# Unit #5: The Normal Distribution and the Central Limit Theorem

4.3, 6.4, 6.5

# Learning Objectives (Normal Distribution)

At the end of this unit, students should be able to:

1. Define the normal distribution, the standard normal distribution, and state the parameters that characterize the these distributions.

2. Identify how the normal distribution changes as a function of the parameters.

3. Describe situations where the normal distribution would be a good model.

4. Calculate probabilities involving the normal distribution using a table and/or R.

5. Define the critical value for a normal distribution

6. Convert a non-standard normal distribution to a standard normal distribution.

# Learning Objectives (CLT)

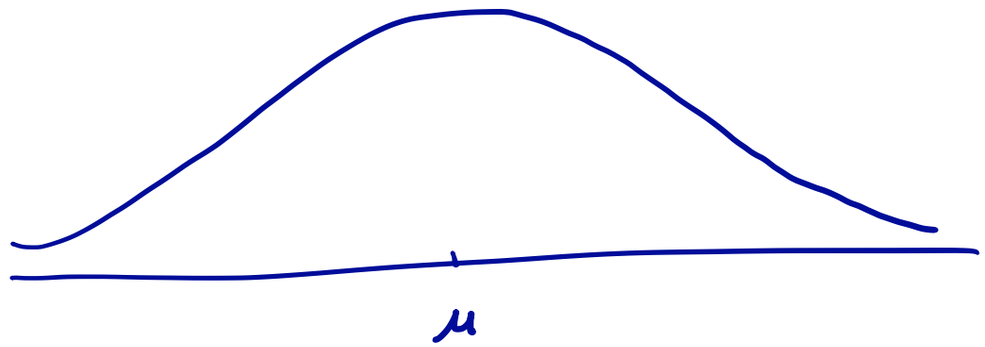
1. Describe and provide examples of statistical inference.
2. Define independent and identically distributed and a simple random sample.
3. Define and provide examples of an *estimator.*
4. Describe why estimators (e.g., the sample mean) has a probability distribution (that is, describe why estimators are random variables).
5. Define the sampling distribution of an estimator.
6. Describe the three features upon which the sampling distribution of an estimator depends.
7. Define the standard error of an estimator.
8. Write R code that illustrates the fact that an estimator (e.g., the mean) has a sampling distribution.
9. Describe the mean and variance of the sample mean of a sample.
10. State the Central Limit Theorem (CLT), which characterizes the distribution of the sample mean.
11. Apply the central limit theorem to answer questions about the mean of a simple random sample.

# The Normal Distribution

The normal distribution (sometimes called the Gaussian distribution) is probably the most important distribution in all of probability and statistics.

Many populations have distributions that can be fit very closely by an appropriate normal (or Gaussian, bell) curve.

Examples: height, weight, and other physical characteristics, scores on various tests, etc.




$X \sim N(0, 3)$

**Definition:** A continuous r.v. *X* is said to have a *normal distribution* with parameters mean, $\mu$ and S.D., $\sigma$ > *0*, if the pdf of *X* is:

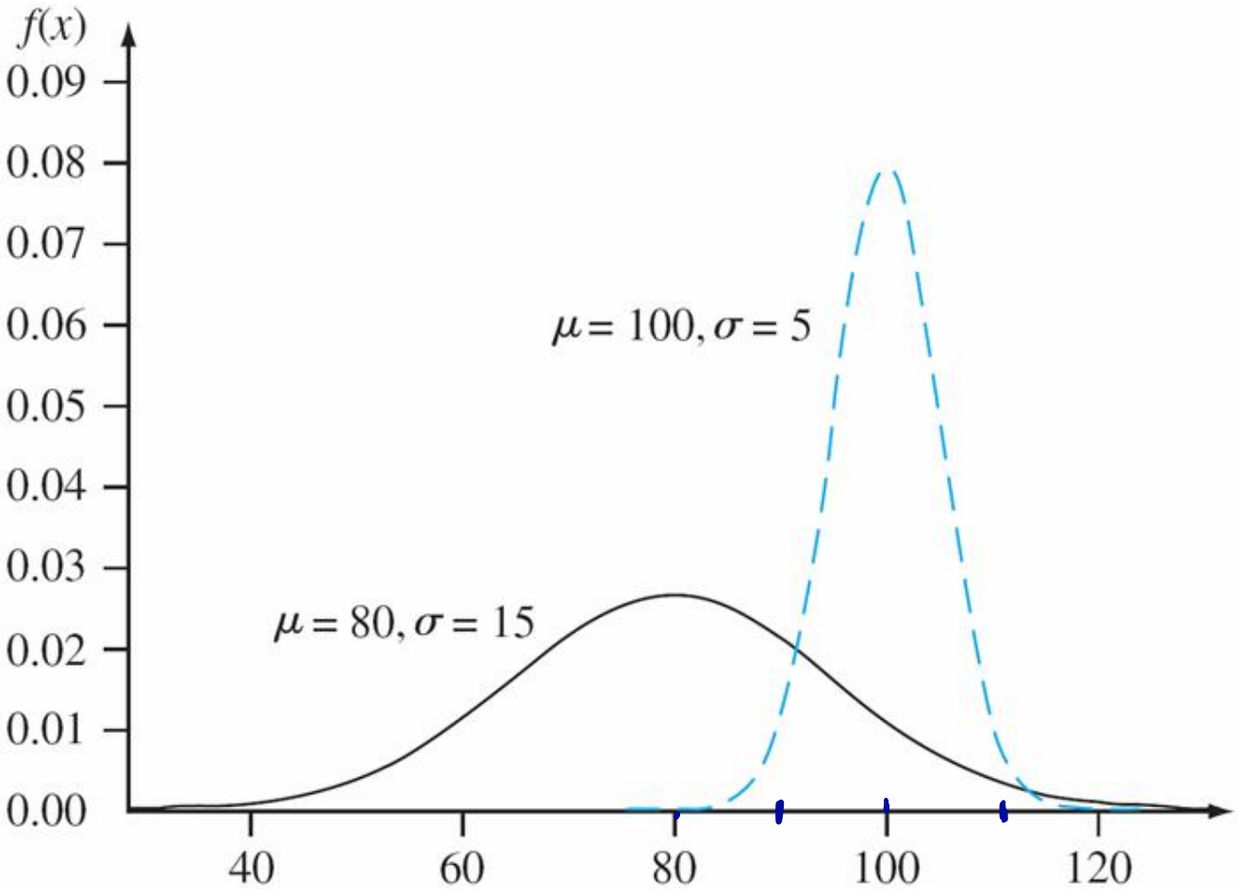
$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}$$

