# Sampling Distributions

# Estimation

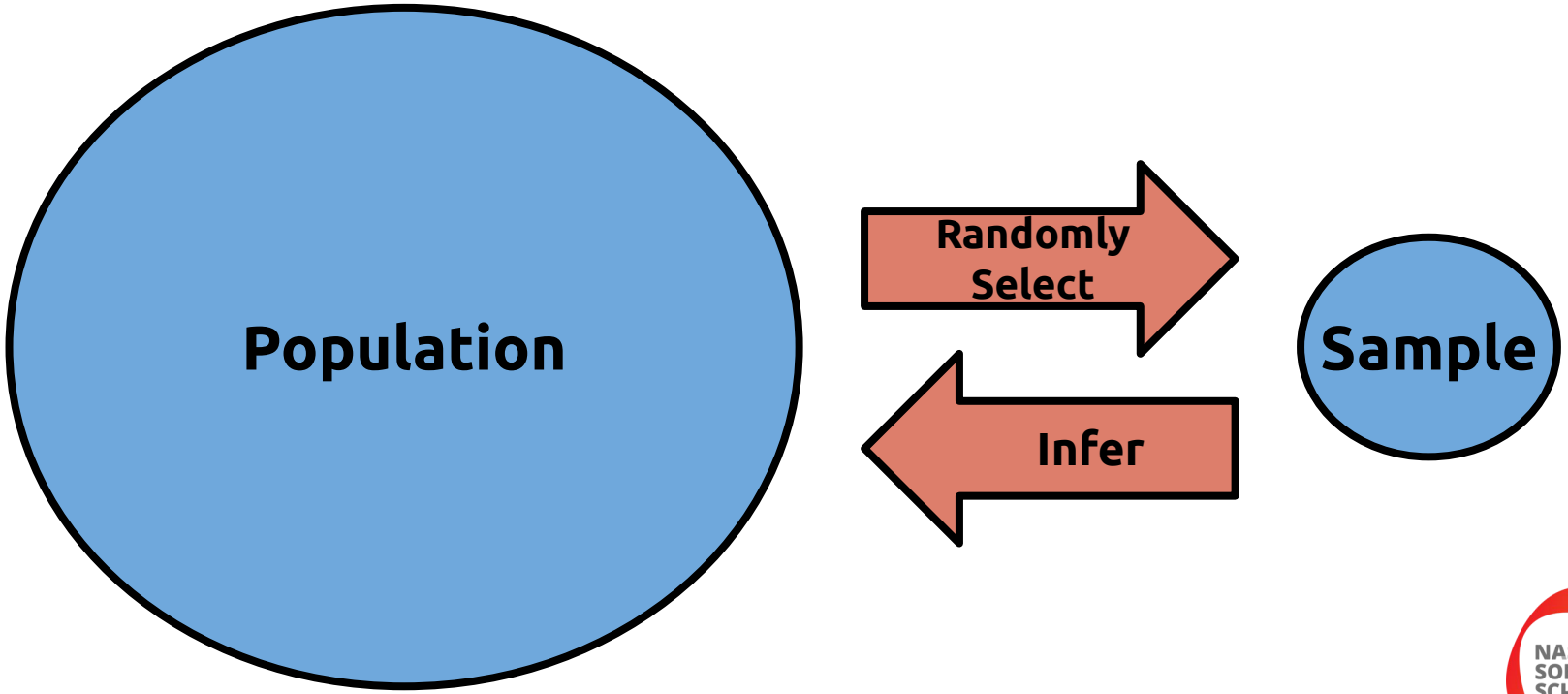
When doing statistics, the goal is often to infer something about a population *parameter* using only a sample from that population.

Examples:

- Estimating the average household income in Putnam County.
- Predicting the percentage of votes a particular candidate will receive in an upcoming election, based on a poll.

NASHVILLE SOFTWARE SCHOOL

# Estimation




NASHVILLE SOFTWARE SCHOOL

# Sampling Distributions

When doing inference, we will have only a single sample to base our estimates off of.

However, we can work to understand the distribution from which our sample came - not the population distribution but the *distribution of all possible samples*.

For this, we can use the Sampling_Distribution.ipynb notebook.

NASHVILLE SOFTWARE SCHOOL

**Population Mean**

$$\mu = 100$$

**Population Standard Deviation**

$$\sigma = 10$$

| Sample Size | Mean of Sample Means | Standard Deviation of Sample Means |
|---|---|---|
| 1 | | |
| 10 | | |
| 25 | | |
| 100 | | |
| 1000 | | |

NASHVILLE SOFTWARE SCHOOL

# Cool Fact:

If *X* is a normally-distributed variable with mean ***μ*** and standard deviation ***σ***, the distribution of sample means of size *n* is also normal with mean ***μ*** and standard deviation $\frac{\sigma}{\sqrt{n}}$

# Cool Fact:

If *X* is a normally-distributed variable with mean ***μ*** and standard deviation ***σ***, the distribution of sample means of size *n* is also normal with mean ***μ*** and standard deviation $\frac{\sigma}{\sqrt{n}}$

The quantity $\frac{\sigma}{\sqrt{n}}$ is called the **standard error of the mean**.

# Really Cool Fact

# Really Cool Fact

**Central Limit Theorem:** Let $X$ be a random variable with mean $\mu$ and standard deviation $\sigma$.

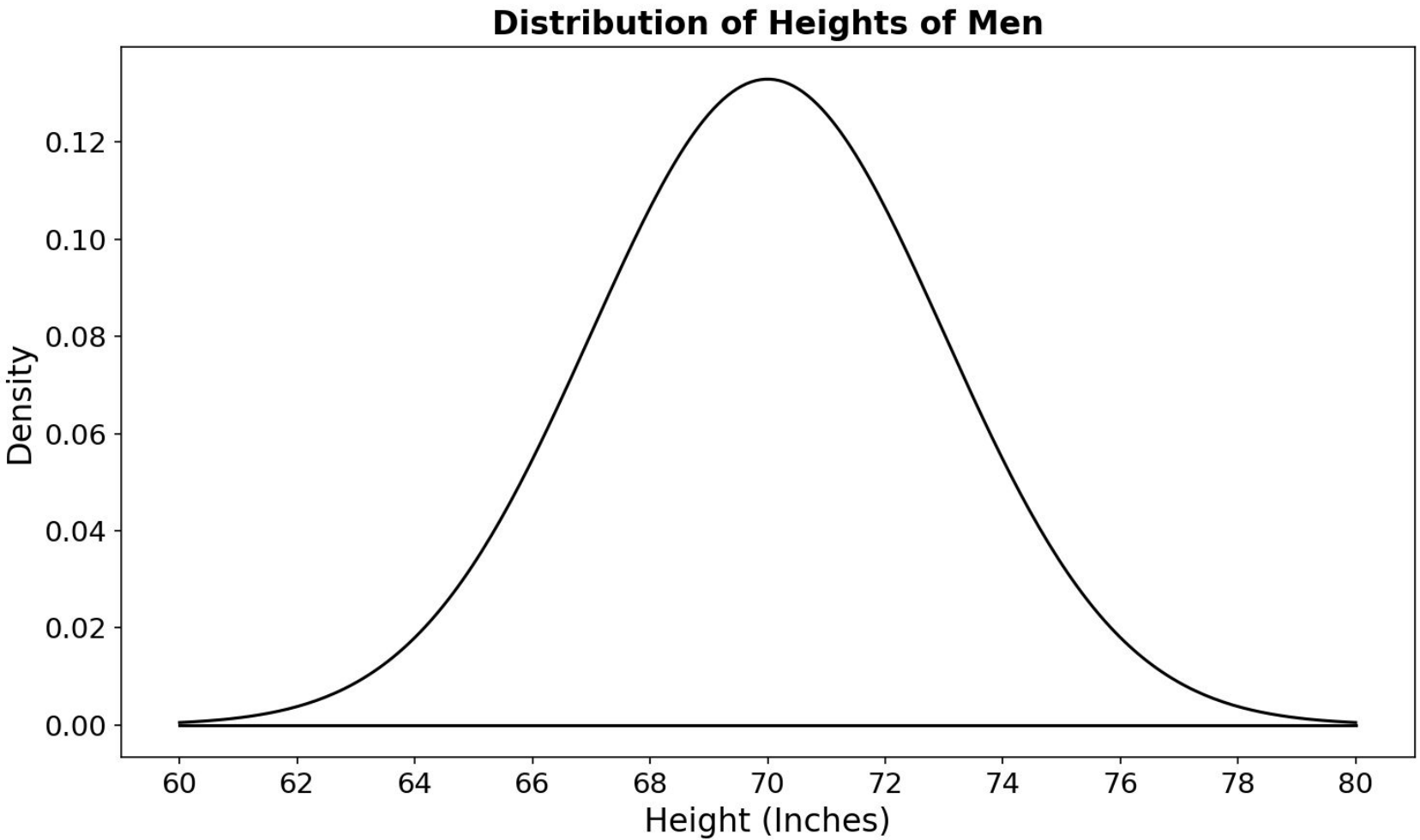
For large enough sample size $n$, (as long as X is reasonably well-behaved), the distribution of sample means is approximately normal with mean $\mu$ and standard deviation $\frac{\sigma}{\sqrt{n}}$

