11

Bootstrap Confidence Intervals

The jackknife and the bootstrap represent a different use of modern computer power: rather than extending classical methodology—from ordinary least squares to generalized linear models, for example—they extend the reach of classical inference.

Chapter 10 focused on standard errors. Here we will take up a more ambitious inferential goal, the bootstrap automation of confidence intervals. The familiar *standard intervals*

$$\hat{\theta} \pm 1.96\,\widehat{\text{se}}, \tag{11.1}$$

for approximate 95% coverage, are immensely useful in practice but often not very accurate. If we observe $\hat{\theta} = 10$ from a Poisson model $\hat{\theta} \sim \text{Poi}(\theta)$, the standard 95% interval (3.8, 16.2) (using $\widehat{\text{se}} = \hat{\theta}^{1/2}$) is a mediocre approximation to the exact interval[1]

$$(5.1, 17.8). \tag{11.2}$$

Standard intervals (11.1) are symmetric around $\hat{\theta}$, this being their main weakness. Poisson distributions grow more variable as θ increases, which is why interval (11.2) extends farther to the right of $\hat{\theta} = 10$ than to the left. Correctly capturing such effects in an automatic way is the goal of bootstrap confidence interval theory.

11.1 Neyman's Construction for One-Parameter Problems

The student score data of Table 3.1 comprised $n = 22$ pairs,

$$x_i = (m_i, v_i), \qquad i = 1, 2, \ldots, 22, \tag{11.3}$$

[1] Using the Neyman construction of Section 11.1, as explained there; see also Table 11.2 in Section 11.4.

where m_i and v_i were student i's scores on the "mechanics" and "vectors" tests. The sample correlation coefficient $\hat{\theta}$ between m_i and v_i was computed to be

$$\hat{\theta} = 0.498. \tag{11.4}$$

Question: What can we infer about the true correlation θ between m and v? Figure 3.2 displayed three possible Bayesian answers. Confidence intervals provide the frequentist solution, by far the most popular in applied practice.

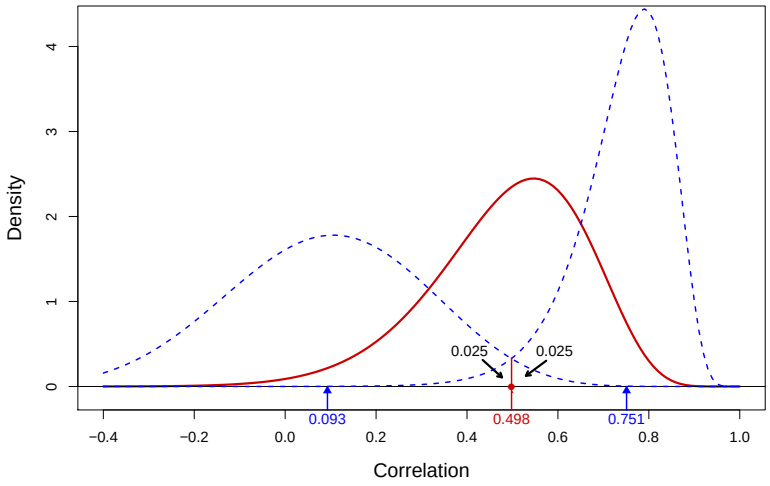

Figure 11.1 The solid curve is the normal correlation coefficient density $f_{\hat{\theta}}(r)$ (3.11) for $\hat{\theta} = 0.498$, the MLE estimate for the student score data; $\hat{\theta}(\text{lo}) = 0.093$ and $\hat{\theta}(\text{up}) = 0.751$ are the endpoints of the 95% confidence interval for θ, with corresponding densities shown by dashed curves. These yield tail areas 0.025 at $\hat{\theta}$ (11.6).

Suppose, first, that we assume a bivariate normal model (5.12) for the pairs (m_i, v_i). In that case the probability density $f_\theta(\hat{\theta})$ for sample correlation $\hat{\theta}$ given true correlation θ has known form (3.11). The solid curve in Figure 11.1 graphs f for $\theta = 0.498$, that is, for θ set equal to the observed value $\hat{\theta}$. In more careful notation, the curve graphs $f_{\hat{\theta}}(r)$ as a function of the dummy variable[2] r taking values in $[-1, 1]$.

[2] This is an example of a parametric bootstrap distribution (10.49), here with $\hat{\mu}$ being $\hat{\theta}$.

Two other curves $f_\theta(r)$ appear in Figure 11.1: for θ equaling

$$\hat{\theta}(\text{lo}) = 0.093 \quad \text{and} \quad \hat{\theta}(\text{up}) = 0.751. \tag{11.5}$$

These were numerically calculated as the solutions to

$$\int_{\hat{\theta}}^{1} f_{\hat{\theta}(\text{lo})}(r)\, dr = 0.025 \quad \text{and} \quad \int_{-1}^{\hat{\theta}} f_{\hat{\theta}(\text{up})}(r)\, dr = 0.025. \tag{11.6}$$

In words, $\hat{\theta}(\text{lo})$ is the smallest value of θ putting probability at least 0.025 above $\hat{\theta} = 0.498$, while $\hat{\theta}(\text{up})$ is the largest value with probability at least 0.025 below $\hat{\theta}$;

$$\theta \in \left[\hat{\theta}(\text{lo}), \hat{\theta}(\text{up})\right] \tag{11.7}$$

is a 95% confidence interval for the true correlation, statement (11.7) holding true with probability 0.95, for every possible value of θ.

We have just described *Neyman's construction* of confidence intervals for one-parameter problems $f_\theta(\hat{\theta})$. (Later we will consider the more difficult situation where there are "nuisance parameters" in addition to the parameter of interest θ.) One of the jewels of classical frequentist inference, it depends on a *pivotal argument*—"ingenious device" number 5 of Section 2.1—to show that it produces genuine confidence intervals, i.e., ones that contain the true parameter value θ at the claimed probability level, 0.95 in Figure 11.1. The argument appears in the chapter endnotes.[†] †1

For the Poisson calculation (11.2) it was necessary to define exactly what "the smallest value of θ putting probability at least 0.025 above $\hat{\theta}$" meant. This was done assuming that, for any θ, half of the probability $f_\theta(\hat{\theta})$ at $\hat{\theta} = 10$ counted as "above," and similarly for calculating the upper limit.

Transformation Invariance

Confidence intervals enjoy the important and useful property of transformation invariance. In the Poisson example (11.2), suppose our interest shifts from parameter θ to parameter $\phi = \log\theta$. The 95% exact interval (11.2) for θ then transforms to the exact 95% interval for ϕ simply by taking logs of the endpoints,

$$(\log(5.1), \log(17.8)) = (1.63, 2.88). \tag{11.8}$$

To state things generally, suppose we observe $\hat{\theta}$ from a family of densities $f_\theta(\hat{\theta})$ and construct a confidence interval $\mathcal{C}(\hat{\theta})$ for θ of coverage level

α ($\alpha = 0.95$ in our examples). Now let parameter ϕ be a monotonic increasing function of θ, say

$$\phi = m(\theta) \tag{11.9}$$

($m(\theta) = \log\theta$ in (11.8)), and likewise $\hat{\phi} = m(\hat{\theta})$ for the point estimate. Then $\mathcal{C}(\hat{\theta})$ maps point by point into $\mathcal{C}^{\phi}(\hat{\phi})$, a level-$\alpha$ confidence interval for ϕ,

$$\mathcal{C}^{\phi}(\hat{\phi}) = \left\{\phi = m(\theta) \text{ for } \theta \in \mathcal{C}\left(\hat{\theta}\right)\right\}. \tag{11.10}$$

This just says that the event $\{\theta \in \mathcal{C}(\hat{\theta})\}$ is the same as the event $\{\phi \in \mathcal{C}^{\phi}(\hat{\phi})\}$, so if the former always occurs with probability α then so must the latter.

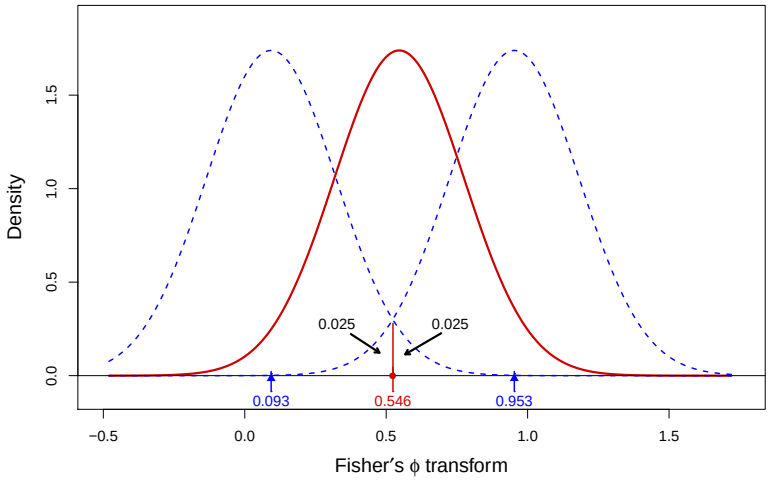

Figure 11.2 The situation in Figure 11.1 after transformation to $\phi = m(\theta)$ according to (11.11). The curves are nearly $N(\phi, \sigma^2)$ with standard deviation $\sigma = 1/\sqrt{19} = 0.229$.

