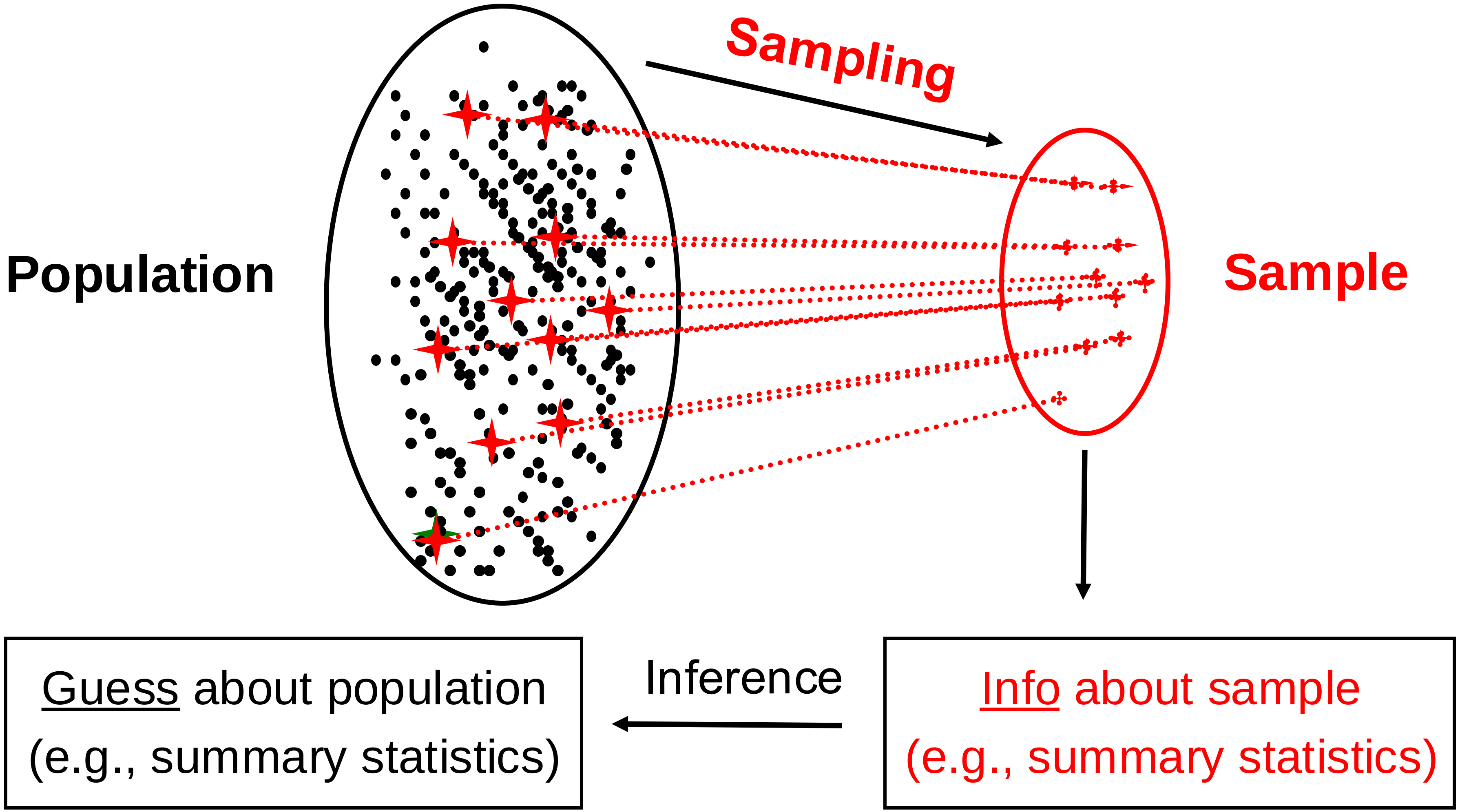
Sampling and CLT

Part 3.1 Populations, Samples, Sample Mean,

Central Limit Theorem, Normal Distribution

But how do we learn about the population?

Q. What can we say about the heights of all US residents?



Sampling

Q. What can we say about the population from a sample?

Population **Parameter**	**Guess Based on** **a Sample**
Population Mean μ	**Sample** Mean $\bar{X}$
Population Variance σ^2	**Sample** Variance S^2
Population Std. Dev. σ	**Sample** Std. Dev. S
Population Proportion p	**Sample** Proportion $\hat{p}$
Population Regression Coefficients	**Sample Regression Coefficients**
Many other **population** parameters…	Many other **sample** statistics…

Sampling

Q. Is the sample mean a good guess for the population mean?

Population Parameter		**Guess Based on a Sample**	
Population Mean	μ	**Sample** Mean	$\bar{X}$

Sampling

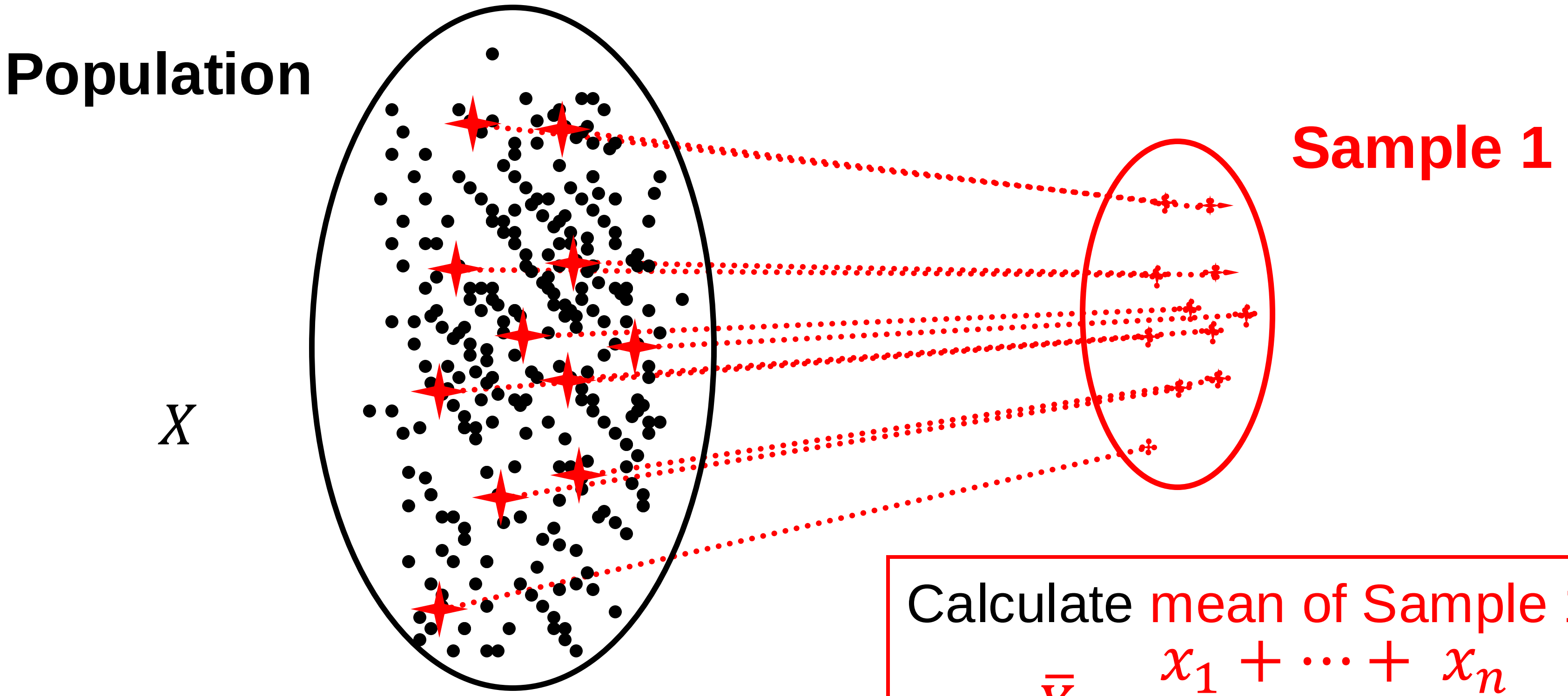
Q. Is the sample mean a good guess for the population mean?

Yes, but we need some theory….

Population Parameter		**Guess Based on a Sample**	
Population Mean	μ	**Sample** Mean	$\bar{X}$