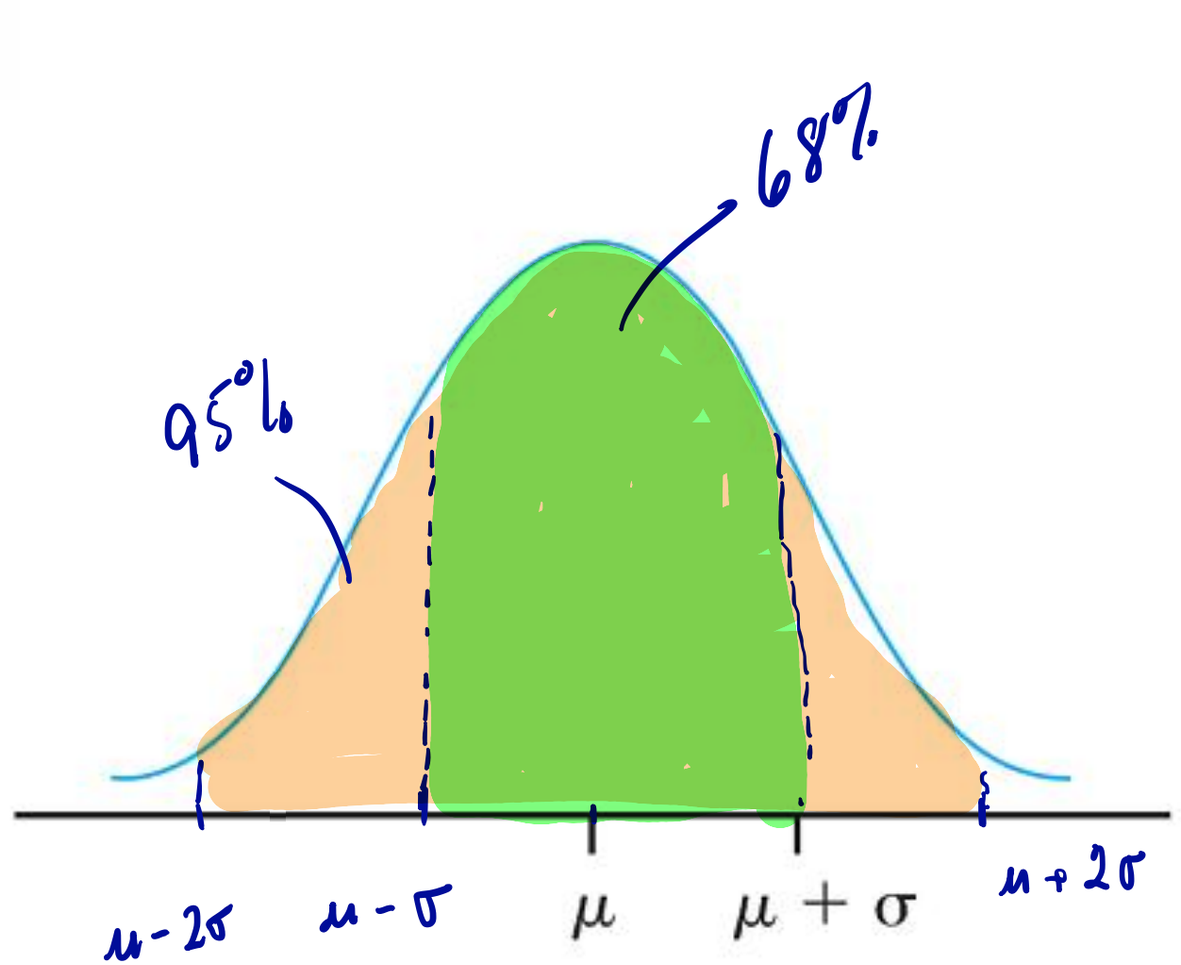
Notation: $X \sim N(\mu, \sigma^2)$

↑ Variance

# The Normal Distribution

The figure below presents graphs of *f* for different parameter pairs:



# The Standard Normal Distribution

**Definition:** The normal distribution with parameter values $\mu = 0$ and $\sigma = 1$ is called the *standard normal distribution*.

A r.v. with this distribution is called a standard normal random variable and is denoted by $Z$. Its pdf is:

$$f(z;0,1) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad z \in \mathbb{R}$$

