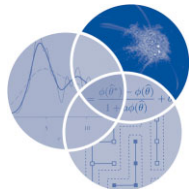
Overview



Bootstrap

Tim Hesterberg*

This article provides an introduction to the bootstrap. The bootstrap provides statistical inferences—standard error and bias estimates, confidence intervals, and hypothesis tests—without assumptions such as Normal distributions or equal variances. As such, bootstrap methods can be remarkably more accurate than classical inferences based on Normal or t distributions. The bootstrap uses the same basic procedure regardless of the statistic being calculated, without requiring the use of application-specific formulae. This article may provide two big surprises for many readers. The first is that the bootstrap shows that common t confidence intervals are woefully inaccurate when populations are skewed, with one-sided coverage levels off by factors of two or more, even for very large samples. The second is that the number of bootstrap samples required is much larger than generally realized. *WIREs Comp Stat* 2011 3 497–526 DOI: 10.1002/wics.182

Keywords: resampling; permutation tests; inference; standard error; bias

INTRODUCTION

We begin with an example of the simplest type of bootstrapping in this section, then discuss the idea behind the bootstrap (Section *Plug-In Principle*), implementation by random sampling (Section *Monte Carlo Sampling—The "Second Bootstrap Principle"*), using the bootstrap to estimate standard error and bias (Section *Bias and Standard Error*), a variety of examples, the central limit theorem and different types of bootstraps (Section *Examples*), the accuracy of the bootstrap (Section *Accuracy of Bootstrap Distributions*), confidence intervals (Section *Bootstrap Confidence Intervals*), hypothesis tests (Section *Hypothesis Testing*), planning clinical trials (Section *Planning Clinical Trials*), the number of bootstrap samples needed and ways to reduce this number (Section *How Many Bootstrap Samples Are Needed*), and conclude with references for additional reading.

Figure 1 shows a normal quantile plot of Arsenic concentrations from 271 wells in Bangladesh, from `http://www.bgs.ac.uk/arsenic/bangladesh/Data/SpecialStudyData.csv` referenced from `statlibhttp://lib.stat.cmu.edu/datasets`. The sample mean and standard deviation are $\overline{x} = 124.5$ and $s = 298$, respectively.

The usual formula standard error for the mean is $s/\sqrt{n} = 18.1$, and usual 95% confidence interval $\overline{x} \pm t_{\alpha/2,n-1}s/\sqrt{n}$ is (88.8, 160.2). This interval may be suspect because of the skewness of the data, in spite of the reasonably large sample size.

We may use the bootstrap[1] for inferences for the mean of this dataset. We draw a *bootstrap sample*, or *resample*, of size n with replacement from the data, and compute the mean. We repeat this process many times, say 10^4 or more. The resulting *bootstrap means* comprise the *bootstrap distribution*, which we use to estimate aspects of the sampling distribution for $\overline{X}$. Figure 2 shows a histogram and Normal quantile plot of the bootstrap distribution. The *bootstrap standard error* is the standard deviation of the bootstrap distribution; in this case the bootstrap standard error is 18.2, quite close to the formula standard error. The mean of the bootstrap means is 124.4, quite close to $\overline{x}$ (the difference is -0.047, to three decimal places). The bootstrap distribution looks quite normal, with some skewness.

This amount of skewness is a cause for concern. This may be counter to the intuition of many readers, who use Normal quantile plots to look at data. This bootstrap distribution corresponds to a sampling distribution, not raw data. This is after the central limit theorem has had its one chance to work, so any deviations from normality here may translate into errors in inferences. We may quantify how badly this amount of skewness affects confidence intervals; we defer this to Section *Bootstrap Confidence Intervals*.

*Correspondence to: timhesterberg@gmail.com
Google, Seattle, USA
DOI: 10.1002/wics.182

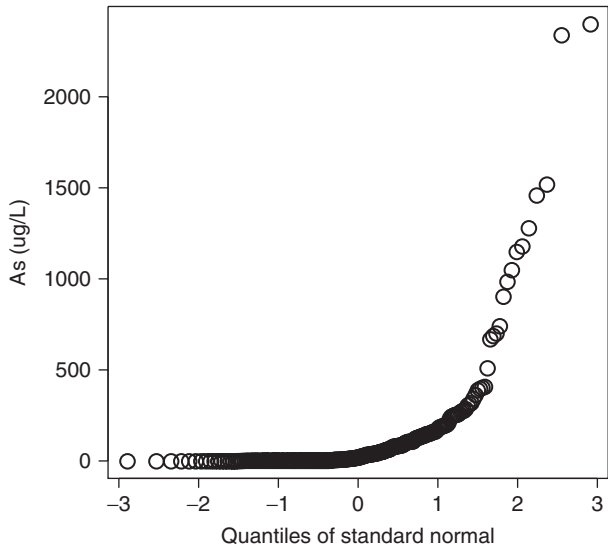


FIGURE 1 | Arsenic concentrations in 271 wells in Bangladesh.

We first discuss the idea behind the bootstrap, and give some idea of its versatility.

PLUG-IN PRINCIPLE

The idea behind the bootstrap is the *plug-in principle*[2]—that if a quantity is unknown, we plug in an estimate for it.

This principle is used all the time in statistics. The standard deviation of a sample mean for *i.i.d.* observations from a population with standard deviation σ is $\sigma/\sqrt{n}$; when σ is unknown we plug in an estimate s to obtain the usual standard error $s/\sqrt{n}$.

What is different in the bootstrap is that we plug in an estimate for the whole population, not just for a numerical summary of the population.

Statistical inference depends on the sampling distribution. The sampling distribution depends on

1. the underlying population(s),
2. the sampling procedure, and
3. the statistic, such as $\overline{X}$.

Conceptually, the sampling distribution is the result of drawing many samples from the population and calculating the statistic for each. The bootstrap principle is to plug in an estimate for the population, then mimic the real life sampling procedure and statistic calculation. The bootstrap distribution depends on

1. an estimate for the population(s),
2. the sampling procedure, and
3. the statistic, such as $\overline{X}$.

The simplest case is when the original data are an *i.i.d.* sample from a single population, and we use the empirical distribution $\hat{F}_n$ to estimate the population, where $\hat{F}_n(u) = (1/n)\sum I(x_i \le u)$. This gives the ordinary nonparametric bootstrap, corresponding to drawing samples of size n without replacement from the original data.

How Useful is the Bootstrap Distribution?

A fundamental question is how well the bootstrap distribution approximates the sampling distribution. We discuss this question in greater detail in Section *Accuracy of Bootstrap Distributions*, but note a few key points here. For most common *estimators* (statistics that are estimates of a population parameter,

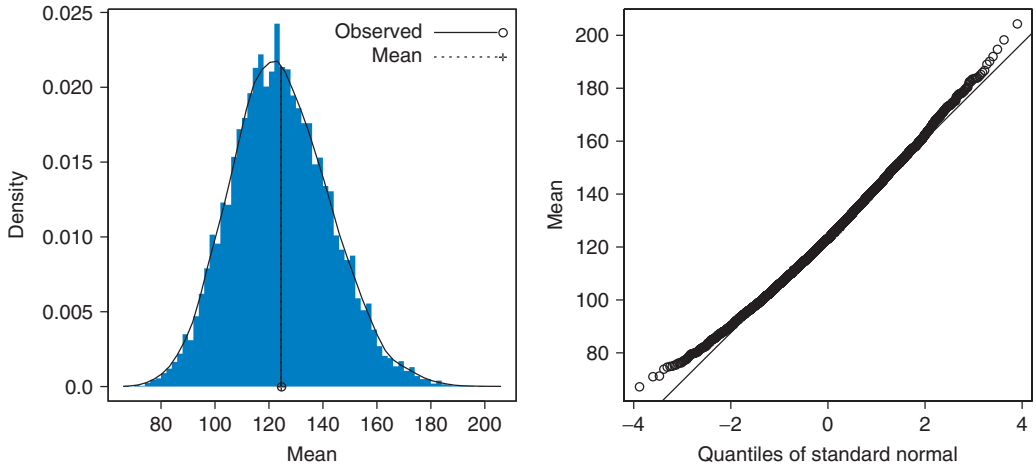


FIGURE 2 | Histogram and Normal quantile plot of the bootstrap distribution for arsenic concentrations.

e.g., $\overline{X}$ is an estimator for μ, whereas a t statistic is not an estimator), and under fairly general distribution assumptions,

center: the center of the bootstrap distribution is *not* an accurate approximation for the center of the sampling distribution. For example, the center of the bootstrap distribution for $\overline{X}$ is centered at approximately $\overline{x}$, the mean of the sample, whereas the sampling distribution is centered at μ.

spread: the spread of the bootstrap distribution does reflect the spread of the sampling distribution.

bias: the bootstrap bias estimate (see below) does reflect the bias of the sampling distribution.

skewness: the skewness of the bootstrap distribution does reflect the skewness of the sampling distribution.

The first point bears emphasis. It means that *the bootstrap is not used to get better parameter estimates* because the bootstrap distributions are centered around statistics $\hat{\theta}$ calculated from the data (e.g., $\overline{x}$ or regression slope $\hat{\beta}$) rather than the unknown population values (e.g., μ or β). Drawing thousands of bootstrap observations from the original data is not like drawing observations from the underlying population, it does not create new data.

Instead, the bootstrap sampling is useful for *quantifying the behavior of a parameter estimate*, such as its standard error, bias, or calculating confidence intervals.

There are exceptions where bootstrap averages are useful for estimation, such as random forests.[3] These are beyond the scope of this article, except that we give a toy example to illustrate the mechanism. Consider the case of simple linear regression, and suppose that there is a strong linear relationship between y and x. However, instead of using linear regression, one uses a step function—the data are split into eight equal-size groups based on x, and the y values in each group are averaged to obtain the altitude for the step. Applying the same procedure to bootstrap samples randomizes the location of the step edges, and averaging across the bootstrap samples smooths the edges of the steps. Hence the bootstrap average is more accurate than the original step function. This is shown in Figure 3. A similar effect holds in random forests, using bootstrap averaging of tree models, which fit higher dimensional data using multivariate analogs of step functions.

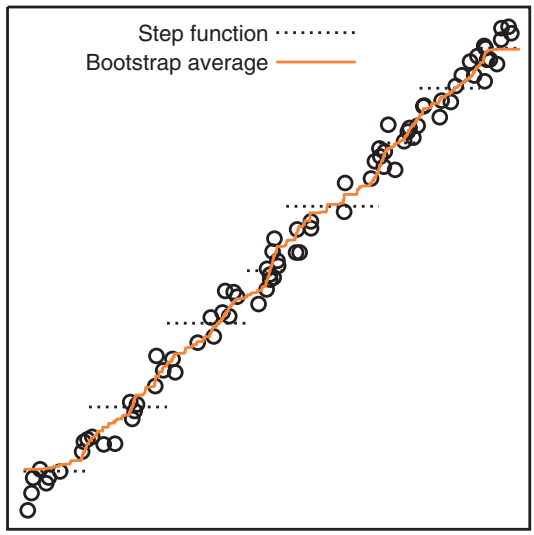


FIGURE 3 | Step function defined by eight equal-size groups, and average across bootstrap samples of step functions.

Other Population Estimates

Other estimates of the population may be used. For example, if there was reason to assume that the arsenic data followed a gamma distribution, we could estimate parameters for the gamma distribution, then draw samples from a gamma distribution with those estimated parameters. This is a *parametric bootstrap*.[4]

In other cases, we may believe that the underlying population is continuous; then rather than draw from the discrete empirical distribution, we may instead draw samples from a density estimated from the data; this is a *smoothed bootstrap*.[4]

Other Sampling Procedures

When the original data were not obtained using an *i.i.d.* sample, the bootstrap sampling should reflect the actual data collection. For example, in stratified sampling applications the bootstrap sampling should be stratified. If the original data are dependent, the bootstrap sampling should reflect the dependence; this may not be straightforward.

There are some cases where the bootstrap sampling should differ from the actual sampling procedure, including:

- regression (Section *Examples*),
- planning clinical trials (Section *Planning Clinical Trials*),
- hypothesis testing (Section *Hypothesis Testing*), and
- small samples (Section *Bootstrap Distributions Are Too Narrow*).

Other Statistics

The bootstrap procedure may be used with a wide variety of statistics—mean, median, trimmed mean, regression coefficients, hazard ratio, x-intercept in a regression, and others—using the same procedure. It does not require problem-specific analytical calculations.

This is a major advantage of the bootstrap. It allows statistical inferences such as confidence intervals to be calculated even for statistics for which there are no easy formulas. It offers hope of reforming statistical practice—away from simple but non-robust estimators like a sample mean or least-squares regression, in favor of robust alternatives.

MONTE CARLO SAMPLING—THE "SECOND BOOTSTRAP PRINCIPLE"

The second bootstrap "principle" is that the bootstrap is implemented by random sampling. This is not actually a principle, but an implementation detail.[4]

Given that we are drawing *i.i.d.* samples of size n from the empirical distribution $\hat{F}_n$, there are at most n^n possible samples ($\binom{2n-1}{n}$ if we disregard the order of observations, and ties in the data can further reduce the number of unique samples). In small samples we could create all possible bootstrap samples, deterministically. In practice n is usually too large for that to be feasible, so we use random sampling.

Let B be the number of bootstrap samples used, e.g., $B = 10^4$. The resulting B statistic values represent a random sample of size B with replacement from the *theoretical bootstrap distribution* consisting of n^n values (including duplicates).

In some cases we can calculate the theoretical bootstrap distribution without simulation. In the arsenic example, parametric bootstrapping from a gamma distribution causes the theoretical bootstrap distribution for the sample mean to be another gamma distribution.

In other cases we can calculate some aspects of the sampling distribution without simulation. In the case of the nonparametric bootstrap when the statistic is the sample mean, the mean and standard deviation of the theoretical bootstrap distribution are $\overline{x}$ and $\hat{\sigma}_{\hat{F}_n}/\sqrt{n}$, respectively, where $\hat{\sigma}^2_{\hat{F}_n} = n^{-1}\sum(x_i - \overline{x})^2$.[4] Note that this differs from the usual sample standard deviation in using a divisor of n instead of $n-1$. We return to this point in Section *Bootstrap Distributions Are Too Narrow*.

The use of Monte Carlo sampling adds additional unwanted variability, that may be reduced by increasing the value of B. We discuss how large B should be in Section *How Many Bootstrap Samples Are Needed*.

BIAS AND STANDARD ERROR

Let $\theta = \theta(F)$ be a parameter of a population, such as the mean, or difference in regression coefficients between sub-populations. Let $\hat{\theta}$ be the corresponding estimate from the data, $\hat{\theta}^*$ be the estimate from a bootstrap sample, $\overline{\hat{\theta}^*} = B^{-1}\sum_{b=1}^{B}\hat{\theta}^*_b$ be the average of B bootstrap estimates, and $s^2_{\hat{\theta}^*} = (B-1)^{-1}\sum_{b=1}^{B}(\hat{\theta}^*_b - \overline{\hat{\theta}^*})^2$ be the sample standard deviation of the bootstrap estimates.

Some bootstrap calculations require that $\hat{\theta}$ be a *functional statistic*, one that depends on the data only through the empirical distribution, not on n. A mean is a functional statistic, whereas the usual sample standard deviation s with divisor $n-1$ is not—repeating each observation twice gives the same empirical distribution but a different s.

The bootstrap bias estimate for a functional statistic is

$$\overline{\hat{\theta}^*} - \hat{\theta}. \tag{1}$$

Note how this relates to the plug-in principle. The bias of a statistic is $E(\hat{\theta}) - \theta$, which for a functional statistic may be expanded as $E_F(\hat{\theta}) - \theta(F)$, the expected value of $\hat{\theta}$ *when sampling from* F minus the value *for population* F. Substituting $\hat{F}$ for the unknown F in both terms yields the theoretical bootstrap bias estimate

$$E_{\hat{F}}(\hat{\theta}^*) - \theta(\hat{F}). \tag{2}$$

The Monte Carlo version of the bias estimate (1) substitutes the sample average of bootstrap statistics for the expected value.

EXAMPLES

In this section we consider some examples, with a particular eye to standard error, bias, and normality of the sampling distribution.

Relative Risk

A major study of the association between blood pressure and cardiovascular disease found that 55 out of 3338 men with high blood pressure died of cardiovascular disease during the study period, compared to 21 out of 2676 with low blood pressure. The estimated relative risk is $\hat{\theta} = \hat{p}_1/\hat{p}_2 = 0.0165/0.0078 = 2.12$.

