Chapter 7: The Central Limit Theorem

Introduction

The central limit theorem (CLT for short) is one of the most powerful and useful ideas in all of statistics. There are two alternative forms of the theorem, and both alternatives are concerned with drawing finite samples size n from a population with a known mean, μ, and a known standard deviation, σ.

- The first alternative says that if we collect samples of size n with a "large enough n," calculate each sample's mean, and create a histogram of those means, then the resulting histogram will tend to have an approximate normal bell shape.

- The second alternative says that if we again collect samples of size n that are "large enough," calculate the sum of each sample and create a histogram, then the resulting histogram will again tend to have a normal bell-shape.

The size of the sample, n, that is required in order to be "large enough" depends on the original population from which the samples are drawn (the sample size should be at least 30 or the data should come from a normal distribution). If the original population is far from normal, then more observations are needed for the sample means or sums to be normal. Sampling is done with replacement.

It would be difficult to overstate the importance of the central limit theorem in statistical theory. Knowing that data, even if its distribution is not normal, behaves in a predictable way is a powerful tool.

7.1: The Central Limit Theorem for Sample Means (Averages)

Recall:

- **Population** - the complete collection of *all* individuals to be studied.
- **Sample** - a subcollection of members selected from a population.
- **Parameter** - a numerical measurement describing some characteristic of a *population*.
- **Statistic** - a numerical measurement describing some characteristic of a *sample*.
- **Sampling variability** - values of sample statistics vary from sample to sample.

Recall our notation:

	(Population) Parameter		(Sample) Statistic	
Proportion	p		$\hat{p}$	"p-hat"
Mean	μ	"mu"	$\bar{x}$	"x-bar"
Standard Deviation	σ	"sigma"	s	

Sampling Distribution of a statistic - the distribution of all values of the statistic when all possible samples of the same size n are taken from the same population.

Means

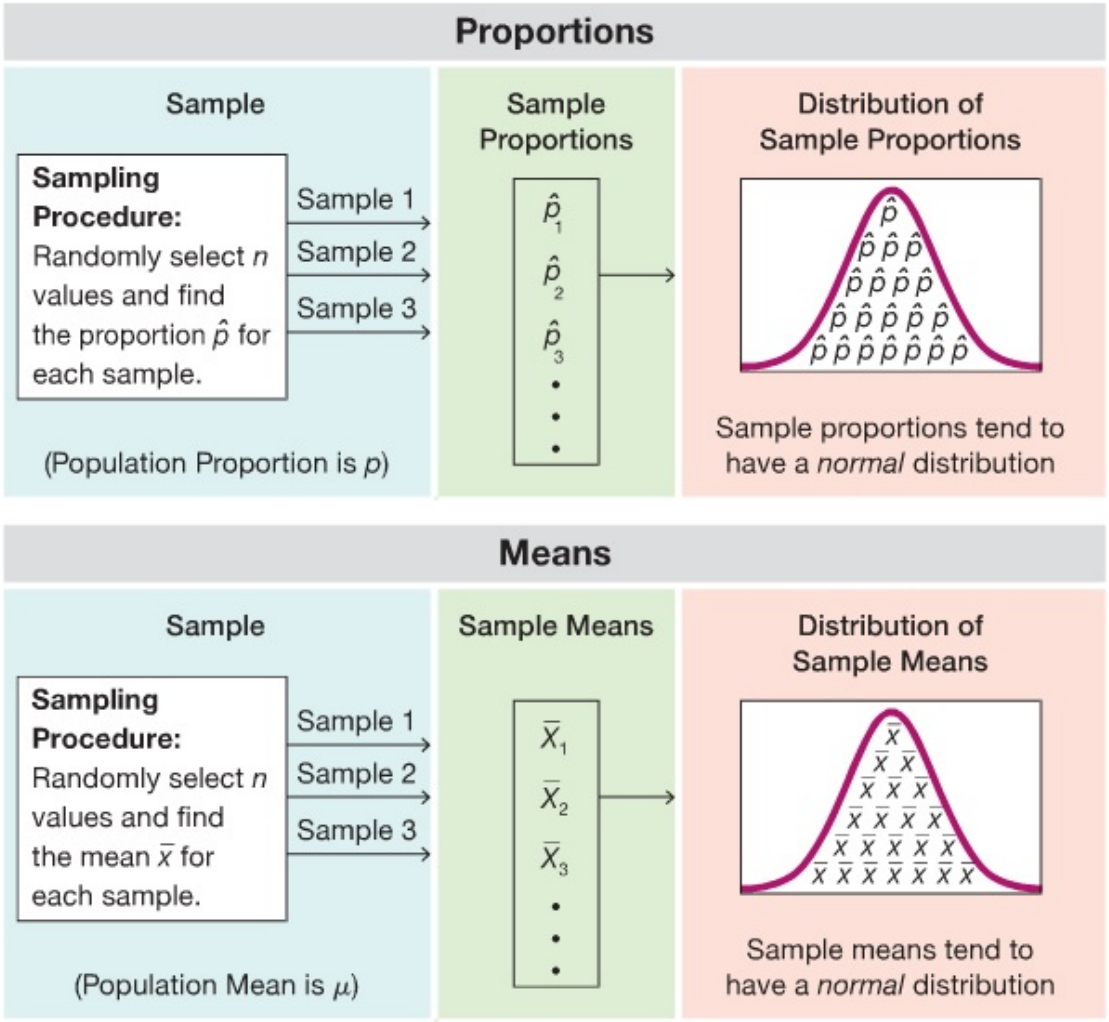
Sampling distribution of the sample mean - the distribution of all possible sample means with all samples having the same sample size n taken from the same population.

