It's Time To Retire the "$n \geq 30$" Rule

Tim Hesterberg*

Abstract

The old rule of using z or t tests or confidence intervals if $n \geq 30$ is a relic of the pre-computer era, and should be discarded in favor of bootstrap-based diagnostics.

The diagnostics will surprise many statisticians, who don't realize how lousy the classical inferences are. For example, 95% confidence intervals should miss 2.5% on each side, and we might expect the actual non-coverage to be within 10% of that. Using a t interval, this requires $n > 5000$ for a moderately-skewed (exponential) population. There are better confidence intervals and tests, bootstrap and others.

The bootstrap also offers pedagogical benefits in teaching sampling distributions and other statistical concepts, offering actual distributions that can be viewed using histograms and other familiar techniques.

Key Words: Central limit theorem, bootstrap, normal distribution, diagnostics, resampling

1. Introduction

Confidence intervals and hypothesis tests based on Normal approximations and t-statistics are common throughout statistics. These are based on asymptotic results, that the distributions of estimators such as a sample mean or regression coefficient approach Normal distributions as sample sizes go to infinity, and that the corresponding t-statistics approach t distributions, if certain regularity conditions hold.

For finite simples, we rely on common rules of thumb, e.g. for a single mean of *i.i.d.* data, if the sample size is at least 30 and the sample is not too skewed, then one may proceed with Normal-based inferences.

But what does "not too skewed" mean? What diagnostics should we use for statistics other than the mean? And what if the sample size is small, or if the sample is noticeably skewed? Well, then one does Normal-based inferences anyway! (With some exceptions.)

And what diagnostic measures should we use in other situations, such as for logistic regression?

I claim in this article that:

- we should replace the "$n \geq 30$ and not too skewed" rule with more effective diagnostics based on the bootstrap,
- these bootstrap diagnostics are easy to apply,
- that the results will surprise many statisticians, showing how inaccurate t-based inferences are,
- that better alternatives to t-based inferences are available.

I'll also comment on two related points:

- 1000 bootstrap samples aren't enough for high-quality diagnostics, and
- while better inferences are available for large samples, more work is needed for small samples.

2. Bootstrap Diagnostics

I'll begin with a quick review of the bootstrap. For successively longer introductions see ((Hesterberg *et al.*, 2003)), ((Efron and Tibshirani, 1993)) or ((Davison and Hinkley, 1997)).

Suppose that $X_1, \ldots, X_n$ is an *i.i.d.* sample from a population F (possibly multivariate), that $\hat{\theta}$ is some estimate of a parameter θ. I assume here that $\hat{\theta}$ is a functional statistic, that depends on the data only through the empirical distribution $\hat{F}_n$ that has probability $1/n$ on each of the observed data points.

In the ordinary nonparametric bootstrap, we draw a sample from the empirical distribution (i.e. a sample with replacement from the data), $X_1^*, \ldots, X_n^*$, and calculate the corresponding statistic $\hat{\theta}^*$. Repeating this many times, say $B = 1000$, we obtain B bootstrap statistics $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ that comprise the bootstrap distribution, which we use for estimating standard errors, confidence intervals, or diagnostics.