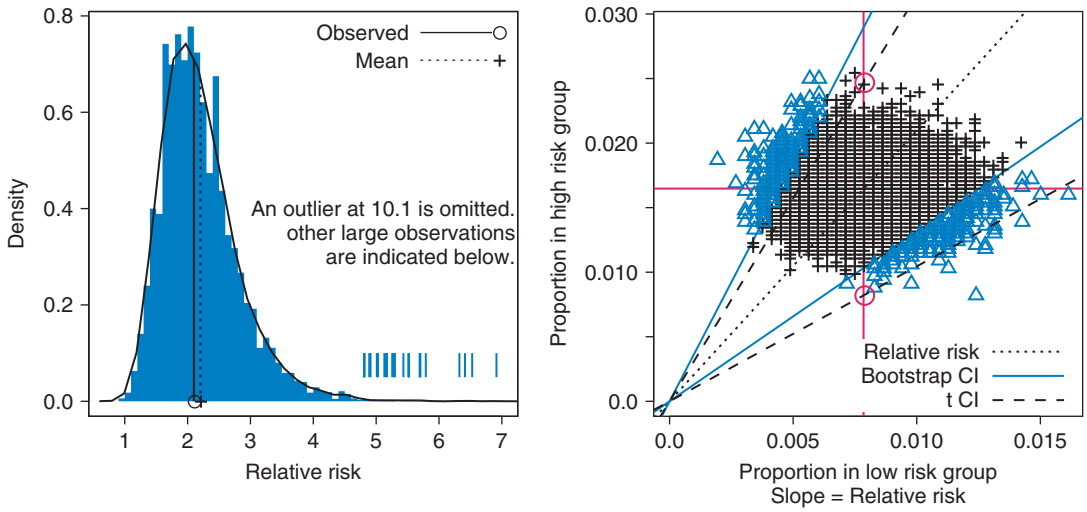


FIGURE 4 | Histogram and scatterplot of the bootstrap distribution for relative risk.

To bootstrap this, we draw samples of size $n_1 = 3338$ with replacement from the first group, independently draw samples of size $n_2 = 2676$ from the second group, and calculate the relative risk $\hat{\theta}^*$. In addition, we record the individual proportions $\hat{p}_1^*$ and $\hat{p}_2^*$. The bootstrap distribution for relative risk is shown in the left panel of Figure 4. It is highly skewed, with a long right tail caused by divisor values relatively close to zero. The standard error, from a sample of 10^4 observations, is 0.6188. The theoretical bootstrap standard error is undefined because some of the $n_1^{n_1} n_2^{n_2}$ bootstrap samples have $\hat{\theta}^*$ undefined because the denominator $\hat{p}_2^*$ is zero; this is not important in practice.

The average of the bootstrap replicates is larger than the original relative risk, indicating bias. The estimated bias is $2.205 - 2.100 = 0.106$, corresponding to 0.17 standard errors. While the bias does not appear large in the figure, this amount of bias can have a huge impact on inferences; a rough calculation suggests that the actual non-coverage of one side of a two-sided 95% confidence interval would be $1 - \Phi(0.17 + 1.96) = 0.0367$ rather than 0.025, or 47% too large.

The right panel of Figure 4 shows the joint bootstrap distribution of $\hat{p}_1^*$ and $\hat{p}_2^*$. Each point corresponds to one bootstrap sample, and the relative risk is the slope of the line between the origin and the point. The original data is at the intersection of horizontal and vertical lines. The solid diagonal lines exclude 2.5% of the bootstrap observations on each side; the corresponding slopes are the endpoints of a 95% bootstrap percentile confidence interval.

The bottom and top dashed diagonal lines are the endpoints of a t interval with standard error obtained using the usual delta method. This interval corresponds to calculating the standard error of residuals above and below the central line (the line with slope $\hat{\theta}$), going up and down 1.96 residual standard errors from the central point (the original data) to the circled points; the endpoints of the interval are the slopes of the lines from the origin to the circled points. A t interval would not be appropriate in the example, because of the bias and skewness.

In practice one would normally do a t interval on a transformed statistic, e.g., log of relative risk, or log-odds-ratio $\log(\hat{p}_1(1-\hat{p}_2)/((1-\hat{p}_1)p_2))$. Figure 5 shows a normal quantile plot for the bootstrap distribution of the log of relative risk. The distribution for log relative risk is much less skewed than is the distribution for relative risk, but still noticeably

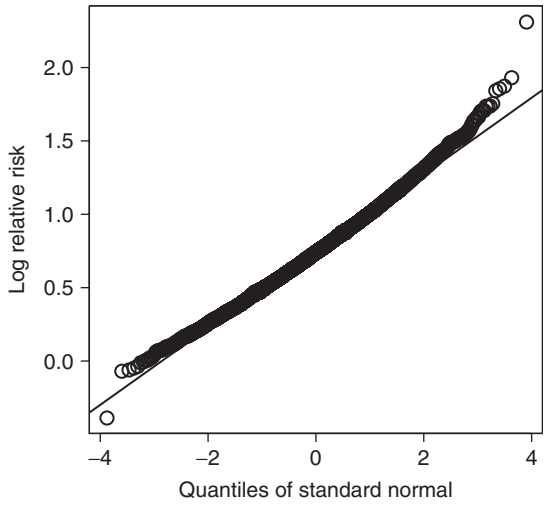


FIGURE 5 | Normal quantile plot for bootstrap distribution for log of relative risk.

skewed. Even with a log transformation, a *t* interval would only be adequate for work where accuracy is not required. We discuss confidence intervals further in Section *Bootstrap Confidence Intervals*.

Linear Regression

The next examples, for linear regression, are based on a dataset from a large pharmaceutical company. The response variable is a pharmacokinetic parameter of interest, and candidate predictors are weight, sex, age, and dose (3 levels—200, 400, and 800). There are 300 observations, one per subject. Our primary interest in this dataset will be to use the bootstrap to investigate the behavior of stepwise regression; however, first we consider some other issues.

A standard linear regression using main effects gives:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	32.0819	4.2053	7.6290	0.0000
wgt	0.2394	0.0372	6.4353	0.0000
sex	−7.2192	1.2306	−5.8666	0.0000
age	−0.1507	0.0367	−4.1120	0.0000
dose	0.0003	0.0018	0.1695	0.8653

The left panel of Figure 6 contains a scatterplot of a PK parameter versus weight, for the 25 males receiving dose = 400, as well as regression lines from 30 bootstrap samples. This is useful for giving a rough idea of variability. A bootstrap percentile confidence interval for mean PK given weight would be the range of the middle 95% of heights of regression lines at a given weight.

The right panel shows all 300 observations, and predictions for the PK/weight relationship using (1) all 300 observations, (2) the main-effects model, and (3) predictions for the "base case", males receiving dose = 400, with weight equal to the average weight for all subjects. In effect this uses the full dataset to improve predictions for a subset, "borrowing strength". There is much less variability than in the left panel, particularly for slope, primarily because of the larger sample size, but also because the addition of an important covariate (age) to the model reduces residual variance.

Note that the *y* values shown are the actual data, not adjusted for differences between the base case and the actual values of sex, age, and dose. The line is higher than most of the observations, because the PK values tend to be higher for males.

In the right panel it is clear that the variation in the regression lines is much smaller than the vertical variation in *y* values. We now turn to this point.

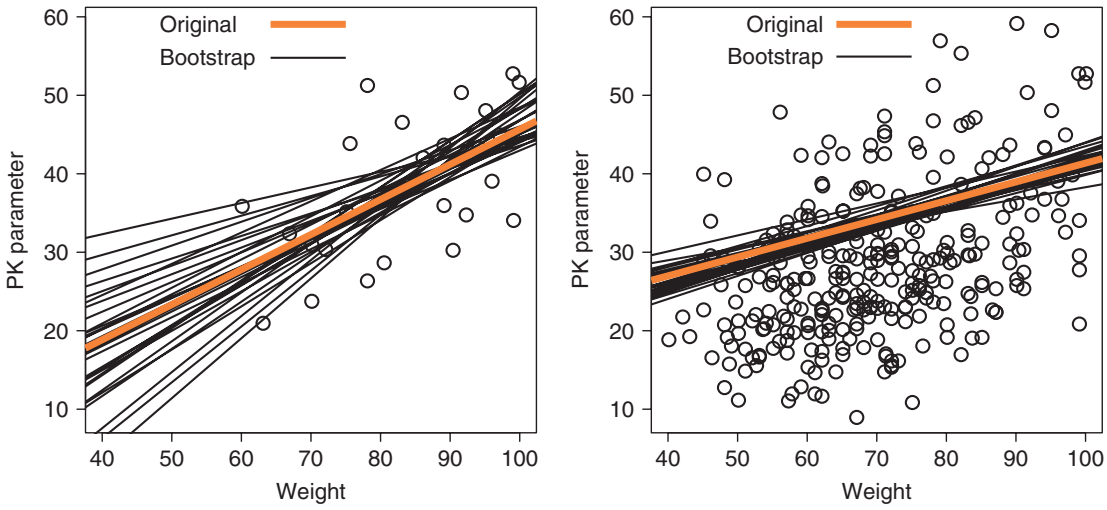


FIGURE 6 | Bootstrap regression lines. Left panel: 25 males receiving dose = 400. The orange line is the least-squares fit for those 25 observations, and black lines are from bootstrap samples of size 25. Right panel: the orange line is the prediction for males receiving dose = 400, based on the main-effects linear regression using all 300 subjects, and the black lines are from bootstrap samples.

Prediction Intervals and Non-Normality

The right panel also hints at the difference between a confidence interval (for mean response given covariates) and a prediction interval (for a new observation). With large n, the regression lines show little variation, but the variation of an individual point above and below the (true) line remains constant regardless of n. Hence as n increases, confidence intervals become narrow, but prediction intervals do not. This is reflected in the standard formulae for confidence intervals:

$$\hat{y} \pm t_\alpha s\sqrt{1/n + (x-\overline{x})^2/S_{xx}} \tag{3}$$

and prediction intervals in the simple linear regression case:

$$\hat{y} \pm t_\alpha s\sqrt{1 + 1/n + (x-\overline{x})^2/S_{xx}}, \tag{4}$$

where s is the residual standard error and $S_{xx} = \sum(x_i - \overline{x})^2$. As $n \to \infty$ the terms inside the square root decrease to zero for a confidence interval but approach 1 for a prediction interval; the prediction interval approaches $\hat{y} \pm z_\alpha s$.

Now, suppose that residuals are not normally distributed. Asymptotically and for reasonably large n the confidence intervals are approximately correct, but prediction intervals are not—the interval $\hat{y} \pm z_\alpha s$ is only correct for normally distributed data. Prediction intervals should approach $(\hat{y} \pm \hat{F}_\epsilon^{-1}(\alpha/2), \hat{y} \pm \hat{F}_\epsilon^{-1}(1-\alpha/2))$ as $n \to \infty$, where $\hat{F}$ is the estimated residual distribution.

In other words, there is no central limit theorem for prediction intervals. The outcome for a new observation depends primarily on a single random value, not an average across a large sample. Equation (4) should only be used after confirming that the residual distribution is approximately normal. And, in the opinion of this author, (4) should not be taught in introductory statistics, to students ill-equipped to understand that it should only be used if residuals are normally distributed.

A bootstrap approach that takes into account both the shape of the residual distribution and the variability in regression lines is outlined below in Section *Prediction Intervals*.

Stepwise Regression

Now consider the case of stepwise regression. We consider models ranging from the intercept-only model to a full second-order model that includes all main effects, all interactions, and quadratic functions of dose, age, and weight. We use forward and backward stepwise regression, with terms added or subtracted to minimize the C_p statistic, using the `step` function of `S-PLUS`. The resulting coefficients and inferences are:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	12.8035	14.1188	0.9068	0.3637
wgt	0.6278	0.1689	3.7181	0.0002
sex	9.2008	7.1634	1.2844	0.1980
age	−0.6583	0.2389	−2.7553	0.0055
I(age^2)	0.0052	0.0024	2.1670	0.0294
wgt:sex	−0.2077	0.0910	−2.2814	0.0218

The sex coefficient is retained in spite of the small t value because including an interaction forces retention of the corresponding main effects.

We use the bootstrap here to check model stability, obtain standard errors, and check for bias.

Model Stability

The stepwise procedure selects a six-term model. We may use the bootstrap to check the stability of the procedure under random sampling (does it consistently select the same model, or is there substantial variation?) and to see which terms are consistently included.

We create bootstrap samples by resampling subjects—whole rows of the data—with replacement. Resampling whole rows preserves covariances between variables.

In 1000 bootstrap samples, only 95 result in the same model as the original data; on average 3.2 terms differ between the original model and the bootstrap models. The original model has six terms; the bootstrap models range from 4 to 12, with an average of 7.9, or 1.9 more than the original data. This suggests that stepwise regression tends to select more terms for random data than for the corresponding population. This in turn suggests that the original six-term model may also be overfitted.

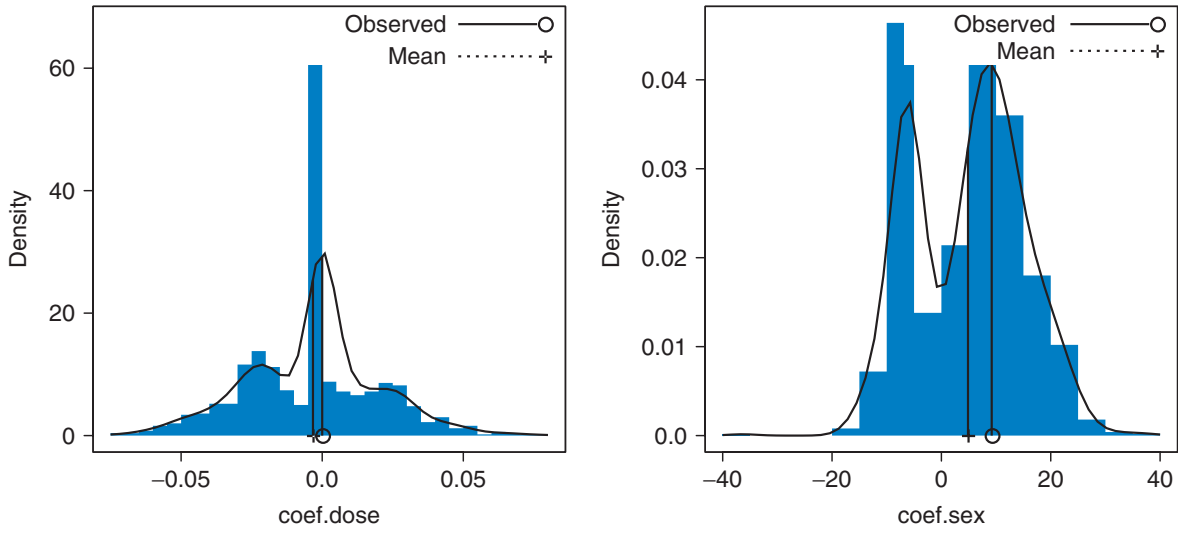


FIGURE 7 | Histograms of bootstrap distributions for dose and sex coefficients in stepwise regression.

Figure 7 shows the bootstrap distributions for two coefficients: dose, and sex. The dose coefficient is usually zero, though it may be positive or negative. This suggests that dose is not very important in determining the response.

The sex coefficient is bimodal, with the modes on opposite sides of zero. It turns out that the sex coefficient is usually negative when the weight–sex interaction is included, otherwise is positive.

Overall, the bootstrap suggests that the original model is not very stable.

For comparison, repeating the experiment with a more stringent criterion for variable inclusion—a modified C_p statistic with double the penalty—results in a more stable model. The original model has the same six terms. Of the bootstrap samples 154 yield the same model, and on average the number of different terms is 2.15. The average number of terms is 5.93, slightly less than for the original data; this suggests that stepwise regression may now be slightly underfitting (though one should not read too much into this).

Standard Errors

At the end of the stepwise procedure, the table of coefficients, standard errors, and *t* values is calculated, ignoring the variable selection process. In particular, the standard errors are calculated under the usual regression assumptions, which assume that the model is fixed from the outset. Call these *nominal standard errors*.

