Republic of Iraq Ministry of Higher Education & Research

University of Anbar

College of Education for Pure Sciences

Department of Mathematics

Lecture Note On Mathematical Statistics 2
B.Sc. in Mathematics
Fourth Stage
Assist. Prof. Dr. Feras Shaker Mahmood

CONFIDENCE INTERVALS FOR PROPORTIONS

We have suggested that the histogram is a good description of how the observations of a random sample are distributed. We might naturally inquire about the accuracy of those relative frequencies (or percentages) associated with the various classes. To illustrate, concerning the weights of $n = 40$ candy bars, we found that the relative frequency of the class interval (22.25, 23.15) was $8/40 = 0.20$, or 20%. If we think of this collection of 40 weights as a random sample observed from a larger population of candy bar weights, how close is 20% to the true percentage (or 0.20 to the true proportion) of weights in that class interval for the entire population of weights for this type of candy bar?

In considering this problem, we generalize it somewhat by treating the class interval (22.25, 23.15) as "success." That is, there is some true probability of success, p—namely, the proportion of the population in that interval. Let Y equal the frequency of measurements in the interval out of the n observations, so that (under the assumptions of independence and constant probability p) Y has the binomial distribution $b(n, p)$. Thus, the problem is to determine the accuracy of the relative frequency Y/n as an estimator of p. We solve this problem by finding, for the unknown p, a confidence interval based on Y/n.

In general, when observing n Bernoulli trials with probability p of success on each trial, we shall find a confidence interval for p based on Y/n, where Y is the number of successes and Y/n is an unbiased point estimator for p.

In Section 5.7, we noted that

$$\frac{Y - np}{\sqrt{np(1-p)}} = \frac{(Y/n) - p}{\sqrt{p(1-p)/n}}$$

has an approximate normal distribution $N(0, 1)$, provided that n is large enough. This means that, for a given probability $1-\alpha$, we can find a $z_{\alpha/2}$ such that

$$P\left[-z_{\alpha/2} \le \frac{(Y/n)-p}{\sqrt{p(1-p)/n}} \le z_{\alpha/2}\right] \approx 1-\alpha.$$

If we proceed as we did when we found a confidence interval for μ we would obtain

$$P\left[\frac{Y}{n} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \le p \le \frac{Y}{n} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right] \approx 1-\alpha.$$

Unfortunately, the unknown parameter p appears in the endpoints of this inequality. There are two ways out of this dilemma. First, we could make an additional approximation, namely, replacing p with Y/n in $p\,(1-p)/n$ in the endpoints. That is, if n is large enough, it is still true that

$$P\left[\frac{Y}{n} - z_{\alpha/2}\sqrt{\frac{(Y/n)(1-Y/n)}{n}} \le p \le \frac{Y}{n} + z_{\alpha/2}\sqrt{\frac{(Y/n)(1-Y/n)}{n}}\right] \approx 1-\alpha.$$

Thus, for large n, if the observed Y equals y, then the interval

$$\left[\frac{y}{n} - z_{\alpha/2}\sqrt{\frac{(y/n)(1-y/n)}{n}}, \frac{y}{n} + z_{\alpha/2}\sqrt{\frac{(y/n)(1-y/n)}{n}}\right]$$

serves as an approximate $100(1-\alpha)\%$ confidence interval for p. Frequently, this interval is written as

$$\frac{y}{n} \pm z_{\alpha/2}\sqrt{\frac{(y/n)(1-y/n)}{n}}$$

for brevity. This formulation clearly notes, as does $\overline{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})$ the reliability of the estimate y/n, namely, that we are $100(1-\alpha)\%$ confident that p is within $z_{\alpha/2}\sqrt{(y/n)(1-y/n)/n}$ of $\widehat{p} = y/n$.

