CHAPTERS 4-6: Sampling Distributions

Read sections 4.1, 4.2.3, 4.4, 5.1.1, 6.1.1

POPULATION vs. SAMPLE:

Population: The entire group of individuals that we want information about.

Sample: A part of the population from which data is collected.

PARAMETER vs. STATISTIC:

Parameer: A numerical value calculated from a population of individuals.
Statistic: A numerical value calculated from a sample of individuals.

Recall the table of population parameters and the statistics that estimate them:

Statistics	Parameters
$\bar{x}$	μ
$\tilde{x}$	$\tilde{\mu}$
s	σ
s^2	σ^2
$\hat{p}$	p

SAMPLING DISTRIBUTION:
The value of a statistic varies from sample-to-sample. In other words, different samples will result in different values of a statistic. **Since the value of a statistic varies from sample-to sample, it is a variable! Therefore, it has a distribution!** The distribution of a statistic is called a **Sampling Distribution.**

How to Construct a Sampling Distribution (conceptually - this cannot be done in practice):

- Take all possible samples of size n from the population.
- Compute the value of the statistic for each sample.
- Display the sampling distribution of the statistic as a table, graph, or equation.

Sampling Variability: The sampling distribution of a statistic has a center and spread. The spread of the sampling distribution is called the **sampling variability.**

Sampling Distribution of $\overline{X}$

- $\mu_{\overline{X}} = \mu$ The sampling distribution for $\overline{X}$ is **always** centered at the same place as the X distribution (population distribution).
- If the data is from a SRS, then $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$ and $\sigma^2_{\overline{X}} = \dfrac{\sigma^2}{n}$.
  - Recall that **sampling without replacement** is valid only if $.05N \geq n$ (i.e., if the population is large) (this "rule of thumb" is more conservative than on p. 179 of your text).
  - If **sampling without replacement** and $.05N < n$ (i.e., if the population is small), then $\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)$. The multiplier $\left(1 - \frac{n-1}{N-1}\right)$ is called a **finite correction factor**.
- When the data are normal, then $\overline{X}$ is normal: $X \sim N(\mu, \sigma)$, then $\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$.

  Example: See Figure 4.20 on page 195 of your text

- **CENTRAL LIMIT THEOREM (CLT)**: For large SRS, $\overline{X} \dot{\sim} \mathrm{N}(\mu, \frac{\sigma}{\sqrt{n}})$
  - Mathematically, the CLT says that $\lim_{n\to\infty}$ distribution $\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right) = N(0,1)$
  - CLT holds even when the data X is NOT normal!
  - For data that is "not far from normality" (i.e. fairly symmetric and uni-modal), then a "large" SRS means $n \geq 15$. The more the population distribution (X dist.) differs from the normal distribution, the larger the sample size needed for normality of $\overline{X}$ (this "rule of thumb" is more liberal than on p. 178 of your text).
  - Even for data that are "far from normal" (i.e. severely skewed), samples with $n \geq 30$ are "large" since the sampling distribution of $\overline{X}$ in this case is approximately normal for essentially any population distribution (this "rule of thumb" is more liberal than on p. 178 of your text).

  CLT Example: See Figure 4.20 on page 195

NOTES:

- μ, σ, and σ^2 (without a subscript) represent μ_X, σ_X, and σ_X^2 (respectively).
- To use the normal tables, you need to **standardize** $\overline{X}$:
  - If $\overline{X} \sim \text{N}(\mu, \frac{\sigma}{\sqrt{n}})$, then $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ where Z $\sim$ N(0, 1).
  - If $\overline{X} \dot{\sim} \text{N}(\mu, \frac{\sigma}{\sqrt{n}})$, then $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ where Z $\dot{\sim}$ N(0, 1).

EXAMPLE: Suppose that X, the service lifetime (in miles) of disc brake pads, is normally distributed, $X \sim N(55000,\ 4500)$. Suppose that a SRS of size $n = 8$ is taken.

- Are there enough data to determine the sampling distribution of $\overline{X}$? If so, what's the sampling distribution?

- Find $P(\overline{X} \leq 51800)$

Sampling Distribution of $\hat{p}$

For binary data, each measurement X_i is denoted as either a 1 (has property of interest) or a 0 (does not have property of interest).

$$\hat{p} = \text{sample proportion of successes} = \frac{\#\text{ of } 1'\text{s}}{n}$$

- $\mu_{\hat{p}} = p$
- If the data is from a SRS, then $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$ and $\sigma^2_{\hat{p}} = \dfrac{p(1-p)}{n}$. Recall that for finite populations, $.05N \geq n$.
- Since $\hat{p}$ is simply a mean, then if we have a large SRS, CLT implies that $\hat{p} \dot{\sim} \text{N}\left(p,\ \sqrt{\frac{p(1-p)}{n}}\right)$
  - $\hat{p}$ never has an *exact* normal distribution because the population distribution is binary and therefore not normal. We completely depend on the CLT to obtain approximate normality of the $\hat{p}$ distribution.
  - Since the data is binary, then the criterion for "large" is different than when the data is continuous. We will consider a sample of binary data large enough so that $\hat{p}$ is normal if
    - $np \geq 10$
    - $n(1-p) \geq 10$.

    When p is too close to either 0 or 1, the sampling distribution of $\hat{p}$ is highly skewed.

If $\hat{p} \dot{\sim} N(p,\ \sqrt{\frac{p(1-p)}{n}})$, then standardize by setting $Z = \dfrac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$ where $Z \dot{\sim} N(0,\ 1)$.

EXAMPLE: Suppose that 20% of families in a certain region of the U.S. live in poverty. A SRS of n=400 families is taken and $\hat{p}$ is calculated from this sample.

- Is the sample large enough to be certain that the sampling distribution of $\hat{p}$ is normal?

- What is the approximate sampling distribution of $\hat{p}$?

- Find $P(\hat{p} > 0.25)$.

Exercises

Sampling distribution for $\bar{x}$ on p203: 4.1, 4.3, 4.5 (interpret "point estimate" as "estimate")
Central Limit Theorem on p214: 4.33, 4.34, 4.35 - 4.41 (odd)
Sampling distribution for $\hat{p}$ on p312: 6.1, 6.3, 6.13a