*Google, 651 N. 34th St., Seattle WA 98103

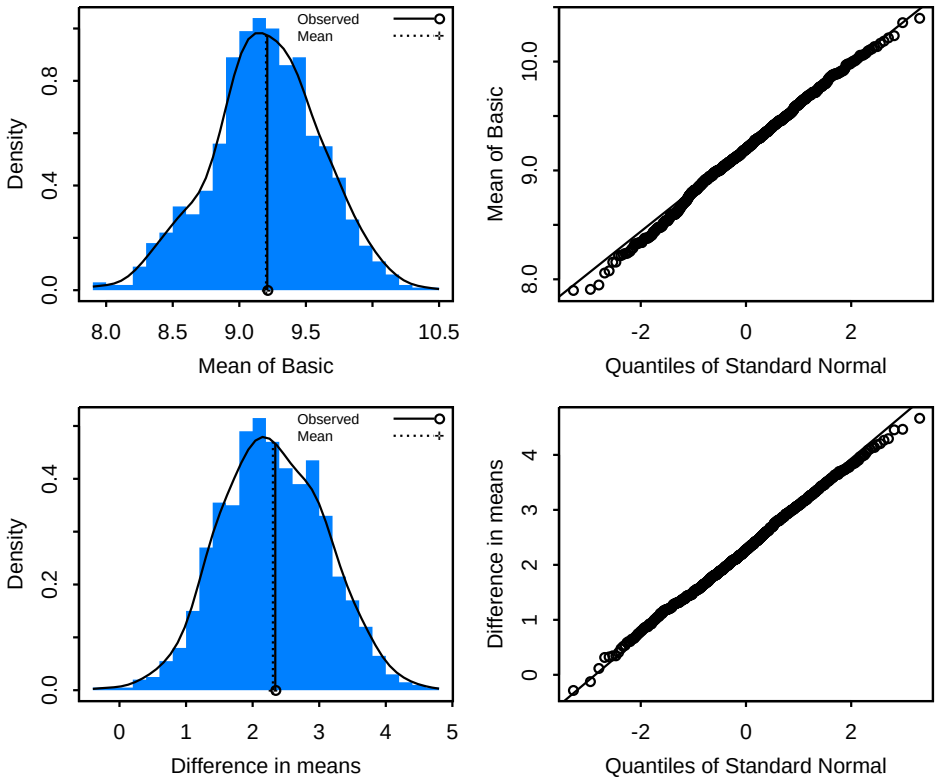

Figure 1: Bootstrap distributions for TV data. The top row gives a histogram and Normal quantile plot of bootstrap distribution for the mean of basic TV commercial times. The second row is for the difference in means between the basic and extended channels.

2.1 TV Means Example

For example, student Barrett Rogers collected data on the number of minutes of commercials per half-hour of basic and extended (extra-cost) cable TV, finding an average of 9.21 minutes of commercials for the basic channels, and 6.87 for the extended channels, based on 10 observations for each (the poor student could only bear to watch 20 random half-hours of TV). The data are 7.0, 10.0, 10.6, 10.2, 8.6, 7.6, 8.2, 10.4, 11.0, 8.5 for the basic channels, and 3.4, 7.8, 9.4, 4.7, 5.4, 7.6, 5.0, 8.0, 7.8, 9.6 for the extended channels. The bootstrap distribution for the mean of the basic times is shown in the top row of Figure 1.

For two-sample problems, we draw samples independently from the two samples, and compute the statistic of interest, e.g. a difference in means or hazard ratio for each pair of bootstrap samples. The bootstrap distribution for the difference in means between the basic and extended channels is shown in Figure 1.

In this case, even though the samples are quite small, the bootstrap suggests that the sampling distributions for the one-sample mean and difference in means are approximately Normal.

2.2 Verizon Means Example

The bootstrap gives a different picture in the next example. The data are shown in Figure 2. The larger "ILEC" sample consists of 1664 observations, repair times with mean 8.4 hours; the smaller "CLEC" sample has 23 observations with mean 16.5 ((Hesterberg *et al.*, 2003)). These correspond to repair times for two groups of customers, and the question of interest is whether the difference in means is different, at a one-sided significance level of 0.01.

The bootstrap distributions are shown in the bottom of Figure 2. The bootstrap distribution for the mean of the larger sample, $n = 1664$, appears approximately normal, but for the smaller sample there is substantial 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 translate into errors in inferences. We may quantify how badly this amount of skewness affects confidence intervals; we defer this to Section 3, in the context of bootstrap t distributions. First we consider additional examples, for statistics other than means.

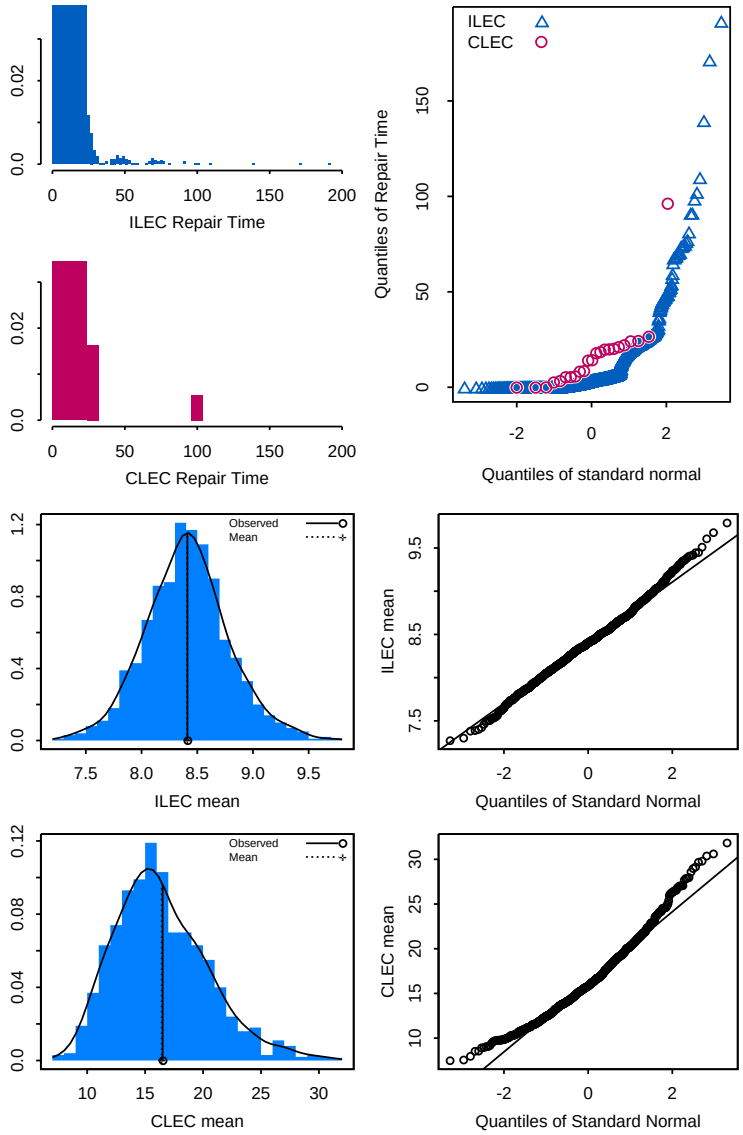

Figure 2: Repair times for Verizon data set; ILEC and CLEC groups ($n = 1664$ and $n = 23$, respectively). Data are in the top panels, and bootstrap distributions for the mean of each group at bottom.

