All random variables have probability distributions, and as statistics are random variables, they too have distributions. The random phenomenon that produces the statistics of interest in this chapter, \( \hat{p} \) and \( \overline{y} \) is random sampling of a population. Consequently, the distribution of the statistic (\( \hat{p} \) or \( \overline{y} \)) is called the *sampling distribution*.

While the concepts and theory behind sampling distributions are sometimes difficult, the sampling distributions of both statistics usually can be approximated by normal distributions.

*Example:* Gallup conducted a poll of 2,475 likely voters between October 30th and November 1st, 2008 (a few days before the 2008 Presidential election). 50.6% of likely voters said that they would vote for Barack Obama. Assume for the moment that it is November 2nd, 2008.

- Let \( p \) denote the proportion of all voters that will vote for Obama.
- The statistic (an estimator of \( p \)) is \( \hat{p} \) and it’s realization was
  \[
  \hat{p} = \frac{1253}{2475} = 0.5063.
  \]
- It’s highly unlikely that \( \hat{p} = p \), though the difference \( \hat{p} - p \) is likely to be small if the sample size is large.

A measure of error is essential to properly interpret the estimate; most polling organizations report the *margin of error*.

It’s impossible to know the error before the election results are tabulated and \( p \) is computed. In fact, it’s extremely rare to ever know the parameter. Consequently, computing \( \hat{p} - p \) as a measure of error is infeasible and some other approach is needed to characterize the accuracy (or error) of a statistic.

One characterization of the error is the standard deviation of the sampling distribution of \( \hat{p} \). In this case, the standard deviation is the average difference between \( p \) and a value of \( \hat{p} \) obtained from random sampling. The standard deviation of \( \hat{p} \) can be estimated from the sample.

The characterization of error using the standard deviation can be criticized because the average error is used and the realized value is but *one* of many possible realizations (each depending on the drawn sample), and the observed realization is highly unlikely to be exactly one standard deviation from \( p \).
Specifically, the estimated expected error is reported. The expected error is the average error if the method were to be used many times (the method being: collect a random sample of size $n$ and compute $\hat{p}$ many times.)

**Definition:** The distribution of $\hat{p}$ across all possible samples of a given size is called the *sampling distribution* of $\hat{p}$.

- The standard deviation of the sampling distribution is the expected (or average) error of the estimator.
- Statistical theory provides a technique for using a single sample to approximate the sampling distribution.

First, recall that the distribution of a discrete random variable consists of 1. the possible realizations of the random variable and, 2. the associated probabilities of each realization. The distribution can be approximated by repeatedly random sampling the population and computing the statistic. The relative frequencies of occurrence of each realized value comprises the approximate sampling distribution.

*Simulating the sampling distribution of $\hat{p}$*

**Example:** Polling before an election is a good example for simulation since the true $p$ is usually known after the election. In the 2008 presidential election, 52.9% of Americans voted for Barack Obama. If $\hat{p}$ is computed from a random sample of $n$ voters drawn from the 131,257,328 voters, then we can simulate the distribution of $\hat{p}$ using the random number table.

- Randomly draw $n$ 3-digit numbers from 000, 001, 002, $\ldots$, 999. If number is in the set \{000, 001, 002, $\ldots$, 528\}, then the simulated vote was for Obama. If the number is not in the set, then the vote is for someone else.

A random number table was used to simulate the polling of $n = 6$ voters. The results from ten repetitions are shown below.
Table 1: Ten repetitions of sampling the 2008 presidential voters. \(\hat{p}\) is the proportion of voters for Obama. \(n = 6\) voters were randomly sampled by simulation.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Random numbers</th>
<th>(\hat{p})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>03939 30763 06138 80062</td>
<td>5/6 = .833</td>
</tr>
<tr>
<td>2</td>
<td>75998 37203 07959 32846</td>
<td>2/6 = .333</td>
</tr>
<tr>
<td>3</td>
<td>94435 97441 90998 25104</td>
<td>3/6 = .5</td>
</tr>
<tr>
<td>4</td>
<td>04362 40989 69167 38894</td>
<td>3/6 = .5</td>
</tr>
<tr>
<td>5</td>
<td>89059 43528 10547 40115</td>
<td>2/6 = .333</td>
</tr>
<tr>
<td>6</td>
<td>87736 04666 75145 49175</td>
<td>.667</td>
</tr>
<tr>
<td>7</td>
<td>76488 80199 15860 07323</td>
<td>.333</td>
</tr>
<tr>
<td>8</td>
<td>36460 53722 66634 25045</td>
<td>.667</td>
</tr>
<tr>
<td>9</td>
<td>13205 69237 21820 20952</td>
<td>.5</td>
</tr>
<tr>
<td>10</td>
<td>11282 43632 49531 78988</td>
<td>.5</td>
</tr>
</tbody>
</table>

- The histogram to the right shows the distribution. The mean value was \(\bar{p} = .515\) and the standard deviation was \(SD(\hat{p}) = .167\).

