

Where a confidence interval comes from

The bootstrap, the sampling distribution and why the simplest bootstrap interval is too narrow in small samples

Selçuk Korkmaz

8 October 2026

IN SHORT

A confidence interval is read off a sampling distribution: the distribution an estimate would have if the study were repeated with new participants. The t interval works with that distribution through a model and a formula. The bootstrap estimates it from the data: it treats the sample as a stand-in for the population, draws new samples from it with replacement, and recomputes the estimate each time. For the 20-person pilot trial of the previous note, the ideal bootstrap distribution, computed exactly, gives a standard error of 3.16 mmHg and a 95% percentile interval of 1.3 to 13.7 mmHg. The t interval is wider, 0.5 to 14.5 mmHg, and in simulated trials with 10 participants per arm the percentile interval covered the true difference only 92% of the time.

Three interactive figures build the bootstrap distribution by resampling, set it against the t distribution and two better bootstrap intervals (BCa and the bootstrap- t), and count how often each interval covers the truth in simulated trials. Tables cover the intervals for the pilot data, their coverage, what to resample in each design and how many resamples are enough, and a checklist covers what to report.

Keywords: bootstrap · confidence intervals · standard error · sampling distribution · resampling · small samples

1 A confidence interval needs a sampling distribution

The previous note showed where a p -value comes from: a null distribution, built there by shuffling the group labels. A confidence interval has a source too. In an earlier note I described the standard error as a standard deviation: the standard deviation of an estimate across hypothetical repeats of the study, each with new participants. Those repeats form the *sampling distribution* of the estimate, and a 95% confidence interval is built from it, by a procedure that would capture the true value in 95% of the repeats if the model's assumptions hold.

The pilot trial of the previous note will serve again. Twenty adults with raised blood pressure were randomised, 10 to a lifestyle programme and 10 to usual care, and the outcome is the fall in systolic blood pressure over 12 weeks (Table 1). The programme arm fell by 9.4 mmHg on average and the usual-care arm by 1.9 mmHg, a difference of 7.5 mmHg.

Table 1. The pilot data: fall in systolic blood pressure over 12 weeks, in mmHg (negative values are rises). The data are made up and are the same as in the previous note.			
ARM	FALL IN EACH PARTICIPANT	MEAN	SD
Programme (<i>n</i> = 10)	21, 17, 15, 14, 13, 7, 6, 2, 2, -3	9.4	7.7
Usual care (<i>n</i> = 10)	11, 10, 9, 5, 2, 1, 0, -1, -9, -9	1.9	7.1

The usual 95% confidence interval for the difference, from Student's *t*, runs from 0.5 to 14.5 mmHg:

$$7.5 \pm 2.10 \times 3.33 = 0.5 \text{ to } 14.5 \text{ mmHg.}$$

Behind the formula is a model: the outcomes in both arms are normally distributed with the same standard deviation (SD). Under that model the difference in means has a normal sampling distribution, whose SD the standard error, 3.33 mmHg, estimates from the data. Because the standard error is itself estimated, the estimation error divided by the standard error follows Student's *t* distribution with 18 degrees of freedom, and that is where the multiplier 2.10 comes from. Equivalently, the interval's limits are the 2.5th and 97.5th percentiles of a *t* distribution centred at the estimate and stretched by the standard error. Every step is a mathematical consequence of the model.

