CS 237: Probability in Computing

Wayne Snyder
Computer Science Department
Boston University

Lecture 14:

- Statistics = Applications of the Central Limit Theorem
- Sampling Theory
- Point Estimates: warmup -- when the population parameters are known
- Confidence Intervals -- when the population parameters are known
- [If time] Introduction to Hypothesis Testing

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

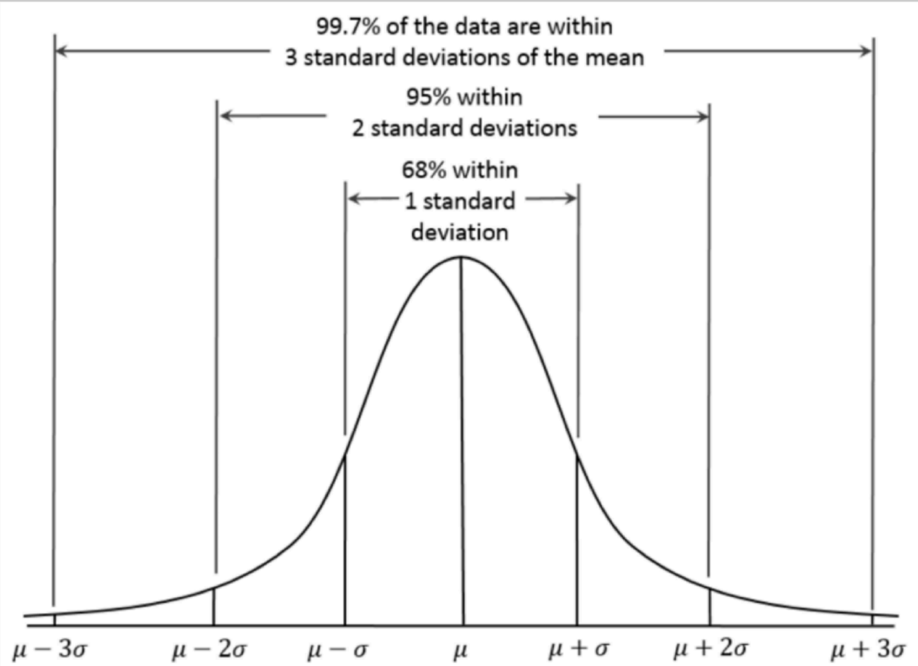
> As n gets large, the random variable $\overline{X}_n$ converges to the distribution $N\left(\mu, \frac{\sigma^2}{n}\right)$.

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
```

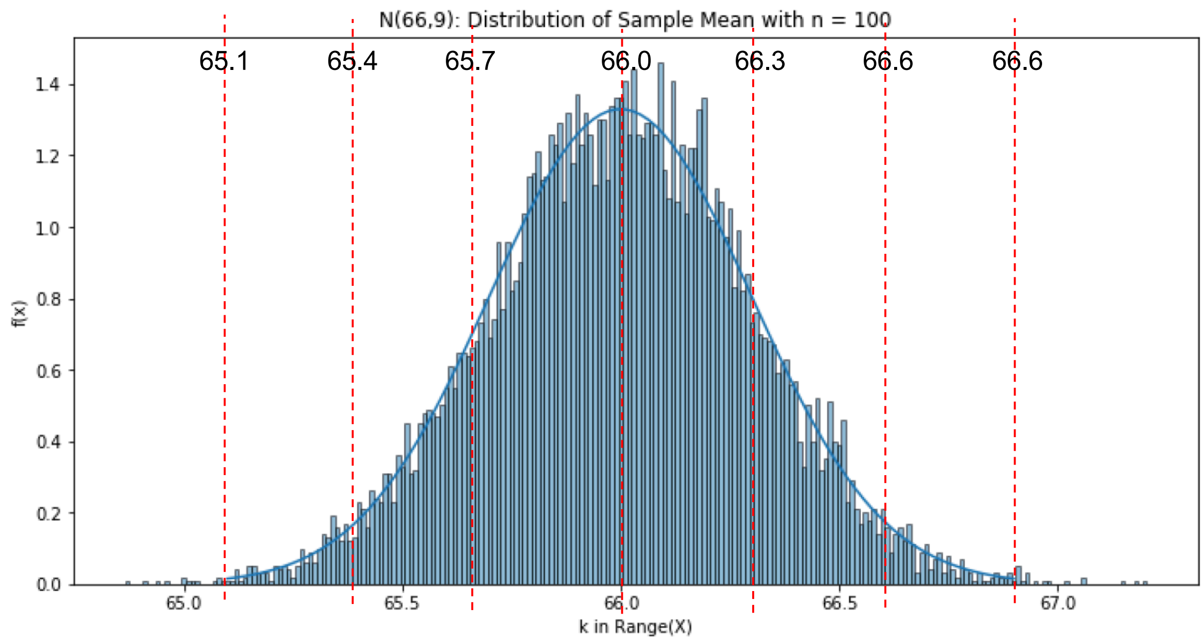
Review: CLT and the Normal Distribution

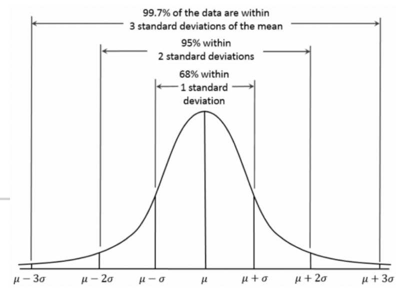
Example: Let X ~ N(66,3^2). We calculated the mean for n = 100, so we should get a standard deviation smaller by a factor of 10:

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

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

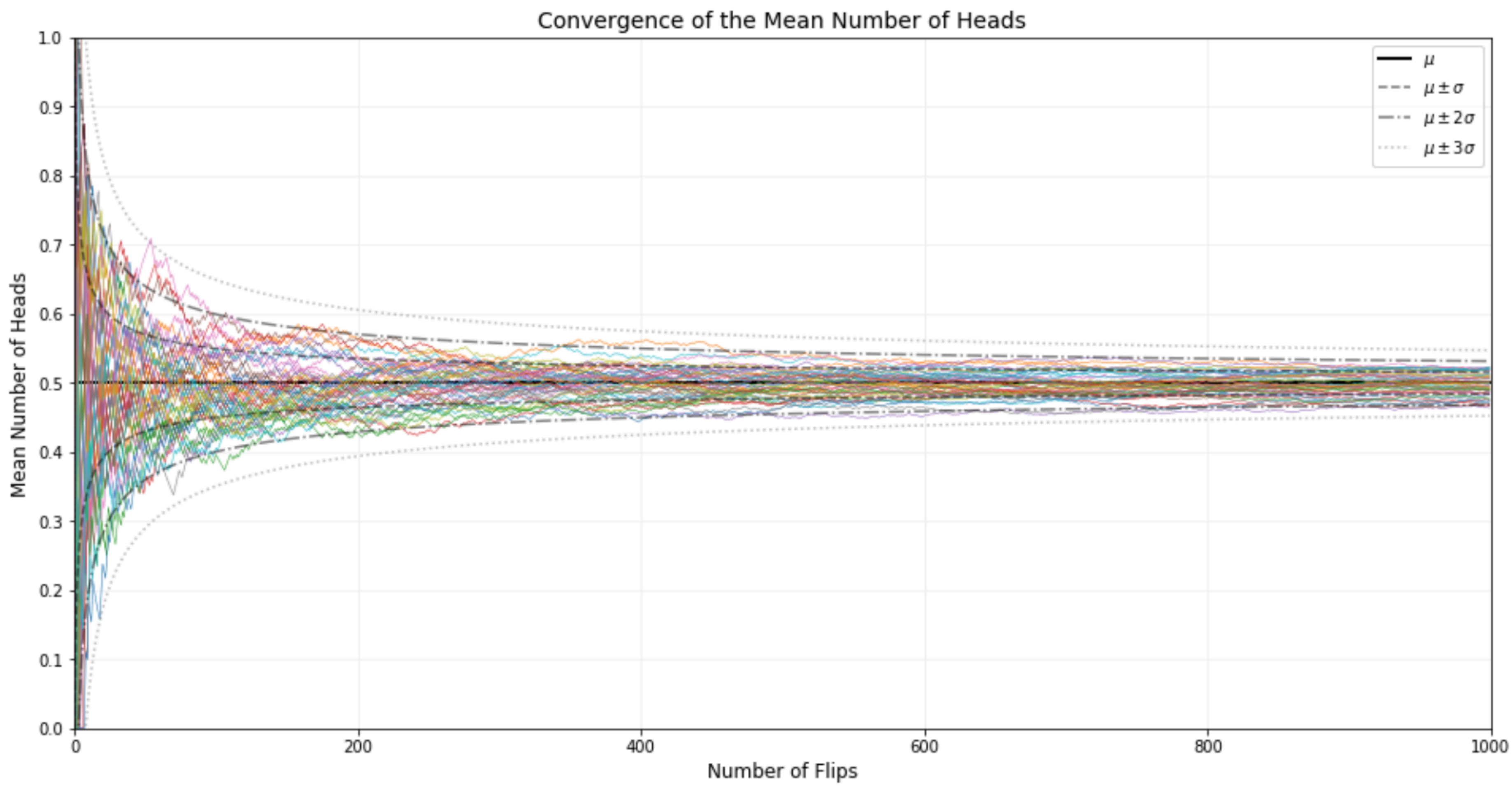
```
mu = 66
sigma = 3
n = 100                    # try for 1, 2, 5, 10, 30, 100
num_trials = 10000
display_sample_mean_normal(mu,sigma,n,num_trials,2)
```





Review: CLT and the Normal Distribution

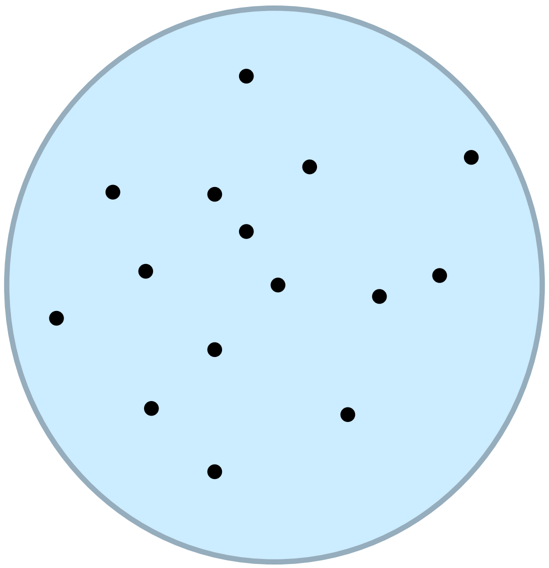
Graphically, you can see this in the experiment with flipping coins:



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:

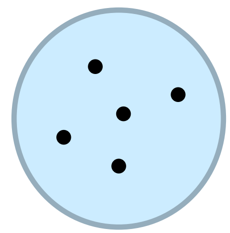
Population X



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

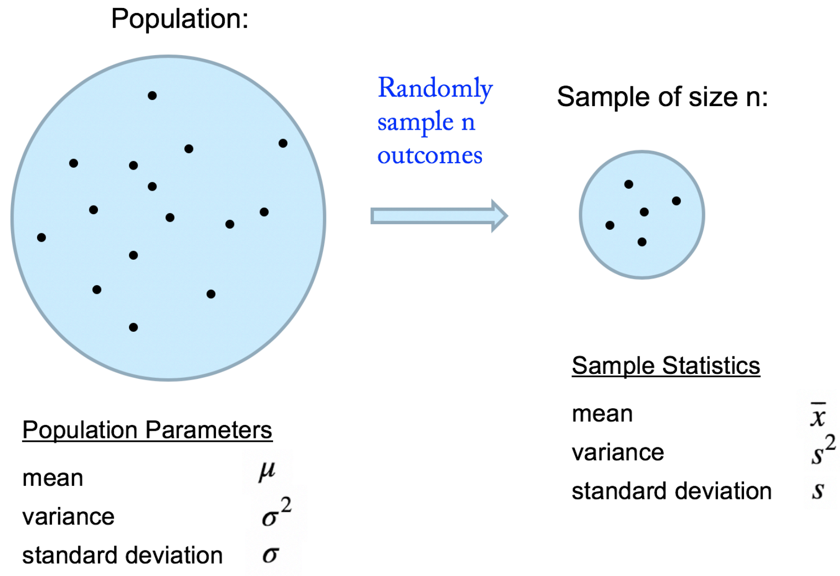
Sampling Theory

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:

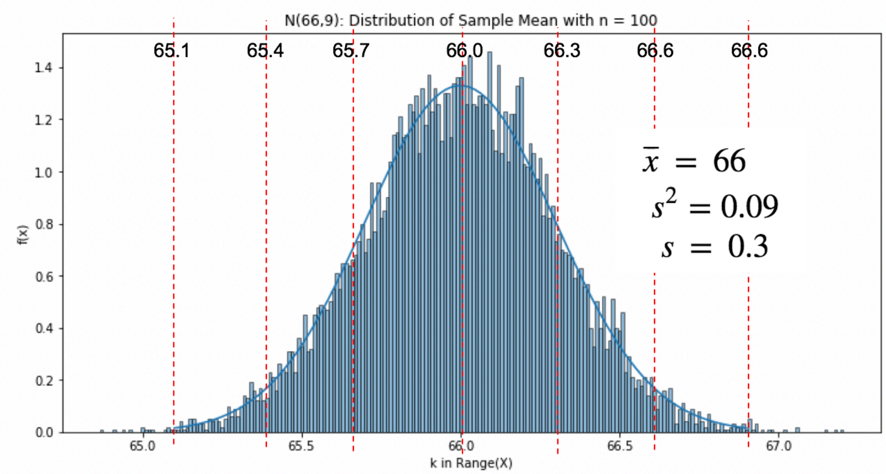
$$\bar{x} = \bar{x}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

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

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



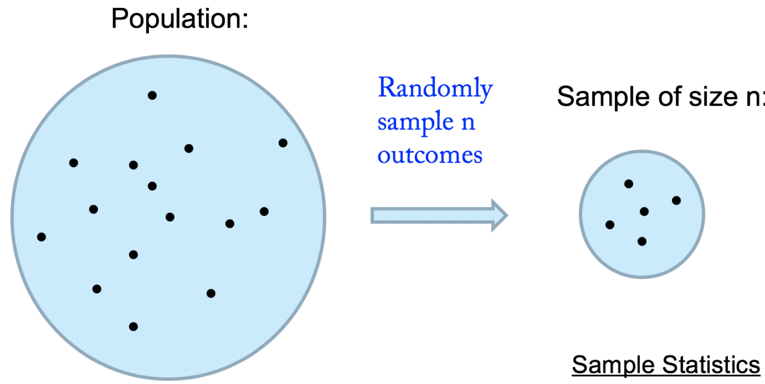

