Introduction

**Review:** A continuous random variable can assume any value between two endpoints.

Many continuous random variables have an approximately normal distribution, which means we can use the distribution of a normal random variable to make statistical inference about the variable.
Example: Heights of randomly selected women

(a) Random sample of 100 women

(b) Sample size increased and class width decreased

(c) Sample size increased and class width decreased further

(d) Normal distribution for the population
Section 6-1: Properties of a Normal Distribution

A normal distribution is a continuous, symmetric, bell-shaped distribution of a variable.

The theoretical shape of a normal distribution is given by the mathematical formula

\[
y = \frac{e^{\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}.
\]

where \( \mu \) and \( \sigma \) are the mean and standard deviations of the probability distribution, respectively.

Review: The \( \mu \) and \( \sigma \) are parameters and hence describe the population.
The shape of a normal distribution is fully characterized by its mean $\mu$ and standard deviation $\sigma$.

- $\mu$ specifies the location of the distribution.
- $\sigma$ specifies the spread/shape of the distribution.

(a) Same means but different standard deviations
(b) Different means but same standard deviations

(c) Different means and different standard deviations
Properties of the Theoretical Normal Distribution

• A normal distribution curve is bell-shaped.
• The mean, median, and mode are equal and are located at the center of the distribution.
• A normal distribution curve is unimodal.
• The curve is symmetric about the mean.
• The curve is continuous (no gaps or holes).
• The curve never touches the x-axis (just approaches it).
• The total area under a normal distribution curve is equal to 1.0 or 100%.
• **Review:** The Empirical Rule applies (68%-95%-99.7%)
The Normal Distribution

- About 68% of the data lies within one standard deviation (σ) of the mean (μ).
- About 95% of the data lies within two standard deviations of the mean.
- About 99.7% of the data lies within three standard deviations of the mean.
The Standard Normal Distribution

Since each normally distributed variable has its own mean and standard deviation, the shape and location of these curves will vary.

To simplify making statistical inference using the normal distribution, we use the **standard normal distribution**.

**Standard normal distribution** - a normal distribution with mean 0 and a standard deviation of 1.
All normally distributed variables can be transformed into a standard normally distributed variable using the formula for the standard score (z-score):

\[ z = \frac{\text{value} - \text{mean}}{\text{standard deviation}} \quad \text{or} \quad z = \frac{X - \mu}{\sigma}. \]

The z-score for an observation is the number of standard deviations the observation lies from the mean.

The letter \( Z \) is used to denote a standard normal random variable.
Example: The mean resting pulse rates for men are normally distributed with a mean of 70 and a standard deviation of 8. A man has a pulse of 80. Find the corresponding z-score.

We know: \( \mu=70, \sigma=8, X=80 \)

\[
z = \frac{X - \mu}{\sigma} = \frac{80 - 70}{8} = \frac{10}{8} = 1.25
\]

The man’s pulse is \( 1.25 \) standard deviations above the mean because \( 1.25 \) is positive.
Example: The average speed of drivers on I-25 is 63 mph with a standard deviation of 5 mph. Assuming the data are normally distributed, find the z-score for the following speeds clocked by police officers:

74 mph \[ z = \]

51 mph \[ z = \]

82 mph \[ z = \]
Finding Areas/Probabilities for the Standard Normal Distribution

Table E in Appendix C gives the area/probability under the standard normal distribution curve to the left of \( z \)-values between -3.49 and 3.49 in increments of .01.

- To find \( P(Z \leq z) \) (the probability that a standard normal random variable \( Z \) is less than or equal to the value \( z \)), we match our value of \( z \) with the left column and top row of Table E (by adding the numbers together). The row and column where these numbers intersect gives us the probability or area under the standard normal curve to the left of \( z \).
- Note: Use .0001 when \( z \leq -3.50 \) and .9999 when \( z \geq 3.50 \).
Examples:

\[ P(Z \leq -3.49) = 0.0002 \]

\[ P(Z \leq -1.31) = 0.0951 \]

\[ P(Z \leq 0) = 0.5000 \]

\[ P(Z \leq 1.96) = 0.9750 \]

\[ P(Z \leq 3.49) = 0.9998 \]

\[ P(Z \leq 1.43) = \]

\[ P(Z \leq -1.82) = \]
**Finding the Area Under the Standard Normal Distribution Curve**

1. To the left of any $z$-value: look up the $z$ value in the table and use the area given.
Examples:

Find the area to the left of $z = 0.84$, i.e., find $P(Z \leq 0.84)$.

Find the area to the left of $z = -1.32$, i.e., find $P(Z \leq -1.32)$. 
2. To the right of any $z$-value: look up the $z$ value and subtract the area from 1.
Examples:

Find the area to the right of \( z = -0.21 \), i.e., find \( P(Z > -0.21) \).

\[
P(Z > -0.21) = 1 - P(Z \leq -0.21) = 1 - 0.4168 = 0.5832
\]

Find the area to the right of \( z = 1.52 \), i.e., find \( P(Z > 1.52) \).
3. Between any two $z$-values: look up both $z$ values and subtract the smaller area from the larger area.
Examples:

Find the area between \( z = -1.71 \) and \( z = 0.45 \), i.e., find

\[
P(-1.71 \leq Z \leq 0.45) = P(Z \leq 0.45) - P(Z \leq -1.71)
\]

\[
= 0.6736 - 0.0436
\]

\[
= 0.63
\]

Find the area between \( z = 1.43 \) and \( z = 3.01 \), i.e., find

\[
P(1.43 \leq Z \leq 3.01).
\]
Note: For a **continuous** random variable $X$ (not a discrete random variable), the probability that $X$ equals a specific number is 0.