2.3 Bushmeat Regression Example

Brashares *et al.* (2004) discuss the relationship between fish supply and the loss of wildlife due to bushmeat hunting in Ghana. Figure 3 shows 30 years of data, for per-capital fish supply and estimated total biomass, based on population estimates of 30 species in national parks, together with a scatterplot of fish supply and relative change in biomass. It is evident that there are greater declines in biomass. when fish supply is smaller. This suggests (and is supported by other evidence) that bushmeat hunting is more prevalent when fish supply is smaller.

The bottom left panel in Figure 3 shows regression lines from 20 bootstrap samples. One quantity of interest is the x-intercept; this gives an estimate of the fish supply that would result in zero average loss of wildlife. The bottom right panel shows the bootstrap distribution for the x-intercept; the distribution is strongly positively skewed, so Normal approximations would be inaccurate.

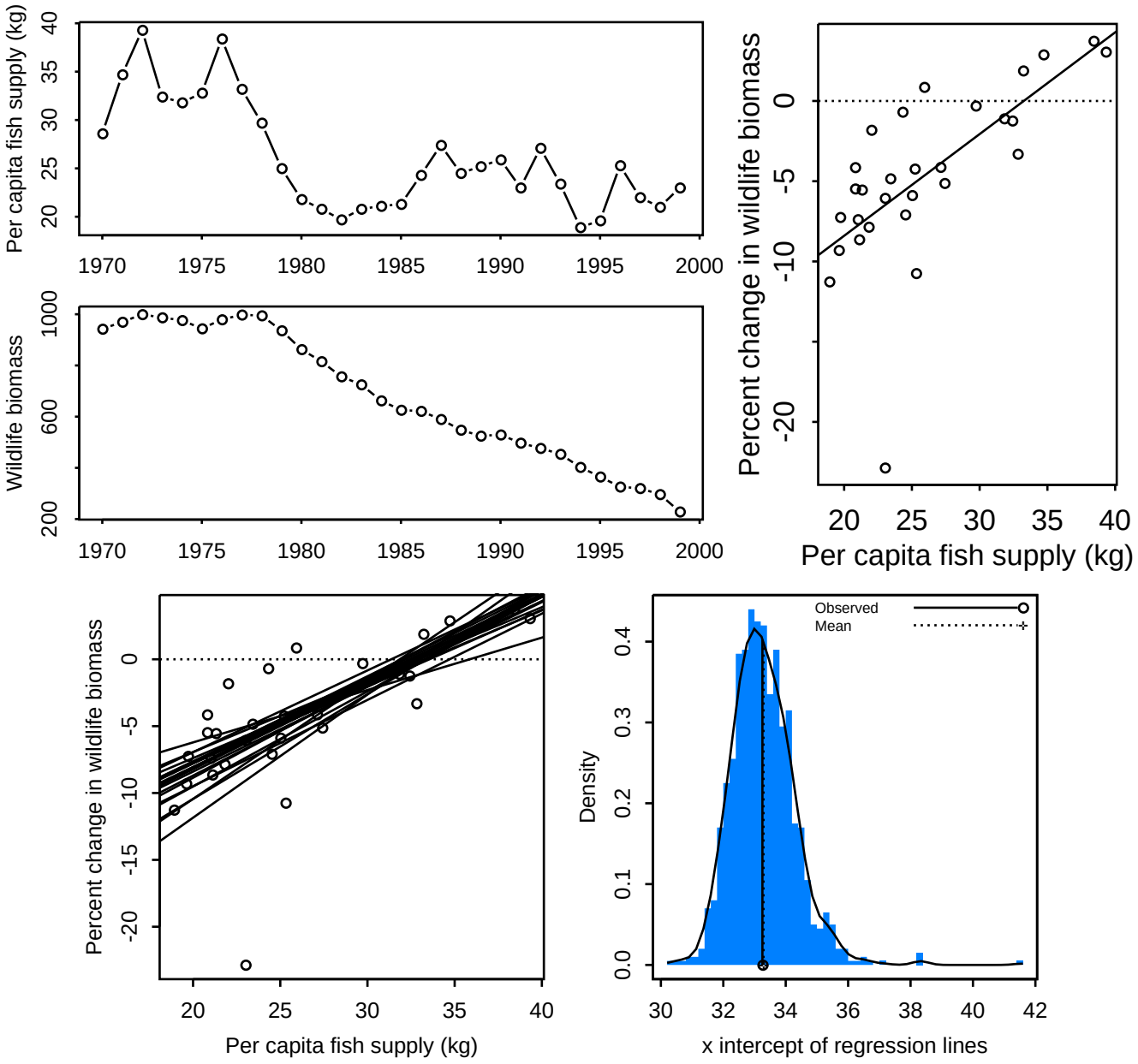

Figure 3: Fish supply and wildlife biomass, over 30 years in Ghana. The bottom panels show bootstrap lines, and bootstrap distribution for the x intercept.

