24

More on confidence intervals

While in Chapter 23 we were solely concerned with confidence intervals for expectations, in this chapter we treat a variety of topics. First, we focus on confidence intervals for the parameter p of the binomial distribution. Then, based on an example, we briefly discuss a general method to construct confidence intervals. One-sided confidence intervals, or upper and lower confidence bounds, are discussed next. At the end of the chapter we investigate the question of how to determine the sample size when a confidence interval of a certain width is desired.

24.1 The probability of success

A common situation is that we observe a random variable X with a $Bin(n, p)$ distribution and use X to estimate p. For example, if we want to estimate the proportion of voters that support candidate G in an election, we take a sample from the voter population and determine the proportion in the sample that supports G. If n individuals are selected at random from the population, where a proportion p supports candidate G, the number of supporters X in the sample is modeled by a $Bin(n, p)$ distribution; we count the supporters of candidate G as "successes." Usually, the sample proportion X/n is taken as an estimator for p.

If we want to make a confidence interval for p, based on the number of successes X in the sample, we need to find statistics L and U (see the definition of confidence intervals on page 343) such that

$$\mathrm{P}(L < p < U) = 1 - \alpha,$$

where L and U are to be based on X only. In general, this problem does not have a solution. However, the method for large n described next, sometimes called "the Wilson method" (see [40]), yields confidence intervals with confidence level approximately $100(1-\alpha)\%$. (How close the true confidence level is to $100(1-\alpha)\%$ depends on the (unknown) p, though it is known that for p near 0 and 1 it is too low. For some details and an alternative for this situation, see Remark 24.1.)

Recall the normal approximation to the binomial distribution, a consequence of the central limit theorem (see page 201 and Exercise 14.5): for large n, the distribution of X is approximately normal and

$$\frac{X - np}{\sqrt{np(1-p)}}$$

is approximately standard normal. By dividing by n in both the numerator and the denominator, we see that this equals:

$$\frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}}.$$

Therefore, for large n

$$\mathrm{P}\left(-z_{\alpha/2} < \frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\alpha/2} \right) \approx 1 - \alpha.$$

Note that the event

$$-z_{\alpha/2} < \frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\alpha/2}$$

is the same as

$$\left(\frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} \right)^2 < \left(z_{\alpha/2}\right)^2$$

or

$$\left(\frac{X}{n} - p\right)^2 - \left(z_{\alpha/2}\right)^2 \frac{p(1-p)}{n} < 0.$$

To derive expressions for L and U we can rewrite the inequality in this statement to obtain the form $L < p < U$, but the resulting formulas are rather awkward. To obtain the confidence interval, we instead substitute the data values directly and then solve for p, which yields the desired result.

Suppose, in a sample of 125 voters, 78 support one candidate. What is the 95% confidence interval for the population proportion p supporting that candidate? The realization of X is $x = 78$ and $n = 125$. We substitute this, together with $z_{\alpha/2} = z_{0.025} = 1.96$, in the last inequality:

$$\left(\frac{78}{125} - p\right)^2 - \frac{(1.96)^2}{125} p(1-p) < 0,$$

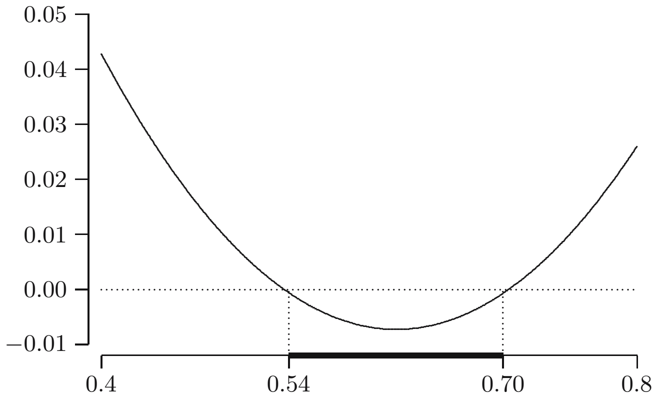

Fig. 24.1. The parabola $1.0307\,p^2 - 1.2787\,p + 0.3894$ and the resulting confidence interval.

