Estimation

Section 7.3

Estimating \( p \) in the Binomial Distribution

Focus Points

• Compute the maximal length of error for proportions using a given level of confidence.

• Compute confidence intervals for \( p \) and interpret the results.

• Interpret poll results.

• Compute the sample size to be used for estimating a proportion \( p \) when we have an estimate for \( p \).
Focus Points

- Compute the sample size to be used for estimating a proportion $p$ when we have no estimate for $p$.

Estimating $p$ in the Binomial Distribution

The binomial distribution is completely determined by the number of trials $n$ and the probability $p$ of success on a single trial.

For most experiments, the number of trials is chosen in advance. Then the distribution is completely determined by $p$. In this section, we will consider the problem of estimating $p$ under the assumption that $n$ has already been selected.

We are employing what are called large-sample methods. We will assume that the normal curve is a good approximation to the binomial distribution, and when necessary, we will use sample estimates for the standard deviation.

Empirical studies have shown that these methods are quite good, provided both

$$np > 5 \quad \text{and} \quad nq > 5, \quad \text{where} \ q = 1 - p$$
Estimating $p$ in the Binomial Distribution

Let $r$ be the number of successes out of $n$ trials in a binomial experiment.

We will take the sample proportion of successes $\hat{p}$ (read "$p$ hat") = $r/n$ as our point estimate for $p$, the population proportion of successes.

Estimating $p$ in the Binomial Distribution

To compute the bounds for the margin of error, we need some information about the distribution of $\hat{p} = r/n$ values for different samples of the same size $n$.

Estimating $p$ in the Binomial Distribution

It turns out that, for large samples, the distribution of values is well approximated by a normal curve with mean $\mu = p$ and standard error $\sigma = \sqrt{pq/n}$.

Since the distribution of $\hat{p} = r/n$ is approximately normal, we use features of the standard normal distribution to find the bounds for the difference $\hat{p} - p$.

Recall that $z_c$ is the number such that an area equal to $c$ under the standard normal curve falls between $-z_c$ and $z_c$.
Then, in terms of the language of probability,

\[ -\frac{\sqrt{n}}{\sqrt{npq}} < \hat{p} - p < \frac{\sqrt{n}}{\sqrt{npq}} \]

Equation (17) says that the chance is \( c \) that the numerical difference between \( \hat{p} \) and \( p \) is between \( -\frac{\sqrt{n}}{\sqrt{npq}} \) and \( \frac{\sqrt{n}}{\sqrt{npq}} \). With the \( c \) confidence level, our estimate \( \hat{p} \) differs from \( p \) by no more than

\[ E = \frac{z}{\sqrt{npq}} \]

As in Section 7.1, we call \( E \) the maximal margin of error.

\[ z = \begin{cases} 1.96 & \text{for } c = 0.95 \\ 1.645 & \text{for } c = 0.90 \end{cases} \]
Estimating \( p \) in the Binomial Distribution

To find a \( c \) confidence interval for \( p \), we will use \( E \) in place of the expression \( \frac{z}{\sqrt{npq}} \) in Equation (17). Then we get

\[
P(-E < \hat{p} - p < E) = c
\]

(19)

Some algebraic manipulation produces the mathematically equivalent statement

\[
P(\hat{p} - E < p < \hat{p} + E) = c
\]

(20)

Equation (20) states that the probability is \( c \) that \( p \) lies in the interval from \( \hat{p} - E \) to \( \hat{p} + E \).

Therefore, the interval from \( \hat{p} - E \) to \( \hat{p} + E \) is the \( c \) confidence interval for \( p \) that we wanted to find.

There is one technical difficulty in computing the \( c \) confidence interval for \( p \).

The expression \( E = \frac{z}{\sqrt{npq}} \) requires that we know the values of \( p \) and \( q \). In most situations, we will not know the actual values of \( p \) or \( q \), so we will use our point estimates

\[
p = \hat{p} \quad \text{and} \quad q = 1 - p = 1 - \hat{p}
\]

to estimate \( E \).
Estimating $p$ in the Binomial Distribution

These estimates are reliable for most practical purposes, since we are dealing with large-sample theory ($np > 5$ and $nq > 5$).

