8.3: Estimation in the Bernoulli Model

Introduction

Recall that an *indicator variable* is a random variable that just takes the values 0 and 1. In applications, an indicator variable indicates which of two complementary events in a random experiment has occurred. Typical examples include

- A manufactured item subject to unavoidable random factors is either defective or acceptable.
- A voter selected from a population either supports a particular candidate or does not.
- A person selected from a population either does or does not have a particular medical condition.
- A student in a class either passes or fails a standardized test.
- A sample of radioactive material either does or does not emit an alpha particle in a specified ten-second period.

Recall also that the distribution of an indicator variable is known as the *Bernoulli distribution*, named for Jacob Bernoulli, and has probability density function given by $\mathbb{P}(X = 1) = p$, $\mathbb{P}(X = 0) = 1 - p$, where $p \in (0, 1)$ is the basic parameter. In the context of the examples above,

- p is the probability that the manufactured item is defective.
- p is the proportion of voters in the population who favor the candidate.
- p is the poportion of persons in the population that have the medical condition.
- p is the probability that a student in the class will pass the exam.
- p is the probability that the material will emit an alpha particle in the specified period.

Recall that the mean and variance of the Bernoulli distribution are $\mathbb{E}(X) = p$ and $\var(X) = p(1 - p)$. Often in statistical applications, p is unknown and must be estimated from sample data. In this section, we will see how to construct interval estimates for the parameter from sample data. A parallel section on Tests in the Bernoulli Model is in the chapter on Hypothesis Testing.

The One-Sample Model

Preliminaries

Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample from the Bernoulli distribution with unknown parameter $p \in [0, 1]$. That is, $\boldsymbol{X}$ is a squence of Bernoulli trials. From the examples in the introduction above, note that often the underlying experiment is to sample at random from a dichotomous population. When the sampling is *with* replacement, $\boldsymbol{X}$ really is a sequence of Bernoulli trials. When the sampling is *without* replacement, the variables are dependent, but the Bernoulli model is still approximately valid if the population size is large compared to the sample size n. For more on these points, see the discussion of sampling with and without replacement in the chapter on Finite Sampling Models.

Note that the sample mean of our data vector $\boldsymbol{X}$, namely

$$M = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{8.3.1}$$

is the sample proportion of objects of the type of interest. By the central limit theorem, the standard score

$$Z = \frac{M - p}{\sqrt{p(1 - p)/n}} \tag{8.3.2}$$

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 the distribution mean μ in the normal model. But of course all of the confidence intervals so constructed are approximate.

As usual, for $r \in (0, 1)$, let $z(r)$ denote the quantile of order r for the standard normal distribution. Values of $z(r)$ can be obtained from the special distribution calculator, or from most statistical software packages.

Basic Confidence Intervals

For $\alpha \in (0,1)$, the following are approximate $1-\alpha$ confidence sets for p:

1. $\left\{p \in [0,1] : M - z(1-\alpha/2)\sqrt{p(1-p)/n} \le p \le M + z(1-\alpha/2)\sqrt{p(1-p)/n}\right\}$
2. $\left\{p \in [0,1] : p \le M + z(1-\alpha)\sqrt{p(1-p)/n}\right\}$
3. $\left\{p \in [0,1] : M - z(1-\alpha)\sqrt{p(1-p)/n} \le p\right\}$

Proof

From our discussion above, $(M-p)/\sqrt{p(1-p)/n}$ has approximately a standard normal distribution. Hence by definition of the quantiles,

1. $\mathbb{P}[-z(1-\alpha/2) \le (M-p)/\sqrt{p(1-p)/n} \le z(1-\alpha/2)] \approx 1-\alpha$
2. $\mathbb{P}[-z(1-\alpha) \le (M-p)/\sqrt{p(1-p)/n}] \approx 1-\alpha$
3. $\mathbb{P}[(M-p)/\sqrt{p(1-p)/n} \le z(1-\alpha)] \approx 1-\alpha$

Solving the inequalities for p in the numerator of $(M-p)/\sqrt{p(1-p)/n}$ for each event gives the corresponding confidence set.

These confidence sets are actually intervals, known as the *Wilson intervals*, in honor of Edwin Wilson.

The confidence sets for p in (1) are intervals. Let

$$U(z) = \frac{n}{n+z^2}\left(M + \frac{z^2}{2n} + z\sqrt{\frac{M(1-M)}{n} + \frac{z^2}{4n^2}}\right) \tag{8.3.3}$$

Then the following have approximate confidecne level $1-\alpha$ for p.