```
mu = 66
sigma = 3
n = 100              # try for 1, 2, 5, 10, 30, 100
num_trials = 10000
display_sample_mean_normal(mu,sigma,n,num_trials,2)
```



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 varial $\bar{x}$) 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 distributic all over again!



<u>Population Parameters</u>

mean	μ
variance	σ^2
standard deviation	σ

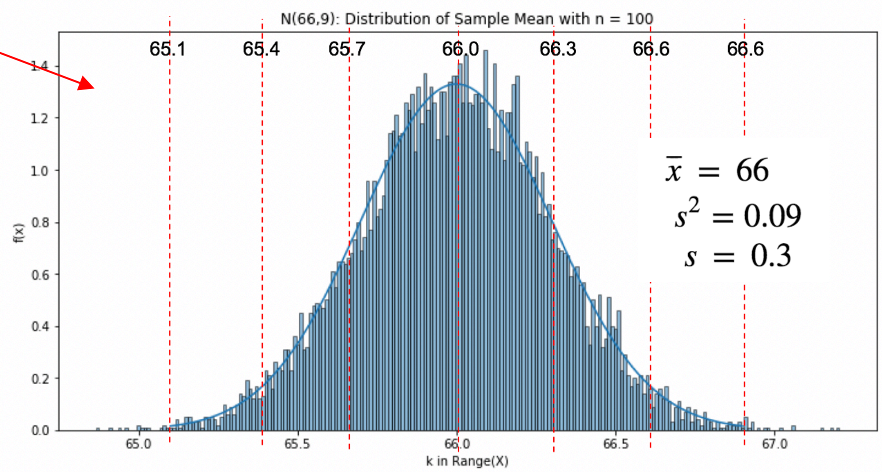
<u>Sample Statistics</u>

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)
```

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

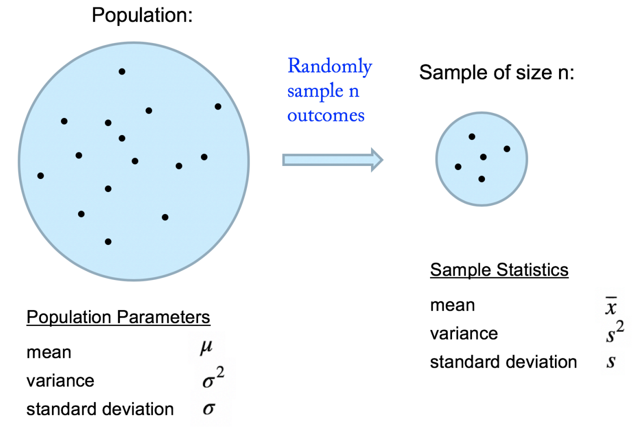


$$\bar{x} = 66$$
$$s^2 = 0.09$$
$$s = 0.3$$

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



$$\sigma = 3 \text{ inches.}$$

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

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.

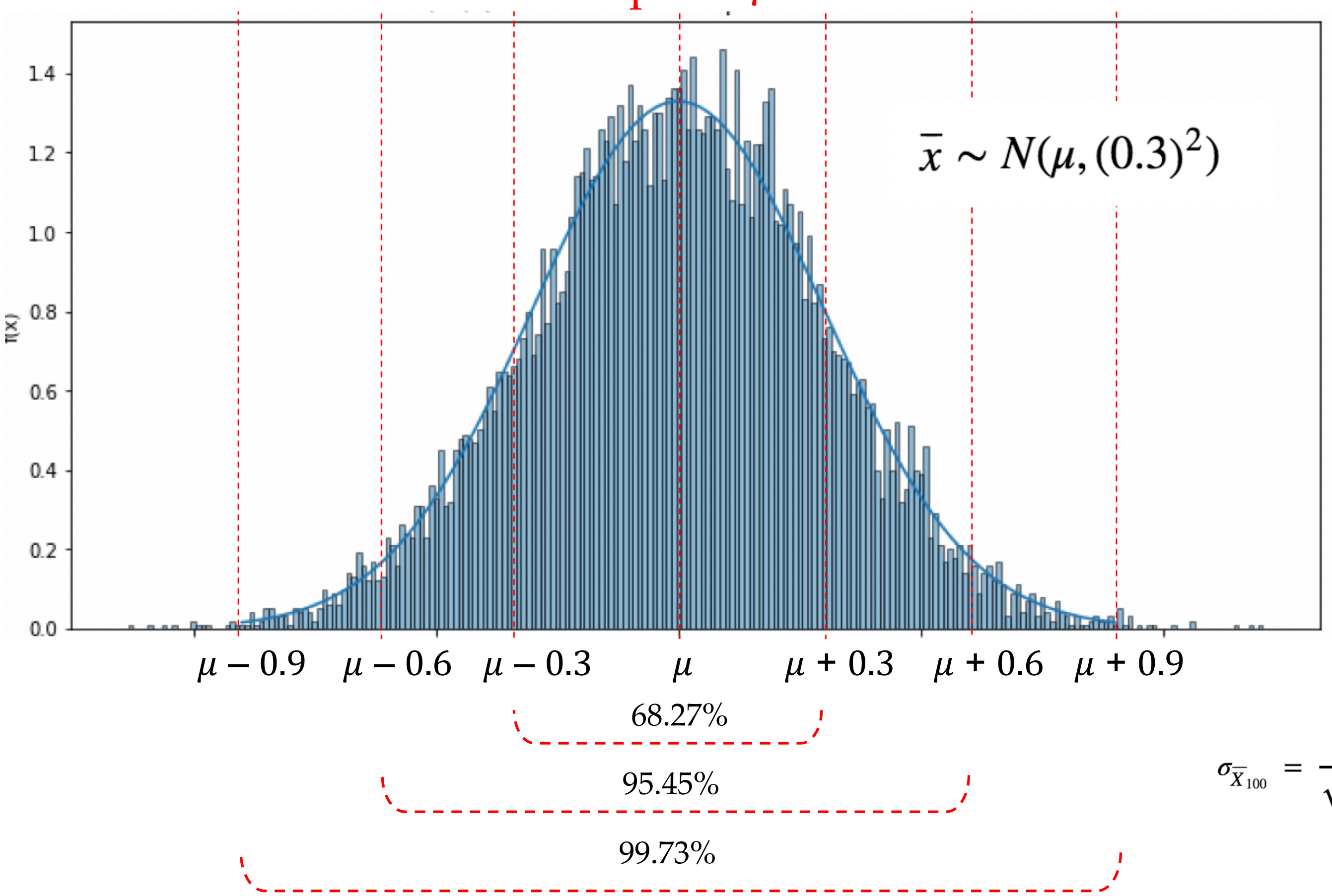
What to do? Sample 100 randomly-selected students (**one trial**) and use the sample mean as your estimate! (Think polling: you ask 100 random people who they voted for.)

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 $\sigma = 3$ 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.

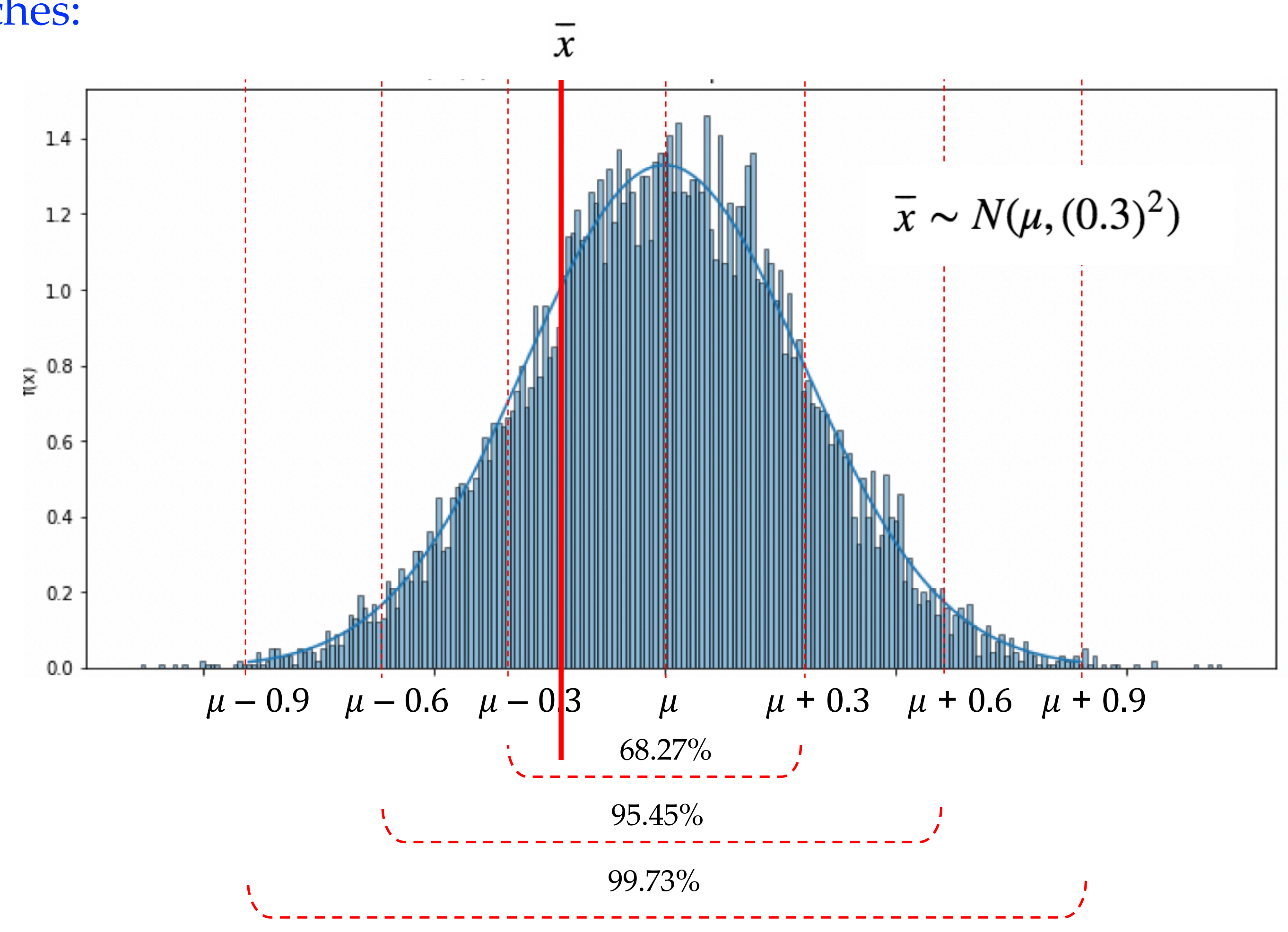
You just don't know where the centerpoint μ is:



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

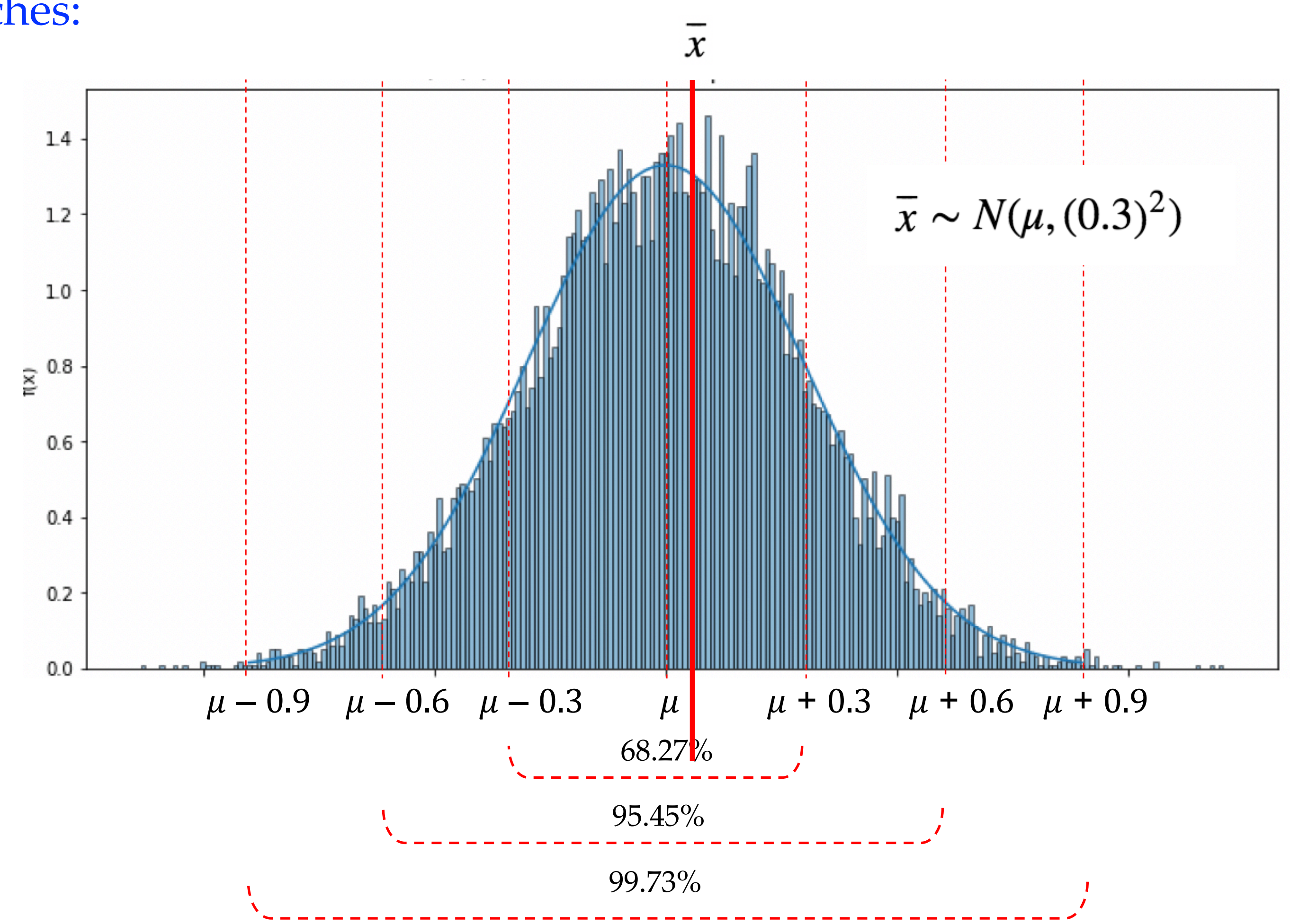
Confidence Intervals When Population Parameters are Known

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:



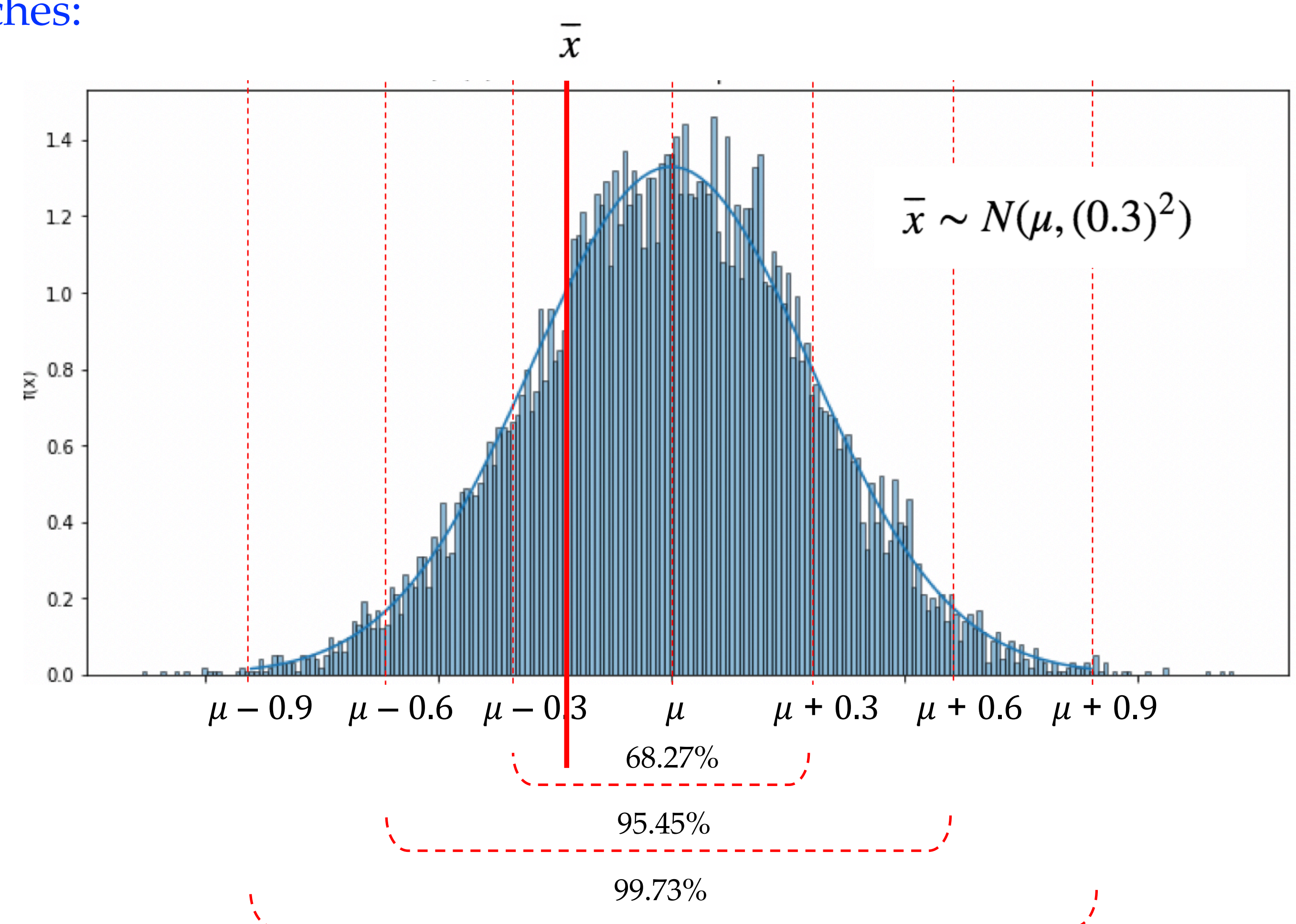
Confidence Intervals When Population Parameters are Known

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:



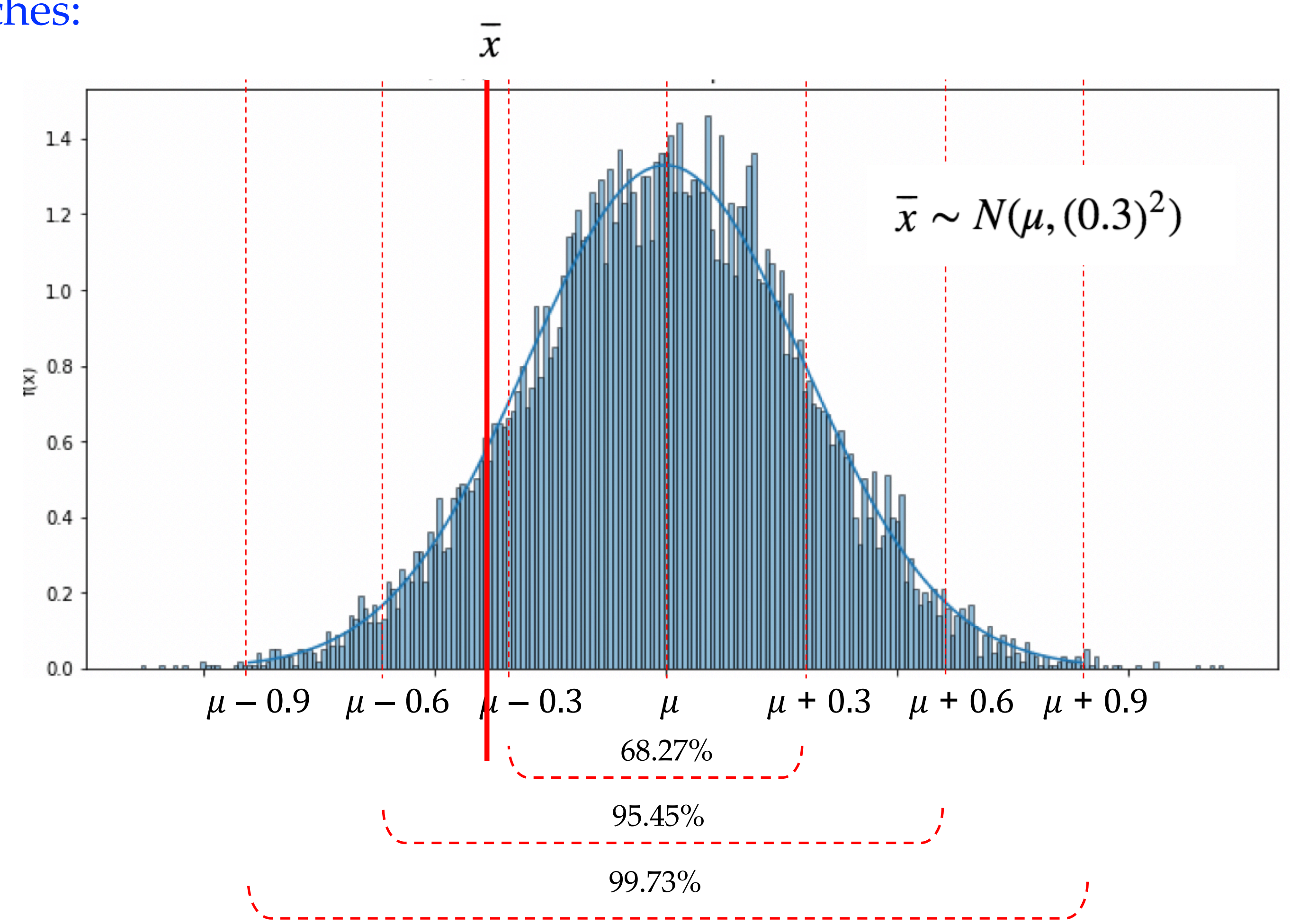
Confidence Intervals When Population Parameters are Known

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:



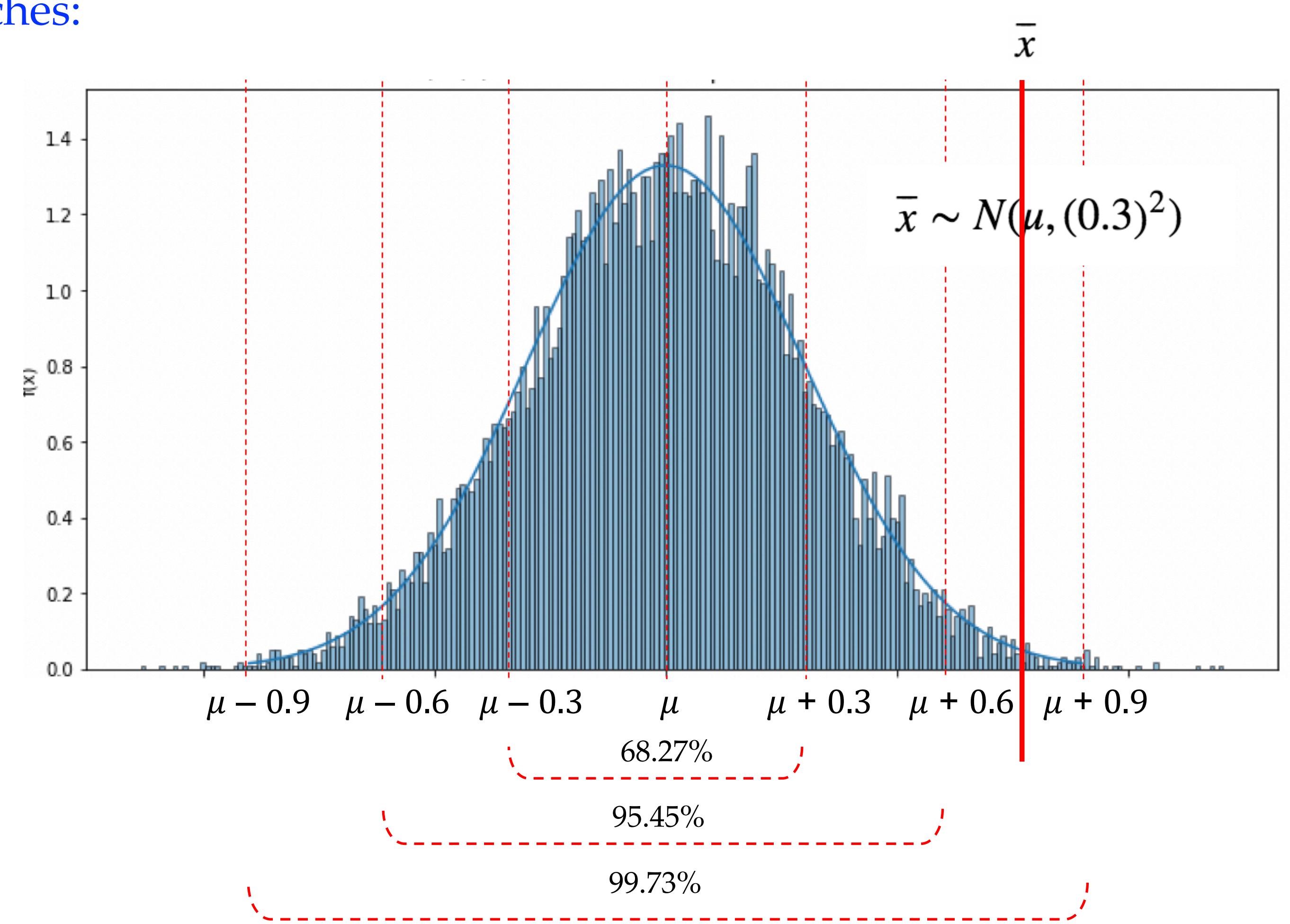
Confidence Intervals When Population Parameters are Known

