Interval Estimation

1 Introduction

Although point estimation is a common way in which estimates are expressed, it leaves room for many questions. For instance, it does not tell us on how much information the estimate is based, nor does it tell us anything about the possible size of the error. Thus, we might have to supplement a point estimate $\hat{\theta}$ of θ with the size of the sample and the value of var$(\hat{\Theta})$ or with some other information about the sampling distribution of $\hat{\Theta}$. As we shall see, this will enable us to appraise the possible size of the error.

Alternatively, we might use **interval estimation**. An interval estimate of θ is an interval of the form $\hat{\theta}_1 < \theta < \hat{\theta}_2$, where $\hat{\theta}_1$ and $\hat{\theta}_2$ are values of appropriate random variables $\hat{\Theta}_1$ and $\hat{\Theta}_2$.

DEFINITION 1. CONFIDENCE INTERVAL. *If $\hat{\theta}_1$ and $\hat{\theta}_2$ are values of the random variables $\hat{\Theta}_1$ and $\hat{\Theta}_2$ such that*

$$P(\hat{\Theta}_1 < \theta < \hat{\Theta}_2) = 1 - \alpha$$

for some specified probability $1 - \alpha$, we refer to the interval

$$\hat{\theta}_1 < \theta < \hat{\theta}_2$$

*as a $(1 - \alpha)100\%$ **confidence interval** for θ. The probability $1 - \alpha$ is called the **degree of confidence**, and the endpoints of the interval are called the lower and upper **confidence limits**.*

For instance, when $\alpha = 0.05$, the degree of confidence is 0.95 and we get a 95% confidence interval.

It should be understood that, like point estimates, interval estimates of a given parameter are not unique. This is illustrated by Exercises 2 and 3 and also in Section 2, where we show that, based on a single random sample, there are various confidence intervals for μ, all having the same degree of confidence $1 - \alpha$. As was the case in point estimation, methods of interval estimation are judged by their various statistical properties. For instance, one desirable property is to have the length of a $(1 - \alpha)100\%$ confidence interval as short as possible; another desirable property is to have the expected length, $E(\hat{\Theta}_2 - \hat{\Theta}_1)$ as small as possible.

2 The Estimation of Means

To illustrate how the possible size of errors can be appraised in point estimation, suppose that the mean of a random sample is to be used to estimate the mean of a normal population with the known variance σ^2. By the theorem, "If χ is the mean of a random sample of size n from a normal population with the mean μ and the variance σ^2, its sampling distribution is a normal distribution with the mean μ and the variance σ^2/n", the sampling distribution of $\overline{X}$ for random samples of size n from a normal population with the mean μ and the variance σ^2 is a normal distribution with

$$\mu_{\overline{x}} = \mu \quad \text{and} \quad \sigma^2_{\overline{x}} = \frac{\sigma^2}{n}$$

Thus, we can write

$$P(|Z| < z_{\alpha/2}) = 1 - \alpha$$

where

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

and $z_{\alpha/2}$ is such that the integral of the standard normal density from $z_{\alpha/2}$ to ∞ equals $\alpha/2$. It follows that

$$P\left(\left|\overline{X} - \mu\right| < z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

or, in words, we have the following theorem.

> **Theorem 1.** If $\overline{X}$, the mean of a random sample of size n from a normal population with the known variance σ^2, is to be used as an estimator of the mean of the population, the probability is $1 - \alpha$ that the error will be less than $z_{\alpha/2} \cdot \dfrac{\sigma}{\sqrt{n}}$.

EXAMPLE 1

A team of efficiency experts intends to use the mean of a random sample of size $n = 150$ to estimate the average mechanical aptitude of assembly-line workers in a large industry (as measured by a certain standardized test). If, based on experience, the efficiency experts can assume that $\sigma = 6.2$ for such data, what can they assert with probability 0.99 about the maximum error of their estimate?

Solution

Substituting $n = 150$, $\sigma = 6.2$, and $z_{0.005} = 2.575$ into the expression for the maximum error, we get

$$2.575 \cdot \frac{6.2}{\sqrt{150}} = 1.30$$

Thus, the efficiency experts can assert with probability 0.99 that their error will be less than 1.30.