For each bootstrap sample, we perform stepwise selection and record the coefficients and nominal standard errors. For the main effects the bootstrap standard errors (standard deviation of bootstrap coefficients) and average of the nominal standard errors are:

	boot SE	avg.nominalSE
Intercept	27.9008	14.0734
wgt	0.5122	0.2022
sex	9.9715	5.4250
age	0.3464	0.2137
dose	0.0229	0.0091

The bootstrap standard errors are much larger than the average of the nominal standard errors.

The bootstrap standard errors reflect additional variability due to model selection, such as the bimodal distribution for the sex coefficient, factors that the nominal standard errors ignore.

This is not to say that one should use the bootstrap standard errors here. At the end of the stepwise variable selection process, it is appropriate to condition on the model, and do inferences accordingly. For example, a confidence interval for the sex coefficient should be conditional on the weight–sex interaction being included in the model.

But it does suggest that the nominal standard errors are optimistic. In fact they are biased downward, even conditional on the model terms, because they are calculated using a formula that depends on residual standard error, which in turn is biased due to model selection.

Bias

Figure 8 shows bootstrap distributions for R^2 (unadjusted) and residual standard deviation. Both show very large bias.

The bias is not surprising—optimizing generally gives biased results. Consider ordinary linear

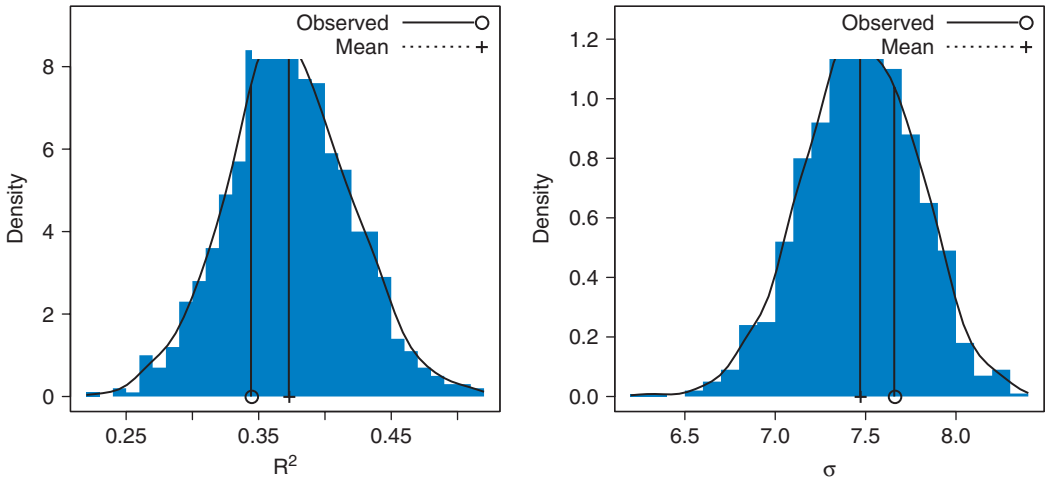


FIGURE 8 | Histograms of bootstrap distributions for R^2 and residual standard deviation in stepwise regression.

regression—unadjusted R^2 is biased. If it were calculated using the true β's instead of estimated $\hat{\beta}$'s it would not be biased. Optimizing $\hat{\beta}$ to minimize residual squared error (and maximize R^2) makes unadjusted R^2 biased.

In classical linear regression, with the model selected in advance, we commonly use adjusted R^2 to counteract the bias. Similarly, we use residual variance calculated using a divisor of $(n - p - 1)$ instead of n, where p is the number of terms in the model.

But in this case it is not only the *values* of the coefficients that are optimized, but *which* terms are included in the model. This is not reflected in the usual formulae. As a result, the residual standard error obtained from the stepwise procedure is biased downward, even using a divisor of $(n - p - 1)$.

Bootstrapping Rows or Residuals

There are two basic ways to bootstrap linear regression models—to resample rows (observations), or residuals.[2,5]

To resample residuals, we fit the initial model $\hat{y}_i = \hat{\beta}_0 + \sum \hat{\beta}_j x_{ij}$, calculate the residuals $r_i = y_i - \hat{y}_i$, then create new bootstrap samples as

$$y_i^* = \hat{y}_i + r_i^* \tag{5}$$

for $i = 1, \ldots, n$, where r_i^* is sampled with replacement from the observed residuals $\{r_1, \ldots, r_n\}$. We keep the original x and $\hat{y}$ values fixed in order to create new bootstrap y^* values.

Resampling rows corresponds to a random effects sampling design—in which x and y are both obtained by random sampling from a joint population. Resampling residuals corresponds to a fixed effects model, in which the x's are fixed by the experimental design and y's are obtained conditional on the x's. So at first glance it would appear appropriate to resample rows when the original data collection has random x's.

However, in classical statistics we commonly use inferences derived using the fixed effects model, even when the x's are actually random. We do inferences conditional on the observed x values. Similarly, in bootstrapping we may resample residuals even when the x's were originally random.

In practice the difference matters most when there are factors with rare levels, or interactions of factors with rare combinations. If resampling rows it is possible that a bootstrap sample may have none of the level or combination, in which case the corresponding term cannot be estimated, and the software may give an error. Or, what is worse, there may be one or two rows with the rare level, enough so the software would not crash, but instead quietly give garbage answers, imprecise because they are based on few observations.

Hence with factors with rare levels, or small samples more generally, it may be preferable to resample residuals.

Resampling residuals implicitly assumes that the residual distribution is the same for every x, that there is no heteroskedasticity. A variation on resampling residuals that allows heteroskedasticity is the *wild bootstrap* or *external bootstrap*,[6] which in its simplest form adds either plus or minus the original residual r_i to each fitted value,

$$y_i^* = \hat{y}_i \pm r_i, \tag{6}$$

with equal probabilities. Hence the expected value of y_i^* is $\hat{y}_i$, and the standard deviation is proportional to r_i. For further discussion see Ref 5.

There are other variations on resampling residuals, such as resampling studentized residuals, or weighted error resampling for non-constant variance.[5]

Prediction Intervals

The idea of resampling residuals provides a way to obtain more accurate prediction intervals. In order to capture both variation in the estimated regression line and residual variation, we may resample both. Variation in the regression line may be obtained by resampling either residuals or rows in order to generate random $\hat{\beta}^*$ values and corresponding $\hat{y}^* = \hat{\beta}_0 + \sum \hat{\beta}_j x_{0j}$, for predictions at x_0. Independently we draw random residuals r^*, and add them to the $\hat{y}^*$. After repeating this many times, the range of the middle 95% of the $(\hat{y}^* + r^*)$ values gives a prediction interval. For further discussion and alternatives see Ref 5.

Logistic Regression

In logistic regression it is straightforward to resample rows of the data, but resampling residuals fails—the y values must be either zero or one, but adding the residual from one observation to the prediction from another yields values anywhere between −1 and 2.

Instead, we keep the x's fixed, and generate y values from the estimated conditional distributions given each x. Let $\hat{p}_i$ be the predicted probability that $y_i = 1$ given x_i. Then

$$y_i^* = \begin{cases} 1 & \text{with probability } \hat{p}_i \\ 0 & \text{with probability } 1 - \hat{p}_i. \end{cases} \tag{7}$$

The *kyphosis* dataset[7] contains observations on 81 children who had corrective spinal surgery, on four variables: Kyphosis (a factor indicating whether a postoperative deformity is present), Age (in months), Number (of vertebrae involved in the operation), and Start (beginning of the range of vertebrae involved).

A logistic regression using main effects gives coefficients:

	Value	Std. Error	t value
(Intercept)	−2.03693352	1.449574526	−1.405194
Age	0.01093048	0.006446256	1.695633
Start	−0.20651005	0.067698863	−3.050421
Number	0.41060119	0.224860819	1.826024

suggesting that `Start` is the most important predictor.

The left panel of Figure 9 shows Kyphosis versus Start, together with predicted curve for the base case with Age = 87 (the median) and Number = 4 (the median). This is a *sunflower plot*,[8,9] in which a flower with $k > 2$ petals represents k duplicate values.

The right panel of of Figure 9 shows predictions from 20 bootstrap curves.

Figure 10 shows the bootstrap distributions for the four regression coefficients. All of the distributions are substantially non-normal. It would not be appropriate to use classical normal-based inferences. Indeed, the printout of regression coefficients above,

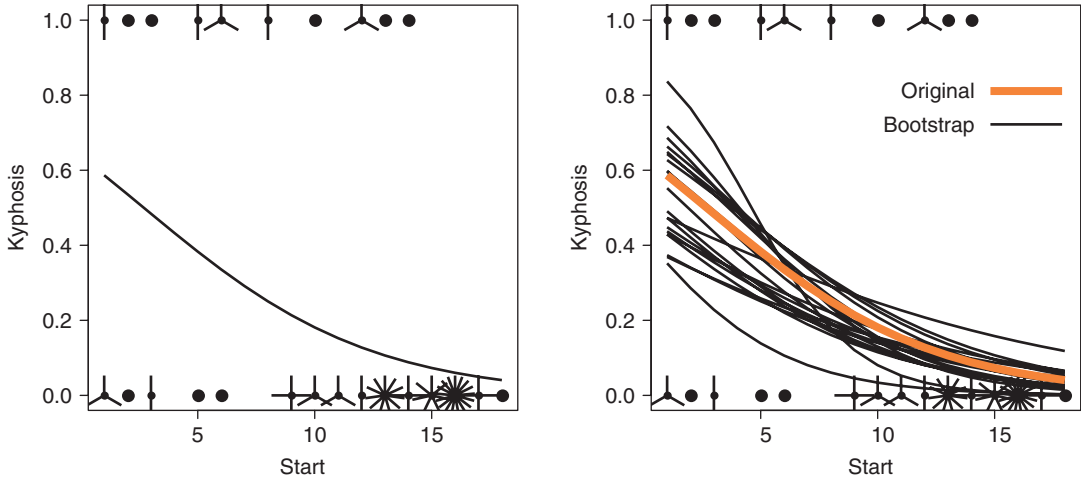


FIGURE 9 | Bootstrap curves for predicted kyphosis, for Age = 87 and Number = 4.

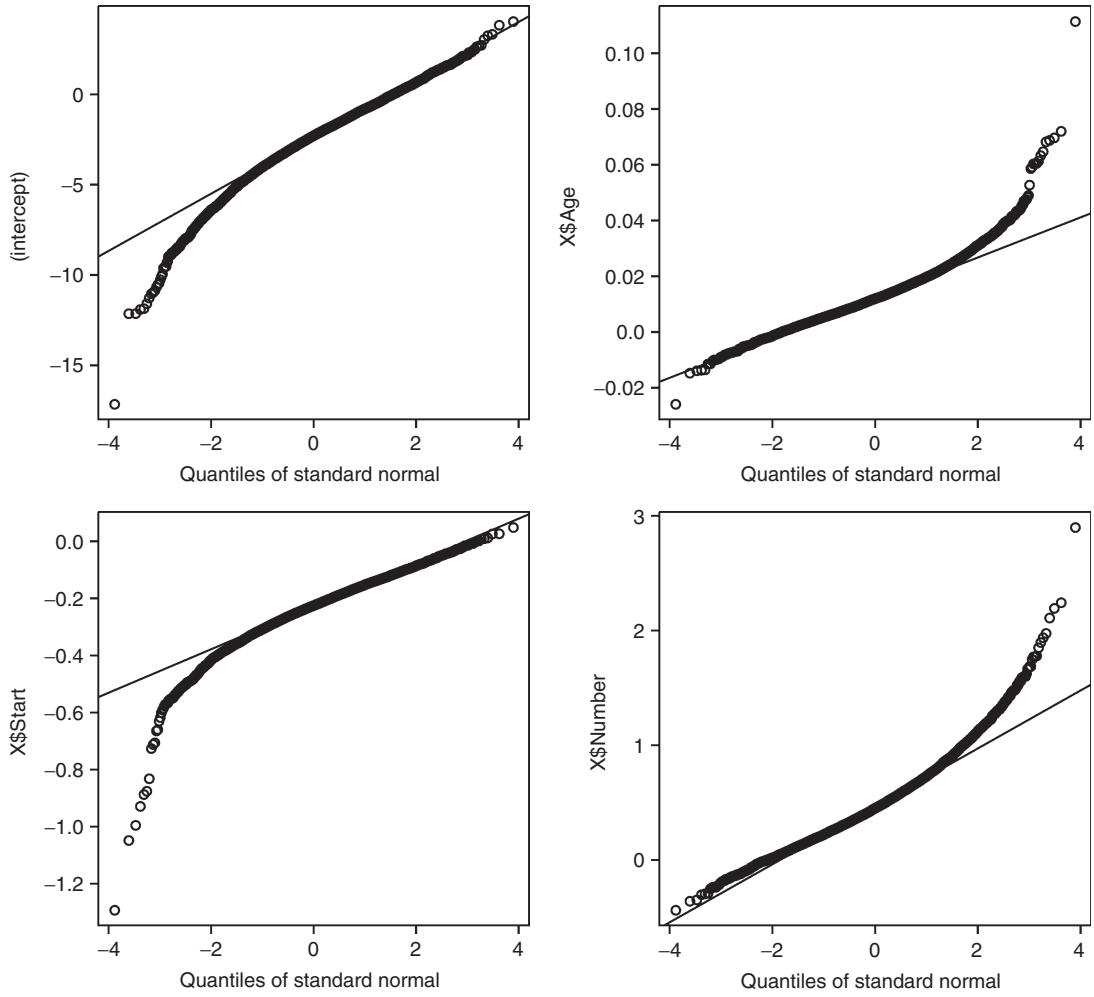


FIGURE 10 | Normal quantile plots of bootstrap distributions for logistic regression coefficients.

from a standard statistical package (S-PLUS) includes *t* values but omits *p* values. Yet it would be tempting for a package user to interpret the *t* coefficients as arising from a *t* distribution; the bootstrap demonstrates that this would be improper. The distributions are so non-normal as to make the utility of standard errors doubtful.

The numerical bootstrap results are:

	Observed	Mean	Bias	SE
(Intercept)	−2.03693	−2.41216	−0.375224	1.737216
Age	0.01093	0.01276	0.001827	0.008017
Start	−0.20651	−0.22991	−0.023405	0.084246
Number	0.41060	0.48335	0.072748	0.274049

The bootstrap standard errors are larger than the classical (asymptotic) standard errors by 20–24%. The distributions are also extremely biased, with absolute bias estimates ranging from 0.22 to 0.28 standard errors.

These results are for the conditional distribution bootstrap, a kind of parametric bootstrap. Repeating the analysis with the nonparametric bootstrap (resampling observations) yields bootstrap distributions that are even longer-tailed, indicating larger biases and standard errors. This reinforces the conclusion that classical normal-based inferences are not appropriate here.

ACCURACY OF BOOTSTRAP DISTRIBUTIONS

How accurate is the bootstrap? This entails two questions:

- How accurate is the theoretical bootstrap?
- How accurately does the Monte Carlo implementation approximate the theoretical bootstrap?

We begin this section with a series of pictures intended to illustrate both questions. We conclude this section with a discussion of cases where the theoretical bootstrap is not accurate, and remedies. In Section *How Many Bootstrap Samples Are Needed* we return to the question of Monte Carlo accuracy.

The treatment in this section is mostly not rigorous. There is a large literature that looks at the first question rigorously and asymptotically; we reference some of that work in other sections, particularly Section *Bootstrap Confidence Intervals* about confidence intervals, and also refer the reader to Refs 10, 11 and some sections of Ref 5, and the references therein.

Large Sample Mean

Figure 11 shows a population, and five samples of size 50 from the population, in the left column. The middle column shows the sampling distribution for the mean, and bootstrap distributions from each sample, based on $B = 1000$ bootstrap samples. Each bootstrap distribution is centered at the statistic ($\overline{x}$) from the corresponding sample rather than being centered at the population mean μ. The spreads and shapes of the bootstrap distributions vary a bit, but not a lot.

This informs what the bootstrap distributions may be used for. The bootstrap does not provide a better estimate of the population parameter μ, because no matter how many bootstrap samples are used, they are centered at $\overline{x}$ (plus random variation), not μ. On the other hand, the bootstrap distributions are useful for estimating the spread and shape of the sampling distribution.

The right column shows five more bootstrap distributions from the first sample; the first four using $B = 1000$ resamples, and the final using $B = 10^4$. These illustrate the Monte Carlo variation in the bootstrap. This variation is much smaller than the variation due to different original samples. For many uses, such as quick and dirty estimation of standard errors or approximate confidence intervals, $B = 1000$ resamples is adequate. However, there is noticeable variability, particularly in the tails of the bootstrap distributions, so when accuracy matters $B = 10^4$ or more samples should be used.

