### 7.1 What is a Sampling Distribution?

**HW:** page 428, #1-8 all, 9, 13, 15, 19, 21-26 all

**Objectives:**
- Identify parameters and statistics in a sample or experiment
- Recognize the fact of sampling variability: a statistic will take different values when you repeat a sample or experiment.
- Understand that the variability of a statistic is controlled by the size of the sample. Statistics from larger samples are less variable

We have to differentiate between a number that describes the sample or a population.

- a number that describes the population. A parameter is a fixed number, but in practice we don’t know its value because we cannot examine the entire population.

\[ p = \text{population proportion} \quad \mu = \text{mean of population} \]

- a number that describes a sample and it can change from sample to sample. We use a statistic to estimate an unknown parameter.

\[ \hat{p} = \text{sample proportion (an estimate of the unknown parameter } p) \]
\[ \bar{x} = \text{mean of sample (an estimate of the mean } \mu \text{ of the population)} \]

**sampling distribution**

The **sampling distribution** is the distribution of values taken by the statistic in all possible samples of the same size from the same population.

**Sampling Variability**

**Sampling Variability** is the natural tendency of randomly drawn samples to differ, one from another. Sampling variability is not an error, just the natural result of random sampling. Statistics attempts to minimize, control, and understand variability so that informed decisions can be made from data despite their variation.

**To decrease spread, increase the # of trials.** Larger samples give smaller spread.

A statistic as an estimator of a parameter may suffer from bias or from high variability. **Bias** means that the center of the sampling distribution is not equal to the true mean.

---

**Statistics helps us to make better decisions by allowing us to assess risk (How serious are the dangers of Hormone Replacement Therapy?), to make predictions (How long will it take those killer African bees to reach the United States?) and to determine how certain things are related (Are short people more prone to heart attacks?).**

In order to answer questions such as these, statisticians collect data from samples. For each sample collected, the results could be different. So how does a statistician arrive at a “good conclusion” if the results are affected by the sample selected? The reasoning of statistical inference rests on asking “How often would this method give a correct answer if I used it very many times?” The purpose of this chapter is to prepare us for the study of statistical inference by looking at the probability distributions of some very common statistics: sample proportions and sample means.
A statistic used to estimate a parameter is **unbiased** if the mean of its sampling distribution is equal to the true value of the parameter being estimated.

\[ \bar{x} = \mu \quad \text{or} \quad \hat{p} = p \]

The **variability of a statistic** is described by the spread of its sampling distribution. This spread is determined primarily by the size of the random sample. Larger samples give smaller spread. The spread of the sampling distribution does not depend on the size of the population, as long as the population is at least 10 times larger than the sample.

**See figure 9.9 page 500 (Bulls Eye)**

- **(a)** High bias, low variability
- **(b)** Low bias, high variability
- **(c)** High bias, high variability
- **(d)** The ideal: low bias, low variability
What proportion of U.S. teens know that 1492 was the year in which Columbus “discovered” America? A Gallup Poll found that 210 out of a random sample of 501 American teens aged 13 to 17 knew this historically important date.

Objectives:
- Recognize when a problem involves a sample proportion \( \hat{p} \).
- Find the mean and standard deviation of the sampling distribution of a sample proportion \( \hat{p} \) for an SRS of size \( n \) from a population having proportion \( p \) of successes.
- Check whether the 10% and Normal conditions are met in a given setting.
- Use Normal approximation to calculate probabilities involving \( \hat{p} \).
- Use the sampling distribution of \( \hat{p} \) to evaluate a claim about a population proportion.

The sample proportion \( \hat{p} = \frac{210}{501} = .42 \) is the statistic that we use to gain information about the unknown population parameter \( p \).

Sample proportions are most often used when we are interested in categorical variables—the proportion of US adults that watch Lost, percent of the adult population that attended church last week.

Shape: As \( n \) increases, the sampling distribution of \( \hat{p} \) becomes approximately Normal. Before you perform Normal calculations, check that the Normal condition is met:

\[ np \geq 10 \quad \text{and} \quad n(1 - p) \geq 10. \]

Center: \( \mu_{\hat{p}} = p \)

Spread: \( \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \), provided the 10% rule is met.

In Chapter 6, we learned that the mean and standard deviation of a binomial random variable \( X \) are

\[ \mu_X = np \quad \quad \sigma_X = \sqrt{np(1-p)} \]

Since \( \hat{p} = \frac{X}{n} = (1/n) \cdot X \), we are just multiplying the random variable \( X \) by a constant \( (1/n) \) to get the random variable \( \hat{p} \). Therefore,

\[ \mu_{\hat{p}} = \frac{1}{n} (np) = p \quad \quad \sigma_{\hat{p}} = \frac{1}{n} \sqrt{np(1-p)} = \sqrt{\frac{np(1-p)}{n^2}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}} \]

\( \hat{p} \) is an unbiased estimator of \( p \) As the sample size increases, the spread decreases.
Memory and Standard Deviation of a Sample Proportion:
Choose an SRS of size n from a large population with population proportion p having some characteristic of interest. Let \( \hat{p} \) be the proportion of the sample having that characteristic.