Proportions

The parameter of interest is the **population proportion,** p.

In each case, we can compute the statistic $\hat{p}$ - the **sample proportion** $= \frac{x}{n}$, the proportion in the sample with responses in the specified category.

Sampling distribution of the sample proportion - the distribution of all possible sample proportions with all samples having the same sample size n taken from the same population.



The Central Limit Theorem and the Sampling Distribution of $\bar{x}$

Suppose a simple random sample of size n is to be taken from a large population in which the variable of interest has mean μ and standard deviation σ. Then the sampling distribution of the sample mean $\bar{x}$ will have the following properties:

- *Shape*: (approximately) normal
- *Center (mean)*: $\mu_{\bar{x}} = \mu$
- *Spread (standard deviation)*: $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

If you draw random samples of size n, then as n increases, the random variable $\overline{X}$ which consists of sample means, tends to be normally distributed and

$$\overline{X} \sim \text{ normal } (\mu_{\bar{x}}, \sigma_{\bar{x}}).$$

$$\overline{X} \sim \text{ normal } \left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

The random variable $\overline{X}$ has a different z-score associated with it from that of the random variable X.

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

Conditions for Normality

1. Representative Sample
2. One of the following:
   (a) The population must be normally distributed, OR
   (b) The sample size needs to be *large* enough, $n \geq 30$

The Central Limit Theorem and the Sampling Distribution of $\hat{p}$

Suppose a simple random sample of size n is to be taken from a large population in which the true population possessing the attribute of interest is p. Then we can predict three things about the sampling distribution of the sample proportion $\hat{p}$:

- *Shape*: (approximately) normal
- *Center (mean)*: $\mu_{\hat{p}} = p$
- *Spread (standard deviation)*: $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$

If you draw random samples of size n, then as n increases, the random variable $\hat{p}$ which consists of sample proportions, tends to be normally distributed and

$$\hat{p} \sim \text{ normal } (\mu_{\hat{p}}, \sigma_{\hat{p}}) .$$

$$\hat{p} \sim \text{ normal } \quad p, \sqrt{\frac{p(1-p)}{n}}\Bigg)\Bigg($$

The random variable $\hat{p}$ has the following z-score associated with it.