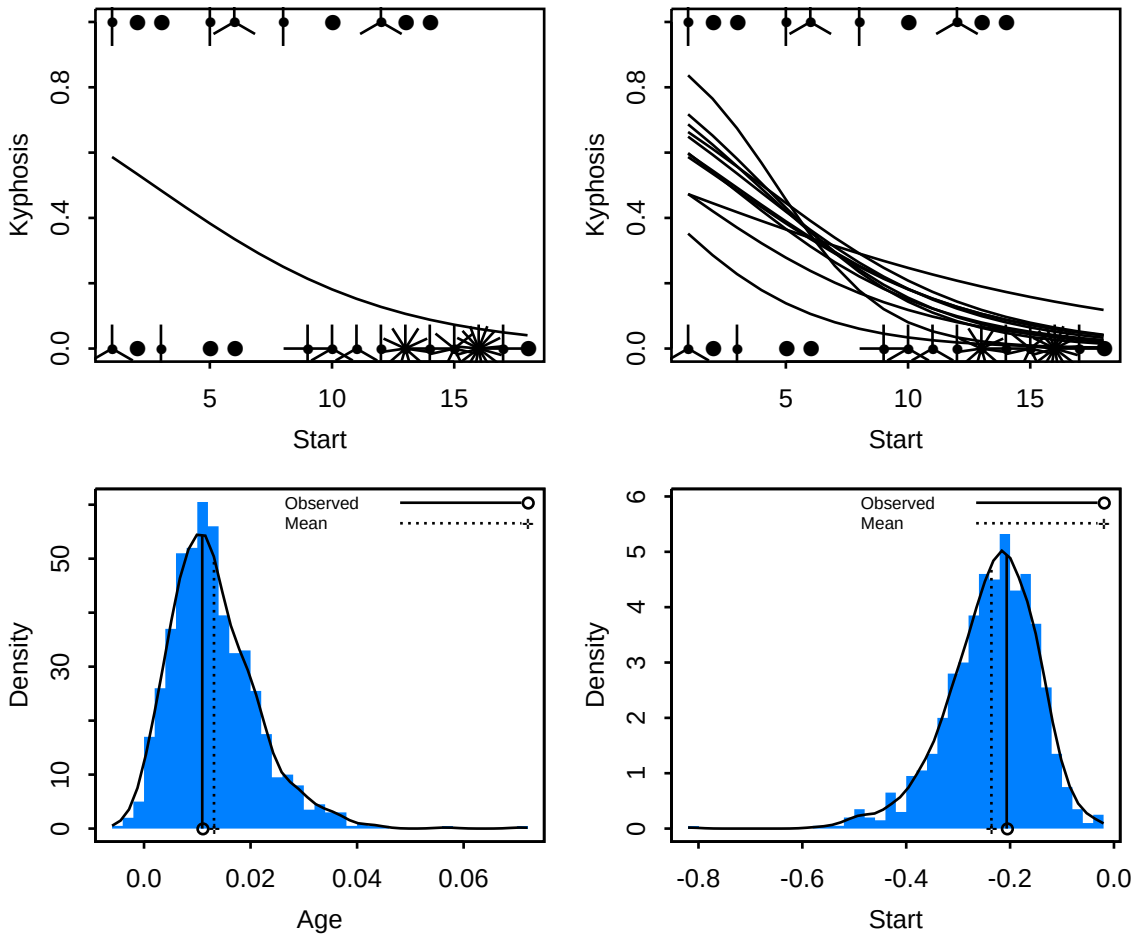

Figure 4: Kyphosis data and bootstrap distributions. The first panel is the response, Kyphosis, against the most informative of three covariates, Start, together with the prediction from logistic regression (when the other two covariates are set at their median values). The second panel shows predictions from 20 bootstrap samples. The bottom panels show bootstrap distributions for two of the four coefficients in logistic regression; the others are Intercept and Age.

2.4 Kyphosis Logistic Regression Example

The Kyphosis data set ((Chambers and Hastie, 1992)) consists of 81 observations on four variables—the response "Kyphosis" is a binary variable indicating whether a postoperative deformity is present, and covariates are Age, Number (of vertebrae involved in the operation) and Start (number of the first vertebrae involved in the operation). We run a logistic regression of Kyphosis against the covariates.