Note the difference between using $B = 1000$ and $B = 10^4$ bootstrap samples. These correspond to drawing samples of size 1000 or 10^4 observations, with replacement, from the theoretical bootstrap distribution. Using more samples reduces random Monte Carlo variation, but does not fundamentally change the bootstrap distribution—it still has the same approximate center, spread, and shape.

Small Sample Mean

Figure 12 is similar to Figure 11, but for a smaller sample size, $n = 9$ (and a different population). As before, the bootstrap distributions are centered at the corresponding sample means, but now the spreads and shapes of the bootstrap distributions vary substantially, because the spreads and shapes of the samples vary substantially. As before, the Monte Carlo variation is small, and may be reduced using $B = 10^4$ or more samples.

It is useful to compare the bootstrap distributions to classical statistical inferences. With classical t intervals of the form $\overline{x} \pm t_{\alpha/2} s/\sqrt{n}$, the confidence interval width varies substantially in small samples as the sample standard deviation s varies. Similarly, the classical standard error $s/\sqrt{n}$ varies. The bootstrap is no different in this regard—bootstrap standard errors and widths of confidence intervals for the mean are proportional to s.

Where the bootstrap does differ from classical inferences is how it handles skewness. The bootstrap percentile interval, and other bootstrap confidence intervals discussed below in Section *Bootstrap Confidence Intervals*, are in general asymmetrical with asymmetry depending on the sample. They estimate the population skewness from the sample skewness. In contrast, classical t intervals assume that the population skewness is zero. In Bayesian terms, the bootstrap uses a noninformative prior for skewness, while classical procedures use a prior with 100% of its mass on skewness = 0.

Which is preferred? In large samples, clearly the bootstrap. In small samples, the classical procedure may be preferred. If the sample size is small, then skewness cannot be estimated accurately from the sample, and it may be better to assume skewness = 0 in spite of the bias, rather than to use an estimate that has high variability.

Sample Median

Now turn to Figure 13, where the statistic is the sample median. Here the bootstrap distributions are

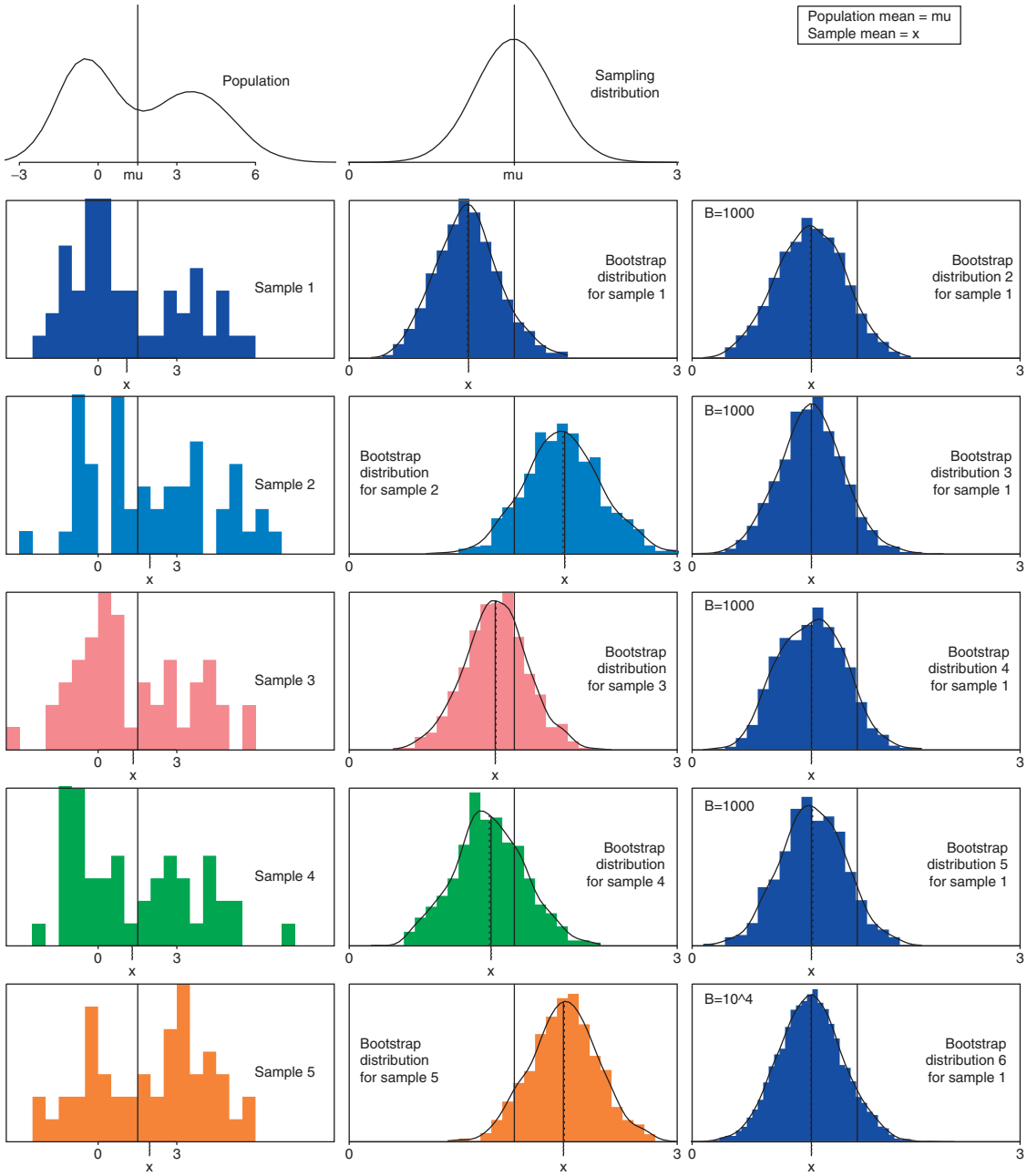


FIGURE 11 | *Bootstrap distribution for the mean,* $n = 50$. The left column shows the population and five samples. The middle column shows the sampling distribution, and bootstrap distributions from each sample. The right column shows five more bootstrap distributions from the first sample, with $B = 1000$ or $B = 10^4$.

poor approximations of the sampling distribution. In contrast, the sampling distribution is continuous, but the bootstrap distributions are discrete, with the only possible values being values in the original sample (here n is odd). The bootstrap distributions are very sensitive to the sizes of gaps among the observations near the center of the sample.

The ordinary bootstrap tends not to work well for statistics such as the median or other quantiles that depend heavily on a small number of observations out of a larger sample.

In the case of the median and other interior quantiles, this can be remedied using a *smoothed bootstrap*,[12,13] drawing samples from a density estimate based on the data, rather than drawing from the data itself. Smoothing is less effective for more extreme quantiles, where the bootstrap distribution would still depend heavily on a small number of

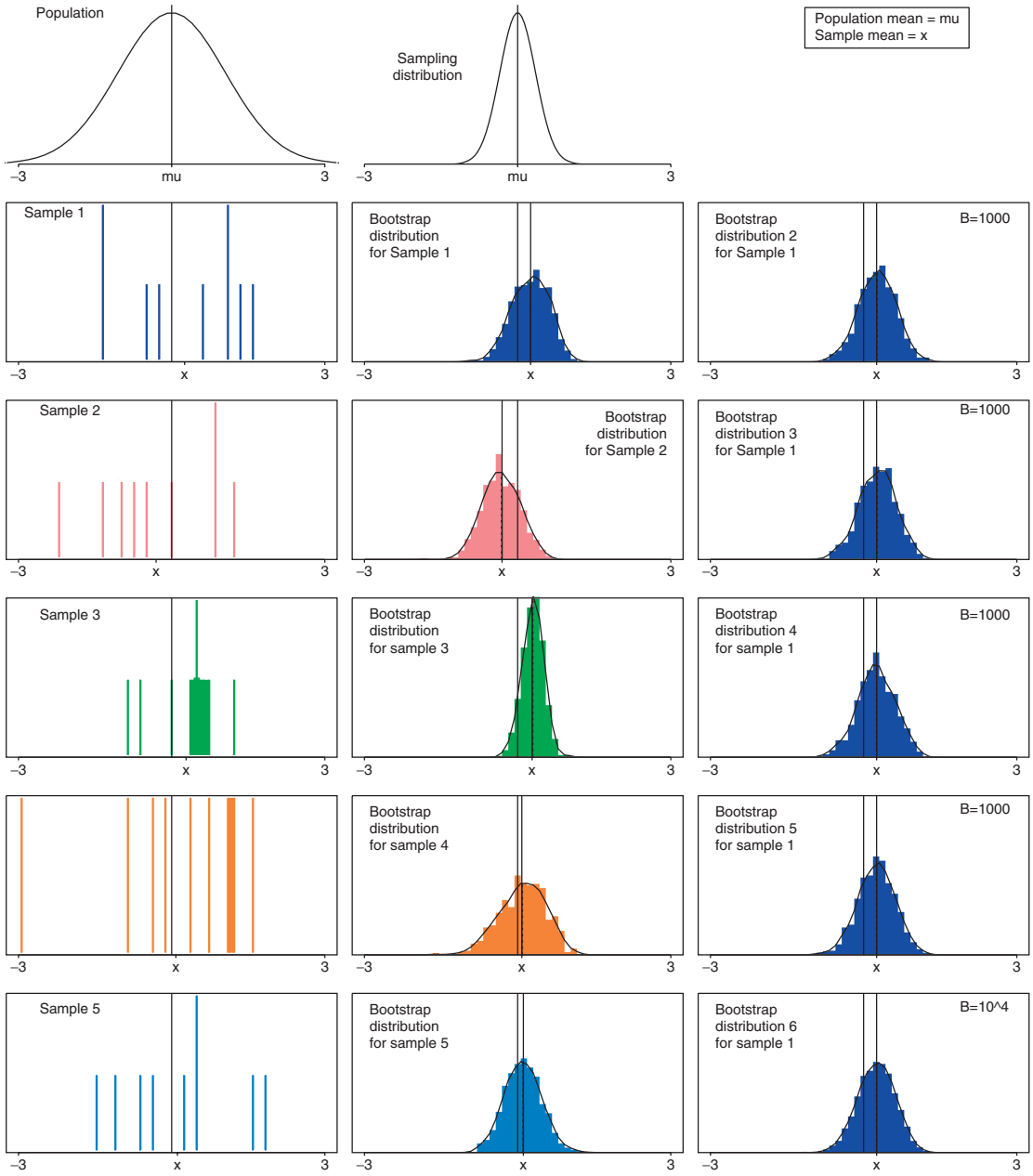


FIGURE 12 | *Bootstrap distributions for the mean, $n = 9$.* The left column shows the population and five samples. The middle column shows the sampling distribution, and bootstrap distributions from each sample. The right column shows five more bootstrap distributions from the first sample, with $B = 1000$ or $B = 10^4$.

observations. In that case it may be necessary to impose additional structure by assuming a parametric family, and perform a parametric bootstrap.

Trust the Data or Impose Structure?

In general, when there is little data, or a statistic of interest depends on a small subset of a larger data set, then it may be appropriate to make additional assumptions such as smoothness or a parametric family. When there is a lot of data you can trust the data to represent the shape of the population; when there is less data you cannot.

Systematic Errors in Bootstrap Distributions

We note three ways that bootstrap distributions are systematically different than sampling distributions. First, as noted above, bootstrap distributions are centered at the statistic $\hat{\theta}$ (plus bias) rather than at the parameter θ (plus bias).

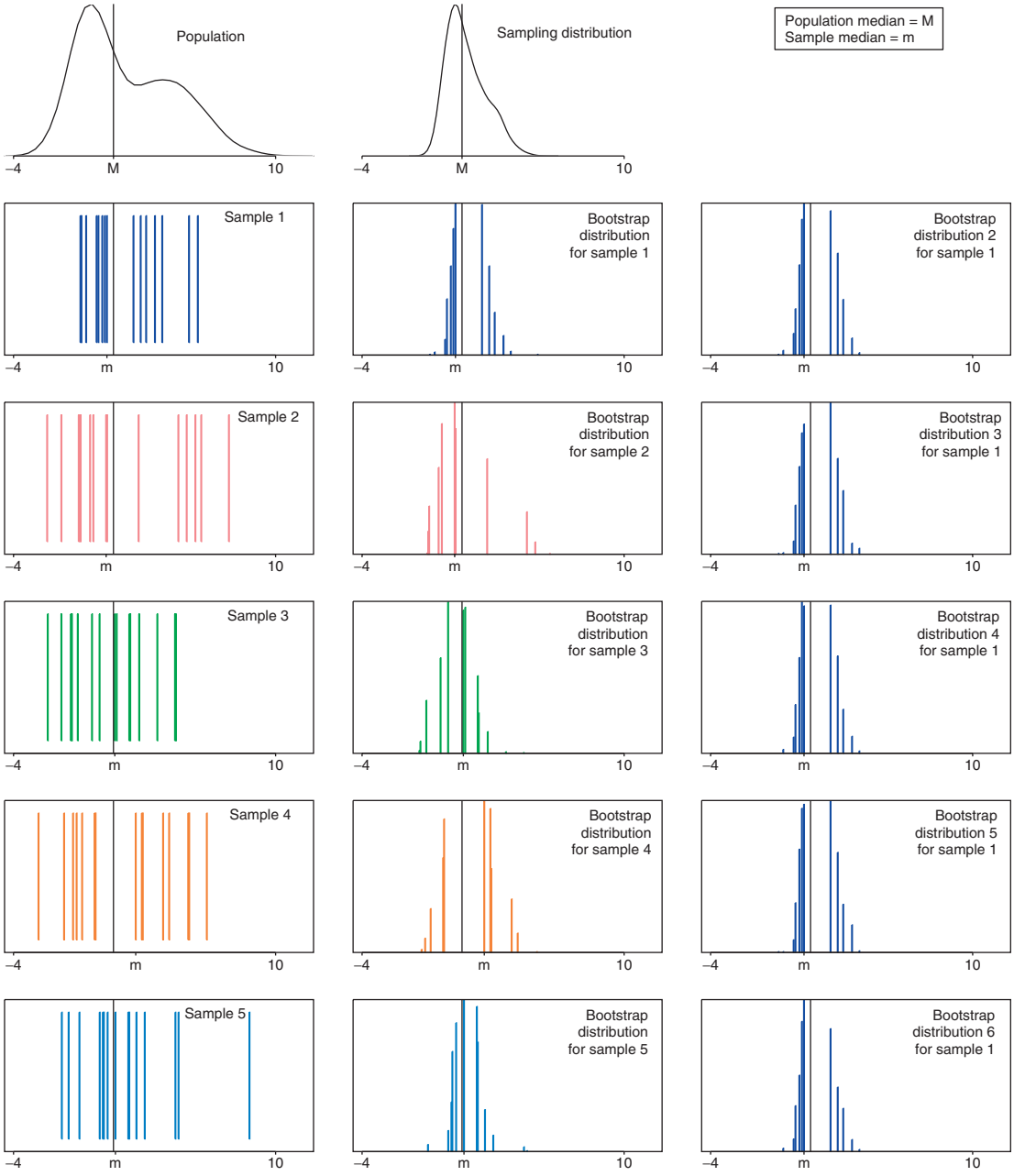


FIGURE 13 | *Bootstrap distributions for the median, n* = 15. The left column shows the population and five samples. The middle column shows the sampling distribution, and bootstrap distributions from each sample. The right column shows five more bootstrap distributions from the first sample.

Second, in many applications there is a relationship between the statistic and its standard error (''acceleration'' in the terminology of Ref 14). For example, the standard error of a binomial proportion $\sqrt{\hat{p}(1-\hat{p})/n}$ depends on $\hat{p}$. Similarly, when sampling from a gamma distribution, the variance of the sample mean depends on the underlying mean. More generally when sampling the mean from positively skewed distributions, samples with larger means tend to give larger standard errors.

When there is acceleration, the bootstrap standard error reflects the standard error corresponding to $\hat{\theta}$, not the true standard deviation of the sampling distribution (corresponding to θ). Suppose the relationship is positive; then when $\hat{\theta} < \theta$ it tends to be true that the estimated standard error is also less than the true standard deviation of the sampling distribution,

and confidence intervals tend to be too short. This is true for t intervals, whether using a formula or bootstrap standard error, and also to a lesser extent for bootstrap percentile intervals. The more accurate intervals discussed in Section *Bootstrap Confidence Intervals* correct for acceleration.