Transformation invariance has an historical resonance with the normal correlation coefficient. Fisher's derivation of $f_{\theta}(\hat{\theta})$ (3.11) in 1915 was a mathematical triumph, but a difficult one to exploit in an era of mechanical computation. Most ingeniously, Fisher suggested instead working with the transformed parameter $\phi = m(\theta)$ where

$$\phi = m(\theta) = \frac{1}{2}\log\left(\frac{1+\theta}{1-\theta}\right), \tag{11.11}$$

and likewise with statistic $\hat{\phi} = m(\hat{\theta})$. Then, to a surprisingly good approximation,

$$\hat{\phi} \doteq \mathcal{N}\left(\phi, \frac{1}{n-3}\right). \tag{11.12}$$

See Figure 11.2, which shows Neyman's construction on the ϕ scale.

In other words, we are back in Fisher's favored situation (4.31), the simple normal translation problem, where

$$\mathcal{C}^{\phi}\left(\hat{\phi}\right) = \hat{\phi} \pm 1.96\frac{1}{\sqrt{n-3}} \tag{11.13}$$

is the "obviously correct" 95% confidence interval[3] for ϕ, closely approximating Neyman's construction. The endpoints of (11.13) are then transformed back to the θ scale according to the inverse transformation

$$\theta = \frac{e^{2\phi} - 1}{e^{2\phi} + 1}, \tag{11.14}$$

giving (almost) the interval $\mathcal{C}(\hat{\theta})$ seen in Figure 11.1, but without the involved computations.

Bayesian confidence statements are inherently transformation invariant. The fact that the Neyman intervals are also invariant, unlike the standard intervals (11.1), has made them more palatable to Bayesian statisticians. Transformation invariance will play a major role in justifying the bootstrap confidence intervals introduced next.

11.2 The Percentile Method

Our goal is to automate the calculation of confidence intervals: given the bootstrap distribution of a statistical estimator $\hat{\theta}$, we want to automatically produce an appropriate confidence interval for the unseen parameter θ. To this end, a series of four increasingly accurate bootstrap confidence interval algorithms will be described.

The first and simplest method is to use the standard interval (11.1), $\hat{\theta} \pm 1.96\widehat{\text{se}}$ for 95% coverage, with $\widehat{\text{se}}$ taken to be the bootstrap standard error $\widehat{\text{se}}_{\text{boot}}$ (10.16). The limitations of this approach become obvious in Figure 11.3, where the histogram shows $B = 2000$ nonparametric bootstrap replications $\hat{\theta}^*$ of the sample correlation coefficient for the student

[3] This is an anachronism. Fisher hated the term "confidence interval" after it was later coined by Neyman for his comprehensive theory. He thought of (11.13) as an example of the *logic of inductive inference*.

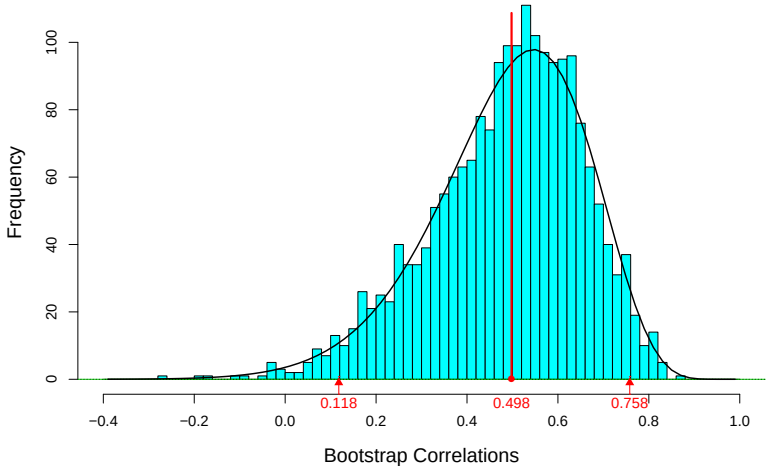

Figure 11.3 Histogram of $B = 2000$ nonparametric bootstrap replications $\hat{\theta}^*$ for the student score sample correlation; the solid curve is the ideal parametric bootstrap distribution $f_{\hat{\theta}}(r)$ as in Figure 11.1. Observed correlation $\hat{\theta} = 0.498$. Small triangles show histogram's 0.025 and 0.975 quantiles.

score data, obtained as in Section 10.2. The standard intervals are justified by taking literally the asymptotic normality of $\hat{\theta}$,

$$\hat{\theta} \dot{\sim} \mathcal{N}(\theta, \sigma^2), \tag{11.15}$$

σ the true standard error.

Relation (11.15) will generally hold for large enough sample size n, but we can see that for the student score data asymptotic normality has *not* yet set in, with the histogram being notably long-tailed to the left. We can't expect good performance from the standard method in this case. (The parametric bootstrap distribution is just as nonnormal, as shown by the smooth curve.)

The *percentile method* uses the shape of the bootstrap distribution to improve upon the standard intervals (11.1). Having generated B bootstrap replications $\hat{\theta}^{*1}, \hat{\theta}^{*2}, \ldots, \hat{\theta}^{*B}$, either nonparametrically as in Section 10.2 or parametrically as in Section 10.4, we use the obvious percentiles of their distribution to define the percentile confidence limits. The histogram in Figure 11.3 has its 0.025 and 0.975 percentiles equal to 0.118 and 0.758,

and these are the endpoints of the central 95% nonparametric percentile interval.

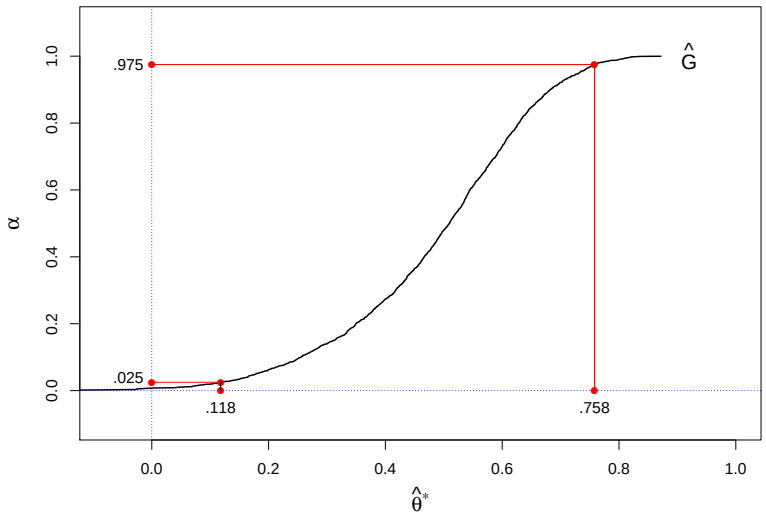

Figure 11.4 A 95% central confidence interval via the percentile method, based on the 2000 nonparametric replications $\hat{\theta}^*$ of Figure 11.3.

We can state things more precisely in terms of the *bootstrap cdf* $\hat{G}(t)$, the proportion of bootstrap samples less than t,

$$\hat{G}(t) = \#\left\{\hat{\theta}^{*b} \le t\right\} \Big/ B. \tag{11.16}$$

The αth percentile point $\hat{\theta}^{*(\alpha)}$ of the bootstrap distribution is given by the inverse function of $\hat{G}$,

$$\hat{\theta}^{*(\alpha)} = \hat{G}^{-1}(\alpha); \tag{11.17}$$

$\hat{\theta}^{*(\alpha)}$ is the value putting proportion α of the bootstrap sample to its left. The level-α upper endpoint of the percentile interval, say $\hat{\theta}_{\%\text{ile}}[\alpha]$, is by definition

$$\hat{\theta}_{\%\text{ile}}[\alpha] = \hat{\theta}^{*(\alpha)} = \hat{G}^{-1}(\alpha). \tag{11.18}$$

In this notation, the 95% central percentile interval is

$$\left(\hat{\theta}_{\%\text{ile}}[.025], \hat{\theta}_{\%\text{ile}}[.975]\right). \tag{11.19}$$

The construction is illustrated in Figure 11.4.

The percentile intervals are transformation invariant. Let $\phi = m(\theta)$ as in (11.9), and likewise $\hat{\phi} = m(\hat{\theta})$ ($m(\cdot)$ monotonically increasing), with bootstrap replications $\hat{\phi}^{*b} = m(\hat{\theta}^{*b})$ for $b = 1, 2, \ldots, B$. The bootstrap percentiles transform in the same way,

$$\hat{\phi}^{*(\alpha)} = m\left(\hat{\theta}^{*(\alpha)}\right), \tag{11.20}$$

so that, as in (11.18),

$$\hat{\phi}_{\%\text{ile}}[\alpha] = m\left(\hat{\theta}_{\%\text{ile}}[\alpha]\right), \tag{11.21}$$

verifying transformation invariance.