Mean: \( \mu_{\hat{p}} = p \)

Standard Deviation: \( \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \)

Formula: \( z = \frac{\hat{p} - p}{\sigma_{\hat{p}}} \) or \( z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \)

Calculator: normalcdf (min, max, \( \mu_{\hat{p}}, \sigma_{\hat{p}} \))

Conditions:

1.) 10% Rule – use \( \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \) only when the population is at least 10 times as large as the sample (10 sample \( \leq \) Pop).

2.) Normal Condition – use the normal approximation \( z = \frac{\hat{p} - p}{\sigma_{\hat{p}}} \) when \( np \geq 10 \) and \( n(1-p) \geq 10 \).

State:
- State what you are looking for.

Plan:
- State parameters and sampling distribution
- Verify Conditions
- Make a picture

Do:
- If conditions are met carry out procedure. Find z-score and the probability

Conclude:
- State your conclusion in the context of the problem.

Example: (4 steps)

a. Records show that 65% of visitors will spend money in the gift shop at a local attraction. If there are 525 visitors on a given day, what is the probability that at least 70% of these visitors will purchase something in the gift shop?

b. A polling organization asks an SRS of 1500 first-year college students how far away their home is. Suppose that 35% of all first-year students actually attend college within 50 miles of home. What is the probability that the random sample of 1500 students will give a result within 2 percentage points of this true value?
7.3 Sample Means and Central Limit Theorem

HW: page 454, #49-63 odd, 64-72 all

Objectives:
- Recognize when a problem involves the mean $\bar{x}$ of a sample.
- Find the mean and standard deviation of the sampling distribution of a sample mean $\bar{x}$ from an SRS of size $n$ when the mean $\mu$ and standard deviation $\sigma$ of the population are known.
- Calculate probabilities involving a sample mean $\bar{x}$ when the population distribution is Normal.
- Understand that $\bar{x}$ has approximately a normal distribution when the sample is large (central limit theorem). Use the normal distribution to calculate probabilities that concern $\bar{x}$.

Sample means are often used when we record quantitative variables — the income of a household, the lifetime of a car’s brake pad, the blood pressure of a patient.

Shape:
- If the population distribution is Normal, then so is the sampling distribution of $\bar{x}$.
- If the population distribution is not Normal, the central limit theorem tells us that the sampling distribution of $\bar{x}$ will be approximately Normal in most cases if $n \geq 30$.

Center: $\mu_{\bar{x}} = \mu$

Spread: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, provided the 10% rule is met.

The Central Limit Theorem
(The Fundamental Theorem of Statistics)
As the sample size, $n$, increases, the mean of $n$ independent values has a sampling distribution that tends toward a Normal model with mean, $\mu$, and standard deviation,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$  

The Central Limit Theorem says that means of repeated samples will tend to follow a Normal Model if the sample size is “large enough”. This is true no matter what shape the population distribution has. **We need a sample of at least 30 for CLT to apply.**
### Mean and Standard Deviation of a Sample Mean

Suppose that $\bar{x}$ is the mean of an SRS of size $n$ drawn from a large population with mean $\mu$ and standard deviation $\sigma$.

**Mean:** $\mu_{\bar{x}} = \mu$

**Standard Deviation is** $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

**Formula:** $z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \iff z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

**Calculator:** normalcdf (min, max, $\mu_{\bar{x}}, \sigma_{\bar{x}}$)

**Conditions:**

1.) **Normal:**
   - Population distribution is Normal OR large sample, $n \geq 30$ (CLT).

2.) **10% Rule** – use $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ only when the population is at least 10 times as large as the sample ($10 \cdot \text{sample} \leq \text{Population}$).

- Averages are less variable than individual observations
- Averages are more normal than individual observations
- The sample mean, $\bar{x}$, is an unbiased estimator of the population mean $\mu$
- The values of $\bar{x}$ are less spread out for larger samples. Their standard deviation decreases at the rate $\sqrt{n}$, for example, you must take a sample four times as large to cut the standard deviation of $\bar{x}$ in half.
- You should use the formula $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ for the standard deviation of $\bar{x}$ only when the population is at least 10 times as large as the sample.

---

**Remember the 4 Step Method**

**State:**
- State what you are looking for.

**Plan:**
- State parameters and sampling distribution
- Verify Conditions
- Make a picture

**Do:**
- If conditions are met carry out procedure. Find $z$-score and the probability

**Conclude:**
- State your conclusion in the context of the problem.

---

**Example:**

*Of 500 people attending an international convention, it was determined that the average distance traveled by the conventioneers was 1917 miles with a standard deviation of 2500 miles. What is the probability that a random sample of 40 of the attendees would have traveled an average distance of 900 miles or less to attend this convention?*