Sampling

How to guess the population mean: μ_X

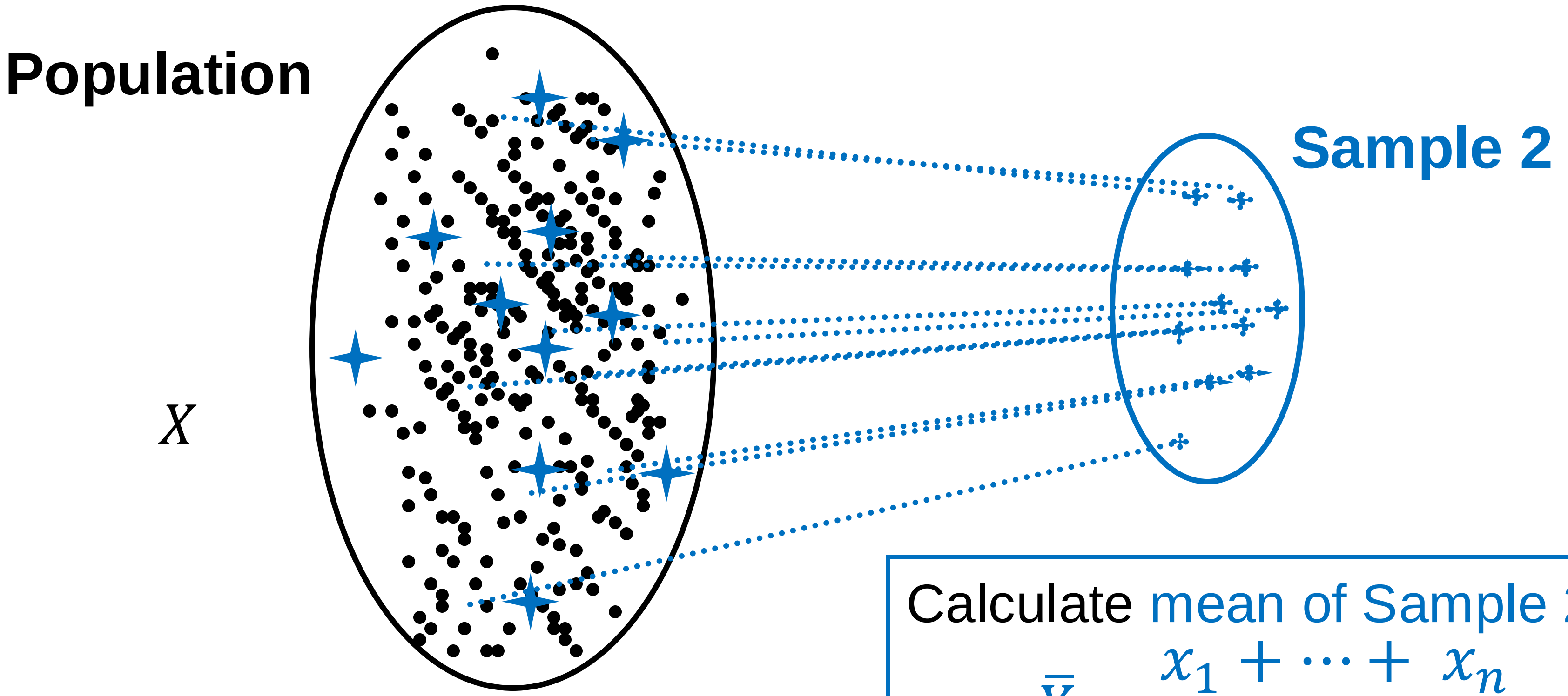


Calculate mean of Sample 1

$$\bar{X} = \frac{x_1 + \cdots + x_n}{n}$$

Sampling

How to guess the population mean: μ_X

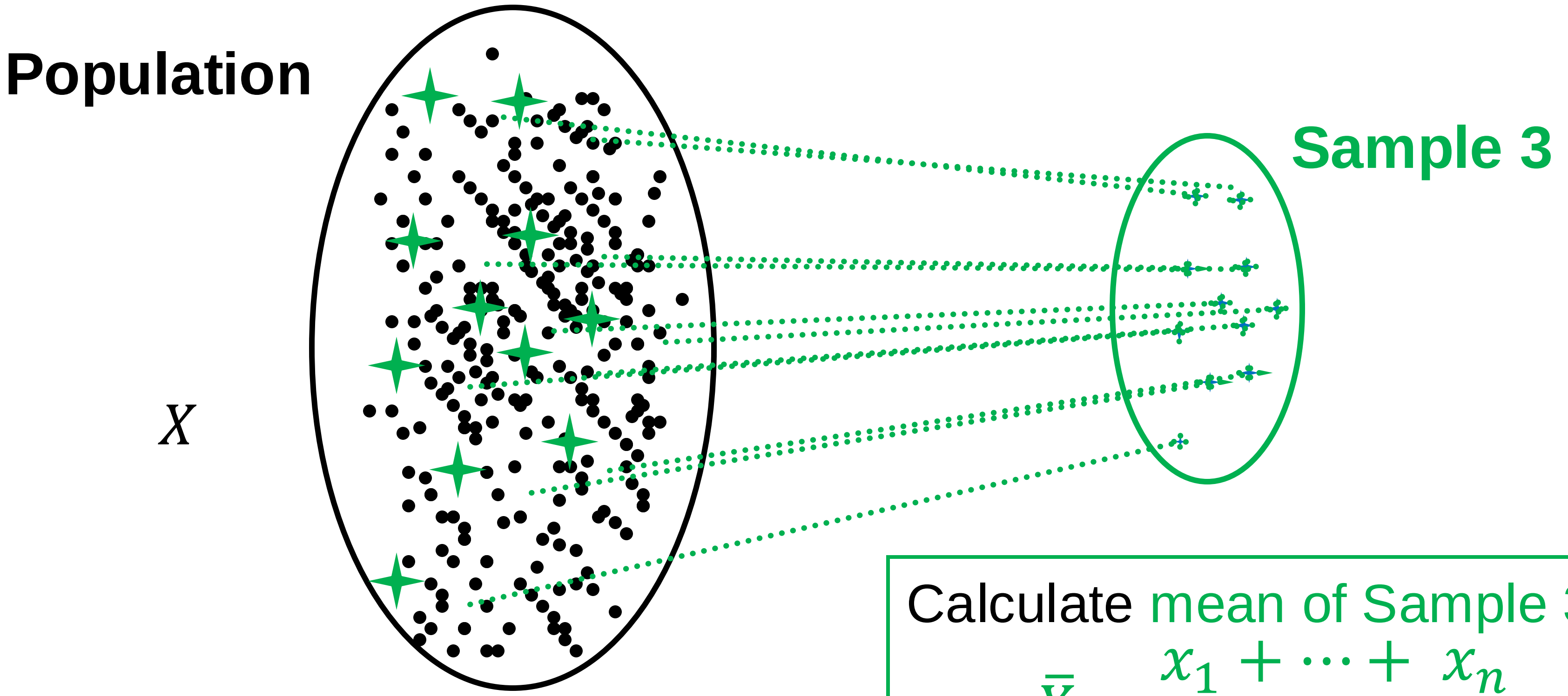


Calculate mean of Sample 2

$$\bar{X} = \frac{x_1 + \cdots + x_n}{n}$$

Sampling

How to guess the population mean: μ_X

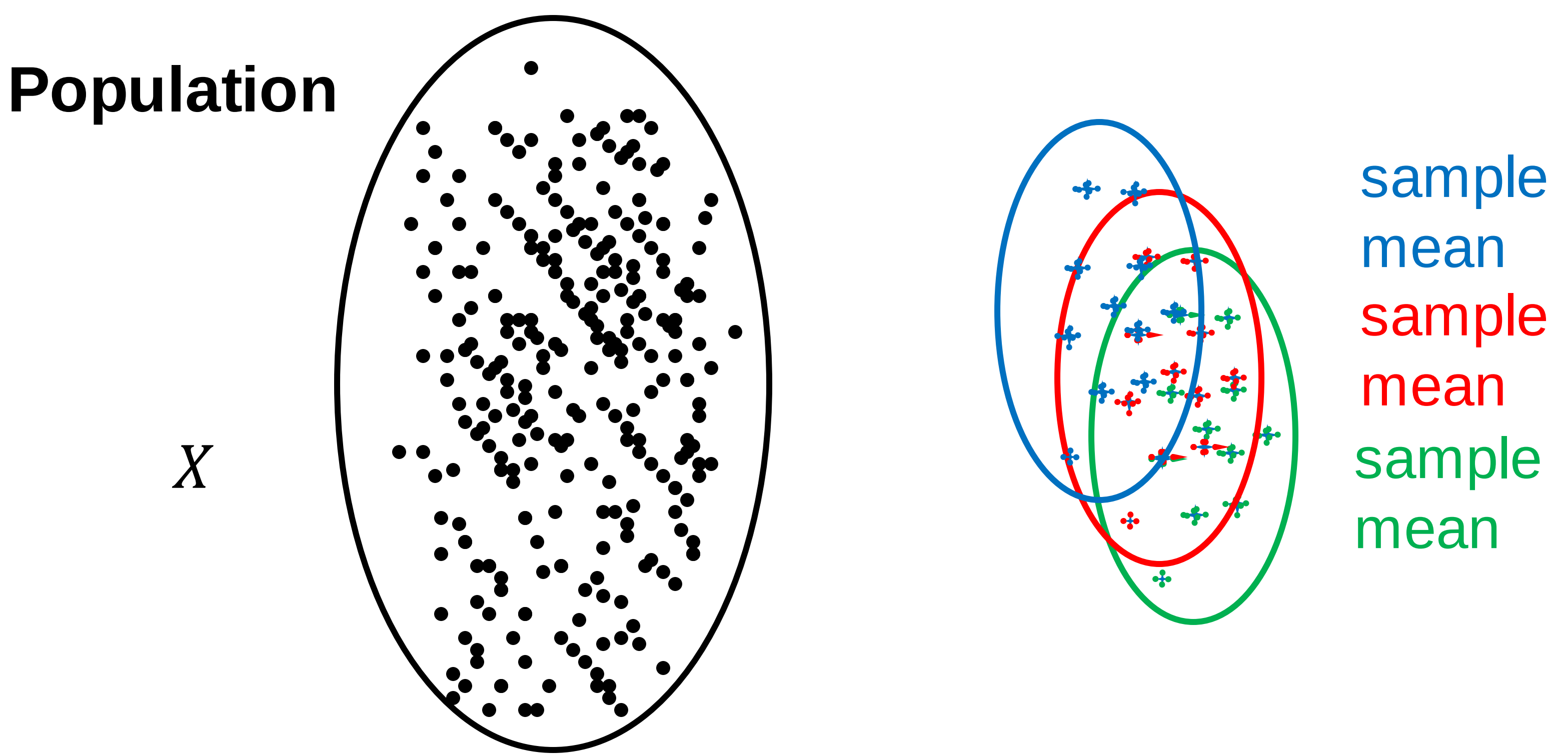


Calculate mean of Sample 3

$$\bar{X} = \frac{x_1 + \cdots + x_n}{n}$$

Sampling

How to guess the population mean: μ_X



Sampling: Fluctuations in $\bar{X}$

How to guess the population mean: μ_X

Since samples are selected at random, the sample mean $\bar{X}$ fluctuates from sample to sample.

- Sample mean $\bar{X}$ is a random variable
- Like any random variable, the sample mean has an expected value and a standard deviation

Why does it matter that $\bar{X}$ is a random variable?

- The distribution of the sample mean can help us understand if $\bar{X}$ a good guess for the population mean.

- How do we figure out the population mean?
  - We need to know the distribution of the sample mean

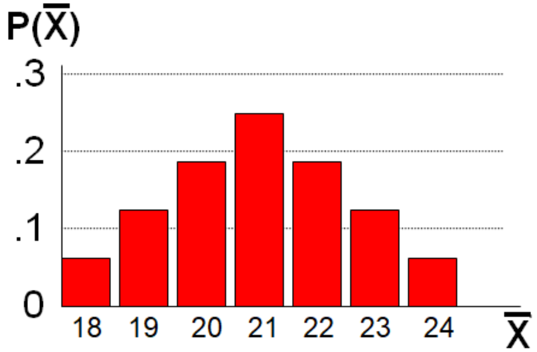
Sampling: Distribution of $\bar{X}$

What is the distribution of $\bar{X}$ if X is normally distributed?

If the population distribution is normal,

$$X \sim N(\mu, \sigma^2)$$

then $\bar{X}$ also follows normal distribution



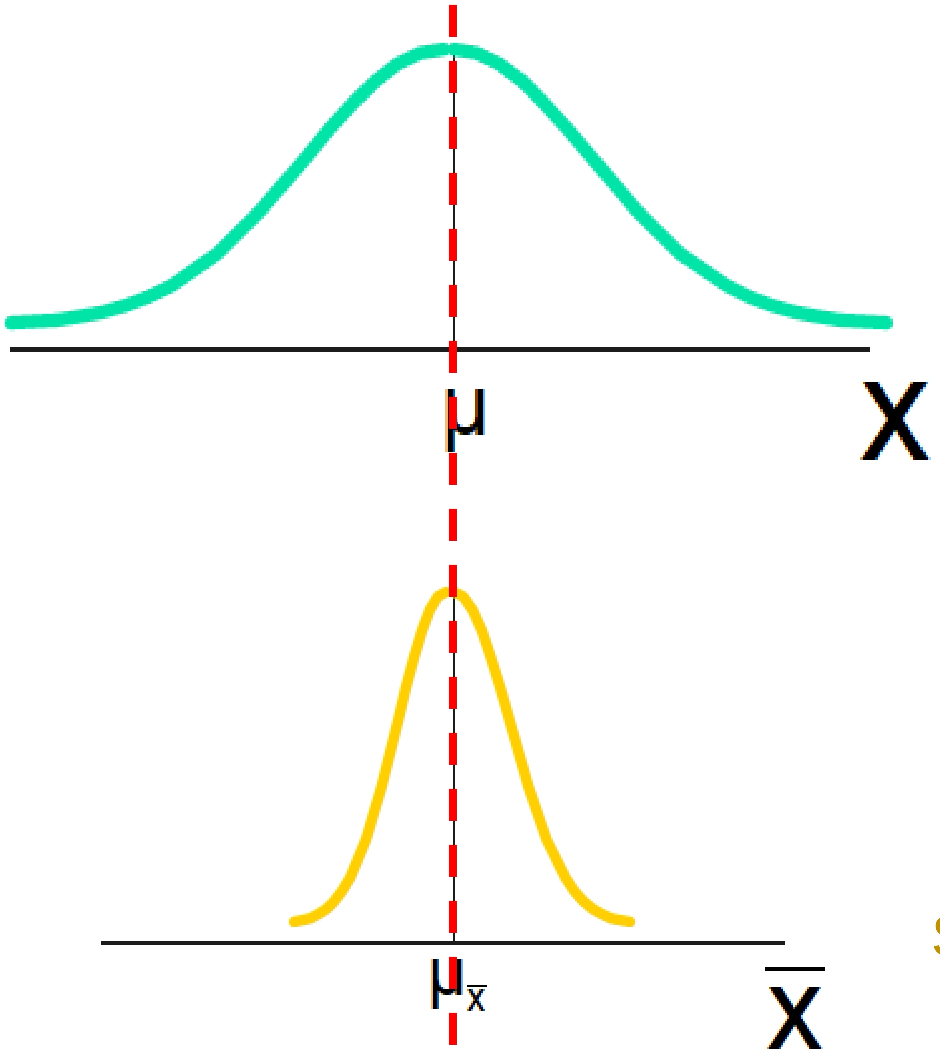
$$\mu_{\bar{X}} = \mu$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

$$\bar{X} \sim N(\mu_{\bar{X}}, \sigma_{\bar{X}}^2) = N\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right) = N\left(\mu, \frac{\sigma^2}{n}\right)$$

Sampling: Distribution of $\bar{X}$

What is the distribution of $\bar{X}$ if X is normally distributed?



Normal
Population Distribution

1. $\mu_{\bar{X}} = \mu$

   $\bar{X}$ is unbiased

2. $\sigma_{\bar{X}} < \sigma$

   less spread out than X

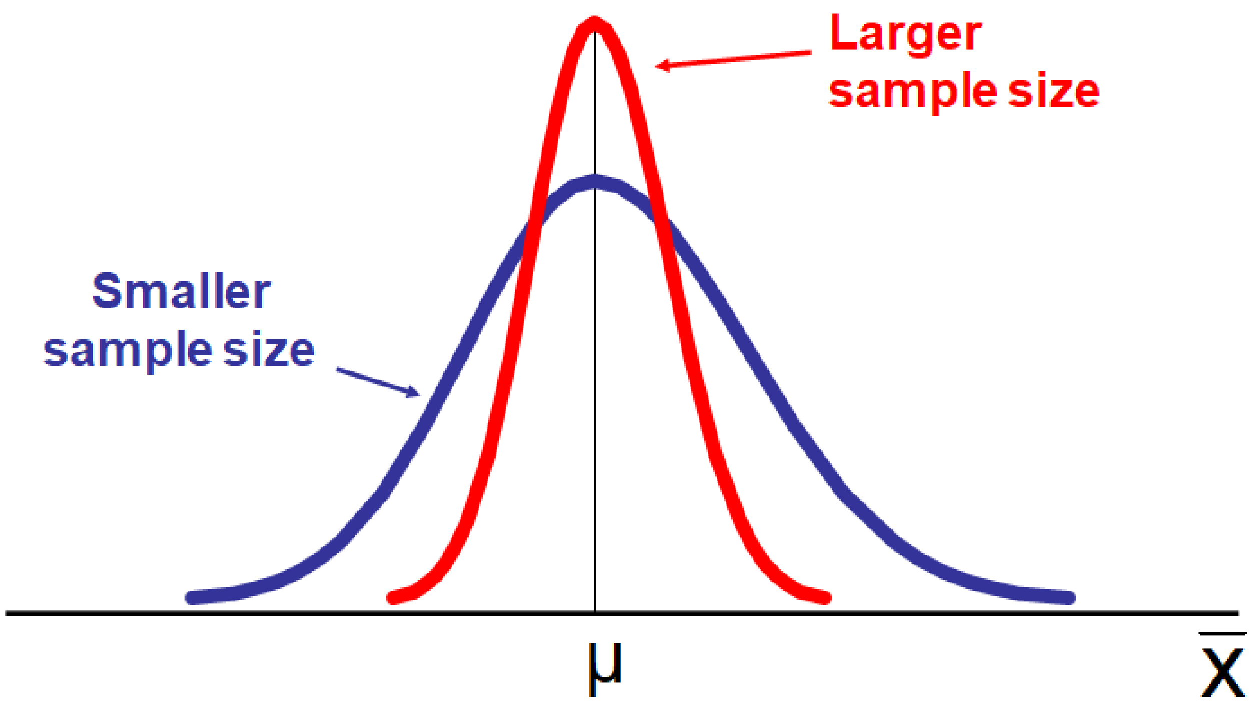
Normal
Sample Mean Distribution

Sampling: Distribution of $\bar{X}$

What is the distribution of $\bar{X}$ if X is normally distributed?

As *n* increases:

- $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ decreases
- $\bar{X}$ becomes a more accurate estimation of μ



Sampling: Distribution of $\bar{X}$

Q. What if X is ***not*** *normally distributed?*

The Central Limit Theorem tells us that the sample means is approximately normal if the sample size is large enough.

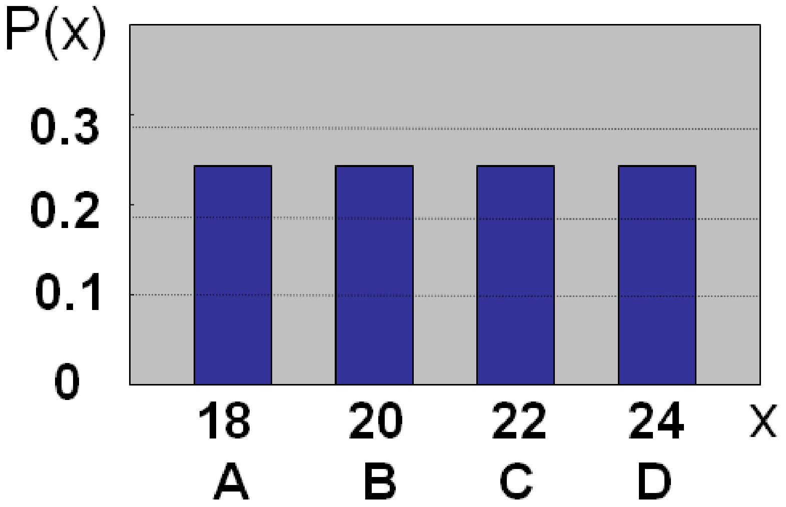
How large is large enough?

- For most population distributions, $n > 30$
- If population distribution is fairly symmetric, $n > 5$
- If population distribution is normal, any n would work

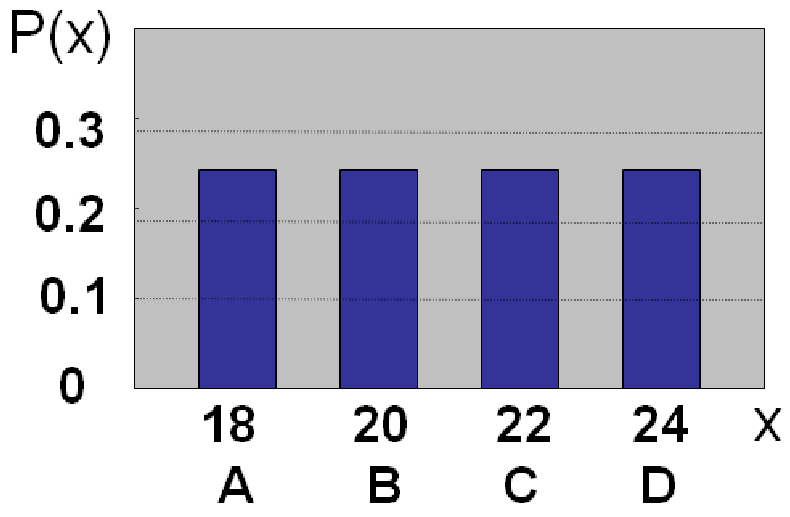
Distribution of $\bar{X}$

Assume there is a population:

- Population size: $N = 4$
- Random variable: X = age
- X = 18, 20, 22, 24



Distribution of $\bar{X}$



Summary measures of X:

- Mean:

$$\mu = \frac{18 + 20 + 22 + 24}{4} = 21$$

- Standard deviation:

$$\sigma^2 = \frac{(18 - 21)^2 + (20 - 21)^2 + (22 - 21)^2 + (24 - 21)^2}{4} = 5$$

$$\sigma = \sqrt{\sigma^2} = 2.24$$

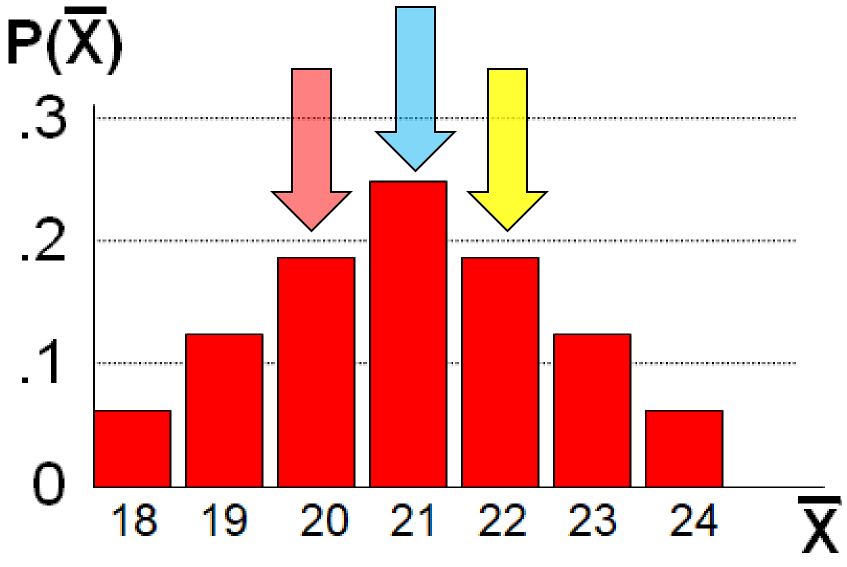
Distribution of $\bar{X}$

Consider all possible samples of size n = 2.

$$\bar{X} = (x_1 + x_2)/2$$

1st Obs	2nd Observation			
	18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