In what sense does the percentile method improve upon the standard intervals? One answer involves transformation invariance. Suppose there exists a monotone transformation $\phi = m(\theta)$ and $\hat{\phi} = m(\hat{\theta})$ such that

$$\hat{\phi} \sim \mathcal{N}(\phi, \sigma^2) \tag{11.22}$$

for every θ, with σ^2 constant. Fisher's transformation (11.11)–(11.12) almost accomplishes this for the normal correlation coefficient.

It would then be true that parametric bootstrap replications would also follow (11.22),

$$\hat{\phi}^* \sim \mathcal{N}\left(\hat{\phi}, \sigma^2\right). \tag{11.23}$$

That is, the bootstrap cdf $\hat{G}^{\phi}$ would be normal with mean $\hat{\phi}$ and variance σ^2. The αth percentile of $\hat{G}^{\phi}$ would equal

$$\hat{\phi}_{\%\text{ile}}[\alpha] = \hat{\phi}^{*(\alpha)} = \hat{\phi} + z^{(\alpha)}\sigma, \tag{11.24}$$

where $z^{(\alpha)}$ denotes the αth percentile of a standard normal distribution,

$$z^{(\alpha)} = \Phi^{-1}(\alpha) \tag{11.25}$$

($z^{(.975)} = 1.96$, $z^{(.025)} = -1.96$, etc.).

In other words, the percentile method would provide Fisher's "obviously correct" intervals for ϕ,

$$\hat{\phi} \pm 1.96\sigma \tag{11.26}$$

for 95% coverage for example. But, because of transformation invariance, the percentile intervals for our original parameter θ would also be exactly correct.

Some comments concerning the percentile method are pertinent.

- The method does not require actually knowing the transformation to normality $\hat{\phi} = m(\hat{\theta})$, it only assumes its existence.
- If a transformation to form (11.22) exists, then the percentile intervals are not only accurate, but also *correct* in the Fisherian sense of giving the logically appropriate inference.† †2
- The justifying assumption for the standard intervals (11.15), $\hat{\theta} \dot{\sim} \mathcal{N}(\theta, \sigma^2)$, becomes more accurate as the sample size n increases (usually with σ decreasing as $1/\sqrt{n}$), but the convergence can be slow in cases like that of the normal correlation coefficient. The broader assumption (11.22), that $m(\hat{\theta}) \dot{\sim} \mathcal{N}(m(\theta), \sigma^2)$ for *some* transformation $m(\cdot)$, speeds up convergence, irrespective of whether or not it holds exactly. Section 11.4 makes this point explicit, in terms of asymptotic rates of convergence.
- The standard method works fine once it is applied on an appropriate scale, as in Figure 11.2. The trouble is that the method is *not* transformation invariant, leaving the statistician the job of finding the correct scale. The percentile method can be thought of as a transformation-invariant version of the standard intervals, an "automatic Fisher" that substitutes massive computations for mathematical ingenuity.
- The method requires bootstrap sample sizes† on the order of $B = 2000$. †3
- The percentile method is not the last word in bootstrap confidence intervals. Two improvements, the "BC" and "BCa" methods, will be discussed in the next section. Table 11.1 compares the various intervals as applied to the student score correlation, $\hat{\theta} = 0.498$.

Table 11.1 *Bootstrap confidence limits for student score correlation, $\hat{\theta} = 0.498$, $n = 22$. Parametric exact limits from Neyman's construction as in Figure 11.1. The BC and BCa methods are discussed in the next two sections; (z_0, a), two constants required for BCa, are $(-0.055, 0.005)$ parametric, and $(0.000, 0.006)$ nonparametric.*

	Parametric		Nonparametric	
	.025	.975	.025	.975
1. Standard	.17	.83	.18	.82
2. Percentile	.11	.77	.13	.76
3. BC	.08	.75	.13	.76
4. BCa	.08	.75	.12	.76
Exact	**.09**	**.75**		

The label "computer-intensive inference" seems especially apt as ap-

plied to bootstrap confidence intervals. Neyman and Fisher's constructions are expanded from a few special theoretically tractable cases to almost any situation where the statistician has a repeatable algorithm. Automation, the replacement of mathematical formulas with wide-ranging computer algorithms, will be a major theme of succeeding chapters.

11.3 Bias-Corrected Confidence Intervals

The ideal form (11.22) for the percentile method, $\hat{\phi}^* \sim \mathcal{N}(\hat{\phi}, \sigma^2)$, says that the transformation $\hat{\phi} = m(\hat{\theta})$ yields an unbiased estimator of constant variance. The improved methods of this section and the next take into account the possibility of bias and changing variance. We begin with bias.

If $\hat{\phi} \sim \mathcal{N}(\phi, \sigma^2)$ for all $\phi = m(\theta)$, as hypothesized in (11.22), then $\hat{\phi}^* \sim \mathcal{N}(\hat{\phi}, \sigma^2)$ and

$$\Pr_*\left\{\hat{\phi}^* \le \hat{\phi}\right\} = 0.50 \tag{11.27}$$

($\Pr_*$ indicating bootstrap probability), in which case the monotonicity of $m(\cdot)$ gives

$$\Pr_*\left\{\hat{\theta}^* \le \hat{\theta}\right\} = 0.50. \tag{11.28}$$

That is, $\hat{\theta}^*$ is *median unbiased*[4] for $\hat{\theta}$, and likewise $\hat{\theta}$ for θ.

We can check that. For a parametric family of densities $f_\theta(\hat{\theta})$, (11.28) implies

$$\int_{-\infty}^{\hat{\theta}} f_{\hat{\theta}}\left(\hat{\theta}^*\right)\, d\hat{\theta}^* = 0.50. \tag{11.29}$$

For the normal correlation coefficient density (3.11), $n = 22$, numerical integration gives

$$\int_{-1}^{.498} f_{.498}\left(\hat{\theta}^*\right)\, d\hat{\theta}^* = 0.478, \tag{11.30}$$

which is not far removed from 0.50, but far enough to have a small impact on proper inference. It suggests that $\hat{\theta}^*$ is biased *upward* relative to $\hat{\theta}$—that's why *less* than half of the bootstrap probability lies below $\hat{\theta}$—and by implication that $\hat{\theta}$ is upwardly biased for estimating θ. Accordingly, confidence intervals should be adjusted a little bit downward. The *bias-corrected percentile method* (BC for short) is a data-based algorithm for making such adjustments.

[4] Median unbiasedness, unlike the usual mean unbiasedness definition, has the advantage of being transformation invariant.

Having simulated B bootstrap replications $\hat{\theta}^{*1}, \hat{\theta}^{*2}, \ldots, \hat{\theta}^{*B}$, parametric or nonparametric, let p_0 be the proportion of replications less than $\hat{\theta}$,

$$p_0 = \#\left\{\hat{\theta}^{*b} \le \hat{\theta}\right\} \Big/ B \tag{11.31}$$

(an estimate of (11.29)), and define the *bias-correction value*

$$z_0 = \Phi^{-1}(p_0), \tag{11.32}$$

where Φ^{-1} is the inverse function of the standard normal cdf. The BC level-α confidence interval endpoint is defined to be

$$\hat{\theta}_{\mathrm{BC}}[\alpha] = \hat{G}^{-1}\left[\Phi\left(2z_0 + z^{(\alpha)}\right)\right], \tag{11.33}$$

where $\hat{G}$ is the bootstrap cdf (11.16) and $z^{(\alpha)} = \Phi^{-1}(\alpha)$ (11.25).

If $p_0 = 0.50$, the median unbiased situation, then $z_0 = 0$ and

$$\hat{\theta}_{\mathrm{BC}}[\alpha] = \hat{G}^{-1}\left[\Phi\left(z^{(\alpha)}\right)\right] = \hat{G}^{-1}(\alpha) = \hat{\theta}_{\%\mathrm{ile}}[\alpha], \tag{11.34}$$

the percentile limit (11.18). Otherwise, a bias correction is made. Taking $p_0 = 0.478$ for the normal correlation example (the value we would get from an infinite number of parametric bootstrap replications) gives bias correction value -0.055. Notice that the BC limits are indeed shifted downward from the parametric percentile limits in Table 11.1. Nonparametric bootstrapping gave p_0 about 0.50 in this case, making the BC limits nearly the same as the percentile limits.