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:



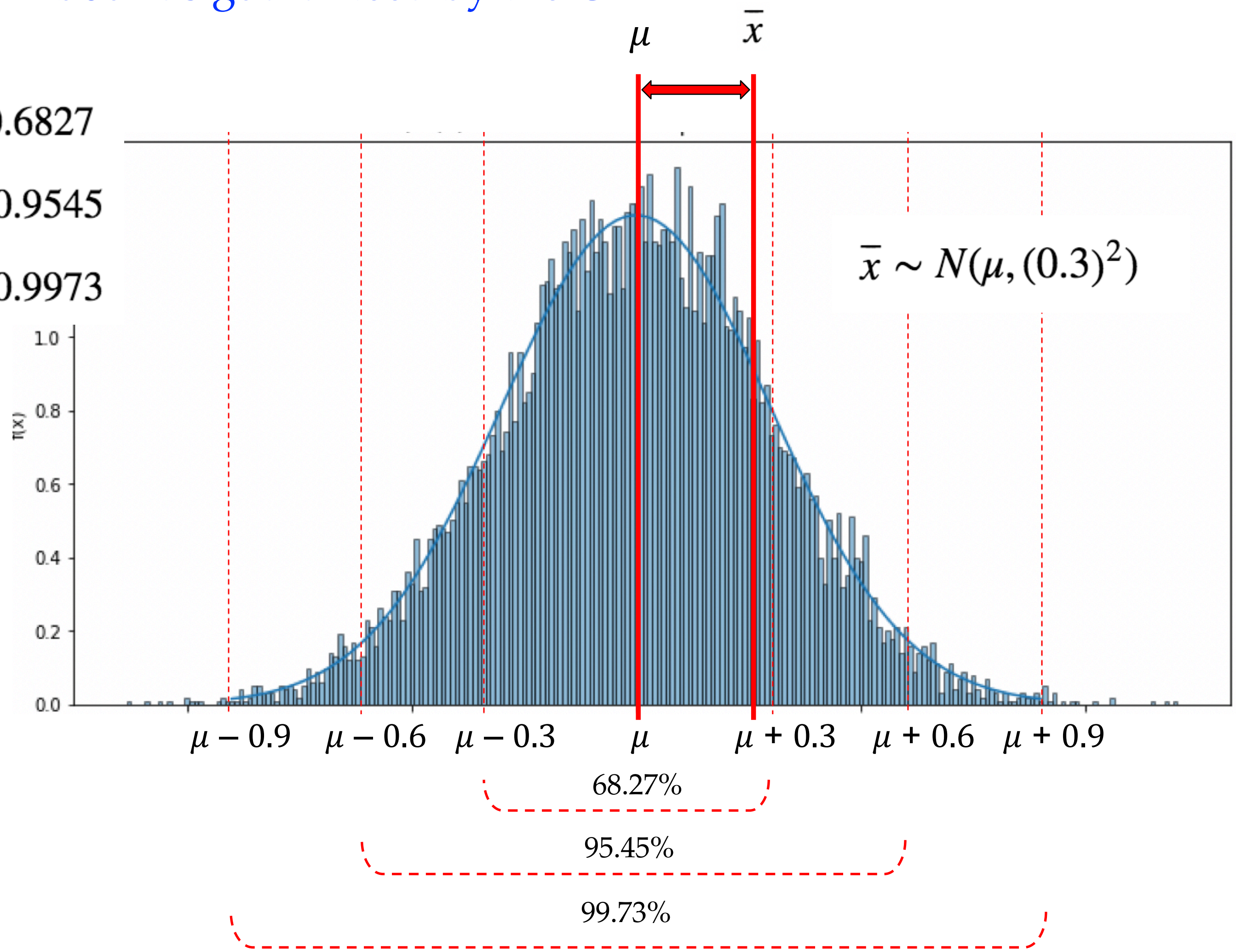
Confidence Intervals When Population Parameters are Known

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:

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

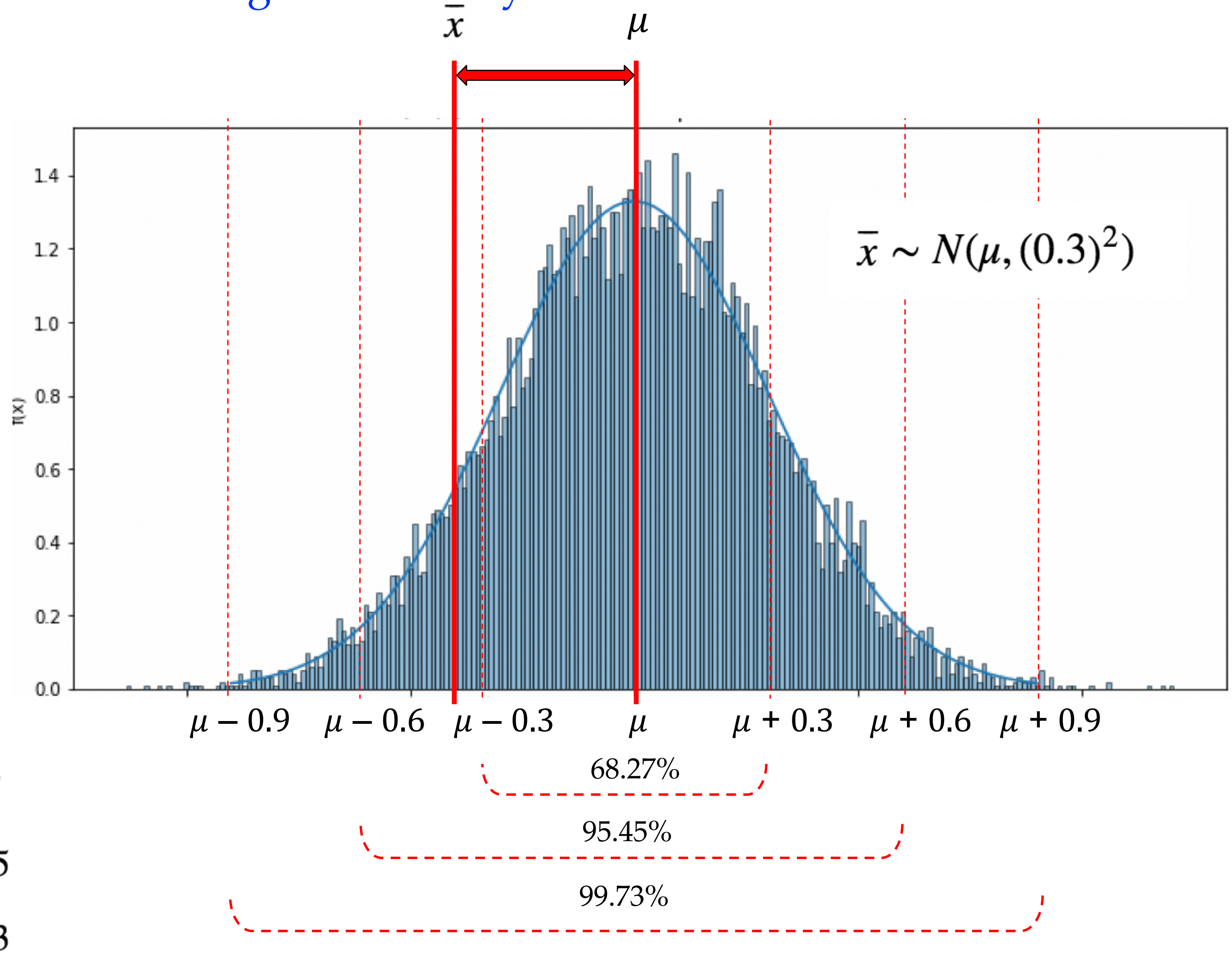
$$P(|\bar{x} - \mu| \leq 2\sigma) = 0.9545$$

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



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:



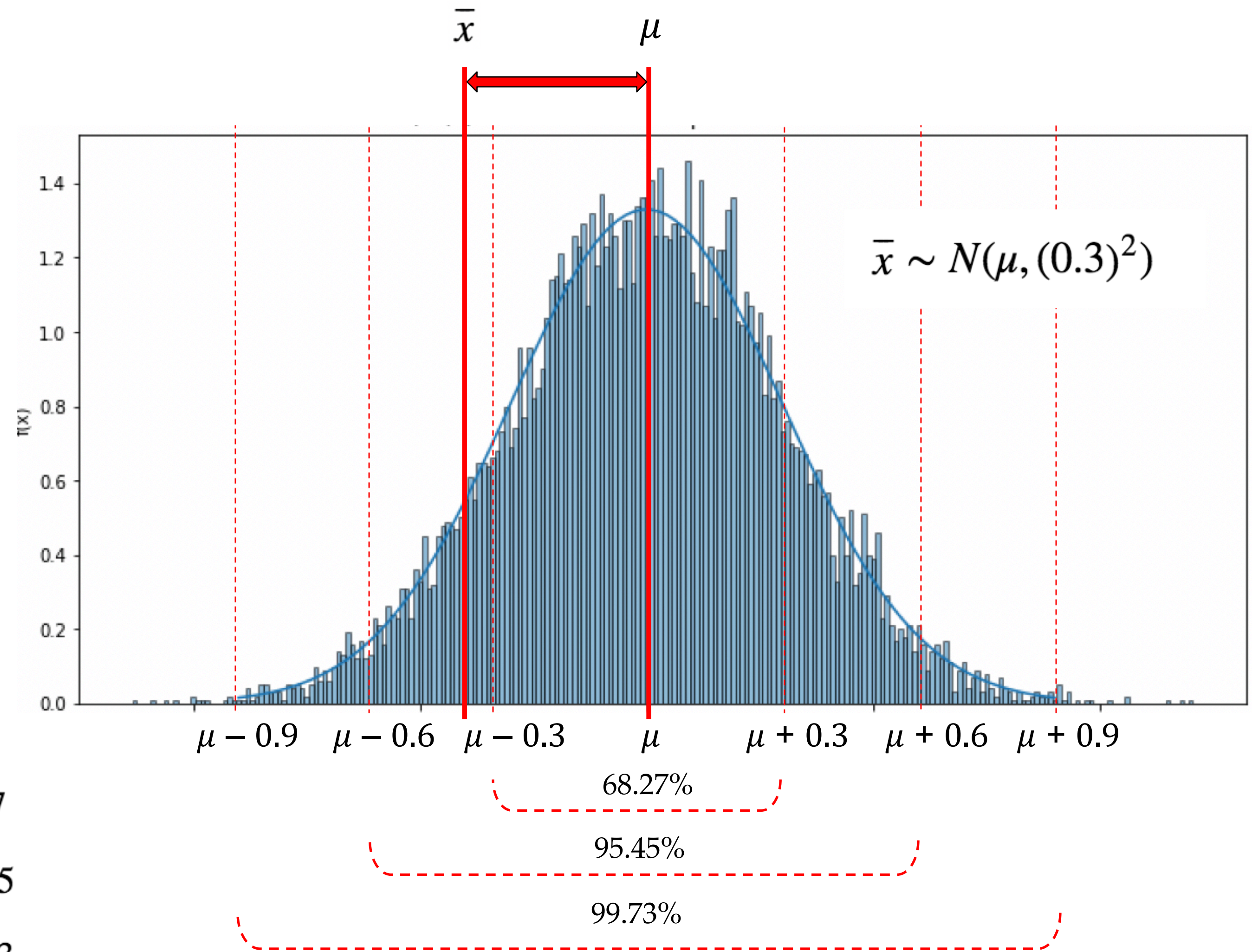
$$P(|\bar{x} - \mu| \leq \sigma) = 0.6827$$

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

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

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 $\bar{x}$:



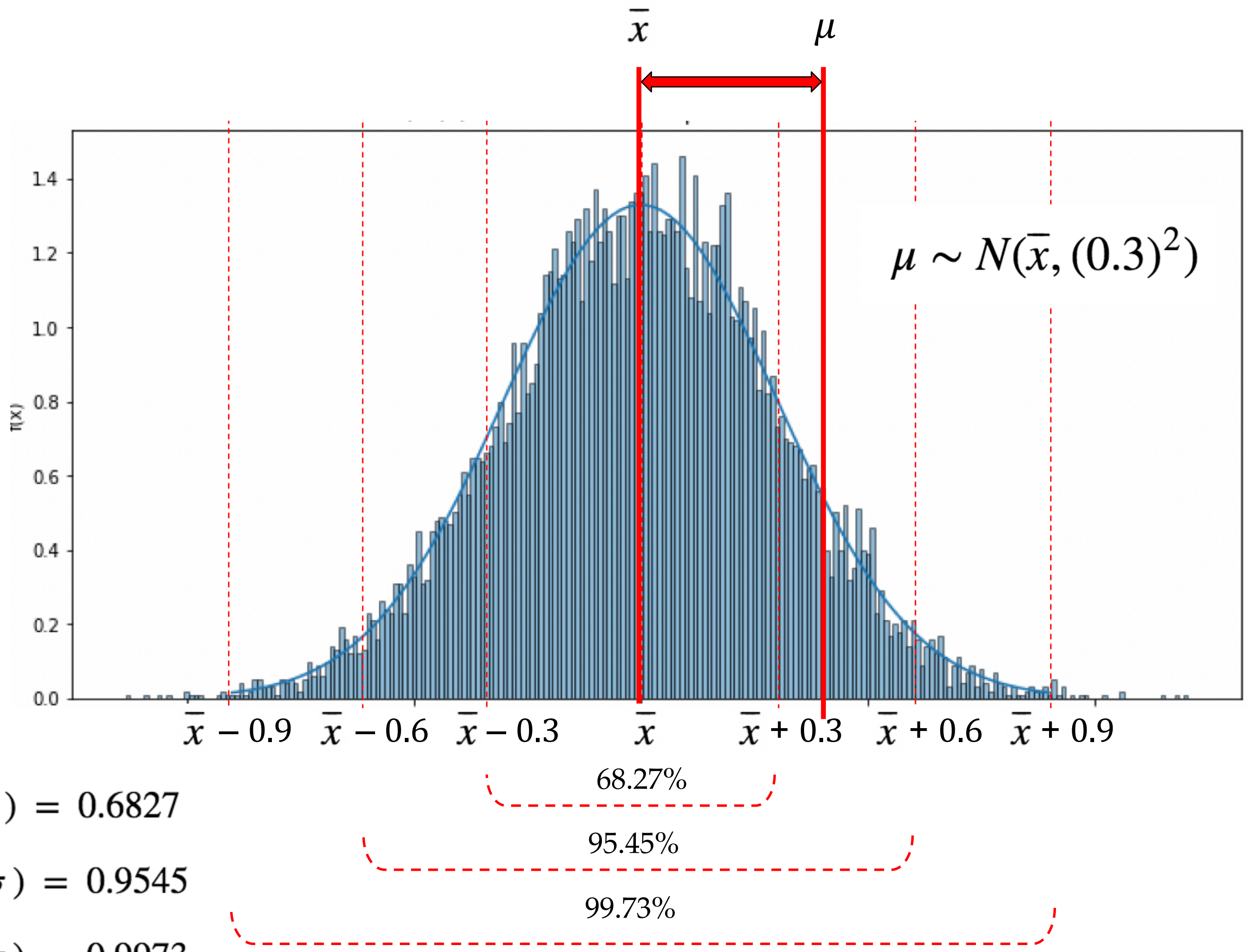
$$P(|\bar{x} - \mu| \le \sigma) = 0.6827$$

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

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

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 $\bar{x}$:



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

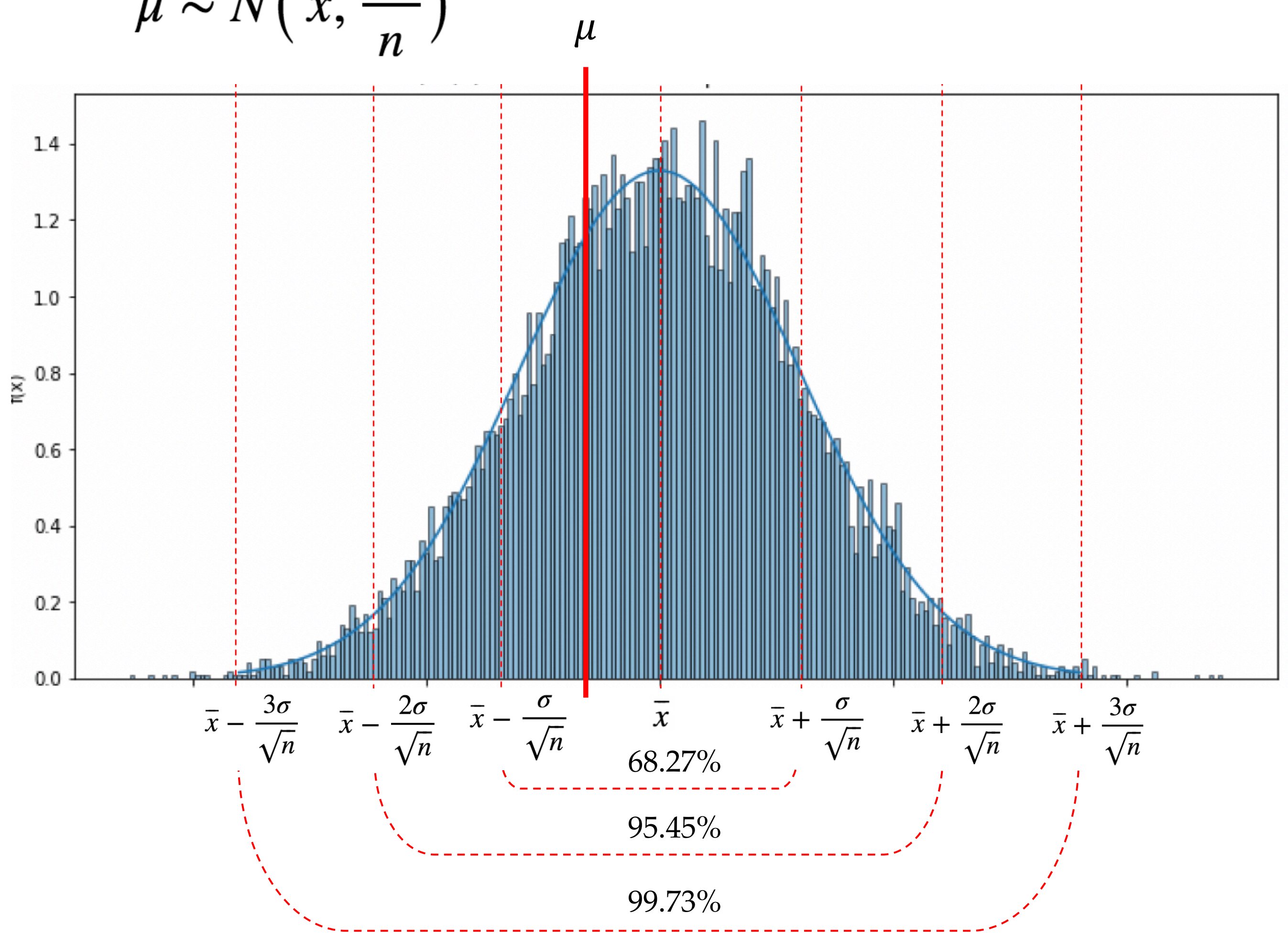
$$P(|\bar{x} - \mu| \leq 2\sigma) = 0.9545$$

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

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).

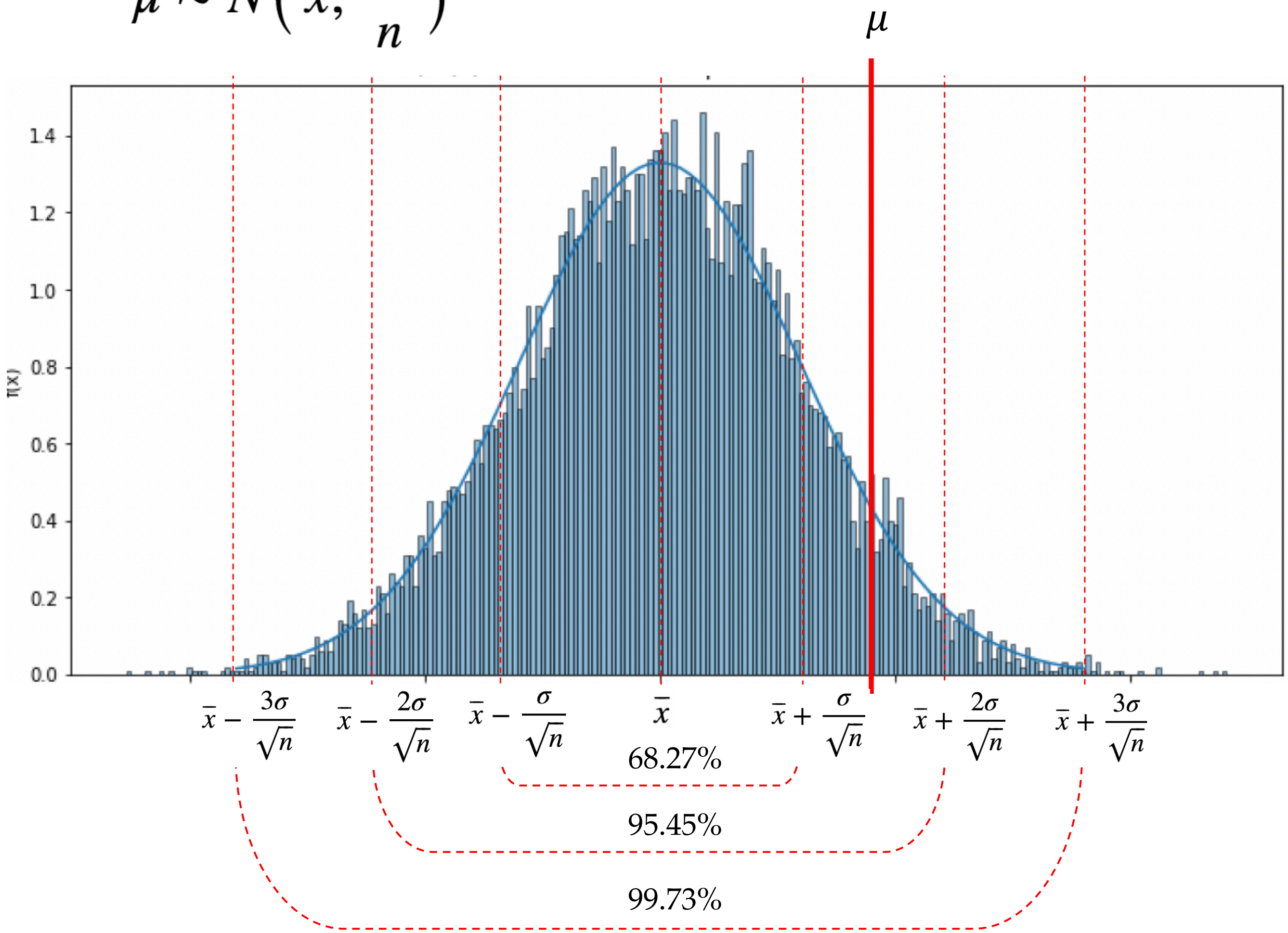
$$\mu \sim N\left(\bar{x}, \frac{\sigma^2}{n}\right)$$



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).

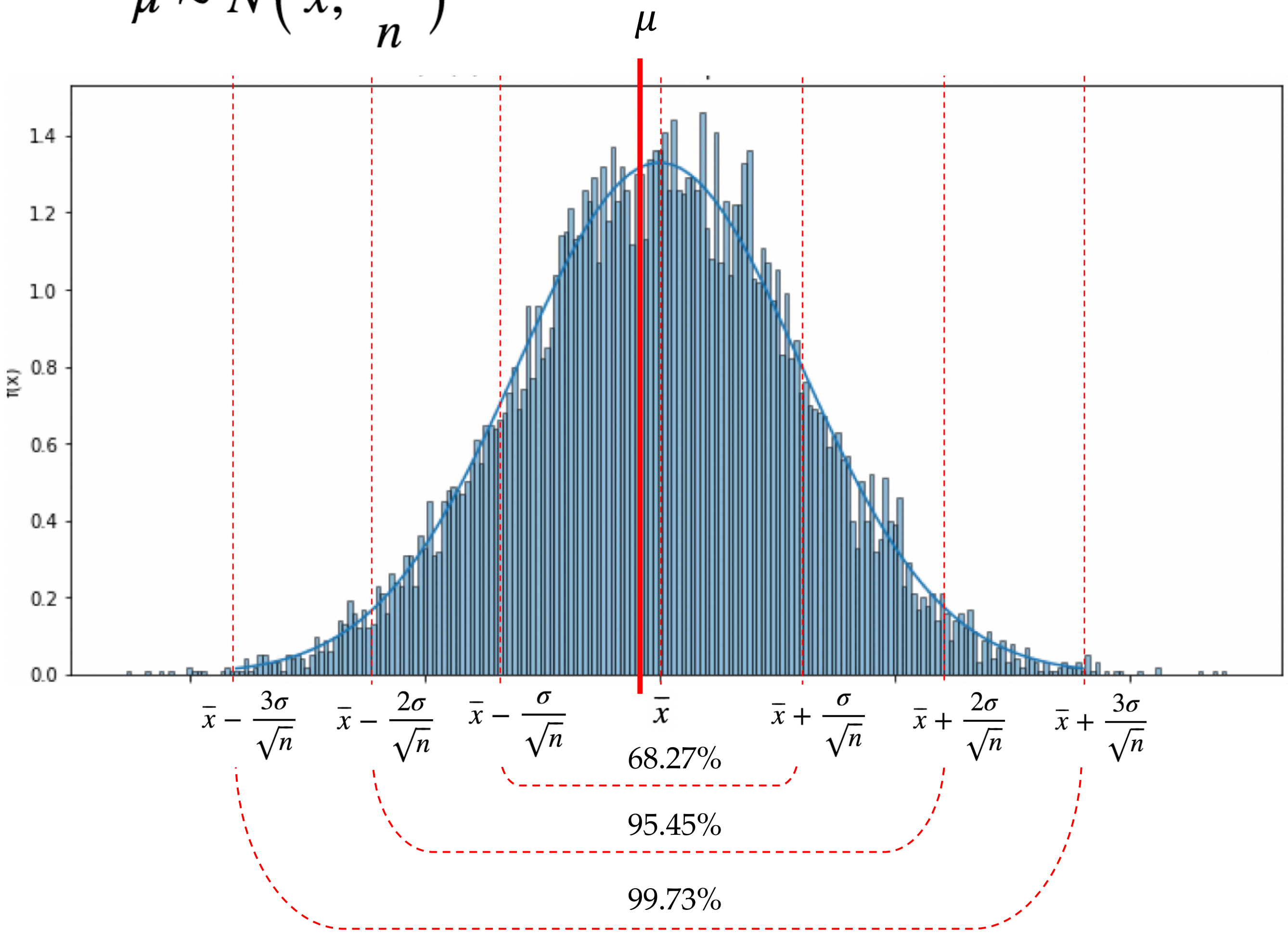
$$\mu \sim N\left(\bar{x}, \frac{\sigma^2}{n}\right)$$



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 $CL = P(\mu - k \cdot \sigma_{\bar{x}} \leq \bar{x} \leq \mu + k \cdot \sigma_{\bar{x}})$;
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 $\sigma = 3$ inches.

