Chapter 7
Sampling Distributions and Point Estimation of Parameters

Part 1: Sampling Distributions, the Central Limit Theorem, Point Estimation & Estimators

Sections 7-1 to 7-2
A random sample is collected on a population to draw conclusions, or make statistical inferences, about the population.

**Definition (Random Sample)**

The random variables \( X_1, X_2, \ldots, X_n \) are a random sample of size \( n \) if...

1) the \( X_i \)'s are independent
2) every \( X \) has the same probability distribution

**Types of statistical inference:**

1. **Parameter estimation** (e.g. estimating \( \mu \)) with a confidence interval
   - For estimating \( \mu \), we collect data and we use the observed sample mean \( \bar{x} \) as a point estimate for \( \mu \) and create a confidence interval to report a likely range in which \( \mu \) lies.

2. **Hypothesis testing** about a population parameter (e.g. \( H_0 : \mu = 50 \))
   - We wish to compare the mean time that women and men spend at the CRWC. \( H_0 : \mu_M = \mu_W \) ? Or perhaps there is evidence against this hypothesis.
Sample Mean $\bar{X}$, a Point Estimate for $\mu$

- The sample mean $\bar{X}$ is used as a point estimate for the population parameter $\mu$. It is a point estimate because it is a single value.

- NOTATION: $\hat{\mu} = \bar{X}$ (a ‘hat’ over a parameter represents an estimator) $\bar{X}$ is the estimator here

- Prior to data collection, $\bar{X}$ is random variable and it is the statistic of interest calculated from the data when estimating $\mu$.

- The value we get for $\bar{X}$ (the sample mean) depends on the specific sample chosen!

- If $\bar{X}$ is random variable, then it has a certain expected value, variance, and distribution. The distribution of the random variable $\bar{X}$ is called the sampling distribution of $\bar{X}$.
Sample-to-Sample Variability

- As stated earlier, there is randomness in the $\bar{X}$ value we get from a random sample. Suppose I want to estimate a population mean height $\mu$ using a sample mean $\bar{X}$.
  - Suppose I randomly select 50 individuals from a population, measure their heights, and find the sample mean $\bar{x} = 5 \text{ foot 6 inches}$
  - Suppose I repeat the process, I again randomly select 50 individuals from a population, measure their heights, and find the sample mean $\bar{x} = 5 \text{ foot 8 inches}$
  - Suppose I repeat the process, I again randomly select 50 individuals from a population, measure their heights, and find the sample mean $\bar{x} = 5 \text{ foot 5 inches}$

- I didn’t do anything wrong in my data collection, this is just SAMPLING VARIABILITY!

[NOTE: In reality, we only take one sample. The above is meant to emphasize the existence of sample-to-sample variability.]
The Sampling Distribution of $\bar{X}$

**Definition (Sampling Distribution)**

The probability distribution of a statistic is called a **sampling distribution**.

- $\bar{X}$ is a statistic calculated from a random sample $X_1, X_2, \ldots, X_n$.

- $\bar{X}$ is a linear combination of random variables.

  $$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{1}{n} X_1 + \frac{1}{n} X_2 + \cdots + \frac{1}{n} X_n$$

- For a random sample $X_1, X_2, \ldots, X_n$ drawn from any distribution with $E(X_i) = \mu$ and $V(X_i) = \sigma^2$ or $X_i \sim ?(\mu, \sigma^2)$, we have

  $$E(\bar{X}) = \mu \quad \text{and} \quad V(\bar{X}) = \frac{\sigma^2}{n}$$

- But a mean and variance does not fully specify a distribution.
  Do we know the probability distribution of $\bar{X}$? ...
The Sampling Distribution of $\bar{X}$

- It turns out that $\bar{X}$ has some predictable behavior...

- If the $X_1, X_2, \ldots, X_n$ are drawn from a normal distribution, or by notation $X_i \sim N(\mu, \sigma^2)$ for all $i$, then

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

for any sample size $n$.

Example

Suppose IQ scores are normally distributed with mean $\mu = 100$ and variance $\sigma^2 = 256$. If $n = 9$ IQ scores are drawn at random from this population, what is the probability that the sample mean is less than 93?

