Lecture 5: Sampling Distribution

Readings: Sections 5.5, 5.6

Introduction

**Parameter**: describes population

**Statistic**: describes the sample; sampling variability

**Sampling distribution** of a statistic:

- A probability distribution that characterize the sampling variability
- The distribution of values taken by a sample statistic, e.g., sample mean, sample proportion, across all possible samples of the same size from the population
- Imagine taking repeated samples of size $n$ from the population and calculating sample statistic for each of them.

**Example 1**: The lifetime of a certain battery follows an exponential distribution with a mean value of 10 hours, i.e. $f(x) = 0.1e^{-0.1x}(x > 0)$.  

<table>
<thead>
<tr>
<th>Sample</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Mean</td>
<td>$\bar{x}_1 = 8.2$</td>
<td>$\bar{x}_2 = 16.4$</td>
<td>$\bar{x}_3 = 7.1$</td>
<td>...</td>
</tr>
</tbody>
</table>

**Example 2**: To check what percent of the resistors manufactured by a certain firm have resistances that exceed 10 ohms ($p$), we take repeated random samples of 100 resistors and compute the proportion of resistors with resistances exceeding 10 ohms ($\hat{p}$).

<table>
<thead>
<tr>
<th>Sample</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample proportion</td>
<td>$\hat{p}_1 = 0.03$</td>
<td>$\hat{p}_2 = 0.02$</td>
<td>$\hat{p}_3 = 0.05$</td>
<td>...</td>
</tr>
</tbody>
</table>

- How are these sample statistics distributed?

**General Properties of Sampling Distribution**

1. The sampling distribution of the statistic is centered at the population parameter estimated by the statistic.

2. The spread of the sampling distribution of the statistic decreases as the sample size increases

3. As the sample size increases, the sampling distribution becomes increasingly Normally distributed

**Sampling Distribution of a Sample Mean**

$$\text{Sample mean } \bar{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$$
– If the population distribution has mean $\mu$ and standard deviation $\sigma$, then the mean and standard deviation of $\bar{X}$ are

$$E(\bar{X}) = \mu_{\bar{X}} = \mu, \ SD(\bar{X}) = \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

– If the population distribution is a normal distribution, then the sampling distribution of $\bar{X}$ is also normal distribution for any sample size $n$

– What is the population distribution is not normal?

* **Central Limit Theorem**: The sampling distribution of $\bar{X}$ can be approximated by a normal distribution, $N(\mu_{\bar{X}} = \mu, \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}})$, when the sample size $n$ is sufficiently large, irrespective of the shape of the population distribution. The larger the sample size $n$, the better the approximation

![Graphs showing Central Limit Theorem](image)

Example 3: In a certain population of fish, the lengths of individual fish follow a normal distribution with mean 54mm and standard deviation 4.5mm.

a. What is the probability that a randomly chosen fish is between 51mm and 60mm long?

b. What is the probability that the mean length of the 4 randomly chosen fish is between 51mm and 60mm?
Example 4: The lifetime of a certain battery follows an exponential distribution with a mean value of 10 hours, i.e. \( f(x) = 0.1e^{-0.1x} (x > 0) \). You take a random sample of 100 batteries. What is the probability that the average lifetime would be between 9 and 11 hours?

Example 5: A surveying instrument makes an error of -2, -1, 0, 1, or 2 feet with equal probabilities when measuring the height of a 200-foot tower. Find the probability that, in 18 independent measurements of the tower, the average of the measurements is between 199 and 201 feet.
Sampling Distribution of a Sample Proportion

- Recall in the binomial experiment, $X =$ the number of successes in $n$ trials and $X \sim \text{binomial}(n, p)$, where $p$ is the population proportion of successes.
- Consider the sample proportion $\hat{p} = X/n$ as an estimate of the population proportion $p$.
- The sampling distribution of $\hat{p}$ has the same shape as the distribution of $X$, but on a non-integer scale since
  $$P(\hat{p} = p^*) = P(X/n = p^*) = P(X = np^*).$$

Example 6: The manufacturer admits that St. John’s Wort Capsules have 2% probability of being defective (containing the wrong amount of active ingredient). Select a random sample of 40 capsules.

- Let $Y$ be the number of defective capsules in the sample, $Y \sim \text{binomial}(n = 40, p = 0.02)$.
- Let $\hat{p} = Y/40$ be the fraction of defective capsules.

<table>
<thead>
<tr>
<th>$Y$</th>
<th>$\hat{p}$</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.000</td>
<td>0.446</td>
</tr>
<tr>
<td>1</td>
<td>0.025</td>
<td>0.364</td>
</tr>
<tr>
<td>2</td>
<td>0.050</td>
<td>0.145</td>
</tr>
<tr>
<td>3</td>
<td>0.075</td>
<td>0.037</td>
</tr>
<tr>
<td>4</td>
<td>0.100</td>
<td>0.007</td>
</tr>
<tr>
<td>$\geq 5$</td>
<td>0.125</td>
<td>$\approx 0.001$</td>
</tr>
</tbody>
</table>
a. What is the probability that \( \hat{p} \) is within 0.02 of the true value?

b. If we select a random sample of 100 capsules, what is the probability that \( \hat{p} \) is within 0.02 of the true value?

- We have learned that binomial random variables can be approximated using normal variables:
  - If \( X \sim \text{Binomial}(n, p) \), then
    
    \[
    X \approx N(\mu = np, \sigma = \sqrt{np(1-p)})
    \]
    
    when \( n \) is large (\( np > 5 \) and \( n(1-p) > 5 \)).

- What about \( \hat{p} \) when \( n \) is large?

- The mean and standard deviation of the sampling distribution of \( \hat{p} \) are
  
  \[
  \mu_{\hat{p}} = p, \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}
  \]

- For a sufficiently large sample size, the sampling distribution of \( \hat{p} \) is approximately normal. That is,

  \[
  \hat{p} \approx N(\mu_{\hat{p}} = p, \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}})
  \]

  - Rule of thumb: use the normal approximation when \( np > 5 \) and \( n(1-p) > 5 \).
  - **Continuity correction**: add or subtract \( \frac{0.5}{n} \) to improve accuracy.

- Note that the mean and standard deviation of \( \hat{p} \) depend on the true population proportion \( p \), which we often don’t know.
  - Use a value based on theory
  - Use \( p = 0.5 \) in the formula
Example 7: 8% of the population are left-handed. If 100 people are randomly selected, what is the probability that less than 5% of the people sampled will be left-handed?
Example 8: Assume that 10% of a certain product manufactured by a firm is defective. How large a sample is needed to be at least 80% certain that the proportion of defective products is between 7% and 13%?