1. Choose a sample size n = 100;
2. $\sigma_x = 0.3$ 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 conf. 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.

Sampling When the Population Parameters are Unknown

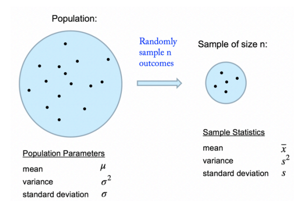
Remember that all this time we have made an **absurd assumption**, that we knew what the population standard deviation was:

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

$$\sigma = 3 \text{ inches.}$$



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

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...

This is universally done in statistics books to "warm-up" to the more realistic situation.

Now that we have applied the basic technique of sampling to confidence intervals it is time to take the training wheels off....

Sampling When the Population Parameters are Unknown

When you don't know the standard deviation of the population, there are three cases where you can still proceed to use the CLT:

➡

(1) When the population has a standard deviation which is **related to the mean by a formula** (e.g., all we studied except the Normal Distribution), **you can simply use the formula with the calculated mean of the sample.**

Example: (Proportions) Yes/No polls assume a Bernoulli population, so the standard deviation is:

$$s = \sqrt{\bar{x} \cdot (1.0 - \bar{x})}$$

(Bernoulli populations are called "Sampling with Proportions" – this is the most common case where we have a formula for the standard deviation.)

Confidence Intervals: Sampling with Proporti

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

When the population is Bernoulli (Yes/No, 1/0, etc.)
we can use the formula `s = p(1-p)` for the variance:

$$s^2 = \bar{x} \cdot (1.0 - \bar{x})$$

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

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Then:

1. Choose a sample size n;
2. Calculate the standard deviation of the sample mean;
3. Choose a confidence level CL (e.g., 95%);
4. Calculate the multiplier k for **s** correspondir $CL = P(\mu - k \cdot s_{\bar{x}} \leq \bar{x} \leq \mu + k \cdot s_{\bar{x}})$
5. Perform random sampling and calculate $\bar{x}$ and $s_{\bar{x}}$;
6. Report your results using the confidence interval corresponding to CL:

"The mean of the population $\bar{x} \pm k \cdot s_{\bar{x}}$ 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 for Proportions

Example -- Poll of BU Students: Should the MA ban on Vaping be continued? (1 = Yes and 0 = No)

1. Choose a sample size n = 50;
2. Choose a confidence level CL = 90 %;
3. Calculate the multiplier k = 1.64;
4. Perform random sampling of students:

```
Sample =
   [1,0,1,0,0,1,0,1,0,0,
    1,0,1,0,1,0,0,1,0,0,
    0,1,1,1,0,1,1,0,0,1... ]
```

5. Calculate the percentage of sample who support the ban: $\bar{x}$ = 0.4667
6. Calculate the sample standard deviation and the standard deviation of the sample mean:

   s = sqrt($\bar{x}$ *(1- $\bar{x}$) = 0.4989; $s_{\bar{x}}$ = 0.4989/sqrt(50) = 0.0706

6. Report your results using the confidence interval corresponding to CL:

“Of 30 BU students polled, 46.67 % +/- 11.61 % support a continued ban on vaping products, with a confidence of 90 %.”

Sampling When the Population Parameters are Unknown

When you don't know the standard deviation of the population, there are three cases:

(1) (Formula) When the population has a standard deviation which is related to the mean by a formula (e.g., all we studied except the Normal Distribution), you can simply use the formula.

(2) (Large Samples) When the population is large (typically, n > 30), by the CLT the distribution of the sample mean is approximately normal, and we can use the sample standard deviation, with one small correction to the formula:

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

$$s^2 = \frac{(X_1 - \bar{x})^2 + (X_2 - \bar{x})^2 + \cdots + (X_n - \bar{x})^2}{n}$$

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

We will change this

Bessel's Correction for Sample Standard Deviation

The formula for the standard deviation has a bias: it **under-estimates** the true standard deviation when applied to samples, because it is an estimate (standard deviation **s** of sample) based on an estimate (mean $\bar{x}$ of the sample):

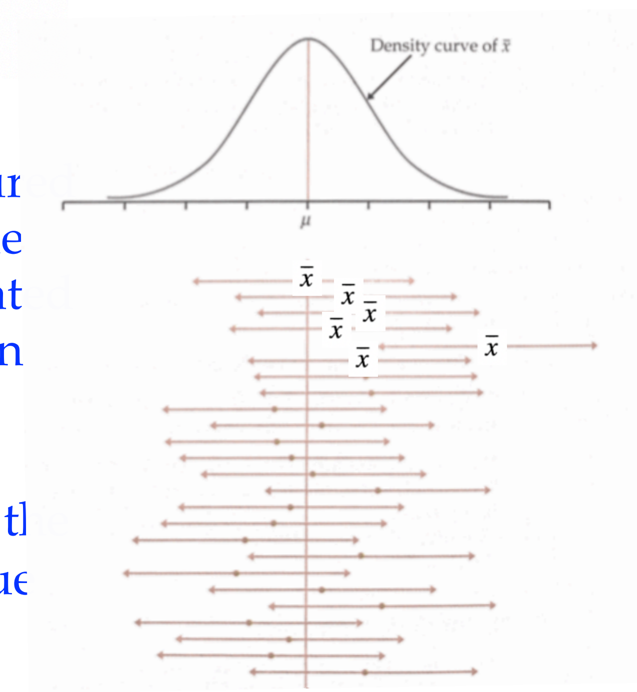
$$\bar{x} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

$$s^2 = \frac{(X_1 - \bar{x})^2 + (X_2 - \bar{x})^2 + \cdots + (X_n - \bar{x})^2}{n}$$

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

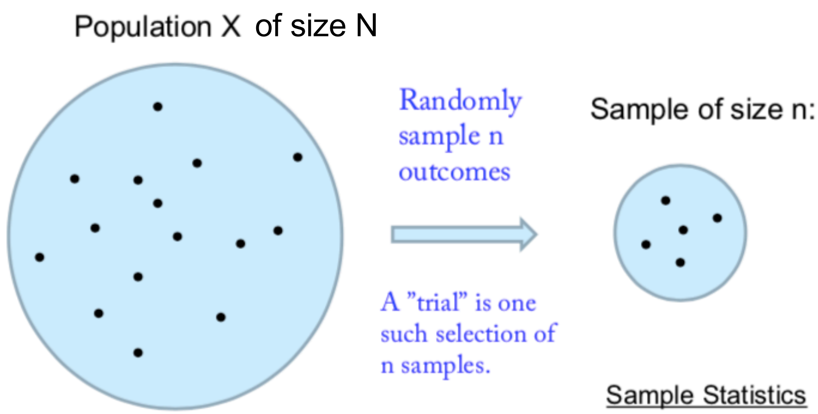
Intuition: Since the sample variance s^2 is being measur against a random value $\bar{x}$, which varies as the sample changes, it is less than the population variance calculat from the mean μ , which is a constant for the duration of the experiment.

Mathematically, it can be shown that the by changing tl denominator to n – 1, we eliminate the bias of the valu calculated.



Bessel's Correction for Standard Deviation

So, there are TWO different formula for the standard deviation:



Population Parameters

mean	μ
variance	σ^2
standard deviation	σ

Sample Statistics

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

$$\text{Population } = \{x_1, \cdots, x_N\}$$

$$\mu = \frac{x_1 + \cdots + x_N}{N}$$

$$\sigma^2 = \frac{(x_1 - \mu)^2 + \cdots + (x_N - \mu)^2}{N}$$

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

$$\sigma_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}}$$

$$\text{Sample } = \{X_1, \cdots, X_n\}$$

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

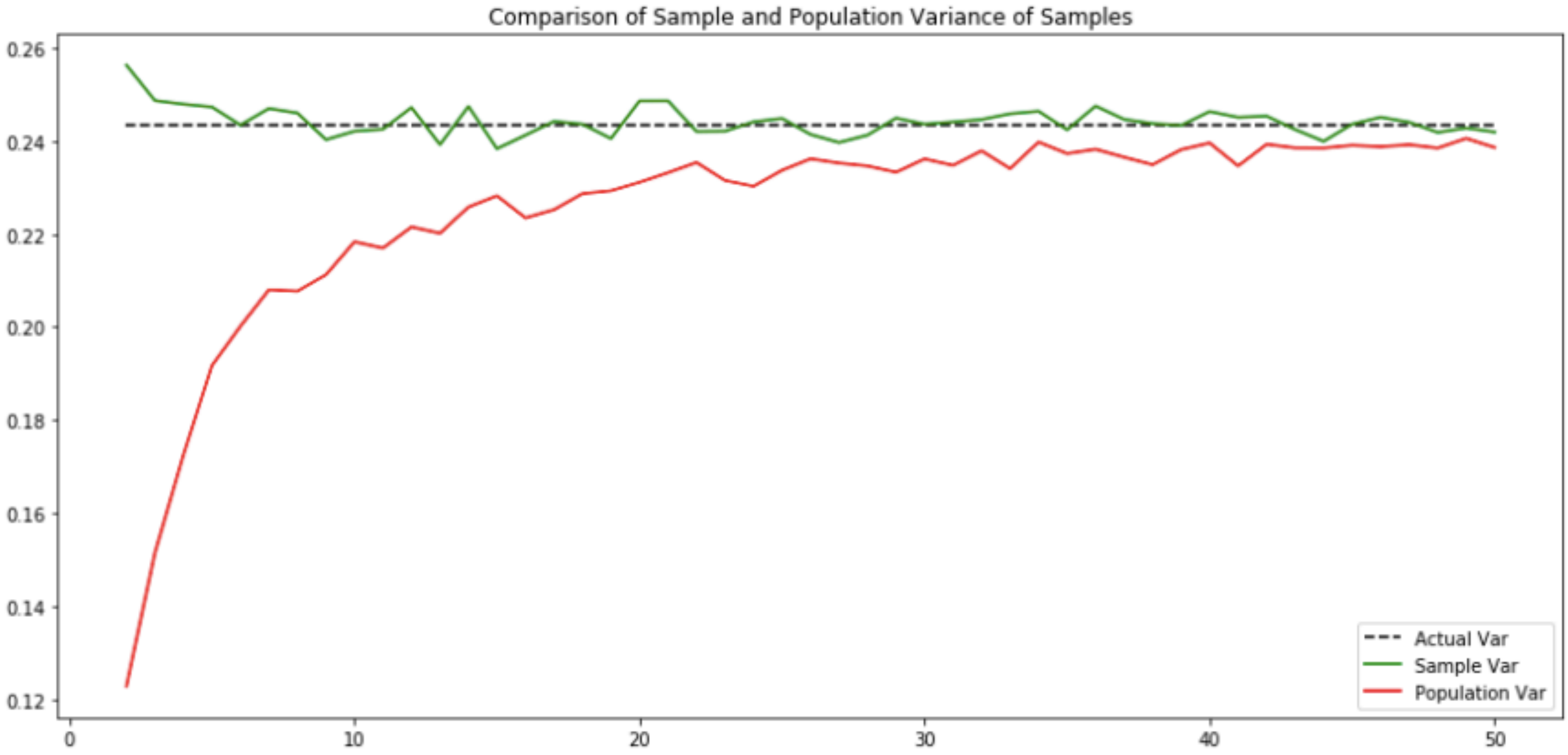
$$s^2 = \frac{(X_1 - \bar{x})^2 + \cdots + (X_n - \bar{x})^2}{n-1}$$

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

$$s_{\bar{x}} = \sqrt{\frac{s^2}{n}}$$

Bessel's Correction for Standard Deviation

This improves the estimate!



Confidence Intervals: Summary of the Procedure for Large Samples

Confidence Intervals Using the Sample Standard Deviation when n > 30.

$$\text{Sample} = \{X_1, \cdots, X_n\}$$

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

$$s^2 = \frac{(X_1 - \bar{x})^2 + \cdots + (X_n - \bar{x})^2}{n-1}$$

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

$$s_{\bar{x}} = \sqrt{\frac{s^2}{n}}$$

We will use **s** as the standard deviation of the sample, calculated using Bessel's Correction (divide by n-1):

Then:

1. Choose a sample size n;
2. Choose a confidence level CL (e.g., 95 %);
3. Calculate the multiplier k correspondir $CL = P(\mu - k \cdot s_{\bar{x}} \leq \bar{x} \leq \mu + k \cdot s_{\bar{x}})$
4. Perform random sampling and calculate $\bar{x}$, **s**, and $s_{\bar{x}}$;
5. Report your results using the confidence interval corresponding to CL:

"The mean of the population $\bar{x} \pm k \cdot s_{\bar{x}}$ 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 Example

Example -- Height of BU Students:

1. Choose a sample size n = 100;
2. Choose a confidence level CL = 95.45%;
3. Calculate the multiplier k = 2;
4. Perform random sampling of 100 students and calculate $\bar{x}$ = 66.13 and the **sample** standard deviation s = 3.45 inches, and then

$$s_{\bar{x}} = \frac{3.45}{\sqrt{100}} = 0.345$$

5. Report your results using the confidence interval corresponding to CL:

"The mean height of BU students is 66.13 +/- 0.69 inches with a confidence of 95.45%."