Figure 4 shows a sunflower plot of Kyphosis against Start (the most informative of the covariates), together with predictions from the logistic regression (with Age and Number set at their median values). The top right figure shows 20 bootstrap curves for this prediction. Considerable variation is evident. One quantity of interest is the sampling distribution of predictions for fixed values of the covariates; here, for larger values of Start, and the other covariates at their median, the bootstrap distribution of the predictions is bounded below by zero and is strongly positively skewed.

The bottom two panels show the bootstrap distributions for two of the regression coefficients, Age and Start; the bootstrap distributions are strongly skewed, so Normal approximations would not be appropriate.

It is interesting to note that the printout for logistic regression from one statistical package, S-PLUS, shows admirable restraint—it gives the coefficients, standard errors, and t statistics, but does not give P-values associated with those t statistics. That is appropriate because the t statistics do not follow t distributions. Unfortunately, not all packages are so restrained.

One common thread in these examples, and other examples of statistics other than a sample mean, is that the sampling distributions are inherently skewed. In that case one should be even more reluctant to rely on a central

limit theorem than in the case of a mean.

3. Bootstrap t

We turn now to the bootstrap distribution for the t statistic. Classical statistical theory indicates that when underlying distributions are normal, the distribution of the sample mean $\overline{X}$ and sample standard deviation s are independent, and the t-statistic $t = (\overline{X} - \mu)/(s/\sqrt{n})$ follows a t-distribution. We may use the bootstrap to diagnose how close the actual distribution is to a t-distribution.

In general, let $\hat{\theta}$ be an estimator and that $s_{\hat{\theta}}$ a standard error for $\hat{\theta}$, and let

$$t = (\hat{\theta} - \theta)/s_{\hat{\theta}} \tag{1}$$

be the t-statistic. The bootstrap analog is

$$t^* = (\hat{\theta}^* - \hat{\theta})/s^*_{\hat{\theta}} \tag{2}$$

where $s^*_{\hat{\theta}}$ is the standard error calculated from a bootstrap sample.

Figure 5 shows bootstrap diagnostics related to t-statistics for the Verizon dataset. The top left panel shows a scatterplot of $\overline{X}^*$ vs s^* for the larger ($n = 1664$) ILEC dataset. Because of the skewness in the data, these quantities are not independent, but instead are strongly correlated. This is because those bootstrap samples with relatively large means typically have more observations from the long right tail, and hence larger standard deviations. The next two panels are histograms of the bootstrap distributions for the t statistic; the distribution for the smaller CLEC example is strongly negatively skewed, and is bimodal (due to the presence or not of the large observation). The distribution for the larger ILEC sample is more Normal, perhaps with mild negative skewness. The final panel is a Normal quantile plot for the bootstrap distribution of the ILEC data set, against showing mild negative skewness.

Note that although the distribution for the mean is positively skewed, the t statistic is negatively skewed. This is because when the mean is small, the numerator of 2 is negative and the denominator tends to relative near zero, so the result has relatively large negative values.

3.1 Numerical diagnostics

A visual inspection suggests that for $n = 1664$ the bootstrap t distribution is only mildly skewed. However, the final panel includes additional notes giving numerical results, that upset any sense of complacency we may have felt after the visual inspection. If the bootstrap t distribution actually follows a t distribution, we would expect 2.5% of the bootstrap samples to fall below $-t_{1663,.025}$ and above $t_{1663,.025}$. The actual proportions are 3.94% and 2.12%.

The 3.94% noncoverage is 58% larger than the nominal 2.5% value. That kind of error is unacceptable. Errors like that could cause a meltdown of our financial system. (Oops.)

If 1664 observations isn't enough for t inferences to be reasonably accurate, then we can toss the old $n \geq 30$ rule out the window.

But is the situation really that bad?

4. Convergence of t statistic, and alternatives

Yes, it is that bad. It takes nearly forever for t inferences to be reasonably accurate when the population is skewed. The central limit theorem operates on geological time scales.

Consider those two elements: "reasonably accurate", and skewed.

I define a "reasonably accurate confidence interval" to have actual non-coverage probabilities within 10% of the nominal values on each side. A 95% confidence interval nominally misses 2.5% of the time on each side. A reasonably accurate interval would miss somewhere between 90% and 110% of that nominal value, or between 2.25% and 2.75% on each side.

Note that it is the non-coverage on each side that matters, not the total non-coverage. Few practical statistical problems are truly two sided—it is nearly always the case that missing one one side has a different practical effect than missing on the other side. Furthermore, a lopsided interval (say with 4% non-coverage on one side and 1% on the other) gives a biased impression of where the the true parameter is likely to be. Errors on the two sides do not cancel out; the overall accuracy should be measured by the sum of absolute errors from the two sides.

Now turn to skewness. It is well-known that when the underlying population is skewed, the distributions of $\overline{X}$ and t converge to Normal and t distributions at the relatively slow rate of $O(1/\sqrt{n})$.