The third systematic error is that bootstrap distributions tend to be too narrow.

Bootstrap Distributions Are Too Narrow

In small samples, bootstrap distributions tend to be too narrow. Consider the case of a sample mean from a single population; in this case the theoretical bootstrap standard error is $\hat{\sigma}/\sqrt{n}$ where $\hat{\sigma}^2 = (1/n)\sum(x_i - \overline{x})^2$.[4] In contrast to the usual sample standard deviation s, this uses a divisor of n rather than $n-1$.

The reason the distributions are too narrow relates to the plug-in principle; when plugging in the empirical distribution $\hat{F}_n$ for use as the population, we are drawing samples from a population with standard deviation $\hat{\sigma}$.

The result is that bootstrap standard errors are too small, by a factor $\sqrt{1-1/n}$ relative to the usual $s/\sqrt{n}$; about 5% too small when $n = 10$, about 1% too small when $n = 50$, etc.

In stratified bootstrap situations the bias depends on the strata sizes rather than on the total sample size.

There are some easy remedies. The first is to draw bootstrap samples of size $n-1$, with replacement from the data of size n. The second, *bootknife sampling*,[15] is a combination of jackknife and bootstrap sampling—first create a jackknife sample by omitting an observation, then draw a bootstrap sample of size n with replacement from the $n-1$ remaining observations. The omission can be random or systematic.

A third remedy is the smoothed bootstrap. Instead of drawing random samples from the discrete distribution $\hat{F}_n$, we draw from a kernel density estimate $\hat{F}_h(x) = n^{-1}\sum\Phi((x-x_i)/h)$, where Φ is the standard normal density (other densities may be used). The original motivation[12,13] was to draw samples from continuous distributions, but it can also be used to correct for the downward bias of bootstrap standard errors.[15] The variance of an observation from $\hat{F}_h$ is $\hat{\sigma}^2 + h^2$. Using $h^2 = s^2/n$ makes the theoretical bootstrap standard error for the mean match the usual formula standard error.[15] For multidimensional data $\mathbf{x}$ the kernel covariance can be $1/n$ times the empirical covariance matrix. For non-normal data it may be appropriate to smooth on a transformed scale; e.g., for failure time data, to take a log transform of the failure times, add normal noise, then transform back to the original scale.

BOOTSTRAP CONFIDENCE INTERVALS

A large number of bootstrap confidence intervals have been proposed in the literature. Reviews of confidence intervals are found in Refs 16–18. Here we focus on five: t intervals with either bootstrap or formula standard errors, bootstrap percentile intervals,[1] bootstrap t intervals,[2,4,19] bootstrap BCa (bias-corrected, accelerated) intervals,[14] and bootstrap tilting.[19–22]

Note that "t intervals with bootstrap standard errors" and "bootstrap t intervals" are different.

Percentile and t intervals are quick-and-dirty intervals, relatively simple to compute, but are not very accurate except for very large samples. They do not properly account for factors such as bias, acceleration, or transformations. They are *first-order correct*—under fairly general circumstances (basically, for asymptotically normal statistics) the one-sided non-coverage levels for nominal $(1-\alpha)$ intervals are $\alpha/2 + O(1/\sqrt{n})$. The $O(1/\sqrt{n})$ errors decrease to zero very slowly.

The BCa, tilting, and bootstrap t intervals are *second-order correct*, with coverage errors $O(1/n)$.

The percentile, BCa, and tilting intervals are *transformation invariant*—they give equivalent results for different transformations of a statistic, e.g., hazard ratio and log-hazard ratio, or relative risk and log relative risk. t intervals are not transformation invariant. Bootstrap t intervals are less sensitive to transformations than are t intervals; the use of different (smooth) transformations has coverage effects of order $O(1/n)$, compared to $O(1/\sqrt{n})$ for t intervals.

Our focus is on one-sided errors because few practical situations are truly two-sided. A nominal 95% interval that misses 2% of the time on the left and 3% of the time on the right should not be considered satisfactory. It is a *biased confidence interval*—both endpoints are too low, so it gives a biased impression about where the true parameter may be. The appropriate way to aggregate one-sided coverage errors is by adding their absolute values, so the biased interval has a total coverage error of $|2-2.5|\% + |3-2.5|\% = 1\%$, not 0%.

t Intervals

A t interval is of the form

$$\hat{\theta} \pm t_{\alpha/2,\nu} s_{\hat{\theta}}. \tag{8}$$

where $s_{\hat{\theta}}$ is a standard error computed using a formula or using the bootstrap, and ν is degrees of freedom,

typically set to $n-1$ (although other values would be better for non-normal distributions).

The bootstrap standard error may be computed using the techniques in Section *Bootstrap Distributions Are Too Narrow*—bootknife, sampling with reduced size, or smoothed bootstrap. This results in slightly wider intervals that are usually more accurate in practice. These techniques have an $O(1/n)$ effect on one-sided coverage errors, which is unimportant for large samples but is important in small samples. For example, for a sample of independent identically distributed observations from a normal distribution, a nominal 95% t interval for the mean using a bootstrap standard error without these corrections would have one-sided coverages errors:

n	Non-coverage	Error
10	0.0302	0.0052
20	0.0277	0.0027
40	0.0264	0.0014
100	0.0256	0.0006

Percentile Intervals

In its simplest form, a 95% bootstrap percentile interval is the range of the middle 95% of a bootstrap distribution.

More formally, bootstrap percentile intervals are of the form

$$(\hat{G}^{-1}(\alpha/2), \hat{G}^{-1}(1-\alpha/2)). \tag{9}$$

where $\hat{G}$ is the estimated bootstrap distribution of $\hat{\theta}^*$.

Variations are possible that improve finite-sample performance.[22] These have received little attention in the bootstrap literature, which tends to focus on asymptotic properties. In particular, the simple bootstrap percentile intervals tend to be too narrow, and the variations give wider intervals with better coverage.

First, the bootknife or other techniques in Section *Bootstrap Distributions Are Too Narrow* may be used.

Second, the percentiles may be adjusted. In a simple situation like the sample mean from a symmetric distribution the interval is similar to the t interval (8) but using quantiles of a normal distribution rather than t distribution, $z_{\alpha/2}$ rather than $t_{\alpha/2,n-1}$. As a result, the interval tends to be too narrow. A correction is to adjust the quantiles based on the difference between a normal and t distribution,

$$(\hat{G}^{-1}(\alpha'/2), \hat{G}^{-1}(1-\alpha'/2)). \tag{10}$$

where $\Phi^{-1}(\alpha'/2) = F^{-1}_{t,n-1}(\alpha/2)$ where Φ is the standard normal distribution and $F_{t,n-1}$ is the t distribution function with $n-1$ degrees of freedom. This gives wider intervals. Extensive simulations[22] show that this gives smaller coverage errors in practice, in a wide variety of applications. The effect on coverage errors is $O(1/n)$, the same order as the bootknife adjustment, but the magnitude of the effect is larger; for example, the errors caused by using z rather than t quantiles in a standard t interval for a normal population are:

n	Non-coverage	Error
10	0.0408	0.0158
20	0.0324	0.0074
40	0.0286	0.0036
100	0.0264	0.0014

For a sample size of 20, this effect alone makes intervals tend to miss $0.0074/0.025 = 30\%$ too often!

A third variation relates to how quantiles are calculated for a finite number B of bootstrap samples. Hyndman and Fan[23] give a family of definitions of quantiles for finite samples, governed by a parameter $0 \le \delta \le 1$. The bth order statistic $\hat{\theta}^*_{(b)}$ is the $(b-\delta)/(B+1-2\delta)$ quantile of the bootstrap distribution, for $b = 1, \ldots, B$. Linear interpolation between adjacent bootstrap statistics is used if the desired quantile is not of the form $(b-\delta)/(B+1-2\delta)$ for some integer b. For bootstrap confidence intervals $\delta = 0$ is preferred, as other choices result in lower coverage probability. The effect on coverage errors is $O(1/B)$.

Bootstrap *t*

The difference between t intervals (possibly using bootstrap standard errors) and *bootstrap t* intervals[2,4,19] is that the former assume that a t statistic follows a t distribution, while the latter estimate the actual distribution using the bootstrap.

Let

$$t = \frac{\hat{\theta} - \theta}{s_{\hat{\theta}}} \tag{11}$$

be a t statistic. Under certain conditions the t statistic follows a t distribution. Those conditions are rarely met in practice.

The bootstrap analog of t is

$$t^* = \frac{\hat{\theta}^* - \hat{\theta}}{s_{\hat{\theta}^*}}. \tag{12}$$

The standard error may be calculated either by formula or bootstrap sampling; in the latter case, calculating each $s_{\hat{\theta}^*}$ requires a second level of bootstrap sampling, with second-level bootstrap samples drawn from each first-level bootstrap sample.

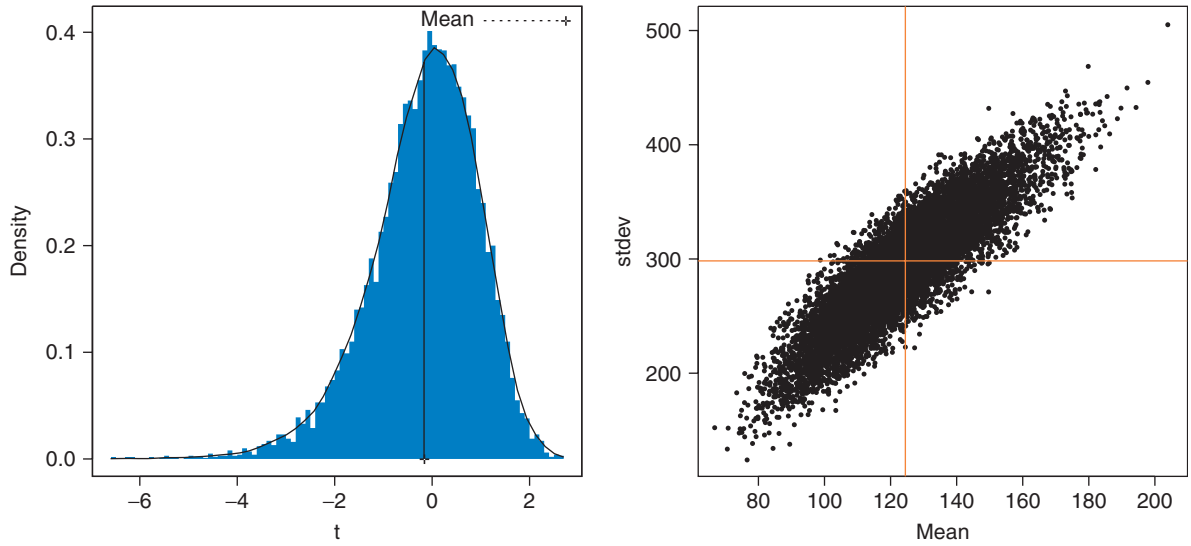


FIGURE 14 | Histogram of bootstrap distribution for the t statistic, and relationship between bootstrap means and standard deviations, of arsenic concentrations.

Figure 14 shows the bootstrap distribution for the t statistic for mean arsenic concentration, where t is the ordinary t statistic $(\overline{x} - \mu)/(s/\sqrt{n})$. In contrast to Figure 2, where the bootstrap distribution for the mean is positively skewed, the distribution for the t statistic is negatively skewed. The reason is that there is positive correlation between $\overline{x}^*$ and s^*, as seen in the right panel of Figure 14, so that a negative numerator in (12) tends to occur with a small denominator.

The bootstrap t interval is based on the identity

$$P(G_t^{-1}(\alpha/2) < \frac{\hat{\theta} - \theta}{s_{\hat{\theta}}} < G_t^{-1}(1 - \alpha/2)) = 1 - \alpha, \tag{13}$$

where G_t is the sampling distribution of t (11). Assuming that t^* (12) has approximately the same distribution as t, we substitute quantiles of the bootstrap distribution for t^*; then solving for θ yields the bootstrap t interval

$$(\hat{\theta} - G_{t^*}^{-1}(1 - \alpha/2)s_{\hat{\theta}}, \hat{\theta} - G_{t^*}^{-1}(\alpha/2)s_{\hat{\theta}}). \tag{14}$$

Note that the right tail of the bootstrap distribution of t^* is used in computing the left side of the confidence interval, and conversely.

The bootstrap t and other intervals for the mean arsenic concentration example described in Section *Introduction* are shown in Table 1.

It is not appropriate to use bootknife or other sampling methods in Section *Bootstrap Distributions Are Too Narrow* with the bootstrap t. The reason we use those methods with the other intervals is because those intervals are too narrow if the plug-in population is narrower, on average, than the parent population. The sampling distribution of a t statistic, in contrast, is invariant under changes in the scale of the parent population. This gives it an automatic correction for the plug-in population being too narrow, and to add bootknife sampling would over-correct.

Efron and Tibshirani[2] note that the bootstrap t is sometimes erratic, and suggest transforming the statistic of interest. Hesterberg[22] observes erratic behavior in small samples. We conjecture the following explanation—that the bootstrap t depends not only on skewness, but also on kurtosis, and kurtosis is hard to estimate from small samples. The bootstrap t does not use a t table, but instead estimates the distribution of the t statistic by simulating from the data. This distribution depends not only on asymmetry caused by skewness, but also on the effective degrees of freedom, that depend on kurtosis—larger kurtosis results in greater variability in standard errors and smaller effective degrees of freedom. In contrast, other second-order-correct intervals depend on skewness, but not (or much less so) on kurtosis, so are less erratic for small samples.

BCa Intervals

The bootstrap BCa interval[14] uses quantiles of the bootstrap distribution, like the percentile interval, but with the percentiles adjusted depending on a bias parameter z_0 and acceleration parameter a. The interval is

$$(G^{-1}(p(\alpha/2)), G^{-1}(p(1 - \alpha/2))), \tag{15}$$

where

$$p(c) = \Phi\left(z_0 + \frac{z_0 + \Phi^{(-1)}(c)}{1 - a(z_0 + \Phi^{(-1)}(c))}\right) \quad (16)$$

is the adjusted probability level for quantiles; it simplifies to c when $z_0 = a = 0$.

The BCa interval is derived by assuming there exists a smooth transformation h such that

$$h(\hat{\theta}) \sim N(h(\theta) + z_0\sigma_h, \sigma_h^2), \quad (17)$$

where $\sigma_h = 1 + ah(\theta)$ and that the same relationship holds for bootstrap samples (substitute $\hat{\theta}^*$ for $\hat{\theta}$, and $\hat{\theta}$ for θ). Some algebra yields the BCa confidence interval. The transformation h cancels out, so need not be estimated.

For the nonparametric bootstrap, the parameter z_0 is usually estimated using the fraction of bootstrap observations that fall below the original observed value,

$$z_0 = \Phi^{(-1)}(\#(\hat{\theta}^* < \hat{\theta})/B) \quad (18)$$

and acceleration parameter based on the skewness of the empirical influence function. One estimate of that skewness is obtained from jackknife samples; let $\hat{\theta}_{(i)}$ be the statistic calculated from the original sample but excluding observation i, and $\overline{\theta}_{(i)}$ be the average of those values, then

$$a = \frac{-\sum_{i=1}^{n}(\hat{\theta}_{(i)} - \overline{\theta}_{(i)})^3}{6(\sum_{i=1}^{n}(\hat{\theta}_{(i)} - \overline{\theta}_{(i)})^2)^{3/2}}. \quad (19)$$

Davison and Hinkley[5] also give expressions for a in the case of stratified sampling, including two-sample applications.

For the arsenic data, $z_0 = 0.0438$ (based on 100,000 replications) and $a = 0.0484$. The 95% interval is then the range from the 0.0436 to 0.988 quantiles of the bootstrap distribution.

The BCa interval has greater Monte Carlo error than the ordinary percentile interval because Monte Carlo error in estimating z_0 propagates into the endpoints, and because typically one of the quantiles is farther in the tail than for the percentile interval, e.g., here the 98.8% quantile is used instead of the 97.5% quantile. In the best case, that $a = z_0 = 0$, this requires a bit more than twice as many bootstrap samples as the percentile interval for comparable Monte Carlo accuracy.