Sampling When the Population Parameters are Unknown

When you don't know the standard deviation of the population, there are three cases:

(1) (Formula) When the population has a standard deviation which is related to the mean by a formula (e.g., all we studied except the Normal Distribution), you can simply use the formula.

(2) (Large Samples) When the population is large (typically, $n > 30$), by the CLT the distribution of the sample mean is approximately normal, and we can use the sample standard deviation, with one small correction to the formula:

(3) **Third**, when sampling with $n <= 30$ from a population known to be Normal, but with unknown mean and standard deviation, you use the sample standard deviation and a slightly different distribution, called the T-Distribution. (Not covered in CS 237.)

Hypothesis Testing

Hypothesis Testing is a probabalistic version of a Refutation of a mathematical hypothesis, or a Proof by Contradiction.

Example of a **Refutation**:

Hypothesis: Any number with four occurrences of the digit 1, two occurrences of 4, two occurrences of 8, and no occurrences of 2 or 6, is a prime number.

Refutation: Nope! 1,197,404,531,881 = 1,299,827 * 921,203

Example of **Proof by Contradiction**:

Theorem: For all integers n, if n^2 is odd, then n is odd.

Proof: Suppose we assume the negation of the theorem:

Hypothesis: $\exists$ n such that n^2 is odd and n is even.

Nope! Because then $\exists$ k such that $n = 2k$ and so $n^2 = (2k)^2 = 4(k)^2$ and hence n^2 is divisible by 2 and even. Therefore, the hypothesis is false, and the theorem (the inverse of the hypothesis) must be true. Q.E.D.

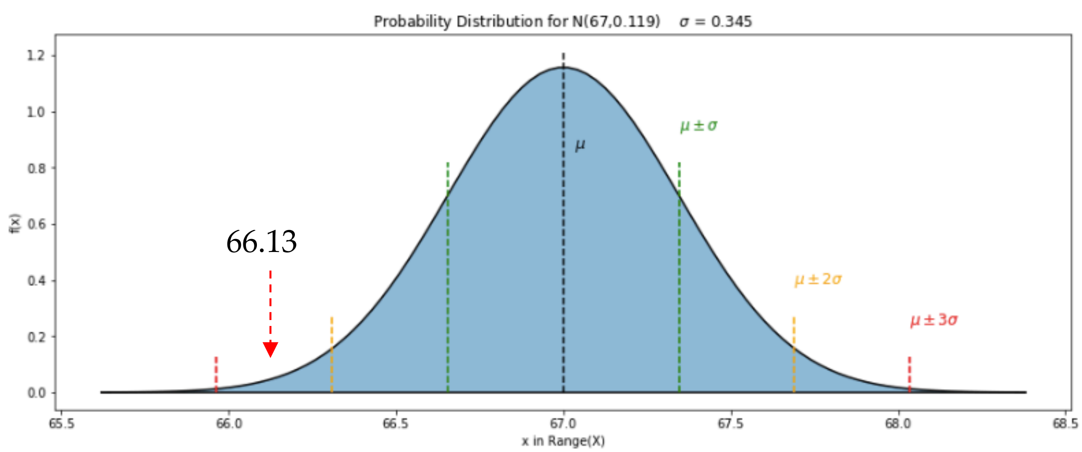
Hypothesis Testing

When we test a hypothesis probabalisticly, instead of absolutely refuting it, we show that the hypothesis is extremely unlikely given the result of our sampling experiment.

Now suppose you had a previous hypothesis about the heights of BU students:

Hypothesis: BU students have a mean height of 67 inches.

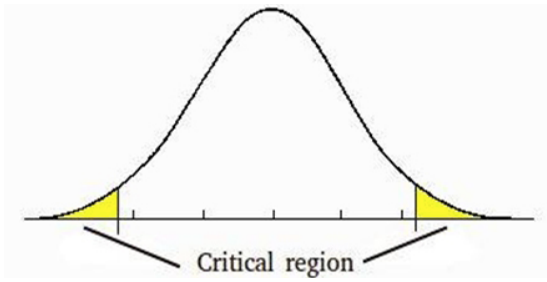
Now we estimated the standard deviation of the population as $s = 3.45$ inches, and when we do our sample mean with n = 100 students, **our hypothesis implies that this s**



But our experiment gives a value of 66.13, which is unlikely! So our hypothesis is very likely to be wrong, and we should reject it. But how to decide? How

Hypothesis Testing: Two-Tailed Tests

When the extreme values could be in either direction (low or high): your hypothesis could be rejected because it is too low, OR because it is too high.



- BU students have a mean height of 68
  - Sam Adams Boston Lager contains 4.75% alcohol

In this case, you state a **Null Hypothesis** about the mean of a population **X**:

H_0 = "$\mu_X = k$." ← This is the hypothesis to reject or not.

And you state (or leave implicit) the **Alternative Hypothesis:**

H_1 = "$\mu_X < k \ or \ k < \mu_X$" or, more simply, H_1 = "$\mu_X \neq k$."

You Reject H_0 if your sample mean is much larger or much smaller than k :

$$\bar{x} \ll k \quad \text{or} \quad \bar{x} \gg k$$

Hypothesis Testing: Two-Tailed Tests

Hypothesis Two-Sided Test:

Step One: State a **Null Hypothesis** making a claim about the mean of a population **X**:

$\mathbf{H_0}$ **=** "$\mu_X = k$." (and $\mathbf{H_1}$ **=** "$\mu_X \neq k$.")

You will either **Reject** this hypothesis or do nothing (**Fail to Reject**).

Step Two. Determine how willing you are to be wrong, i.e., define the **Level of Significance** $\boldsymbol{\alpha}$ of the test:

α = probability you are wrong if you Reject $\mathbf{H_0}$ when it is actually correct.

Example:

1. $\mathbf{H_0}$**:** BU students have a mean height of 67 inches (k = 67).

2. $\alpha = 0.01$ (I am willing to be wrong 1% of the time)

Hypothesis Testing: Two-Tailed Tests

Hypothesis Two-Sided Test:

Step Three. Do the sampling experiment to find a sample mean $\bar{x}$ and the standard deviation of the sampling distribution s .

Example:

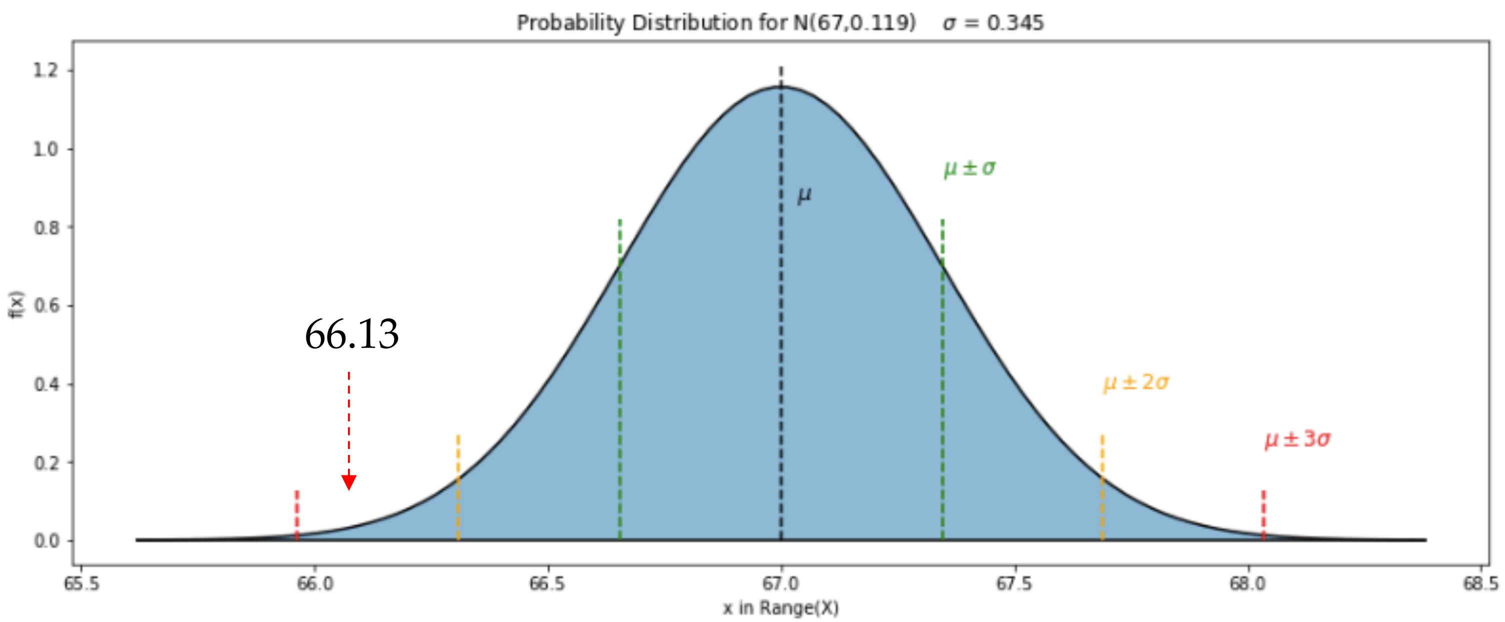
3. We perform the sampling experiment for n = 100, and find $\bar{x}$ = 66.13 and s = 3.45.

Hypothesis Testing: Two-Tailed Tests

Hypothesis Two-Sided Test:

Now, at this point, using the hypothesis that the mean should be 67 inches, and the fact that the standard deviation of the sampling distribution is s = 0.345, according to the hypothesis, we should have a sampling distribution of

$$\bar{X} = N(67, 0.345^2)$$



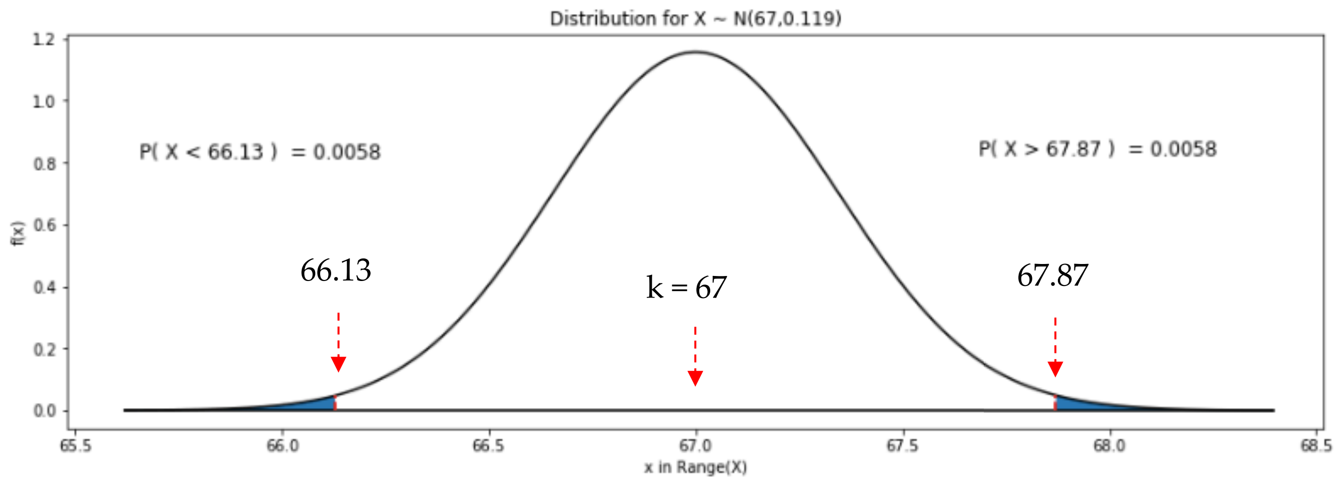
The question is, of course, how likely our actual value of 66.13 is under this assumption!