A more general transformation argument motivates the BC definition (11.33). Suppose there exists a monotone transformation $\phi = m(\theta)$ and $\hat{\phi} = m(\hat{\theta})$ such that for any θ

$$\hat{\phi} \sim \mathcal{N}(\phi - z_0\sigma, \sigma^2), \tag{11.35}$$

with z_0 and σ fixed constants. Then the BC endpoints are accurate, i.e., have the claimed coverage probabilities, and are also "obviously correct" in the Fisherian sense. See the chapter endnotes† for proof and discussion. †4

As before, the statistican does not need to know the transformation $m(\cdot)$ that leads to $\hat{\phi} \sim \mathcal{N}(\phi - z_0\sigma, \sigma^2)$, only that it exists. It is a broader target than $\hat{\phi} \sim \mathcal{N}(\phi, \sigma^2)$ (11.22), making the BC method better justified than the percentile method, irrespective of whether or not such a transformation exists. There is no extra computational burden: the bootstrap replications $\{\hat{\theta}^{*b},\ b = 1, 2, \ldots, B\}$, parametric or nonparametric, provide $\hat{G}$ (11.16) and z_0 (11.31)–(11.32), giving $\hat{\theta}_{\mathrm{BC}}[\alpha]$ from (11.33).

11.4 Second-Order Accuracy

Coverage errors of the standard confidence intervals typically decrease at order $O(1/\sqrt{n})$ in the sample size n: having calculated $\hat{\theta}_{\text{stan}}[\alpha] = \hat{\theta} + z^{(\alpha)}\hat{\sigma}$ for an iid sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, we can expect the actual coverage probability to be

$$\Pr_\theta \left\{\theta \le \hat{\theta}_{\text{stan}}[\alpha]\right\} \doteq \alpha + c_1/\sqrt{n}, \tag{11.36}$$

where c_1 depends on the problem at hand; (11.36) defines "first-order accuracy." It can connote painfully slow convergence to the nominal coverage level α, requiring sample size $4n$ to cut the error in half.

A *second-order accurate* method, say $\hat{\theta}_{\text{2nd}}[\alpha]$, makes errors of order only $O(1/n)$,

$$\Pr_\theta \left\{\theta \le \hat{\theta}_{\text{2nd}}[\alpha]\right\} \doteq \alpha + c_2/n. \tag{11.37}$$

The improvement is more than theoretical. In practical problems like that of Table 11.1, second-order accurate methods—BCa, defined in the following, is one such—often provide nearly the claimed coverage probabilities, even in small-size samples.

Neither the percentile method nor the BC method is second-order accurate (although, as in Table 11.1, they tend to be more accurate than the standard intervals). The difficulty for $\hat{\theta}_{\text{BC}}[\alpha]$ lies in the ideal form (11.35), $\hat{\phi} \sim \mathcal{N}(\phi - z_0\sigma, \sigma^2)$, where it is assumed $\hat{\phi} = m(\hat{\theta})$ has *constant* standard error σ. Instead, we now postulate the existence of a monotone transformation $\phi = m(\theta)$ and $\hat{\phi} = m(\hat{\theta})$ less restrictive than (11.35),

$$\hat{\phi} \sim \mathcal{N}(\phi - z_0\sigma_\phi, \sigma_\phi^2), \qquad \sigma_\phi = 1 + a\phi. \tag{11.38}$$

†5 Here the "acceleration"† a is a small constant describing how the standard deviation of $\hat{\phi}$ varies with ϕ. If $a = 0$ we are back in situation (11.34)[5], but if not, an amendment to the BC formula (11.33) is required.

The *BCa method* ("bias-corrected and accelerated") takes its level-α confidence limit to be

$$\hat{\theta}_{\text{BCa}}[\alpha] = \hat{G}^{-1}\left[\Phi\left(z_0 + \frac{z_0 + z^{(\alpha)}}{1 - a(z_0 + z^{(\alpha)})}\right)\right]. \tag{11.39}$$

A still more elaborate transformation argument shows that, if there exists a monotone transformation $\phi = m(\theta)$ and constants z_0 and a yielding

[5] This assumes $\sigma_0 = 1$ on the right side of (11.38), which can always be achieved by further transforming ϕ to ϕ/σ.

(11.38), then the BCa limits have their claimed coverage probabilities and, moreover, are correct in the Fisherian sense.

BCa makes three corrections to the standard intervals (11.1): for nonnormality of $\hat{\theta}$ (through using the bootstrap percentiles rather than just the bootstrap standard error); for bias (through the bias correction value z_0); and for nonconstant standard error (through a). Notice that if $a = 0$ then BCa (11.39) reduces to BC (11.33). If $z_0 = 0$ then BC reduces to the percentile method (11.18); and if $\hat{G}$, the bootstrap histogram, is normal, then (11.18) reduces to the standard interval (11.1). All three of the corrections, for nonnormality, bias, and acceleration, can have substantial effects in practice and are necessary to achieve second-order accuracy. A great deal of theoretical effort was devoted to verifying the second-order accuracy and BCa intervals under reasonably general assumptions.[6]

Table 11.2 *Nominal 95% central confidence intervals for Poisson parameter θ having observed $\hat{\theta} = 10$; actual tail areas above and below $\hat{\theta} = 10$ defined as in Figure 11.1 (atom of probability split at 10). For instance, lower standard limit 3.80 actually puts probability 0.004 above 10, rather than nominal value 0.025. Bias correction value z_0* (11.32) *and acceleration a* (11.38) *both equal 0.050.*

	Nominal limits		Tail areas	
	.025	.975	Above	Below
1. Standard	3.80	16.20	.004	.055
2. %ile	4.18	16.73	.007	.042
3. BC	4.41	17.10	.010	.036
4. BCa	5.02	17.96	.023	.023
Exact	**5.08**	**17.82**	**.025**	**.025**

The advantages of increased accuracy are not limited to large sample sizes. Table 11.2 returns to our original example of observing $\hat{\theta} = 10$ from Poisson model $\hat{\theta} \sim \text{Poi}(\theta)$. According to Neyman's construction, the 0.95 exact limits give tail areas 0.025 in both the above and below directions, as in Figure 11.1, and this is nearly matched by the BCa limits. However the standard limits are much too conservative at the left end and anti-conservative at the right.

[6] The mathematical side of statistics has also been affected by electronic computation, where it is called upon to establish the properties of general-purpose computer algorithms such as the bootstrap. Asymptotic analysis in particular has been aggressively developed, the verification of second-order accuracy being a nice success story.

Table 11.3 *95% nominal confidence intervals for the parametric and nonparametric eigenratio examples of Figures 10.2 and 10.6.*

	Parametric		Nonparametric	
	.025	.975	.025	.975
1. Standard	.556	.829	.545	.840
2. %ile	.542	.815	.517	.818
3. BC	.523	.828	.507	.813
4. BCa	.555	.820	.523	.828
	($z_0 = -.029$, $a = .058$)		($z_0 = -.049$, $a = .051$)	

Bootstrap confidence limits continue to provide better inferences in the vast majority of situations too complicated for exact analysis. One such situation is examined in Table 11.3. It relates to the eigenratio example illustrated in Figures 10.2–10.6. In this case the nonnormality and bias corrections stretch the bootstrap intervals to the left, but the acceleration effect pulls right, partially canceling out the net change from the standard intervals.

The percentile and BC methods are completely automatic, and can be applied whenever a sufficiently large number of bootstrap replications are available. The same cannot be said of BCa. A drawback of the BCa method is that the acceleration a is not a function of the bootstrap distribution and must be computed separately. Often this is straightforward:

- For one-parameter exponential families such as the Poisson, a equals z_0.
- In one-sample nonparametric problems, a can be estimated from the jackknife resamples $\hat{\theta}_{(i)}$ (10.5),

$$\hat{a} = \frac{1}{6} \frac{\sum_{i=1}^{n} \left(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}\right)^3}{\left[\sum_{i=1}^{n} \left(\hat{\theta}_i - \hat{\theta}_{(\cdot)}\right)^2\right]^{1.5}}. \tag{11.40}$$

- The *abc method* computes a in multiparameter exponential families (5.54), as does the resampling-based `R` algorithm `accel`.

Confidence intervals require the number of bootstrap replications B to be on the order of 2000, rather than the 200 or fewer needed for standard errors; the corrections made to the standard intervals are more delicate than standard errors and require greater accuracy.

There is one more cautionary note to sound concerning nuisance parameters: biases can easily get out of hand when the parameter vector μ is

high-dimensional. Suppose we observe

$$x_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, 1) \qquad \text{for } i = 1, 2, \ldots, n, \tag{11.41}$$

and wish to set a confidence interval for $\theta = \sum_1^n \mu_i^2$. The MLE $\hat{\theta} = \sum_1^n x_i^2$ will be sharply biased upward if n is at all large. To be specific, if $n = 10$ and $\hat{\theta} = 20$, we compute† †6

$$z_0 = \Phi^{-1}(0.156) = -1.01. \tag{11.42}$$

This makes[7] $\hat{\theta}_{\text{BC}}[.025]$ (11.33) equal a ludicrously small bootstrap percentile,

$$\hat{G}^{-1}(0.000034), \tag{11.43}$$

a warning sign against the BC or BCa intervals, which work most dependably for $|z_0|$ and $|a|$ small, say ≤ 0.2.