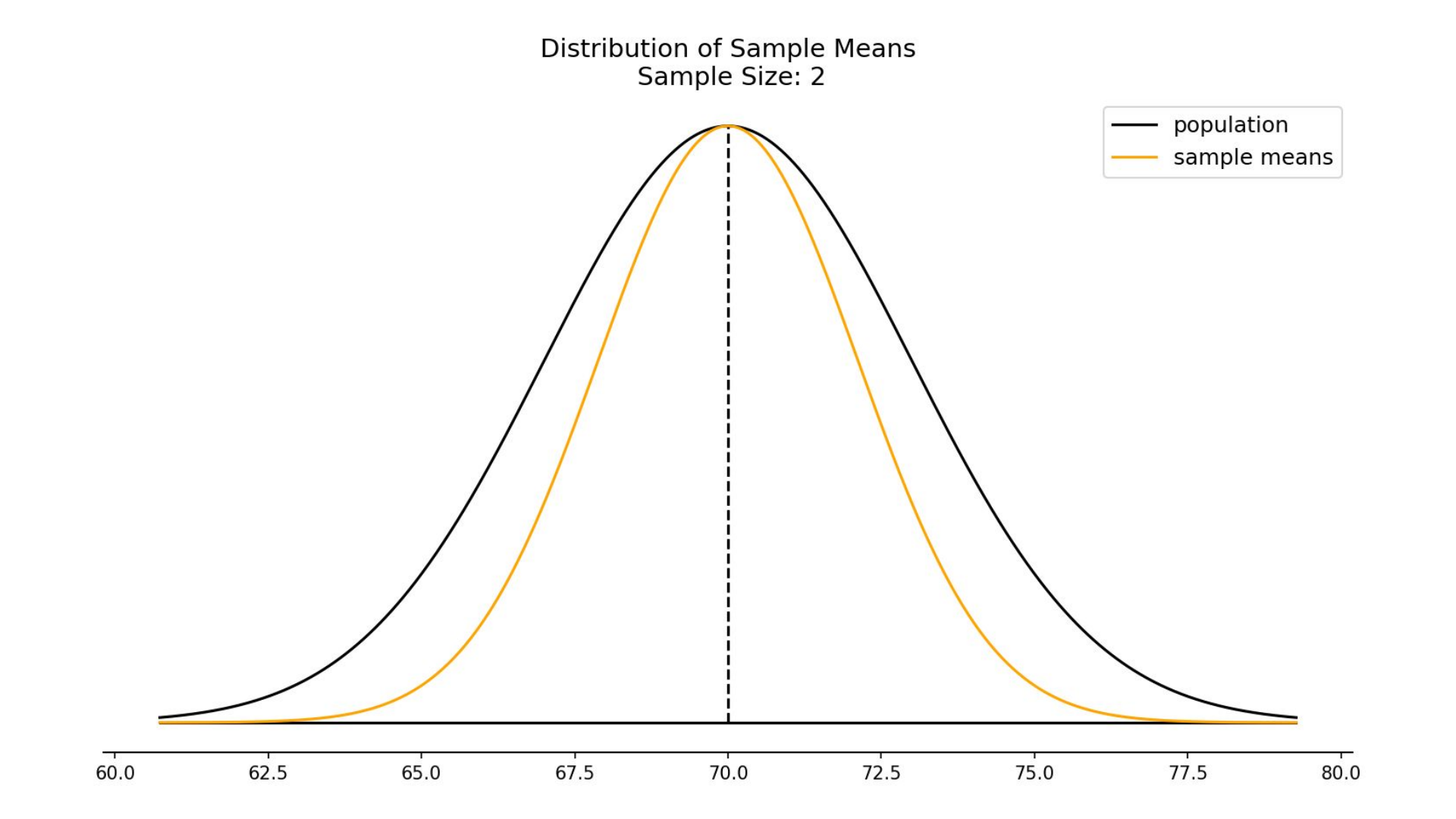
# Really Cool Fact

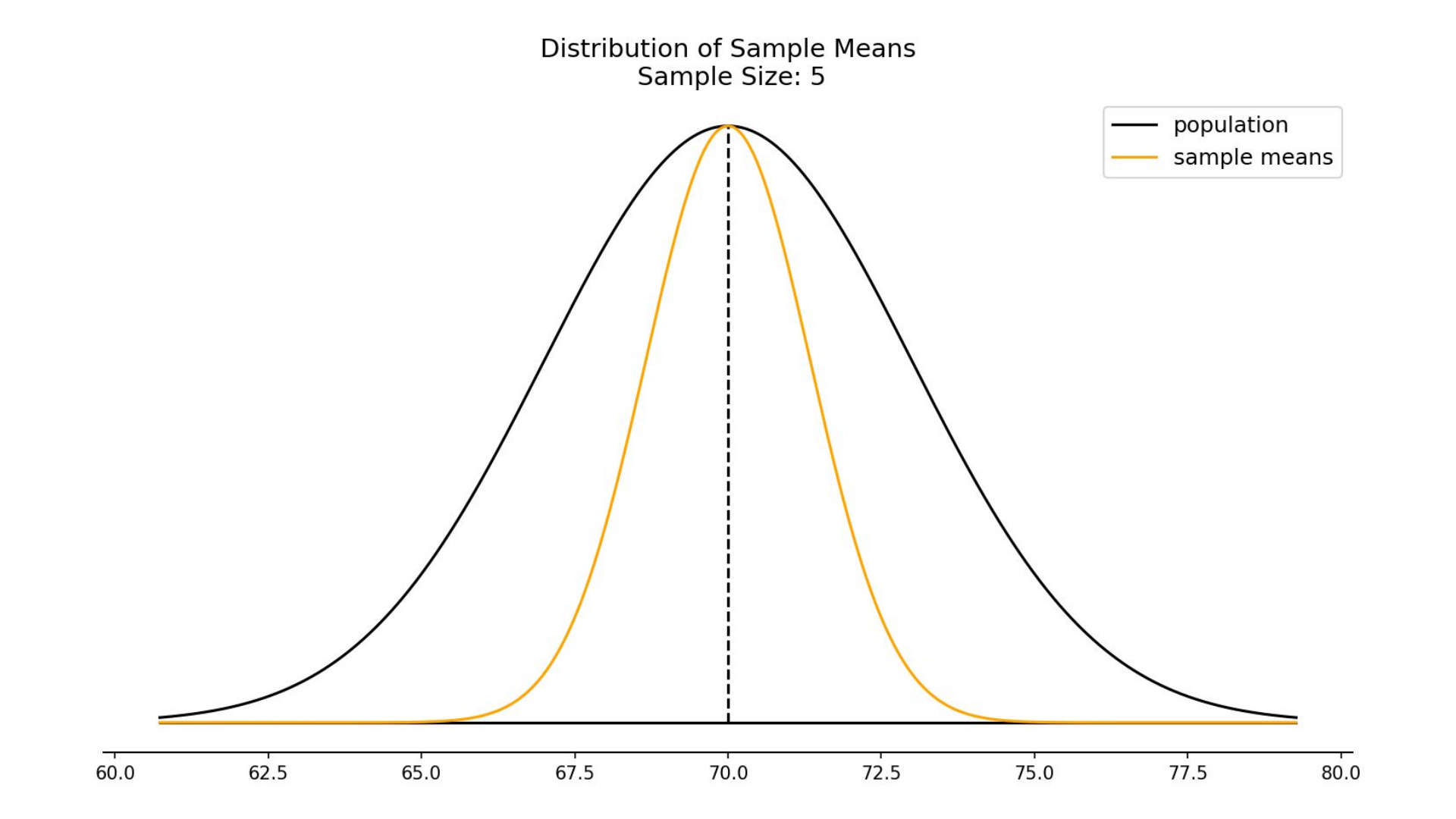
**Central Limit Theorem:** Let $X$ be a random variable with mean $\mu$ and standard deviation $\sigma$.

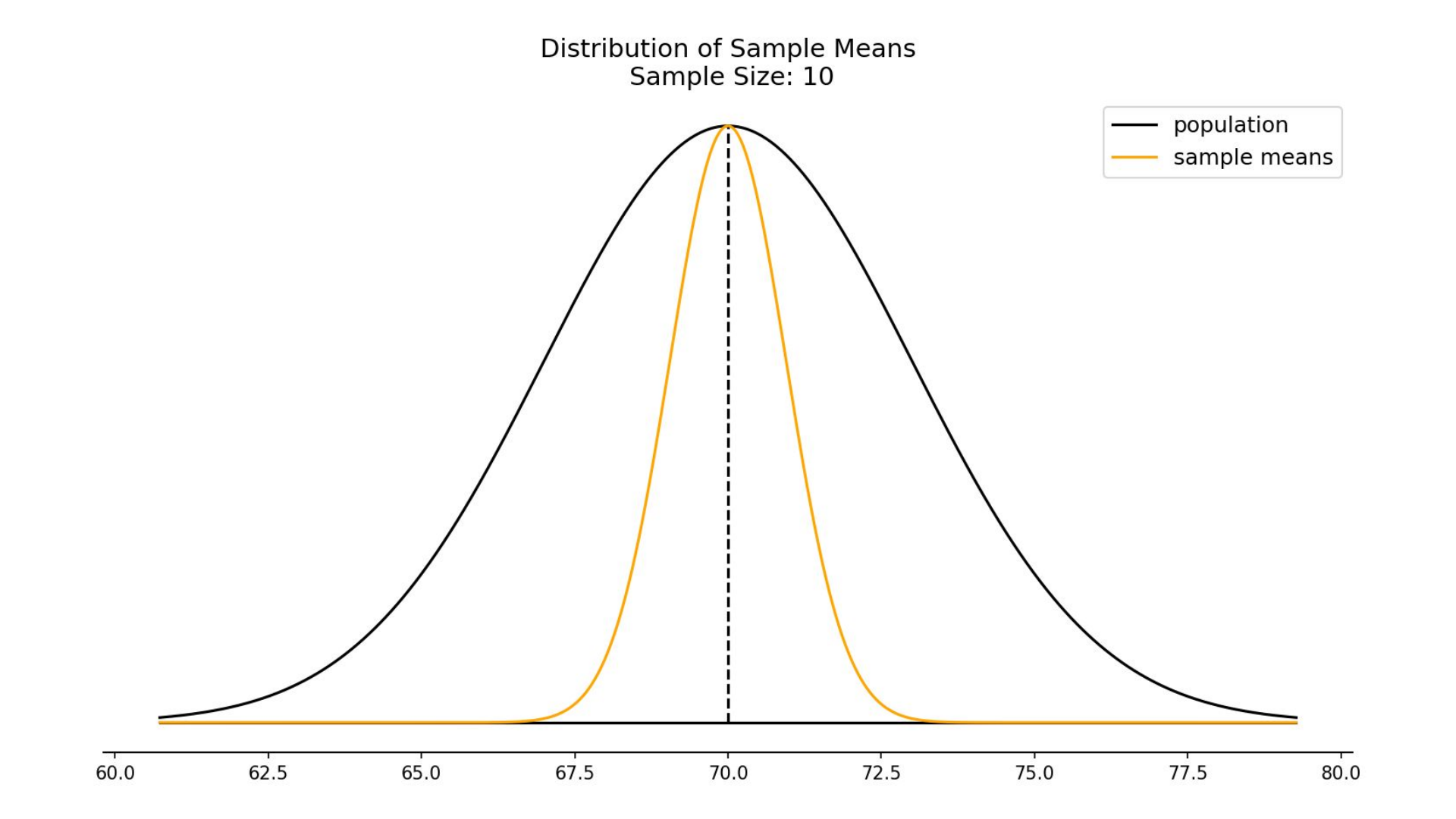
For large enough sample size $n$, (as long as X is reasonably well-behaved), the distribution of sample means is approximately normal with mean $\mu$ and standard deviation $\frac{\sigma}{\sqrt{n}}$

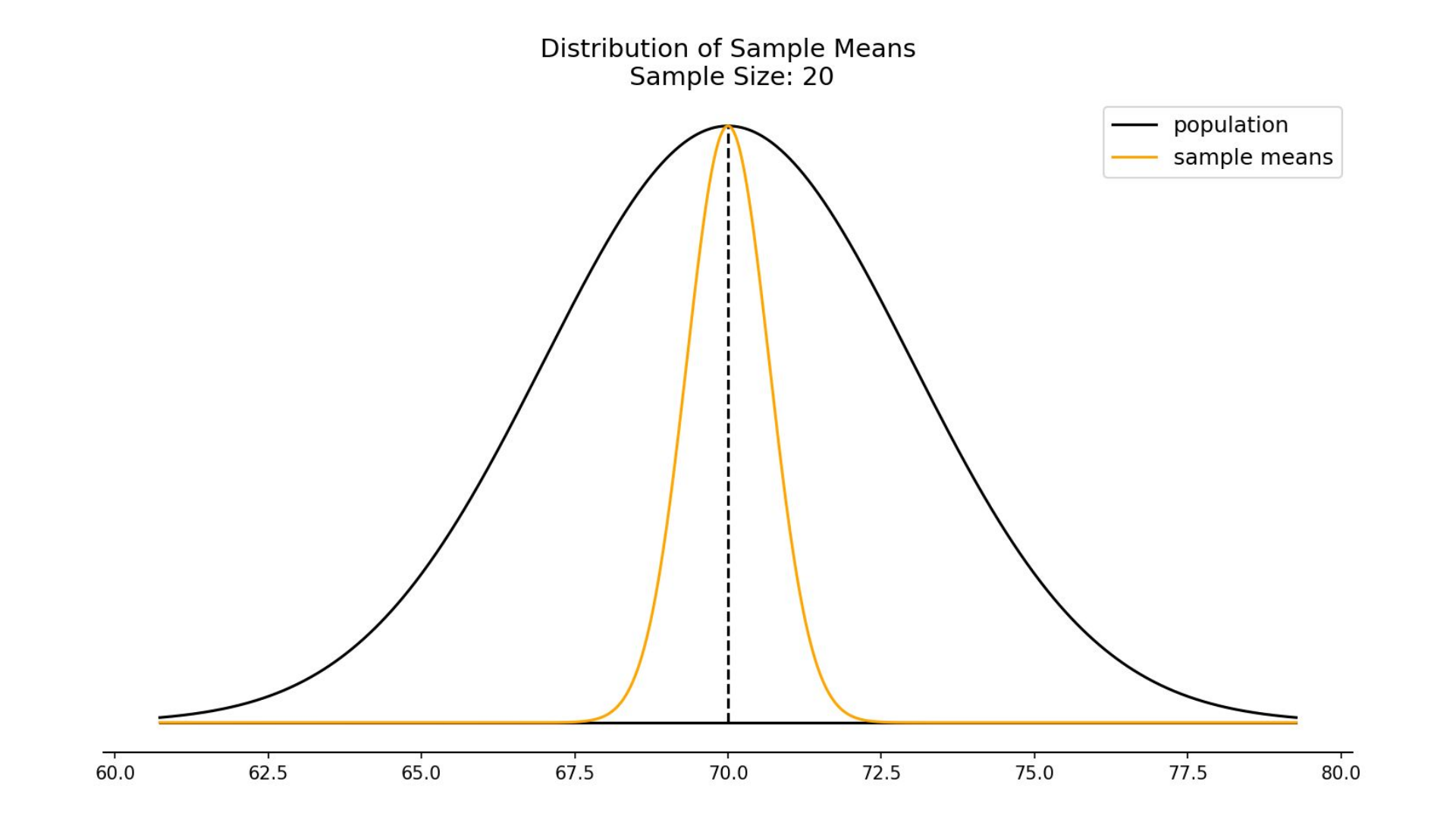
(In a lot of cases, sample size $n = 30$ is good enough for the central limit theorem to kick in).

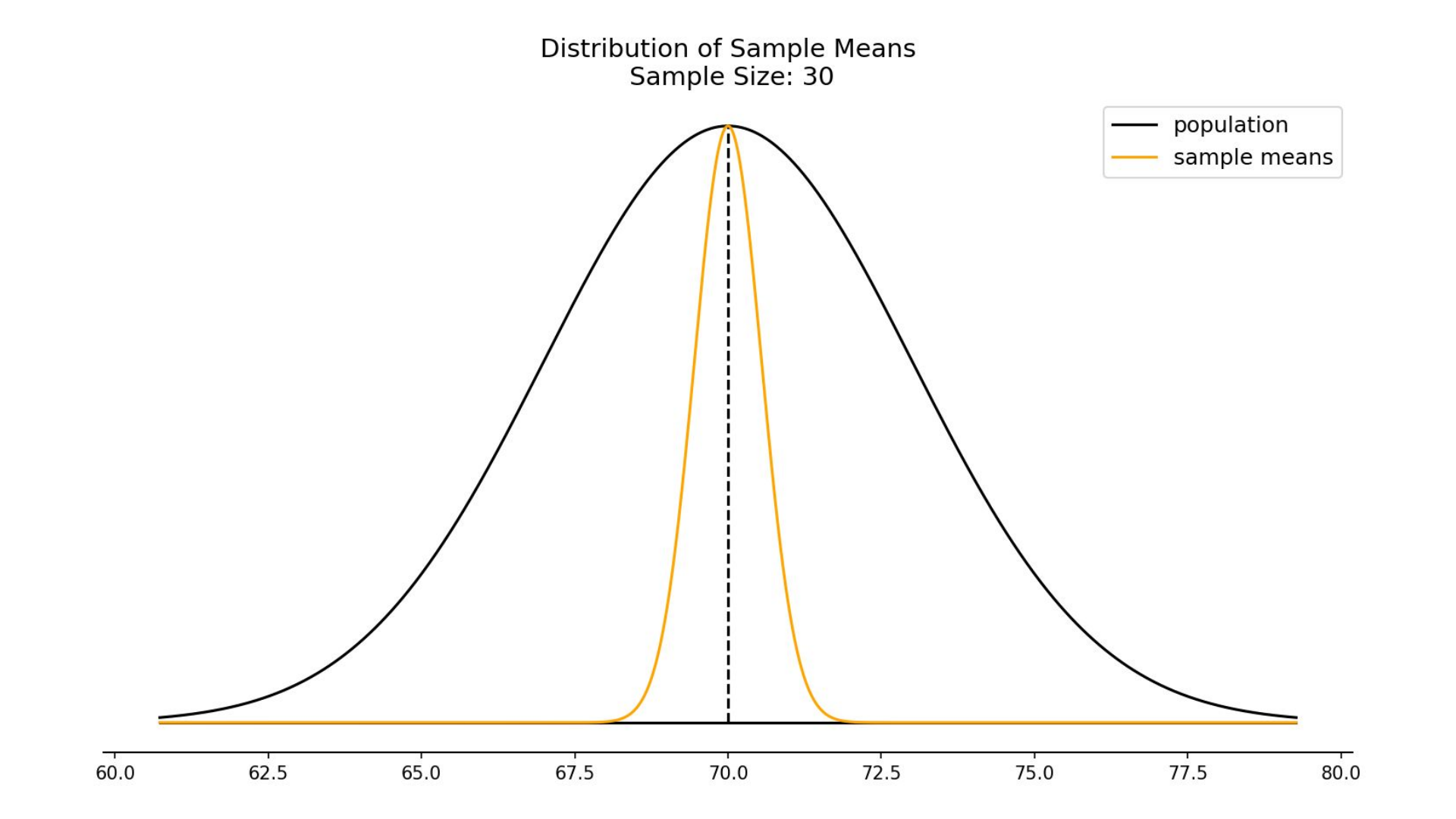

Distribution of Heights of Men
Density
0.12
0.10
0.08
0.06
0.04
0.02
0.00
60
62
64
66
68
70
72
74
76
78
80
Height (Inches)