or, working out squares and products and grouping terms:

$$1.0307\,p^2 - 1.2787\,p + 0.3894 < 0.$$

This quadratic form describes a parabola, which is depicted in Figure 24.1. Also, for other values of n and x there always results a quadratic inequality like this, with a positive coefficient for p^2 and a similar picture. For the confidence interval we need to find the values where the parabola intersects the horizontal axis. The solutions we find are:

$$p_{1,2} = \frac{-(-1.2787) \pm \sqrt{(-1.2787)^2 - 4 \cdot 1.0307 \cdot 0.3894}}{2 \cdot 1.0307} = 0.6203 \pm 0.0835;$$

hence, $l = 0.54$ and $u = 0.70$, so the resulting confidence interval is $(0.54, 0.70)$.

Quick exercise 24.1 Suppose in another election we find 80 supporters in a sample of 200. Suppose we use $\alpha = 0.0456$ for which $z_{\alpha/2} = 2$. Construct the corresponding confidence interval for p.

Remark 24.1 (Coverage probabilities and an alternative method). Because of the discrete nature of the binomial distribution, the probability that the confidence interval covers the true parameter value depends on p. As a function of p it typically oscillates in a sawtooth-like manner around $1 - \alpha$, being too high for some values and too low for others. This is something that cannot be escaped from; the phenomenon is present in every method. In an average sense, the method treated in the text yields coverage probabilities close to $1 - \alpha$, though for arbitrarily high values of n it is possible to find p's for which the actual coverage is several percentage points too low. The low coverage occurs for p's near 0 and 1.

An alternative is the method proposed by Agresti and Coull, which overall is more conservative than the Wilson method (in fact, the Agresti-Coull interval contains the Wilson interval as a proper subset). Especially for p near 0 or 1 this method yields conservative confidence intervals. Define

$$\tilde{X} = X + \frac{(z_{\alpha/2})^2}{2} \quad \text{and} \quad \tilde{n} = n + (z_{\alpha/2})^2,$$

and $\tilde{p} = \tilde{X}/\tilde{n}$. The approximate $100(1-\alpha)\%$ confidence interval is then given by

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

For a clear survey paper on confidence intervals for p we recommend Brown et al. [4].

24.2 Is there a general method?

We have now seen a number of examples of confidence intervals, and while it should be clear to you that in each of these cases the resulting intervals *are* valid confidence intervals, you may wonder how we go about finding confidence intervals in new situations. One could ask: is there a general method? We first consider an example.

A confidence interval for the minimum lifetime

Suppose we have a random sample $X_1, \ldots, X_n$ from a *shifted* exponential distribution, that is, $X_i = \delta + Y_i$, where $Y_1, \ldots, Y_n$ are a random sample from an $Exp(1)$ distribution. This type of random variable is sometimes used to model lifetimes; a minimum lifetime is guaranteed, but otherwise the lifetime has an exponential distribution. The unknown parameter δ represents the minimum lifetime, and the probability density of the X_i is positive only for values greater than δ.

To derive information about δ it is natural to use the smallest observed value $T = \min\{X_1, \ldots, X_n\}$. This is also the maximum likelihood estimator for δ; see Exercise 21.6. Writing

$$T = \min\{\delta + Y_1, \ldots, \delta + Y_n\} = \delta + \min\{Y_1, \ldots, Y_n\}$$

and observing that $M = \min\{Y_1, \ldots, Y_n\}$ has an $Exp(n)$ distribution (see Exercise 8.18), we find for the distribution function of T: $F_T(a) = 0$ for $a < \delta$ and

$$\begin{aligned} F_T(a) = \mathrm{P}(T \le a) = \mathrm{P}(\delta + M \le a) &= \mathrm{P}(M \le a - \delta) \\ &= 1 - \mathrm{e}^{-n(a-\delta)} \quad \text{for } a \ge \delta. \end{aligned} \tag{24.1}$$