A second way to solve for p in the inequality in Equation 7.3-1 is to note that

$$\frac{|Y/n - p|}{\sqrt{p(1-p)/n}} \le z_{\alpha/2}$$

is equivalent to

$$H(p) = \left(\frac{Y}{n} - p\right)^2 - \frac{z_{\alpha/2}^2\, p(1-p)}{n} \le 0.$$

But $H(p)$ is a quadratic expression in p. Thus, we can find those values of p for which $H(p) \le 0$ by finding the two zeros of $H(p)$. Letting $\widehat{p} = Y/n$ and $z_0 = z_{\alpha/2}$ in

$$H(p) = \left(1 + \frac{z_0^2}{n}\right)p^2 - \left(2\widehat{p} + \frac{z_0^2}{n}\right)p + \widehat{p}^{\,2}.$$

By the quadratic formula, the zeros of $H(p)$ are, after simplifications,

$$\frac{\widehat{p} + z_0^2/(2n) \pm z_0\sqrt{\widehat{p}\,(1-\widehat{p}\,)/n + z_0^2/(4n^2)}}{1 + z_0^2/n},$$

and these zeros give the endpoints for an approximate $100(1-\alpha)\%$ confidence interval for p. If n is large, $z_0^2/(2n)$, $z_0^2/(4n^2)$, and z_0^2/n are small. Thus, the confidence intervals given by Equations 7.3-2 and 7.3-4 are approximately equal when n is large.

Example

Let us return to the example of the histogram of the candy bar weights, with $n = 40$ and $y/n = 8/40 = 0.20$. If $1-\alpha = 0.90$, so that $z_{\alpha/2} = 1.645$, then, using Equation 7.3-2, we find that the endpoints

$$0.20 \pm 1.645\sqrt{\frac{(0.20)(0.80)}{40}}$$

serve as an approximate 90% confidence interval for the true fraction p. That is, $[0.096, 0.304]$, which is the same as $[9.6\%, 30.4\%]$, is an approximate 90% confidence interval for the percentage of weights of the entire population in the interval $(22.25, 23.15)$. If we had used the endpoints , the confidence interval would be $[0.117, 0.321]$. Because of the small sample size, there is a nonnegligible difference in these intervals. If the sample size had been $n = 400$ and $y = 80$, so that $y/n = 80/400 = 0.20$, the two 90% confidence intervals would have been $[0.167, 0.233]$ and $[0.169, 0.235]$, respectively, which differ very little.

Example

A possible gubernatorial candidate wants to assess initial support among the voters before making an announcement about her candidacy. If the fraction p of voters who are favorable, without any advance publicity, is around 0.15, the candidate will enter the race. From a poll of n voters selected at random, the candidate would like the estimate y/n to be within 0.03 of p. That is, the decision will be based on a 95% confidence interval of the form $y/n \pm 0.03$. Since the candidate has no idea about the magnitude of p, a consulting statistician formulates the equation

$$n = \frac{(1.96)^2}{4(0.03)^2} = 1067.11.$$

Thus, the sample size should be around 1068 to achieve the desired reliability and accuracy. Suppose that 1068 voters around the state were selected at random and interviewed and $y = 214$ express support for the candidate. Then $\widehat{p} = 214/1068 = 0.20$ is a point estimate of p, and an approximate 95% confidence interval for p is

$$0.20 \pm 1.96\sqrt{(0.20)(0.80)/n}, \qquad \text{or} \qquad 0.20 \pm 0.024.$$

That is, we are 95% confident that p belongs to the interval $[0.176, 0.224]$. On the basis of this sample, the candidate decided to run for office. Note that, for a confidence coefficient of 95%, we found a sample size so that the maximum error of the estimate would be 0.03. From the data that were collected, the maximum error of the estimate is only 0.024. We ended up with a smaller error because we found the sample size assuming that $p = 0.50$, while, in fact, p is closer to 0.20. ■

Suppose that you want to estimate the proportion p of a student body that favors a new policy. How large should the sample be? If p is close to 1/2 and you want to be 95% confident that the maximum error of the estimate is $\varepsilon = 0.02$, then

$$n = \frac{(1.96)^2}{4(0.02)^2} = 2401.$$

Such a sample size makes sense at a large university. However, if you are a student at a small college, the entire enrollment could be less than 2401. Thus, we now give a procedure that can be used to determine the sample size when the population is not so large relative to the desired sample size.