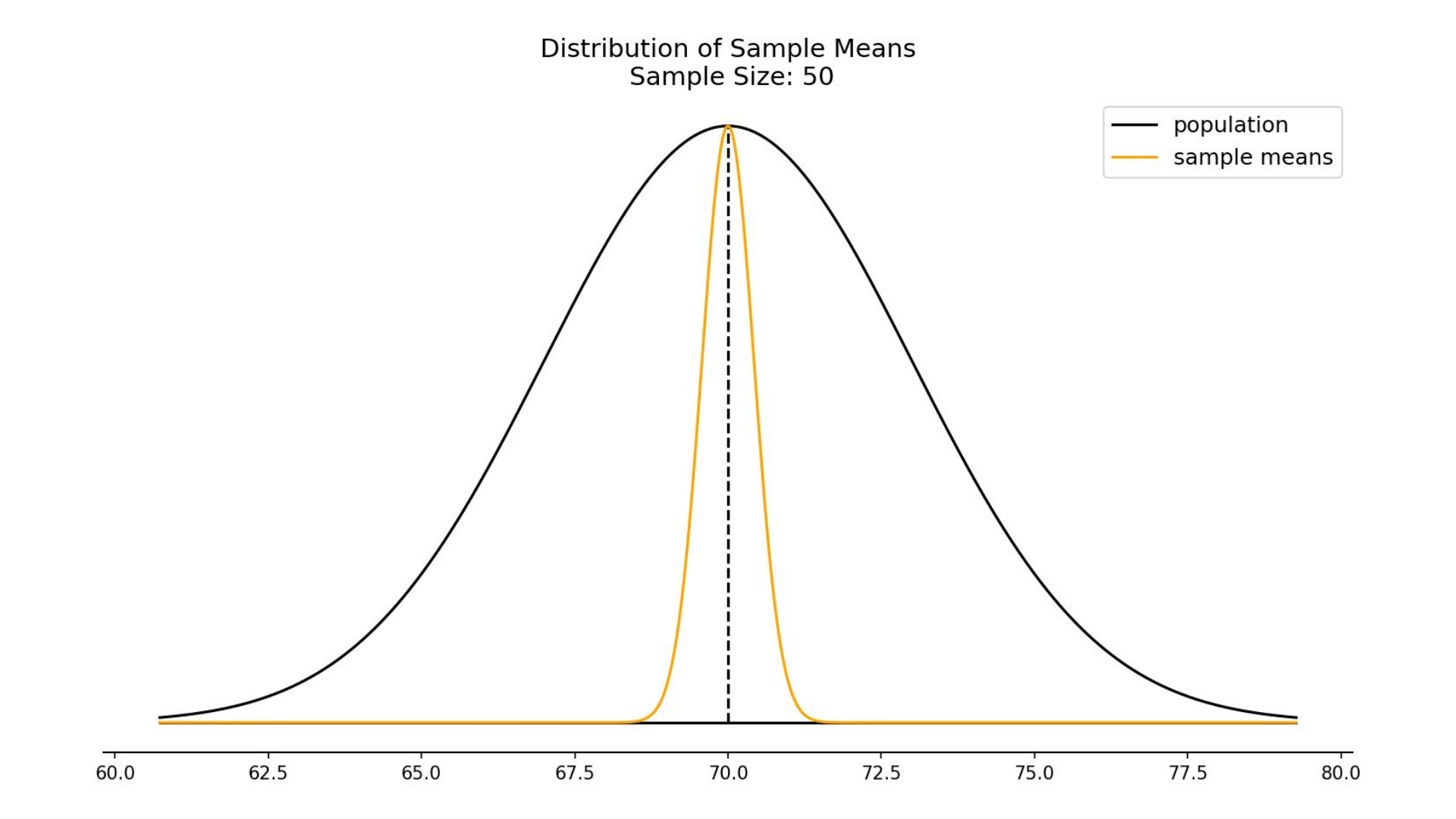

Distribution of Sample Means
Sample Size: 2
population
sample means
60.0
62.5
65.0
67.5
70.0
72.5
75.0
77.5
80.0

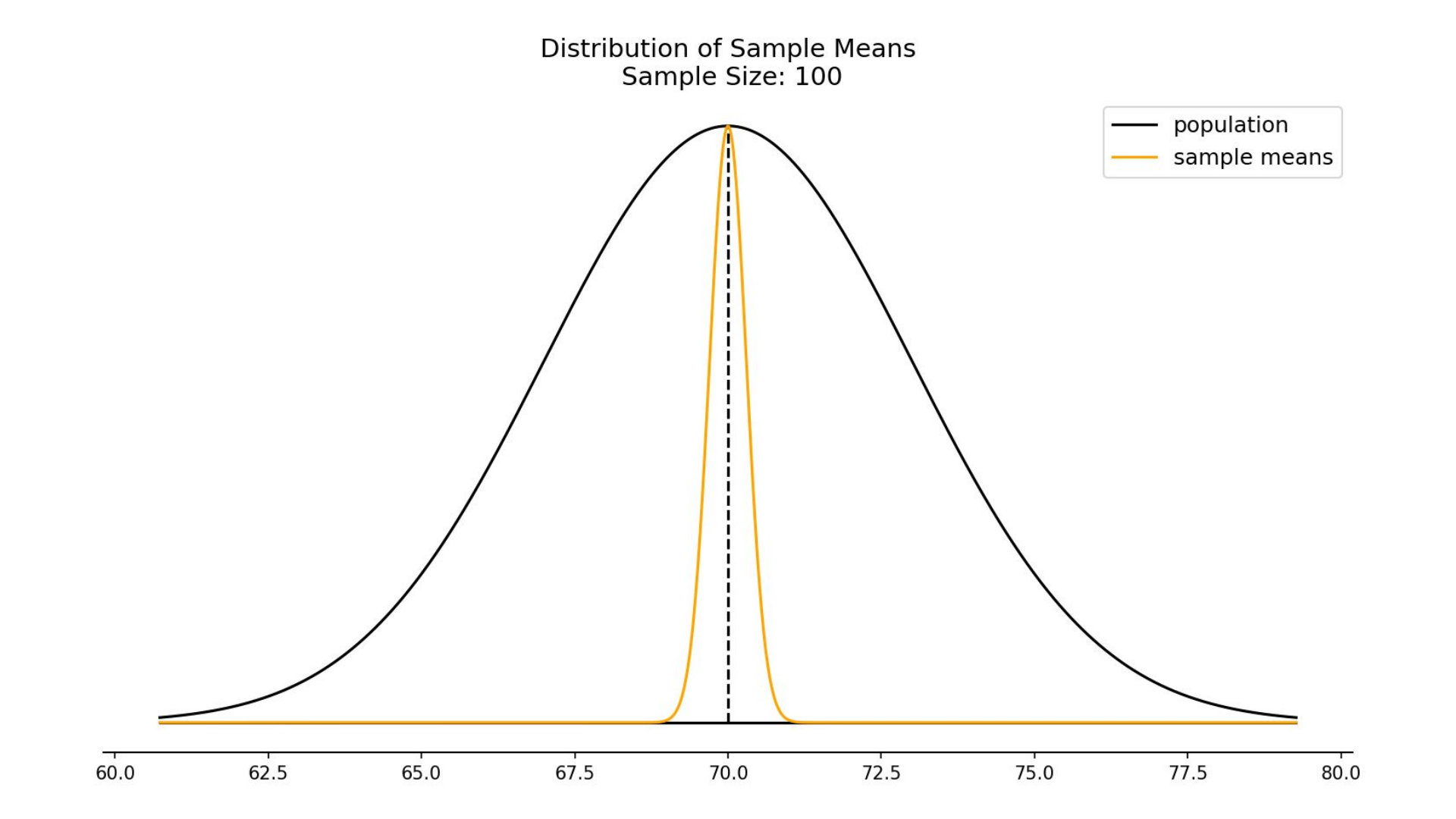

Distribution of Sample Means
Sample Size: 5
population
sample means
60.0
62.5
65.0
67.5
70.0
72.5
75.0
77.5
80.0


Distribution of Sample Means
Sample Size: 10
population
sample means
60.0
62.5
65.0
67.5
70.0
72.5
75.0
77.5
80.0

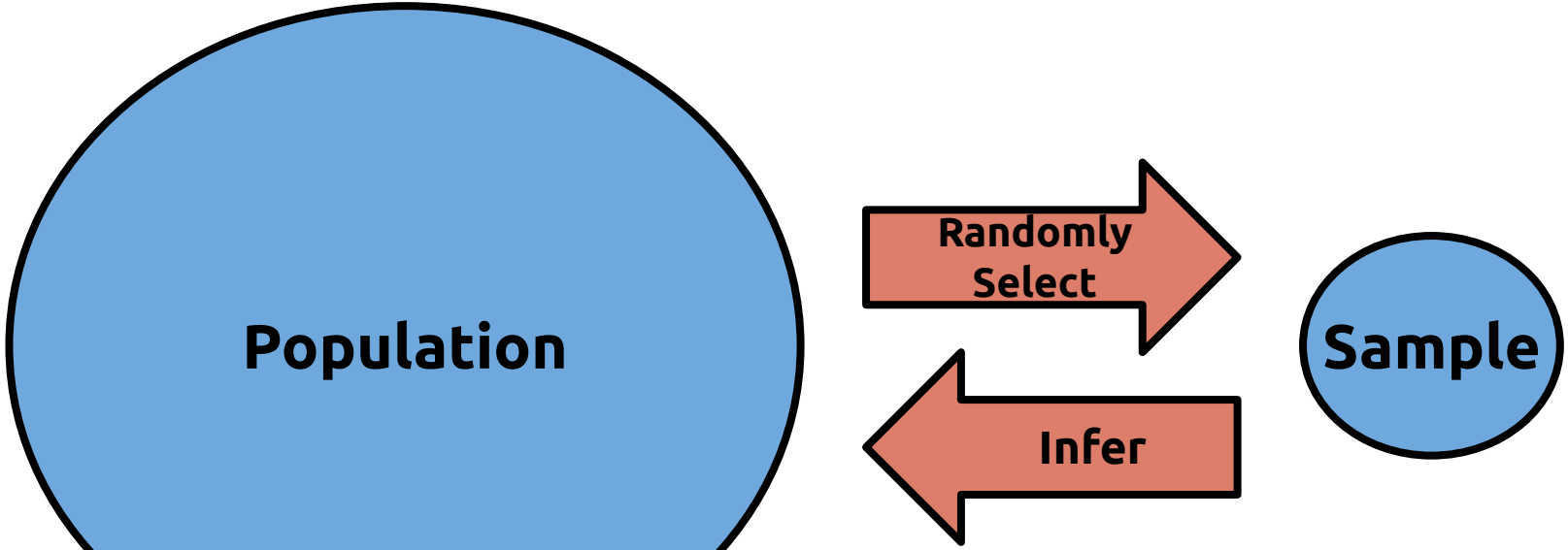

Distribution of Sample Means
Sample Size: 20
population
sample means
60.0
62.5
65.0
67.5
70.0
72.5
75.0
77.5
80.0


Distribution of Sample Means
Sample Size: 30
population
sample means
60.0
62.5
65.0
67.5
70.0
72.5
75.0
77.5
80.0


Distribution of Sample Means
Sample Size: 50
population
sample means
60.0
62.5
65.0
67.5
70.0
72.5
75.0
77.5
80.0


Distribution of Sample Means
Sample Size: 100
population
sample means
60.0
62.5
65.0
67.5
70.0
72.5
75.0
77.5
80.0

The heights of men are approximately normally distributed with mean 70 inches and standard deviation of 3 inches




$$\mu = 70$$

$$\sigma = 3$$

NASHVILLE SOFTWARE SCHOOL

The heights of men are approximately normally distributed with mean 70 inches and standard deviation of 3 inches