Suppose now that these efficiency experts actually collect the necessary data and get $\bar{x} = 69.5$. Can they still assert with probability 0.99 that the error of their estimate, $\bar{x} = 69.5$, is less than 1.30? After all, $\bar{x} = 69.5$ differs from the true (population) mean by less than 1.30 or it does not, and they have no way of knowing whether it is one or the other. Actually, they can, but it must be understood that the 0.99 probability applies to the method that they used to get their estimate and calculate the maximum error (collecting the sample data, determining the value of $\bar{x}$, and using the formula of Theorem 1) and not directly to the parameter that they are trying to estimate.

To clarify this distinction, it has become the custom to use the word "confidence" here instead of "probability." *In general, we make probability statements about future values of random variables* (*say, the potential error of an estimate*) *and confidence statements once the data have been obtained*. Accordingly, we should have said in our example that the efficiency experts can be 99% confident that the error of their estimate, $\bar{x} = 69.5$, is less than 1.30.

To construct a confidence-interval formula for estimating the mean of a normal population with the known variance σ^2, we return to the probability

$$P\left(|\overline{X} - \mu| < z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

the previous page, which we now write as

$$P\left(\overline{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

From this result, we have the following theorem.

Theorem 2. If $\bar{x}$ is the value of the mean of a random sample of size n from a normal population with the known variance σ^2, then

$$\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

is a $(1 - \alpha)100\%$ confidence interval for the mean of the population.

EXAMPLE 2

If a random sample of size $n = 20$ from a normal population with the variance $\sigma^2 = 225$ has the mean $\bar{x} = 64.3$, construct a 95% confidence interval for the population mean μ.

Solution

Substituting $n = 20$, $\bar{x} = 64.3$, $\sigma = 15$, and $z_{0.025} = 1.96$ into the confidence-interval formula of Theorem 2, we get

$$64.3 - 1.96 \cdot \frac{15}{\sqrt{20}} < \mu < 64.3 + 1.96 \cdot \frac{15}{\sqrt{20}}$$

which reduces to

$$57.7 < \mu < 70.9$$

As we pointed out earlier, confidence-interval formulas are not unique. This may be seen by changing the confidence-interval formula of Theorem 2 to

$$\bar{x} - z_{2\alpha/3} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/3} \cdot \frac{\sigma}{\sqrt{n}}$$

or to the **one-sided** $(1-\alpha)100\%$ **confidence-interval** formula

$$\mu < \bar{x} + z_{\alpha} \cdot \frac{\sigma}{\sqrt{n}}$$

Alternatively, we could base a confidence interval for μ on the sample median or, say, the midrange.

Strictly speaking, Theorems 1 and 2 require that we are dealing with a random sample from a normal population with the known variance σ^2. However, by virtue of the central limit theorem, these results can also be used for random samples from nonnormal populations provided that n is sufficiently large; that is, $n \geqq 30$. In that case, we may also substitute for σ the value of the sample standard deviation.

EXAMPLE 3

An industrial designer wants to determine the average amount of time it takes an adult to assemble an "easy-to-assemble" toy. Use the following data (in minutes), a random sample, to construct a 95% confidence interval for the mean of the population sampled:

17	13	18	19	17	21	29	22	16	28	21	15
26	23	24	20	8	17	17	21	32	18	25	22
16	10	20	22	19	14	30	22	12	24	28	11

Solution

Substituting $n = 36, \bar{x} = 19.92, z_{0.025} = 1.96$, and $s = 5.73$ for σ into the confidence-interval formula of Theorem 2, we get

$$19.92 - 1.96 \cdot \frac{5.73}{\sqrt{36}} < \mu < 19.92 + 1.96 \cdot \frac{5.73}{\sqrt{36}}$$

Thus, the 95% confidence limits are 18.05 and 21.79 minutes.

When we are dealing with a random sample from a normal population, $n < 30$, and σ is unknown, Theorems 1 and 2 cannot be used. Instead, we make use of the fact that

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

is a random variable having the t distribution with $n-1$ degrees of freedom. Substituting $\dfrac{\overline{X} - \mu}{S/\sqrt{n}}$ for T in

$$P(-t_{\alpha/2,n-1} < T < t_{\alpha/2,n-1}) = 1 - \alpha$$

we get the following confidence interval for μ.