# The Standard Normal Distribution

We use special notation to denote the cdf of the standard normal curve:

$$F(z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$$

$X \sim N(\mu, \sigma^2)$: $F(x) =$ pnorm$(x, \mu, \sigma)$

# The Standard Normal Distribution

Note:

1. The standard normal distribution <u>rarely</u> occurs naturally.

2. Instead, it is a *reference distribution* from which information about other normal distributions can be obtained via a simple formula.

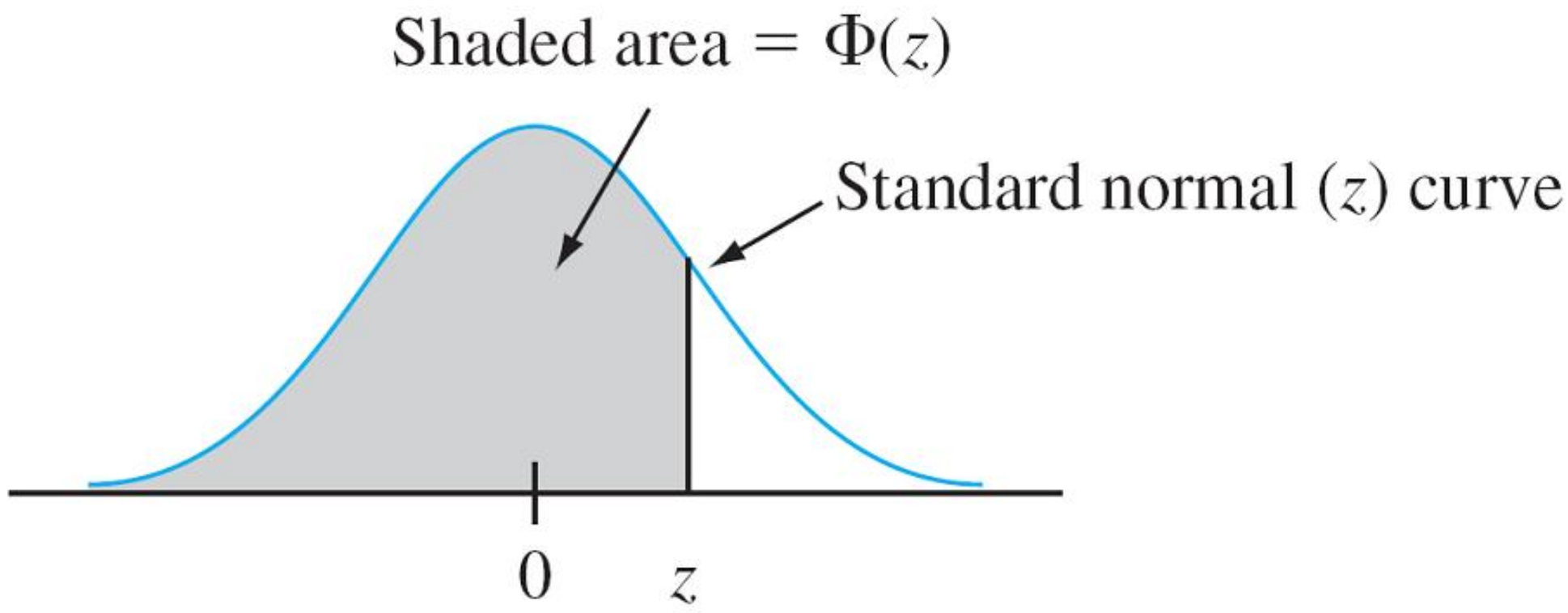
3. These probabilities can then be found "normal tables".