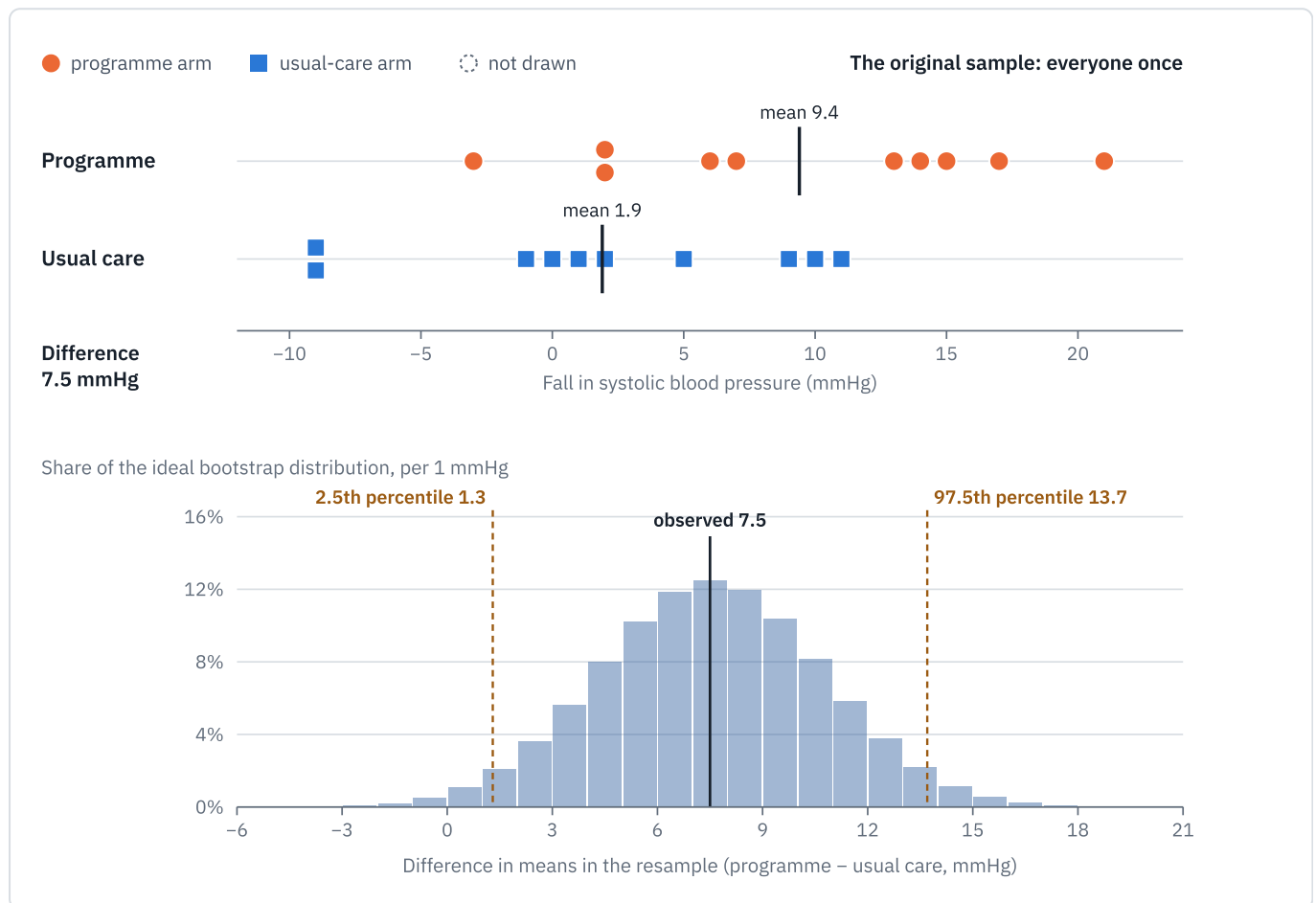
Repeating the trial to see the real sampling distribution is the one thing we cannot do. The bootstrap, which Efron introduced in 1979^[1], does the next best thing.

2 Resampling the participants

The idea is the plug-in principle^[2]. The sampling distribution describes what happens when samples are drawn from the population. We do not have the population, but we have a sample from it, and the sample is our best picture of it. So we draw samples from the sample. Drawing 10 programme participants at random *with replacement*, so that some appear twice or more and others not at all, imitates recruiting 10 new participants from a population that looks like the 10 we have. Do the same for the usual-care arm, compute the difference in means, and repeat many times. The spread of the differences estimates the spread of the sampling distribution.

Bootstrap for a difference between two groups.

1. Draw 10 participants with replacement from the programme arm, and separately 10 from the usual-care arm, so that each arm keeps its size ^[2].
2. Compute the statistic for this resample, T^* (here the difference in means).
3. Repeat B times.
4. The SD of the B values of T^* is the *bootstrap standard error*. Their 2.5th and 97.5th percentiles are the *95% percentile interval*.



The ideal bootstrap distribution, which ever more resamples approach, computed exactly: its standard deviation, the bootstrap standard error, is 3.16 mmHg, and its middle 95% runs from 1.3 to 13.7 mmHg (the percentile interval). The t interval is 0.5 to 14.5 mmHg. The top panel shows the original sample.

Figure 1. Building the bootstrap distribution by resampling. Top: the 20 participants of Table 1, placed by their fall in systolic blood pressure. In a resample, a participant drawn more than once is marked with the number of draws ($\times 2$, $\times 3$, ...), a participant not drawn is shown as a dashed outline, and the vertical marks are the two arm means of the resample. Bottom: the difference in means (programme minus usual care) for every resample so far, as a share of them, in 1 mmHg bins; the dashed lines are the 2.5th and 97.5th percentiles, the ends of the 95% percentile interval. On the interactive page the buttons draw one, 100 or 1,000 resamples, or show the ideal bootstrap distribution, which ever more resamples approach and which can be computed exactly here; the static copy shows the ideal distribution, with the original sample at the top. Its standard deviation is 3.16 mmHg and its middle 95% runs from 1.3 to 13.7 mmHg.

Figure 1 lets you do it. Each resample is one imitation of the trial, run on a population made of copies of the participants we have. A resample never contains a value that is not in the sample, so the largest difference the bootstrap can produce is 30 mmHg (all 10 programme draws the participant who fell by 21, all 10 usual-care draws a participant who rose by 9), and in the bootstrap's world nobody responds more strongly than the people already observed.

