USING SIMULATION METHODS IN STATISTICS INSTRUCTION: EVALUATING ESTIMATORS OF VARIABILITY

Bruce E. TRUMBO and Eric A. SUESS, California State University, Hayward
Contact—Bruce E. Trumbo, Statistics Department, California State
University, Hayward, CA 94542 USA (btrumbo@csuhayward.edu)

Key Words: Simulation, Estimator evaluation, Range estimate of sigma, Teaching, Bootstrap confidence interval (parametric), Control chart.

Abstract:

Statistics instruction for the current job market must include both theoretical principles and modern computer methods. Combining these objectives, we use numerical and graphical output from computer simulations to illustrate principles of evaluating estimators (e.g., bias, MSE). *Main example:* properties of the range of a small normal sample as an estimator of population standard deviation (used in industrial control charts). *Software:* S-Plus, R, Minitab. *Level:* Introductory engineering statistics through first-year graduate mathematical statistics.

1. Introduction

Advances in computer hardware and statistical software continue to change statistical practice. Employers expect graduates of statistics programs to have a variety of computational skills in addition to a solid background in statistical theory and methodology. Statistics educators have responded to this expectation by adding software-oriented courses to statistics programs and by integrating computer methods for data analysis into basic applied courses and courses on methodology.

In theoretical textbooks, various authors have also shown the importance of computational methods. Notable examples are Feller (1957), Lindgren (1962), and Rice (1995). These uses of computation can go beyond data analysis to illuminate traditional principles and explore new methods of modeling and inference.

We believe that computational methods should play an increasingly important role in theoretical courses. Here we illustrate simulation methods that are essentially parametric bootstraps to investigate properties of estimators of variability. To facilitate classroom use we include S-Plus computer code and suggest some simulation projects and theoretical exercises for students.

While we have found that S-Plus and R are convenient software packages for the kinds of computation we do in mathematical statistics courses, most of the examples in this article could be restructured for Minitab. The results in this paper have been computed using S-Plus 2000 (Release 2) running under Windows, but the code we show would run in R with little or no modification. R has the advantage of being free.

2. Biases and Mean Squared Errors of Variance Estimators

We begin with a summary of facts about estimating the variance of a normal population. Depending on the level of the course, students might be asked to provide analytic proofs of all, some, or none of these results. Our purpose here is to illustrate these results by simulating some of the quantities involved.

Let X_1, X_2, ..., X_n be a random sample from a normal population with unknown mean μ and variance θ. The maximum likelihood estimator of θ is

$$\hat{\theta}_M = \Sigma(X_i - X)/n = Q/n\,.$$

Because Q/θ is distributed CHISQ($n-1$), we have E(Q/n) = $n-1$ and $E(\hat{\theta}_M)$ = $[(n-1)/n]\theta$, so that $\hat{\theta}_M$ is a biased estimator of θ, and the bias is more serious for smaller sample sizes. The unbiased estimator $\hat{\theta}_U$ = $Q/(n-1)$ is commonly used in practice.

One criterion for the “goodness” of an estimator $\hat{\theta}$ of a parameter θ is that it have a small mean squared error

$$\mathrm{MSE}(\theta) = E[(\hat{\theta} - \theta)^2] = V(\hat{\theta}) - [b(\hat{\theta})]^2\,,$$

where $b(\hat{\theta}) = E(\hat{\theta}) - \theta$ is the bias. Among unbiased estimators of θ (for which $b(\hat{\theta})$ = 0), one can show that $\hat{\theta}_U$ has the smallest variance—that is, it is UMVUE. Mathematical statistics courses often dwell on the elegant theoretical results for finding UMVUEs. However, in estimating θ,

$$\mathrm{MSE}(\hat{\theta}_U) = V(\hat{\theta}_U) > \mathrm{MSE}(\hat{\theta}_M)\,.$$

Furthermore, among constant multiples (in n) of Q, the smallest MST is achieved by $\hat{\theta}_A$ = $Q(n-1)$. (See Lindgren, 1962).