4. This can also be computed with a single command in R.

$\Phi(z) =$ pnorm$(z)$

# The Standard Normal Distribution

The figure below illustrates the probabilities found in a normal table (this can easily be found online):



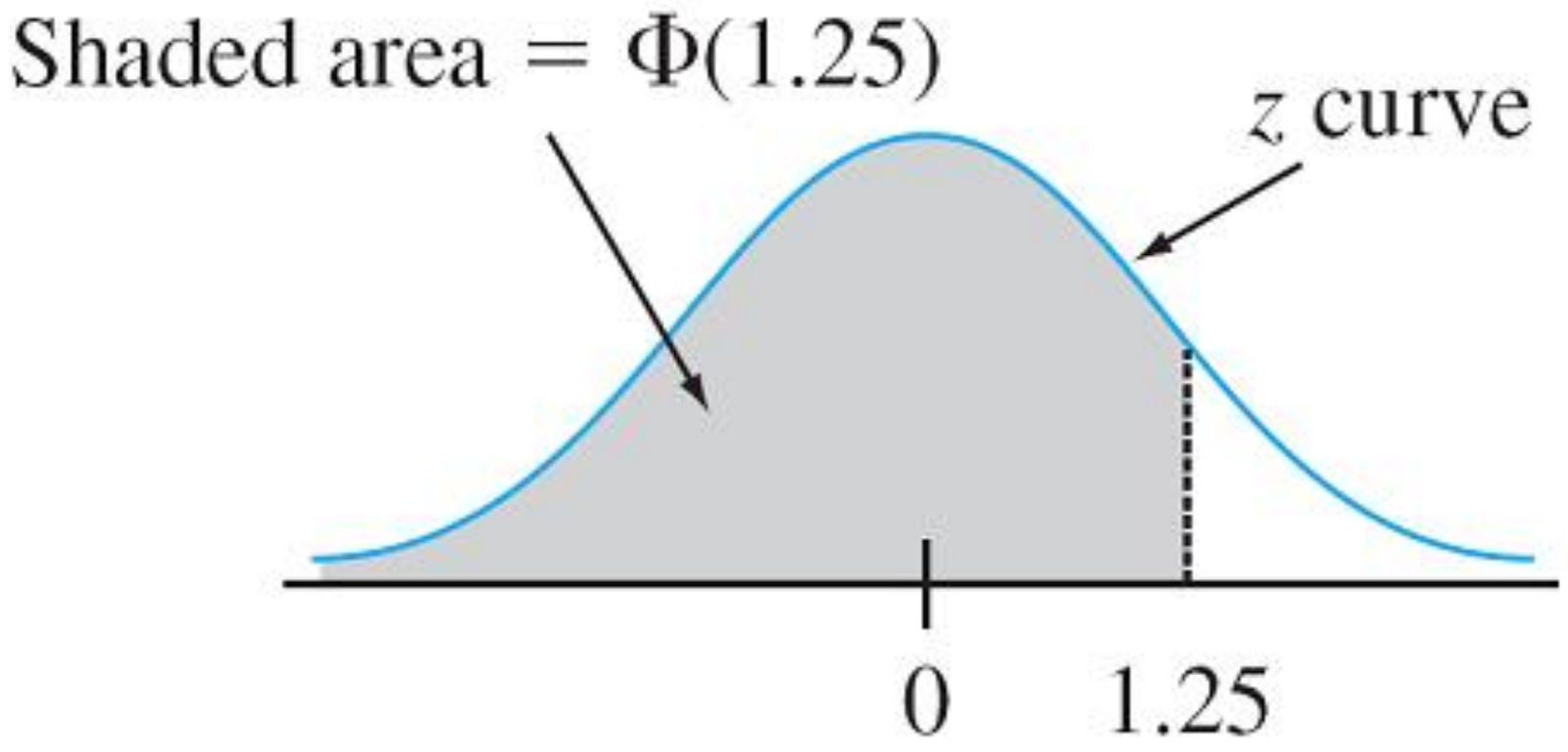
x is the $p^{th}$ % tile iff $F(x) = p/100$ } quantiles in R: qnorm(p)

# The Standard Normal Distribution

$\approx 0.8944$

$P(Z \leq 1.25) = \Phi(1.25)$, a probability that is tabulated in a normal table. What is this probability?

The figure below illustrates this probability:



# The Standard Normal Distribution

Examples:

1. $P(Z \geq 1.25)$ $= 1 - \Phi(1.25) = 1 - 0.8944 = 0.106$

2. Why does $P(Z \leq -1.25) = P(Z \geq 1.25)$? What is $\Phi(-1.25)$? Symmetry

3. How do we calculate $P(-.38 \leq Z \leq 1.25)$?

$\Phi(1.25) - \Phi(-0.38)$

# The Standard Normal Distribution

The *99th* percentile of the standard normal distribution is that value of *z* such that the area under the *z* curve to the left of the value is *0.99.*

Tables give, for fixed *z,* the area under the standard normal curve to the left of *z;* now we have the area and want the value of *z*.