1. The two-sided interval $[U[-z(1-\alpha/2)], U[z(1-\alpha/2)]]$.
2. The upper bound $U[z(1-\alpha)]$.
3. The lower bound $U[-z(1-\alpha)]$.

Proof

This follows by solving the inequalities in (1) for p. For each inequality, we can isolate the square root term, and then square both sides. This gives quadratic inequalities, which can be solved using the quadratic formula.

As usual, the *equal-tailed* confidence interval in (a) is not the only two-sided $1-\alpha$ confidence interval for p. We can divide the α probability between the left and right tails of the standard normal distribution in any way that we please.

For α, $r \in (0,1)$, an approximate two-sided $1-\alpha$ confidence interval for p is $[U[z(\alpha - r\alpha)], U[z(1-r\alpha)]]$ where U is the function in (2).

Proof

As in the proof of (1),

$$\mathbb{P}\left[z(\alpha - r\alpha) \le \frac{M-p}{\sqrt{p(1-p)/n}} \le z(1-r\alpha)\right] \approx 1-\alpha \tag{8.3.4}$$

Solving for p with the help of the quadratic formula gives the result.

In practice, the equal-tailed $1-\alpha$ confidence interval in part (a) of (2), obtained by setting $r = \frac{1}{2}$, is the one that is always used. As $r \uparrow 1$, the right enpoint converges to the $1-\alpha$ confidence upper bound in part (b), and as $r \downarrow 0$ the left endpoint converges to the $1-\alpha$ confidence lower bound in part (c).

Simplified Confidence Intervals

Simplified approximate $1-\alpha$ confidence intervals for p can be obtained by replacing the distribution mean p by the sample mean M in the extreme parts of the inequalities in (1).

For $\alpha \in (0, 1)$, the following have approximate confidence level $1 - \alpha$ for p:

1. The two-sided interval with endpoints $M \pm z(1 - \alpha/2)\sqrt{M(1 - M)/n}$.
2. The upper bound $M + z(1 - \alpha)\sqrt{M(1 - M)/n}$.
3. The lower bound $M - z(1 - \alpha)\sqrt{M(1 - M)/n}$.

Proof

As noted, these results follows from the confidence set in (1) by replacing p with M in the expression $\sqrt{p(1 - p)/n}$.

These confidence intervals are known as *Wald intervals*, in honor of Abraham Wald.. Note that the Wald interval can also be obtained from the Wilson intervals in (2) by assuming that n is large compared to z, so that $n/(n + z^2) \approx 1$, $z^2/2n \approx 0$, and $z^2/4n^2 \approx 0$. Note that this interval in (c) 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.

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 and compare the proportion of successful intervals to the theoretical confidence level.

As always, the equal-tailed interval in (4) is not the only two-sided, $1 - \alpha$ confidence interval.

For α, $r \in (0, 1)$, an approximate two-sided $1 - \alpha$ confidence interval for p is

$$\left[M - z(1 - r\alpha)\sqrt{\frac{M(1 - M)}{n}}, M - z(\alpha - r\alpha)\sqrt{\frac{M(1 - M)}{n}}\right] \tag{8.3.5}$$

The interval with smallest length is the equal-tail interval with $r = \frac{1}{2}$.

Conservative Confidence Intervals

Note that the function $p \mapsto p(1 - p)$ on the interval $[0, 1]$ is maximized when $p = \frac{1}{2}$ and thus the maximum value is $\frac{1}{4}$. We can obtain conservative confidence intervals for p from the basic confidence intervals by using this fact.

For $\alpha \in (0, 1)$, the following have approximate confidence level at least $1 - \alpha$ for p:

1. The two-sided interval with endpoints $M \pm z(1 - \alpha/2)\frac{1}{2\sqrt{n}}$.
2. The upper bound $M + z(1 - \alpha)\frac{1}{2\sqrt{n}}$.
3. The lower bound $M - z(1 - \alpha)\frac{1}{2\sqrt{n}}$.

Proof

As noted, these results follows from the confidence sets in (1) by replacing p with $\frac{1}{2}$ in the expression $\sqrt{p(1 - p)/n}$.

