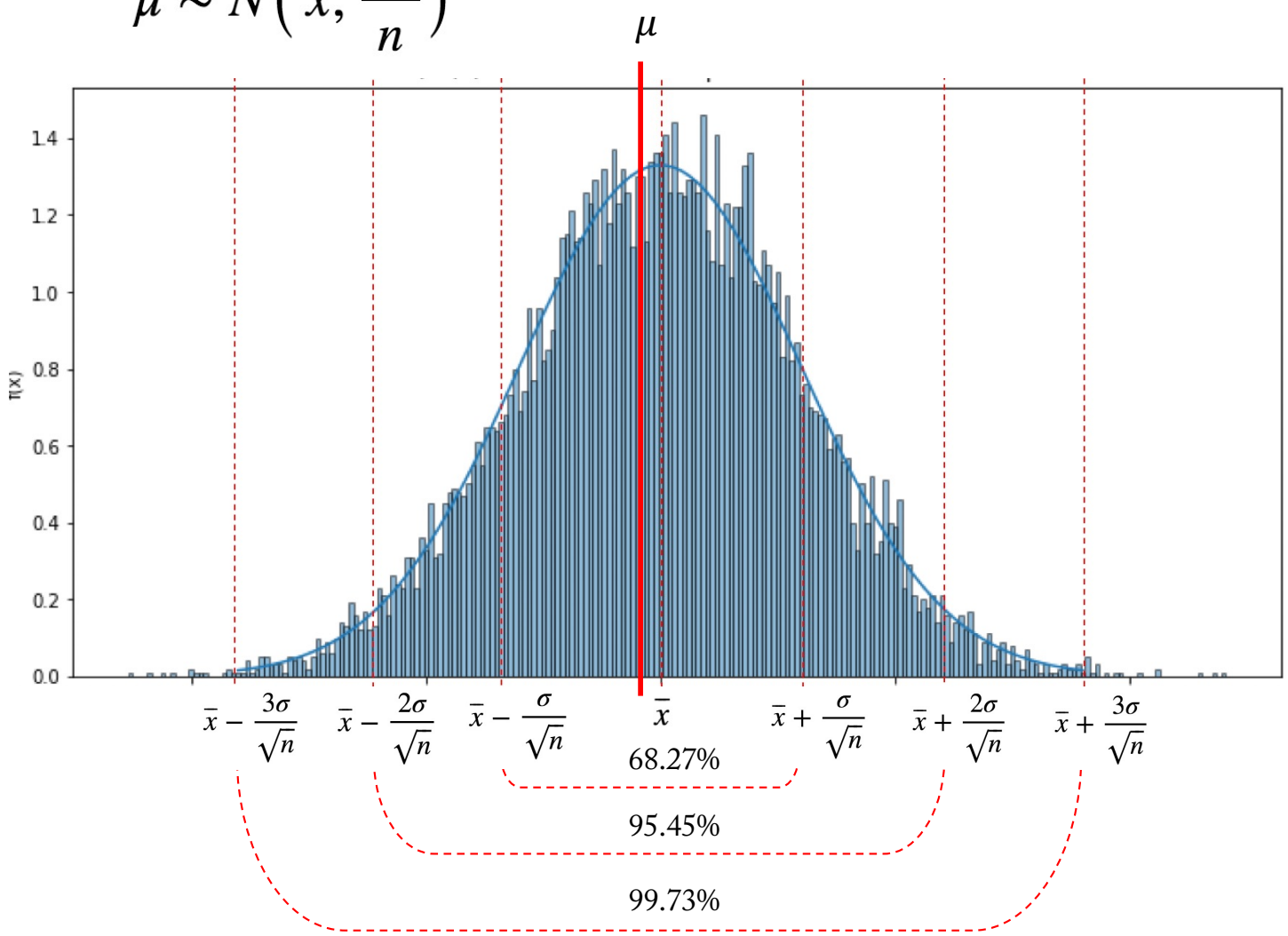
$$\mu \sim N\left(\bar{x}, \frac{\sigma^2}{n}\right)$$



Confidence Intervals When Population Parameters are Known

So we can **pretend** that the population mean is normally distributed around the sample mean (not true in general, but for one sample, it is effectively the same thing).

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



Confidence Intervals When the Population Std Dev is Known

Confidence Intervals Using the Population Standard Deviation:

Let σ be the standard deviation of the population…..