- Using R, thousands of random samples can be simulated for \(n = 6\) or any other size. Simulated sampling distributions of \(\hat{p}\) are shown for sample sizes of 6, 25, and 100 in the histograms below.
• The center of each distribution is $p = 0.529$.

• The standard deviations of the distributions decrease as $n$ increases.

• The distributions are appear to be approximately normal.

The sampling distribution of $\hat{p}$

Henceforth, it is assumed that $\hat{p}$ is computed from a random sample of size $n$ drawn from an infinitely large population with population proportion $p$.

1. The mean of the sampling distribution of $\hat{p}$ is

   $$E(\hat{p}) = p.$$  

   De Veaux et al. write $\mu(\hat{p}) = p$.

2. The standard deviation of the sampling distribution of $\hat{p}$ is

   $$\sigma(\hat{p}) = \text{SD}(\hat{p}) = \sqrt{pq/n}$$

   where $q = 1 - p$. 

135
3. If $n$ is sufficiently large, then the distribution of $\hat{p}$ is approximately $N(p, \sqrt{pq/n})$.

Shorthand notation:

$$\hat{p} \sim N(p, \sqrt{pq/n}).$$

If the population size is very large, then items 1-3 are approximately correct. Specifically,

- The expected value of $\hat{p}$ is $E(\hat{p}) = p$.

- The standard deviation of $\hat{p}$ is well-approximated by $\sqrt{pq/n}$ provided that the 10% condition is valid (the sample size is no more than 10% of the population size).\(^1\)

- The sampling distribution of $\hat{p}$ is approximately normal provided that the success/failure condition holds. The success/failure condition states that $np \geq 10$ and $nq \geq 10$.

**Example:** The sampling distribution of $\hat{p}$ computed from a sample of $n = 100$ voters is determined by noting that the true proportion of voters who voted for Barack Obama is $p = .529$. The 10% and success/failure conditions are satisfied since $n = 100$ is much less than the population of 119 million voters. Thus, the normal distribution approximation is appropriate, and

$$\hat{p} \sim N(\mu(\hat{p}), \sigma(\hat{p}))$$

where

$$\mu(\hat{p}) = p = .529,$$

and

$$\sigma(\hat{p}) = \sqrt{pq/n} = \sqrt{.529 \cdot .471/100} = .050.$$ 

- The model is used now to approximate the probability that a realized sample proportion will be less than .50. Since $\hat{p} \sim N(.53, .050)$,

$$P(\hat{p} < .5) = P \left( \frac{\hat{p} - p}{\sigma(\hat{p})} < \frac{.5 - .53}{.050} \right) = P(Z < -.6) = .2743.$$ 

**Sampling distribution of the sample mean:** The sample mean is the principal statistic for characterizing the distribution of a quantitative variable over a population.

- The population mean $\mu$ is estimated by the sample mean $\bar{y}$.

\(^1\)If the 10% condition is satisfied then the sample can be viewed as independent realizations of $n$ Bernoulli trials.
• Estimation error is characterized by the standard deviation of the sampling distribution of $\overline{y}$ as it is the expected difference between $\overline{y}$ and $\mu$.

• We’ll begin with a sampling experiment that generates approximate sampling distributions before developing the main theoretical results.

*Example:* The survival times (days) of 2843 Australian patients diagnosed with AIDS before 1 July 1991 are shown in a histogram to the right.²

1. These observations will be treated as a population which will be sampled to estimate its mean.

2. The population mean and standard deviation are $\mu = 405.93$ and $\sigma = 363.94$ days, respectively.

3. The distributions of the sample mean computed from SRS of sizes $n = 3, 5, 10, \text{ and } 25$ are to be investigated by taking 10,000 repeated samples of each size. Each sample yields a sample mean for a total of 10,000 $\overline{y}$’s.

4. Histograms of sample means for each sample size are shown below. The centers are all approximately 406 days. The spread decreases as $n$ increases, and the shapes of the distributions become more normal-like in appearance as $n$ increases.

²Source: Dr P. J. Solomon and the Australian National Centre in HIV Epidemiology and Clinical Research.
The Central Limit Theorem explains why the sampling distribution of the sample mean appears to be increasingly more normal as the sample size increases.

Central Limit Theorem: The sampling distribution of the sample mean computed from a simple random sample converges to a normal distribution as the sample size $n$ tends to infinity.

The Central Limit Theorem predicts the sampling distribution for (almost) all quantitative variables measured on any population. To which normal distribution is the sampling distribution of the mean converging?
1. The mean of the sampling distribution of $\bar{y}$ is the population mean $\mu$. Hence, $\mu(\bar{y}) = \mu$.