For the sake of argument, I consider an exponential population, because it is well-known, and convenient for simulations. Note that while an exponential distribution appears very skewed, its numerical skewness of 2 is not that large. Its right tail declines exponentially fast. Distributions with polynomial tails, such as F distributions

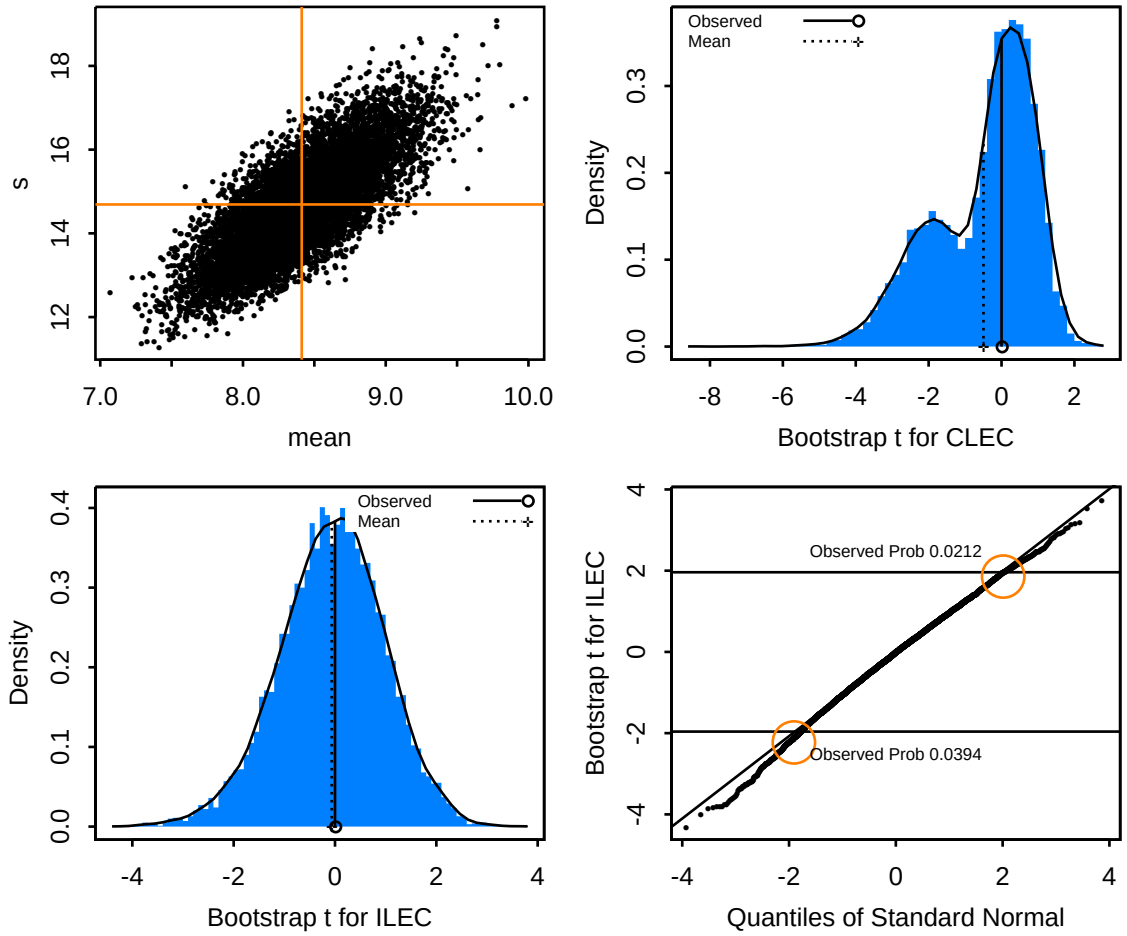

Figure 5: Bootstrap distributions of the t statistic, for Verizon data. The top left panel shows a scatterplot of $\overline{X}^*$ vs s^* for the larger ($n = 1664$) ILEC dataset. The remaining panels are histograms and Normal quantile plots of the bootstrap distribution of the t statistic, for CLEC and ILEC datasets.

with small denominator degrees of freedom, can appear to be closer to Normal, but have higher skewness. And in practice many distributions appear to exhibit polynomial tail behavior.

Figure 6 shows simulation results when the population is exponential, for t confidence intervals and some alternatives. The first panel gives a general idea of the convergence of the t interval on each side, as well as that of an alternative Johnson's skewness-adjusted t-statistic $\overline{x} + \frac{s}{\sqrt{n}}(t_\alpha + \frac{s_3}{6\sqrt{n}}(1 + 2t_\alpha^2)$ where s_3 is the sample skewness ((Johnson, 1978)). The Johnson adjusted interval converges to the nominal much faster than does the t interval. The second panel is a rescaled version, showing that the unadjusted t-statistic reaches reasonable accuracy only when n exceeds 5000. The third panel shows that the Johnson procedure reaches reasonable accuracy at about $n = 220$—not great, but much better than 5000. The final panel shows the convergence of those and two additional intervals, bootstrap t intervals ((Efron, 1981, 1982; Efron and Tibshirani, 1993)), and the ABC interval ((DiCiccio and Efron, 1992)), a non-sampling approximation to a bootstrap BCa interval ((Efron, 1987)). These, and a number of other bootstrap and non-bootstrap confidence intervals, are "second-order correct", with coverage converging at the rate $O(1/n)$ ((Efron and Tibshirani, 1993)). In contrast, t intervals are only "first-order correct".

