

3. Estimation in the Bernoulli Model

Preliminaries

Suppose that $X = (X_1, X_2, \dots, X_n)$ is a random sample from the [Bernoulli distribution](#) with unknown success parameter $p \in (0, 1)$. Thus, these are [independent random variables](#) taking the values 1 and 0 with probabilities p and $1 - p$ respectively. Recall that the mean and variance of the Bernoulli distribution are $\mathbb{E}(X) = p$ and $\text{var}(X) = p(1 - p)$.

Usually, this model arises in one of the following contexts:

1. There is an *event* of interest in a basic experiment, with unknown probability p . We replicate the experiment n times and define $X_i = 1$ if and only if the event occurred on the i^{th} run.
2. We have a population of objects of several different types; p is the unknown proportion of objects of a particular type of interest. We select n objects at random from the population and let $X_i = 1$ if and only if the i^{th} object is of the type of interest. When the sampling is *with* replacement, these variables really do form a random sample from the Bernoulli distribution. When the sampling is *without* replacement, the variables are dependent, but the Bernoulli model may still be approximately valid. For more on these points, see the discussion of [sampling with and without replacement](#) in the chapter on [Finite Sampling Models](#).

In this section, we will construct confidence intervals for p . A parallel section on [Tests in the Bernoulli Model](#) is in the chapter on [Hypothesis Testing](#). Note that the sample mean of our data vector X

$$M = \frac{1}{n} \sum_{i=1}^n X_i$$

is the sample proportion of objects of the type of interest. By the [central limit theorem](#), the standard score

$$Z = \frac{M - p}{\sqrt{\frac{p(1-p)}{n}}}$$

has approximately a standard normal distribution and hence is (approximately) a pivot variable for p . For a given sample size n , the distribution of Z is closest to normal when p is near $\frac{1}{2}$ and farthest from normal when p is near 0 or 1 (extreme). Because the pivot variable is (approximately) normally distributed, the construction of confidence intervals for p in this model is similar to the construction of [confidence intervals for \$\mu\$ in the normal model](#).

As usual, for $r \in (0, 1)$, let $z(r)$ denote the [quantile](#) of order r for the standard normal distribution. For

selected values of r , $z(r)$ can be obtained from the last row of the [table of the \$t\$ distribution](#), from the [table of the standard normal distribution](#), from the [quantile applet](#), or from most statistical software packages.

Approximate Confidence Intervals for p

Basic Confidence Intervals

1. Use the pivot variable to show that for any $r \in (0, 1)$ and any $\alpha \in (0, 1)$, an approximate $1 - \alpha$ confidence set for p is

$$\left\{ p \in [0, 1] : M - z(1 - r\alpha) \sqrt{\frac{p(1-p)}{n}} \leq p \leq M - z(\alpha - r\alpha) \sqrt{\frac{p(1-p)}{n}} \right\}$$

As usual, r is the proportion of the significance level $1 - \alpha$ in the right tail of the distribution of the pivot variable, and $1 - r$ is the proportion of the significance level $1 - \alpha$ in the left tail of the distribution of the pivot variable.

2. Use the quadratic formula to show that the confidence set in [Exercise 1](#) is actually an interval of the form $[U(z(1 - r\alpha)), U(z(\alpha - r\alpha))]$ where

$$U(z) = \frac{n}{n + z^2} \left(M + \frac{z^2}{2n} - z \sqrt{\frac{M(1-M)}{n} + \frac{z^2}{4n^2}} \right)$$

As usual, the most important special cases are the equal-tailed $1 - \alpha$ confidence interval, obtained by setting $r = \frac{1}{2}$, the $1 - \alpha$ confidence upper bound, obtained by setting $r = 0$, and the $1 - \alpha$ confidence lower bound obtained by setting $r = 1$.