THEOREM 3. If $\overline{x}$ and s are the values of the mean and the standard deviation of a random sample of size n from a normal population, then

$$\overline{x} - t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}} < \mu < \overline{x} + t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}$$

is a $(1-\alpha)100\%$ confidence interval for the mean of the population.

Since this confidence-interval formula is used mainly when n is small, less than 30, we refer to it as a small-sample confidence interval for μ.

EXAMPLE 4

A paint manufacturer wants to determine the average drying time of a new interior wall paint. If for 12 test areas of equal size he obtained a mean drying time of 66.3 minutes and a standard deviation of 8.4 minutes, construct a 95% confidence interval for the true mean μ.

Solution

Substituting $\overline{x} = 66.3, s = 8.4$, and $t_{0.025,11} = 2.201$ (from Table IV of "Statistical Tables"), the 95% confidence interval for μ becomes

$$66.3 - 2.201 \cdot \frac{8.4}{\sqrt{12}} < \mu < 66.3 + 2.201 \cdot \frac{8.4}{\sqrt{12}}$$

or simply

$$61.0 < \mu < 71.6$$

This means that we can assert with 95% confidence that the interval from 61.0 minutes to 71.6 minutes contains the true average drying time of the paint.

The method by which we constructed confidence intervals in this section consisted essentially of finding a suitable random variable whose values are determined by the sample data as well as the population parameters, yet whose distribution does not involve the parameter we are trying to estimate. This was the case, for example, when we used the random variable

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

whose values cannot be calculated without knowledge of μ, but whose distribution for random samples from normal populations, the standard normal distribution, does not involve μ. This method of confidence-interval construction is called the **pivotal method** and it is widely used, but there exist more general methods, such as the one discussed in the book by Mood, Graybill, and Boes referred to at the end of this chapter.

3 The Estimation of Differences Between Means

For independent random samples from normal populations

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

has the standard normal distribution. If we substitute this expression for Z into

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

the pivotal method yields the following confidence-interval formula for $\mu_1 - \mu_2$.

THEOREM 4. If $\overline{x}_1$ and $\overline{x}_2$ are the values of the means of independent random samples of sizes n_1 and n_2 from normal populations with the known variances σ_1^2 and σ_2^2, then

$$(\overline{x}_1 - x_2) - z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\overline{x}_1 - \overline{x}_2) + z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

is a $(1 - \alpha)100\%$ confidence interval for the difference between the two population means.

By virtue of the central limit theorem, this confidence-interval formula can also be used for independent random samples from nonnormal populations with known variances when n_1 and n_2 are large, that is, when $n_1 \geqq 30$ and $n_2 \geqq 30$.

EXAMPLE 5

Construct a 94% confidence interval for the difference between the mean lifetimes of two kinds of light bulbs, given that a random sample of 40 light bulbs of the first kind lasted on the average 418 hours of continuous use and 50 light bulbs of the second kind lasted on the average 402 hours of continuous use. The population standard deviations are known to be $\sigma_1 = 26$ and $\sigma_2 = 22$.

Solution

For $\alpha = 0.06$, we find from Table III of "Statistical Tables" that $z_{0.03} = 1.88$. Therefore, the 94% confidence interval for $\mu_1 - \mu_2$ is

$$(418 - 402) - 1.88 \cdot \sqrt{\frac{26^2}{40} + \frac{22^2}{50}} < \mu_1 - \mu_2 < (418 - 402) + 1.88 \cdot \sqrt{\frac{26^2}{40} + \frac{22^2}{50}}$$

which reduces to

$$6.3 < \mu_1 - \mu_2 < 25.7$$

Hence, we are 94% confident that the interval from 6.3 to 25.7 hours contains the actual difference between the mean lifetimes of the two kinds of light bulbs. The fact that both confidence limits are positive suggests that on the average the first kind of light bulb is superior to the second kind.