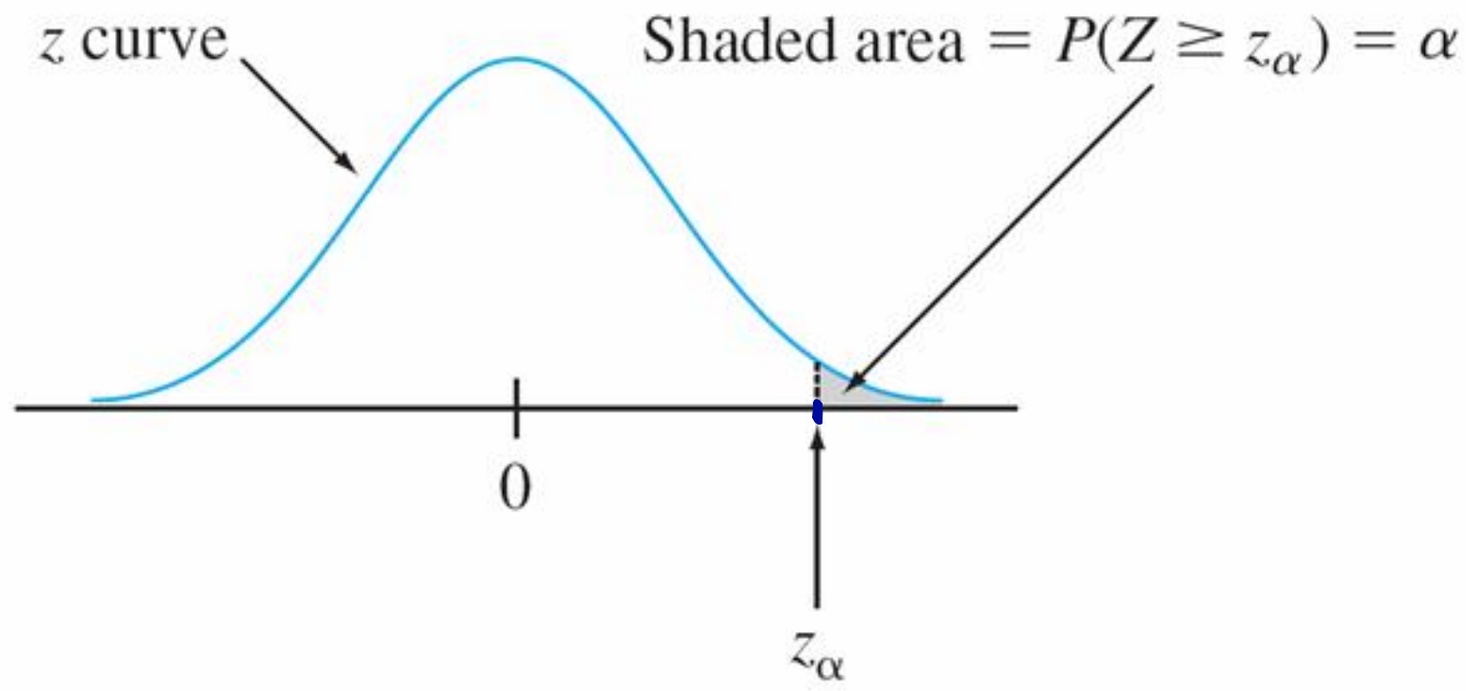
This is the "inverse" problem to $P(Z \leq z) = ?$

How can the table be used for this?

# Critical Values

In statistical inference, we need the *z* values that give certain tail areas under the standard normal curve.

There, this notation will be standard: $z_\alpha$ will denote the *z* value for which  of the area under the *z* curve lies to the right of $z_\alpha$.



# Non-Standard Normal Distributions

When $X \sim N(\mu, \sigma^2)$, probabilities involving $X$ are computed by "standardizing." The **standardized variable** is:

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

Standardized variable

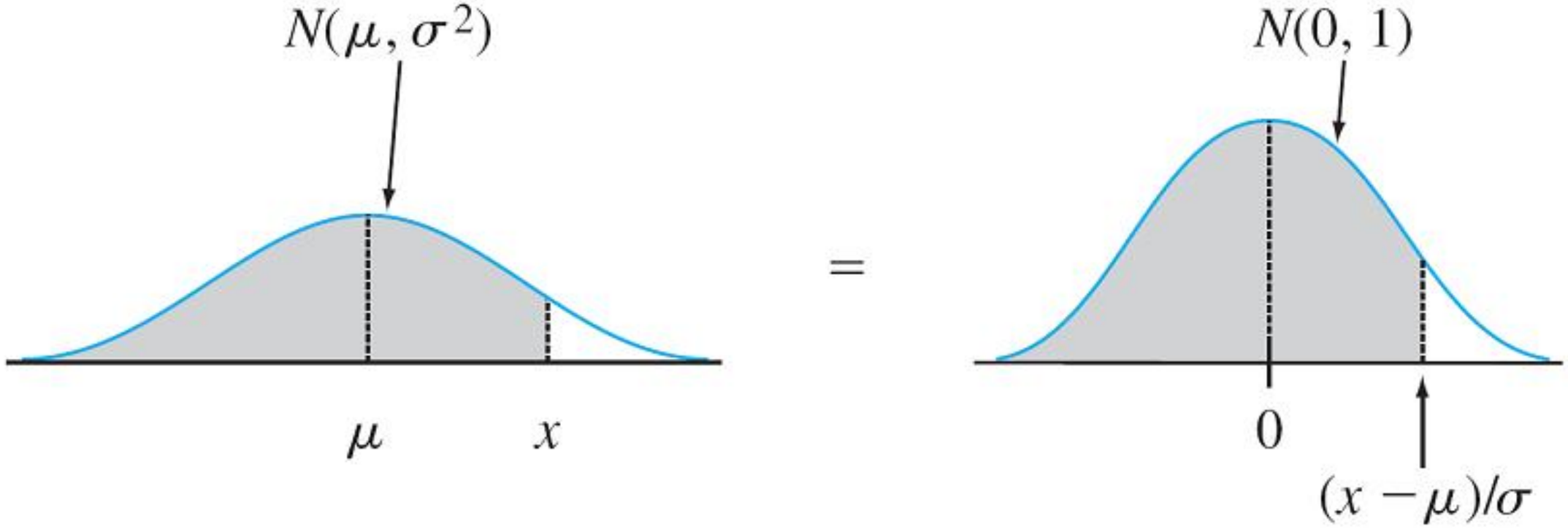
**Proposition:** If $X$ has a normal distribution with mean $\mu$ and standard deviation $\sigma$, then

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

is distributed standard normal.

# Non-Standard Normal Distributions

Why do we standardize normal random variables?




Equality of nonstandard and standard normal curve areas

# Non-Standard Normal Distributions

Example: The time that it takes a driver to react to the brake lights on a decelerating vehicle is critical in helping to avoid rear-end collisions.

Research suggests that reaction time for an in-traffic response to a brake signal from standard brake lights can be modeled with a normal distribution having mean value 1.25 sec and standard deviation of 0.46 sec.

$$X \sim N(1.25, 0.46^2)$$

What is the probability that reaction time is between 1.00 sec and 1.75 sec?

$$P(1 \le X \le 1.75) = P\left(\underbrace{\frac{1-1.25}{0.46}}_{a} \le Z \le \underbrace{\frac{1.75-1.25}{0.46}}_{b}\right)$$

$$= \Phi(b) - \Phi(a)$$

$X_{1,1}, \ldots, X_n \to \bar{X}$ ; $X'_{1,1}, \ldots, X'_n \to \bar{X}'$ ; $\cdots$ ; $X''_{1,1}, \ldots, X''_n \to \bar{X}''$

m samples → ∞

# Statistical Inference: Motivating Examples

Soon, we will be focusing on making "statistical inference" about the true mean of a population by using sample datasets. The normal distribution is widely used in statistical inference.

Examples?

1.) Infer the mean profit of a company producing smart phones based on profits in some time period.

2.) Infer the proportion of people that have a certain disease in a given town.

# Random Samples

**Definition**: The r.v.'s $X_1, X_2, \ldots, X_n$ are said to form a (simple) **random sample** of size $n$ if:

1. the $X_i$s are independent

2. each $X_i$ has the same probability dist.

We say that these $X_i$'s are: independent and identically dist. (iid)

ex. $\longrightarrow X_i \overset{iid}{\sim} N(\mu, \sigma^2)$, $i = 1, \ldots, n$

# Estimators and Their Distributions

We use **estimators** to summarize our i.i.d. sample.

Examples?

1.) $\bar{X}$ is an estimator of the population mean $\mu$. ($\hat{\mu}$)

2.) $\hat{p}$ = sample proportion is an estimator of the pop. proportion $p$.

3.) $\hat{\sigma} = s$ is an estimator of the pop. s.d. $\sigma$.

$$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

4.) $Y = \beta_0 + \beta_1 X$. estimators will be $\hat{\beta}_0$ and $\hat{\beta}_1$

# Estimators and Their Distributions

We use *estimators* to summarize our i.i.d. sample. Any estimator, including the ~~*same*~~ sample *mean* $\bar{X}$ is a random variable (since it is based on a random sample).

This means that $\bar{X}$ has a distribution of it's own, which is referred to as **sampling distribution of the sample mean.** This sampling distribution depends on:

1.) The method of sampling
2.) The population dist.
3.) The sample size $n$.

Ex. Let $\bar{X}$ est $\mu$. ($\mu$ = pop mean) $\bar{X}$ has a dist w/ mean $E(\bar{X}) = \mu$ and $Var(\bar{X}) = \sigma^2$. Then $\sigma$ is the standard error.