A more general warning would be against blind trust in maximum likelihood estimates in high dimensions. Computing z_0 is a wise precaution even if it is not used for BC or BCa purposes, in case it alerts one to dangerous biases.

Confidence intervals for classical applications were most often based on the standard method (11.1) (with $\widehat{\text{se}}$ estimated by the delta method) except in a few especially simple situations such as the Poisson. Second-order accurate intervals are very much a computer-age development, with both the algorithms and the inferential theory presupposing high-speed electronic computation.

11.5 Bootstrap-*t* Intervals

The initial breakthrough on exact confidence intervals came in the form of Student's t distribution in 1908. Suppose we independently observe data from two possibly different normal distributions, $\boldsymbol{x} = (x_1, x_2, \ldots, x_{n_x})$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_{n_y})$,

$$x_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu_x, \sigma^2) \quad \text{and} \quad y_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu_y, \sigma^2), \tag{11.44}$$

and wish to form a 0.95 central confidence interval for

$$\theta = \mu_y - \mu_x. \tag{11.45}$$

The obvious estimate is

$$\hat{\theta} = \bar{y} - \bar{x}, \tag{11.46}$$

[7] Also $\hat{\theta}_{\text{BCa}}[.025]$, a is zero in this model.

but its distribution depends on the nuisance parameter σ^2.

Student's masterstroke was to base inference about θ on the *pivotal quantity*

$$t = \frac{\hat{\theta} - \theta}{\widehat{\text{se}}} \tag{11.47}$$

where $\widehat{\text{se}}^2$ is an unbiased estimate of σ^2,

$$\widehat{\text{se}}^2 = \left(\frac{1}{n_x} + \frac{1}{n_y}\right) \frac{\sum_1^{n_x}(x_i - \bar{x})^2 + \sum_1^{n_y}(y_i - \bar{y})^2}{n_x + n_y - 2}; \tag{11.48}$$

t then has the "Student's t distribution" with $\text{df} = n_x + n_y - 2$ degrees of freedom if $\mu_x = \mu_y$, no matter what σ^2 may be.

Letting $t_{\text{df}}^{(\alpha)}$ represent the 100αth percentile of a t_{df} distribution yields

$$\hat{\theta}_t[\alpha] = \hat{\theta} - \widehat{\text{se}} \cdot t_{\text{df}}^{(1-\alpha)} \tag{11.49}$$

as the upper level-α interval of a Student's t confidence limit. Applied to the difference between the **AML** and **ALL** scores in Figure 1.4, the central 0.95 Student's t interval for $\theta = E\{\texttt{AML}\} - E\{\texttt{ALL}\}$ was calculated to be

$$\left(\hat{\theta}_t[.025], \hat{\theta}_t[.975]\right) = (.062, .314). \tag{11.50}$$

Here $n_x = 47$, $n_y = 25$, and $\text{df} = 70$.

Student's theory depends on the normality assumptions of (11.44). The *bootstrap-t* approach is to accept (or pretend) that t in (11.47) is pivotal, but to estimate its distribution via bootstrap resampling. Nonparametric bootstrap samples are drawn separately from $\boldsymbol{x}$ and $\boldsymbol{y}$,

$$\boldsymbol{x}^* = (x_1^*, x_2^*, \ldots, x_{n_x}^*) \quad \text{and} \quad \boldsymbol{y}^* = (y_1^*, y_2^*, \ldots, y_{n_y}^*), \tag{11.51}$$

from which we calculate $\hat{\theta}^*$ and $\widehat{\text{se}}^*$, (11.46) and (11.48), giving

$$t^* = \frac{\hat{\theta}^* - \hat{\theta}}{\widehat{\text{se}}^*}, \tag{11.52}$$

with $\hat{\theta}$ playing the role of θ, as appropriate in the bootstrap world. Replications $\{t^{*b},\ b = 1, 2, \ldots, B\}$ provide estimated percentiles $t^{*(\alpha)}$ and corresponding confidence limits

$$\hat{\theta}_t^*[\alpha] = \hat{\theta} - \widehat{\text{se}} \cdot t^{*(1-\alpha)}. \tag{11.53}$$

For the **AML–ALL** example, the t^* distribution differed only slightly from a t_{70} distribution; the resulting 0.95 interval was $(0.072, 0.323)$, nearly

the same as (11.50), lending credence to the original normality assumptions.

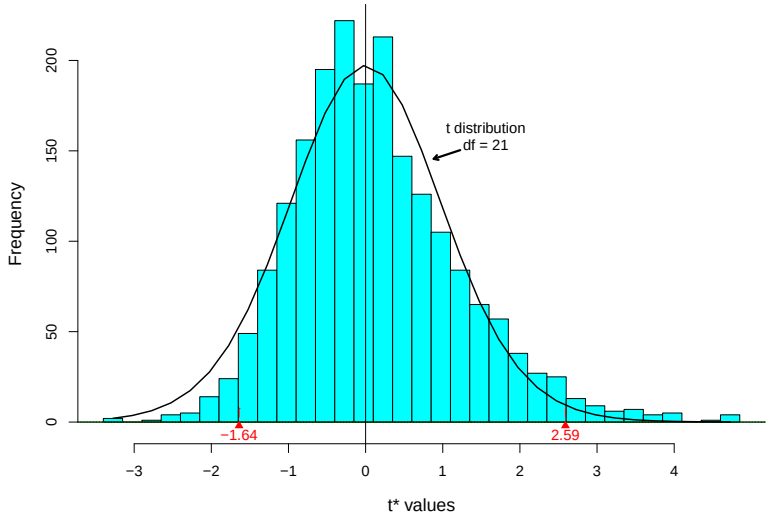

Figure 11.5 $B = 2000$ nonparametric replications of bootstrap-t statistic for the student score correlation; small triangles show 0.025 and 0.975 percentile points. The histogram is sharply skewed to the right; the solid curve is Student's t density for 21 degrees of freedom.

Returning to the student score correlation example of Table 11.1, we can apply bootstrap-t methods by still taking $t = (\hat{\theta} - \theta)/\widehat{\text{se}}$ to be notionally pivotal, but now with θ the true correlation, $\hat{\theta}$ the sample correlation, and $\widehat{\text{se}}$ the approximate standard error $(1 - \hat{\theta}^2)/\sqrt{19}$. Figure 11.5 shows the histogram of $B = 2000$ nonparametric bootstrap replications $t^* = (\hat{\theta}^* - \hat{\theta})/\widehat{\text{se}}^*$. These gave bootstrap percentiles

$$\left(t^{*(.025)}, t^{*(.975)}\right) = (-1.64, 2.59) \tag{11.54}$$

(which might be compared with $(-2.08, 2.08)$ for a standard t_{21} distribution), and 0.95 interval $(0.051, 0.781)$ from (11.53), somewhat out of place compared with the other entries in the right panel of Table 11.1.

Bootstrap-t intervals are *not* transformation invariant. This means they can perform poorly or well depending on the scale of application. If performed on Fisher's scale (11.11) they agree well with exact intervals for

the correlation coefficient. A practical difficulty is the requirement of a formula for $\widehat{\text{se}}$.

Nevertheless, the idea of estimating the actual distribution of a proposed pivotal quantity has great appeal to the modern statistical spirit. Calculating the percentiles of the original Student t distribution was a multi-year project in the early twentieth century. Now we can afford to calculate our own special "t table" for each new application. Spending such computational wealth wisely, while not losing one's inferential footing, is the central task and goal of twenty-first-century statisticians.

11.6 Objective Bayes Intervals and the Confidence Distribution

Interval estimates are ubiquitous. They play a major role in the scientific discourse of a hundred disciplines, from physics, astronomy, and biology to medicine and the social sciences. Neyman-style frequentist confidence intervals dominate the literature, but there have been influential Bayesian and Fisherian developments as well, as discussed next.

Given a one-parameter family of densities $f_\theta(\hat\theta)$ and a prior density $g(\theta)$, Bayes' rule (3.5) produces the posterior density of θ,

$$g(\theta|\hat\theta) = g(\theta)\, f_\theta(\hat\theta)/f(\hat\theta), \tag{11.55}$$

where $f(\hat\theta)$ is the marginal density $\int f_\theta(\hat\theta)g(\theta)d\theta$. The Bayes 0.95 *credible interval* $C(\theta|\hat\theta)$ spans the central 0.95 region of $g(\theta|\hat\theta)$, say

$$C(\theta|\hat\theta) = (a(\hat\theta), b(\hat\theta)), \tag{11.56}$$

with

$$\int_{a(\hat\theta)}^{b(\hat\theta)} g(\theta|\hat\theta)\, d\theta = 0.95, \tag{11.57}$$

and with posterior probability 0.025 in each tail region.