Random resamples estimate the *ideal bootstrap distribution*, the one that infinitely many resamples would give. For a difference in means it can be computed exactly. Each arm can be resampled in 10^{10} equally likely ordered ways, and the distribution of the sum of 10 draws follows by convolving the distribution of one draw with itself 10 times. The ideal distribution has an SD, the bootstrap standard error, of 3.16 mmHg, and its middle 95% runs from 1.3 to 13.7 mmHg.

Two features stand out. First, the bootstrap distribution is centred at the observed difference of 7.5 mmHg. It is not centred at zero, as the null distribution of the previous note was, and not at the true difference, which nobody knows. It describes how far an estimate tends to stray, not where the truth is: its job here is to estimate uncertainty, not to change the estimate ^[2]. The two resampling methods answer different questions: shuffling draws without replacement, mixes the arms and builds the distribution the statistic would have if the programme did nothing; the bootstrap draws with replacement within each arm and estimates the distribution the estimate actually has. Second, the bootstrap answer is narrower than the t-test's: a standard error of 3.16 mmHg against 3.33, and an interval of 1.3 to 13.7 against 0.5 to 14.5.

3 Why the simplest interval is too narrow

Two well-understood causes account for both gaps ^[2].

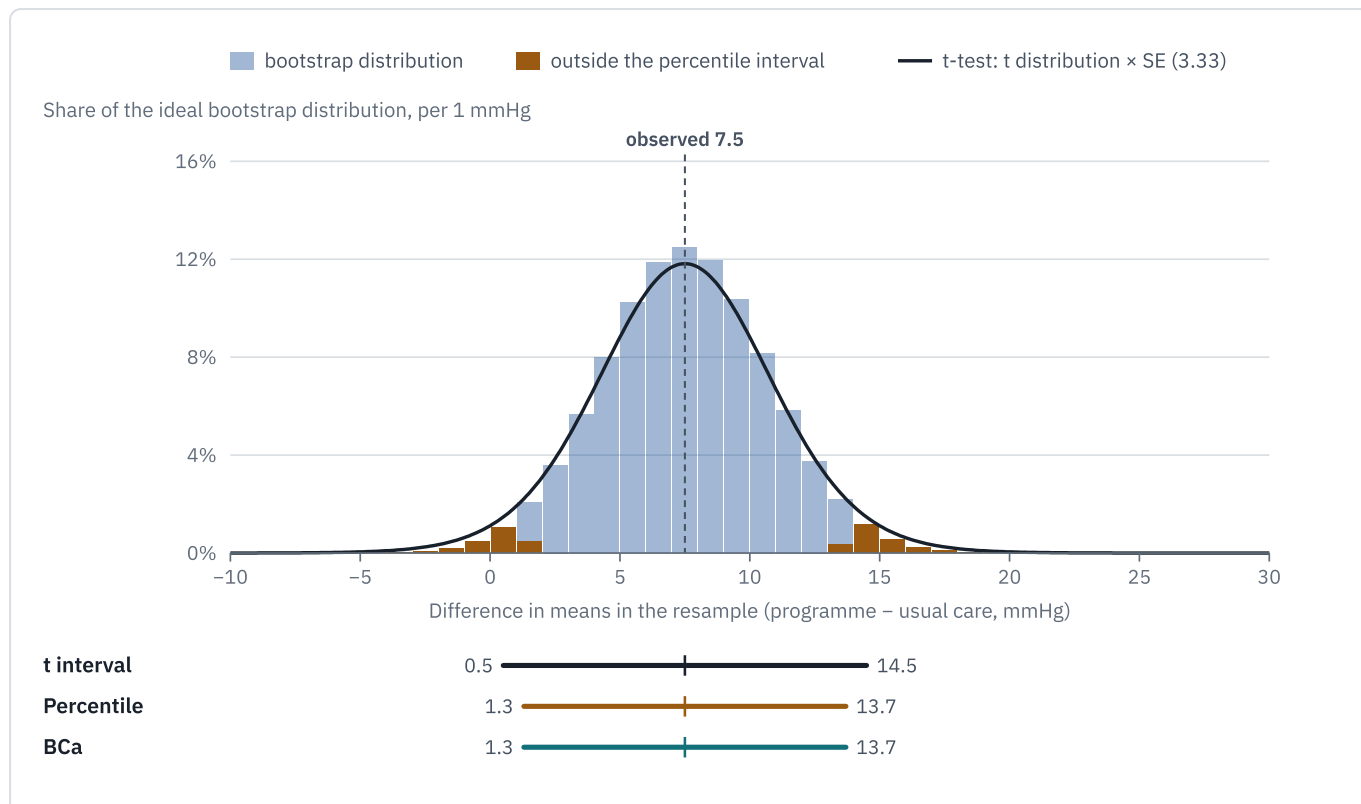
- **The bootstrap's SD divides by n .** Treated as a population, a sample has variance $\Sigma(x - \bar{x})^2/n$, not the usual s^2 , which divides by $n - 1$. So the bootstrap standard error of a mean is smaller than the formula's by the factor $\sqrt{(n - 1)/n}$. When two groups are compared, the factor depends on the size of each group, not on the total ^[2]. With 10 per arm it is $\sqrt{9/10} = 0.949$, and $3.33 \times 0.949 = 3.16$.
- **The percentile interval makes no allowance for the uncertainty in the SD.** The bootstrap distribution of the difference is close to normal, so its 2.5th and 97.5th percentiles lie about 1.96 bootstrap standard errors either side of the estimate: $7.5 \pm 1.96 \times 3.16 = 1.3$ to 13.7 mmHg. The t interval multiplies by 2.10 instead of 1.96, because its standard error is itself an estimate from only 20 values, and the t distribution allows for that. For symmetric data, the percentile interval behaves like an interval that uses z where t belongs and divides by n instead of $n - 1$ ^[2].