To construct a $(1 - \alpha)100\%$ confidence interval for the difference between two means when $n_1 \geqq 30$, $n_2 \geqq 30$, but σ_1 and σ_2 are unknown, we simply substitute

s_1 and s_2 for σ_1 and σ_2 and proceed as before. When σ_1 and σ_2 are unknown and either or both of the samples are small, the procedure for estimating the difference between the means of two normal populations is not straightforward unless it can be assumed that $\sigma_1 = \sigma_2$. If $\sigma_1 = \sigma_2 = \sigma$, then

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

is a random variable having the standard normal distribution, and σ^2 can be estimated by **pooling** the squared deviations from the means of the two samples. In Exercise 9 the reader will be asked to verify that the resulting **pooled estimator**

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

is, indeed, an unbiased estimator of σ^2. Now, by the two theorems, "If $\overline{X}$ and S^2 are the mean and the variance of a random sample of size n from a normal population with the mean μ and the standard deviation σ, then **1.** $\overline{X}$ and S^2 are independent; **2.** the random variable $\dfrac{(n-1)S^2}{\sigma^2}$ has a chi-square distribution with $n-1$ degrees of freedom. If $X_1, X_2, \ldots, X_n$ are independent random variables having chi-square distributions with $\nu_1, \nu_2, \ldots, \nu_n$ degrees of freedom, then $Y = \sum_{i=1}^{n} X_i$ has the chi-square distribution with $\nu_1 + \nu_2 + \cdots + \nu_n$ degrees of freedom" the independent random variables

$$\frac{(n_1 - 1)S_1^2}{\sigma^2} \quad \text{and} \quad \frac{(n_2 - 1)S_2^2}{\sigma^2}$$

have chi-square distributions with $n_1 - 1$ and $n_2 - 1$ degrees of freedom, and their sum

$$Y = \frac{(n_1 - 1)S_1^2}{\sigma^2} + \frac{(n_2 - 1)S_2^2}{\sigma^2} = \frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2}$$

has a chi-square distribution with $n_1 + n_2 - 2$ degrees of freedom. Since it can be shown that the above random variables Z and Y are independent (see references at the end of this chapter)

$$\begin{aligned} T &= \frac{Z}{\sqrt{\dfrac{Y}{n_1 + n_2 - 2}}} \\ &= \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \end{aligned}$$

has a t distribution with $n_1 + n_2 - 2$ degrees of freedom. Substituting this expression for T into

$$P(-t_{\alpha/2,n-1} < T < t_{\alpha/2,n-1}) = 1 - \alpha$$

we arrive at the following $(1 - \alpha)100\%$ confidence interval for $\mu_1 - \mu_2$.

THEOREM 5. If $\overline{x}_1, \overline{x}_2, s_1$, and s_2 are the values of the means and the standard deviations of independent random samples of sizes n_1 and n_2 from normal populations with equal variances, then

$$(\overline{x}_1 - \overline{x}_2) - t_{\alpha/2, n_1+n_2-2} \cdot s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2$$
$$< (\overline{x}_1 - \overline{x}_2) + t_{\alpha/2, n_1+n_2-2} \cdot s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

is a $(1-\alpha)100\%$ confidence interval for the difference between the two population means.

Since this confidence-interval formula is used mainly when n_1 and/or n_2 are small, less than 30, we refer to it as a small-sample confidence interval for $\mu_1 - \mu_2$.

EXAMPLE 6

A study has been made to compare the nicotine contents of two brands of cigarettes. Ten cigarettes of Brand A had an average nicotine content of 3.1 milligrams with a standard deviation of 0.5 milligram, while eight cigarettes of Brand B had an average nicotine content of 2.7 milligrams with a standard deviation of 0.7 milligram. Assuming that the two sets of data are independent random samples from normal populations with equal variances, construct a 95% confidence interval for the difference between the mean nicotine contents of the two brands of cigarettes.