Hypothesis Testing: Two-Tailed Tests

Hypothesis Two-Sided Test:

Step Four: Calculate the **p-value** of the sample m $\bar{x}$ n , the probability that the random $\bar{X}$ ariable would be farther away from k (our hypothesis value for $\bar{x}$ ie mean) than is:

$$P(|\bar{X} - k| > |\bar{x} - k|)$$



$\bar{x}$

The p-value is the probability of seeing the value or a value even $\bar{x}$ ore unlikely, if H_0 were true. **Because we have a two-tailed test**, we have to calculate how far is from the hypothesized value k and multiply by 2:

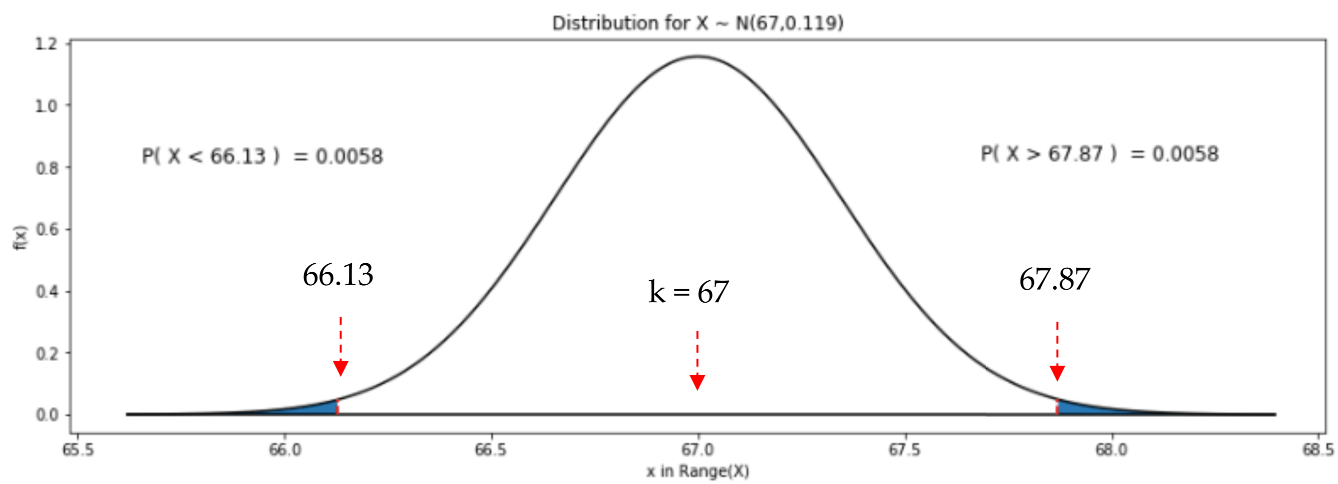
$$2 * P(\bar{X} < \bar{x}) \quad \text{if} \quad \bar{x} < k \qquad\qquad 2 * P(\bar{X} > \bar{x}) \quad \text{if} \quad \bar{x} > k$$

Hypothesis Testing: Two-Tailed Tests

Hypothesis Two-Sided Test:

Step Four: Calculate the **p-value** of the sample m $\bar{x}$ n , the probability that the random $\bar{X}$ ariable would be farther away from k (our hypothesis value for $\bar{x}$ ie mean) than is:

$$P(|\bar{X} - k| > |\bar{x} - k|)$$



Example: Since 66.123 < 67, we calculate the p-value = 0.0117 from the left side:

```
In [31]:  1  2 * norm.cdf(x=66.13,loc=67,scale=0.345)

Out[31]: 0.011677627373262023
```

Hypothesis Testing: Two-Tailed Tests

Hypothesis Two-Sided Test:

Step Five. If the **p-value < α** , Reject, otherwise Fail to Reject.

Example: Clearly we must Fail to Reject, since $0.0117 > 0.01$! We can not reject the hypothesis on the basis of the data!

Some things to notice:

(1) If we had set the level of significance at 95%, we would have Rejected! It is important, therefore, to set your parameters **before** doing the test!

(2) This is precisely the same thing as if we asked "Is 67 inside the 99% confidence interval for our result?" using techniques from last lecture.

Hypothesis Testing: One-Tailed Tests

When the extreme values are considered in one direction only, you have either an **Upper One-Tailed Test** or a **Lower One-Tailed Test**:

Example of hypothesis for an **Upper One-Tailed Test**:

- I claim Trevor does not have ESP: his chance of guessing the color of a card I hold hidden from him is 0.5 (if he does *much* better I'll reject my hypothesis!)

Example of hypothesis for a **Lower One-Tailed Test**:

- Seagate claims its disk drives last an average of 10,000 hours before failing (if we find the mean is *much* lower we may reject their claim).

Hypothesis Testing: One-Tailed Tests

One-Tailed: When the extreme values are considered in one direction only, you have either an **Upper One-Tailed Test** or a **Lower One-Tailed Test**:

In these cases, you again state a **Null Hypothesis** about the mean of a population **X**:

H_0 = "$\mu_X = k$." ← This is the hypothesis to reject or not.

And you state (or leave implicit) the **Alternative Hypothesis:**

For Lower: **H_1 = "$\mu_X < k$"** **For Upper:** **H_1 = "$k < \mu_X$"**

You Reject H_0 if your sample mean is very different than k:

For Lower: $\bar{x} \ll k$ **For Upper:** $k \ll \bar{x}$

[The main difference here is that you don't multiply by 2 when calculating the p-value.]

Hypothesis Testing: One-Tailed Tests

Hypothesis Upper One-Tailed Test:

1. State a **Null Hypothesis** which makes a claim about the mean of a population **X**:

$$\mathbf{H_0} = \text{“}\mu_X = k.\text{”} \qquad\qquad (\text{and } \mathbf{H_1} = \text{“}k < \mu_X\text{”})$$

You will either **Reject** this hypothesis or do nothing (**Fail to Reject**).

2. Determine how willing you are to be wrong, i.e., define the **Level of Significance** $\boldsymbol{\alpha}$ of the test: α = probability you are wrong if you Reject $\mathbf{H_0}$ when it is actually correct.

$\bar{x}$

3. Determine a sample size n, take a random sample of size n, and determine the sample mean . . Establish the standard deviation, either using the (known) population standard deviation or $\bar{x}$ .e sample standard deviation (more on this later).

$\bar{x}$

$\bar{x}$

4. Calculate the **p-value** of the mean x , the probability that the random variable X would be larger than k : P(X >) The p-value represents the probability of seeing the value or a value even more unlikely (i.e., larger), if H_0 were true.

Hypothesis Testing: One-Tailed Tests

Hypothesis Lower One-Tailed Test:

1. State a **Null Hypothesis** which makes a claim about the mean of a population **X**:

H_0 = "$\mu_X = k$."   (and H_1 = "$\mu_X < k$")

You will either **Reject** this hypothesis or do nothing (**Fail to Reject**).

2. Determine how willing you are to be wrong, i.e., define the **Level of Significance *α*** of the test

α = probability you are wrong if you Reject H_0 when it is actually correct.

3. Det $\bar{x}$ mine a sample size n, take a random sample of size n, and determine the sample mean . . Establish the standard deviation, either using the (known) population standard deviation or the sample standard deviation (more on this later).

$\bar{x}$

$\bar{x}$ $\bar{x}$

4. Calcul $\bar{x}$ e the **p-value** of the mean , the probability that the random variable X would be **smaller** than : P(X <). The p-value represents the probability of seeing the value or a value even more unlikely (i.e., even smaller), if H_0 were true.

Hypothesis Testing: One-Tailed Tests

Example: Upper One-Tailed Test:

Trevor claims that he has ESP. I disagree. My hypothesis is that Trevor does not have ESP. The question is whether he can guess correctly much **more** than half the time, so this is an upper one-tailed test.

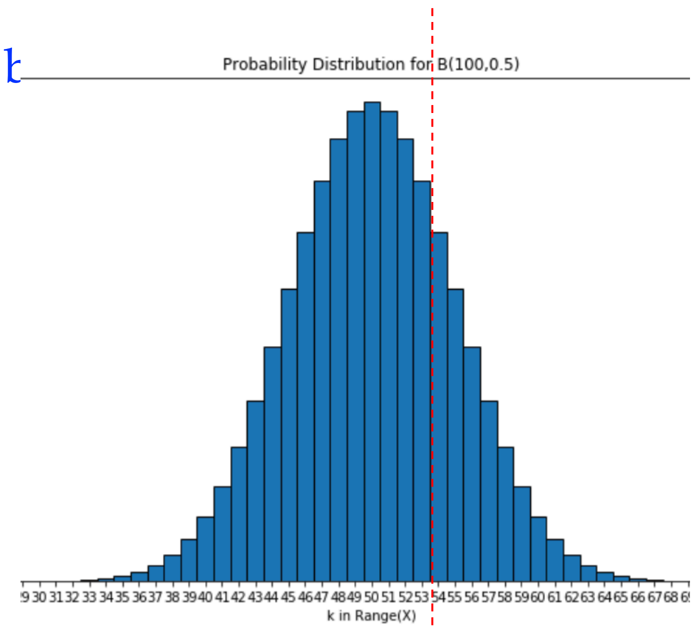
To test, I draw 100 cards from a deck (with replacement) and he guesses the color. The level of significance will be 5%.

H_0 = “Trevor's average number of correct cards is 50, because he is randomly guessing.”
H_1 = “Trevor will guess many more than 50 correct, b

In the experiment, he gets 54 cards correct.

Note that the best model of this experiment is a Binomial experiment, not Normal. Since this is an u

$$P(X \geq 54) = \sum_{i=54}^{100} \binom{100}{i} (0.5)^i (0.5)^{100-i} = 0.2431.$$



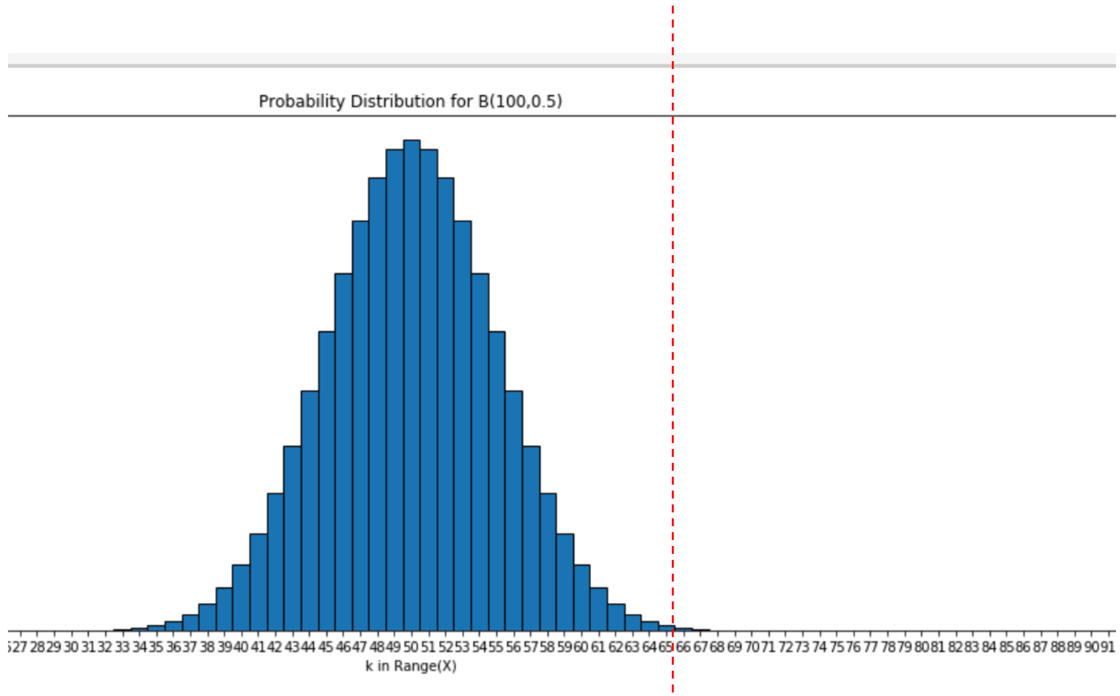
Since 0.2431 > 0.05, we fail to reject H_0

Hypothesis Testing: One-Tailed Tests

But what if he had guessed 68 of them correctly?

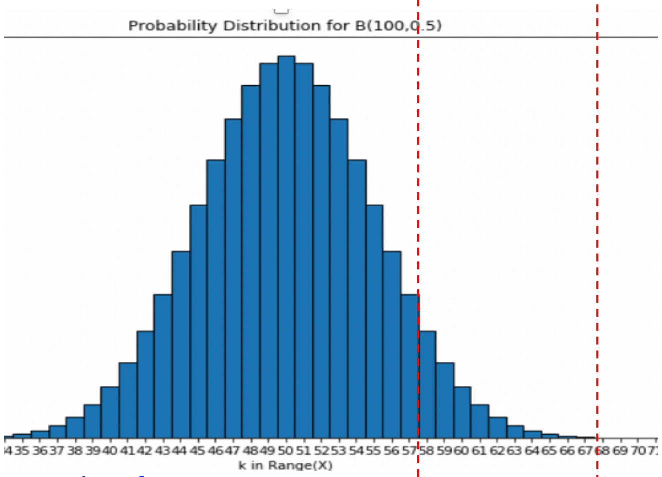
P(X>=68) = 0.0002044

Since 0.0002 < 0.05, we Reject my hypothesis that Trevor does not have ESP, because he did something very, very unlikely!



Hypothesis Testing: One-Tailed Tests

Here is a table of how probable it is that Trevor guessed ≥ k cards correctly, if in fact he were simply guessing with probability 0.5 of success; these the “p-values” of the outcome of the test:




```
xbar = 50:  0.460205381306
xbar = 51:  0.382176717201
xbar = 52:  0.308649706795
xbar = 53:  0.242059206804
xbar = 54:  0.184100808663
xbar = 55:  0.135626512037
xbar = 56:  0.0966739522478
xbar = 57:  0.0666053096036
xbar = 58:  0.04431304005 7
xbar = 59:  0.0284439668205
xbar = 60:  0.0176001001089
xbar = 61:  0.0104893678389
xbar = 62:  0.00601648786268
xbar = 63:  0.00331856025796
xbar = 64:  0.00175882086149
xbar = 65:  0.000894965195743
xbar = 66:  0.000436859918456
xbar = 67:  0.000204388583713
xbar = 68:  9.15716124412e-05
xbar = 69:  3.9250698228e-05
xbar = 70:  1.60800076479e-05
```

Reject at 5% Level of Significance

Reject at 1% Level of Significant