Together, $0.949 \times 1.96 / 2.10 = 0.885$: the percentile interval is about 11% narrower than the t interval, 12.4 mmHg wide against 14.0. Both factors fade as the arms grow. For the same calculation with 5, 10, 20 and 40 per arm, the ratio is 0.76, 0.89, 0.94 and 0.97. So, for a mean or a difference in means, the instinct to reach for the bootstrap because a sample is small has it backwards: the simplest bootstrap interval is at its worst in small samples ^[2].

4 Better bootstrap intervals

The percentile interval is the simplest of several ways to turn a bootstrap distribution into an interval ^[3]. Two others are widely available, for example in R's boot package, which implements the methods of Davison and Hinkley ^[4]:

- **BCa** (bias-corrected and accelerated), introduced by Efron in 1987 ^[5]. It reads the interval off the bootstrap distribution, but at shifted percentiles. A bias correction comes from the share of the bootstrap distribution below the estimate, and an acceleration constant adjusts for skewness ^[2,5]. It has no adjustment for the narrowness described in Section 3 ^[2]. For the pilot data the shifts are tiny, and BCa gives 1.3 to 13.7 mmHg again.
- **Bootstrap-t** (the studentised bootstrap). Instead of the difference itself, it bootstraps a t statistic: each resample's difference minus the observed one, divided by that resample's own standard error. The 2.5th and 97.5th percentiles of these values take the place of ± 2.10 from the t table ^[2]. Because every resample estimates its own standard error, the method sees how uncertain the standard error is, as the t interval does. For the pilot data it gives 0.5 to 14.6 mmHg, close to the t interval.



Data of Table 1: difference in means 7.5 mmHg. Bootstrap standard error 3.16 mmHg, against 3.33 from the t -test formula. 95% intervals: ■ percentile 1.3 to 13.7, ■ BCa 1.3 to 13.7, ■ t 0.5 to 14.5 mmHg.

Figure 2. The bootstrap distribution, the t distribution and three intervals. Bars: the ideal bootstrap distribution of the statistic, computed exactly, as a share in 1 mmHg bins (for medians, one bar for each possible value, 0.5 mmHg apart); the parts outside the 95% percentile interval are shaded. Curve (difference in means only): the distribution the t interval reads its limits from, a t distribution with 18 degrees of freedom centred at the estimate and stretched by the t -test's standard error (3.33 mmHg for the data of Table 1; 4.80 mmHg with the 45 mmHg value). Strictly, the t distribution describes the estimation error divided by its estimated standard error; under the normal model the difference in means itself has a normal sampling distribution. Below the axis: the 95% t , percentile and BCa intervals (for medians, the percentile interval only). On the interactive page the buttons switch to a data set in which the largest fall in the programme arm is 45 mmHg instead of 21, and to the difference in medians, for which there is no t interval; the static copy shows the data of Table 1 and the difference in means. With the 45 mmHg value, the bootstrap standard error is 4.55 mmHg and the intervals are 1.8 to 19.5 (percentile), 2.8 to 21.5 (BCa) and -0.2 to 20.0 mmHg (t); for the difference in medians the percentile interval is -2.5 to 15.5 mmHg in both data sets.

Figure 2 sets the ideal bootstrap distribution against the curve the t interval reads its limits from. For the pilot data they have the same centre and a similar shape, and the bootstrap distribution is the narrower of the two, for the reasons of Section 3. Table 2 collects the intervals, including the one from inverting the permutation test in the previous note.

Table 2. Six 95% confidence intervals for the difference in mean fall in systolic blood pressure (programme minus usual care), in mmHg. The percentile and BCa intervals come from the ideal bootstrap distribution, computed exactly; the bootstrap- t interval from 10 million random resamples. The inverted permutation interval assumes a constant shift δ (the programme lowers everyone's blood pressure by the same amount) and collects the values of δ for which the exact permutation test of the raw difference in means, run after subtracting δ from each programme value, gives a two-sided p above 0.05 (previous note, Section 7).