Next, we solve

$$\mathrm{P}(c_l < T < c_u) = 1 - \alpha$$

by requiring

$$\mathrm{P}(T \leq c_l) = \mathrm{P}(T \geq c_u) = \tfrac{1}{2}\alpha.$$

Using (24.1) we find the following equations:

$$1 - \mathrm{e}^{-n(c_l - \delta)} = \tfrac{1}{2}\alpha \quad \text{and} \quad \mathrm{e}^{-n(c_u - \delta)} = \tfrac{1}{2}\alpha$$

whose solutions are

$$c_l = \delta - \frac{1}{n}\ln\left(1 - \tfrac{1}{2}\alpha\right) \quad \text{and} \quad c_u = \delta - \frac{1}{n}\ln\left(\tfrac{1}{2}\alpha\right).$$

Both c_l and c_u are values larger than δ, because the logarithms are negative. We have found that, whatever the value of δ:

$$\mathrm{P}\left(\delta - \frac{1}{n}\ln\left(1 - \tfrac{1}{2}\alpha\right) < T < \delta - \frac{1}{n}\ln\left(\tfrac{1}{2}\alpha\right)\right) = 1 - \alpha.$$

By rearranging the inequalities, we see this is equivalent to

$$\mathrm{P}\left(T + \frac{1}{n}\ln\left(\tfrac{1}{2}\alpha\right) < \delta < T + \frac{1}{n}\ln\left(1 - \tfrac{1}{2}\alpha\right)\right) = 1 - \alpha,$$

and therefore a $100(1-\alpha)\%$ confidence interval for δ is given by

$$\left(t + \frac{1}{n}\ln\left(\tfrac{1}{2}\alpha\right),\, t + \frac{1}{n}\ln\left(1 - \tfrac{1}{2}\alpha\right)\right). \tag{24.2}$$

For $\alpha = 0.05$ this becomes:

$$\left(t - \frac{3.69}{n},\, t - \frac{0.0253}{n}\right).$$

Quick exercise 24.2 Suppose you have a dataset of size 15 from a shifted *Exp*(1) distribution, whose minimum value is 23.5. What is the 99% confidence interval for δ?

Looking back at the example, we see that the confidence interval could be constructed because we know that $T - \delta = M$ has an exponential distribution. There are many more examples of this type: some function $g(T, \theta)$ of a sample statistic T and the unknown parameter θ has a known distribution. However, this still does not cover all the ways to construct confidence intervals (see also the following remark).

Remark 24.2 (About a general method). Suppose $X_1, \ldots, X_n$ is a random sample from some distribution depending on some unknown parameter θ and let T be a sample statistic. One possible choice is to select a T that is an estimator for θ, but this is not necessary. In each case, the distribution of T depends on θ, just as that of $X_1, \ldots, X_n$ does. In some cases it might be possible to find functions $g(\theta)$ and $h(\theta)$ such that

$$\mathrm{P}(g(\theta) < T < h(\theta)) = 1 - \alpha \quad \text{for every value of } \theta. \tag{24.3}$$

If this is so, then confidence statements about θ can be made. In more special cases, for example if g and h are strictly increasing, the inequalities $g(\theta) < T < h(\theta)$ can be rewritten as

$$h^{-1}(T) < \theta < g^{-1}(T),$$

and then (24.3) is equivalent to

$$\mathrm{P}\big(h^{-1}(T) < \theta < g^{-1}(T)\big) = 1 - \alpha \quad \text{for every value of } \theta.$$

Checking with the confidence interval definition, we see that the last statement implies that $(h^{-1}(t), g^{-1}(t))$ is a $100(1-\alpha)\%$ confidence interval for θ.