16 sample means
→ 16 values of $\bar{X}$



Distribution of sample mean $\bar{X}$

Distribution of X vs. Distribution of $\bar{X}$

Population distribution:

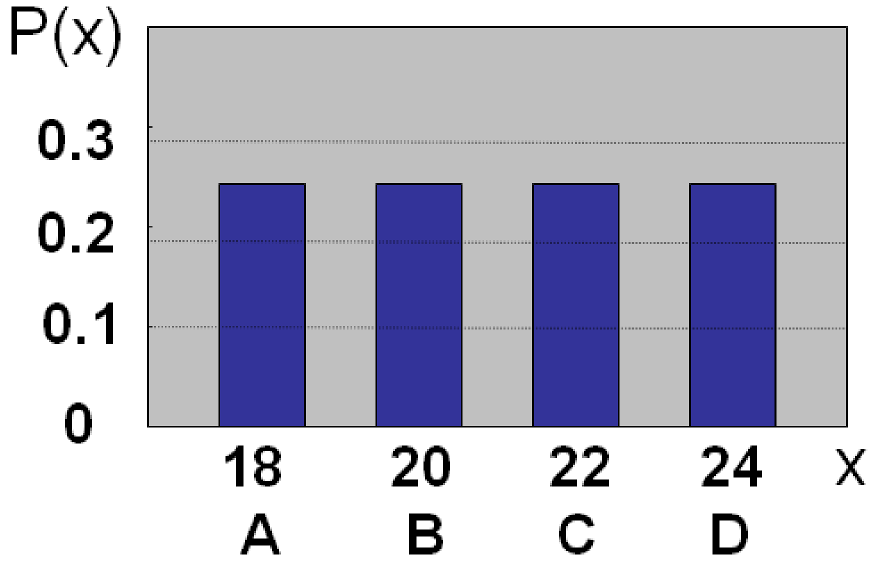
- Distribution of X
- $N = 4$
- Mean $\mu = 21$
- Std. dev. $\sigma = 2.24$

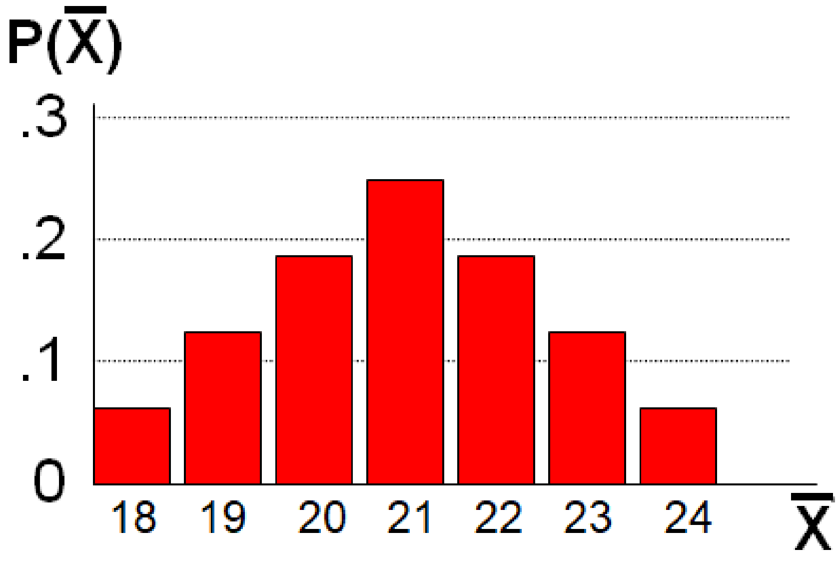
$=$

$>$

Sample mean distribution:

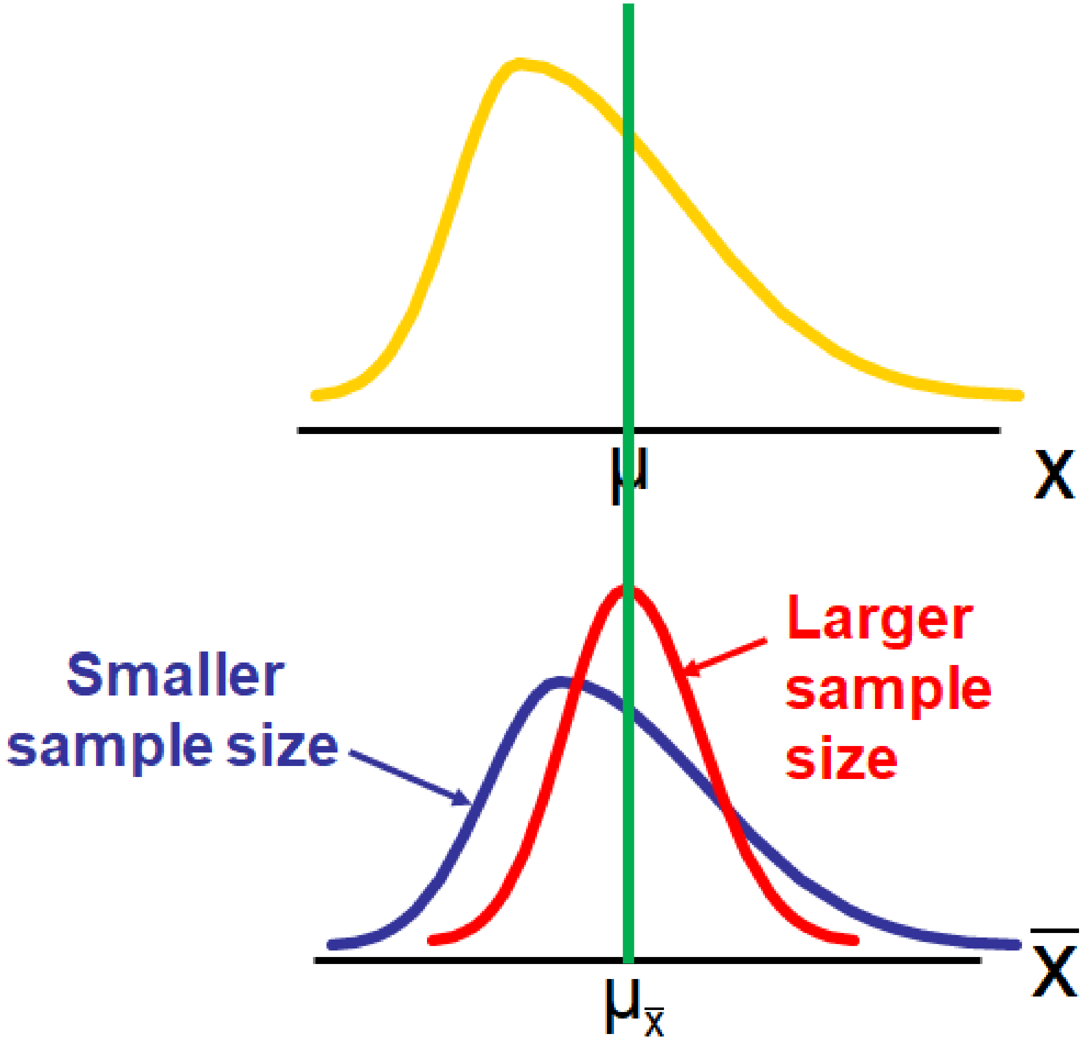
- Distribution of $\bar{X}$
- $n = 16$
- Mean $\mu = 21$
- Std. dev. $\sigma = 1.58$





Distribution of $\bar{X}$

Q. What if X is **not** *normally distributed?*



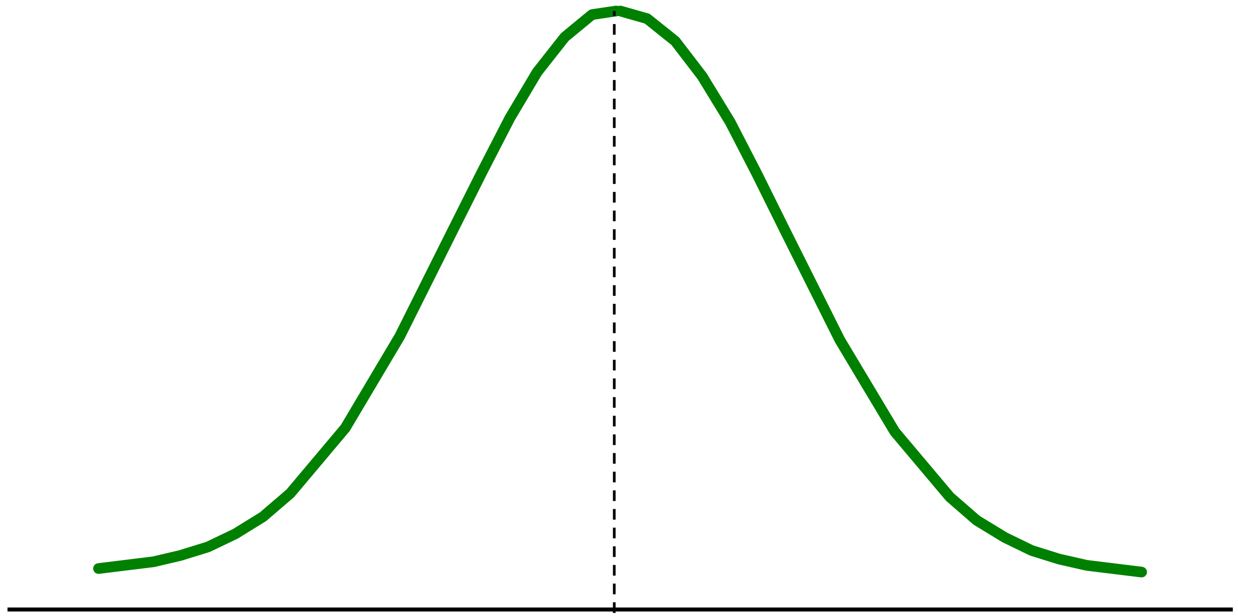
Population Distribution

Sample Mean Distribution
(becomes normal
as *n* goes up)

Normal Distribution

… the single most important distribution in Statistics.

I’s also called “Gaussian” or “bell curve”.



Normal Distribution

Main motivation:

- The average of a large sample of measurements drawn from a population is normally distributed (Central Limit Theorem)

Examples:

- Number/size of objects produced by machines
- Errors in measurements

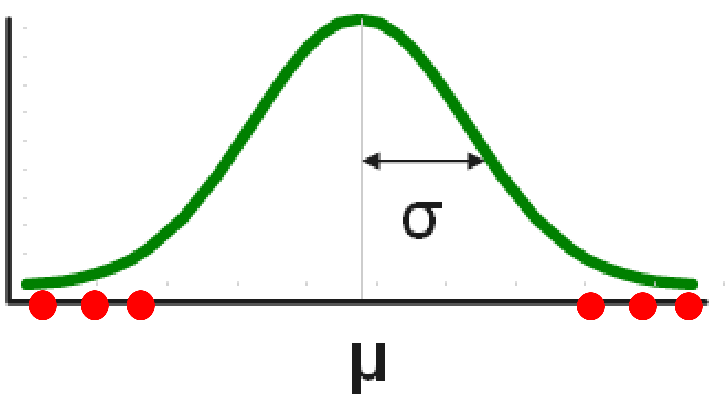
Mathematical models that use normally distributed random variables:

- Regression models
- Some theories in financial economics

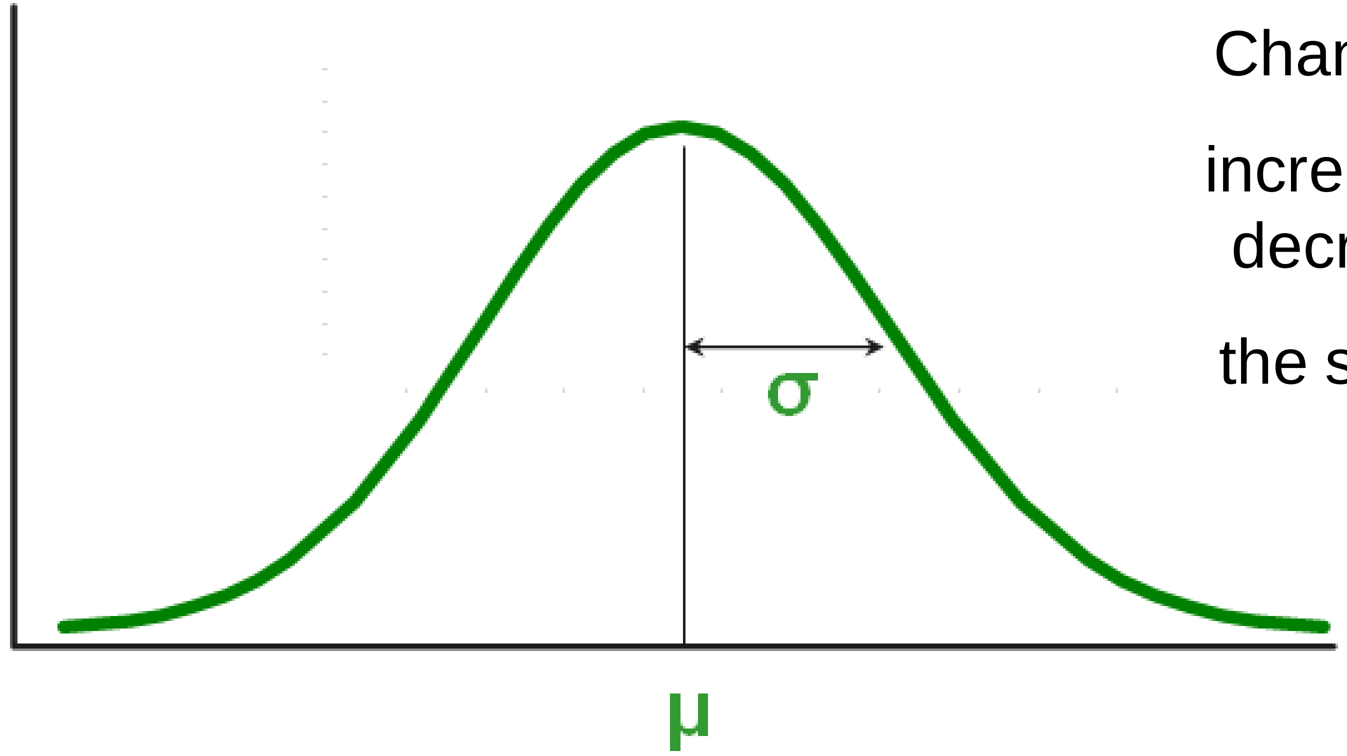
Normal Distribution: $N(\mu, \sigma^2)$

… is a continuous random variable with a distribution:

- Bell-shaped
- Symmetrical
- Mean = Median = Mode
- Unbounded
- Location is determined by the mean μ
- Spread is determined by the standard deviation σ
- Formula for probability density function is very complicated