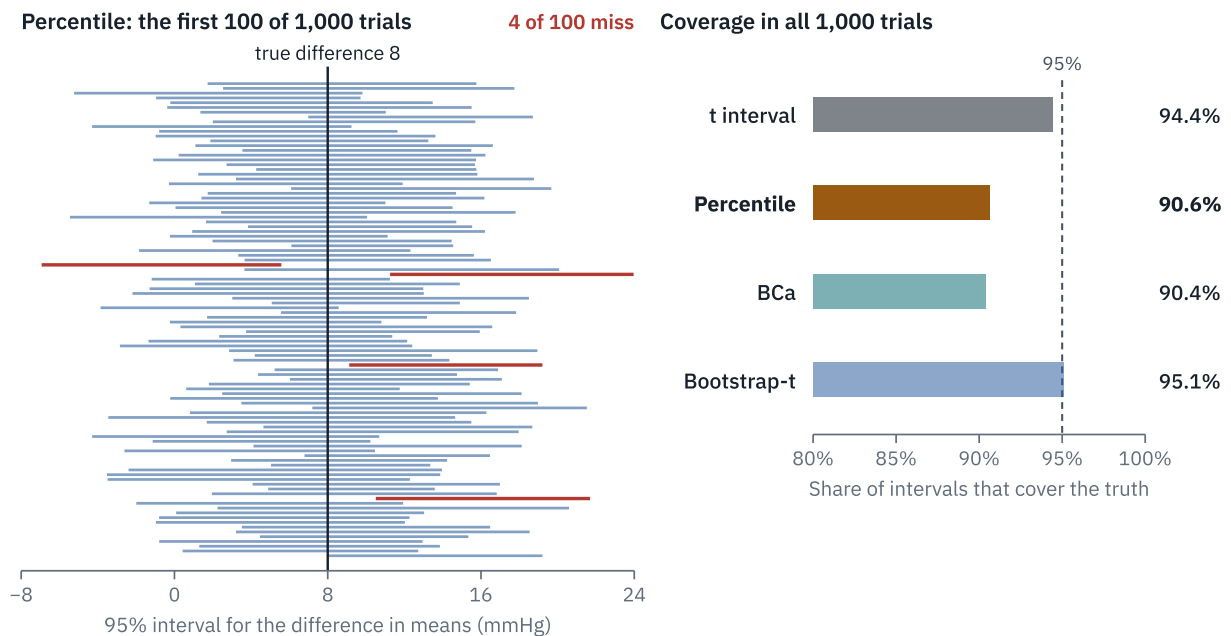
METHOD	DATA OF TABLE 1 (DIFFERENCE 7.5)	LARGEST FALL 45 MMHG (DIFFERENCE 9.9)
Student's t	0.5 to 14.5	-0.2 to 20.0
Welch's t	0.5 to 14.5	-0.4 to 20.2
Inverted permutation test (previous note)	0.5 to 14.5	0.6 to 20.0
Bootstrap percentile	1.3 to 13.7	1.8 to 19.5
Bootstrap BCa	1.3 to 13.7	2.8 to 21.5
Bootstrap- t	0.5 to 14.6	1.7 to 23.7

Now switch Figure 2 to the second data set, in which the largest fall in the programme arm is 45 mmHg instead of 21. The bootstrap distribution becomes skewed to the right. In 35% of resamples the 45 is not drawn at all, in 39% it is drawn once, and in 26% twice or more, and each copy adds about 3.7 mmHg to the difference. The percentile and BCa intervals follow the skew and move to the right, and the bootstrap- t stretches its upper limit to 23.7 mmHg. Across the six methods the upper limit now ranges from 19.5 to 23.7 mmHg. Each method makes its own guess about how far the upper tail of the population reaches, and ten participants, one of whom carries much of the information about that tail, cannot settle the question. As in the previous note, the honest reading is that the difference in means rests heavily on one person.

Switch the statistic to the difference in medians, for which there is no simple standard-error formula: the situation the bootstrap is often recommended for. Its bootstrap distribution is lumpy. The median of 10 resampled values is the average of the 5th and 6th smallest, so for the data of Table 1 the difference in medians can take only 85 values, and the most likely one, 5.0 mmHg, carries 4.5% of the distribution. The percentile interval runs from -2.5 to 15.5 mmHg in both data sets, because the median ignores how far the largest value lies from the rest. The bootstrap tends not to work well for medians and other quantiles in small samples, which depend heavily on a few observations near the middle, though it does better in large samples ^[2].

5 How often do the intervals cover the truth?

A 95% interval method should capture the true value in 95% of repeated studies, and that can be checked by simulation: generate many trials with a known true difference, compute each interval, and count how often it contains the truth. Figure 3 does this for 1,000 trials at a time, with a true difference of 8 mmHg and an SD of 8 mmHg in each arm, and Table 3 shows the long run from 10,000 trials per row.