While the standard t interval does converge to the right coverage values, it does so only slowly.

5. Bootstrap sample sizes - the 64,000 sample question

The Verizon bootstrap t distributions are based on 10^4 bootstrap samples. The larger one-sided non-coverage estimate there is 0.0394; The standard error of that estimate is is $\sqrt{0.0394 * (1 - 0.0394)/10^4} = 0.0019$, so the value 0.0394 is over 7 standard errors away from the nominal value.

We note that the number of bootstrap samples needed for highly accurate numerical diagnostics is much larger than traditional recommendations for bootstrap sample sizes. Efron and Tibshirani (1993) suggest that $B = 200$, or

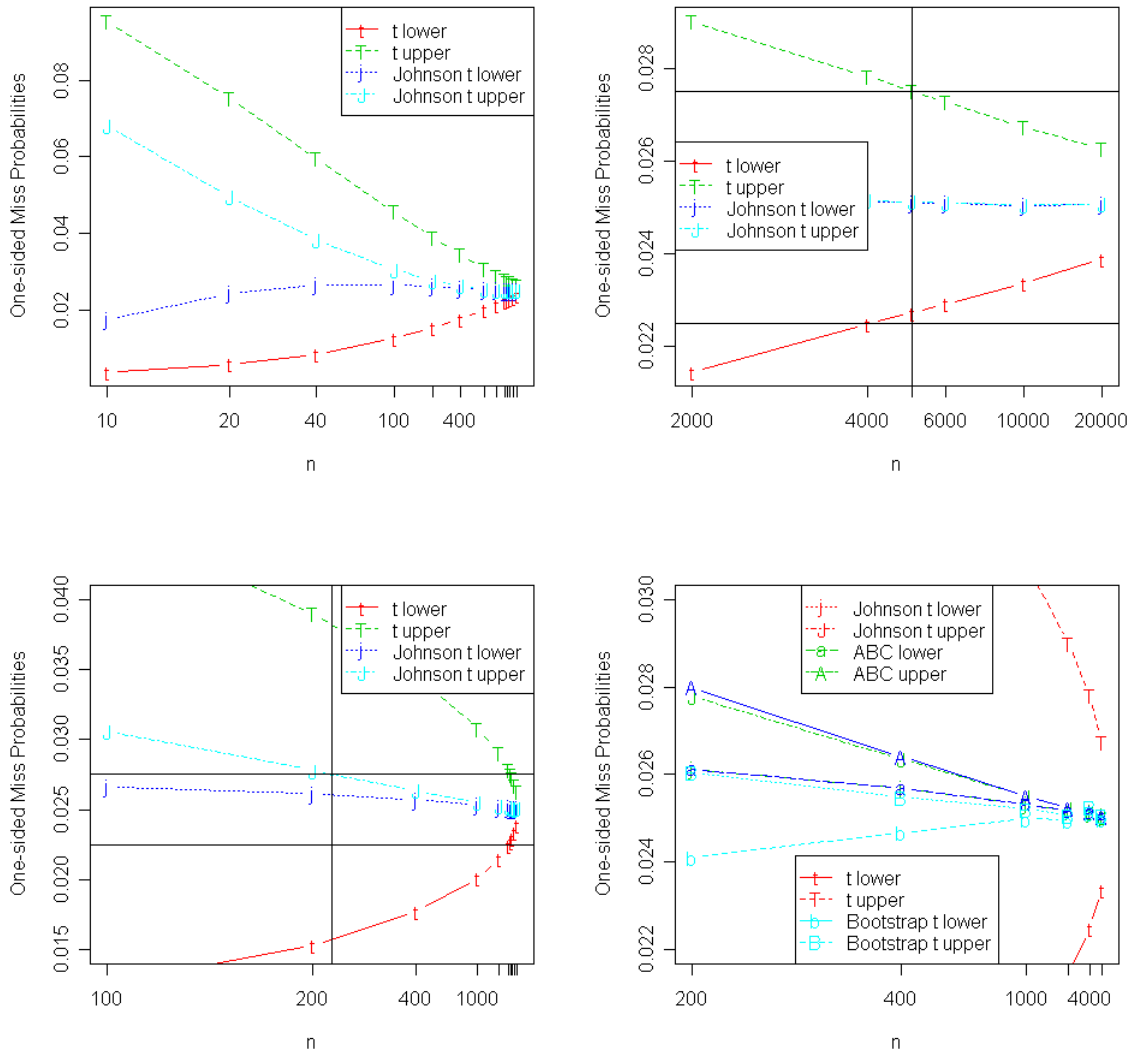

Figure 6: Simulation results for convergence of t intervals and alternatives

even as few as $B = 25$, suffices for estimating standard errors, and that $B = 1000$ is enough for confidence intervals.