2. The standard deviation of the sampling distribution of $\bar{y}$ is: $\sigma/\sqrt{n}$. Hence, $\sigma(\bar{y}) = \text{SD}(\bar{y}) = \sigma/\sqrt{n}$.

In summary, as the sample size $n$ tends to infinity, the sampling distribution of $\bar{y}$ tends to a $N(\mu, \sigma/\sqrt{n})$ distribution. It’s necessary to know the sampling distribution of a statistic for most of the formal inferential methods to come. The Central Limit Theory provides a simple approximation to the actual sampling distribution—namely, a normal distribution with appropriate mean and standard deviation. For $\bar{y}$, the approximation is

$$\bar{y} \sim N(\mu, \sigma/\sqrt{n}).$$

The approximation holds for a sufficiently large sample size regardless of the shape of the population distribution. For virtually any population shape, the sample mean from a reasonably large sample will have an approximately normal distribution.

The normal approximation is accurate whenever

- The population distribution is itself normal. In this case, the sampling distribution of $\bar{y}$ is exactly normal for any sample size.
- The sample size is at least 30. If there are any outliers, an even larger sample size may be needed.

**Necessary conditions:** To be confident that the normal approximation is accurate, the following conditions must hold:

1. The data are a random sample the population.
2. The sample size is at least 30.

**Example:** The mean age of the AIDS patients was reported to be 37.4 years and the standard deviation reported to be 10.1 years. To verify this statement, suppose a random sample of $n = 36$ patients is obtained and their ages recorded.

1. What are $\mu, \sigma, \bar{y},$ and $s$?\(^3\)

2. What proportion of ages are less than or equal to 21? Let $X$ denote the age of a random sampled AIDS patient. The proportion of ages less than or equal to 21 is the same as $P(X \leq 21)$ where is the age of a randomly chosen AIDS patient. Without the full population, it is impossible to answer the question exactly. The question can be

\(^3\)Presumably, $\mu = 37.4$ and $\sigma = 10.1$. $\bar{y}$ and $s$ are not known.
answered approximately by proceeding as if the distribution of age over the population was normal:

\[ P(X \leq 21) = P\left( Z \leq \frac{21 - 37.4}{10.1} \right) \]
\[ = P(Z \leq -1.623) \]
\[ = .052. \]

3. What does the Central Limit Theorem say about the sampling distribution of \( \bar{y} \)\(^4\)?

4. What is the probability that the sample mean will be less or equal to than 35?

\[ P(\bar{y} \leq 35) = P\left( Z \leq \frac{35 - 37.4}{1.683} \right) \]
\[ = P(Z \leq -1.425) \]
\[ = .077. \]

5. Suppose that the age distribution is skewed to right. Is the calculation of the proportion of ages less than or equal to 21 accurate?\(^5\) Is the calculation involving the sample mean accurate?\(^6\)

6. Suppose that the sample mean age was \( \bar{y} = 34 \). Is there reason to doubt the reported population mean value of \( \mu = 37.4 \)?

A measure of statistical evidence against a population mean of 37.4 is obtained by computing the probability of obtaining a sample mean as small or smaller than 34. If this probability is very small, then the sample mean, and the data, contradict the supposition that 37.4 is \( \mu \). The computation is:

\[ P(\bar{y} \leq 34) = P\left( Z \leq \frac{34 - 37.4}{1.683} \right) \]
\[ = P(Z \leq -2.020) \]
\[ = .0217. \]

The result is statistically significant in the sense that the sample data contradict the reported value: it is improbable (the probability is approximately .02) to observe a sample mean as small or smaller than 34 if the population mean truly were 37.4.

7. What assumptions are underlying the previous calculation?\(^7\)

\(^4\) \( \bar{y} \sim N(\mu, \sigma(\bar{y})) \) where \( \mu = 37.4 \) and \( \sigma(\bar{y}) = 10.1/\sqrt{36} = 1.683. \)
\(^5\) No because the normal distribution probably will not be an accurate approximation of the age distribution. In fact, the actual proportion is .0204.
\(^6\) Yes, because \( n \geq 30 \) implies that the normal distribution is an accurate approximation of the distribution of the sample mean.
\(^7\) The sample mean was computed from a random sample.
8. Would a sample mean as small or smaller than 34 have been more or less likely to have occurred by chance if the sample size had been $n = 72$ (instead of $n = 36$)?

$^8$Smaller, since $\sigma(\bar{y}) = \sigma/\sqrt{n}$, and $P(\bar{y} \leq 34) = P\left(Z \leq \sqrt{n}\frac{34 - \mu}{\sigma}\right)$. 