$$z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}.$$

Conditions for Normality

1. Representative sample
2. X, the number of successes, follows a binomial distribution
3. Both np and nq are at least 5 (at least 5 successes and at least 5 failures)

Sampling Distribution		Categorical Variable	Quantitative Variable
	Sample Statistic	$\hat{p}$ (sample proportion)	$\bar{x}$ (sample mean)
	Mean	$\mu_{\hat{p}} = p$ (population proportion)	$\mu_{\bar{x}} = \mu$ (population mean)
	Standard Deviation	$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
	Shape for large sample size(s)	Approximately normal	Normal or approximately normal
	When is n large?	When $np \geq 5$ and $n(1-p) \geq 5$	When population is normal or $n \geq 30$
	z-score	$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$	$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

Notes:

- It's IMPORTANT to distinguish clearly between parameters and statistics.
  - A parameter is associated with a population. We typically do not know the parameter value in real life, though we may perform calculations assuming a particular parameter value.
  - A statistic is associated with a sample. This varies from sample to sample.

- Notice the Central Limit Theorem specifies *three* things about the distribution of a sample mean: shape, center (mean), and spread (standard deviation).

- The first step in any CLT problem is to identify which version of the result to use. Determining this involves asking whether the question is about a sample proportion or a sample mean.
  - Often the question itself will use the word *proportion* or *mean*, though not always.

7.3: Using the Central Limit Theorem

It is important for you to understand when to use the central limit theorem. If you are being asked to find the probability of the MEAN, use the CLT for the mean. If you are being asked to find the probability of a PROPORTION, use the CLT for proportions.

If you are being asked to find the probability of an INDIVIDUAL VALUE, do not use the CLT. Use the distribution of its random variable.

Sampling for a Long, Long Time: The Law of Large Numbers

The sampling distribution for sample means tells us how much variability in possible sample means to expect for different sample sizes. There is a much simpler technical result called the **Law of Large Numbers**, which guarantees that the sample mean $\bar{x}$ will eventually get "close" to the population mean μ, *no matter how small a difference* you use to define *close*.

The larger n gets, the smaller the standard deviation gets.

In practice, the Law of Large Numbers says that for any specific population, the larger the sample size, the more you can count on $\bar{x}$ to be an accurate representation of μ.

Example 1. Assume that cans of Dr. Pepper are filled so that the actual amounts have a mean of 12.00 oz and a standard deviation of 1.5 oz. Find the probability that a sample of 36 cans will have a mean amount of at least 12.35 oz.

Example 2. Suppose the population proportion of people who never wear a seatbelt is 30%. In a random sample of 200 drivers, what is the probability that less than 50 individuals say that they never wear a seatbelt when driving.

Example 3. The Centers for Disease Control and Prevention reported that 20.9% of American adults smoked regularly in 2004. Treat this as the parameter value for the current population of American adults. $n = 100$

(a) What symbol represents the population proportion, 0.209?

(b) Describe the sampling disribution of a sample 100 American adults.

(c) Calculate the probability that the sample proportion who smoke will exceed 0.25.

Example 4. A simple random sample of size $n = 49$ is obtained from a population with $\mu = 80$ and $\sigma = 14$.

(a) What is $P(x > 83)$?

(b) What is $P(\bar{x} \leq 75.8)$?

(c) What is $P(78.3 \leq \bar{x} \leq 85.1)$?

Example 5. A blind taste test is done to compare Cola 1 and Cola 2. Among 75 participants surveyed, 60% said they prefer Cola 1. Assuming a 50% chance of a participant choosing Cola 1, what is the probability the proportion observed in a taste test is less than 60%?

Example 6. Based on tests of the Chevrolet Cobalt, engineers have found that the miles per gallon in highway driving are normally distributed, with a mean of 32 miles per gallon and a standard deviation 3.5 miles per gallon.

(a) What is the probability that a randomly selected Cobalt gets more than 34 miles per gallon?

(b) Ten Cobalts are randomly selected and the miles per gallon for each car are recorded. What is the probability that the mean miles per gallon exceeds 34 miles per gallon?