Normal Distribution $N(\mu, \sigma^2)$: Shape



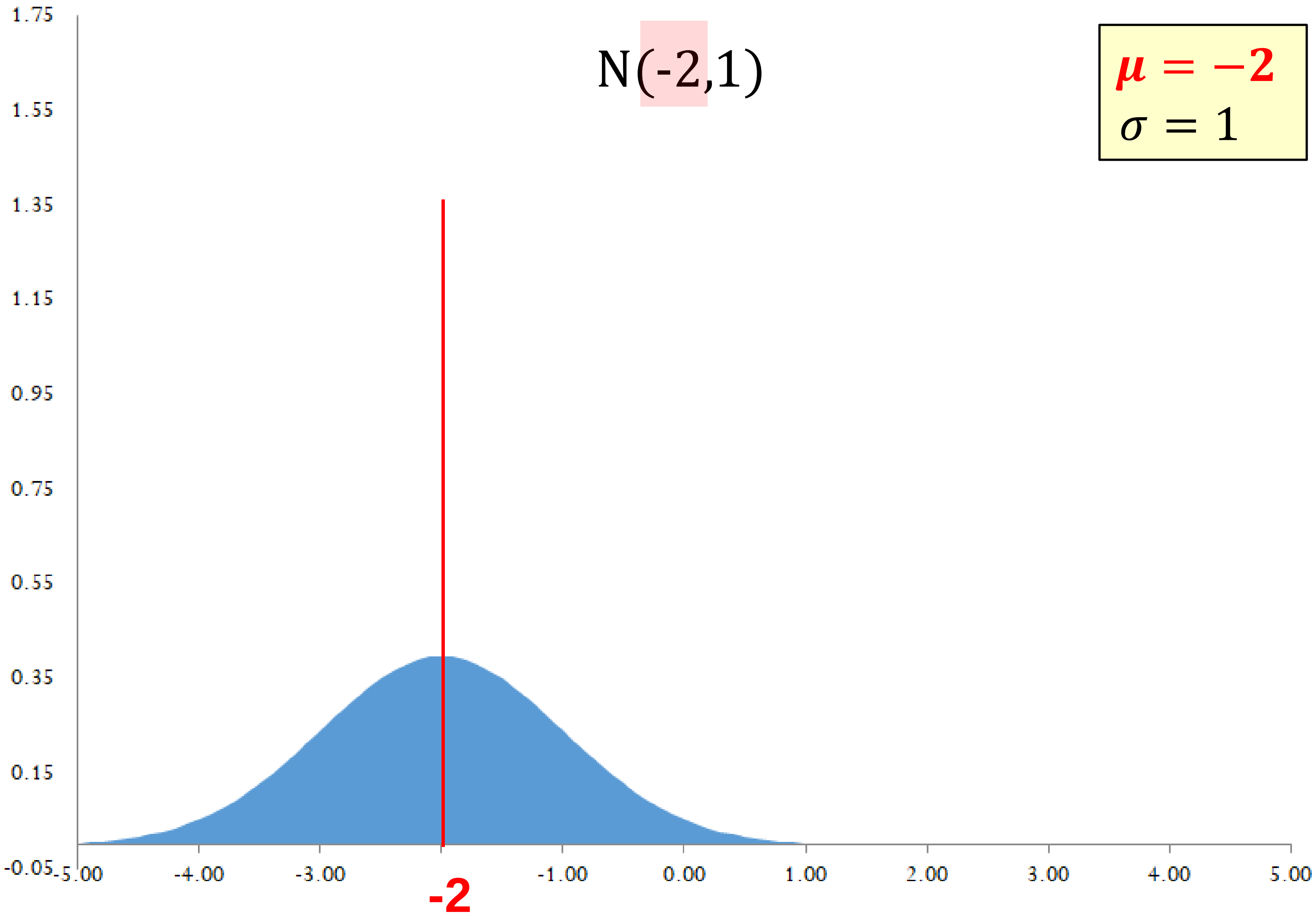
Changing σ

increases or decreases

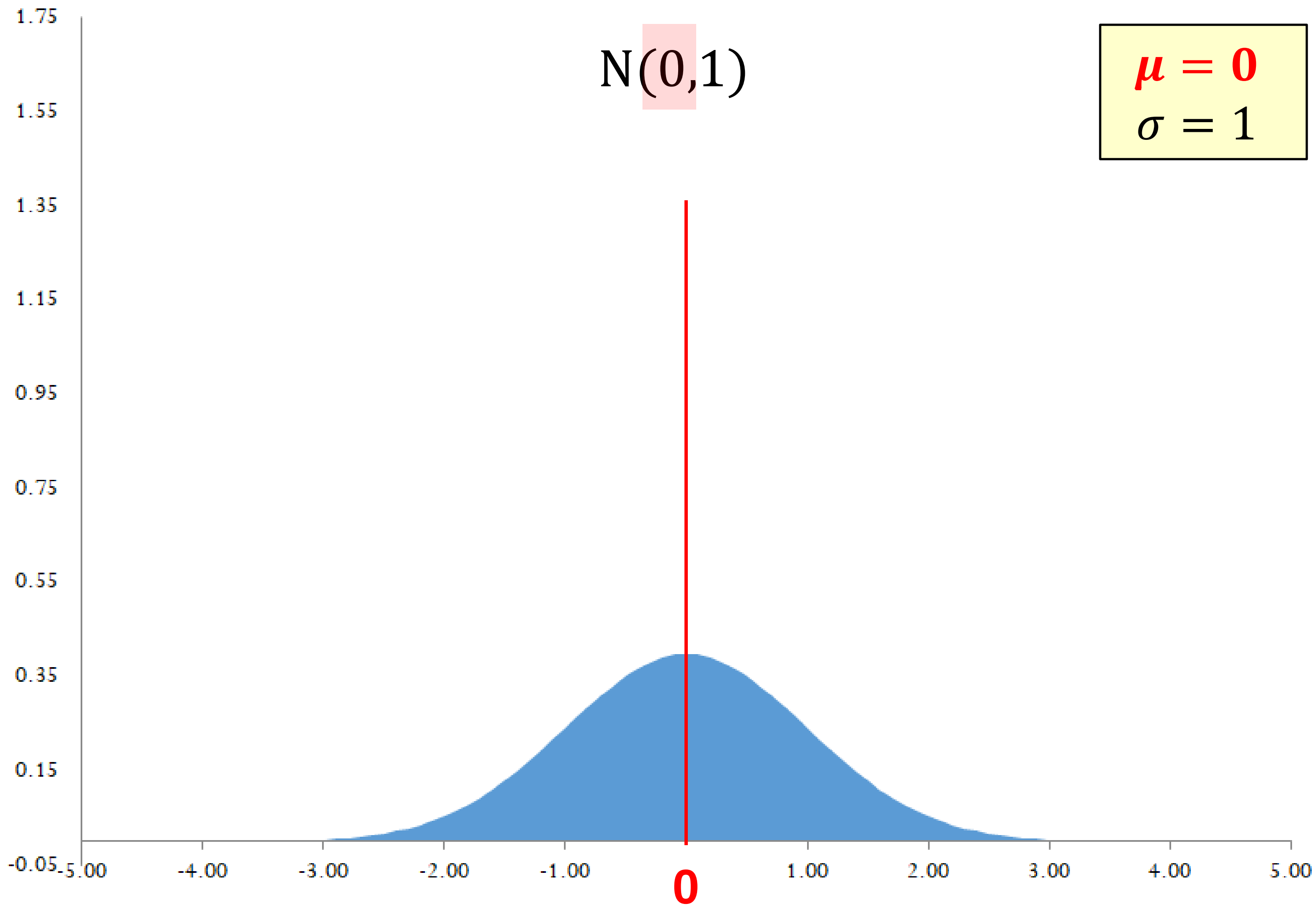
the spread.

Changing μ shifts the distribution left or right.

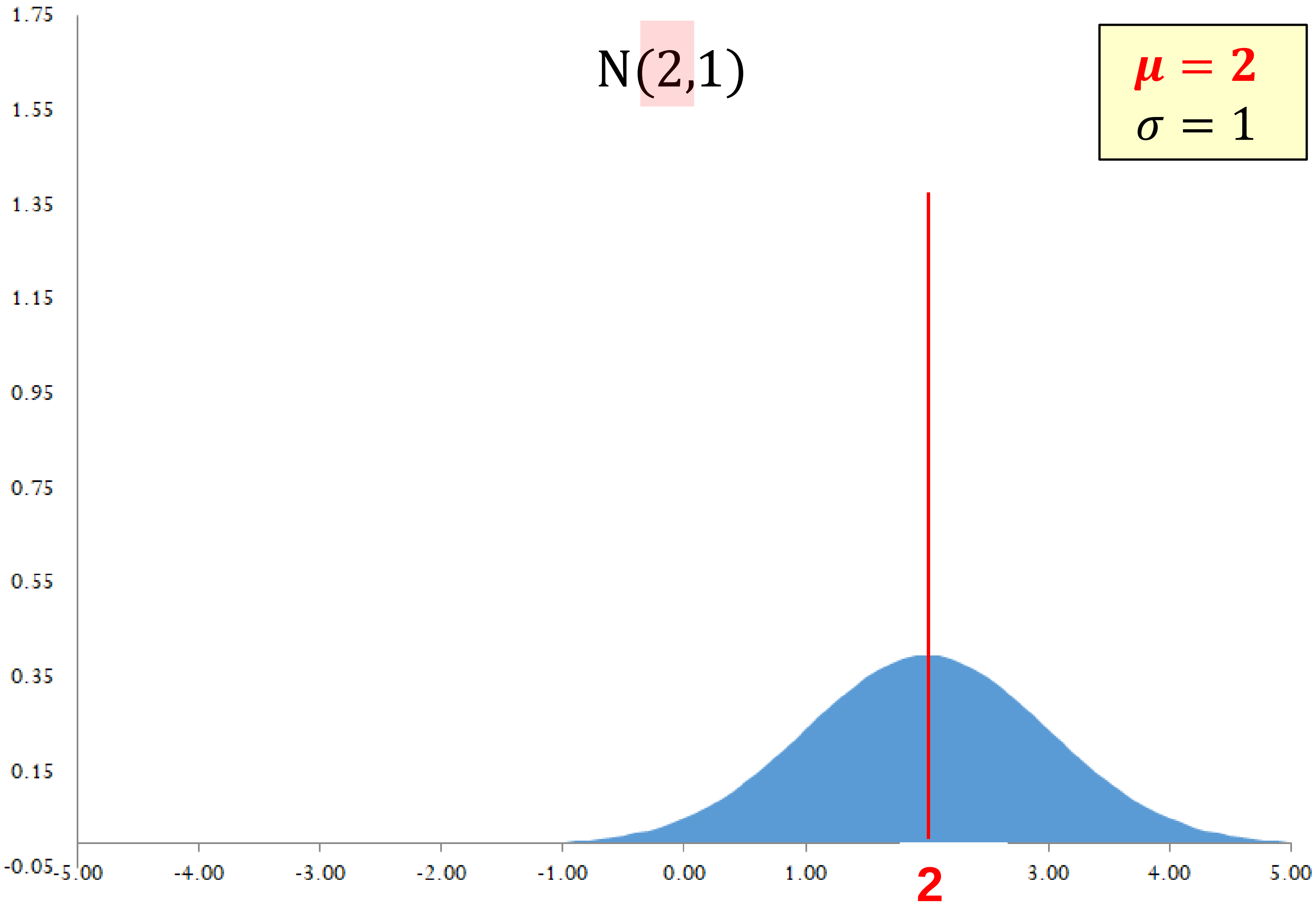
Normal Distribution $N(\mu, \sigma^2)$: Shape



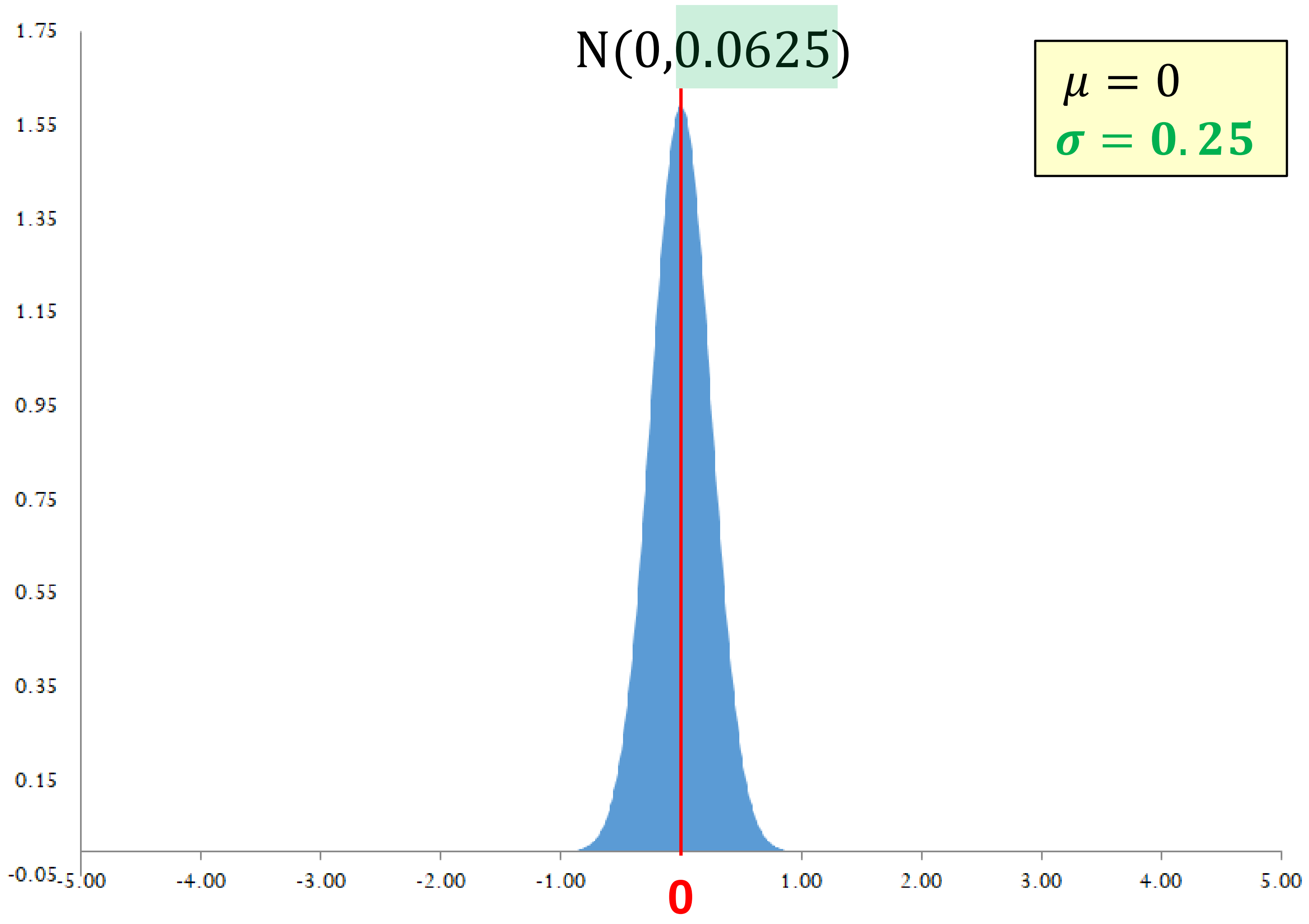
Normal Distribution $N(\mu, \sigma^2)$: Shape



Normal Distribution $N(\mu, \sigma^2)$: Shape



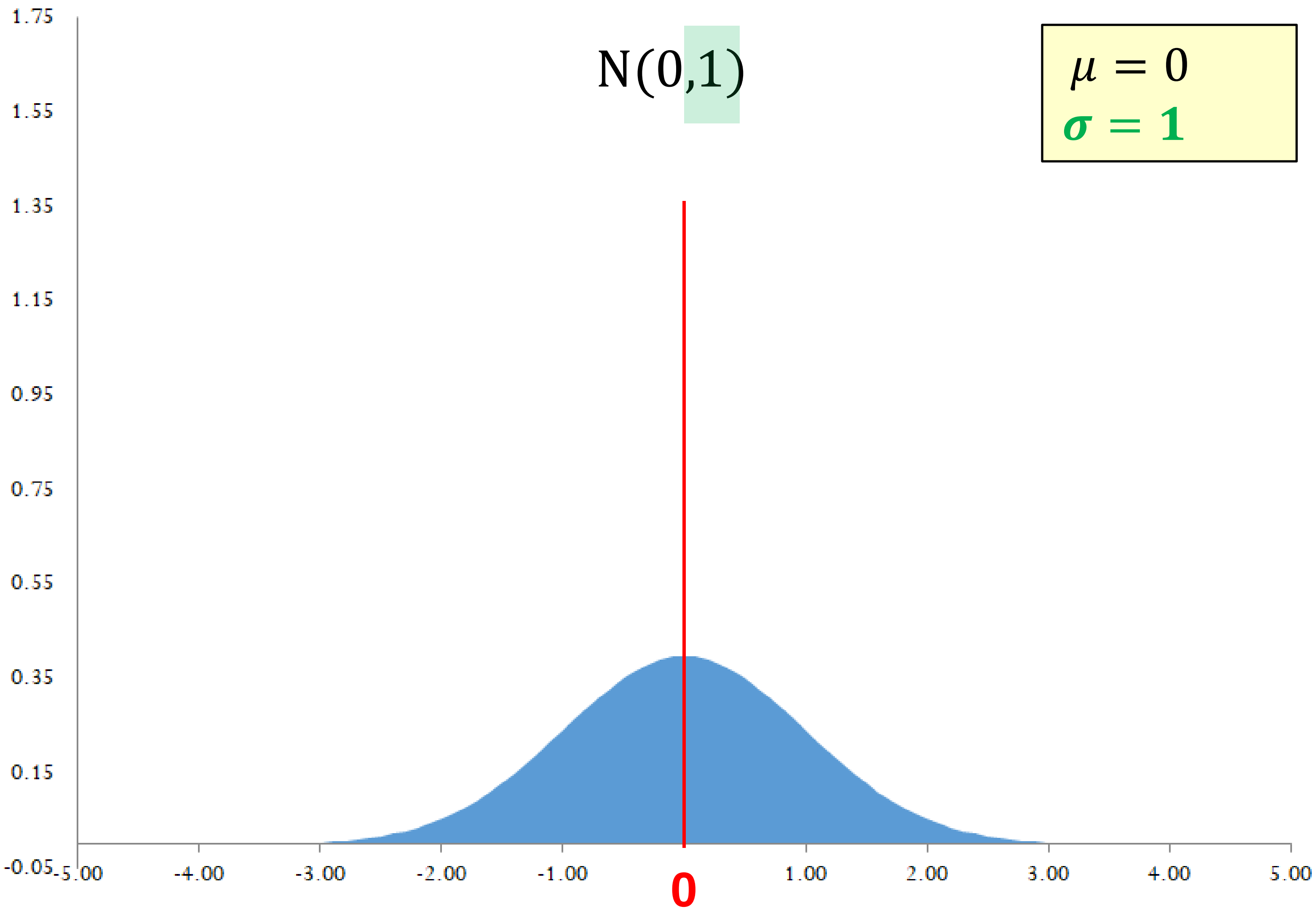
Normal Distribution $N(\mu, \sigma^2)$: Shape