When the non-coverage is 2.5%, and with a "reasonable accuracy" margin of error of one tenth of that, in situations where high accuracy is needed one would like to reduce the standard error to say one-fourth of that margin of accuracy, 0.025/10/4. This requires approximately $B = 64000$ bootstrap samples. OK, actually 62,400—but 64,000 has a nicer ring to it.

6. Summary

The "$n \geq 30$ rule is a relic of the pre-computer era; it is time to retire it.

The central limit theorem operates on geological time scales.

Why aren't these issues known? I believe a prime reason is that people haven't been aware of the use of the bootstrap for diagnostics.

There are additional reasons. A big reason is computational. The computer simulations that produced Figure 6 would have taken around 20,000 hours in 1981, relatively early in the bootstrap era and shortly after the development of the Johnson adjustment. Even now, 64,000 bootstrap samples is a lot, too many to run routinely. But even 1000 bootstrap samples can yield approximate diagnostics.

Second, we need good alternatives. For a simple mean, Johnson's procedure is reasonably good if the sample size is somewhat larger, but may not be good for small samples, where it is hard to estimate skewness accurately. I

believe we need a semi-Bayesian approach, in which estimated skewness is shrunken toward zero, with more shrinkage for smaller samples.

For comparison, t intervals are Bayesian with the prior distribution for skewness a point mass at zero.

For statistics other than the mean, the ABC interval and a variety of bootstrap and non-bootstrap intervals are second-order correct, but under-cover for small samples. Many of them, when applied to a single mean when the population is really Normal and the data symmetric, yield confidence intervals of approximately $\overline{x} \pm z_{\alpha/2}\hat{\sigma}/\sqrt{n}$, where $\hat{\sigma}^2 = n^{-1}\sum(x_i - \overline{x})^2$. They are too short, shorter than t intervals by two factors—using $z_{\alpha/2}$ in place of $t_{\alpha/2,n-1}$, and using $\hat{\sigma}$ in place of s. These terms are $O(1/n)$—they don't affect asymptotic second-order correctness, but they really matter for small n. See simulation comparisons, and some adjustments for these factors, in ((Hesterberg, 1999)).

I should also mention that there are situations where skewness doesn't matter. For comparing two means, for example, if the sample sizes are equal and the populations have the same s and skewness, then the skewness cancels and the t distribution is symmetric.

Aside from cases with such structural cancellation. For large samples, say $n > 100$, one should clearly use a second-order correct procedure, unless the sample size is extremely large—e.g. greater than 5000, or more with more skewness or inherently-skewed statistics.

For medium to small sample sizes, second-order procedures should probably be used, but with some care, and adjustments for the factors that make them too short, and perhaps shrinking the skewness adjustments toward zero. Further work is needed to quantify these recommendations.

A final reason is inertia—people are used to the simple diagnostic. This where statistics education comes in. We should teach students the bootstrap diagnostics, rather than the old rule. This has other benefits—doing bootstrapping, and seeing pictures of bootstrap distributions, will help give them a better understanding of statistical concepts such as sampling distributions, the central limit theorem, standard error, bias, and P-values.

References

Brashares, J. S., Arcese, P., Sam, M. K., Coppolillo, P. B., Sinclair, A. R. E. and Balmford, A. (2004) Bushmeat hunting, wildlife declines, and fish supply in west africa. *Science*, **306**, 1180–1183.

Chambers, J. and Hastie, T. (1992) *Statistical Models in S*. Wadsworth, California.

Davison, A. and Hinkley, D. (1997) *Bootstrap Methods and their Applications*. Cambridge University Press.

DiCiccio, T. and Efron, B. (1992) More accurate confidence intervals in exponential families. *Biometrika*, **79**, 231–245.

Efron, B. (1981) Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics*, **9**, 139 – 172.

Efron, B. (1982) *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.

Efron, B. (1987) Better bootstrap confidence intervals (with discussion). *Journal of the American Statistical Association*, **82**, 171 – 200.

Efron, B. and Tibshirani, R. J. (1993) *An Introduction to the Bootstrap*. Chapman and Hall.

Hesterberg, T., Monaghan, S., Moore, D. S., Clipson, A. and Epstein, R. (2003) *Bootstrap Methods and Permutation Tests*. W. H. Freeman. URL `http://bcs.whfreeman.com/ips5e/content/cat_080/pdf/moore14.pdf`. Chapter for *The Practice of Business Statistics* by Moore, McCabe, Duckworth, and Sclove.

Hesterberg, T. C. (1999) Bootstrap tilting confidence intervals. Research Department Technical Report 84, MathSoft, Inc. URL `http://www.insightful.com/Hesterberg/articles/tech84-tiltingCI.pdf`.

Johnson, N. J. (1978) Modified t tests and confidence intervals for asymmetrical populations. *Journal of the American Statistical Association*, **73**, 536–544.