Confidence intervals, of course, require no prior information, making them eminently useful in day-to-day applied practice. The Bayesian equivalents are credible intervals based on uninformative priors, Section 3.2. "Matching priors," those whose credible intervals nearly match Neyman confidence intervals, have been of particular interest. Jeffreys' prior (3.17),

$$\begin{aligned} g(\theta) &= \mathcal{I}_\theta^{1/2}, \\ \mathcal{I}_\theta &= \int \left[\frac{\partial}{\partial\theta}\log f_\theta(\hat\theta)\right]^2 f_\theta(\hat\theta)\, d\hat\theta, \end{aligned} \tag{11.58}$$

provides a generally accurate matching prior for one-parameter problems. Figure 3.2 illustrates this for the student score correlation, where the credible interval $(0.093, 0.750)$ is a near-exact match to the Neyman 0.95 interval of Figure 11.1.

Difficulties begin with multiparameter families $f_\mu(x)$ (5.1): we wish to construct an interval estimate for a one-dimensional function $\theta = t(\mu)$ of the p-dimensional parameter vector μ, and must somehow remove the effects of the $p-1$ "nuisance parameters." In a few rare situations, including the normal theory correlation coefficient, this can be done exactly. Pivotal methods do the job for Student's t construction. Bootstrap confidence intervals greatly extend the reach of such methods, at a cost of greatly increased computation.

Bayesians get rid of nuisance parameters by integrating them out of the posterior density $g(\mu|\boldsymbol{x}) = g(\mu)f_\mu(\boldsymbol{x})/f(\boldsymbol{x})$ (3.6) ($\boldsymbol{x}$ now representing all the data, "$\boldsymbol{x}$" equaling $(\boldsymbol{x}, \boldsymbol{y})$ for the Student t setup (11.44)). That is, we calculate[8] the marginal density of $\theta = t(\mu)$ given $\boldsymbol{x}$, and call it $h(\theta|\boldsymbol{x})$. A credible interval for θ, $\mathcal{C}(\theta|\boldsymbol{x})$, is then constructed as in (11.56)–(11.57), with $h(\theta|\boldsymbol{x})$ playing the role of $g(\theta|\hat{\theta})$. This leaves us the knotty problem of choosing an uninformative multidimensional prior $g(\mu)$. We will return to the question after first discussing fiducial methods, a uniquely Fisherian device.

Fiducial constructions begin with what seems like an obviously incorrect interpretation of pivotality. We rewrite the Student t pivotal $t = (\hat{\theta} - \theta)/\widehat{\text{se}}$ (11.47) as

$$\theta = \hat{\theta} - \widehat{\text{se}} \cdot t, \tag{11.59}$$

where t has a Student's t distribution with df degrees of freedom, $t \sim t_{\text{df}}$. Having observed the data $(\boldsymbol{x}, \boldsymbol{y})$ (11.44), fiducial theory assigns θ the distribution implied by (11.59), as if $\hat{\theta}$ and $\widehat{\text{se}}$ were *fixed* at their calculated values while t was distributed as t_{df}. Then $\hat{\theta}_t[\alpha]$ (11.49), the Student t level-α confidence limit, is the 100αth percentile of θ's fiducial distribution.

We seem to have achieved a Bayesian posterior conclusion without any prior assumptions.[9] The historical development here is confused by Fisher's refusal to accept Neyman's confidence interval theory, as well as his disparagement of Bayesian ideas. As events worked out, all of Fisher's immense prestige was not enough to save fiducial theory from the scrapheap of failed statistical methods.

[8] Often a difficult calculation, as discussed in Chapter 13.

[9] "Enjoying the Bayesian omelette without breaking the Bayesian eggs," in L. J. Savage's words.

And yet, in Arthur Koestler's words, "The history of ideas is filled with barren truths and fertile errors." Fisher's underlying rationale went something like this: $\hat{\theta}$ and $\widehat{\text{se}}$ exhaust the information about θ available from the data, after which there remains an irreducible component of randomness described by t. This is an idea of substantial inferential appeal, and one that can be rephrased in more general terms discussed next that bear on the question of uninformative priors.

By definition, an upper confidence limit $\hat{\theta}_{\boldsymbol{x}}[\alpha]$ satisfies

$$\Pr\left\{\theta \le \hat{\theta}_{\boldsymbol{x}}[\alpha]\right\} = \alpha \tag{11.60}$$

(where now we have indicated the observed data $\boldsymbol{x}$ in the notation), and so

$$\Pr\left\{\hat{\theta}_{\boldsymbol{x}}[\alpha] \le \theta \le \hat{\theta}_{\boldsymbol{x}}[\alpha+\epsilon]\right\} = \epsilon. \tag{11.61}$$

We can consider $\hat{\theta}_{\boldsymbol{x}}[\alpha]$ as a one-to-one function between α in $(0,1)$ and θ a point in its parameter space Θ (assuming that $\hat{\theta}_{\boldsymbol{x}}[\alpha]$ is smoothly increasing in α). Letting ϵ go to zero in (11.61) determines the *confidence density* of θ, say $\tilde{g}_{\boldsymbol{x}}(\theta)$,

$$\tilde{g}_{\boldsymbol{x}}(\theta) = d\alpha/d\theta, \tag{11.62}$$

the local derivative of probability at location θ for the unknown parameter, the derivative being taken at $\theta = \hat{\theta}_{\boldsymbol{x}}[\alpha]$.

Integrating $\tilde{g}_{\boldsymbol{x}}(\theta)$ recovers α as a function of θ. Let $\theta_1 = \hat{\theta}_{\boldsymbol{x}}[\alpha_1]$ and $\theta_2 = \hat{\theta}_{\boldsymbol{x}}[\alpha_2]$ for any two values $\alpha_1 < \alpha_2$ in $(0,1)$. Then

$$\begin{aligned}\int_{\theta_1}^{\theta_2} \tilde{g}_{\boldsymbol{x}}(\theta)\, d\theta &= \int_{\theta_1}^{\theta_2} \frac{d\alpha}{d\theta}\, d\theta = \alpha_2 - \alpha_1 \\ &= \Pr\{\theta_1 \le \theta \le \theta_2\},\end{aligned} \tag{11.63}$$

as in (11.60). There is nothing controversial about (11.63) as long as we remember that the random quantity in $\Pr\{\theta_1 \le \theta \le \theta_2\}$ is not θ but rather the interval (θ_1, θ_2), which varies as a function of $\boldsymbol{x}$. Forgetting this leads to the textbook error of attributing Bayesian properties to frequentist results: "There is 0.95 probability that θ is in its 0.95 confidence interval," etc.

This is exactly what the fiducial argument does.[10] Whether or not one accepts (11.63), there is an immediate connection with *matching priors*.

[10] Fiducial and confidence densities agree, as can be seen in the Student t situation (11.59), at least in the somewhat limited catalog of cases Fisher thought appropriate for fiducial calculations.

Suppose prior $g(\mu)$ gives a perfect match to the confidence interval system $\hat{\theta}_{\boldsymbol{x}}[\alpha]$. Then, by definition, its posterior density $h(\theta|\boldsymbol{x})$ must satisfy

$$\int_{-\infty}^{\hat{\theta}_x[\alpha]} h(\theta|\boldsymbol{x})\, d\theta = \alpha = \int_{-\infty}^{\hat{\theta}_x[\alpha]} \tilde{g}_{\boldsymbol{x}}(\theta)\, d\theta \tag{11.64}$$

for $0 < \alpha < 1$. But this implies $h(\theta|\boldsymbol{x})$ equals $\tilde{g}_{\boldsymbol{x}}(\theta)$ for all θ. That is, *the confidence density $\tilde{g}_{\boldsymbol{x}}(\theta)$ is the posterior density of θ given x for any matching prior.*

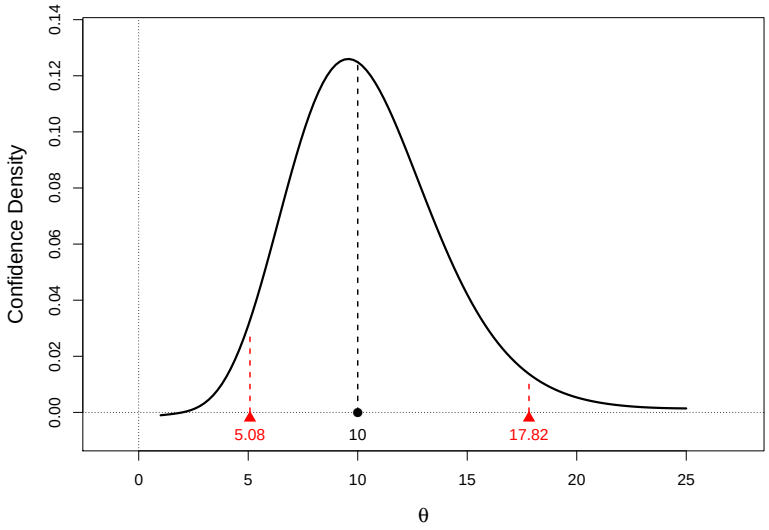

Figure 11.6 Confidence density (11.62) for Poisson parameter θ having observed $\hat{\theta} = 10$. There is area 0.95 under the curve between 5.08 and 17.82, as in Table 11.2, and areas 0.025 in each tail.