24.3 One-sided confidence intervals

Suppose you are in charge of a power plant that generates and sells electricity, and you are about to buy a shipment of coal, say a shipment of the Daw Mill coal identified as 258GB41 earlier. You plan to buy the shipment if you are confident that the gross calorific content exceeds 31.00 MJ/kg. At the end of Section 23.2 we obtained for the gross calorific content the 95% confidence interval $(30.946, 31.067)$: based on the data we are 95% confident that the gross calorific content is higher than 30.946 and lower than 31.067.

In the present situation, however, we are *only* interested in the lower bound: we would prefer a confidence statement of the type "we are 95% confident that the gross calorific content exceeds 31.00." Modifying equation (23.4) we find

$$\mathrm{P}\left(\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} < t_{n-1,\alpha}\right) = 1 - \alpha,$$

which is equivalent to

$$\mathrm{P}\left(\bar{X}_n - t_{n-1,\alpha}\frac{S_n}{\sqrt{n}} < \mu\right) = 1 - \alpha.$$

We conclude that

$$\left(\bar{x}_n - t_{n-1,\alpha}\frac{s_n}{\sqrt{n}}, \infty\right)$$

is a $100(1-\alpha)\%$ *one-sided confidence interval* for μ. For the Daw Mill coal, using $\alpha = 0.05$, with $t_{21,0.05} = 1.721$ this results in:

$$\left(31.012 - 1.721\frac{0.1294}{\sqrt{22}}, \infty\right) = (30.964, \infty).$$

We see that because "all uncertainty may be put on one side," the lower bound in the one-sided interval is higher than that in the two-sided one, though still below 31.00. Other situations may require a confidence *upper bound*. For example, if the calorific value is below a certain number you can try to negotiate a lower the price.

The definition of confidence intervals (page 343) can be extended to include one-sided confidence intervals as well. If we have a sample statistic L_n such that

$$\mathrm{P}(L_n < \theta) = \gamma$$

for every value of the parameter of interest θ, then

$$(l_n, \infty)$$

is called a $100\gamma\%$ *one-sided confidence interval for* θ. The number l_n is sometimes called a $100\gamma\%$ *lower confidence bound for* θ. Similary, U_n with $\mathrm{P}(\theta < U_n) = \gamma$ for every value of θ, yields the one-sided confidence interval $(-\infty, u_n)$, and u_n is called a $100\gamma\%$ *upper confidence bound.*

Quick exercise 24.3 Determine the 99% upper confidence bound for the gross calorific value of the Daw Mill coal.

24.4 Determining the sample size

The narrower the confidence interval the better (why?). As a general principle, we know that more accurate statements can be made if we have more measurements. Sometimes, an accuracy requirement is set, even before data are collected, and the corresponding sample size is to be determined. We provide an example of how to do this and note that this generally can be done, but the actual computation varies with the type of confidence interval.

Consider the question of the calorific content of coal once more. We have a shipment of coal to test and we want to obtain a 95% confidence interval, but it should not be wider than 0.05 MJ/kg, i.e., the lower and upper bound should not differ more than 0.05. How many measurements do we need?

We answer this question for the case when ISO method 1928 is used, whence we may assume that measurements are normally distributed with standard deviation $\sigma = 0.1$. When the desired confidence level is $1 - \alpha$, the width of the confidence interval will be

$$2 \cdot z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

Requiring that this is at most w means finding the smallest n that satisfies

$$2 z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq w$$

or

$$n \geq \left(\frac{2z_{\alpha/2}\sigma}{w}\right)^2.$$

For the example: $w = 0.05$, $\sigma = 0.1$, and $z_{0.025} = 1.96$; so

$$n \geq \left(\frac{2 \cdot 1.96 \cdot 0.1}{0.05}\right)^2 = 61.4,$$

that is, we should perform at least 62 measurements.