**Question:** If we randomly select a man, what is the probability that he is between 69 and 71 inches tall?

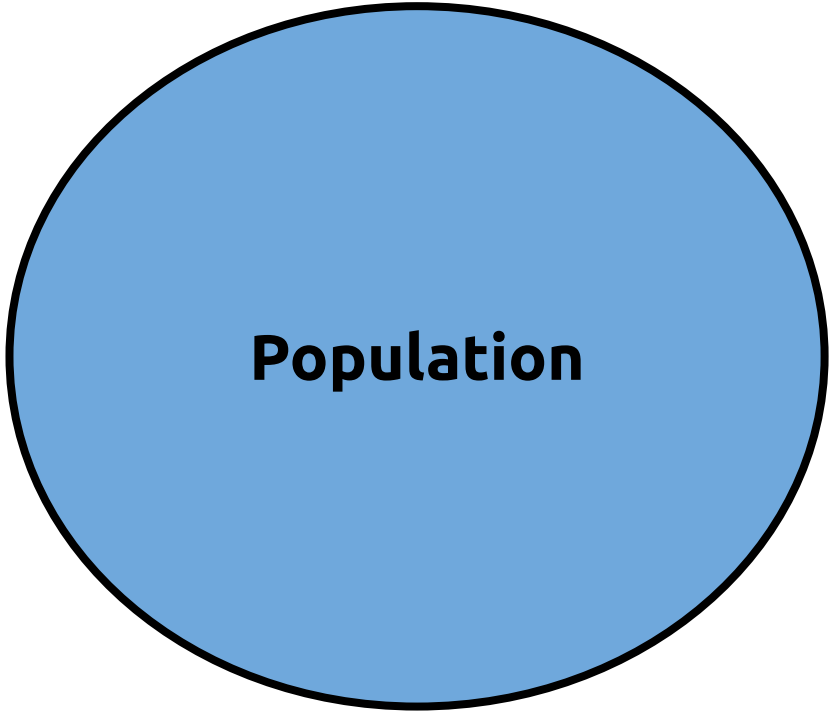
The heights of men are approximately normally distributed with mean 70 inches and standard deviation of 3 inches

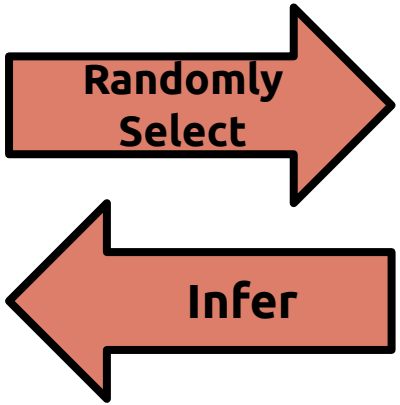
**Question:** If we randomly select a group of 10 men, what is the probability that their average height is between 69 and 71 inches?

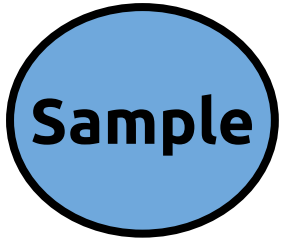
The heights of men are approximately normally distributed with mean 70 inches and standard deviation of 3 inches

**Question:** If we randomly select a group of 100 men, what is the probability that their average height is between 69 and 71 inches?

**Difficult Question:** Say we have a sample of size 20 from a normal distribution with **standard deviation** of 10, and the sample has a sample mean of 50.



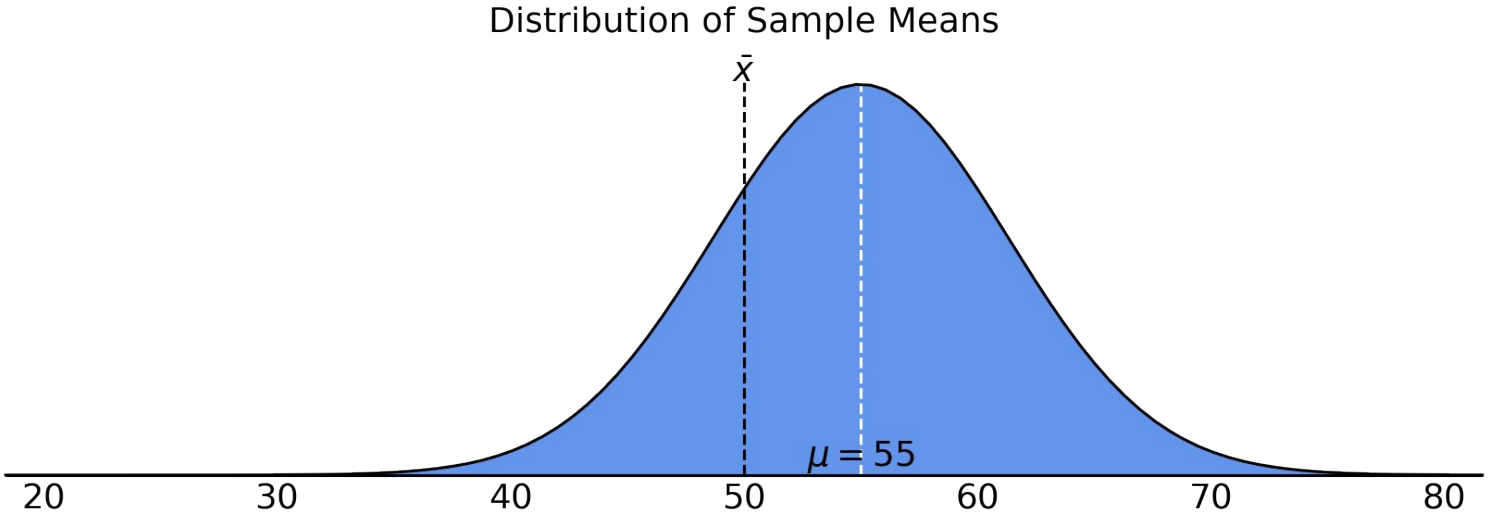

$$\bar{x} = 50$$
$$n = 20$$

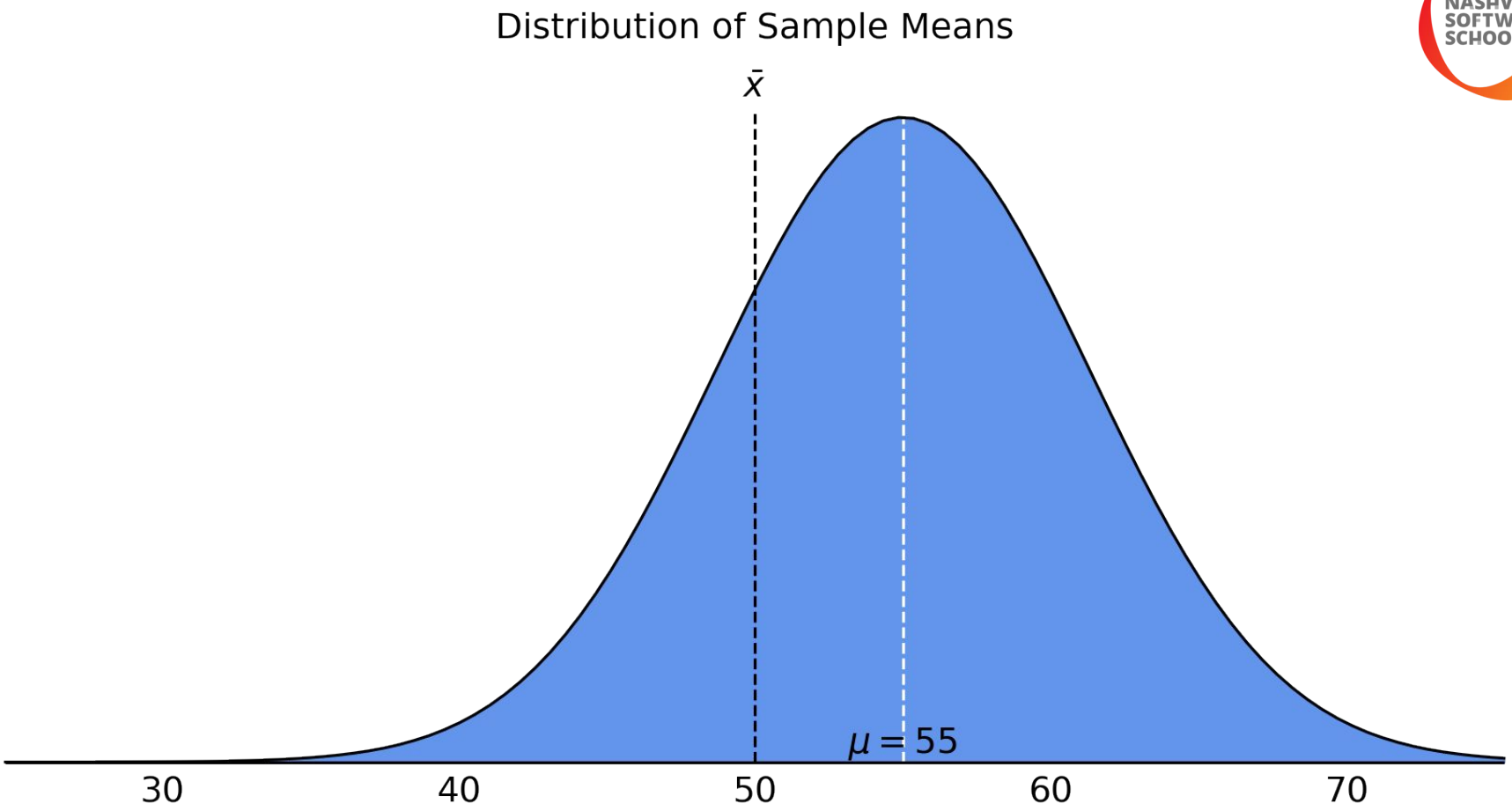
$$\mu = ?$$
$$\sigma = 10$$

NASHVILLE SOFTWARE SCHOOL

**Difficult Question:** Say we have a sample of size 20 from a normal distribution with **standard deviation** of 10, and the sample has a sample mean of 50.

Is it possible that the *population* mean is actually 55? Does it seem plausible that the population mean is actually 55?

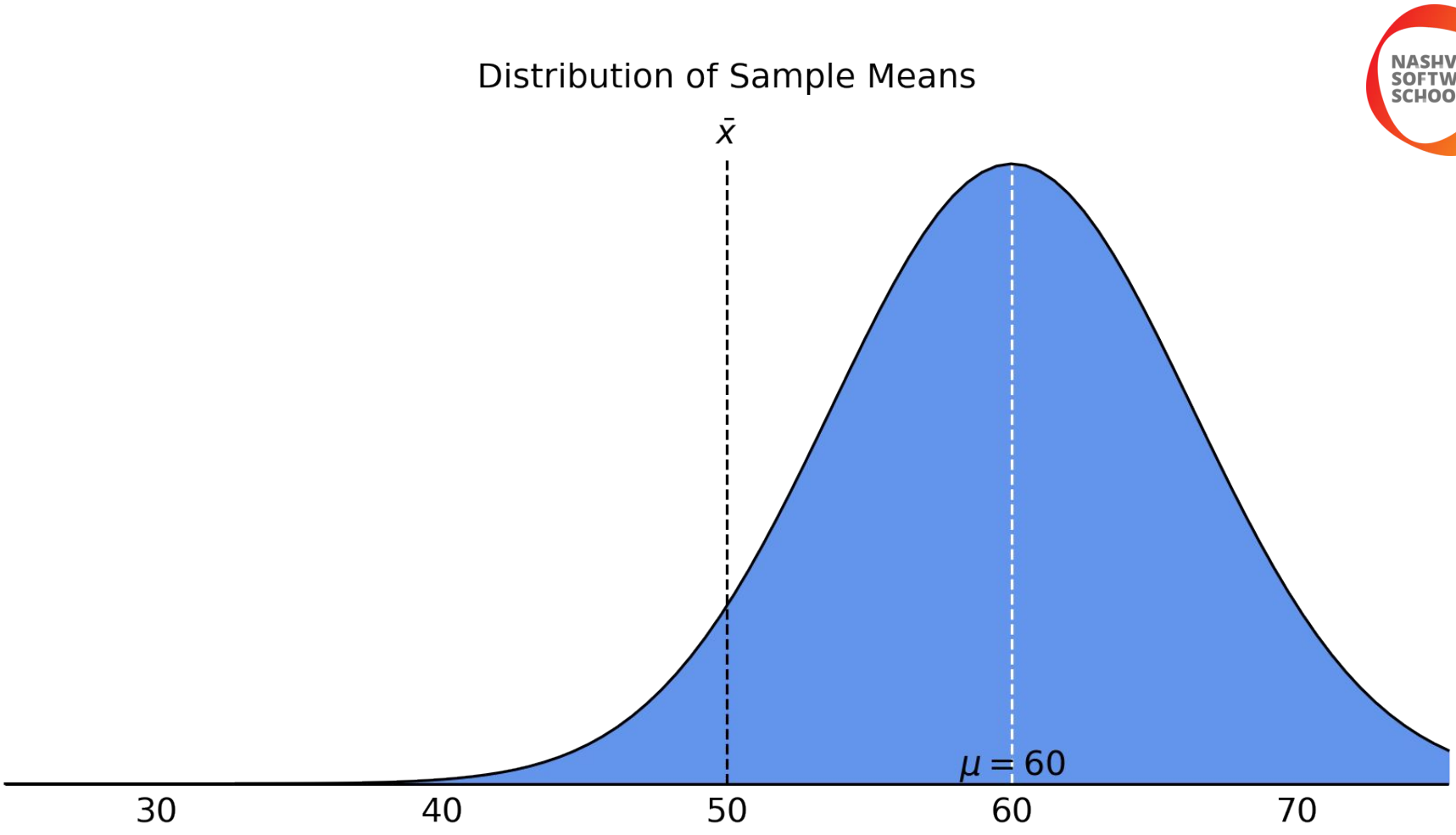



Distribution of Sample Means
$\bar{x}$
$\mu = 55$
30
40
50
60
70

**Difficult Question:** Say we have a sample of size 20 from a normal distribution with **standard deviation** of 10, and the sample has a sample mean of 50.