Bootstrap Tilting Intervals

In parametric statistics, the left endpoint of a confidence interval for a parameter θ is the value that makes the sampling distribution have probability 2.5% of exceeding the observed value, $P_{\theta_{\text{left}}}(\hat{\theta}^* > \hat{\theta}) = 0.025$. Bootstrap tilting[19] borrows this idea.

The idea behind bootstrap tilting[19] is to create a one-parameter family of populations that includes the empirical distribution function, to find the member of that family that has 2.5% (or 97.5%) of the bootstrap distribution exceeding the observed value, and let the left (right) endpoint of the interval be the parameter of interest calculated from that population. The family is restricted to have support on the empirical data, with varying probabilities on the observations.

In effect, the procedure turns a nonparametric problem into a parametric problem, then uses classical one-parametric techniques. This turns out to be accurate statistically, giving second-order-accurate confidence intervals,[20] and with a clever implementation[19,21] requires a very small number of bootstrap samples, roughly 17 times fewer than a bootstrap percentile interval, but there are implementation difficulties and different implementations can very dramatically in small-sample accuracy.[22]

The following sections on bootstrap tilting may be skipped by most readers, but is useful background for Section *Planning Clinical Trials* on planning clinical trials.

Tilting For a Sample Mean

For example, given i.i.d. observations $(x_1, \ldots, x_n)$ when the parameter of interest is the population mean, one suitable family is the exponential tilting family, which places probabilities

$$p_i = c\exp(\tau x_i) \quad (20)$$

on observation i, where τ is a *tilting parameter*, and c is a normalizing constant (depending on τ) such that $\sum_i p_i = 1$.

$\tau = 0$ gives equal probabilities $p_i = 1/n$, corresponding to the empirical distribution function, and about half of the bootstrap distribution is below the observed $\overline{x}$. $\tau < 0$ places higher probabilities on smaller observations; sampling with these probabilities is more likely to give samples with smaller observations, and smaller bootstrap means, so more of the bootstrap distribution is below $\overline{x}$. We find the values of τ for which only 2.5% of the bootstrap distribution is above $\overline{x}$; The left endpoint of the confidence interval is the mean of the corresponding weighted population, $\theta_{\text{left}} = \sum_{i=1}^{n} p_i x_i$.

Similarly, the right endpoint is $\sum_{i=1}^{n} p_i x_i$ when p_i is computed using the τ that puts 97.5% of the bootstrap distribution to the right of $\overline{x}$.

Another suitable family is the maximum likelihood family, with probability

$$p_i = \frac{c}{1 - \tau(x_i - \overline{x})} \tag{21}$$

on observation i.

Importance Sampling Implementation

Conceptually, finding the right value of τ requires trial and error; for any given τ, we calculate $\mathbf{p} = (p_1, \ldots, p_n)$, draw bootstrap samples with those probabilities, calculate the bootstrap statistics, and calculate the fraction of those statistics that are above $\hat{\theta}$, then repeat with a different τ until the fraction is 2.5%. This is expensive, and the fraction varies due to random sampling.

In practice we use an importance sampling implementation. Instead of sampling with unequal probabilities, we sample with equal probabilities, then reweight the bootstrap samples by the relative likelihood of the sample under weighted and ordinary bootstrap sampling. The likelihood for a bootstrap sample is

$$l(x_{1^*}, \ldots, x_{n^*}) = \prod w_{i^*} \tag{22}$$

compared to $(1/n)^n$ for ordinary bootstrap sampling. Let $w_b = (\prod w_{i^*})/(1/n)^n = \prod n w_{i^*}$ be the relative likelihood for bootstrap sample b. We estimate the probability by

$$\hat{P}_{\mathbf{p}}(\hat{\theta}^* > \hat{\theta}) = B^{-1} \sum_{b=1}^{B} w_b I(\hat{\theta}_b^* > \hat{\theta}) = B^{-1} \sum_{b \in R} w_b, \tag{23}$$

where R is the subset of $\{1, \ldots, B\}$ with $\hat{\theta}_b^* > \hat{\theta}$.

In practice we also worry about ties, cases with $\hat{\theta}^* = \hat{\theta}$; let E be the subset with $\hat{\theta}_b^* = \hat{\theta}$. We numerically find τ to solve $0.025B = \sum_{b \in R} w_b + (1/2) \sum_{b \in E} w_b$.

Similar calculations are done for the τ used for the right endpoint; solve $0.025B = \sum_{b \in L} w_b + (1/2) \sum_{b \in E} w_b$ where L is the subset of $\{1, \ldots, B\}$ with $\hat{\theta}_b^* < \hat{\theta}$.

In any case, after finding τ, the endpoint of the interval is the weighted mean for the empirical distribution with probabilities calculated using τ.

Tilting for Nonlinear Statistics

The procedure can be generalized to statistics other than the mean using a least-favorable single-parameter family, one for which inference within the family is not easier, asymptotically, than for the original problem.[19] This is best done in terms of derivatives. Let $F_{\mathbf{p}}$ denote a weighted distribution with probability p_i on original data point x_i, $\theta(\mathbf{p}) = \theta(F_{\mathbf{p}})$ be the parameter for the weighted distribution (e.g., weighted mean, or weighted regression coefficient), and $\mathbf{p}_0 = (1/n, \ldots, 1/n)$ correspond to the original equal-probability empirical distribution function. The gradient of $\theta(\mathbf{p})$ is

$$U_i(\mathbf{p}) = \lim_{\epsilon \to 0} \epsilon^{-1}(\theta(\mathbf{p} + \epsilon(\delta_i - \mathbf{p})) - \theta(\mathbf{p})), \tag{24}$$

where δ_i is the vector with 1 in position i and 0 elsewhere. When evaluated at $\mathbf{p}_0$ these derivatives are known as the empirical influence function, or infinitesimal jackknife.

Four least-favorable families found in the tilting literature are:

$$\begin{aligned}
\mathcal{F}_1 &: p_i = c \exp(\tau U_i(\mathbf{p}_0)) \\
\mathcal{F}_2 &: p_i = c \exp(\tau U_i(\mathbf{p})) \\
\mathcal{F}_3 &: p_i = c(1 - \tau U_i(\mathbf{p}_0))^{-1} \\
\mathcal{F}_4 &: p_i = c(1 - \tau U_i(\mathbf{p}))^{-1},
\end{aligned} \tag{25}$$

each indexed by a tilting parameter τ, where each c normalizes the probabilities to add to 1.

$\mathcal{F}_1$ and $\mathcal{F}_2$ are well-known as "exponential tilting", and coincide with (20) if θ is a mean. Similarly, $\mathcal{F}_3$ and $\mathcal{F}_4$ are maximum likelihood tilting and coincide with (21) for a mean. $\mathcal{F}_2$ and $\mathcal{F}_4$ minimize the backward and forward Kullback–Leibler distances between $\mathbf{p}$ and $\mathbf{p}_0$, respectively, subject to $p_i \geq 0$, $\sum p_i = 1$, and $\theta(\mathbf{p}) = A$; varying A results in solutions of the form given in (25). $\mathcal{F}_4$ also maximizes the likelihood $\prod p_i$ subject to the same constraints.

As in the case of the sample mean, having selected a family, we find the value of τ for which 2.5% (95%) of the bootstrap distribution is to the right of the observed $\hat{\theta}$; the left (right) endpoint of the confidence interval is then the parameter calculated for the weighted distribution with probability p_i on x_i.

All four families result in second-order-accurate confidence intervals,[20] but the finite-sample performance differs, sometimes dramatically for smaller samples.[22] The differences in coverage between the four families are $O(1/n)$, similar to the adjustments discussed in Section *Percentile Intervals*. The fixed-derivative versions $\mathcal{F}_1$ and $\mathcal{F}_3$ are easier to work with, but have inferior statistical properties; they are shorter, have actual coverage probability lower than the nominal confidence, and for sufficiently high nominal confidence levels the actual coverage can decrease as the nominal confidence increases.

TABLE 1 | Confidence Intervals for Mean Arsenic Concentration, Based on 100,000 Bootstrap Samples, Using Ordinary Nonparametric and Bootknife Resampling

	95% Interval	Asymmetry
Formula *t*	(88.8, 160.2)	±35.7
Ordinary Bootstrap		
t w boot SE	(88.7, 160.2)	±35.8
Percentile	(91.5, 162.4)	(−33.0, 38.0)
Bootstrap *t*	(94.4, 172.6)	(−30.1, 48.1)
BCa	(95.2, 169.1)	(−29.3, 44.6)
Tilting	(95.2, 169.4)	(−29.3, 44.9)
Bootknife		
t w boot SE	(88.7, 160.3)	±35.8
Percentile	(91.5, 162.6)	(−32.9, 38.1)
BCa	(95.4, 169.3)	(−29.1, 44.8)
Tilting	(95.2, 169.4)	(−29.3, 45.0)

The "asymmetry" column is obtained by subtracting the observed mean. The "*t* w boot SE" interval is a *t* interval using a bootstrap standard error.

Similarly, exponential tilting is more convenient numerically, but maximum likelihood has better statistical properties, producing wider confidence intervals closer to the desired coverage levels. Overall, the maximum likelihood version with changing derivatives $\mathcal{F}_4$ gives the widest intervals with highest and usually most accurate coverage.

Confidence Intervals for Mean Arsenic Concentration

Table 1 shows 95% confidence intervals for the mean arsenic concentration example described in Section *Introduction*.

The intervals vary dramatically, particularly in the degree of asymmetry. The *t* intervals are symmetric about $\bar{x}$. The bootstrap *t* interval reaches much farther to the right, and is much wider. The percentile interval is asymmetrical, longer on the right side, to a lesser extent than other asymmetrical intervals. While it is asymmetrical, it is not asymmetrical enough for good accuracy. While preferable to the *t* intervals, it is not as accurate as the second-order-accurate procedures

The *t* intervals assume that the underlying population is normal, which is not true here. Still, the common practice with a sample size as large as 271 would be to use *t* intervals anyway. The bootstrap can help answer whether that is reasonable, by giving an idea what the actual non-coverage is for a 95% *t* interval. Table 2 shows what nominal coverage levels would be needed for the bootstrap *t* and BCa intervals to coincide with the actual endpoints of the

TABLE 2 | Actual Non-Coverage of Nominal 95% *t* Intervals, as Estimated From Second-Order-Accurate Intervals

Estimated using	Left	Right
Bootstrap *t*	0.0089	0.062
BCa	0.0061	0.052

A *t* interval would miss more than twice too often on the right side. The actual non-coverage should be 0.025 on each side.

t interval—in other words, what the bootstrap *t* and BCa intervals think is the actual non-coverage of the *t* intervals. The discrepancies are striking. On the left side, the *t* interval should miss 2.5% of the time; it actually misses only about a third or fourth that often, according to the bootstrap *t* and BCa intervals. On the right side, it should miss 2.5% of the time, but actually misses somewhere between 5.2 and 6.2%, according to the BCa and bootstrap *t* procedures. This suggests that the *t* interval is severely biased, with both endpoints systematically lower than they should be.

Implications for Other Situations

The *t* intervals are badly biased in the arsenic example. What does this imply for other situations?

On the one hand, the arsenic data are quite skewed, relative to most data observed in practice. On the other hand, the sample size is large. What can we say about other combinations of sample size and population skewness?

For comparison, samples of size 47 from an exponential population are comparable to the arsenic data, in the sense that the sampling distribution for the mean is equally skewed. A quick simulation with 10^6 samples of exponential data with $n = 47$ shows that the actual non-coverage of 95% *t* intervals is 0.0089 on the left and 0.0567 on the right, comparable to the bootstrap estimates above. This shows that for a distribution with only moderate skewness, like the exponential distribution, $n = 30$ is not nearly enough to use *t* intervals; that even $n = 47$ results in non-coverage probabilities that are off by factors of about 3 and 2, on the two sides. Reducing the errors in non-coverage to a more reasonable 10% of the desired value, i.e., that the actual one-sided non-coverage probabilities are between 2.25 and 2.75% on each side for a nominal 95% interval, would require around $n = 5000$ for an exponential distribution.

Even for distributions that are not particularly skewed, say 1/4 the skewness of an exponential distribution (e.g., a gamma distribution with shape = 16), the sample size would need to be around 470 to reduce the errors in non-coverage to 10% of the desired values.

To obtain reasonable accuracy for smaller sample sizes requires the use of more accurate confidence intervals, either a second-order-accurate bootstrap interval, or comparable second-order-accurate non-bootstrap interval. Two general second-order-accurate procedures that do not require sampling are *ABC*[24] and *automatic percentile*[25] intervals, which are approximations for BCa and tilting intervals, respectively.

The current practice of statistics, using normal and t intervals with skewed data, systematically produces confidence intervals with endpoints that are too low (for positively skewed data).

Similarly, hypothesis tests are systematically biased; for positively skewed data they reject $H_0: \theta = \theta_0$ too often for cases with $\hat{\theta} < \theta_0$, and too little for $\hat{\theta} > \theta_0$. The primary reason is acceleration—when $\hat{\theta} < \theta_0$ then acceleration makes it likely that $s < \sigma$, and the t interval does not correct for this, so improperly rejects H_0.

Comparing Intervals

t intervals and bootstrap percentile intervals are quick-and-dirty intervals, suitable for rough approximations, but should not be used where accuracy is needed.

Among the others, I recommend the BCa in most cases, provided that the number of bootstrap samples B is very large.

In my experience with extensive simulations, the bootstrap t is the most accurate in terms of coverage probabilities. However, it achieves this at high cost—the interval is longer on average than the BCa and tilting intervals, often much longer. Adjusting the nominal coverage level of the BCa and tilting intervals upward gives comparable coverage to bootstrap t with shorter length. And the lengths of bootstrap t intervals vary much more than the others. I conjecture that this is because bootstrap t intervals are sensitive to the kurtosis of the bootstrap distribution, which is hard to estimate accurately from reasonable-sized samples. In contrast, BCa and tilting intervals depend primarily on mean, standard deviation, and skewness of the bootstrap distribution.

Also, the bootstrap t is computationally expensive if the standard error is obtained by bootstrapping. If $s_{\hat{\theta}}$ is calculated by bootstrapping, then $s_{\hat{\theta}^*}$ is calculated using a second level of bootstrapping—drawing bootstrap samples from each first-level bootstrap sample (requiring a total of $B + BB_2$ bootstrap samples, if B_2 second-level bootstrap samples from each of B first-level bootstrap samples).

The primary advantage of bootstrap tilting over BCa is that it requires many fewer bootstrap replications, typically by a factor of 37 for a 95% confidence interval. The disadvantages of tilting are that the small-sample properties of the fixed-derivative versions $\mathcal{F}_1$ and $\mathcal{F}_3$ are not particularly good, while the more rigorous $\mathcal{F}_2$ and $\mathcal{F}_4$ are harder to implement reliably.

HYPOTHESIS TESTING

An important point in bootstrap hypothesis testing is that sampling should be done in a way that is consistent with the null distribution.

We describe here three bootstrap hypothesis testing procedures: pooling for two-sample tests, bootstrap tilting, and bootstrap t.

The first is for two-sample problems, such as comparing two means. Suppose that the null hypothesis is that $\theta_1 = \theta_2$, and that one is willing to assume that if the null hypothesis is true that the two populations are the same. Then one may pool the data, draw samples of size n_1 and n_2 with replacement from the pooled data, and compute a test statistic such as $\hat{\theta}_1 - \hat{\theta}_2$ or a t statistic. Let T^* be the bootstrap test statistic, and T_0 the observed value of the test statistic. P-value is the fraction of time that the T^* exceeds T_0.

In practice we add 1 to the numerator and denominator when computing the fraction—the one-sided P-value for the one-sided alternative hypothesis $\hat{\theta}_1 - \hat{\theta}_2 > 0$ is $(\#(T^* > T_0) + 1)/(B + 1)$. The lower one-sided P-value is $(\#(T^* < T_0) + 1)/(B + 1)$, and the two-sided P-value is two times the smaller of the one-sided P-values.

This procedure is similar to the two-sample permutation test, which pools the data and draws n_1 observations without replacement for the first sample and allots the remaining n_2 observations to be the second sample. The permutation test is preferred. For example, suppose there is one outlier in the combined sample; every pair of permutation samples has exactly one copy of the outlier, while the bootstrap samples may have $0, 1, 2, \ldots$ copies. This adds extra variability not present in the original data, and detracts from the accuracy of the resulting P-values.