In case σ is unknown, we somehow have to estimate it, and then the method can only give an indication of the required sample size. The standard deviation as we (afterwards) estimate it from the data may turn out to be quite different, and the obtained confidence interval may be smaller or larger than intended.

QUICK EXERCISE 24.4 What is the required sample size if we want the 99% confidence interval to be 0.05 MJ/kg wide?

24.5 Solutions to the quick exercises

24.1 We need to solve

$$\left(\frac{80}{200} - p\right)^2 - \frac{(2)^2}{200}\, p(1-p) < 0, \quad \text{or} \quad 1.02\, p^2 - 0.82p + 0.16 < 0.$$

The solutions are:

$$p_{1,2} = \frac{-(-0.82) \pm \sqrt{(-0.82)^2 - 4 \cdot 1.02 \cdot 0.16}}{2 \cdot 1.02} = 0.4020 \pm 0.0686,$$

so the confidence interval is $(0.33, 0.47)$.

24.2 We should substitute $n = 15$, $t = 23.5$, and $\alpha = 0.01$ into:

$$\left(t + \frac{1}{n} \ln\left(\tfrac{1}{2}\alpha\right),\ t + \frac{1}{n} \ln\left(1 - \tfrac{1}{2}\alpha\right)\right),$$

which yields

$$\left(23.5 - \frac{5.30}{15},\ 23.5 - \frac{0.0050}{15}\right) = (23.1467,\ 23.4997).$$

24.3 The upper confidence bound is given by

$$u_n = \bar{x}_n + t_{21,0.01} \frac{s_n}{\sqrt{22}},$$

where $\bar{x}_n = 31.012$, $t_{21,0.01} = 2.518$, and $s_n = 0.1294$. Substitution yields $u_n = 31.081$.

24.4 The confidence level changes to 99%, so we use $z_{0.005} = 2.576$ instead of 1.96 in the computation:

$$n \geq \left(\frac{2 \cdot 2.576 \cdot 0.1}{0.05} \right)^2 = 106.2,$$

so we need at least 107 measurements.

24.6 Exercises

24.1 ⊡ Of a series of 100 (independent and identical) chemical experiments, 70 were concluded succesfully. Construct a 90% confidence interval for the success probability of this type of experiment.

24.2 In January 2002 the Euro was introduced and soon after stories started to circulate that some of the Euro coins would not be fair coins, because the "national side" of some coins would be too heavy or too light (see, for example, the *New Scientist* of January 4, 2002, but also national newspapers of that date).

a. A French 1 Euro coin was tossed six times, resulting in 1 heads and 5 tails. Is it reasonable to use the Wilson method, introduced in Section 24.1, to construct a confidence interval for p?

b. A Belgian 1 Euro coin was tossed 250 times: 140 heads and 110 tails. Construct a 95% confidence interval for the probability of getting heads with this coin.

24.3 In Exercise 23.1, what sample size is needed if we want a 99% confidence interval for μ at most 1 ml wide?

24.4 ⊡ Recall Exercise 23.3 and the 10 bags of cement that should each weigh 94 kg. The average weight was 93.5 kg, with sample standard deviation 0.75.

a. Based on these data, how many bags would you need to sample to make a 90% confidence interval that is 0.1 kg wide?

b. Suppose you actually do measure the required number of bags and construct a new confidence interval. Is it guaranteed to be at most 0.1 kg wide?

24.5 Suppose we want to make a 95% confidence interval for the probability of getting heads with a Dutch 1 Euro coin, and it should be at most 0.01 wide. To determine the required sample size, we note that the probability of getting heads is about 0.5. Furthermore, if X has a $Bin(n, p)$ distribution, with n large and $p \approx 0.5$, then

$$\frac{X - np}{\sqrt{n/4}} \quad \text{is approximately standard normal.}$$

a. Use this statement to derive that the width of the 95% confidence interval for p is approximately

$$\frac{z_{0.025}}{\sqrt{n}}.$$

Use this width to determine how large n should be.

b. The coin is thrown the number of times just computed, resulting in 19 477 times heads. Construct the 95% confidence interval and check whether the required accuracy is attained.