Is it possible that the *population* mean is actually 55? Does it seem plausible that the population mean is actually 55?

Is it possible that the *population* mean is actually 60? Does it seem plausible that the population mean is actually 60?

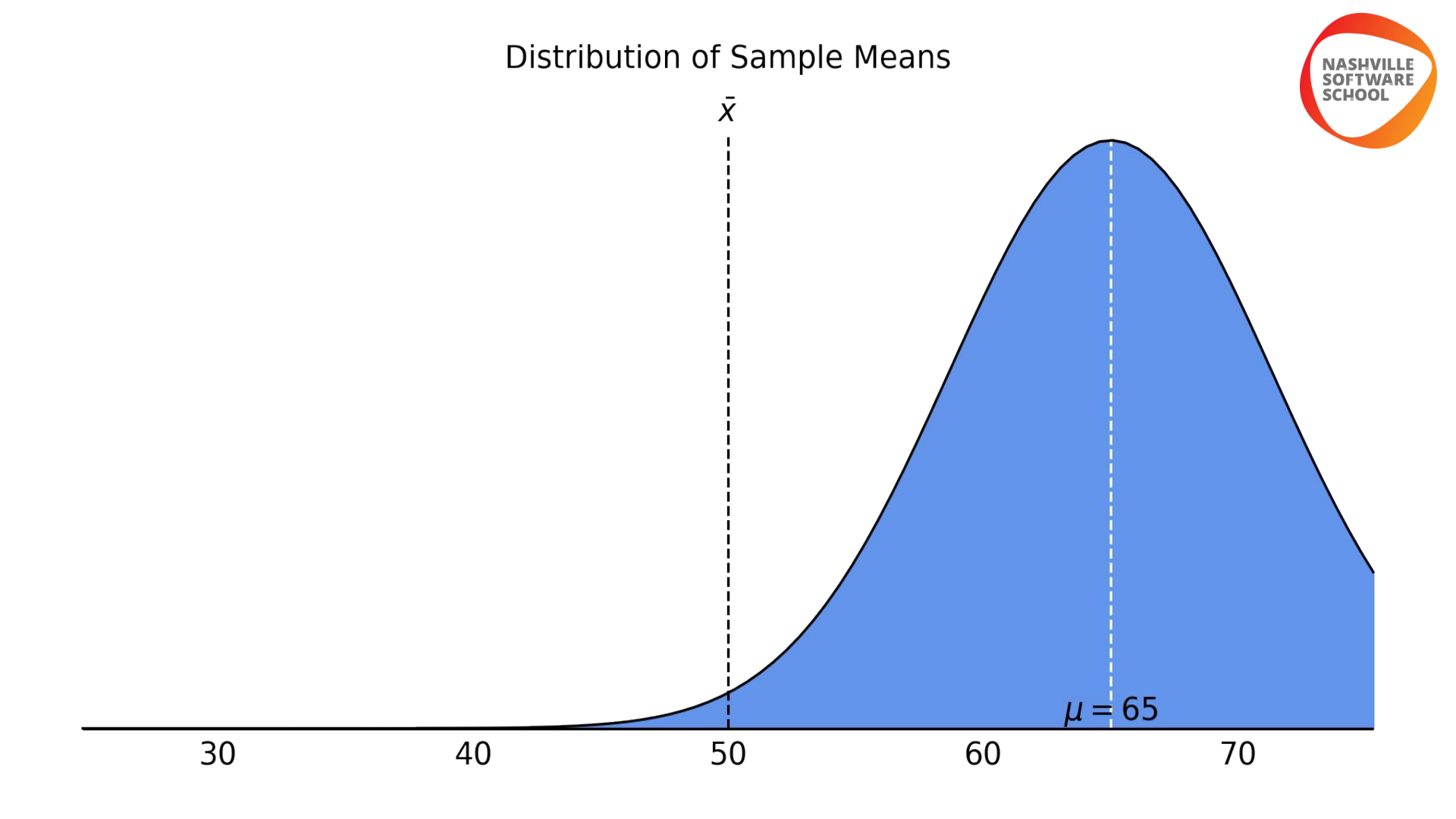

Distribution of Sample Means
$\bar{x}$
$\mu = 60$
30
40
50
60
70

**Difficult Question:** Say we have a sample of size 20 from a normal distribution with **standard deviation** of 10, and the sample has a sample mean of 50.

Is it possible that the *population* mean is actually 55? Does it seem plausible that the population mean is actually 55?

Is it possible that the *population* mean is actually 60? Does it seem plausible that the population mean is actually 60?

Is it possible that the *population* mean is actually 65? Does it seem plausible that the population mean is actually 65?


Distribution of Sample Means
$\bar{x}$
$\mu = 65$
30
40
50
60
70

# Cool Fact:

If *X* is a normally-distributed variable with mean ***μ*** and standard deviation ***σ***, the distribution of sample means of size *n* is also normal with mean ***μ*** and standard deviation $\frac{\sigma}{\sqrt{n}}$

# Cool Fact:

If *X* is a normally-distributed variable with mean ***μ*** and standard deviation ***σ***, the distribution of sample means of size *n* is also normal with mean ***μ*** and standard deviation $\frac{\sigma}{\sqrt{n}}$

This means that $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ approximately follows a standard normal distribution (mean 0 and standard deviation 1).

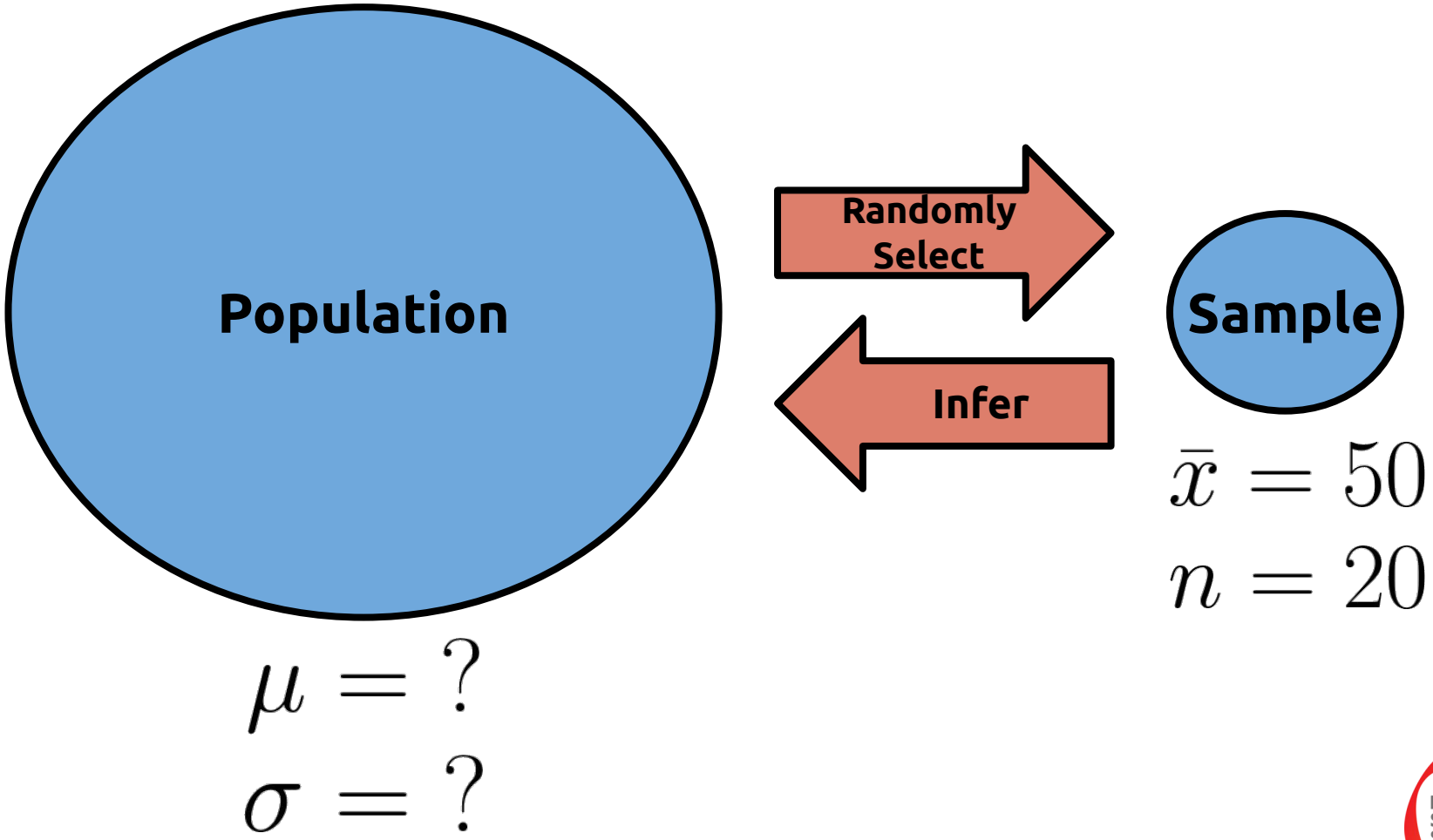
# Why do we care about all of this?

Knowing something about the distribution of sample means, we can determine what a “plausible” range of values for $\mu$.

More specifically, knowing about the standard deviation of the sample means and the shape of its distribution (normal), helps us determine a plausible range of values.

The key thing was knowing the **standard error** - the standard deviation of the distribution of sample means - is equal to $\frac{\sigma}{\sqrt{n}}$ , where $\sigma$ is the population standard deviation.

But wait - why would we know the population standard deviation but not the mean?




NASHVILLE SOFTWARE SCHOOL

If we don’t know the population standard deviation, how could we estimate it using just our single sample?

If we don’t know the population standard deviation, how could we estimate it using just our single sample?

By using the sample standard deviation, *s*.

If we don’t know the population standard deviation, how could we estimate it using just our single sample?

By using the sample standard deviation, *s*.

$$\frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$ follows a standard normal distribution

If we don't know the population standard deviation, how could we estimate it using just our single sample?

By using the sample standard deviation, *s*.

$$\frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

follows a standard normal distribution

$$\frac{\bar{x} - \mu}{s / \sqrt{n}}$$

follows a Student's *t*-distribution with n-1 degrees of freedom

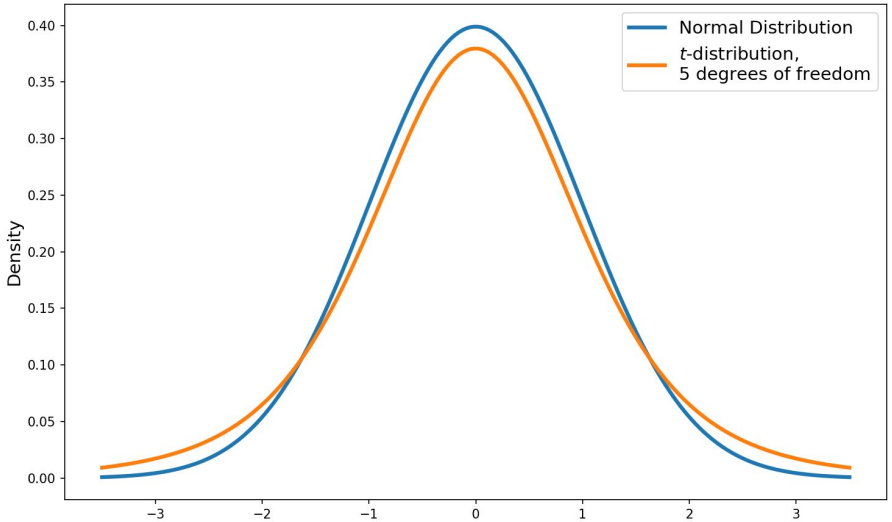
# Estimation

The family of Student's *t*-distribution is named after statistician William Sealy Gosset, who published his research under the pseudonym "Student".

Gosset worked for the Guinness Brewery where he worked on determining the quality of raw materials. Gosset was interested in the problem of small samples, as he would sometimes have to draw inferences from samples with as few as 3 observations.