Solution

First we substitute $n_1 = 10, n_2 = 8, s_1 = 0.5$, and $s_2 = 0.7$ into the formula for s_p, and we get

$$s_p = \sqrt{\frac{9(0.25)+7(0.49)}{16}} = 0.596$$

Then, substituting this value together with $n_1 = 10$, $n_2 = 8$, $\overline{x}_1 = 3.1$, $\overline{x}_2 = 2.7$, and $t_{0.025,16} = 2.120$ (from Table IV of "Statistical Tables") into the confidence-interval formula of Theorem 5, we find that the required 95% confidence interval is

$$(3.1-2.7) - 2.120(0.596)\sqrt{\frac{1}{10}+\frac{1}{8}} < \mu_1 - \mu_2$$
$$< (3.1-2.7) + 2.120(0.596)\sqrt{\frac{1}{10}+\frac{1}{8}}$$

which reduces to

$$-0.20 < \mu_1 - \mu_2 < 1.00$$

Thus, the 95% confidence limits are -0.20 and 1.00 milligrams; but observe that since this includes $\mu_1 - \mu_2 = 0$, we cannot conclude that there is a real difference between the average nicotine contents of the two brands of cigarettes.

Exercises

1. If x is a value of a random variable having an exponential distribution, find k so that the interval from 0 to kx is a $(1-\alpha)100\%$ confidence interval for the parameter θ.

2. If x_1 and x_2 are the values of a random sample of size 2 from a population having a uniform density with $\alpha = 0$ and $\beta = \theta$, find k so that

$$0 < \theta < k(x_1 + x_2)$$

is a $(1-\alpha)100\%$ confidence interval for θ when

(a) $\alpha \leqq \frac{1}{2}$; **(b)** $\alpha > \frac{1}{2}$.

3. This question has been intentionally omitted for this edition.

4. Show that the $(1-\alpha)100\%$ confidence interval

$$\overline{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

is shorter than the $(1-\alpha)100\%$ confidence interval

$$\overline{x} - z_{2\alpha/3} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + z_{\alpha/3} \cdot \frac{\sigma}{\sqrt{n}}$$

5. Show that among all $(1-\alpha)100\%$ confidence intervals of the form

$$\overline{x} - z_{k\alpha} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + z_{(1-k)\alpha} \cdot \frac{\sigma}{\sqrt{n}}$$

the one with $k = 0.5$ is the shortest.

6. Show that if $\overline{x}$ is used as a point estimate of μ and σ is known, the probability is $1-\alpha$ that $|\overline{x}-\mu|$, the absolute value of our error, will not exceed a specified amount e when

$$n = \left[z_{\alpha/2} \cdot \frac{\sigma}{e} \right]^2$$

(If it turns out that $n < 30$, this formula cannot be used unless it is reasonable to assume that we are sampling a normal population.)

7. Modify Theorem 1 so that it can be used to appraise the maximum error when σ^2 is unknown. (Note that this method can be used only after the data have been obtained.)

8. State a theorem analogous to Theorem 1, which enables us to appraise the maximum error in using $\overline{x}_1 - \overline{x}_2$ as an estimate of $\mu_1 - \mu_2$ under the conditions of Theorem 4.

9. Show that S_p^2 is an unbiased estimator of σ^2 and find its variance under the conditions of Theorem 5.

10. This question has been intentionally omitted for this edition.

4 The Estimation of Proportions

In many problems we must estimate proportions, probabilities, percentages, or rates, such as the proportion of defectives in a large shipment of transistors, the probability that a car stopped at a road block will have faulty lights, the percentage of schoolchildren with I.Q.'s over 115, or the mortality rate of a disease. In many of these it is reasonable to assume that we are sampling a binomial population and, hence, that our problem is to estimate the binomial parameter θ. Thus, we can make use of the fact that for large n the binomial distribution can be approximated with a normal distribution; that is,

$$Z = \frac{X - n\theta}{\sqrt{n\theta(1-\theta)}}$$

can be treated as a random variable having approximately the standard normal distribution. Substituting this expression for Z into