The standard deviation of this distribution is called **the standard error of the estimator.**

# Distribution of the Sample Mean

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean value $\mu$ and standard deviation $\sigma$. Then:

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum E(X_i)$$

$$= \frac{1}{n}\sum \mu = \frac{1}{n} \cdot n\mu = \mu$$

$$Var(\bar{X}) = \ldots = \frac{\sigma^2}{n}$$

The standard deviation of the sample mean is:

$$s.e.(\bar{X}) = s.d.(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

This is also called the standard error of the mean.

# Distribution of the Sample Mean

Great, but what is the *distribution* of the sample mean?

# Distribution of the Sample Mean (Normal Population)

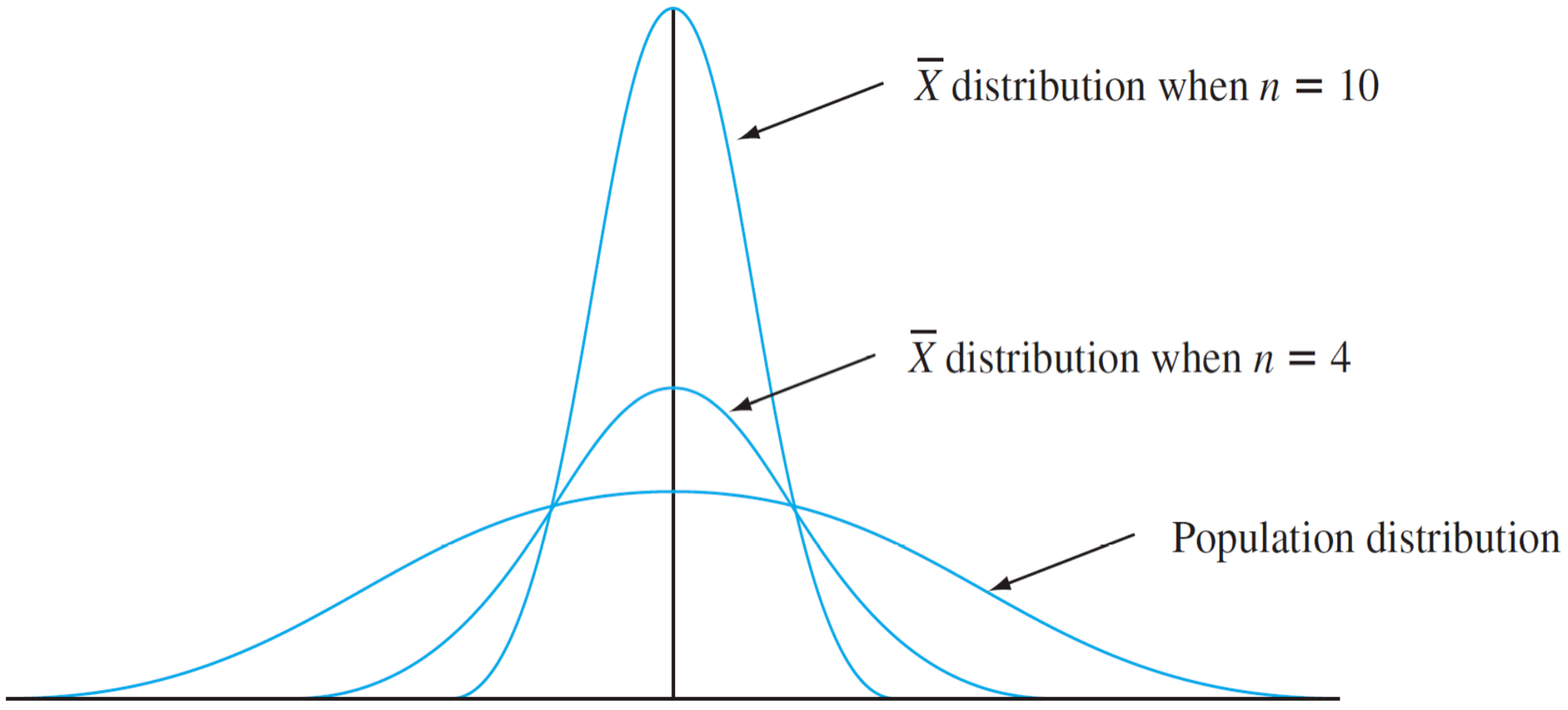
Proposition:

Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. For any $n$,

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

We know everything there is to know about the distribution of the sample mean when the population distribution is normal.

# Distribution of the Sample Mean (Normal Population)



# The Central Limit Theorem

But what if the underlying distribution of the $X_i$'s is not normal?

# The Central Limit Theorem

**Important:** When the population distribution is non-normal, averaging produces a distribution more bell-shaped than the one being sampled.

A reasonable conjecture is that if $n$ is large, a suitable normal curve will approximate the actual distribution of  the sample mean.