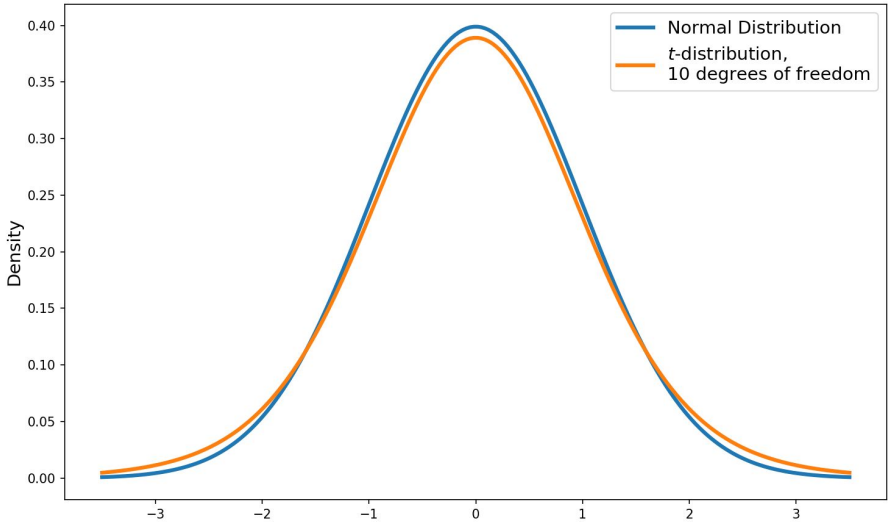
# Estimation

The family of Student's *t*-distributions is a family of symmetric, bell-shaped distributions, which are similar to normal distributions, but have wider tails. That is, more extreme observations are more common. This family is parametrized by the number of degrees of freedom.



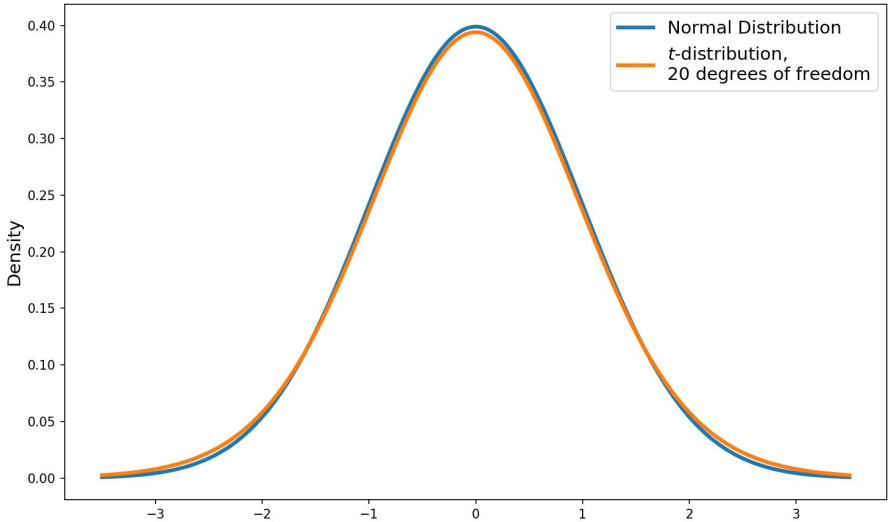
# Estimation

The family of Student's *t*-distributions is a family of symmetric, bell-shaped distributions, which are similar to normal distributions, but have wider tails. That is, more extreme observations are more common. This family is parametrized by the number of degrees of freedom.



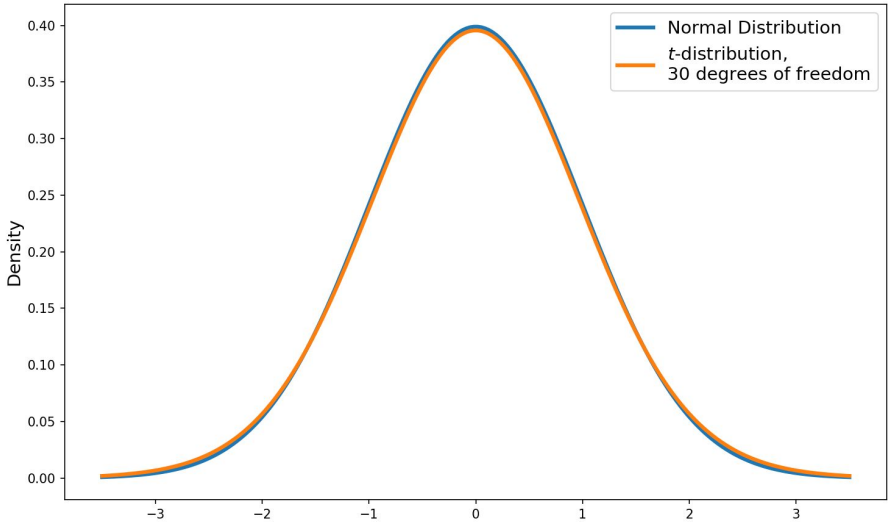
# Estimation

The family of Student's *t*-distributions is a family of symmetric, bell-shaped distributions, which are similar to normal distributions, but have wider tails. That is, more extreme observations are more common. This family is parametrized by the number of degrees of freedom.



# Estimation

The family of Student's *t*-distributions is a family of symmetric, bell-shaped distributions, which are similar to normal distributions, but have wider tails. That is, more extreme observations are more common. This family is parametrized by the number of degrees of freedom.



# Confidence Intervals

**Goal:** For a given sample, construct an interval around the sample mean just wide enough so that for 95% of samples, the interval constructed in this way will contain the true population mean $\mu$.

NASHVILLE SOFTWARE SCHOOL

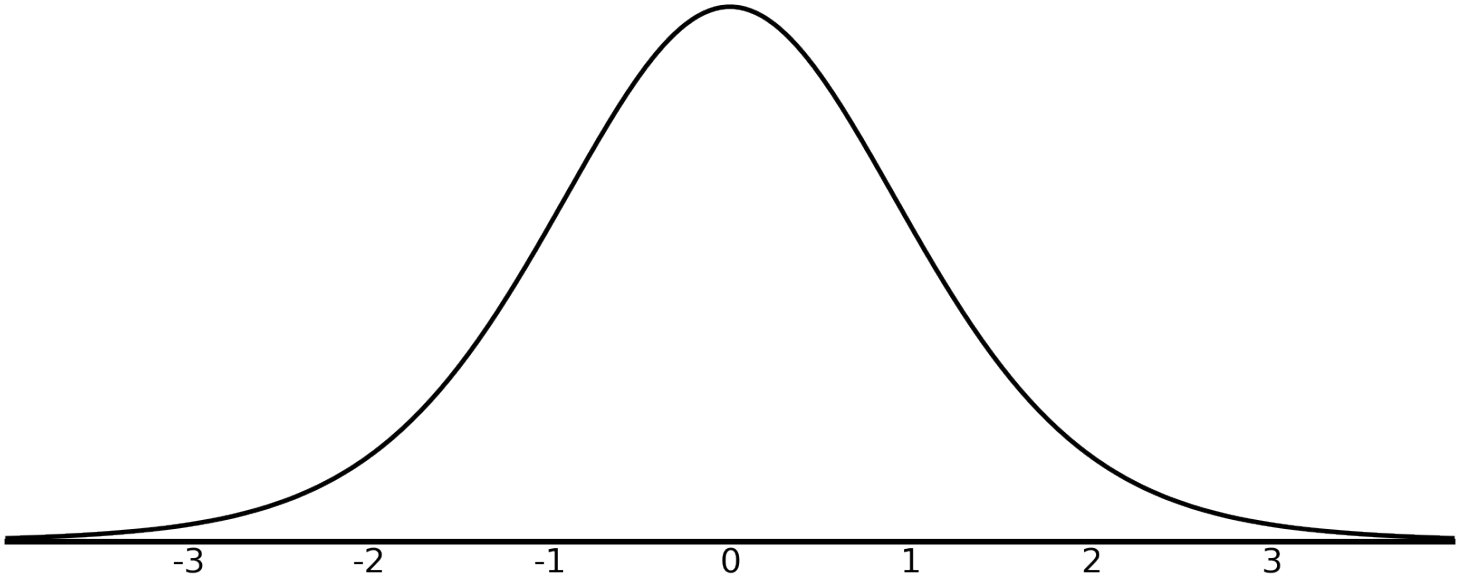
# Confidence Intervals

**Goal:** For a given sample, construct an interval around the sample mean just wide enough so that for 95% of samples, the interval constructed in this way will contain the true population mean $\mu$.

To do this, forget about our single sample temporarily and think about *all* possible samples.

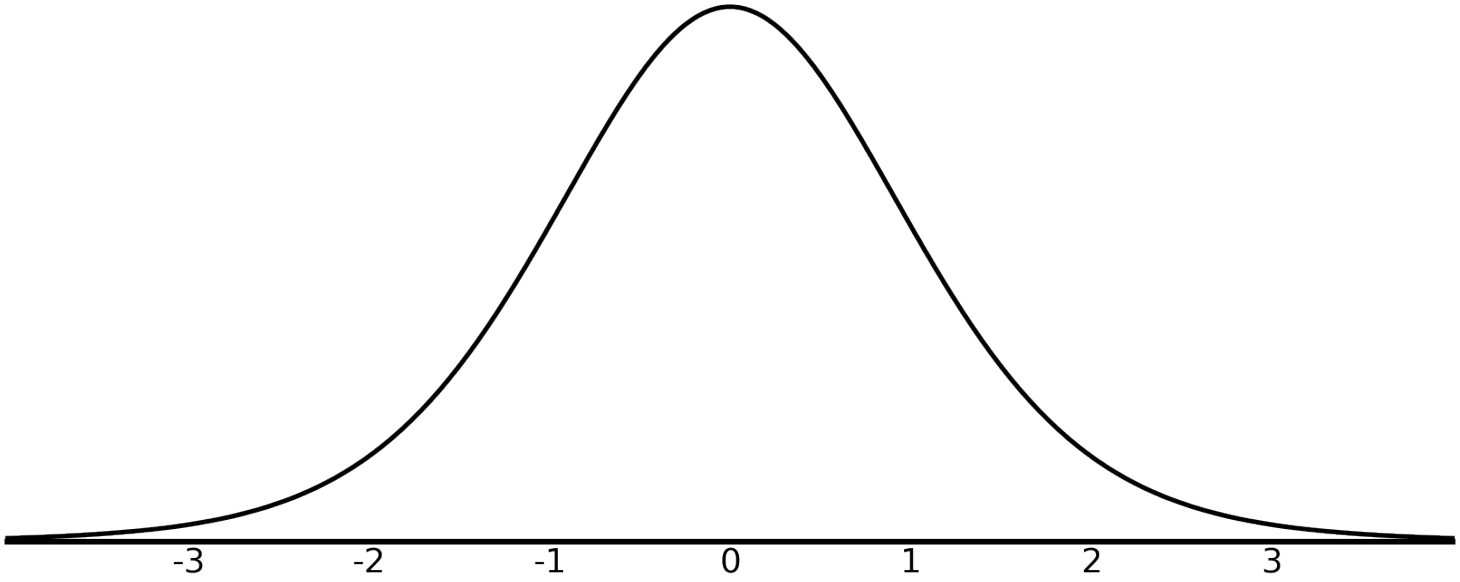
NASHVILLE SOFTWARE SCHOOL

# Distribution of $\frac{\bar{x} - \mu}{s/\sqrt{n}}$




NASHVILLE SOFTWARE SCHOOL

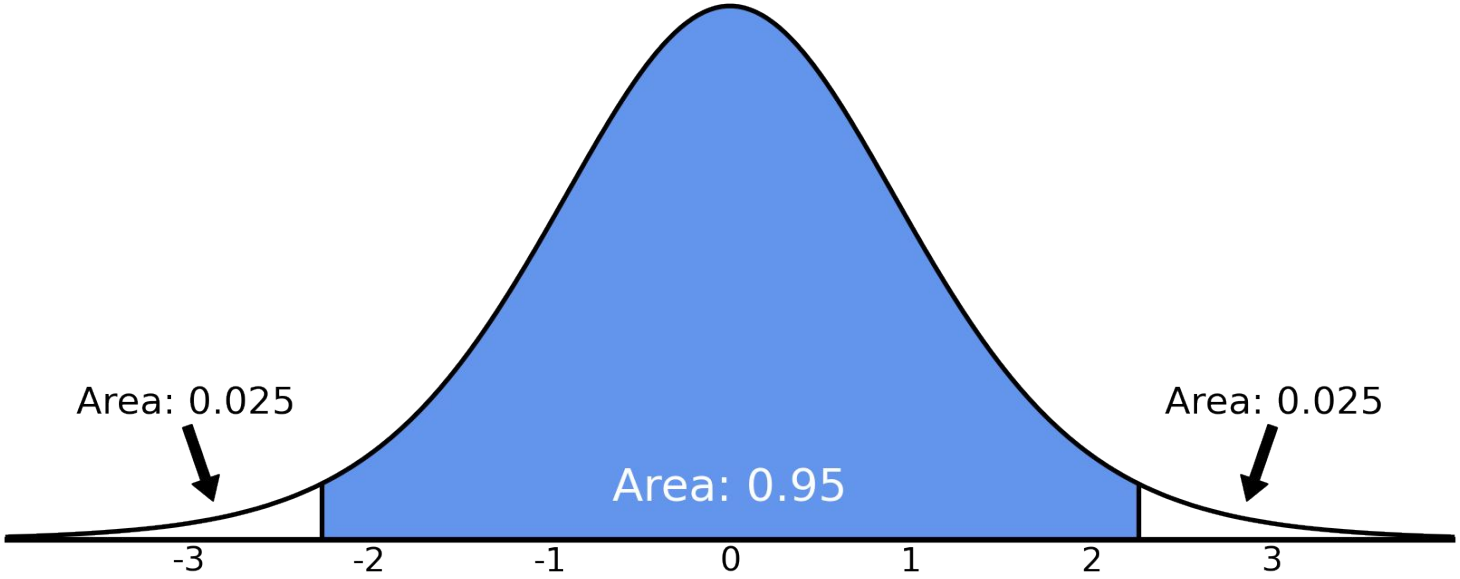
Distribution of $\frac{\bar{x} - \mu}{s/\sqrt{n}}$