Simplified Confidence Intervals

A simplified approximate $1 - \alpha$ confidence interval for p can be obtained by replacing the distribution parameter p by the point estimate M in the extreme parts of the inequality in [Exercise 1](#):

$$\left[M - z(1 - r\alpha) \sqrt{\frac{M(1-M)}{n}}, M - z(\alpha - r\alpha) \sqrt{\frac{M(1-M)}{n}} \right]$$

3. Show that an approximate $1 - \alpha$ level confidence lower bound for p is

$$M - z(1 - \alpha) \sqrt{\frac{M(1-M)}{n}}$$

4. Show that an approximate $1 - \alpha$ level confidence upper bound for p is

$$M - z(\alpha) \sqrt{\frac{M(1-M)}{n}} = M + z(1 - \alpha) \sqrt{\frac{M(1-M)}{n}}$$

5. Of the two-sided $1 - \alpha$ confidence intervals in [Exercise 1](#), show that the one with smallest length is the **equal-tailed** interval obtained by letting $r = \frac{1}{2}$

$$\left[M - z\left(1 - \frac{\alpha}{2}\right) \sqrt{\frac{M(1-M)}{n}}, M + z\left(1 - \frac{\alpha}{2}\right) \sqrt{\frac{M(1-M)}{n}} \right]$$

Note that this interval is **symmetric** about the sample proportion M but that the length of the interval, as well as the center is random. This is the two-sided interval that is normally used.

6. Use the simulation of the **proportion estimation experiment** to explore the procedure. Use various values of p and various confidence levels, sample sizes, and interval types. For each configuration, run the experiment 1000 times with an update frequency of 10 and note how well the proportion of successful intervals approximates the theoretical confidence level.

Conservative Confidence Intervals

7. Show that the variance of the Bernoulli distribution is maximized when $p = \frac{1}{2}$ and thus the maximum variance is $\frac{1}{4}$.

8. Use the pivot variable to show that for any $r \in [0, 1]$ and any $\alpha \in (0, 1)$, a conservative $1 - \alpha$ confidence interval for p is

$$\left[M - z(1 - r\alpha) \frac{1}{2\sqrt{n}}, M - z(\alpha - r\alpha) \frac{1}{2\sqrt{n}} \right]$$

9. Show that a conservative $1 - \alpha$ level confidence lower bound for p is

$$M - z(1 - \alpha) \frac{1}{2\sqrt{n}}$$

10. Show that a conservative $1 - \alpha$ level confidence upper bound for p is

$$M - z(\alpha) \frac{1}{2\sqrt{n}} = M + z(1 - \alpha) \frac{1}{2\sqrt{n}}$$

11. Of the two-sided $1 - \alpha$ confidence intervals in [Exercise 8](#), show that the one with smallest length is the **equal-tailed** interval obtained by letting $r = \frac{1}{2}$

$$\left[M - z\left(1 - \frac{\alpha}{2}\right) \frac{1}{2\sqrt{n}}, M + z\left(1 - \frac{\alpha}{2}\right) \frac{1}{2\sqrt{n}} \right]$$

Note that this interval is **symmetric** about the sample proportion M and that the length of the interval is deterministic. This is the conservative two-sided interval that is normally used. Of course, the conservative confidence intervals will be larger than the approximate confidence intervals. The conservative estimate can be used to design the experiment.

12. Show that a conservative estimate of the sample size n needed to estimate p with confidence $1 - \alpha$ and margin of error d is as follows, where $z_\alpha = z\left(1 - \frac{\alpha}{2}\right)$ for the two-sided interval and $z_\alpha = z(1 - \alpha)$ for the confidence upper or lower bound:

$$n = \left\lceil \frac{z_\alpha^2}{4d^2} \right\rceil$$

Computational Exercises

13. In a pole of 1000 registered voters in a certain district, 427 prefer candidate X. Construct the 95% two-sided confidence interval for the proportion of all registered voters in the district that prefer X.
14. A coin is tossed 500 times and results in 302 heads. Construct the 95% confidence lower bound for the probability of heads. Do you believe that the coin is fair?
15. A sample of 400 memory chips from a production line are tested, and 30 are defective. Construct the conservative 90% two-sided confidence interval for the proportion of defective chips.
16. A drug company wants to estimate the proportion of persons who will experience an adverse reaction to a certain new drug. The company wants a two-sided interval with margin of error 0.03 with 95% confidence. How large should the sample be?
17. An advertising agency wants to construct a 99% confidence lower bound for the proportion of dentists who recommend a certain brand of toothpaste. The margin of error is to be 0.02. How large should the sample be?
18. The [Buffon trial data](#) set gives the results of 104 repetitions of [Buffon's needle experiment](#). Theoretically, the data should correspond to Bernoulli trials with $p = \frac{2}{\pi}$, but because real students dropped the needle, the true value of p is unknown. Construct a 95% confidence interval for p . Do you believe that p is the theoretical value?