Now suppose that one is not willing to assume that the two distributions are the same. Then bootstrap tilting hypothesis testing[5,26,27] may be suitable. Tilting may also be used in one-sample and other contexts. The idea is to find a version of the empirical distribution function(s) with unequal probabilities that satisfy the null hypothesis (by maximizing likelihood or minimizing Kullback–Leibler distance subject to the null hypothesis), then draw samples from the unequal-probability empirical distributions, and let

the P-value be the fraction of times the bootstrap test statistic exceeds the observed test statistic. As in the case of confidence intervals, importance sampling may be used in place of sampling with unequal probabilities, see Section *Bootstrap Confidence Intervals*. There are close connections to empirical likelihood.[28]

Bootstrap tilting hypothesis tests reject H_0 if bootstrap tilting confidence intervals exclude the null hypothesis value.

The third general-purpose bootstrap testing procedure is related to bootstrap t confidence intervals. A t statistic is calculated for the observed data, and the P-value for the statistic is calculated not by reference to the Students t distribution, but rather by reference to the bootstrap distribution for the t statistic. In this case the bootstrap sampling need not be done consistently with the null hypothesis, because t statistics are approximately pivotal—their distribution is approximately the same independent of θ.

PLANNING CLINICAL TRIALS

The usual bootstrap procedure is to draw samples of size n from the empirical data, or more generally to plug in an estimate for the population and draw samples using the sampling mechanism actually used in practice. In planning clinical trials we may modify this in two ways:

- try other sampling procedures, such as different sample sizes or stratification, and/or
- plug-in alternate population estimates.

For example, given training data of size n, to estimate standard errors or confidence interval width that would result from a possible clinical trial of size N, we may draw bootstrap samples of size N with replacement from the data.

Similarly, we may estimate the effects of different sampling mechanisms, such as stratified sampling, or case–control allocation to arms, even if pilot data were obtained in other ways.

For example, we consider preliminary results from a clinical trial to evaluate the efficacy of maintenance chemotherapy for acute myelogenous leukemia (AML).[29,30] After achieving remission through chemotherapy, the patients were assigned to a treatment group receiving maintenance chemotherapy and a control group that did not. The goal was to see if maintenance chemotherapy prolonged the time until relapse. The data are in Table 3. There are 11 subjects in the treatment group and 12 in the control group.

TABLE 3 | Leukemia Data

Group	Length of Complete Remission (in Weeks)
Maintained	9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+
Nonmaintained	5, 5, 8, 8, 12, 16+, 23, 27, 30, 33, 43, 45

A Cox proportional hazards regression, using Breslow's method of breaking ties, yields a log-hazard ratio of 0.904 and standard error 0.512:

```
      coef  exp(coef) se(coef)    z     p
group 0.904     2.47     0.512 1.77 0.078
```

An ordinary bootstrap with $B = 10^4$ results in eleven samples with complete separation—where the minimum observed relapse time in the treatment group exceeds the maximum observed relapse in the control group, giving an infinite estimated hazard ratio. A stratified bootstrap reduces the number of samples with complete separation to three. Here stratification is preferred (even if the original allocation were not stratified) in order to condition on the actual sample sizes, and prevent imbalance in the bootstrap samples. Omitting the three observations results in a slightly long-tailed bootstrap distribution, with standard error 0.523, slightly larger than the formula standard error.

Drawing 50 observations from each group results in a bootstrap distribution for log-hazard ratio that is nearly exactly normal with almost no bias, no samples with separation (they are still possible, but unlikely), and a standard error of 0.221. Surprisingly, this 10% less than obtained by extrapolating the original formula standard error at the rate $1/\sqrt{n}$, $0.512/\sqrt{100/23} = 0.246$, and 12% less than obtained by extrapolating the original bootstrap standard error. Similar results are obtained using Efron's method for handling ties, and from a smoothed bootstrap with a small amount of noise added to the remission times. The fact that the reduction in standard error is 10--12% greater than expected may be because censored observations have a less serious impact with larger sample sizes.

"What if" Analyses—Alternate Population Estimates

In planning clinical trials it is often of interest to do "what if" analyses, perturbing various inputs. For example, how might the results differ under sampling from populations with a log-hazard ratio of zero, or 0.5?

This should be done by reweighting observations.[31,32] This is a version of bootstrap tilting[19,21,31,33] and is closely related to empirical likelihood.[34]

Consider first a simple example—sampling the difference in two means, $\hat{\theta} = \bar{x}_1 - \bar{x}_2$. In order to sample from populations with different values of θ, it is natural to consider perturbing the data, shifting one or both samples, e.g., adding $\theta - \hat{\theta}$ to each value in sample 1.

Perturbing the data does not generalize well to other situations. Furthermore, perturbing the data would often give incorrect answers. Suppose that the observations represent positive skewed observations such as survival times, with a mode at zero. Shifting one of the samples to the left would give negative times; to the right would make the mode nonzero. More subtle, but very important, is that shifting ignores the mean–variance relationships for skewed populations—increasing the mean should also increase the variance. For positive data like survival times, perturbing the data by multiplying one of the samples by a factor avoids the most obvious problems, but assumes a particular mean–variance relationship—that variance is proportional to the square of the mean.

It is also unclear how one would perturb the data in multivariate applications when some variables are categorical.

Instead, we suggest using a weighted version of the empirical data, maximizing the likelihood of the observed data subject to the weighted distributions satisfying desired constraints. To satisfy $\mu_1 - \mu_2 = \theta_0$, for example, we maximize

$$\prod_{i=1}^{n_1} w_{1i} \prod_{i=1}^{n_2} w_{2i} \tag{26}$$

subject to constraints on weights (given here for two samples):

$$\begin{aligned} w_{1i} > 0, i = 1, \ldots, n_1 \qquad (27)\\ w_{2i} > 0, i = 1, \ldots, n_2 \\ \sum_{i=1}^{n_1} w_{1i} = 1 \\ \sum_{i=1}^{n_2} w_{2i} = 1 \end{aligned}$$

and the constraint specific to comparing two means:

$$\sum_{i=1}^{n_1} w_{1i} x_{1i} - \sum_{i=1}^{n_2} w_{2i} x_{2i} = \theta_0. \tag{28}$$

For other statistics we replace (28) with the more general

$$\theta(\hat{F}_{n,w}) = \theta_0, \tag{29}$$

where $\hat{F}_{n,w}$ is the weighted empirical distribution (with obvious generalization to multiple samples or strata).

The computational tools used for empirical likelihood[34] and bootstrap tilting[19,21] are useful in determining the weights.

The bootstrap sampling is from the weighted empirical distributions, i.e., the data are sampled with unequal probabilities.

Figure 15 shows this idea applied to the leukemia data. The top left shows Kaplan–Meier survival curves for the original data, and top right shows the bootstrap distribution for the log-hazard ratio, using 50 observations in each group. The bottom left shows weights chosen to maximize (26), subject to (28) and a log-hazard ratio equal to 0.5. In order to reduce the ratio from its original value of 0.904, the treatment group gets high weights early and low weights later (the weighted distribution has a higher probability of early events) while the control group gets the converse. Censored observations get roughly the average weight of the remaining noncensored observations in the same group. The middle left shows the resulting weighted survival estimates, and middle right the corresponding bootstrap distribution. In this case both bootstraps are nearly normal, and the standard errors are very similar −0.221 for the ordinary bootstrap and 0.212 for the weighted bootstrap, both with 50 observations per group.

HOW MANY BOOTSTRAP SAMPLES ARE NEEDED

We suggested in Section *Accuracy of Bootstrap Distributions* that 1000 bootstrap samples are enough for rough approximations, but that more are needed for greater accuracy. In this section we give details. The focus here is on Monte Carlo accuracy—how well the usual random-sampling implementation of the bootstrap approximates the theoretical bootstrap distribution.

A bootstrap distribution based on B random samples corresponds to drawing B observations with replacement from the theoretical bootstrap distribution. Quantities such as the mean, standard deviation, or quantiles of the bootstrap distribution converge to their theoretical counterparts at the rate $O(1/\sqrt{B})$, in probability.

Efron and Tibshirani[2] suggest that $B = 200$, or even as few as $B = 25$, suffices for estimating standard

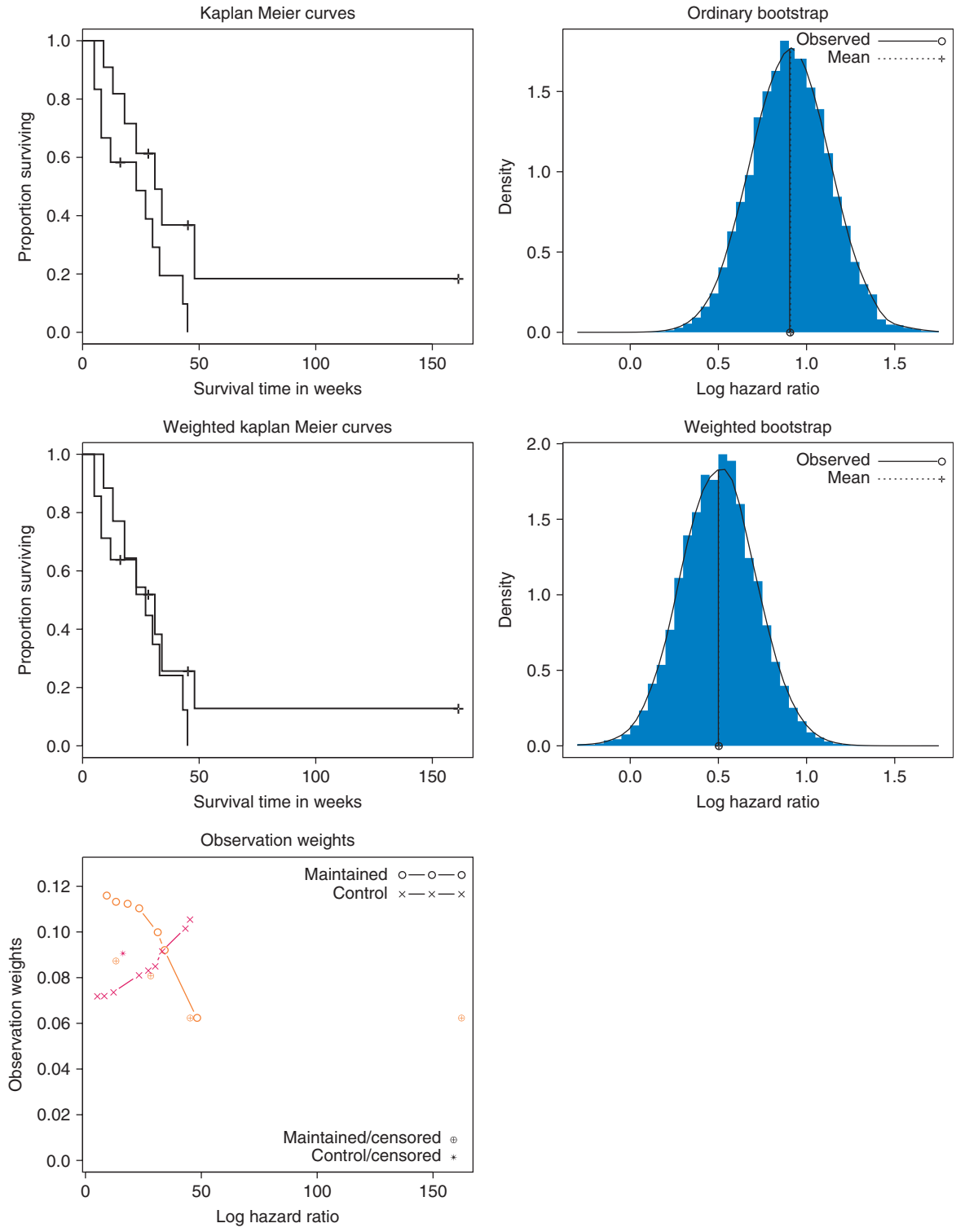


FIGURE 15 | Survival curves and bootstrap distribution for log-hazard ratio, original and perturbed (weighted) to a log-hazard ratio of 0.5.

errors, and that $B = 1000$ is enough for confidence intervals.

We argue that larger sizes are appropriate, on two grounds. First, those criteria were developed when computers were much slower; with faster computers it is much easier to take more samples.

Second, those criteria were developed using arguments that combine the random variation

due to the original sample with the random variation due to bootstrap sampling. For example, Efron and Tibshirani[2] indicate that $\mathrm{cv}(\hat{\mathrm{se}}_B) \doteq \left\{\mathrm{cv}(\hat{\mathrm{se}}_\infty)^2 + (E(\Delta)+2)/(4B)\right\}^{1/2}$, where cv is coefficient of variation, $\mathrm{cv}(Y) = \sigma_Y/E(Y)$, se_B and se_∞ are bootstrap standard errors using B or ∞ replications, respectively, and Δ relates to the kurtosis of the bootstrap distribution; it is zero for normal distributions. Even relatively small values of B make the ratio $\mathrm{cv}(\hat{\mathrm{se}}_B)/\mathrm{cv}(\hat{\mathrm{se}}_\infty)$ not much larger than 1.

We feel that the variation in bootstrap answers *conditional on the data* is more relevant. This is particularly true in clinical trial applications, where

- reproducibility is important—two people analyzing the same data should get (almost exactly) the same results, with random variation between their answers minimized, and
- the data may be very expensive—there is little point in wasting the value of expensive data by introducing extraneous variation using B too small. Given the choice between reducing variation in the ultimate results by gathering more data or by increasing B, it would be cheaper to increase B, at least until B is quite large.

Conditional on the data, $\mathrm{cv}(\hat{\mathrm{se}}_B) \doteq \sqrt{(\delta+2)/(4B)}$, where δ is the kurtosis of the theoretical bootstrap distribution (conditional on the data). When δ is zero (usually approximately true), this simplifies to $\mathrm{cv}(\hat{\mathrm{se}}_B) \doteq 1/\sqrt{2B}$.

To determine how large B should be, we consider the effect on confidence intervals. Consider a t interval of the form $\hat{\theta} \pm t_{\alpha/2}\mathrm{se}_B$. Suppose that such an interval using se_∞ would be approximately correct, with one-sided non-coverage $\alpha/2$. Then the actual non-coverage using se_B in place of se_∞ would be $F_{t,n-1}((\mathrm{se}_B/\mathrm{se}_\infty)F^{-1}_{t,n-1}(\alpha/2))$. For n large and $\alpha = 0.05$, to have the actually one-sided non-coverage fall within 10% of the desired value (between 0.0225 and 0.0275) requires that $\mathrm{se}_B/\mathrm{se}_\infty$ be between $\Phi^{-1}(0.025 \times 1.1)/\Phi^{-1}(0.025) = 0.979$ and $\Phi^{-1}(0.025 \times 0.9)/\Phi^{-1}(0.025) = 1.023$. To have 95% confidence of no more than 10% error requires that $1.96/\sqrt{2B} \le 0.022$, or $B \ge 0.5(1.96/0.022)^2 = 3970$, or about 4000 bootstrap samples.

To satisfy the more stringent criterion of 95% confidence that the non-coverage error is less than 1% of 0.025 would require approximately 400,000 bootstrap samples. With modern computers this is not unreasonable, unless the statistic is particularly slow to compute.

Consider also bootstrap confidence intervals based on quantiles. The simple bootstrap percentile confidence interval is the range from the $\alpha/2$ to $1-\alpha/2$ quantiles of the bootstrap distribution. Let $G^{-1}_\infty(c)$ be the c quantile of the theoretical bootstrap distribution, and the number of bootstrap statistics falling below this quantile is approximately binomial with parameters B and c (the proportion parameter may differ slightly due to the discreteness of the bootstrap distribution). For finite B, the one-sided error has standard error approximately $\sqrt{c(1-c)/B}$. For $c = 0.025$, to reduce 1.96 standard errors to $c/10$ requires $B \ge (10/0.025)^2 1.96^2 0.025 \times 0.975 = 14980$, about 15,000 bootstrap samples.

The more stringent criterion of a 1% error would require approximately 1.5 million bootstrap samples.

The bootstrap BCa confidence interval has greater Monte Carlo error, because it requires estimating a bias parameter using the proportion of bootstrap samples falling below the original $\hat{\theta}$ (and the variance of a binomial proportion $\sqrt{p(1-p)/B}$ is greatest for $p \doteq 0.5$). It requires B about twice as large as the bootstrap percentile interval for equivalent Monte Carlo accuracy—30,000 bootstrap samples to satisfy the 10% criterion.