**ANSWER:** Find $P(\bar{X} < 93)$ (next page).
Example

Suppose IQ scores are normally distributed with mean $\mu = 100$ and variance $\sigma^2 = 256$. If $n = 9$ IQ scores are drawn at random from this population, what is the probability that the sample mean is less than 98?

**ANSWER:** Find $P(\bar{X} < 93)$.

We first need a distribution for $\bar{X}$ (it follows a normal distribution!), and then we'll use it to create a $Z$ random variable and use the $Z$-table.
The Sampling Distribution of $\bar{X}$

- The graphic below shows how the variability in $\bar{X}$ decreases as $n$ increases. Recall $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$. 

![Diagram showing the sampling distribution of $\bar{X}$ with different sample sizes]
The Sampling Distribution of $\bar{X}$

Notation:

- $E(\bar{X}) = \mu_{\bar{X}} = E(X) = \mu$
- $V(\bar{X}) = \sigma^2_{\bar{X}} = \frac{V(X)}{n} = \frac{\sigma^2}{n}$

Terminology:

- The term **standard deviation** refers to the population standard deviation, or $\sqrt{V(X)} = \sigma$, and...

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

- The term **standard error** is a value related to $\bar{X}$ and is also more fully stated as the **standard error of the sample mean** and it is the square root of the variance of $\bar{X}$.

Std. Error of $\bar{X}$ is $\sqrt{V(\bar{X})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$

And then...

$$Z = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$
The Sampling Distribution of $\bar{X}$

- Even when $X_i$ are NOT drawn from a normal distribution, it turns out that $\bar{X}$ has some predictable behavior...

- If the $X_1, X_2, \ldots, X_n$ were NOT drawn from a normal distribution, or by notation $X_i \sim \mathcal{N}(\mu, \sigma^2)$ for all $i$, then $\bar{X}$ is approximately normally distributed as long as $n$ is large enough or

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

  for $n > 25$ or 30.

- Thus, $\bar{X}$ follows a normal distribution!!! (for a sufficiently large $n$)
  This is an incredibly useful result for calculating probabilities for $\bar{X}$!!
Example (Probability for $\bar{X}$, Flaws in a copper wire)

Let $X$ denote the number of flaws in a 1 inch length of copper wire. The probability mass function of $X$ is presented in the following table:

<table>
<thead>
<tr>
<th>$x$</th>
<th>$P(X = x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.48</td>
</tr>
<tr>
<td>1</td>
<td>0.39</td>
</tr>
<tr>
<td>2</td>
<td>0.12</td>
</tr>
<tr>
<td>3</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Suppose $n = 100$ wires are sampled from this population. What is the probability that the average number of flaws per wire in the sample is less than 0.5? (i.e. find $P(\bar{X} < 0.5)$... next page)
The Sampling Distribution of $\bar{X}$

Example (Probability for $\bar{X}$, Flaws in a copper wire)

ANSWER:  

$$P(\bar{X} < 0.5) = \frac{12}{26}$$
Central Limit Theorem (CLT)

Definition (Central Limit Theorem)
Let $X_1, X_2, \ldots, X_n$ be a random sample drawn from any population (or distribution) with mean $\mu$ and variance $\sigma^2$. If the sample size is *sufficiently large*, then $\bar{X}$ follows an approximate normal distribution.

We write: $\bar{X} \xrightarrow{d} N(\mu, \frac{\sigma^2}{n})$ as $n \to \infty$

Or: $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \xrightarrow{d} N(0, 1)$ as $n \to \infty$

If the random sample is drawn from a non-normal population, then $\bar{X}$ is approximately normal for sufficient large $n$ (at least 25 or 30) and the approximation gets better and better as $n$ increases.

**NOTE:** If the original ‘parent population’ from which the sample was drawn is normal, then $\bar{X}$ follows a normal distribution for any $n$ (a linear combination of normals is normal), and the CLT is not needed to achieve normality.
The Sampling Distribution of \( \bar{X} \) (simulation)

Let’s simulate this situation...