Then:

1. Choose a sample size n;
2. Calculate the standard deviation of the sample mean: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
3. Choose a confidence level CL (e.g., 95.45%);
4. Calculate the multiplier k for s corresponding to $CL = P(\mu - k \cdot \sigma_{\bar{x}} \leq \bar{x} \leq \mu + k \cdot \sigma_{\bar{x}})$
5. Perform random sampling of n samples and calculate $\bar{x}$
6. Report your results using the confidence interval corresponding to CL:

"The mean of the population is $\bar{x} \pm k \cdot \sigma_{\bar{x}}$ with a confidence of CL."

```
In [3]:  1  norm.interval(alpha=0.95,loc=0,scale=1)
Out[3]: (-1.959963984540054, 1.959963984540054)
```

Confidence Intervals When Population Parameters are Known

Example -- Height of BU Students:

Suppose we know that the height of BU students has standard deviation σ = 3 inches.

1. Choose a sample size n = 100;
2. σ_x = 0.3 inches
3. Choose a confidence level CL = 95.45%;
4. Calculate the multiplier k = 2;
5. Perform random sampling of 100 students and calculate $\bar{x}$ = 66.134 inches ;
6. Report your results using the confidence interval corresponding to CL:

"The mean height of BU students is 66.134 +/- 0.6 inches with a confidence of 95.45%."

or change the confidence level if you wish:

"The mean height of BU students is 66.134 +/- 0.9 inches with a confidence of 99.73%."

Confidence Intervals When Population Parameters are Known

Caveat: There is a one-to-one correspondence between confidence levels and k, but unfortunately these do not correspond to nice, round numbers on each side. So just be aware of whether you want, for example, "two standard deviations" or "95%" (which are different). Also realize that "95.45%" is an approximation of "two standard deviations":

```
#c. Find P(-k<X<k) for standard normal
CL = norm.cdf(x=2,loc=0,scale=1) - norm.cdf(x=-2,loc=0,scale=1)
print("CL for k = 2:  " + str(CL))
CL = norm.cdf(x=3,loc=0,scale=1) - norm.cdf(x=-3,loc=0,scale=1)
print("CL for k = 3:  " + str(CL))

#f give the endpoints of the range for the central alpha percent
#  of the distribution
print("\n90%: " + str(norm.interval(alpha=0.90, loc=0, scale=1)))
print("95%: " + str(norm.interval(alpha=0.95, loc=0, scale=1)))
print("99%: " + str(norm.interval(alpha=0.99, loc=0, scale=1)))
```

```
CL for k = 2:  0.954499736104
CL for k = 3:  0.997300203937

90%: (-1.6448536269514729, 1.6448536269514722)
95%: (-1.959963984540054, 1.959963984540054)
99%: (-2.5758293035489004, 2.5758293035489004)
```

Sampling When the Population Parameters are Unknown

When the population parameters (mean, standard deviation) are unknown, you have no choice but to use the standard deviation of the sample in place of the (unknown) standard deviation of the population.

There are **three** important cases to consider:

First, you can use the standard deviation of the sample when n > 30 (large samples).

Second, when the population is Bernoulli (yes/no, male/female,1/0, vote for A/vote for B), then the standard deviation is derived from the mean of the sample using the formulae:

$$X \sim Bernoulli(p)$$

$$p = \bar{x} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

$$s = \sqrt{p(1-p)}$$

Then you divide as usual to find the std of the sample mean:

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

This is called **Sampling with Proportions** in most textbooks. You can think of it as sampling a Bernoulli, and simply use the parameters above in the preceding results with the normal distribution, OR you can think of the whole sample as a Binomial, and use the Binomial directly.

Sampling When the Population Parameters are Unknown

When the population parameters (mean, standard deviation) are unknown, you have no choice but to use the standard deviation of the sample in place of the (unknown) standard deviation of the population.

There are **three** important cases to consider:

First, you can use the standard deviation of the sample when n > 30 (large samples).

Second, when sampling proportions, use $s = \sqrt{\bar{x}(1-\bar{x})}$.

Third, when sampling with n <= 30 from a population known to be Normal, but with unknown mean and standard deviation, you can use a slightly different formula for the sample standard deviation and a slightly different distribution, called the T-Distribution.