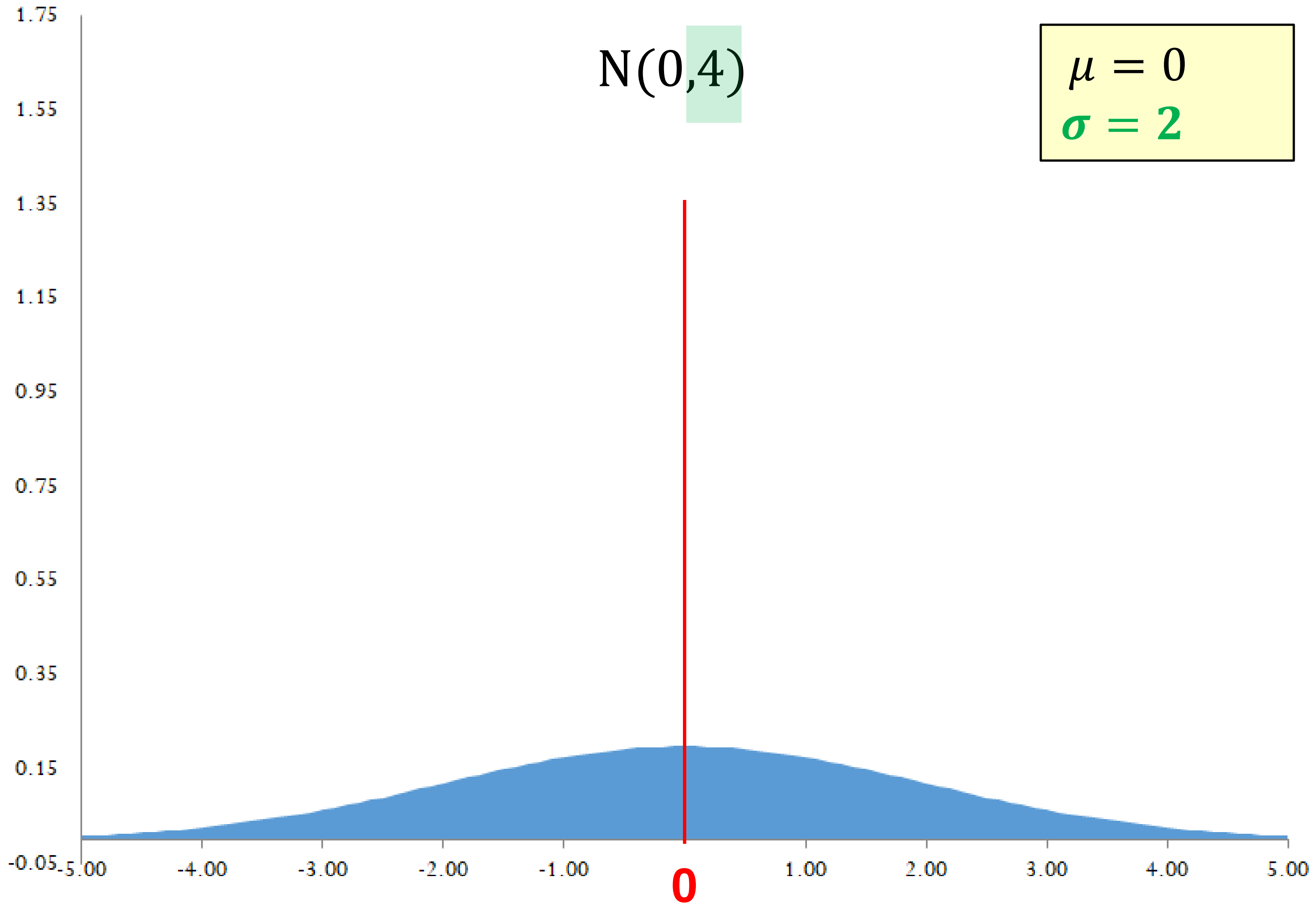
$\mu = 0$

$\boldsymbol{\sigma = 0.25}$

Normal Distribution $N(\mu, \sigma^2)$: Shape



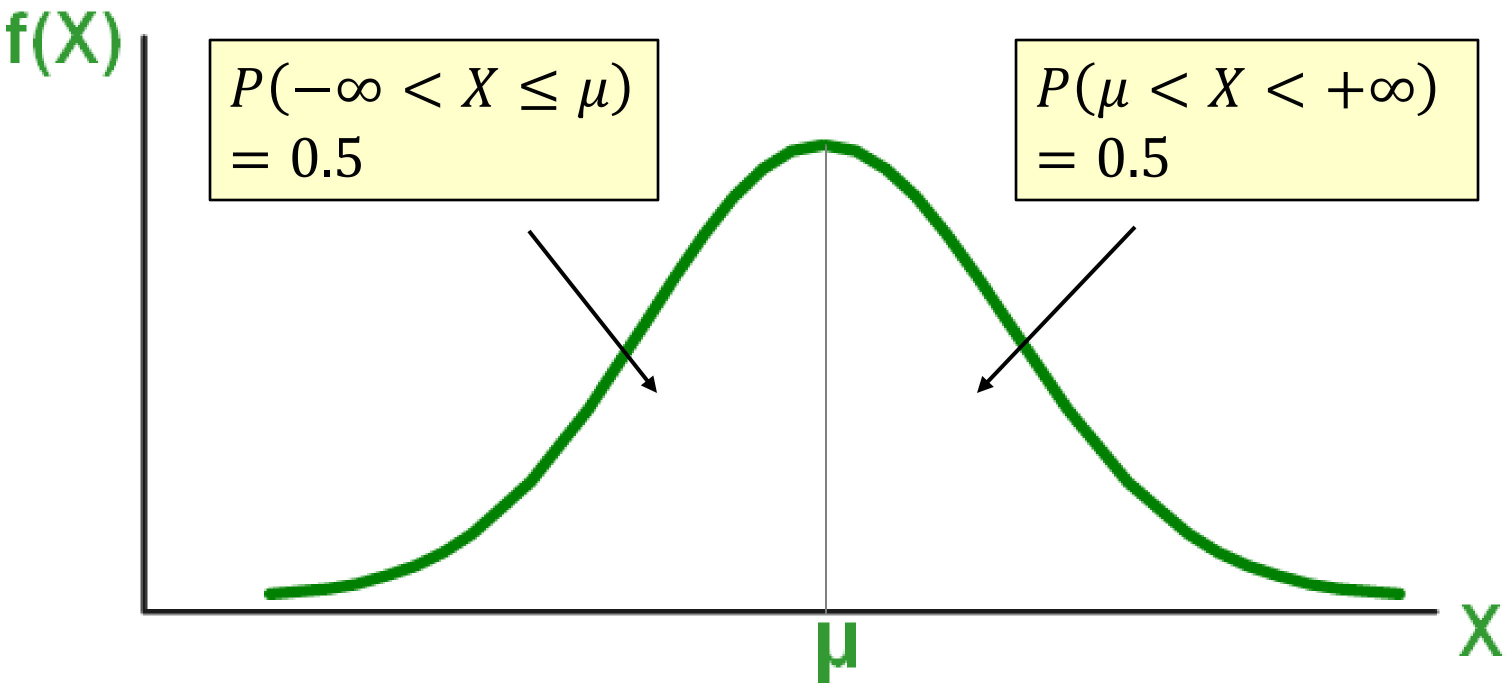
Normal Distribution $N(\mu, \sigma^2)$: Shape



Probability as Area Under the Curve

Recall:

- Area under the curve is 1.0
- The curve is symmetric



$$P(-\infty < X \le \mu) = 0.5$$

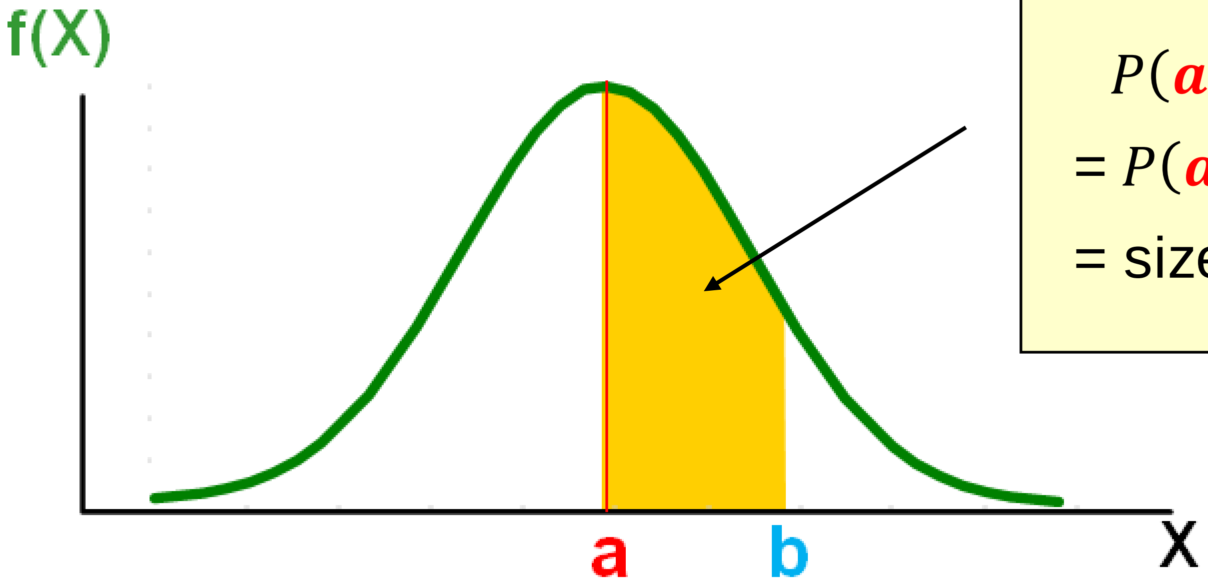
$$P(\mu < X < +\infty) = 0.5$$

$$P(-\infty < X < +\infty) = 1$$

Probability as Area Under the Curve

Recall: if X is a continuous random variable, then

- $P(a < X < b) =$ area between a and b under the curve defined by the probability density function.



$P(\boldsymbol{a} < X < \boldsymbol{b})$

$= P(\boldsymbol{a} \leq X \leq \boldsymbol{b})$

= size of this area

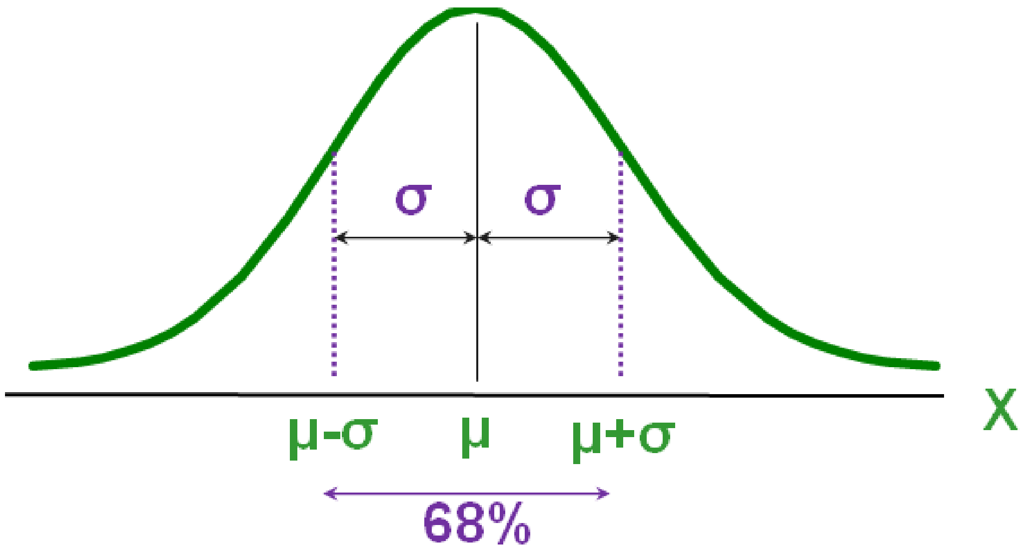
For any single value a : $P(X = \boldsymbol{a}) = \mathbf{0}$

1-2-3 Rule for Normal Distribution

A normal distribution is a symmetric curve centered at the mean; there is:

- 68% chance of being within one std. deviation of the mean

$$P(\mu - \sigma < X < \mu + \sigma) = 0.68$$



1-2-3 Rule for Normal Distribution

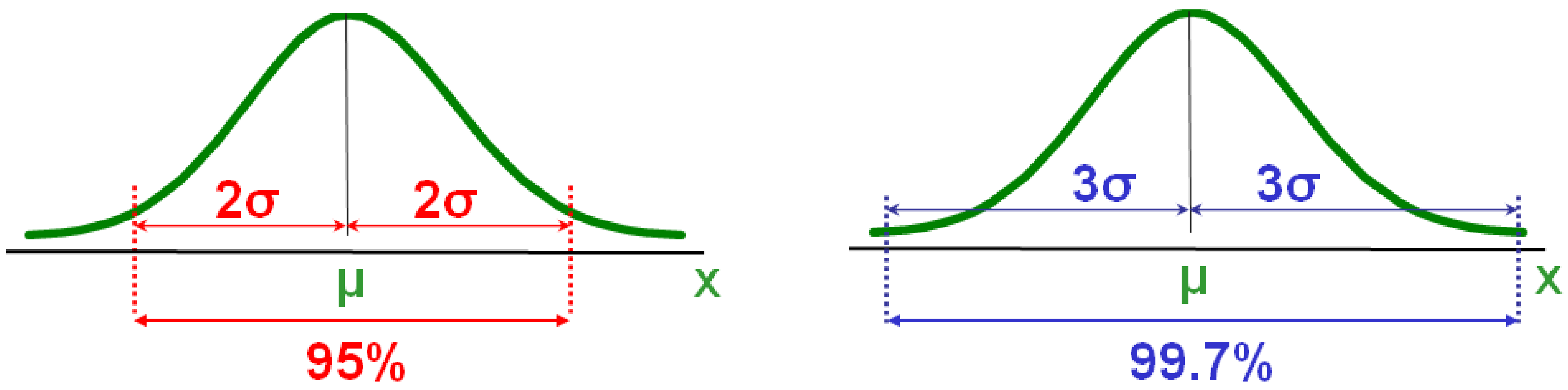
A normal distribution is a symmetric curve centered at the mean; there is:

- 95% chance of being within two std. deviation of the mean

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.95$$

- 99.7% chance of being within three std. deviation of the mean

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.997$$



Standardization: Convert X to Z

Any normally distributed variable $X \sim N(\mu, \sigma^2)$ can be related to

a *standardized* normal variable $Z \sim N(0,1)$ with mean = 0 and std = 1

If $X \sim N(\mu, \sigma^2)$ then $\boxed{Z = \frac{X-\mu}{\sigma}} \sim N(0,1)$

For any normal distribution: only need to care about

the distance from the value of X to the mean μ, i.e. $(X - \mu)$

using σ as the unit to measure that distance

$$\boxed{P(X < a) \;=\; P\left(\frac{X-\mu}{\sigma} < \frac{a-\mu}{\sigma}\right) = P\left(\mathbf{Z} < \frac{a-\mu}{\sigma}\right)}$$

$$\boxed{P(X > b) \;=\; P\left(\frac{X-\mu}{\sigma} > \frac{b-\mu}{\sigma}\right) = P\left(\mathbf{Z} > \frac{b-\mu}{\sigma}\right)}$$