Figure 11.6 graphs the confidence density for $\hat{\theta} \sim \text{Poi}(\theta)$ having observed $\hat{\theta} = 10$. This was obtained by numerically differentiating α as a function of θ (11.62),

$$\alpha = \Pr\{10 \le \text{Poi}(\theta)\}, \tag{11.65}$$

"$\le$" including splitting the atom of probability at 10. According to Table 11.2, $\tilde{g}_{10}(\theta)$ has area 0.95 between 5.08 and 17.82, and area 0.025 in each tail. Whatever its provenance, the graph delivers a striking picture of the uncertainty in the unknown value of θ.

Bootstrap confidence intervals provide easily computable confidence densities. Let $\hat{G}(\theta)$ be the bootstrap cdf and $\hat{g}(\theta)$ its density function (obtained by differentiating a smoothed version of $\hat{G}(\theta)$ when $\hat{G}$ is based on B bootstrap replications). The percentile confidence limits $\hat{\theta}[\alpha] = \hat{G}^{-1}(\alpha)$ (11.16) have $\alpha = \hat{G}(\theta)$, giving

$$\tilde{g}_{\boldsymbol{x}}(\theta) = \hat{g}(\theta). \tag{11.66}$$

(It is helpful to picture this in Figure 11.4.) For the percentile method, the bootstrap density *is* the confidence density.

For the BCa intervals (11.39), the confidence density is obtained by reweighting $\hat{g}(\theta)$,

$$\tilde{g}_{\boldsymbol{x}}(\theta) = c\,w(\theta)\hat{g}(\theta), \tag{11.67}$$

†7 where†

$$w(\theta) = \frac{\varphi\left[z_\theta/(1+az_\theta) - z_0\right]}{(1+az_\theta)^2\varphi(z_\theta + z_0)}, \quad \text{with } z_\theta = \Phi^{-1}\hat{G}(\theta) - z_0. \tag{11.68}$$

Here φ is the standard normal density, Φ its cdf, and c the constant that makes $\tilde{g}_{\boldsymbol{x}}(\theta)$ integrate to 1. In the usual case where the bootstrap cdf is estimated from replications $\hat{\theta}^{*b}$, $b = 1, 2, \ldots, B$ (either parametric or nonparametric), the BCa confidence density is a reweighted version of $\hat{g}(\theta)$. Define

$$W_b = w\left(\hat{\theta}^{*b}\right) \Big/ \sum_{i=1}^{B} w\left(\hat{\theta}^{*i}\right). \tag{11.69}$$

Then the BCa confidence density is the discrete density putting weight W_b on $\hat{\theta}^{*b}$.

Figure 11.7 returns to the student score data, $n = 22$ students, five scores each, modeled normally as in Figure 10.6,

$$x_i \overset{\text{iid}}{\sim} \mathcal{N}_5(\lambda, \boldsymbol{\Sigma}) \qquad \text{for } i = 1, 2, \ldots, 22. \tag{11.70}$$

This is a $p = 20$-dimensional parametric family: 5 expectations, 5 variances, 10 covariances. The parameter of interest was taken to be

$$\theta = \text{maximum eigenvalue of } \boldsymbol{\Sigma}. \tag{11.71}$$

It had MLE $\hat{\theta} = 683$, this being the maximum eigenvalue of the MLE sample covariance matrix $\hat{\boldsymbol{\Sigma}}$ (dividing each sum of squares by 22 rather than 21).

$B = 8000$ parametric bootstrap replications[11] $\hat{\theta}^{*b}$ gave percentile and

[11] $B = 2000$ would have been enough for most purposes, but $B = 8000$ gave a sharper picture of the different curves.

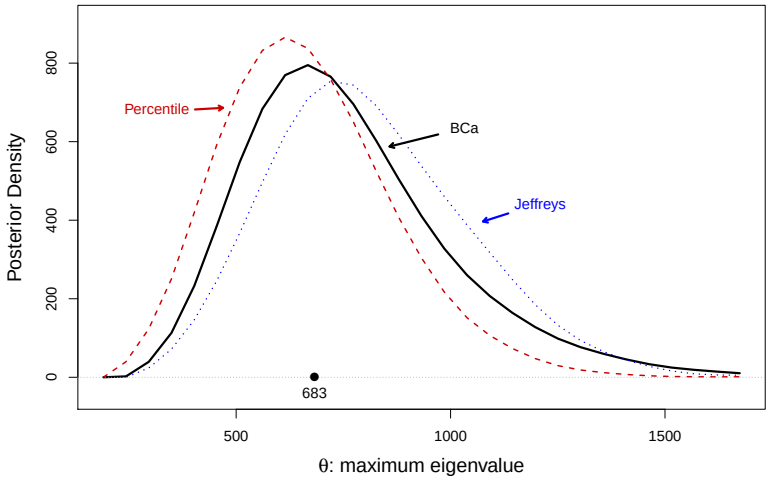

Figure 11.7 Confidence densities for the maximum eigenvalue parameter (11.71), using a multivariate normal model (11.70) for the student score data. The dashed red curve is the percentile method, solid black the BCa (with $(z_0, a) = (0.178, 0.093)$). The dotted blue curve is the Bayes posterior density for θ, using Jeffreys' prior (11.72).

BCa confidence densities as shown. In this case the weights W_b (11.69) increased with $\hat{\theta}^{*b}$, pushing the BCa density to the right. Also shown is the Bayes posterior density† for θ starting from Jeffreys' multiparameter prior †8
density

$$g^{\text{Jeff}}(\mu) = |\mathcal{I}_\mu|^{1/2}, \tag{11.72}$$

where $\mathcal{I}_\mu$ is the Fisher information matrix (5.26). It isn't truly uninformative here, moving its credible limits upward from the second-order accurate BCa confidence limits. Formula (11.72) is discussed further in Chapter 13.

Bayesian data analysis has the attractive property that, after examining the data, we can express our remaining uncertainty in the language of probability. Fiducial and confidence densities provide something similar for confidence intervals, at least partially freeing the frequentist from the interpretive limitations of Neyman's intervals.

11.7 Notes and Details

Fisher's theory of fiducial inference (1930) preceded Neyman's approach, formalized in (1937), which was presented as an attempt to put interval estimation on a firm probabilistic basis, as opposed to the mysteries of fiducialism. The result was an elegant theory of exact and optimal intervals, phrased in hard-edged frequentistic terms. Readers familiar with the theory will know that Neyman's construction—a favorite name in the physics literature—as pictured in Figure 11.1, requires some conditions on the family of densities $f_\theta(\hat\theta)$ to yield optimal intervals, a sufficient condition being monotone likelihood ratios.

Bootstrap confidence intervals, Efron (1979, 1987), are neither exact nor optimal, but aim instead for wide applicability combined with near-exact accuracy. Second-order acuracy of BCa intervals was established by Hall (1988). BCa is emphatically a child of the computer age, routinely requiring $B = 2000$ or more bootstrap replications per use. Shortcut methods are available. The "abc method" (DiCiccio and Efron, 1992) needs only 1% as much computation, at the expense of requiring smoothness properties for $\theta = t(\mu)$, and a less automatic coding of the exponential family setting for individual situations. In other words, it is less convenient.

†1 [p. 183] *Neyman's construction.* For any given value of θ, let $(\theta^{(.025)}, \theta^{(.975)})$ denote the central 95% interval of density $f_\theta(\hat\theta)$, satisfying

$$\int_{-\infty}^{\theta^{(.025)}} f_\theta\left(\hat\theta\right)\, d\hat\theta = 0.025 \quad \text{and} \quad \int_{-\infty}^{\theta^{(.975)}} f_\theta\left(\hat\theta\right)\, d\hat\theta = 0.975; \tag{11.73}$$

and let $I_\theta(\hat\theta)$ be the indicator function for $\hat\theta \in (\theta^{(.025)}, \theta^{(.975)})$,

$$I_\theta\left(\hat\theta\right) = \begin{cases} 1 & \text{if } \theta^{(.025)} < \hat\theta < \theta^{(.975)} \\ 0 & \text{otherwise.} \end{cases} \tag{11.74}$$

By definition, $I_\theta(\hat\theta)$ has a two-point probability distribution,

$$I_\theta\left(\hat\theta\right) = \begin{cases} 1 & \text{probability } 0.95 \\ 0 & \text{probability } 0.05. \end{cases} \tag{11.75}$$

This makes $I_\theta(\hat\theta)$ a pivotal statistic, one whose distribution does not depend upon θ.