Let N equal the size of a population, and assume that N_1 individuals in the population have a certain characteristic C (e.g., favor a new policy). Let $p = N_1/N$, the proportion with this characteristic. Then $1-p = 1-N_1/N$. If we take a sample of size n without replacement, then X, the number of observations with the characteristic C, has a hypergeometric distribution. The mean and variance of X are, respectively,

$$\mu = n\left(\frac{N_1}{N}\right) = np$$

and

$$\sigma^2 = n\left(\frac{N_1}{N}\right)\left(1 - \frac{N_1}{N}\right)\left(\frac{N-n}{N-1}\right) = np(1-p)\left(\frac{N-n}{N-1}\right).$$

The mean and variance of X/n are, respectively,

$$E\left(\frac{X}{n}\right) = \frac{\mu}{n} = p$$

and

$$\mathrm{Var}\left(\frac{X}{n}\right) = \frac{\sigma^2}{n^2} = \frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right).$$

To find an approximate confidence interval for p, we can use the normal approximation:

$$P\left[-z_{\alpha/2} \leq \frac{\dfrac{X}{n} - p}{\sqrt{\dfrac{p(1-p)}{n}\left(\dfrac{N-n}{N-1}\right)}} \leq z_{\alpha/2}\right] \approx 1 - \alpha.$$

Thus,

$$1 - \alpha \approx P\left[\frac{X}{n} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)} \leq p \leq \frac{X}{n} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)}\right].$$

Replacing p under the radical with $\widehat{p} = x/n$, we find that an approximate $1 - \alpha$ confidence interval for p is

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}\,(1-\widehat{p}\,)}{n}\left(\frac{N-n}{N-1}\right)}.$$

This is similar to the confidence interval for p when the distribution of X is $b(n, p)$. If N is large relative to n, then

$$\frac{N-n}{N-1} = \frac{1-n/N}{1-1/N} \approx 1,$$

so in this case the two intervals are essentially equal.

Suppose now that we are interested in determining the sample size n that is required to have $1-\alpha$ confidence that the maximum error of the estimate of p is ε. We let

$$\varepsilon = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)}$$

and solve for n. After some simplification, we obtain

$$\begin{aligned} n &= \frac{Nz_{\alpha/2}^2 p(1-p)}{(N-1)\varepsilon^2 + z_{\alpha/2}^2 p(1-p)} \\ &= \frac{z_{\alpha/2}^2 p(1-p)/\varepsilon^2}{\dfrac{N-1}{N} + \dfrac{z_{\alpha/2}^2 p(1-p)/\varepsilon^2}{N}}. \end{aligned}$$

If we let

$$m = \frac{z_{\alpha/2}^2 p^*(1-p^*)}{\varepsilon^2},$$

which is the n value , then we choose

$$n = \frac{m}{1 + \dfrac{m-1}{N}}$$

for our sample size n.

If we know nothing about p, we set $p^* = 1/2$ to determine m. For example, if the size of the student body is $N = 4000$ and $1 - \alpha = 0.95$, $\varepsilon = 0.02$, and we let $p^* = 1/2$, then $m = 2401$ and

$$n = \frac{2401}{1 + 2400/4000} = 1501,$$

rounded up to the nearest integer. Thus, we would sample approximately 37.5% of the student body.

Example

Suppose that a college of $N = 3000$ students is interested in assessing student support for a new form for teacher evaluation. To estimate the proportion p in favor of the new form, how large a sample is required so that the maximum error of the estimate of p is $\varepsilon = 0.03$ with 95% confidence? If we assume that p is completely unknown, we use $p^* = 1/2$ to obtain

$$m = \frac{(1.96)^2}{4(0.03)^2} = 1068,$$

rounded up to the nearest integer. Thus, the desired sample size is

$$n = \frac{1068}{1 + 1067/3000} = 788,$$

rounded up to the nearest integer.

Thanking for your Intention

Ass. Prof. Dr. Feras Shaker Mahmood