Relationship Between $N(\mu, \sigma^2)$ and $N(0,1)$

For any normal variable:

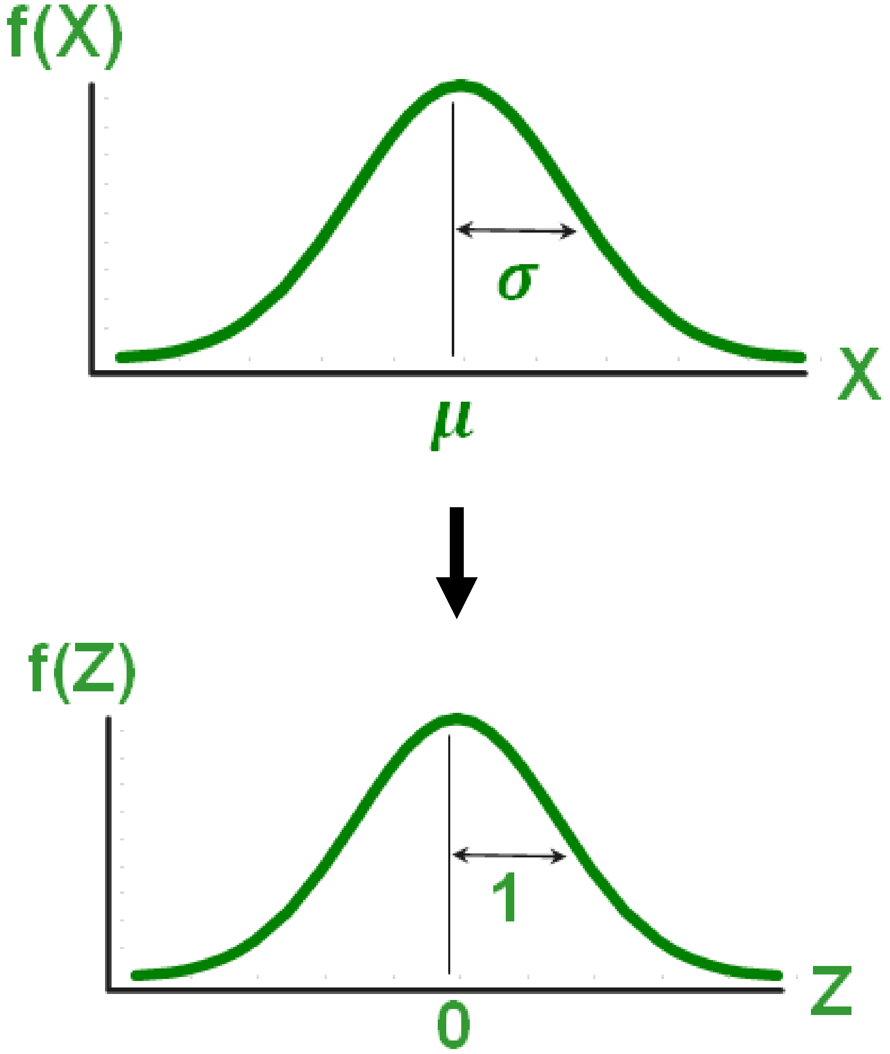
$$X \sim N(\mu, \sigma^2)$$

Define new random variable:

$$Z = \frac{X - \mu}{\sigma}$$

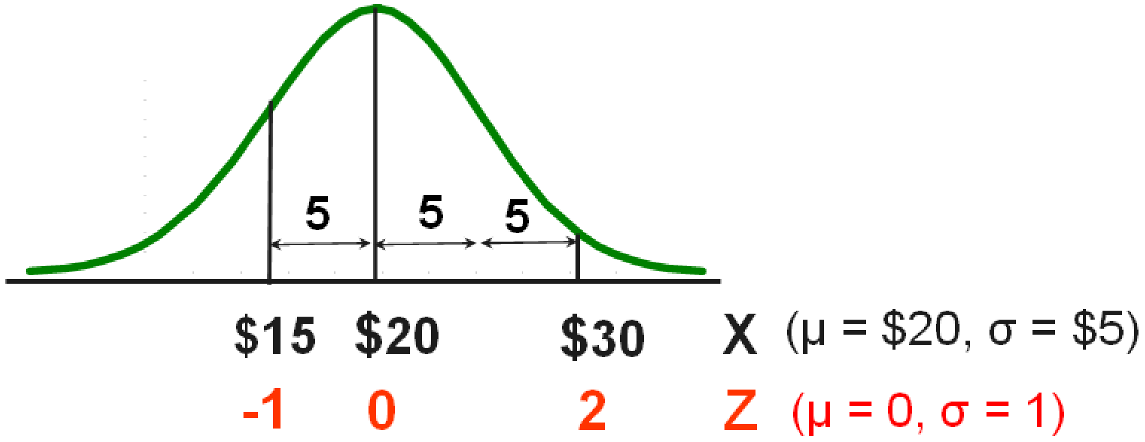
Then Z is a standard normal variable:

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



What does Z tell us?

- X is a normally distributed random variable with mean $\mu = \$20$ and standard deviation $\sigma = \$5$
- For $X = \$30$, the Z value is $Z = \frac{X - \mu}{\sigma} = \frac{30 - 20}{5} = 2$
  - $X = \$30$ is two standard deviations above the mean ($20)
- For $X = \$15$, the Z value is $Z = \frac{X - \mu}{\sigma} = \frac{15 - 20}{5} = -1$
  - $X = \$15$ is one standard deviation below the mean ($20)



Normal Probabilities in 3 Steps

$\ldots$ *if X is normally distributed.*

If $X \sim N(\mu, \sigma^2)$, we can find $P(X > a)$, $P(X < a)$ and $P(a < X < b)$:

Step 1: Convert X-values to Z-values

Step 2: Sketch a bell-shaped curve in terms of Z

Step 3: Use the 1-2-3 rule

Example 1: Wait Time on the Phone

… X *is normally distributed*.

Director of customer support studied the time customers spent on hold waiting for a representative to become available.

- Time spent on hold is normally distributed
    - Mean = 18 mins
    - Standard deviation = 4 mins

Q1. P(a customer waits less than 6 mins)

Q2. P(a customer waits more than 10 mins)

Q3. P(a customer waits more than 22 mins)

Example 1: Wait Time on the Phone

… X is normally distributed.

Time spent on hold, X is a normal variable with $\mu = 18$ mins and $\sigma = 4$ mins.

Q1. P(a customer waits less than 6 mins)

$$P(X < 6) = P\left(\frac{X-\mu}{\sigma} < \frac{6-\mu}{\sigma}\right)$$

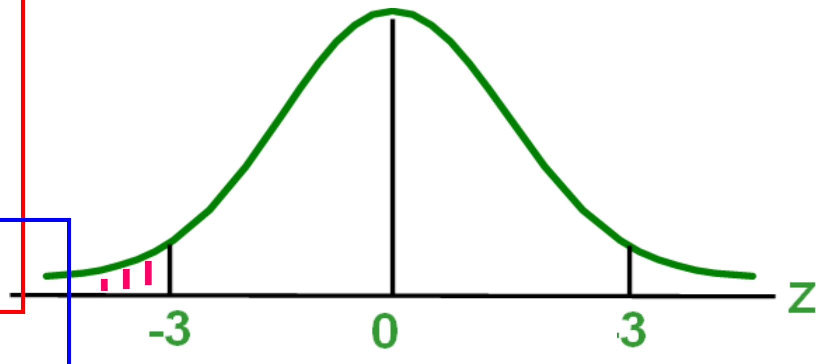
$$= P\left(Z < \frac{6-\mu}{\sigma}\right)$$

$$= P\left(Z < \frac{6-18}{4}\right)$$

$$= P(Z < -3)$$

$$= \frac{1-99.7\%}{2} = 0.0015$$

Standardization ($X \rightarrow Z$)



1-2-3 Rule

Example 1: Wait Time on the Phone

... X is normally distributed.

Time spent on hold, X is a normal variable with $\mu = 18$ mins and $\sigma = 4$ mins.

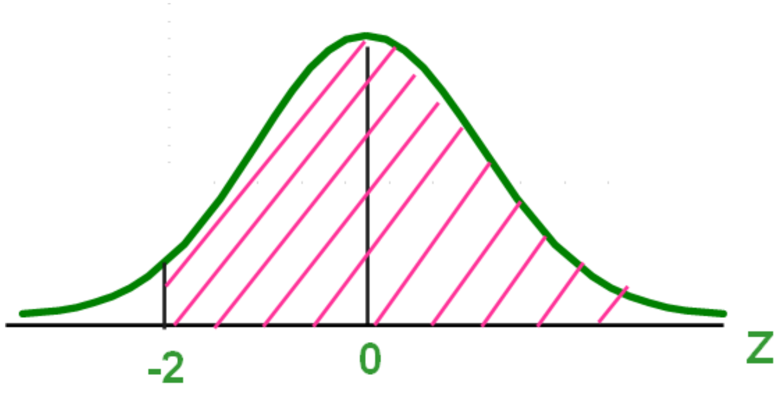
Q2. P(a customer waits more than 10 mins) =

$$P(X > 10) = P\left(\frac{X-\mu}{\sigma} > \frac{10-\mu}{\sigma}\right)$$

$$= P\left(Z > \frac{10-\mu}{\sigma}\right)$$

$$= P\left(Z > \frac{10-18}{4}\right)$$

$$= P(Z > -2)$$



Example 1: Wait Time on the Phone

… X is normally distributed.

Time spent on hold, X is a normal variable with $\mu = 18$ mins and $\sigma = 4$ mins.

Q2. P(a customer waits more than 10 mins) =

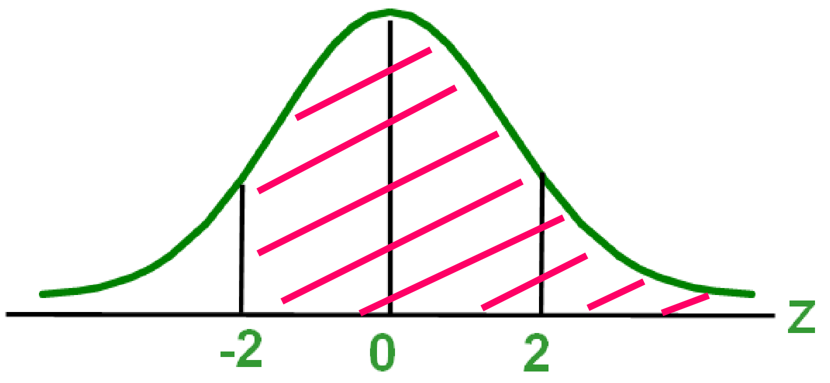
$$P(X > 10) = P\left(\frac{X-\mu}{\sigma} > \frac{10-\mu}{\sigma}\right)$$

$$= P\left(Z > \frac{10-\mu}{\sigma}\right)$$

$$= P\left(Z > \frac{10-18}{4}\right)$$

$$= P(Z > -2)$$

$$= 1 - P(Z < -2) = 1 - \frac{1-95\%}{2} = 0.9750$$



Example 1: Wait Time on the Phone

… X is normally distributed.

Time spent on hold, X is a normal variable with $\mu = 18$ mins and $\sigma = 4$ mins.

Q3. P(a customer waits more than 22 mins) =

$$P(X > 22) = P\left(\frac{X-\mu}{\sigma} > \frac{22-\mu}{\sigma}\right)$$

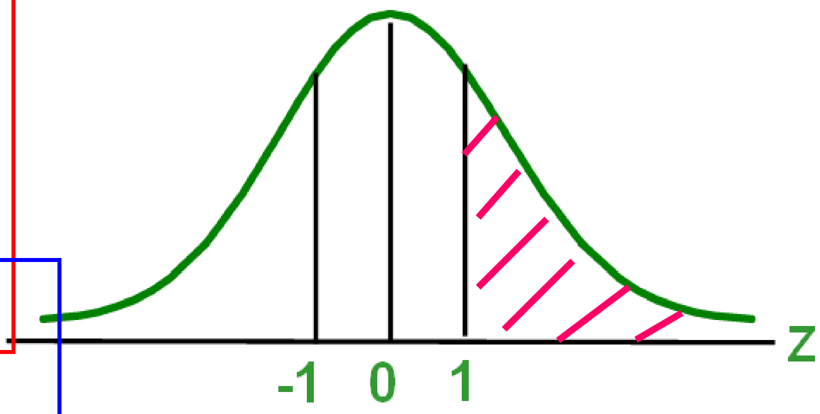
$$= P\left(Z > \frac{22-\mu}{\sigma}\right)$$

$$= P\left(Z > \frac{22-18}{4}\right)$$

$$= P(Z > 1)$$

$$= \frac{1-68\%}{2} = 0.16$$

Standardization ($X \rightarrow Z$)



1-2-3 Rule

Example 2: Image Download Times

... X is normally distributed.

Download time X is a normal variable with $\mu = 18$ seconds and $\sigma = 5$ seconds.

Q1. P(Downloading time is between 18 and 23) =

Step 1: $P(18 < X < 23)$

$$= P\left(\frac{18 - \mu}{\sigma} < \frac{X - \mu}{\sigma} < \frac{23 - \mu}{\sigma}\right)$$

$$= P\left(\frac{18 - \mu}{\sigma} < Z < \frac{23 - \mu}{\sigma}\right)$$

$$= P\left(\frac{18 - 18}{5} < Z < \frac{23 - 18}{5}\right)$$

$$= P(0 < Z < 1)$$

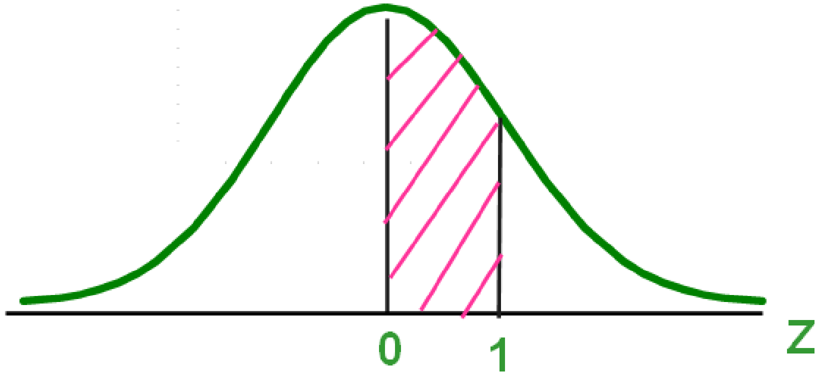
Example 2: Image Download Times

… X is normally distributed.

Download time X is a normal variable with $\mu = 18$ seconds and $\sigma = 5$ seconds.