- Let $c$ be some number (like 1 or 7). Then $P(X = c) = 0$.

Intuitive explanation:
- If we had ten numbers, 1, 2, ..., 10, and $X$ is the number we see, $P(X = 2) = 1/10$.
- If we had one thousand numbers, 1, 2, ..., 1000, and $X$ is the number we see, $P(X = 2) = 1/1000$.
- For a continuous random variable, there are an infinite number of possibilities, so $P(X = c) = 1/\infty = 0$. (This is bad mathematics, but it gets the point across).
Thus, for a continuous random variable:

\[
P(Z \leq z) = P(Z < z) \text{ and} \]

\[
P(Z \geq z) = P(Z > z) \text{ and} \]

\[
P(z_1 \leq Z \leq z_2) = P(z_1 < Z \leq z_2) = P(z_1 \leq Z < z_2) = P(z_1 < Z < z_2)\]
Section 6-2: Applications of the Normal Distribution

Finding Probabilities for Normal Distributions

To find probabilities for any normal distribution with mean $\mu$ and standard deviation $\sigma$, we simply transform the values on the original scale (call these $x$, $x_1$, or $x_2$) using the formula

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}} \quad \text{or} \quad z = \frac{x - \mu}{\sigma},$$

and then use the techniques from the previous section.
Example: An American household generates an average of 28 pounds per month of newspaper for garbage, with a standard deviation of 2 pounds. Assume the amount of newspaper garbage is normally distributed. If a household is selected at random, find the probability of generating between 27 and 31 pounds of newspaper garbage per month.

\[ X = \text{garbage}. \text{ We want } P(27 < X < 31). \]

Turn into a Z score so we can use the standard normal distribution.

\[
P\left(\frac{27 - \mu}{\sigma} < \frac{X - \mu}{\sigma} < \frac{31 - \mu}{\sigma}\right) = P\left(\frac{27 - 28}{2} < Z < \frac{31 - 28}{2}\right)
\]

\[
Z = \frac{X - \mu}{\sigma} \quad \text{and}
\]

\[
= P(-0.5 < Z < 1.5) = P(Z < 1.5) - P(Z < -0.5) = 0.9332 - 0.3085 = 0.6247
\]
Example: College women’s heights are normally distributed with $\mu = 65$ inches and $\sigma = 2.7$ inches. What is the probability that a randomly selected college woman is 62 inches or shorter?

$X=\text{height}$ and $Z = \frac{X - \mu}{\sigma} = \frac{X - 65}{2.7}$.

$$P(X \leq 62) = P\left( \frac{X - \mu}{\sigma} \leq \frac{62 - \mu}{\sigma} \right)$$

$$= P\left( Z \leq \frac{62 - 65}{2.7} \right)$$

We want $$= P(Z \leq -1.11)$$

$$= 0.1335$$
**Example:** What is the probability that a randomly selected college woman is between 63 and 69 inches tall?
Sometimes we need to find the value $X$ that corresponds to a certain percentile. (e.g., what value corresponds to the 98th percentile?)

To solve this problem:

1. Find the $z$ that corresponds to this percentile for the standard normal distribution.
   a. The area to the left of this $z$-score has the area closest to this percentile.

2. $X = \mu + z \cdot \sigma$. 
**Example:** Mensa is a society of high-IQ people whose members have a score on an IQ test at the 98\textsuperscript{th} percentile or higher. The mean IQ score is 100 with a standard deviation of 16. What IQ score do you need to get into Mensa?

1. First determine what z-score is related to the 98\textsuperscript{th} percentile (to be in the 98\textsuperscript{th} percentile means that 98\% of IQ test takers scored at or below your score) (i.e. 0.980).

   \[
   Z = 2.06
   \]

2. What is the IQ for the 98\textsuperscript{th} percentile using the formula for \( X \) given above?

   \[
   Z = \mu + Z \cdot \sigma = 100 + 2.06(16) = 132.96
   \]
**Example:** College women’s heights are normally distributed with a $\mu = 65$ and $\sigma = 2.7$. What is the 54th percentile of this population?

1. Find Z-score for 54th percentile:

2. Find corresponding value for X:
**Example**: An American household generates an average of 28 pounds per month of newspaper for garbage with a standard deviation of 2 pounds. Assume the amount of newspaper garbage is normally distributed. What is the 34\(^{th}\) percentile for the amount of newspaper garbage American households produce per month?
Determining Normality

The easiest way to determine if a distribution is bell-shaped is to draw a histogram for the data and check its shape. If the histogram is NOT approximately bell-shaped then the distribution is NOT normally distributed.