Normal outcomes, 10 per arm, 1,000 simulated trials: the 95% interval covered the true difference of 8 mmHg in 94.4% (t), 90.6% (Percentile), 90.4% (BCa), 95.1% (bootstrap- t). Average widths: 14.9, 13.2, 13.3, 15.4 mmHg in the same order.

Figure 3. How often do the intervals cover the true difference? Each of 1,000 simulated trials draws participants for both arms from the same distribution with an SD of 8 mmHg, shifted by a true difference of 8 mmHg, and computes four 95% intervals for the difference in means: the t interval and, from the same 999 resamples, the percentile, BCa and bootstrap- t intervals. Left: the first 100 intervals of the chosen method, with those that miss the true difference in red. Right: the share of all 1,000 intervals of each method that cover it; the dashed line is the nominal 95%. On the interactive page the buttons switch between normal and right-skewed (log-normal) outcomes, between 5, 10, 20 and 40 participants per arm, and between the methods shown on the left; the static copy shows normal outcomes, 10 per arm and the percentile interval.

Table 3. Coverage of four 95% intervals for a difference in means between two arms of equal size: the share of 10,000 simulated trials per row in which the interval contained the true difference (simulation standard error about 0.2 percentage points near 95% and 0.4 near 85%). Each trial draws both arms from the same distribution with an SD of 8 mmHg: normal, or skewed, $8 \times (e^Z - e^{0.5}) / \sqrt{(e - 1)e}$ with Z standard normal (a log-normal distribution with $\sigma = 1$ on the log scale, skewness about 6). All four methods shift with the data, so the true difference was set to zero without loss. The t interval is Student's; the percentile, BCa (bias correction from the share of resamples below the estimate, acceleration from the empirical influence values) and bootstrap- t intervals use the same 1,999 resamples within each arm, with percentiles by linear interpolation; a resample with a standard error of zero (every draw in each arm the same participant, which is possible but very rare) would count as an infinitely large t value in the direction of its difference. Python 3.9 with NumPy 2.0.2 and SciPy 1.13.1; four runs of 2,500 trials, seeds 11, 22, 33 and 44.

OUTCOMES	PER ARM	T INTERVAL	PERCENTILE	BCA	BOOTSTRAP- T
Normal	5	95%	88%	88%	96%
Normal	10	95%	92%	92%	95%
Normal	20	95%	94%	94%	95%
Normal	40	95%	94%	94%	95%
Skewed	5	97%	87%	83%	97%
Skewed	10	96%	90%	85%	92%
Skewed	20	96%	92%	87%	91%
Skewed	40	96%	93%	89%	91%

- **Normal outcomes.** The t interval covers 95% of the time at every size, as theory says it should. The percentile interval covers 88% of the time with 5 per arm, 92% with 10 and 94% with 20 or 40, as Section 3 predicts, and BCa does no better in these scenarios. The bootstrap- t covers about 95% throughout, but with 5 per arm its intervals are on average wider than the t interval's (25.8 against 22.5 mmHg). Hesterberg's simulations for a single mean tell the same story: the percentile interval does poorly in small normal samples and the bootstrap- t does very well ^[2].
- **Skewed outcomes.** The t interval is, if anything, conservative, at 96% to 97%. The percentile interval covers 87% to 93%, BCa only 83% to 89%, and the bootstrap- t 97% with 5 per arm, where its intervals are on average almost twice as wide as the t interval's (34.2 against 18.3 mmHg), then 91% to 92%.

One feature of this setting helps to read the skewed rows. When both arms have the same shape and the same size, the sampling distribution of the difference in means is exactly symmetric, however skewed each arm is: the errors in the two arm means have the same distribution, and the difference of two such errors is as likely to be positive as negative. So the true bias correction and acceleration that BCa estimates are both zero here, and any asymmetry that BCa or the bootstrap- t builds into an interval comes from the sample alone. With strongly skewed outcomes and 10 or 20 per arm, those sample-based estimates vary a lot from trial to trial. In a further check with 3,000 trials at each of these two sizes, the bias correction alone lowered coverage by 1 to 1.5 percentage points below the percentile interval's, and the acceleration by about 3.5 more. That is a finding

about these simulated scenarios, not a general verdict on BCa: where the sampling distribution really is skewed, as for a single mean from a skewed population, adjusting for skewness is the point of the method, though that problem is much harder for every method ^[2].

So in these scenarios, a difference in means between two equal arms of the same shape, the t interval is hard to beat, the percentile interval is too narrow with 10 per arm, and the bootstrap is no remedy for a small sample. Its strengths lie elsewhere.

6 What the bootstrap assumes, and what to resample

The bootstrap needs no normal distribution, but it is not assumption-free. Its essential assumption is that the sample can stand in for the population. That needs observations, or units of resampling, that are independent of one another, and enough of them for their spread and shape to resemble the population's. With 10 per arm, the sample's tails say little about the population's tails, as the 45 mmHg example showed.