Student's *t*-distribution with n-1 degrees of freedom

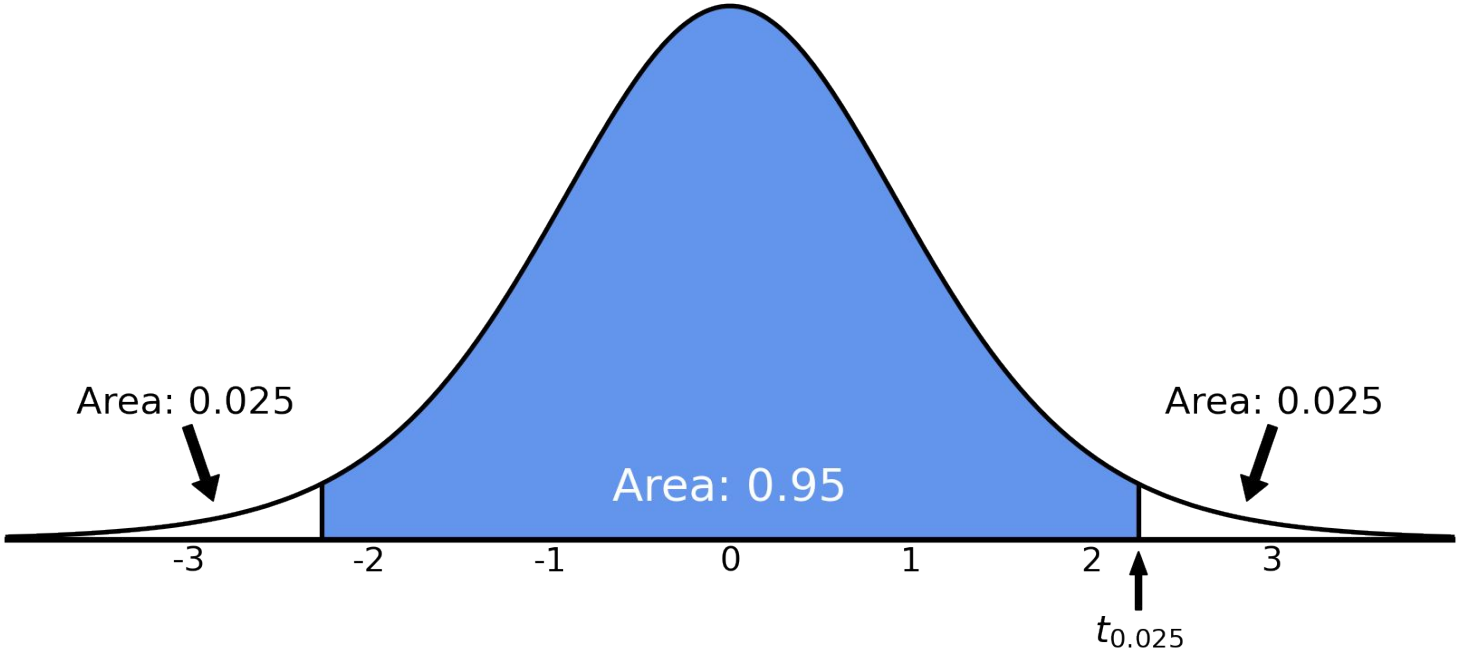
NASHVILLE SOFTWARE SCHOOL

# Distribution of $\frac{\bar{x} - \mu}{s/\sqrt{n}}$



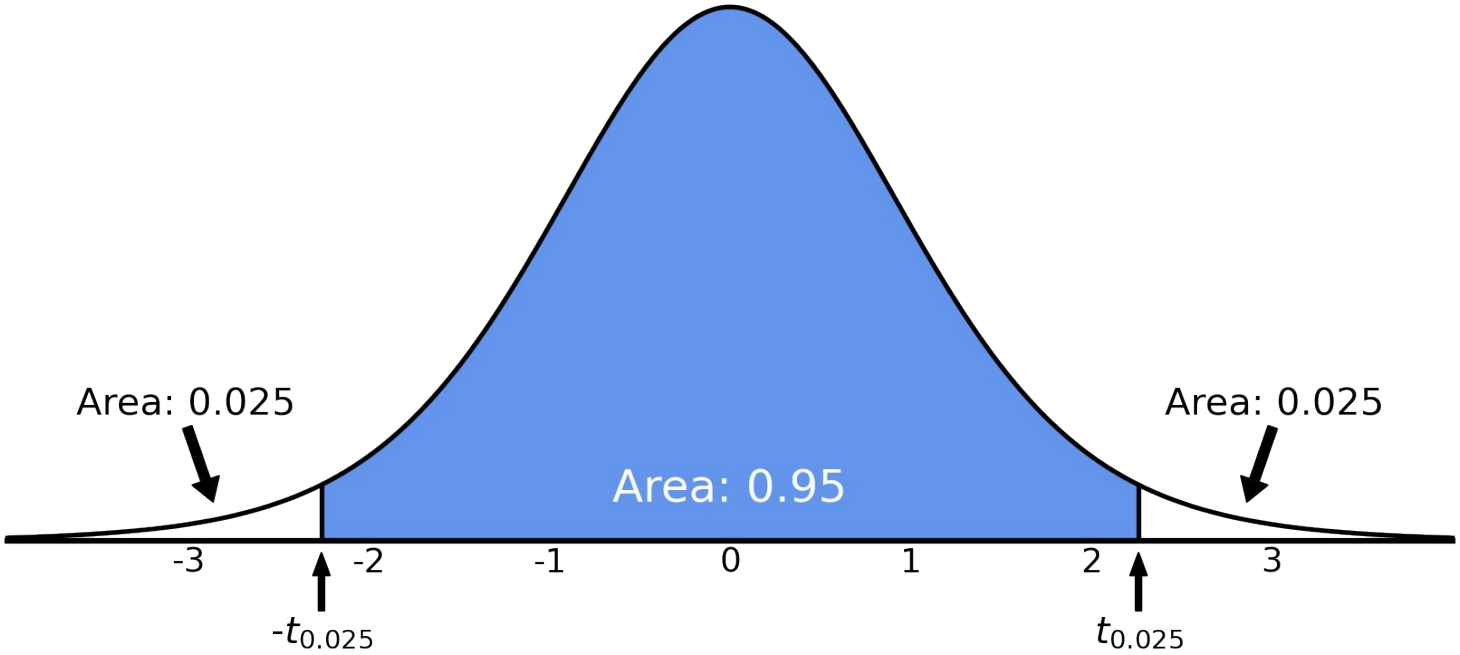

NASHVILLE SOFTWARE SCHOOL

# Distribution of $\frac{\bar{x} - \mu}{s/\sqrt{n}}$



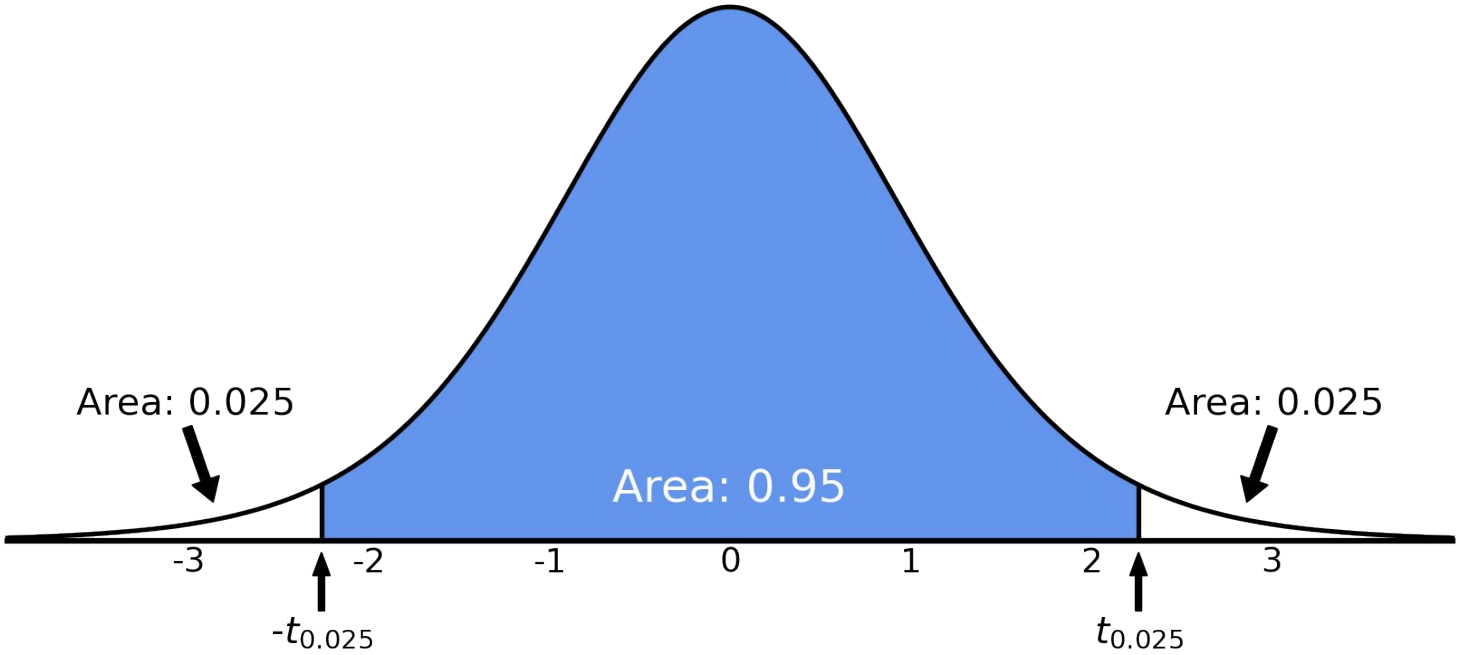

NASHVILLE SOFTWARE SCHOOL

# Distribution of $\frac{\bar{x} - \mu}{s/\sqrt{n}}$




NASHVILLE SOFTWARE SCHOOL

Distribution of $\frac{\bar{x} - \mu}{s/\sqrt{n}}$




For 95% of samples, $-t_{0.025} < \frac{\bar{x} - \mu}{s/\sqrt{n}} < t_{0.025}$

NASHVILLE SOFTWARE SCHOOL

For 95% of samples, $-t_{0.025} < \dfrac{\bar{x} - \mu}{s/\sqrt{n}} < t_{0.025}$

For 95% of samples, $-t_{0.025} < \frac{\bar{x} - \mu}{s/\sqrt{n}} < t_{0.025}$

For 95% of samples, $-t_{0.025} \cdot \frac{s}{\sqrt{n}} < \bar{x} - \mu < t_{0.025} \cdot \frac{s}{\sqrt{n}}$

For 95% of samples, $-t_{0.025} < \frac{\bar{x} - \mu}{s/\sqrt{n}} < t_{0.025}$

For 95% of samples, $-t_{0.025} \cdot \frac{s}{\sqrt{n}} < \bar{x} - \mu < t_{0.025} \cdot \frac{s}{\sqrt{n}}$

For 95% of samples, $-\bar{x} - t_{0.025} \cdot \frac{s}{\sqrt{n}} < -\mu < -\bar{x} + t_{0.025} \cdot \frac{s}{\sqrt{n}}$

For 95% of samples, $-t_{0.025} < \frac{\bar{x} - \mu}{s/\sqrt{n}} < t_{0.025}$

For 95% of samples, $-t_{0.025} \cdot \frac{s}{\sqrt{n}} < \bar{x} - \mu < t_{0.025} \cdot \frac{s}{\sqrt{n}}$

For 95% of samples, $-\bar{x} - t_{0.025} \cdot \frac{s}{\sqrt{n}} < -\mu < -\bar{x} + t_{0.025} \cdot \frac{s}{\sqrt{n}}$

For 95% of samples, $\bar{x} - t_{0.025} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{0.025} \cdot \frac{s}{\sqrt{n}}$

For 95% of samples,

$$\bar{x} - t_{0.025} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{0.025} \cdot \frac{s}{\sqrt{n}}$$

NASHVILLE SOFTWARE SCHOOL

For 95% of samples,

$$\bar{x} - t_{0.025} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{0.025} \cdot \frac{s}{\sqrt{n}}$$

Hence, we can use this recipe to build a 95% confidence interval:

NASHVILLE SOFTWARE SCHOOL

For 95% of samples,

$$\bar{x} - t_{0.025} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{0.025} \cdot \frac{s}{\sqrt{n}}$$

Hence, we can use this recipe to build a 95% confidence interval:

$$\bar{x} \pm t_{0.025} \cdot \frac{s}{\sqrt{n}}$$

NASHVILLE SOFTWARE SCHOOL

For 95% of samples,

$$\bar{x} - t_{0.025} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{0.025} \cdot \frac{s}{\sqrt{n}}$$

Hence, we can use this recipe to build a 95% confidence interval:

$$\bar{x} \pm t_{0.025} \cdot \frac{s}{\sqrt{n}}$$

Point Estimate

NASHVILLE SOFTWARE SCHOOL

For 95% of samples,

$$\bar{x} - t_{0.025} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{0.025} \cdot \frac{s}{\sqrt{n}}$$

Hence, we can use this recipe to build a 95% confidence interval:

$$\bar{x} \pm t_{0.025} \cdot \frac{s}{\sqrt{n}}$$

Point Estimate

Margin of Error

NASHVILLE SOFTWARE SCHOOL

# Estimation - Proportions

What does the sampling distribution for a proportion look like?