For convenient reference, we’ll summarize the information about confidence intervals for $p$, the probability of success in a binomial distribution.

---

Example 6 – Confidence Interval for $p$

Let’s return to our flu shot experiment described at the beginning of this section.

Suppose that 800 students were selected at random from a student body of 20,000 and given shots to prevent a certain type of flu.

All 800 students were exposed to the flu, and 600 of them did not get the flu.

Let $p$ represent the probability that the shot will be successful for any single student selected at random from the entire population of 20,000. Let $q$ be the probability that the shot is not successful.
Example 6(a) – Confidence Interval for $p$

What is the number of trials $n$? What is the value of $r$?

Solution:
Since each of the 800 students receiving the shot may be thought of as a trial, then $n = 800$, and $r = 600$ is the number of successful trials.

Example 6(b) – Confidence Interval for $p$

What are the point estimates for $p$ and $q$?

Solution:
We estimate $p$ by the sample point estimate
\[
\hat{p} = \frac{r}{n} = \frac{600}{800} = 0.75
\]

We estimate $q$ by
\[
\hat{q} = 1 - \hat{p} = 1 - 0.75 = 0.25
\]

Example 6(c) – Confidence Interval for $p$

Check Requirements Would it seem that the number of trials is large enough to justify a normal approximation to the binomial?

Solution:
Since $n = 800$, $p = 0.75$, and $q = 0.25$, then
\[
np = (800)(0.75) = 600 > 5 \quad \text{and} \quad np = (800)(0.25) = 200 > 5
\]

A normal approximation is certainly justified.
Example 6(d) – Confidence Interval for $p$

Find a 99% confidence interval for $p$.

Solution:

$z_{0.99} = 2.58$ (Table 5(b) of Appendix II)

$$E = z_{0.99} \sqrt{\frac{p(1-p)}{n}} = 2.58 \sqrt{\frac{0.75(0.25)}{800}} = 0.0395$$

Example 6(d) – Solution

The 99% confidence interval is then

$$\hat{p} - E < p < \hat{p} + E$$

$$0.75 - 0.0395 < p < 0.75 + 0.0395$$

$$0.71 < p < 0.79$$

Interpretation: We are 99% confident that the probability a flu shot will be effective for a student selected at random is between 0.71 and 0.79.

Interpreting Results from a Poll
Interpreting Results from a Poll

Newspapers frequently report the results of an opinion poll. In articles that give more information, a statement about the margin of error accompanies the poll results.

In most polls, the margin of error is given for a 95% confidence interval.

Sample Size for Estimating $p$

Suppose you want to specify the maximal margin of error in advance for a confidence interval for $p$ at a given confidence level $c$.

What sample size do you need?

The answer depends on whether or not you have a preliminary estimate for the population probability of success $p$ in a binomial distribution.
Sample Size for Estimating $p$

Procedure:

Example 7 – Sample Size for Estimating $p$

A company is in the business of selling wholesale popcorn to grocery stores. The company buys directly from farmers. A buyer for the company is examining a large amount of corn from a certain farmer. Before the purchase is made, the buyer wants to estimate $p$, the probability that a kernel will pop.

Suppose a random sample of $n$ kernels is taken and $r$ of these kernels pop.

The buyer wants to be 95% sure that the point estimate $\hat{p} = r/n$ for $p$ will be in error either way by less than 0.01.

a. If no preliminary study is made to estimate $p$, how large a sample should the buyer use?
Example 7 – Solution

In this case, we use Equation (22) with $z_{0.95} = 1.96$ (see Table 7-2) and $E = 0.01$.

![Image](image.png)

Some Levels of Confidence and Their Corresponding Critical Values

Table 7-2

The buyer would need a sample of $n = 9604$ kernels.

Example 7 – Sample Size for Estimating $p$

(b) A preliminary study showed that $p$ was approximately 0.86. If the buyer uses the results of the preliminary study, how large a sample should he use?

Solution:

In this case, we use Equation (21) with $p = 0.86$.

Again, from Table 7-2, $z_{0.95} = 1.96$, and from the problem, $E = 0.01$.

$$n = \frac{p(1-p)}{E^2}$$
The sample size should be at least \( n = 4626 \). This sample is less than half the sample size necessary without the preliminary study.