Where those conditions hold, the bootstrap earns its keep with statistics that have no easy standard-error formula: a median in a large sample, a ratio, a correlation, an area under a ROC curve, or a summary of a whole analysis. It is also a useful sensitivity check on a formula-based interval: compute both and compare them ^[2]. Agreement can be reassuring, but it does not show that either interval has its nominal coverage, because both rest on the same sample and neither can see tails that the sample does not contain. Disagreement is a reason to find out why.

The general rule for what to resample is to resample in the way the data were drawn, keeping fixed what the design fixed, such as the size of each group ^[2]. Table 4 applies it to common designs. It mirrors the rule of the previous note, to shuffle only what was randomised. The entries are starting points, not guarantees: with few clusters, complex survey designs or long-range dependence, the resampling scheme needs more care, and whether a scheme gives consistent standard errors can depend on the model assumed for the data ^[6].

Table 4. Resample what was sampled.

DESIGN	WHAT TO RESAMPLE
Two independent groups	participants within each group, keeping the group sizes ^[2]
Stratified sampling or randomisation	participants within each stratum ^[2]
Matched pairs, crossover, before and after	whole pairs or whole participants, keeping each one's measurements together
Clustered data (practices, wards, schools)	whole clusters, with all their members ^[6]
Repeated measures over time	whole participants, with all their measurements
A time series	blocks of consecutive observations, not single observations ^[7]
Regression	whole rows (cases); or residuals, added back to the fitted values, when the <i>x</i> values were fixed by design ^[2]
Any analysis with data-driven steps (variable selection, tuning, imputation, outlier removal)	the raw data, rerunning every step in each resample

7 How many resamples?

Random resamples estimate the ideal bootstrap distribution with Monte Carlo error, just as random shuffles estimated the exact permutation *p*-value in the previous note. The error matters most for the ends of an interval, which depend on the few resamples in each tail. Table 5 shows how much the bootstrap standard error and the ends of the percentile and bootstrap-*t* intervals for the pilot data move from one run to the next.

Table 5. How precise are bootstrap results from *B* random resamples? For the pilot data: the central range that holds 95% of the results of repeated runs, each with *B* resamples (percentiles as in R's boot package), and the ideal values in the last row.

RESAMPLES, <i>B</i>	BOOTSTRAP STANDARD ERROR	PERCENTILE INTERVAL, LOWER END	PERCENTILE INTERVAL, UPPER END	BOOTSTRAP- <i>t</i> , LOWER END	BOOTSTRAP- <i>t</i> , UPPER END
999	3.02 to 3.30	0.8 to 1.8	13.2 to 14.3	−0.2 to 1.2	14.0 to 15.4
9,999	3.12 to 3.20	1.2 to 1.5	13.5 to 13.9	0.3 to 0.8	14.4 to 14.9
99,999	3.14 to 3.17	1.3 to 1.4	13.6 to 13.8	0.5 to 0.6	14.6 to 14.7
Ideal	3.16	1.3	13.7	0.5	14.6

With 999 resamples the lower end of the percentile interval can come out anywhere from 0.8 to 1.8 mmHg, and with 9,999 from 1.2 to 1.5 mmHg. The bootstrap-*t* interval moves more, because its ends depend on the tails of a distribution of ratios. Hesterberg recommends 10,000 resamples for routine use and more when accuracy matters, and calculates that at least 15,000 are needed for

Monte Carlo variation to have only a small effect on percentile and bootstrap-t intervals^[2]. As with shuffles, choose B so that the Monte Carlo error is small next to the precision you report, and give the seed so that others can reproduce the numbers.

8 How to report a bootstrap interval

Rule of thumb. Name the statistic, say what was resampled and within what, name the interval method, and give the number of resamples, the software and the seed, alongside the estimate.

A bootstrap interval is still a confidence interval, and the reporting advice of the earlier note applies to it. The methods section needs enough extra detail for someone else to get the same numbers:

- **The statistic:** difference in means, difference in medians, a ratio, or another.
- **What was resampled**, and within what: participants within arms, strata, clusters or pairs, matching the design.
- **The interval method.** "Bootstrap 95% CI" alone does not say which: percentile, BCa and bootstrap-t can differ by several mmHg, as in Table 2.
- **The number of resamples**, the software with its version, and the random seed.
- **With small samples**, why a bootstrap interval was used instead of, or alongside, a formula-based one.

8.1 Sentences to adapt

- **Methods.** "The 95% confidence interval for the difference in mean fall in systolic blood pressure (programme minus usual care) was a bootstrap-t interval from 9,999 bootstrap resamples, drawn with replacement separately within each arm (R 4.5.2, boot package 1.3-32, seed 2026)."
- **Results.** "Systolic blood pressure fell 7.5 mmHg more in the programme group than in the usual-care group (95% CI 0.6 to 14.4, bootstrap-t; the t interval was 0.5 to 14.5)."