That is, if we take a sample from a population and calculate the proportion of observations in our sample that meet a particular condition, what does the distribution of the possible sample proportions look like?

Eg. When we conduct a poll, we might find the proportion of respondents planning to vote for a particular candidate.

NASHVILLE SOFTWARE SCHOOL

# Estimation - Proportions

Say we have a sample size of *n* and the the true population proportion is *p*.

Let $\hat{p}$ be the sample proportion.

We know that $n \cdot \hat{p}$ follows a binomial distribution with $n = n$ and $p = p$.

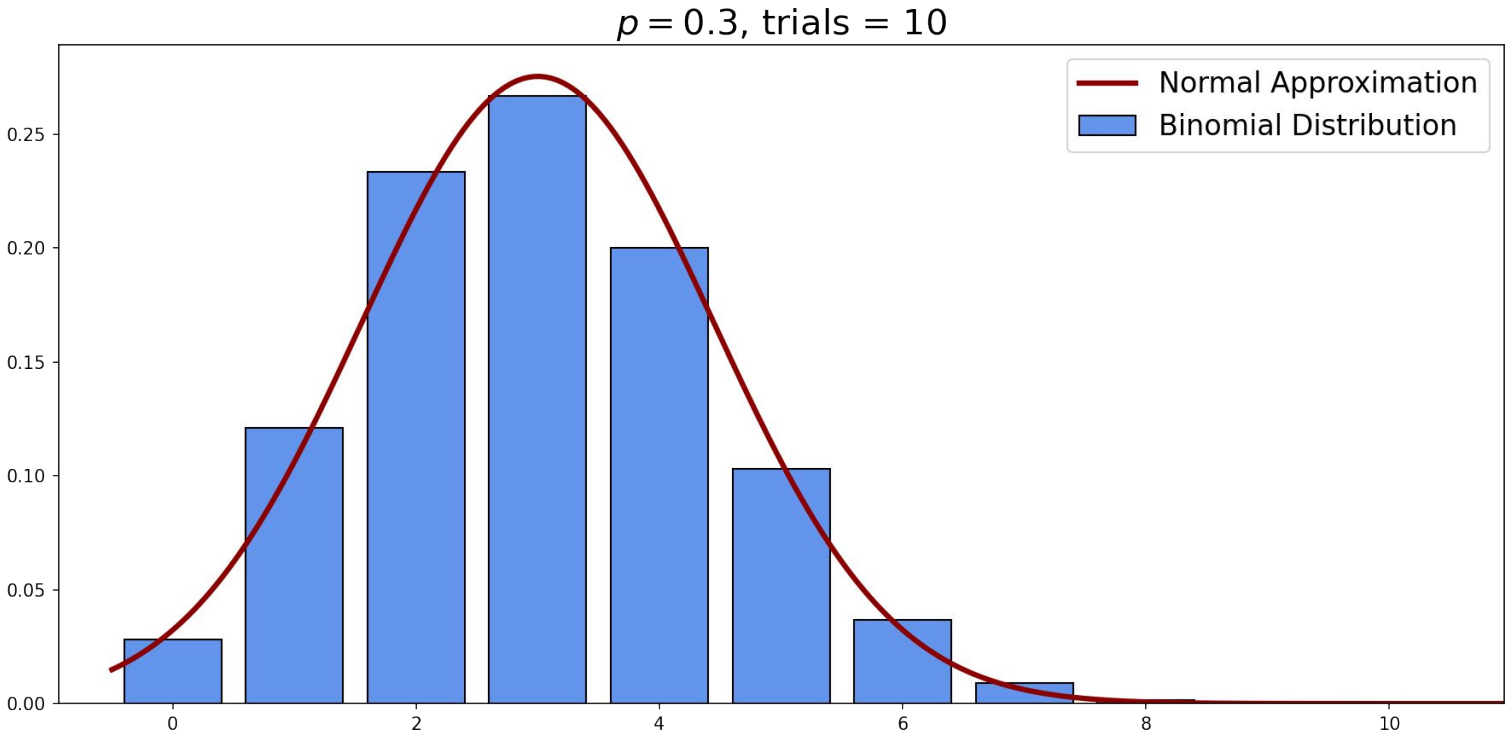
We also know that this distribution has mean $np$ and standard deviation $\sqrt{n \cdot p \cdot (1 - p)}$

NASHVILLE SOFTWARE SCHOOL

**Fun Fact:** As *n* increases, the normal distribution approximates the binomial distribution.

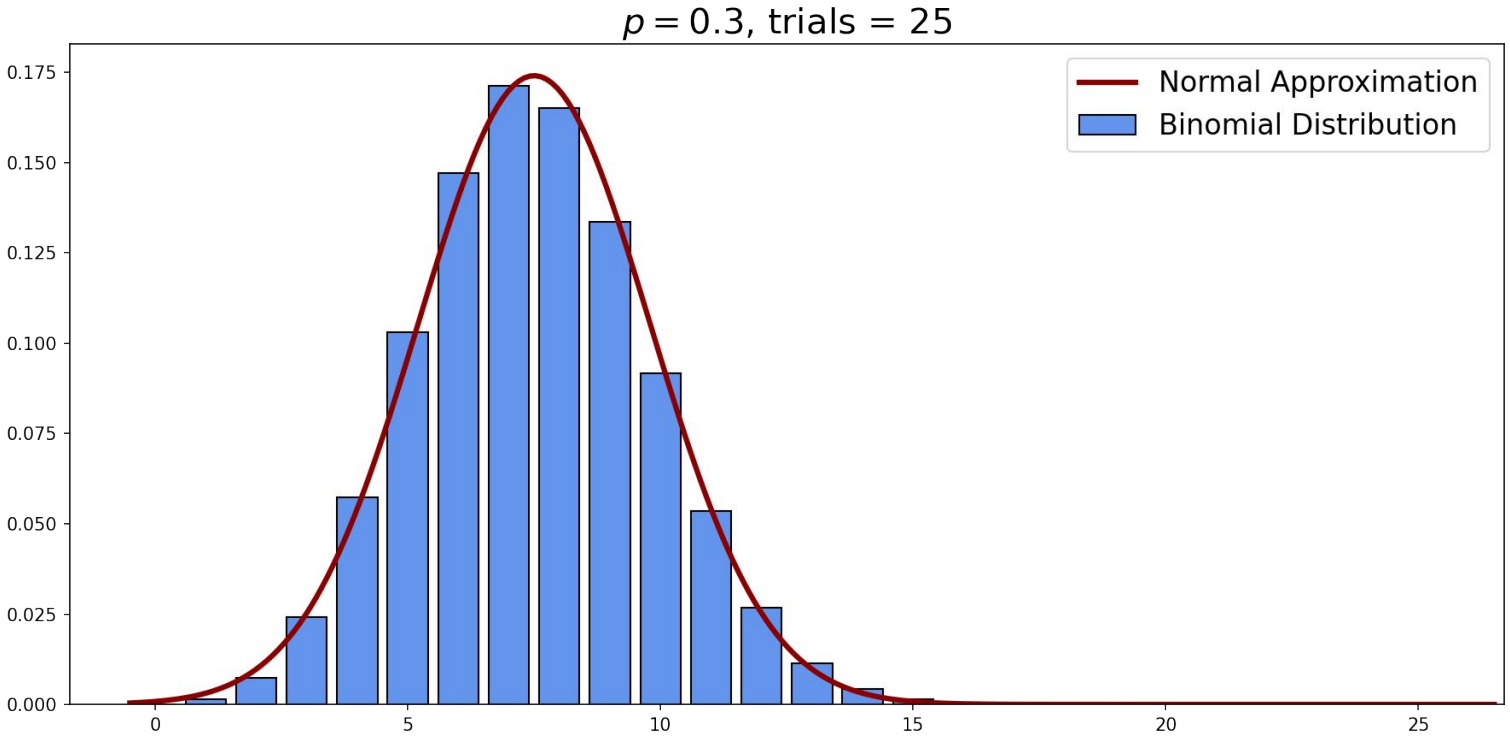
NASHVILLE SOFTWARE SCHOOL

**Fun Fact:** As *n* increases, the normal distribution approximates the binomial distribution.



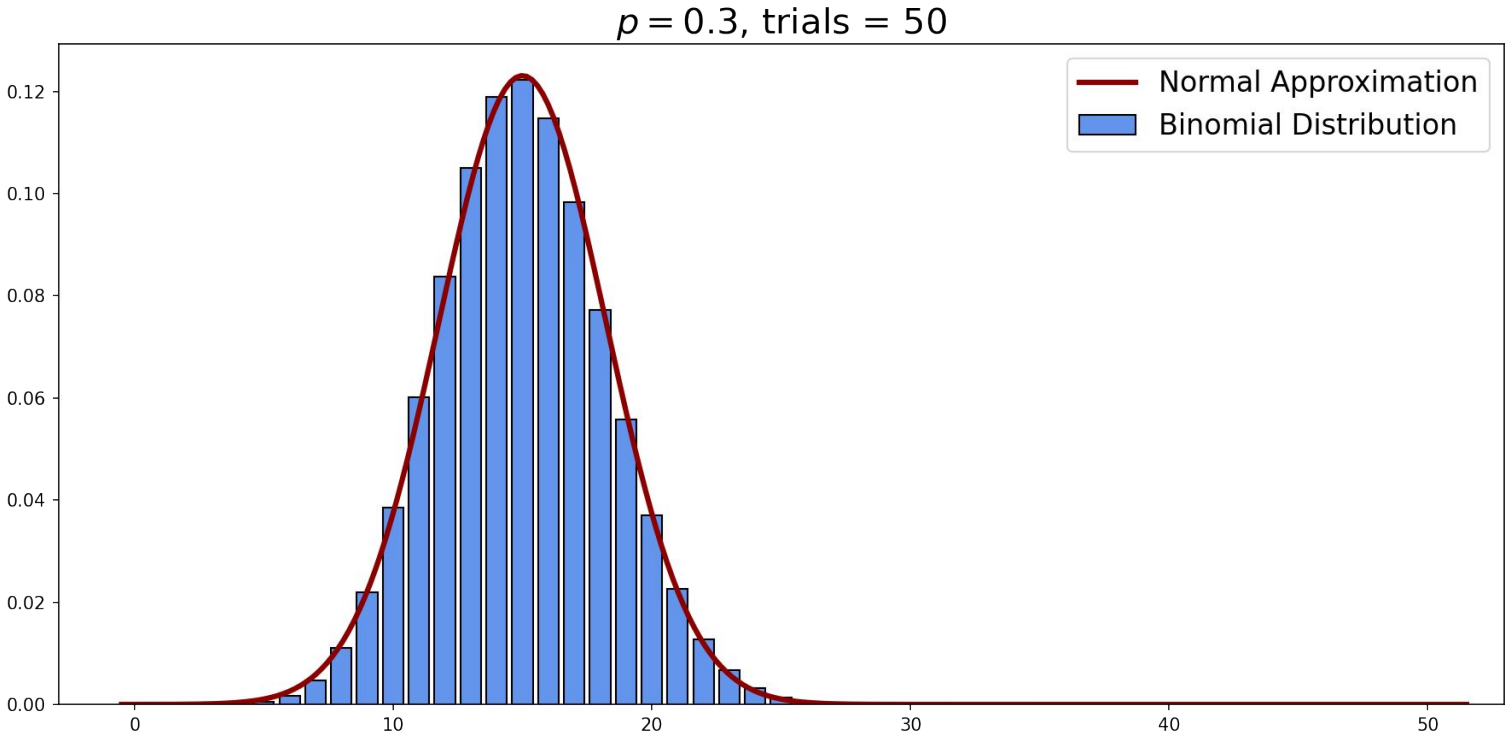

NASHVILLE SOFTWARE SCHOOL

**Fun Fact:** As *n* increases, the normal distribution approximates the binomial distribution.




NASHVILLE SOFTWARE SCHOOL

**Fun Fact:** As *n* increases, the normal distribution approximates the binomial distribution.




NASHVILLE SOFTWARE SCHOOL

# Estimation - Proportions

If *np ≥ 10* and *n(1-p) ≥ 10*, the distribution of $n \cdot \hat{p}$ is approximately normal with mean *np* and standard deviation $\sqrt{n \cdot p \cdot (1-p)}$

Divide by *n* and we can see that the distribution of $\hat{p}$ is approximately normal with mean *p* and standard deviation $\sqrt{\frac{p(1-p)}{n}}$

NASHVILLE SOFTWARE SCHOOL

# Estimation - Proportions

Recall that we can build a confidence interval if we know about the variability of the distribution of test statistics.

And now we know that for $\hat{p}$ , this has a standard deviation of $\sqrt{\frac{p(1-p)}{n}}$

But, there is a problem - it relies on using the population proportion - the quantity we are trying to approximate.

NASHVILLE SOFTWARE SCHOOL

# Estimation - Proportions

Just sub in the sample proportion!

$$\sqrt{\frac{p(1-p)}{n}} \quad \Rightarrow \quad \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

NASHVILLE SOFTWARE SCHOOL

# Estimation - Proportions

**Confidence Interval for the Proportion:**

$$\hat{p} \pm z_\alpha \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

NASHVILLE SOFTWARE SCHOOL

# Estimation - Proportions

**Confidence Interval for the Proportion:**

$$\hat{p} \pm z_\alpha \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Point Estimate

NASHVILLE SOFTWARE SCHOOL

# Estimation - Proportions

**Confidence Interval for the Proportion:**

$$\hat{p} \pm z_{\alpha} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Point Estimate

Critical Value (from the standard normal distribution)

NASHVILLE SOFTWARE SCHOOL