Note that the confidence interval in (a) is symmetric about the sample proportion M and that the length of the interval is deterministic. Of course, the conservative confidence intervals will be larger than the approximate simplified confidence intervals in (4). The conservative estimate can be used to design the experiment. Recall that the *margin of error* is the distance between the sample proportion M and an endpoint of the confidence interval.

A conservative estimate of the sample size n needed to estimate p with confidence $1 - \alpha$ and margin of error d is

$$n = \left\lceil \frac{z_\alpha^2}{4d^2} \right\rceil \tag{8.3.6}$$

where $z_\alpha = z(1 - \alpha/2)$ for the two-sided interval and $z_\alpha = z(1 - \alpha)$ for the confidence upper or lower bound.

Proof

With confidence level $1 - \alpha$, the margin of error is $z_\alpha \frac{1}{2\sqrt{n}}$. Setting this equal to the prescribed value d and solving gives the result.

As always, the equal-tailed interval in (7) is not the only two-sided, conservative, $1-\alpha$ confidence interval.

For α, $r \in (0,1)$, an approximate two-sided, conservative $1-\alpha$ confidence interval for p is

$$\left[M - z(1-r\alpha)\frac{1}{2\sqrt{n}}, M - z(\alpha - r\alpha)\frac{1}{2\sqrt{n}}\right] \tag{8.3.7}$$

The interval with smallest length is the equal-tail interval with $r = \frac{1}{2}$.

The Two-Sample Model

Preliminaries

Often we have two underlying Bernoulli distributions, with parameters p_1, $p_2 \in [0,1]$ and we would like to estimate the difference $p_1 - p_2$. This problem could arise in the following typical examples:

- In a quality control setting, suppose that p_1 is the proportion of defective items produced under one set of manufacturing conditions while p_2 is the proportion of defectives under a different set of conditions.
- In an election, suppose that p_1 is the proportion of voters who favor a particular candidate at one point in the campaign, while p_2 is the proportion of voters who favor the candidate at a later point (perhaps after a scandal has erupted).
- Suppose that p_1 is the proportion of students who pass a certain standardized test with the usual test preparation methods while p_2 is the proportion of students who pass the test with a new set of preparation methods.
- Suppose that p_1 is the proportion of unvaccinated persons in a certain population who contract a certain disease, while p_2 is the proportion of vaccinated person who contract the disease.

Note that several of these examples can be thought of as *treatment-control* problems. Of course, we could construct interval estimates I_1 for p_1 and I_2 for p_2 separately, as in the subsections above. But as we noted in the Introduction, if these two intervals have confidence level $1-\alpha$, then the product set $I_1 \times I_2$ has confidence level $(1-\alpha)^2$ for (p_1, p_2). So if $p_1 - p_2$ is our parameter of interest, we will use a different approach.

Simplified Confidence Intervals

Suppose now that $\boldsymbol{X} = (X_1, X_2, \ldots, X_{n_1})$ is a random sample of size n_1 from the Bernoulli distribution with parameter p_1, and $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_{n_2})$ is a random sample of size n_2 from the Bernoulli distribution with parameter p_2. We assume that the samples $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent. Let

$$M_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i, \quad M_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i \tag{8.3.8}$$

denote the sample means (sample proportions) for the samples $\boldsymbol{X}$ and $\boldsymbol{Y}$. A natural point estimate for $p_1 - p_2$, and the building block for our interval estimate, is $M_1 - M_2$. As noted in the one-sample model, if n_i is large, M_i has an approximate normal distribution with mean p_i and variance $p_i(1-p_i)/n_i$ for $i \in \{1,2\}$. Since the samples are independent, so are the sample means. Hence $M_1 - M_2$ has an approximate normal distribution with mean $p_1 - p_2$ and variance $p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2$. We now have all the tools we need for a simplified, approximate confidence interval for $p_1 - p_2$.

For $\alpha \in (0,1)$, the following have approximate confidence level $1-\alpha$ for $p_1 - p_2$:

1. The two-sided interval with endpoints $(M_1 - M_2) \pm z(1-\alpha/2)\sqrt{M_1(1-M_1)/n_1 + M_2(1-M_2)/n_2}$.
2. The lower bound $(M_1 - M_2) - z(1-\alpha)\sqrt{M_1(1-M_1)/n_1 + M_2(1-M_2)/n_2}$.
3. The upper bound $(M_1 - M_2) + z(1-\alpha)\sqrt{M_1(1-M_1)/n_1 + M_2(1-M_2)/n_2}$.