The boot package does the resampling in a few lines. The interval in the results sentence is the one this code gives; it differs from the ideal 0.5 to 14.6 by Monte Carlo error of the size shown in Table 5.

```

library(boot)
prog <- c(21, 17, 15, 14, 13, 7, 6, 2, 2, -3)
usual <- c(11, 10, 9, 5, 2, 1, 0, -1, -9, -9)
d <- data.frame(fall = c(prog, usual), arm = rep(c("programme", "usual care"), each =
10))

# The difference in means, and its estimated variance (for the bootstrap-t), in one
resample
stat <- function(data, i) {
  x <- data[i, ]
  p <- x$fall[x$arm == "programme"]; u <- x$fall[x$arm == "usual care"]
  c(mean(p) - mean(u), var(p) / length(p) + var(u) / length(u))
}
set.seed(2026)
b <- boot(d, stat, R = 9999, strata = factor(d$arm)) # resample within each arm
sd(b$t[, 1]) # bootstrap SE: 3.11 (ideal 3.16)
boot.ci(b, type = c("perc", "stud")) # percentile 1.5 to 13.6;
# bootstrap-t 0.59 to 14.41
t.test(prog, usual, var.equal = TRUE)$conf.int # 0.51 to 14.49

```

9 Five traps

9.1 Reaching for the bootstrap because the sample is small

The simplest bootstrap interval is too narrow precisely when samples are small (Section 3): with 10 per arm, the percentile interval covered the true difference 92% of the time and the t interval 95% (Table 3). A small sample is a reason for caution with any method, not a reason to switch to the bootstrap^[2].

9.2 Resampling the wrong unit

Resampling individual participants in a cluster-randomised trial, or individual measurements in a repeated-measures study, treats correlated observations as independent. The resamples then vary less than real repeats of the study would, and the interval is too narrow. Follow Table 4.

9.3 Leaving steps out of the resampling

If the analysis made choices with the data, such as selecting variables, tuning a model, removing outliers or imputing missing values, each resample must make them again. Otherwise the bootstrap measures the variability of a simpler procedure than the one that was used, and the interval is too narrow. The same applies to bootstrap estimates of a prediction model's performance, where leaving steps out leaks information from the evaluation into the fitting (see my note on leakage).

9.4 "A bootstrap 95% CI" without saying which

For the data with the 45 mmHg value, the percentile, BCa and bootstrap-t intervals have upper ends of 19.5, 21.5 and 23.7 mmHg (Table 2). A reader cannot judge an interval without knowing which method made it, how many resamples it used and what was resampled.

9.5 Treating the bootstrap as a better estimate

The bootstrap's main job is to describe the estimate's uncertainty, not to replace the estimate^[2]. For a difference in means the ideal bootstrap distribution is centred exactly at the estimate, so the mean of random resamples differs from it only by Monte Carlo noise; for other statistics the two can differ, and the gap is a bootstrap estimate of bias. Correcting for it is possible, but bias estimates can be highly variable^[2]. Report the estimate from the data, and if you correct it for bias, say so and report both.

10 A checklist for authors and reviewers

- ☐ The statistic is named, and the estimate is reported with its interval.
- ☐ What was resampled matches the design: each arm kept at its size, and strata, clusters or pairs kept intact.
- ☐ Every data-driven step of the analysis was repeated in each resample.
- ☐ The interval method is named: percentile, BCa, bootstrap-t or another.
- ☐ The number of resamples is given, with the software and the seed, and it is large enough for the precision reported.
- ☐ With small samples, the bootstrap interval was compared with a formula-based one, and the choice is explained.
- ☐ The text does not claim that the bootstrap needs no assumptions or makes up for a small sample.

The bootstrap turns the sampling distribution into something you can watch being built, from the only sample you have. Every confidence interval has a sampling distribution behind it, whether it comes from a formula or from resampling, and an interval means only as much as the sampling distribution behind it.

References

- [1] Efron B. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*. 1979;7(1):1-26. doi:10.1214/aos/1176344552
- [2] Hesterberg TC. What teachers should know about the bootstrap: resampling in the undergraduate statistics curriculum. *The American Statistician*. 2015;69(4):371-386. doi:10.1080/00031305.2015.1089789
- [3] DiCiccio TJ, Efron B. Bootstrap confidence intervals. *Statistical Science*. 1996;11(3):189-228. doi:10.1214/ss/1032280214
- [4] Davison AC, Hinkley DV. *Bootstrap Methods and Their Application*. Cambridge University Press; 1997. doi:10.1017/CBO9780511802843

- [5] Efron B. Better bootstrap confidence intervals. *Journal of the American Statistical Association*. 1987;82(397):171–185. doi:10.1080/01621459.1987.10478410
- [6] Field CA, Welsh AH. Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2007;69(3):369–390. doi:10.1111/j.1467-9868.2007.00593.x
- [7] Künsch HR. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*. 1989;17(3):1217–1241. doi:10.1214/aos/1176347265