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

we get

$$P\left(-z_{\alpha/2} < \frac{X - n\theta}{\sqrt{n\theta(1-\theta)}} < z_{\alpha/2} \right) = 1 - \alpha$$

and the two inequalities

$$-z_{\alpha/2} < \frac{x - n\theta}{\sqrt{n\theta(1-\theta)}} \quad \text{and} \quad \frac{x - n\theta}{\sqrt{n\theta(1-\theta)}} < z_{\alpha/2}$$

whose solution will yield $(1-\alpha)100\%$ confidence limits for θ. Leaving the details of this to the reader in Exercise 11, let us give here instead a large-sample approximation by rewriting $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha$, with $\dfrac{X - n\theta}{\sqrt{n\theta(1-\theta)}}$ substituted for Z, as

$$P\left(\hat{\Theta} - z_{\alpha/2} \cdot \sqrt{\frac{\theta(1-\theta)}{n}} < \theta < \hat{\Theta} + z_{\alpha/2} \cdot \sqrt{\frac{\theta(1-\theta)}{n}}\right) = 1-\alpha$$

where $\hat{\Theta} = \dfrac{X}{n}$. Then, if we substitute $\hat{\theta}$ for θ inside the radicals, which is a further approximation, we obtain the following theorem.

THEOREM 6. If X is a binomial random variable with the parameters n and θ, n is large, and $\hat{\theta} = \dfrac{x}{n}$, then

$$\hat{\theta} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} < \theta < \hat{\theta} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$$

is an approximate $(1-\alpha)100\%$ confidence interval for θ.

EXAMPLE 7

In a random sample, 136 of 400 persons given a flu vaccine experienced some discomfort. Construct a 95% confidence interval for the true proportion of persons who will experience some discomfort from the vaccine.

Solution

Substituting $n = 400, \hat{\theta} = \frac{136}{400} = 0.34$, and $z_{0.025} = 1.96$ into the confidence-interval formula of Theorem 6, we get

$$0.34 - 1.96\sqrt{\frac{(0.34)(0.66)}{400}} < \theta < 0.34 + 1.96\sqrt{\frac{(0.34)(0.66)}{400}}$$
$$0.294 < \theta < 0.386$$

or, rounding to two decimals, $0.29 < \theta < 0.39$.

Using the same approximations that led to Theorem 6, we can also obtain the following theorem.

THEOREM 7. If $\hat{\theta} = \frac{x}{n}$ is used as an estimate of θ, we can assert with $(1 - \alpha)100\%$ confidence that the error is less than

$$z_{\alpha/2} \cdot \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$$

EXAMPLE 8

A study is made to determine the proportion of voters in a sizable community who favor the construction of a nuclear power plant. If 140 of 400 voters selected at random favor the project and we use $\hat{\theta} = \frac{140}{400} = 0.35$ as an estimate of the actual proportion of all voters in the community who favor the project, what can we say with 99% confidence about the maximum error?

Solution

Substituting $n = 400, \hat{\theta} = 0.35$, and $z_{0.005} = 2.575$ into the formula of Theorem 7, we get

$$2.575 \cdot \sqrt{\frac{(0.35)(0.65)}{400}} = 0.061$$

or 0.06 rounded to two decimals. Thus, if we use $\hat{\theta} = 0.35$ as an estimate of the actual proportion of voters in the community who favor the project, we can assert with 99% confidence that the error is less than 0.06.

---

5 The Estimation of Differences Between Proportions

In many problems we must estimate the difference between the binomial parameters θ_1 and θ_2 on the basis of independent random samples of sizes n_1 and n_2 from two binomial populations. This would be the case, for example, if we want to estimate the difference between the proportions of male and female voters who favor a certain candidate for governor of Illinois.

If the respective numbers of successes are X_1 and X_2 and the corresponding sample proportions are denoted by $\hat{\Theta}_1 = \frac{X_1}{n_1}$ and $\hat{\Theta}_2 = \frac{X_2}{n_2}$, let us investigate the sampling distribution of $\hat{\Theta}_1 - \hat{\Theta}_2$, which is an obvious estimator of $\theta_1 - \theta_2$. Let's take

$$E(\hat{\Theta}_1 - \hat{\Theta}_2) = \theta_1 - \theta_2$$

and

$$\text{var}(\hat{\Theta}_1 - \hat{\Theta}_2) = \frac{\theta_1(1-\theta_1)}{n_1} + \frac{\theta_2(1-\theta_2)}{n_2}$$