Neyman's construction takes the confidence interval $\mathcal{C}(\hat\theta)$ corresponding to observed value $\hat\theta$ to be

$$\mathcal{C}\left(\hat\theta\right) = \left\{\theta : I_\theta\left(\hat\theta\right) = 1\right\}. \tag{11.76}$$

Then $\mathcal{C}(\hat{\theta})$ has the desired coverage property

$$\Pr_\theta\left\{\theta \in \mathcal{C}\left(\hat{\theta}\right)\right\} = \Pr_\theta\left\{I_\theta\left(\hat{\theta}\right) = 1\right\} = 0.95 \tag{11.77}$$

for any choice of the true parameter θ. (For the normal theory correlation density of $f_\theta(\hat{\theta})$, $\hat{\theta}(.025)$ and $\hat{\theta}(.975)$ are increasing functions of θ. This makes our previous construction (11.6) agree with (11.76).) The construction applies quite generally, as long as we are able to define acceptance regions of the sample space having the desired target probability content for every choice of θ. This can be challenging in multiparameter families.

†2 [p. 189] *Fisherian correctness.* Fisher, arguing against the Neyman paradigm, pointed out that confidence intervals could be accurate without being correct: having observed $x_i \overset{\text{iid}}{\sim} \mathcal{N}(\theta, 1)$ for $i = 1, 2, \ldots, 20$, the standard 0.95 interval based on just the first 10 observations would provide exact 0.95 coverage while giving obviously incorrect inferences for θ. If we can reduce the situation to form (11.22), the percentile method intervals satisfy Fisher's "logic of inductive inference" for correctness, as at (4.31).

†3 [p. 189] *Bootstrap sample sizes.* Why we need bootstrap sample sizes on the order of $B = 2000$ for confidence interval construction can be seen in the estimation of the bias correction value z_0 (11.32). The delta-method standard error of $z_0 = \Phi^{-1}(p_0)$ is calculated to be

$$\frac{1}{\varphi(z_0)}\left[\frac{p_0(1-p_0)}{B}\right]^{1/2}, \tag{11.78}$$

with $\varphi(z)$ the standard normal density. With $p_0 \doteq 0.5$ and $z_0 \doteq 0$ this is about $1.25/B^{1/2}$, equaling 0.028 at $B = 2000$, a none-too-small error for use in the BC formula (11.33) or the BCa formula (11.39).

†4 [p. 191] *The acceleration a.* This a appears in (11.38) as $d\sigma_\phi/d\phi$, the rate of change of $\hat{\phi}$'s standard deviation as a function of its expectation. In one-parameter exponential families it turns out that this is one-third of $d\sigma_\theta/d\theta$; that is, the transformation to normality $\phi = m(\theta)$ also decreases the instability of the standard deviation, though not to zero.

The variance of the score function $\dot{l}_x(\theta)$ determines the standard deviation of the MLE $\hat{\theta}$ (4.17)–(4.18). In one-parameter exponential families, one-sixth the *skewness* of $\dot{l}_x(\theta)$ gives a. The skewness connection can be seen at work in estimate (11.40). In multivariate exponential families (5.50), the skewness must be evaluated in the "least favorable" direction, discussed further in Chapter 13. The R algorithm `accel` (book web site) uses B parametric bootstrap replications $(\hat{\beta}^{*b}, \hat{\theta}^{*b})$ to estimate a. The percentile and BC intervals require only the replications $\hat{\theta}^{*b}$, while BCa also

requires knowledge of the underlying exponential family. See Sections 4, 6, and 7 of Efron (1987).

†5 [p. 192] *BCa accuracy and correctness.* The BCa confidence limit $\hat{\theta}_{\mathrm{BCa}}[\alpha]$ (11.39) is transformation invariant. Define

$$z[\alpha] = z_0 + \frac{z_0 + z^{(\alpha)}}{1 - a(z_0 + z^{(\alpha)})}, \tag{11.79}$$

so $\hat{\theta}_{\mathrm{BCa}}[\alpha] = \hat{G}^{-1}\{\Phi[z[\alpha]]\}$. For a monotone increasing transformation $\phi = m(\theta)$, $\hat{\phi} = m(\hat{\theta})$, and $\hat{\phi}^* = m(r)$, the bootstrap cdf $\hat{H}$ of $\hat{\phi}^*$ satisfies $\hat{H}^{-1}(\alpha) = m[\hat{G}^{-1}(\alpha)]$ since $\hat{\phi}^{*(\alpha)} = m(\hat{\theta}^{*(\alpha)})$ for the bootstrap percentiles. Therefore

$$\hat{\phi}_{\mathrm{BCa}}[\alpha] = \hat{H}^{-1}\left\{\Phi\left(z[\alpha]\right)\right\} = m\left(\hat{G}^{-1}\left\{\Phi\left(z[\alpha]\right)\right\}\right) = m\left(\hat{\theta}_{\mathrm{BCa}}[\alpha]\right), \tag{11.80}$$

verifying transformation invariance. (Notice that $z_0 = \Phi^{-1}[\hat{G}(\hat{\theta})]$ equals $\Phi^{-1}[\hat{H}(\hat{\phi})]$ and is also transformation invariant, as is a, as discussed previously.)

Exact confidence intervals are transformation invariant, adding considerably to their inferential appeal. For approximate intervals, transformation invariance means that if we can demonstrate good behavior on any one scale then it remains good on all scales. The model (11.38) to the ϕ scale can be re-expressed as

$$\left\{1 + a\hat{\phi}\right\} = \{1 + a\phi\}\{1 + a(Z - z_0)\}, \tag{11.81}$$

where Z is a standard normal variate, $Z \sim \mathcal{N}(0, 1)$.

Taking logarithms,

$$\hat{\gamma} = \gamma + U, \tag{11.82}$$

where $\hat{\gamma} = \log\{1 + a\hat{\phi}\}$, $\gamma = \log\{1 + a\phi\}$, and U is the random variable $\log\{1 + a(Z - z_0)\}$; (11.82) represents the simplest kind of translation model, where the unknown value of γ rigidly shifts the distribution of U. The obvious confidence limit for γ,

$$\hat{\gamma}[\alpha] = \hat{\gamma} - U^{(1-\alpha)}, \tag{11.83}$$

where $U^{(1-\alpha)}$ is the $100(1-\alpha)$th percentile of U, is then accurate, and also "correct," according to Fisher's (admittedly vague) logic of inductive inference. It is an algebraic exercise, given in Section 3 of Efron (1987), to reverse the transformations $\theta \to \phi \to \gamma$ and recover $\hat{\theta}_{\mathrm{BCa}}[\alpha]$ (11.39). Setting $a = 0$ shows the accuracy and correctness of $\hat{\theta}_{\mathrm{BC}}[\alpha]$ (11.33).

†6 [p. 195] *Equation* (11.42). Model (11.41) makes $\hat{\theta} = \sum x_i^2$ a *noncentral chi-square variable* with noncentrality parameter $\theta = \sum \mu_i^2$ and n degrees of freedom, written as $\hat{\theta} \sim \chi^2_{\theta,n}$. With $\hat{\theta} = 20$ and $n = 10$, the parametric bootstrap distribution is $r \sim \chi^2_{20,10}$. Numerical evaluation gives $\Pr\{\chi^2_{20,10} \le 20\} = 0.156$, leading to (11.42).

Efron (1985) concerns confidence intervals for parameters $\theta = t(\boldsymbol{\mu})$ in model (11.41), where third-order accurate confidence intervals can be calculated. The acceleration a equals zero for such problems, making the BC intervals second-order accurate. In practice, the BC intervals usually perform well, and are a reasonable choice if the accleration a is unavailable.

†7 [p. 202] *BCa confidence density* (11.68). Define

$$z_\theta = \Phi^{-1}\left[\hat{G}(\theta)\right] - z_0 = \frac{z_0 + z^{(\alpha)}}{1 - a\left(z_0 + z^{(\alpha)}\right)}, \tag{11.84}$$

so that

$$z^{(\alpha)} = \frac{z_\theta}{1 + a z_\theta} - z_0 \quad \text{and} \quad \alpha = \Phi\left(\frac{z_\theta}{1 + a z_\theta} - z_0\right). \tag{11.85}$$

Here we are thinking of α and θ as functionally related by $\theta = \hat{\theta}_{\mathrm{BCa}}[\alpha]$. Differentiation yields

$$\begin{aligned} \frac{d\alpha}{dz_\theta} &= \frac{\varphi\left(\frac{z_\theta}{1+az_\theta} - z_0\right)}{(1 + a z_\theta)^2}, \\ \frac{dz_\theta}{d\theta} &= \frac{\varphi\left(\frac{z_\theta}{1+az_\theta} - z_0\right)}{(1 + a z_\theta)^2 \varphi(z_\theta + z_0)}\, \hat{g}(\theta), \end{aligned} \tag{11.86}$$

which together give $d\alpha/d\theta$, verifying (11.68).

The name "confidence density" seems to appear first in Efron (1993), though the idea is familiar in the fiducial literature. An ambitious frequentist theory of confidence distributions is developed in Xie and Singh (2013).

†8 [p. 203] *Jeffreys' prior.* Formula (11.72) is discussed further in Chapter 13, in the more general context of uninformative prior distributions. The theory of matching priors was initiated by Welch and Peers (1963), another important reference being Tibshirani (1989).