On the other hand, the bootstrap tilting interval requires about 17 times fewer bootstrap samples for the same Monte Carlo accuracy as the simple percentile interval, so that about 1000 bootstrap samples would suffice to satisfy the 10% criterion.

In summary, to have 95% probability that the actual one-sided non-coverage for a 95% bootstrap interval falls within 10% of the desired value, between 0.0225 and 0.0275, conditional on the data, requires about 1000 samples for a bootstrap tilting interval, 4000 for a t interval using a bootstrap standard error, 15,000 for a bootstrap percentile interval, and 30,000 for a bootstrap BCa interval.

Figure 16 shows the Monte Carlo variability of a number of bootstrap confidence interval procedures, for various combinations of sample size, statistic, and underlying data; these are representative of a larger collection of examples in Ref 22. The panels show the variability due to Monte Carlo sampling with a finite bootstrap sample size B, conditional on the data.

Figure 16 is based on 2000 randomly generated datasets for each sample size, distribution, and statistic. For each dataset, and for each value of B, two sets of bootstrap samples are created and intervals calculated using all methods. For each method, a sample variance is calculated using the usual unbiased sample variance (based on two observations). The estimate of Monte Carlo variability is then the average across the 2000 datasets of these unbiased sample variances. The result is the ''within-group'' component of variance (due to Monte Carlo variability) and

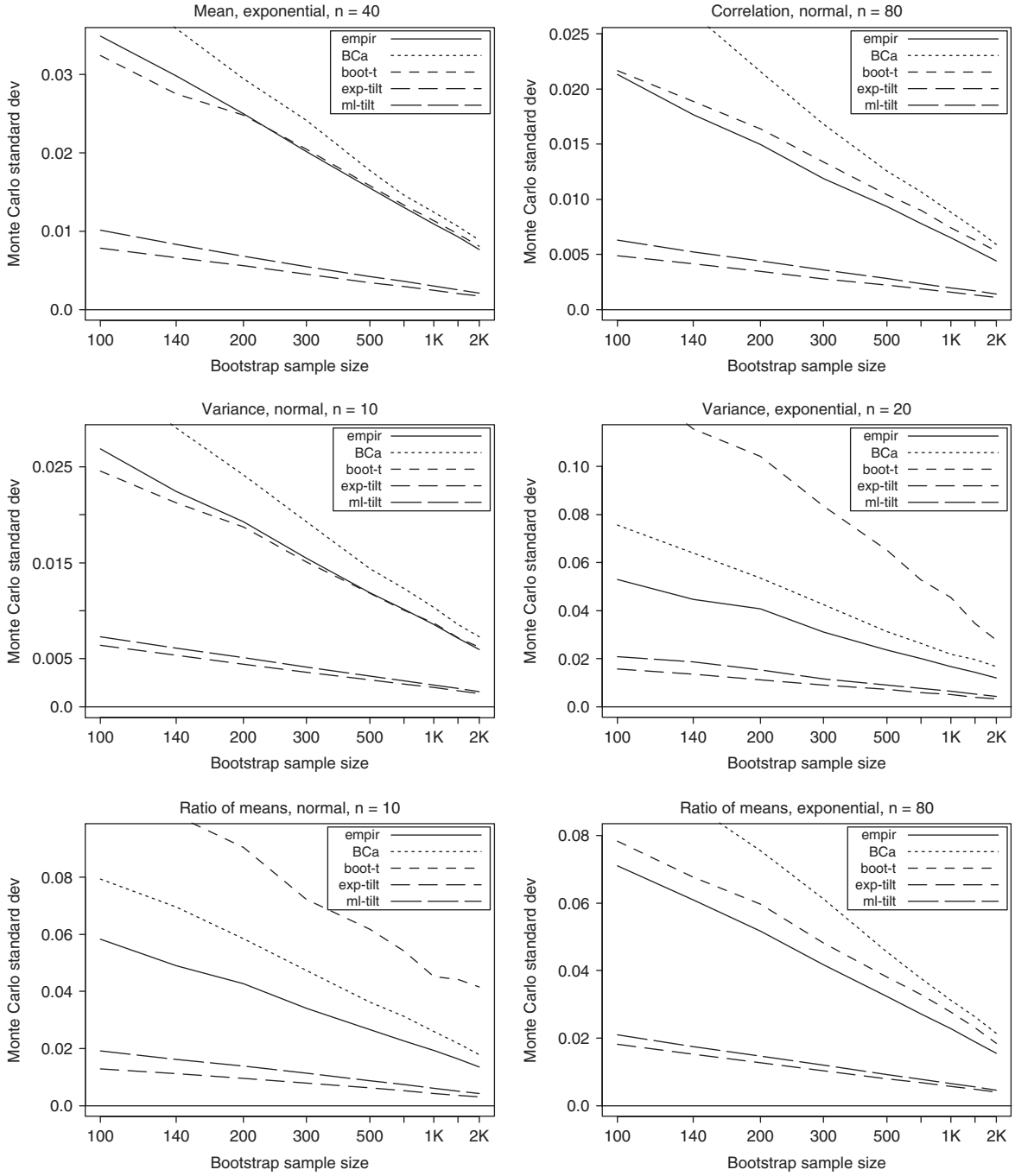


FIGURE 16 | Monte Carlo variability for confidence intervals.

excludes the "between-group" component (due to differences between datasets).

Assessing Monte Carlo Variation

To assess Monte Carlo variation in practice, there are two options. The first is to use asymptotic formulae. For example, the bootstrap estimate of bias (1) depends on the sample mean of the bootstrap statistics; the usual formula for standard error of a sample mean is $\text{se}_B/\sqrt{B}$, where se_B is the sample standard deviation of the bootstrap statistics. The standard error of a bootstrap proportion $\hat{p}$ is $\sqrt{\hat{p}(1-\hat{p})/B}$. The standard error of a bootstrap standard error se_B is $\text{se}_B\sqrt{(\delta+2)/(4B)}$.

The other alternative is to resample the bootstrap values. Given B *i.i.d.* observations $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*$ from

the theoretical bootstrap distribution, and a summary statistic Q (e.g., standard error, bias estimate, or endpoint of a confidence interval), we may draw B_2 bootstrap samples of size B from the B observations, and calculate the summary statistics $Q_1^*, Q_2^*, \ldots, Q_{B_2}^*$. The sample standard deviation of the Q^*s is the Monte Carlo standard error.

Variance Reduction

There are a number of techniques that can be used to reduce the Monte Carlo variation.

The *balanced bootstrap*,[35] in which each of the n observations is included exactly B times in the B bootstrap samples, is useful for bootstrap bias estimates but of little value otherwise.

Antithetic variates[36] is moderately helpful for bias estimation but of little value otherwise.

Importance sampling[37,38] is particularly useful for estimating tail quantiles, as for bootstrap percentile and BCa intervals. For nonlinear statistics one should use a defensive mixture Distribution.[39,40]

Control variates[36,39,41,42] are moderately to extremely useful for bias and standard error estimation and can be combined with importance sampling.[43] They are most effective in large samples for statistics that are approximately linear.

Concomitants[42,44] are moderately to extremely useful for quantiles and can be combined with importance sampling.[45] They are most effective in large samples for statistics that are approximately linear; linear approximations tailored to a tail of interest can dramatically improve the accuracy.[46]

Quasi-random sampling[47] can be very useful for small n and large B; the convergence rate is $O(\log(B)^n B^{-1})$ compared to $O(B^{-1/2})$ for Monte Carlo methods.

Analytical approximations for bootstrap distributions are available in some situations, including analytical approximations for bootstrap tilting and BCa intervals,[20,24] and saddlepoint approximations.[48–52]

ADDITIONAL TOPICS

Some topics that are beyond the scope of this article[a] include bootstrapping dependent data (time series, mixed effects models), cross-validation and bootstrap-validation (bootstrapping prediction errors, and classification errors), Bayesian bootstrap, bootstrap likelihoods. Refs 2 and 5 are good starting points for these topics, with the exception of mixed effects models. Ref 2 is an introduction to the bootstrap written for upper-level undergraduate or beginning graduate students. Ref 5 is the best general-purpose reference for the bootstrap for statistical practitioners. Ref 10 looks at asymptotic properties of various bootstrap methods. The author's website `http://home.comcast.net/~timhesterberg/bootstrap` has resources for teaching statistics using the bootstrap, and some technical reports, particularly on computational aspects of bootstrapping.

NOTE

[a]This article is a minor revision of Ref 53.

REFERENCES

1. Efron B. Bootstrap methods: another look at the jackknife (with discussion). *Ann Stat* 1979, 7:1–26.
2. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman and Hall; 1993.
3. Breiman L. Random forests. *Mach Learn* 2001, 45: 5–32.
4. Efron B. *The Jackknife, the Bootstrap and Other Resampling Plans. National Science Foundation–Conference Board of the Mathematical Sciences Monograph 38*. Philadelphia: Society for Industrial and Applied Mathematics; 1982.
5. Davison A, Hinkley D. *Bootstrap Methods and Their Applications*. Cambridge University Press; 1997.
6. Wu CFJ. Jackknife, bootstrap, and other resampling methods in regression analysis (with discussion). *Ann Stat* 1986, 14:1261–1350.
7. Chambers J, Hastie T. *Statistical Models in S*. Pacific Grove, CA: Wadsworth; 1992.
8. Chambers JM, Cleveland WS, Kleiner B, Tukey PA. *Graphical Methods for Data Analysis*. Wadsworth; 1983.
9. Ruckstuhl A, Stahel W, Maechler M, Hesterberg T. Sunflower. Statlib. Available at: http://lib.stat.cmu.edu/S/sunflower. (Accessed 1995).
10. Hall P. *The Bootstrap and Edgeworth Expansion*. New York: Springer; 1992.
11. Shao J, Tu D. *The Jackknife and Bootstrap*. New York: Springer-Verlag; 1995.
12. Silverman B, Young G. The bootstrap: to smooth or not to smooth. *Biometrika* 1987, 74:469–479.
13. Hall P, DiCiccio T, Romano J. On smoothing and the bootstrap. *Ann Stat* 1989, 17:692–704.

14. Efron B. Better bootstrap confidence intervals (with discussion). *J Am Stat Assoc* 1987, 82:171–200.
15. Hesterberg TC. Unbiasing the bootstrap-bootknife sampling vs. smoothing. *Proceedings of the Section on Statistics & the Environment*. American Statistical Association; 2004, 2924–2930.
16. DiCiccio TJ, Romano JP. A review of bootstrap confidence intervals (with discussion). *J R Stat Soc B* 1988, 50:338–354.
17. Hall P. Theoretical comparison of bootstrap confidence intervals (with discussion). *Ann Stat* 1988, 16:927–985.
18. DiCiccio T, Efron B. Bootstrap confidence intervals (with discussion). *Stat Sci* 1996, 11:189–228.
19. Efron B. Nonparametric standard errors and confidence intervals. *Can J Stat* 1981, 9:139–172.
20. DiCiccio TJ, Romano JP. Nonparametric confidence limits by resampling methods and least favorable families. *Int Stat Rev* 1990, 58:59–76.
21. Hesterberg TC. Bootstrap tilting confidence intervals and hypothesis tests. In: Berk K, Pourahmadi M, eds. *Computer Science and Statistics: Proceedings of the 31st Symposium on the Interface*, vol 31. Fairfax Station, VA: Interface Foundation of North America; 1999, 389–393.
22. Hesterberg TC. Bootstrap tilting confidence intervals. Technical Report 84, Research Department, MathSoft, Inc.; 1999.
23. Hyndman RJ, Fan Y. Sample quantiles in statistical packages. *Am Stat* 1996, 50:361–364.
24. DiCiccio T, Efron B. More accurate confidence intervals in exponential families. *Biometrika* 1992, 79:231–245.
25. DiCiccio TJ, Martin MA, Young GA. Analytic approximations to bootstrap distribution functions using saddlepoint methods. Technical Report 356, Department of Statistics, Stanford University; 1990.
26. Efron B. Censored data and the bootstrap. *J Am Stat Assoc* 1981, 76:312–319.
27. Hinkley DV. Bootstrap significance tests. *Bull Int Stat Inst* 1989, 53:65–74.
28. Owen A. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 1988, 75:237–249.
29. Embury SH, Elias L, Heller PH, Hood CE, Greenberg PL, Schrier SL. Remission maintenance therapy in acute myelogenous leukemia. *West J Med* 1977, 126: 267–272.
30. Insightful. *S-PLUS® 8 Guide to Statistics*. 1700 Westlake Ave N., Suite 500, Seattle; 2007.
31. Hesterberg TC. Bootstrap tilting diagnostics. *Proceedings of the Statistical Computing Section*; 2001.
32. Hesterberg TC. Resampling for planning clinical trials-using S+Resample. Statistical Methods in Biopharmacy, Paris. Available at: http://home.comcast.net/~timhesterberg/articles/Paris05-ResampleClinical.pdf. (Accessed 2011).
33. Hall P, Presnell B. Intentionally biased bootstrap methods. *J R Stat Soc B* 1999, 61:143–158.
34. Owen A. *Empirical Likelihood*. Chapman & Hall/CRC Press; 2001.
35. Gleason JR. Algorithms for balanced bootstrap simulations. *Am St* 1988, 42:263–266.
36. Therneau TM. Variance reduction techniques for the bootstrap. Technical Report No. 200, PhD thesis, Department of Statistics, Stanford University; 1983.
37. Johns MV. Importance sampling for bootstrap confidence intervals. *J Am Stat Assoc* 1988, 83:701–714.
38. Davison AC. Discussion of paper by D. V. Hinkley. *J R Stat Soc B* 1986, 50:356–357.
39. Hesterberg TC. Advances in importance sampling. PhD thesis, Statistics Department, Stanford University; 1988.
40. Hesterberg TC. Weighted average importance sampling and defensive mixture distributions. *Technometrics* 1995, 37:185–194.
41. Davison AC, Hinkley DV, Schechtman E. Efficient bootstrap simulation. *Biometrika* 1986, 73:555–566.
42. Efron B. More efficient bootstrap computations. *J Am Stat Assoc* 1990, 85:79–89.
43. Hesterberg TC. Control variates and importance sampling for efficient bootstrap simulations. *Stat Comput* 1996, 6:147–157.
44. Do KA, Hall P. Distribution estimation using concomitants of order statistics, with application to Monte Carlo simulations for the bootstrap. *J R Stat Soc B* 1992, 54:595–607.
45. Hesterberg TC. Fast bootstrapping by combining importance sampling and concomitants. *Computing Science and Statistics*, 1997, 29:72–78.
46. Hesterberg TC. Tail-specific linear approximations for efficient bootstrap simulations. *J Comput Graph Stat* 1995, 4:113–133.
47. Do KA, Hall P. Quasi-random sampling for the bootstrap. *Stat Comput* 1991, 1:13–22.
48. Tingley M, Field C. Small-sample confidence intervals. *J Am Stat Assoc* 1990, 85:427–434.
49. Daniels HE, Young GA. Saddlepoint approximation for the studentized mean, with an application to the bootstrap. *Biometrika* 1991, 78:169–179.
50. Wang S. General saddlepoint approximations in the bootstrap. *Stat Prob Lett* 1992, 13:61–66.
51. DiCiccio TJ, Martin MA, Young GA. Analytical approximations to bootstrap distributions functions using saddlepoint methods. *Stat Sin* 1994, 4:281.
52. Canty AJ, Davison AC. Implementation of saddlepoint approximations to bootstrap distributions. In: Billard L, Fisher NI, eds. *Computing Science and Statistics; Proceedings of the 28th Symposium on the Interface*, vol 28. Fairfax Station, VA: Interface Foundation of North America; 1997, 248–253.
53. Hesterberg TC. Bootstrap. In: D'Agostino R, Sullivan L, Massaro J, eds. *Wiley Encyclopedia of Clinical Trials*. John Wiley & Sons; 2007.

FURTHER READING

Chernick MR. *Bootstrap Methods: A Practitioner's Guide*. New York: John Wiley & Sons; 1999. (An extensive bibliography, with roughly 1700 references related to the bootstrap.)
Hesterberg T, Monaghan S, Moore DS, Clipson A, Epstein R. *Bootstrap Methods and Permutation Tests*. W. H. Freeman. Chapter for The Practice of Business Statistics by Moore, McCabe, Duckworth, and Sclove; 2003. Available at: http://bcs.whfreeman.com/pbs/cat_160/PBS18.pdf. (An introduction to the bootstrap written for introductory statistics students.) (Accessed 2011).