24.6 ⊞ Environmentalists have taken 16 samples from the wastewater of a chemical plant and measured the concentration of a certain carcinogenic substance. They found $\bar{x}_{16} = 2.24$ (ppm) and $s_{16}^2 = 1.12$, and want to use these data in a lawsuit against the plant. It may be assumed that the data are a realization of a normal random sample.

a. Construct the 97.5% one-sided confidence interval that the environmentalists made to convince the judge that the concentration exceeds legal limits.

b. The plant management uses the same data to construct a 97.5% one-sided confidence interval to show that concentrations are not too high. Construct this interval as well.

24.7 Consider once more the Rutherford-Geiger data as given in Section 23.4. Knowing that the number of α-particle emissions during an interval has a Poisson distribution, we may see the data as observations from a $Pois(\mu)$ distribution. The central limit theorem tells us that the average $\bar{X}_n$ of a large number of independent $Pois(\mu)$ approximately has a normal distribution and hence that

$$\frac{\bar{X}_n - \mu}{\sqrt{\mu}/\sqrt{n}}$$

has a distribution that is approximately $N(0, 1)$.

a. Show that the large sample 95% confidence interval contains those values of μ for which

$$(\bar{x}_n - \mu)^2 \le (1.96)^2 \frac{\mu}{n}.$$

b. Use the result from **a** to construct the large sample 95% confidence interval based on the Rutherford-Geiger data.

c. Compare the result with that of Exercise 23.9 **b**. Is this surprising?

24.8 ⊡ Recall Exercise 23.5 about the 1500 m speed-skating results in the 2002 Winter Olympic Games. If there were no outer lane advantage, the number out of the 23 completed races won by skaters starting in the outer lane would have a $Bin(23, p)$ distribution with $p = 1/2$, because of the lane assignment by lottery.

a. Of the 23 races, 15 were won by the skater starting in the outer lane. Use this information to construct a 95% confidence interval for p by means of the Wilson method. If you think that $n = 23$ is probably too small to use a method based on the central limit theorem, we agree. We should be careful with conclusions we draw from this confidence interval.

b. The question posed earlier "Is there an outer lane advantage?" implies that a one-sided confidence interval is more suitable. Construct the appropriate 95% one-sided confidence interval for p by first constructing a 90% two-sided confidence interval.

24.9 ⊞ Suppose we have a dataset $x_1, \ldots, x_{12}$ that may be modeled as the realization of a random sample $X_1, \ldots, X_{12}$ from a $U(0, \theta)$ distribution, with θ unknown. Let $M = \max\{X_1, \ldots, X_{12}\}$.

a. Show that for $0 \leq t \leq 1$

$$\mathrm{P}\left(\frac{M}{\theta} \leq t\right) = t^{12}.$$

b. Use $\alpha = 0.1$ and solve

$$\mathrm{P}\left(\frac{M}{\theta} \leq c_l\right) = \mathrm{P}\left(\frac{M}{\theta} \leq c_u\right) = \tfrac{1}{2}\alpha.$$

c. Suppose the realization of M is $m = 3$. Construct the 90% confidence interval for θ.

d. Derive the general expression for a confidence interval of level $1 - \alpha$ based on a sample of size n.

24.10 Suppose we have a dataset $x_1, \ldots, x_n$ that may be modeled as the realization of a random sample $X_1, \ldots, X_n$ from an $Exp(\lambda)$ distribution, where λ is unknown. Let $S_n = X_1 + \cdots + X_n$.

a. Check that λS_n has a $Gam(n, 1)$ distribution.

b. The following quantiles of the $Gam(20, 1)$ distribution are given: $q_{0.05} = 13.25$ and $q_{0.95} = 27.88$. Use these to construct a 90% confidence interval for λ when $n = 20$.