Q1. P(Downloading time is between 18 and 23) =

Step 2:



Step 3: $\mathrm{P}(18 < X < 23) \quad = P(0 < Z < 1)$

$$= \frac{68\%}{2} \quad = 34\%$$

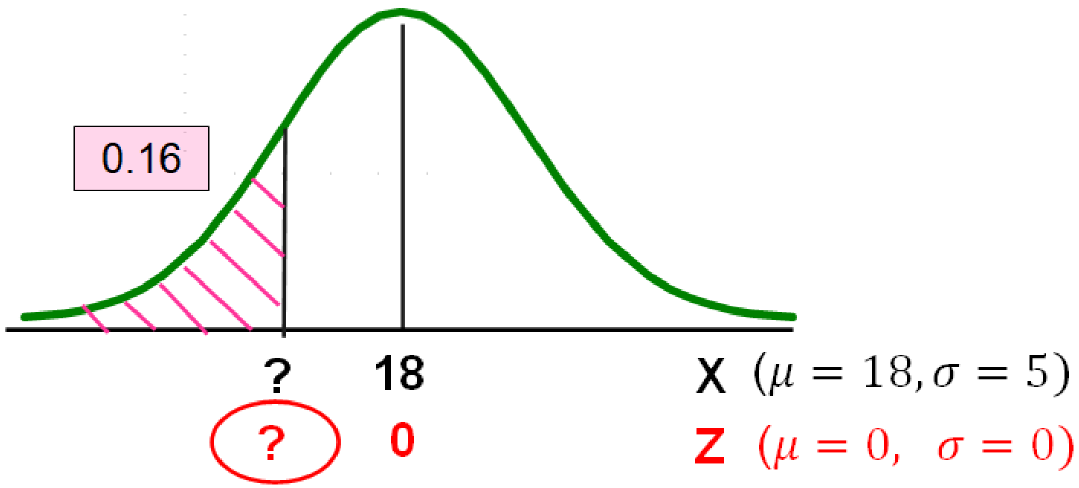
Example 2: Image Download Times

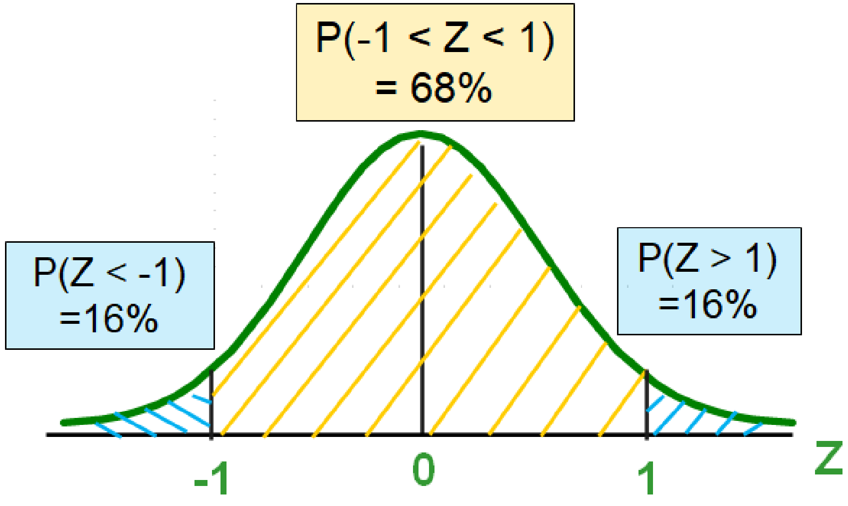
… X is normally distributed.

Download time X is a normal variable with $\mu = 18$ seconds and $\sigma = 5$ seconds.

Q2. Find X such that 16% of download times are less than X.

Step 1: Find the Z value for the known probability.





Example 2: Image Download Times

… X is normally distributed.

Download time X is a normal variable with $\mu = 18$ seconds and $\sigma = 5$ seconds.

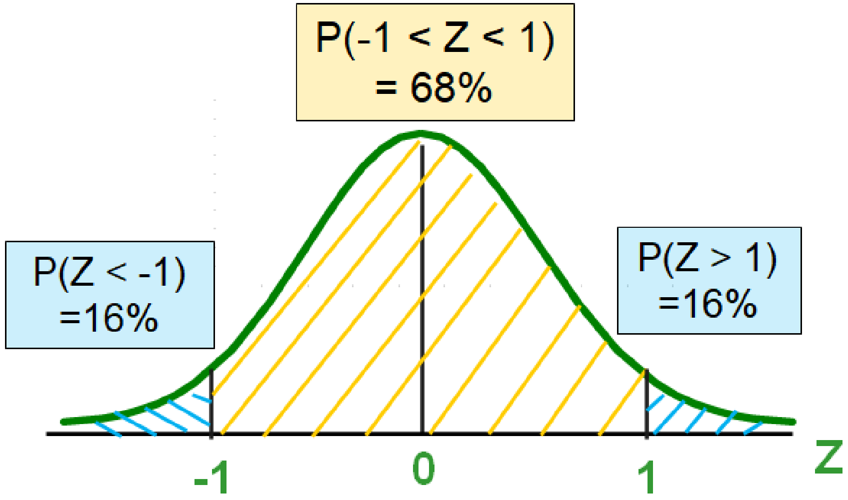
Q2. Find X such that 16% of download times are less than X.

Step 2: Convert $Z = -1$ to X-value.

$$Z = \frac{X-\mu}{\sigma} = -1 \longrightarrow X = \mu + (-1)\cdot\sigma$$

$$= 18 + (-1)\cdot 5$$

$$= 13$$



16% of the values are less than 13 seconds

Distribution of the Sample Mean ($\bar{X}$)

$\ldots$ *review* $\ldots$ *what if we don't know the distribution?*

- Observations $X_1, X_2, \ldots, X_n$

- $\bar{X}$ is a random variable:

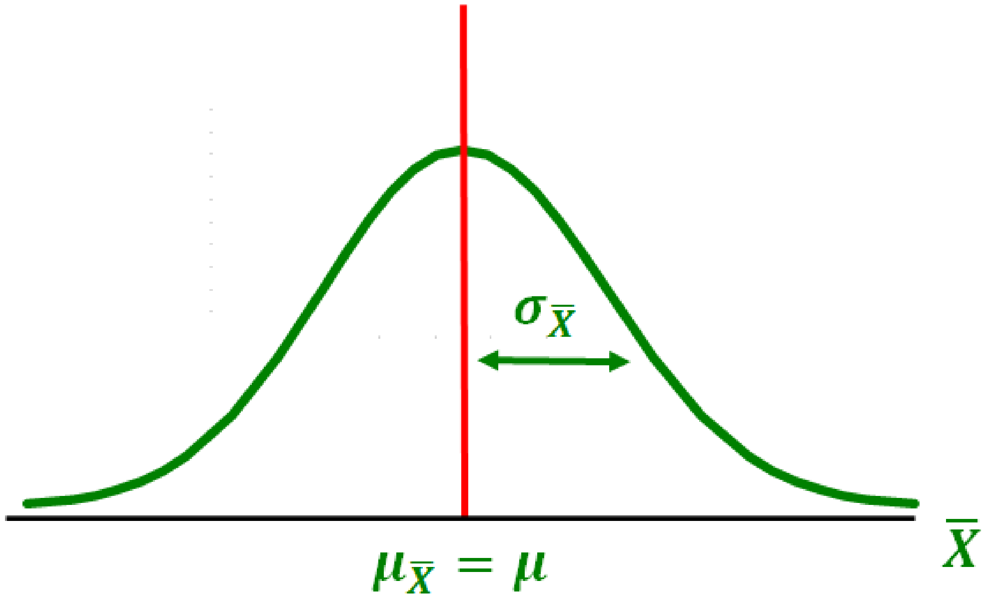
$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

- For any population distribution:

$$\mu_{\bar{X}} = \mu \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} < \sigma$$

- Central Limit Theorem: when *n* is large

$$\bar{X} \sim N\left(\mu_{\bar{X}},\ \sigma_{\bar{X}}^2\right) = N\left(\mu,\ \frac{\sigma^2}{n}\right)$$



Z-Value for the Sample Mean ($\bar{X}$)

Z-value for the distribution of $\bar{X}$

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

where $\bar{X}$ = sample mean

μ = population mean

σ = population standard deviation

n = sample size

Features of $\bar{X}$:

$$\mu_{\bar{X}} = \mu$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

$$\bar{X} \sim N\left(\mu_{\bar{X}}, \sigma_{\bar{X}}^2\right)$$

Example 3

… X *is* NOT *normally distributed*.

Suppose a population has mean $\mu = 8$ and standard deviation $\sigma = 3$. A random sample of size $n = 36$ is selected.

Q. What is the probability that the sample mean is between 6.5 and 9?

Mean: $\mu_{\bar{X}} = \mu = 8$

Standard deviation: $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{3}{\sqrt{36}} = \dfrac{3}{6} = 0.5$

Example 3

… X is NOT *normally distributed.*

$$P(6.5 < \bar{X} < 9)$$

$$= P\left(\frac{6.5 - \mu_{\bar{X}}}{\sigma_{\bar{X}}} < \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} < \frac{9 - \mu_{\bar{X}}}{\sigma_{\bar{X}}}\right)$$

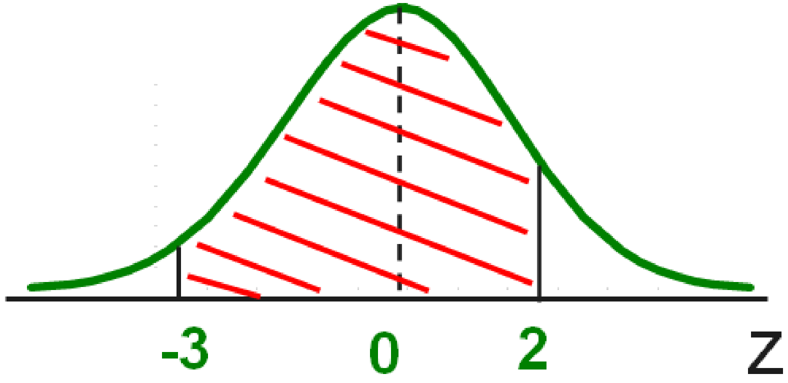
$$= P\left(\frac{6.5 - 8}{0.5} < Z < \frac{9 - 8}{0.5}\right)$$

$$= P(-3 < Z < 2)$$

$$\mu_{\bar{X}} = \mu = 8$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = 0.5$$

$$\bar{X} \sim N\left(\mu_{\bar{X}}, \sigma_{\bar{X}}^2\right) = N(8, 0.5^2)$$



Example 3

... X is NOT *normally distributed.*

$$\mu_{\bar{X}} = \mu = 8$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = 0.5$$

$$\bar{X} \sim N\left(\mu_{\bar{X}}, \sigma_{\bar{X}}^2\right) = N(8, 0.5^2)$$

$$\mathrm{P}(6.5 < \bar{X} < 9)$$

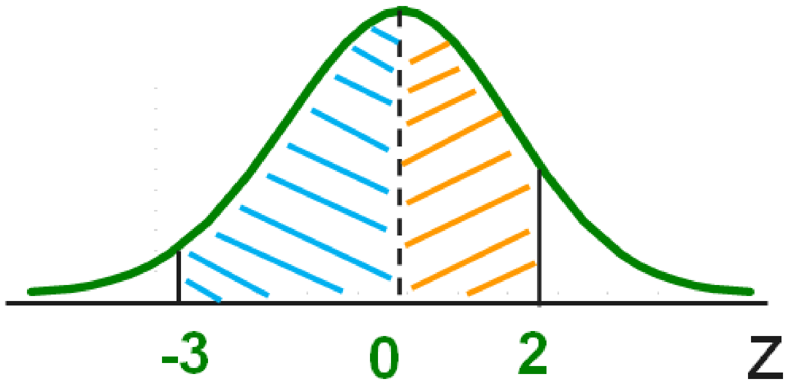
$$= \mathrm{P}\left(\frac{6.5 - \mu_{\bar{X}}}{\sigma_{\bar{X}}} < \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} < \frac{9 - \mu_{\bar{X}}}{\sigma_{\bar{X}}}\right)$$

$$= \mathrm{P}\left(\frac{6.5 - 8}{0.5} < Z < \frac{9 - 8}{0.5}\right)$$

$$= \mathrm{P}(-3 < \mathrm{Z} < 2)$$

$$= \underline{\mathrm{P}(-3 < \mathrm{Z} < 0)} + \underline{\mathrm{P}(0 < \mathrm{Z} < 2)}$$

$$= \frac{99.7\%}{2} + \frac{95\%}{2} \quad = 0.974$$



Example 4

… X *is* NOT *normally distributed*.

Suppose a population has mean $\mu = 368$ and standard deviation $\sigma = 15$. A random sample of size $n = 25$ is selected.

Q. What is a symmetrically distributed interval around μ that includes 95% of all sample means?

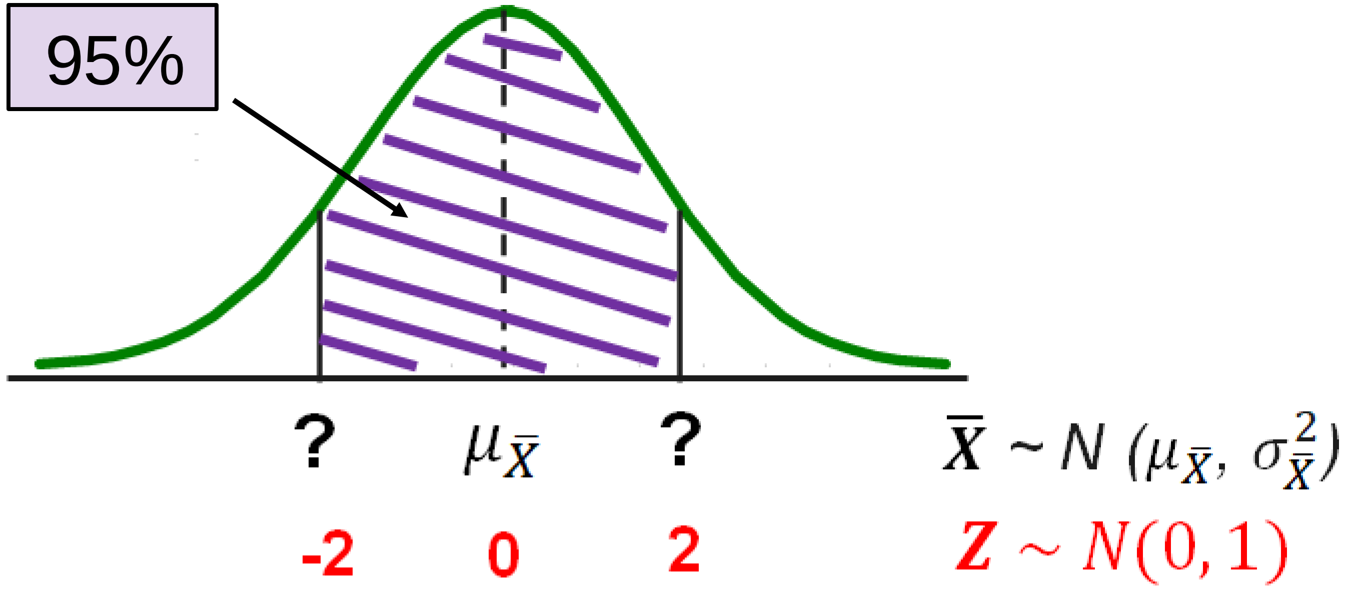
Example 4

... *X is* NOT *normally distributed.*

Suppose a population has mean $\mu = 368$ and standard deviation $\sigma = 15$. A random sample of size $n = 25$ is selected.

Q. What is a symmetrically distributed interval around μ that includes 95% of all sample means?

Step 1: Find the Z value for the known probability.