The skewness of a data set can be checked by using the Pearson’s index, PI, of skewness.

\[ PI = \frac{3(\bar{X} - \text{median})}{s} \]

If the index is greater than or equal to +1 or less than or equal to -1, it can be concluded that the data are significantly skewed.
If the index is smaller than -1, then the distribution is negatively skewed. If the index is greater than +1, then the distribution is positively skewed.

The data should also be checked for outliers since they can have a big effect on normality.

If the histogram is bell-shaped, the PI is between -1 and +1, and there are no outliers then it can be concluded that the distribution is normally distributed.
Example: A survey of 18 high-technology firms showed the number of days’ inventory they had on hand. Determine if the data are normally distributed.
For these data, the sample mean is 79.5, the median is 77.5, and the sample standard deviation is 40.5. Check for skewness using Pearson’s skewness index.

*Based on the PI, which is 0.148, the distribution for the data is not significantly skewed.*

\[
PI = \frac{3(\bar{X} - \text{median})}{s} = \frac{3(79.5 - 77.5)}{40.5} = 0.148
\]
**Example:** Consider a data set with a mean of 126.5, a median of 144 and a standard deviation of 38.65, find the PI and interpret it.

$$PI =$$
Section 6-3: The Central Limit Theorem

If we took numerous random samples and calculated the sample means, what distribution would the sample means have? This distribution is known as the sampling distribution of the sample means.

**Sampling distribution of sample means** – the distribution of the sample means calculated from all possible random samples of size $n$ from a population.
If the samples are randomly selected, the sample means will be somewhat different from the population mean $\mu$. These differences are due to sampling error.

**Sampling error** - the difference between the sample measure and the corresponding population measure due to the fact that the sample is not a perfect representation of the population.

Let’s look at a simulation through the Sampling Distribution (CLT) Experiment available at

http://www.socr.ucla.edu/htmls/SOCR_Experiments.html.

What do you notice about the sampling distribution for the mean?
Properties of the Distribution of Sample Means

When all possible samples of a specific size are selected with replacement from a population, the distribution of the sample means for a variable has three important properties:

1. The mean of the sample means will be the same as the population mean and is given by

   \[ \mu_{\bar{X}} = \mu. \]

2. The standard deviation of the sample means (known as the standard error of the mean) is given by

   \[ \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}. \]

Example: See pg 331 and 332 for a well-worked out example.
The Central Limit Theorem

As the sample size $n$ increases without limit, the shape of the distribution of the sample means taken with replacement from a population with mean $\mu$ and standard deviation $\sigma$ will approach a normal distribution. As previously shown, this distribution will have a mean $\mu$ and a standard deviation $\sigma/\sqrt{n}$.

**Reader’s Digest Version**: The distribution of the sample mean gets closer and closer to a normal distribution as the sample size increases. This distribution has a mean $\mu_{\bar{X}} = \mu$ and a standard deviation $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.

This result applies no matter what the shape of the probability distribution from which the samples are taken.
Some Important Points to Remember About the Central Limit Theorem:

1. *If the distribution of the population is normal*, then the distribution of the sample means will **ALWAYS** be normal.

2. If the sample size is large enough (typically \( n \geq 30 \)), then the distribution of the sample means will be **approximately normal**. The larger the sample, the better the approximation will be.
**Why is the Central Limit Theorem so useful?**

Even when we don’t know the distribution of the population, if the sample size is sufficiently large, then we can use the properties of the normal distribution to make statistical inference about the population mean! So essentially, if the sample size is large enough we can make statistical inference about the population mean even if we don’t know anything else about the population!

If the population size is really large, then the results about the sampling distribution of the mean are approximately correct even if sampling without replacement.
When the sample size is sufficiently large (30+), the central limit theorem can be used to answer questions about sample means in the same way that a normal distribution can be used to answer questions about individual data. However, our z-score formula changes slightly to

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

since the standard deviation of $\bar{X}$ is $\sigma / \sqrt{n}$. 
**Example:** The average yearly cost per household of owning a dog is $186.80 with a standard deviation of $32.00. Suppose that we randomly select 50 households that own a dog. What is the probability that the sample mean for these 50 households is less than $175.00?

We want to know $P(\bar{x} < 175)$ and have $\mu=186.80$, $\sigma=32$, $n=50$. To use our standard normal distribution, we need to convert to a $Z$-score:

$$P(\bar{x} < 175) = P\left( \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < \frac{175 - \mu}{\sigma/\sqrt{n}} \right) = P\left( Z < \frac{175 - 186.80}{32/\sqrt{50}} \right)$$

$$= P(Z < -2.61) = 0.0045$$
What is the probability that the sample mean for these 50 households is between $195 and $200?

What is the probability that the sample mean for these 50 households is between $180 and $190?
Tips:

1. Look for keywords about whether the data are normally distributed or the sample size is relatively large.
2. Note whether you are looking for a probability or a percentile.
3. The formula $z = (x - \mu) / \sigma$ is used to gain information about an individual data value when the variable is normally distributed.
4. The formula $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ is used to gain information when applying the central limit theorem about a sample mean when the population is normally distributed or when the sample size is 30 or more.