Proof

As noted above, if n_1 and n_2 are large,

$$\frac{(M_1 - M_2) - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} \tag{8.3.9}$$

has approximatle a standard normal distribution, and hence so does

$$Z = \frac{(M_1 - M_2) - (p_1 - p_2)}{\sqrt{M_1(1 - M_1)/n_1 + M_2(1 - M_2)/n_2}} \tag{8.3.10}$$

1. $\mathbb{P}[-z(1 - \alpha/2) \le Z \le z(1 - \alpha/2)] \approx 1 - \alpha$. Solving for $p_1 - p_2$ gives the two-sided confidence interval.
2. $\mathbb{P}[-z(1 - \alpha) \le Z] \approx 1 - \alpha$. Solving for $p_1 - p_2$ gives the confidence upper bound.
3. $\mathbb{P}[Z \le z(1 - \alpha/2)] \approx 1 - \alpha$. Solving for $p_1 - p_2$ gives the confidence lower bound.

As always, the equal-tailed interval in (a) is not the only approximate two-sided $1 - \alpha$ confidence interval.

For α, $r \in (0, 1)$, an approximate $1 - \alpha$ confidence set for $p_1 - p_2$ is

$$\left[(M_1 - M_2) - z(1 - r\alpha)\sqrt{M_1(1 - M_1)/n_1 + M_2(1 - M_2)/n_2}\,, (M_1 - M_2) - z(\alpha - r\alpha)\sqrt{M_1(1 - M_1)/n_1 + M_2(1 - M_2)/n_2}\right] \tag{8.3.11}$$

Proof

As noted in the proof of the previous theorem,

$$Z = \frac{(M_1 - M_2) - (p_1 - p_2)}{\sqrt{M_1(1 - M_1)/n_1 + M_2(1 - M_2)/n_2}} \tag{8.3.12}$$

has approximately a standard normal distribution if n_1 and n_2 are large. Hence $\mathbb{P}[-z(\alpha - r\alpha) \le Z \le z(1 - r\alpha)] \approx 1 - \alpha$. Solving for $p_1 - p_2$ gives the two-sided confidence interval.

Conservative Confidence Intervals

Once again, $p \mapsto p(1 - p)$ is maximized when $p = \frac{1}{2}$ with maximum value $\frac{1}{4}$. We can use this to construct approximate conservative confidence intervals for $p_1 - p_2$.

For $\alpha \in (0, 1)$, the following have approximate confidence level at least $1 - \alpha$ for $p_1 - p_2$:

1. The two-sided interval with endpoints $(M_1 - M_2) \pm \frac{1}{2} z(1 - \alpha/2)\sqrt{1/n_1 + 1/n_2}$.
2. The lower bound $(M_1 - M_2) - \frac{1}{2} z(1 - \alpha)\sqrt{1/n_1 + 1/n_2}$.
3. The upper bound $(M_1 - M_2) + \frac{1}{2} z(1 - \alpha)\sqrt{1/n_1 + 1/n_2}$.

Proof

These results follow from the previous theorem by replacing $M_1(1 - M_1)$ and $M_2(1 - M_2)$ each with $\frac{1}{4}$.

Computational Exercises

In a poll 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.

Answer

$(0.396, 0.458)$

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?

Answer

0.579. No, the coin is almost certainly not fair.

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.

Answer

$(0.034, 0.116)$

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?

Answer

1068

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?

Answer

3382

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 $p = 2/\pi$, 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?

Answer

$(0.433, 0.634)$The theoretical value is approximately 0.637, which is not in the confidence interval.

A manufacturing facility has two production lines for a certain item. In a sample of 150 items from line 1, 12 are defective. From a sample of 130 items from line 2, 10 are defective. Construct the two-sided 95% confidence interval for $p_1 - p_2$, where p_i is the proportion of defective items from line i, for $i \in \{1, 2\}$

Answer

$[-0.050, 0.056]$

The vaccine for influenza is tailored each year to match the predicted dominant strain of influenza. Suppose that of 500 unvaccinated persons, 45 contracted the flu in a certain time period. Of 300 vaccinated persons, 20 contracted the flu in the same time period. Construct the two-sided 99% confidence interval for $p_1 - p_2$, where p_1 is the incidence of flu in the unvaccinated population and p_2 the incidence of flu in the vaccinated population.