Example 4

… *X is* NOT *normally distributed.*

Q. What is a symmetrically distributed interval around μ that includes 95% of all sample means?

Step 2: Convert $Z = 2$ and -2 to $\bar{X}$

$$\text{Upper bound of } \bar{X} = \mu_{\bar{X}} + Z \cdot \sigma_{\bar{X}}$$

$$= \mu \;\; + Z \cdot \frac{\sigma}{\sqrt{n}}$$

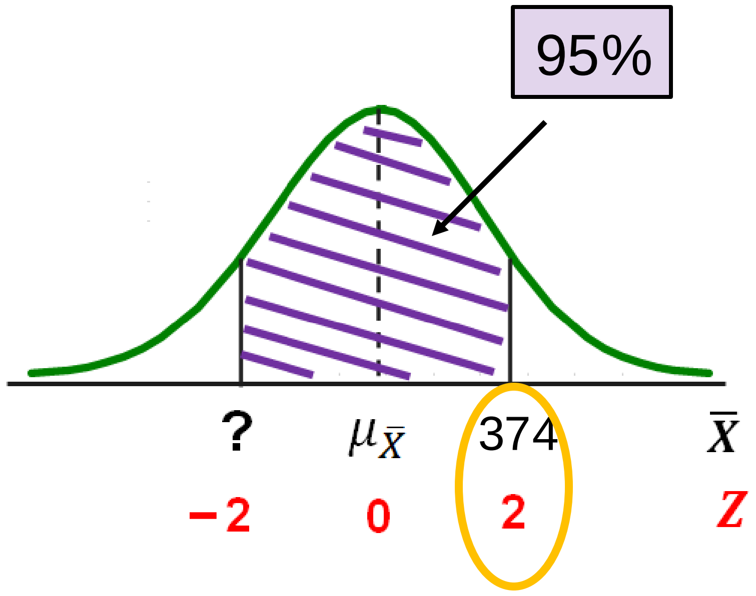
$$= \mu \;\; + 2 \cdot \frac{\sigma}{\sqrt{n}}$$

$$= 368 \;\; + 2 \cdot \frac{15}{\sqrt{25}} \;\; = 374$$

$\mu = 368$

$\sigma = 15$

$\mathrm{n} = 25$



Example 4

… *X is* NOT *normally distributed.*

Q. What is a symmetrically distributed interval around μ that includes 95% of all sample means?

Step 2: Convert $Z = 2$ and -2 to $\bar{X}$

$$\text{Lower bound of } \bar{X} = \mu_{\bar{X}} + Z \cdot \sigma_{\bar{X}}$$

$$= \mu + Z \cdot \frac{\sigma}{\sqrt{n}}$$

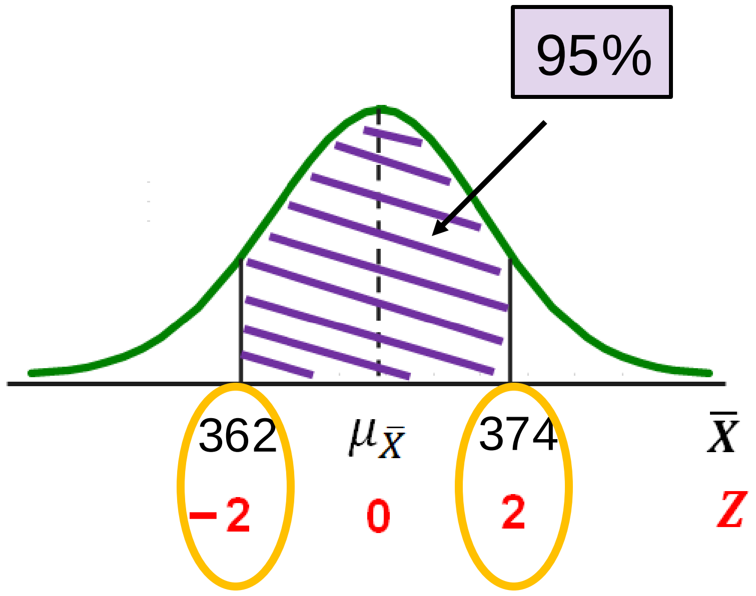
$$= \mu + (-2) \cdot \frac{\sigma}{\sqrt{n}}$$

$$= 368 + (-2) \cdot \frac{15}{\sqrt{25}} = 362$$

$\mu = 368$

$\sigma = 15$

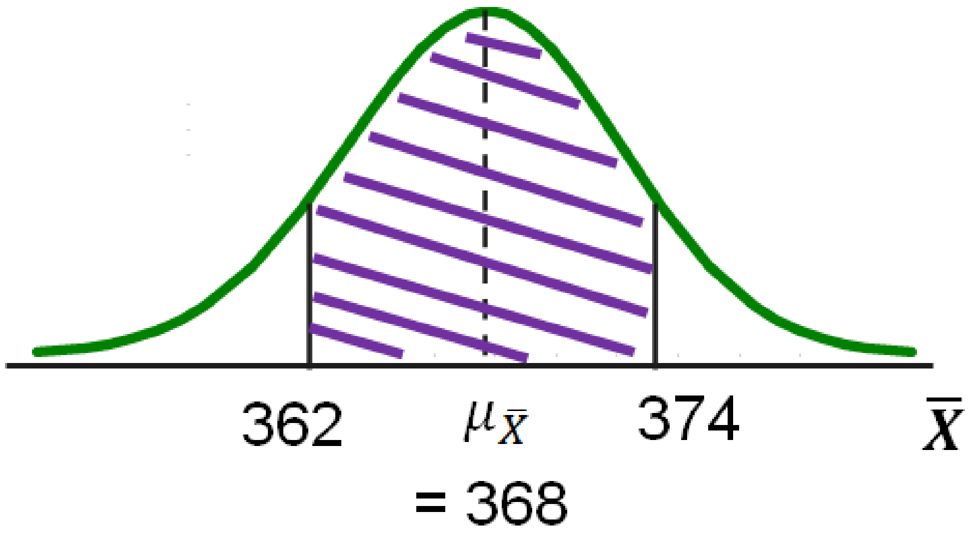
$\text{n} = 25$



Example 4

... *X is* NOT *normally distributed.*

Q. What is a symmetrically distributed interval around μ that includes 95% of all sample means?



Answer: When taking all different samples with size = 25,

95% of all those sample means are between 362 and 374.

Central Limit Theorem in Practice

… in practice we do not know the population statistics like μ *and* σ*.*

The CLT says:

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

Q. How do we use the CLT when we don't know μ or σ?

Use the sample estimate!

Population Statistic ⟵ Sample Estimate

Population Parameter		Guess Based on a Sample	
Population Mean	μ	Sample Mean	$\bar{X}$
Population Variance	σ^2	Sample Variance	S^2
Population Std. Dev.	σ	Sample Std. Dev.	S

Example 5: Vending Machine

… *X is* NOT *normally distributed and we do not know the distribution*.

For a period of 144 days, daily observations have been conducted about the number of candy bars sold from a vending machine. Using these observations:

- Mean of the sample = 258
- Standard deviation of the sample = 60

Q. For a 144-day period, find:

P(**average** number of candy bars sold > 263) =

Example 5: Vending Machine

… X *is* NOT *normally distributed and we do not know the distribution*.

For a period of 144 days, daily observations have been conducted about the number of candy bars sold from a vending machine. Using these observations:

- Mean of the sample = 258
- Standard deviation of the sample = 60

$\bar{X} = 258$,
Good guess of
population mean μ

Q. For a 144-day period, find:

P(**average** number of candy bars sold > 263) =

$S = 60$,
Good guess of
population std. dev. σ

Example 5: Vending Machine

… *X is* NOT *normally distributed and we do not know the distribution.*

$n = 144,$ $\mu = 258,$ $\sigma = 60$

P(**average** number of candy bars sold > 263) =

For $\bar{X}$:
$$\begin{cases} \text{mean } \mu_{\bar{X}} = \mu = 258 \\ \text{standard deviation } \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{60}{\sqrt{144}} = \frac{60}{12} = 5 \\ \bar{X} \sim N(\mu_{\bar{X}}, \sigma_{\bar{X}}{}^2) = N(258, (5)^2) \end{cases}$$

$$P(\bar{X} > 263) = P\left(\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} > \frac{263 - \mu_{\bar{X}}}{\sigma_{\bar{X}}}\right)$$

$$= P\left(Z > \frac{263 - 258}{5}\right) = P(Z > 1) = 0.16$$

Proportions: a special case

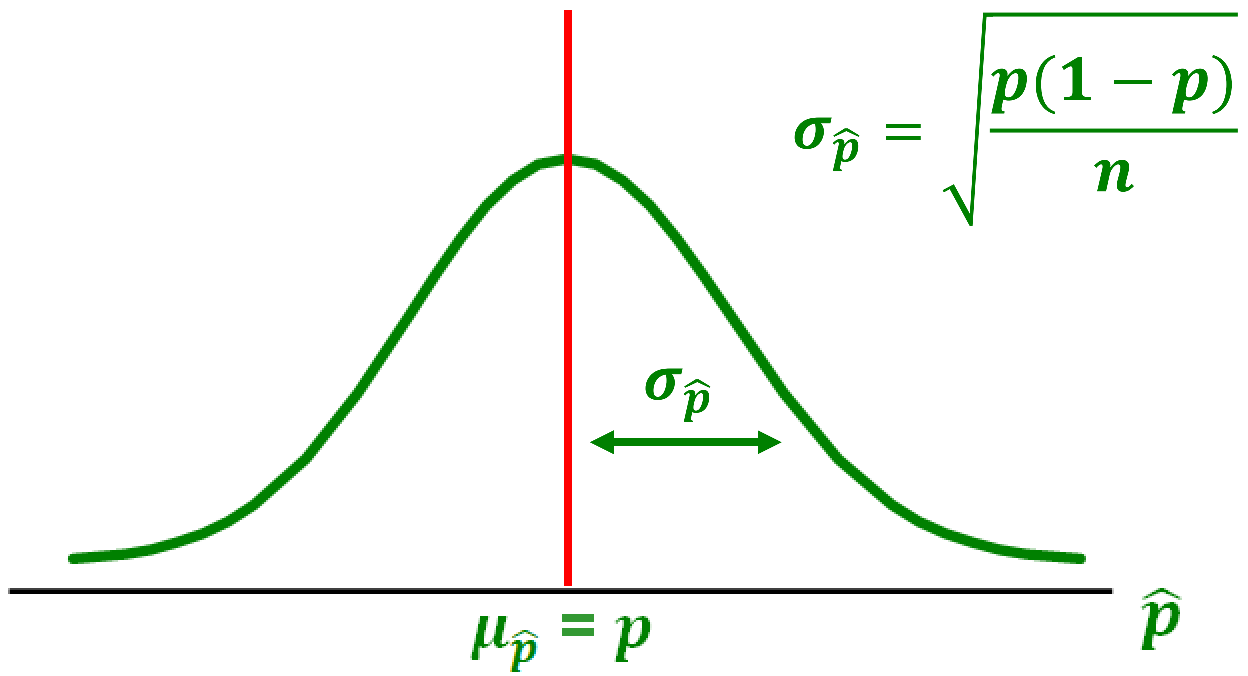
We collect a random sample of size *n*. Each person in the sample may or may not have the characteristics of interest.

$$X_i = \begin{cases} 1 & \text{if person } i \text{ has the characteristic} \qquad \text{(answer “Yes”)} \\ 0 & \text{if not} \qquad \text{(answer “No”)} \end{cases}$$

Each X_i is a binary random variable

$$\text{Prob}\ (X_i = 1) = p \qquad\qquad \text{Prob}\ (X_i = 0) = 1 - p$$

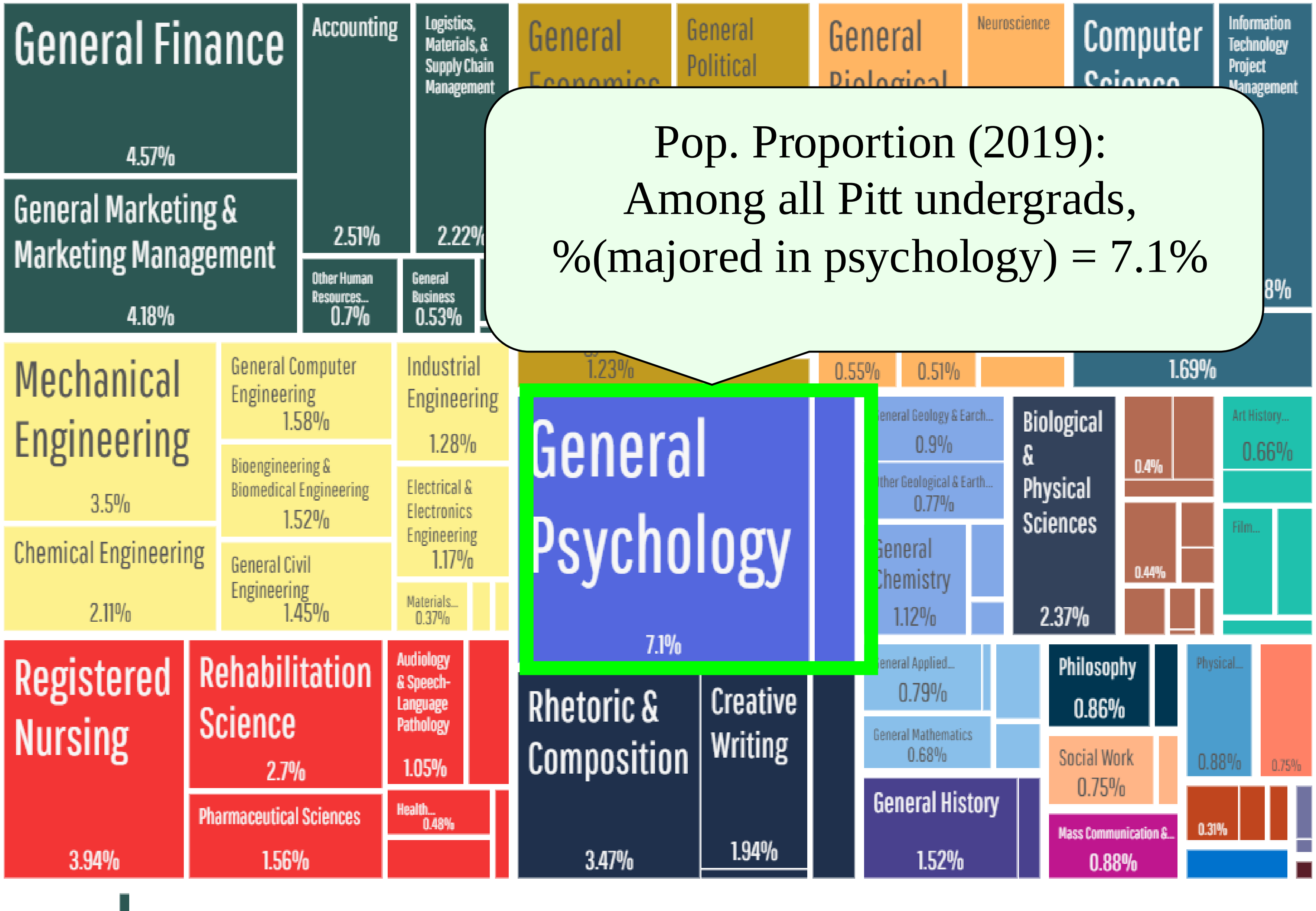
Distribution of Sample Proprotion ($\hat{p}$)



When sample is large enough

$$\hat{p} \sim N\left(\mu_{\hat{p}},\ \sigma_{\hat{p}}^2\right) = N\left(p, \frac{p(1-p)}{n}\right)$$

Majors of Pitt Graduates With a Bachelor’s Degree



Pop. Proportion (2019):
Among all Pitt undergrads,
%(majored in psychology) = 7.1%

Example 6: Psychology Students

… X *is* NOT *normally distributed*.

Q. From 240 randomly selected students, what is the probability that the number of students who will obtain a psychology degree is between 17 and 28?

Let $\hat{p}$ = sample proportion of a sample with size 240

Population

$p = 0.071$

Sample size

$n = 240$

$\mu_{\hat{p}} = p = 0.071$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = 0.015$$

$np > 5$

$n(1-p) > 5$

$\hat{p} \sim N\left(0.071, 0.015^2\right)$

Example 6: Psychology Students

… X *is* NOT *normally distributed*.

$$\mu_{\hat{p}} = 0.071$$
$$\sigma_{\hat{p}} = 0.015$$
$$\hat{p} \sim N(0.071, 0.015^2)$$

P(# of students who will obtain a psychology degree is between 17 and 28) =

$$\cancel{\mathrm{P}(17 < \hat{p} < 28)}$$

$$\mathrm{P}\left(\frac{17}{240} < \hat{p} < \frac{28}{240}\right)$$

proportion
out of a sample of 240

Example 6: Psychology Students

… X *is* NOT *normally distributed*.

$$\mu_{\hat{p}} = 0.071$$
$$\sigma_{\hat{p}} = 0.015$$
$$\hat{p} \sim N(0.071, 0.015^2)$$

P(# of students who will obtain a psychology degree is between 17 and 28) =

$$\mathrm{P}\left(\frac{17}{240} < \hat{p} < \frac{28}{240}\right) = \mathrm{P}(0.071 < \hat{p} < 0.117)$$

$$= \mathrm{P}\left(\frac{0.071 - \mu_{\hat{p}}}{\sigma_{\hat{p}}} < \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} < \frac{0.117 - \mu_{\hat{p}}}{\sigma_{\hat{p}}}\right)$$

$$= \mathrm{P}\left(\frac{0.071 - 0.071}{0.015} < Z < \frac{0.117 - 0.071}{0.015}\right)$$

$$= \mathrm{P}(0 < Z < 3.07)$$

Example 6: Psychology Students

… X *is* NOT *normally distributed*.

$$\mu_{\hat{p}} = 0.071$$
$$\sigma_{\hat{p}} = 0.015$$
$$\hat{p} \sim N(0.071, 0.015^2)$$

P(# of students who will obtain a psychology degree is between 17 and 28)

$$\mathrm{P}\left(\frac{17}{240} < \hat{p} < \frac{28}{240}\right) = \mathrm{P}(0 < Z < 3.07)$$

Accurate answer

$$= \mathrm{P}(Z < 3.07) - \mathrm{P}(Z < 0)$$

$$= \mathrm{norm.s.dist}(3.07, \mathrm{true}) - 0.5 = 0.9989 - 0.5 = 0.4989$$

Hands-on answer:

$$\approx \mathrm{P}(0 < Z < 3) = \frac{99.7\%}{2} = 0.4985$$