The formal statement of this result is one of the most important theorems in probability: **Central Limit Theorem!**

# The Central Limit Theorem

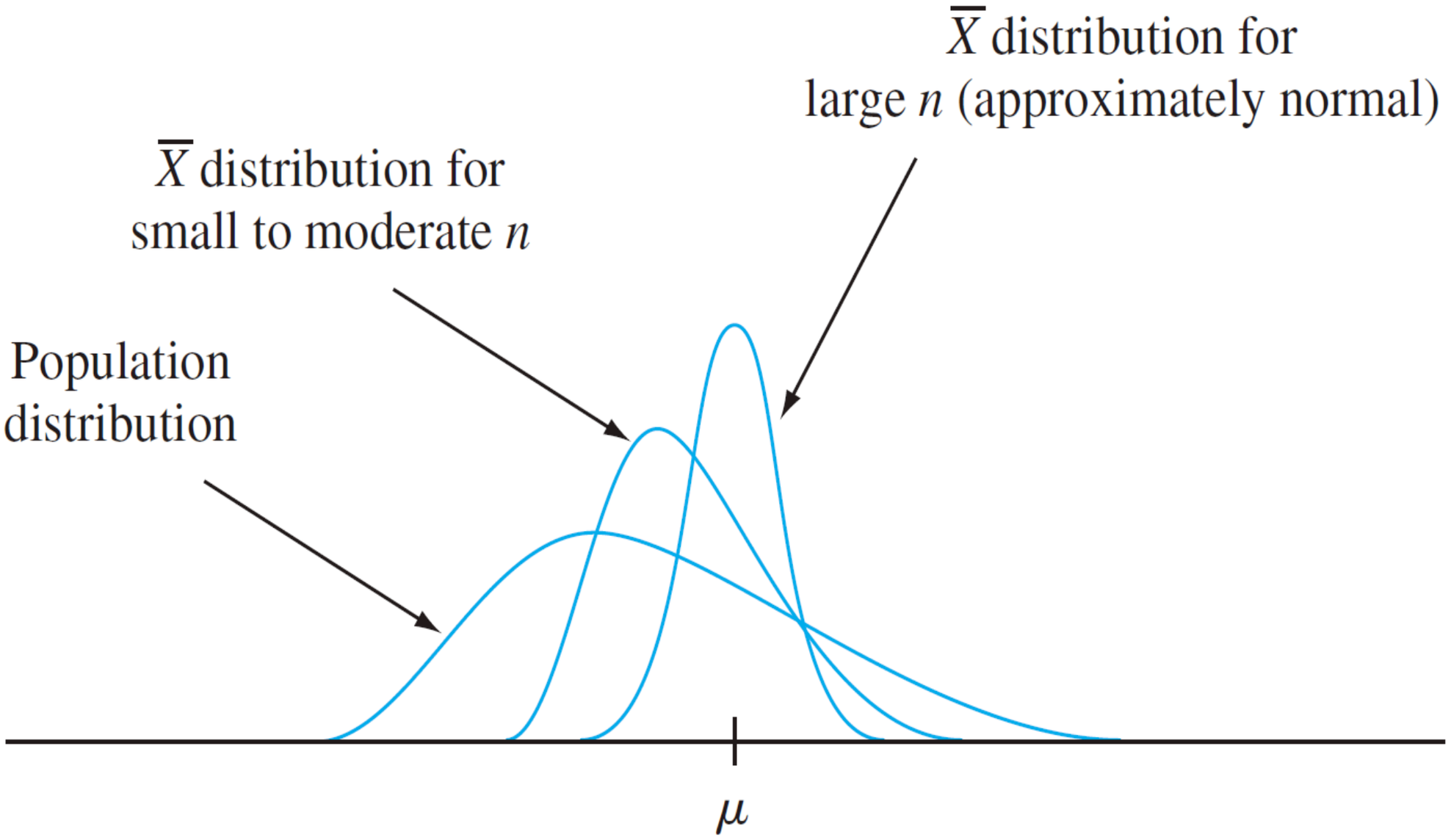
**Theorem**

The Central Limit Theorem (CLT): Let $X_1, \ldots, X_n$ be iid from some pop. w/ $E(X_i) = \mu$, and $\text{Var}(X_i) = \sigma^2$. Then, for a large enough $n$,

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

The larger the value of *n*, the better the approximation! Typical rule of thumb: $n \geq 30$

# The Central Limit Theorem



# The Central Limit Theorem

Example: The amount of impurity in a batch of a chemical product is a random variable with mean value 4.0 g and standard deviation 1.5 g. (unknown distribution)

If 50 batches are independently prepared, what is the (approximate) probability that the average amount of impurity in these 50 batches is between 3.5 and 3.8 g?

# The Central Limit Theorem

The CLT provides insight into why many random variables have probability distributions that are approximately normal.

For example, the measurement error in a scientific experiment can be thought of as the sum of a number of underlying perturbations and errors of small magnitude.

A practical difficulty in applying the CLT is in knowing when *n* is sufficiently large. The problem is that the accuracy of the approximation for a particular *n* depends on the shape of the original underlying distribution being sampled.