- **Case 1**: Original population is normally distributed

1. Choose a sample of size \( n \) from a normal distribution
2. Compute \( \bar{x} \)
3. Plot the \( \bar{x} \) on our frequency histogram
4. Do steps 1-3 many times, such as 1000 times
5. Draw a histogram of the 1000 \( \bar{x} \) values
   (to see the sampling distribution of \( \bar{X} \))

See applet at:

Case 1: Original population is normally distributed (with \( n=2 \))

The empirical distribution for \( \bar{X}_{n=2} \) is in the lower plot (in blue). Its mean is very close to the parent population mean \( \mu = 16 \), and its s.d. of 3.59 is very close to the theoretical \( \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{2}} = 3.54 \).
Case 1: Original population is normally distributed (with n=25)

The empirical distribution for $\bar{X}_{n=25}$ is in the lower plot (in blue). Its mean is very close to the parent population mean $\mu = 16$, and its s.d. of 1.0 is the same as the theoretical $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{25}} = 1$. 
RESULT - If the parent population (the one you are drawing from) is normal, then $\bar{X}$ will follow a normal distribution for any sample size $n$ with known mean and variance as shown below.

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$
The Sampling Distribution of $\bar{X}$ (simulation)

Let’s simulate this situation...

- Case 2: Original population is NOT normally distributed...

1. Choose a sample of size $n$ from a NON-normal distribution
2. Compute $\bar{x}$
3. Plot the $\bar{x}$ on our frequency histogram
4. Do steps 1-3 many time, such as 1000 times
5. Draw a histogram of the 1000 $\bar{x}$ values
   (to see the sampling distribution of $\bar{X}$)

See applet at:
Case 2: Original population is NOT normally distributed (with right-skewed parent population and $n=10$)

The empirical distribution for $\bar{X}_{n=10}$ is in the lower plot (in blue). Its bell-shaped with a mean equal to the parent population mean $\mu = 8.08$. Its s.d. of 1.96 is very close to the theoretical $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{6.22}{\sqrt{10}} = 1.97$. 
The Sampling Distribution of $\bar{X}$ (simulation)

- Case 2: Original population is NOT normally distributed (with very non-normal parent population and $n=2$)

FAIL!!!! The empirical distribution for $\bar{X}_{n=2}$ is in the lower plot (in blue) and it is not normally distributed. This is just too small of a sample size to overcome the very non-normal parent population.
The Sampling Distribution of $\bar{X}$ (simulation)

- Case 2: Original population is NOT normally distributed (with very non-normal parent population and $n=25$)

The empirical distribution for $\bar{X}_{n=25}$ is in the lower plot (in blue). Its bell-shaped with a mean close to the parent population mean $\mu = 16.92$. Its s.d. of 2.46 is very close to the theoretical $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{12.29}{\sqrt{25}} = 2.458$. 
RESULT - If the parent population (the one you are drawing from) is NOT normal, then $\bar{X}$ will follow an approximate normal distribution for sufficiently large $n$ (we’ll say $n > 25$ or $30$).

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

This is the Central Limit Theorem.
The approximation improves as $n$ increases.
The Sampling Distribution of $\bar{X}$

A couple comments:

- Averages are less variable than individual observations.

- The distribution for $\bar{X}$ has *less variability* than the distribution for $X$.

- The distribution of our estimator $\bar{X}_n$ is squeezed closer to, or is tighter, around the thing we’re trying to estimate as $n$ increases.

- For some non-normal distributions, the approximation is pretty good for $n$ lower than 25 or 30, so it depends on the parent population from which we are drawing.
The Sampling Distribution of $\overline{X}$

- The next graphic shows 3 different original populations (one nearly normal, two that are not), and the sampling distribution for $\overline{X}$ based on a sample of size $n = 5$ and size $n = 30$.

- The three original distributions are on the far left (one that is nearly symmetric and bell-shaped, one that is right skewed, and one that is highly right skewed).

- The graphic emphasizes the concept that the normal approximation becomes better as $n$ increases.
The Sampling Distribution of $\bar{X}$

As shown in: Navidi, W. 'Statistics for Engineers and Scientists', McGraw Hill, 2006
The Sampling Distribution of $\bar{X}$

- The variability of $\bar{X}$ decreases as $n$ increases
  Recall: $V(\bar{X}) = \frac{\sigma^2}{n}$.

- If the original population has a shape that’s closer to normal, smaller $n$ is sufficient for $\bar{X}$ to be normal.

- The normal approximation gets better with larger $n$ when you’re starting with a non-normal population.

- Even when $X$ has a very non-normal distribution, $\bar{X}$ still has a normal distribution with a large enough $n$. 