The S-Plus script below simulates m = 50,000 samples of size $n = 5$ from a normal population with (arbitrarily) mean $\mu = 150$ and standard deviation $\sigma = \theta^{-1/2} = 10$, in m rows and n columns of the matrix `Dta`. Each row of the matrix is considered as a sample of size n. (The function `rnorm` simulates a vector of mn independent observations from the desired population and `matrix` formats the vector into an $m \times n$ matrix.) Then we find m-component column vectors of observations from $\hat{\theta}_U$, $\hat{\theta}_M$ and $\hat{\theta}_A$. Finally, we compute sample means and MSEs of these vectors. For our purposes, these simulated values adequately approximate the corresponding theoretical ones. As shown below, we obtained $\text{MSE}(\hat{\theta}_U) \approx 5003.439 = 5.00\theta$, $\text{MSE}(\hat{\theta}_M) \approx 3.61\theta$, and $\text{MSE}(\hat{\theta}_M) \approx 3.34\theta$.

```
m <- 50000;  n <- 5
mu <- 150;  sigma <- 10; theta <- sigma^2
x <- rnorm(m*n, mu, sigma)
Dta <- matrix(x, nrow=m, ncol=n)
th.u <- rowVars(Dta)
mean(th.u);  mean((th.u-theta)^2)
th.m <- ((n-1)/n)*th.u
mean(th.m);  mean((th.m-theta)^2)
th.a <- (n/(n+1))*th.m
mean(th.a);  mean((th.a-theta)^2)
```

```
> mean(th.u)            > mean(th.m)
[1] 99.83805            [1] 79.87044
> mean((th.u-theta)^2)  > mean((th.m-theta)^2)
[1] 5003.493            [1] 3607.418

> mean(th.a)
[1] 66.5587
> mean((th.a-theta)^2)
[1] 3342.083
```

See Fig. 1 for histograms of the simulated distributions of these three estimates of θ. Sample code:

```
hist(th.u[th.u<400], nclass=30,
  xlab="Q/(n-1)).
```

These distributions are severely right-skewed, so MSE is minimized for a more extremely negatively biased estimator $\hat{\theta}_A$ than may be desirable in some applications. Also the population standard deviation is often estimated by taking the square root of one of these estimators. Used in this way even $\hat{\theta}_U$ gives a negatively biased estimate of s, as we see in the next section.

Technical notes. We used `set.seed(1212)` for the run shown above. If you use the same seed and software we did, you will get exactly the same results. Other simulation runs give similar results. Larger values of m give better approximations. In R: for `rowVars(Dta)` substitute `apply(Dta, 1, var)`, which also works in S-Plus but more slowly.

Suggested student exercises

1. Change the program above to include the estimator $Q/(n + 2)$, for which the approximate MSE is 3.5θ. This provides a rough indication that the denominator $n + 1$ gives the smallest MSE.

2. Compare the simulated quantiles of Q/θ with the exact quantiles of CHISQ($n - 1$). Code:

```
probs <- c(.025,.25,.5,.75,.975)
((n-1)/theta)\,*\,quantile(th.u,\,probs)
qchisq(probs,\,n-1)
```

3. Use the simulated quantiles of the distribution of $\hat{\theta}_U$ to make an approximate 95% confidence interval for θ based on five observations that give $\hat{\theta}_U = 89.1$. This is one kind of nonparametric bootstrap confidence interval (see Rice, 1995). Compare your result with the 95% conficence interval based on the exact distribution of $\hat{\theta}_U$.

4. Which of the estimators $\hat{\theta}_U$, $\hat{\theta}_M$ and $\hat{\theta}_A$ has the smallest mean absolute error, defined as $E(|\hat{\theta} - \theta|)$?

5. Repeat the main simulation and selected exercises for n = 4, 6, and 10.

2. Estimators of Standard Deviation

In some industrial applications where the sample size n is small, it is customary to estimate the standard deviation σ of a normal population by a suitable multiple of the sample range R rather than by the sample standard deviation $S = \hat{\theta}_U^{1/2}$. One particularly common use of range-based estimates of σ is in control charts. Here we use simulation to show that such estimates are reasonable for small sample sizes.

Usually the constant K_n is chosen so that $R_U = R/K_n$ is an unbiased estimator of σ.