Using Computer Simulation to Investigate Relationships Between the Sample Mean and Standard Deviation

Bruce E. Trumbo **Christopher M. Fraser**

Eric A. Suess

1. Introduction

Computer simulation is an important tool for probability modeling and statistical analysis. Some probability computations of great practical importance are too difficult to solve by "doing the math." Modern methods of statistical analysis such as bootstrapping and Gibbs sampling[1] depend on computer randomization and simulation.

Simulation methods can also be used in research and instruction to explore known relationships and sometimes even to discover unexpected ones. Our goal is to illustrate this flavor of exploration and discovery. We don't claim to reveal any research breakthroughs, but some of the things we will show you were new to us and we hope they will tempt you to try doing some simulations of your own.

2. When Are the Sample Mean and Standard Deviation Independent?

Perhaps one of the most important and least intuitive theorems of classical statistics states that, for data from a normal population, the sample mean $\bar{X}$ and standard deviation s are independent.[2]

One reason this fact is important is that the derivation of Student's t-distribution depends on it. In practice, if $\bar{X}$ and s are not independent, then tests of hypothesis and confidence intervals based on the t-distribution may give misleading results.

However, it seems strange that $\bar{X}$ and s would be independent. Each of these statistics arises from somewhat different manipulations of the *same data* $X_1, X_2, \ldots, X_n$. The definition and "computation formula" for s even involve $\bar{X}$:

$$s = \sqrt{\frac{\Sigma (X_i - \bar{X})^2}{n-1}} = \sqrt{\frac{\Sigma X_i^2 - n\bar{X}^2}{n-1}}$$

Sometimes proofs satisfy the intuition as well as the intellect, but the standard proofs of this independence use such methods as multivariate transformations, matrix manipulations, and moment generating functions, which have little intuitive appeal for many students.

Intuitive suspicions about the claimed independence of $\bar{X}$ and s are not entirely misguided. It is only for *normal* data that these statistics are independent. And in many real-life situations normality is more a theoretical ideal than a fact.

In the following sections we explore the independence of the sample mean and standard deviation for simulated data from three distributions:

Bruce Trumbo and Eric Suess are faculty members in the Statistics Department at California State University, Hayward. Currently, Chris Fraser is a PhD student at Purdue University; when this work was done, he was a student in the MS program at CSU Hayward.

Display 1: Minitab Code For the Simulation of Section 1

```
MTB > name c1 'x1' c2 'x2' c3 'x3' c4 'x4' c5 'x5'
MTB > name c6 'XBAR' c7 'SD'
MTB >   # name the columns for data and statistics
MTB > random 10000 'x1' - 'x5';
SUBC>  normal 100 10.
MTB >   # generate 10000 samples of size 5
MTB > rmean 'x1' - 'x5' 'XBAR'
MTB > rstdev 'x1' - 'x5' 'SD'
MTB >   # compute mean and std deviation of each row
MTB > describe 'XBAR' 'SD'  # descriptive statistics
MTB > corr 'XBAR' 'SD'      # correlation
MTB > plot 'XBAR' * 'SD';   # scatter plot
SUBC>  symbol;              # change plotting symbol
SUBC>    type 5.            #   from default to dot
```

normal, exponential, and uniform. The results for the three distributions are remarkably different.

We use Minitab for the simple simulations in this article because it is widely available and easy to understand. S-Plus is also a suitable software package for our simulations.[3] Practical suggestions for doing these simulations can be found in Section 6.

3. A Simulation With Normal Data

In this section we illustrate the independence of the sample mean and standard deviation for samples from a normal distribution.

Simulation plan. In this simulation we repeatedly take samples of size $n = 5$ from a normal population with $\mu = 100$ and $\sigma = 10$. (Any mean and standard deviation would do.)

Each simulated sample runs across one *row* of the Minitab worksheet in columns labeled 'x1', 'x2', ..., 'x5'. For each sample (row), we put $\bar{x}$ in the sixth column labeled 'XBAR' and s in the seventh column 'SD'.

We repeat (iterate) this sampling procedure $m = 10{,}000$ times. Thus we will use 10,000 rows in the worksheet. With so many iterations, operations involving the *columns* 'XBAR' and 'SD' simulate the corresponding theoretical results with satisfactory accuracy. For example:

- The mean of 'XBAR' simulates $E(\bar{X})$,
- The mean of 'SD' simulates E(s),
- The sample correlation of these two columns simulates the theoretical correlation $\rho(\bar{x}, s)$
- The scatter plot of these two columns indicates the nature of the joint distribution of $\bar{X}$ and s.

Minitab "program." Display 1 shows the Minitab commands necessary to carry out the planned simulation. Comments follow the #-signs; these comments are not used by Minitab and you need not type them when you try this simulation for yourself.

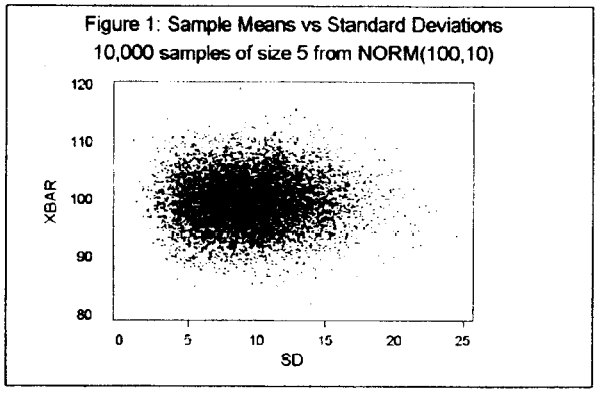
Simulation results. Display 2 shows the contents of the first two rows of the Minitab worksheet after one of our runs of the program in Display 1. You should verify the entries in the columns 'XBAR' and 'SD' for yourself.

As a "reality check" of the performance of the program, we look at means and standard deviations of the 'XBAR' column for three runs of this simulation ($m = 10{,}000$ each). The results are $E(\bar{X}) \approx$ 100.02, 99.948, and 100.03 (the symbol $\approx$ indicates *simulated as*). These results are very close to the true value $\mu = 100$. Also $SD(\bar{X}) \approx 4.46, 4.52, 4.49$, which are not far from the true value $\sigma/\sqrt{n} = 10/\sqrt{5} = 4.472$.

More to the point of our demonstration, we obtained the following simulated correlations

Display 2: Example From Minitab Worksheet

c1	c2	c3	c4	c5	c6	c7
x1	x2	x3	x4	x5	XBAR	SD
84.310	100.736	108.214	99.551	103.787	99.3196	9.03325
102.621	94.627	87.565	109.659	111.761	101.247	10.1760



between 'XBAR' and 'SD': $\rho \approx -0.010$, 0.011, −0.009, all very close to the 0 value one would expect for independent $\bar{X}$ and s. Of course, it is possible for *associated* data to have $\rho = 0$. But none of our scatter plots showed evidence of association. See Figure 1 for one of these plots.

Certainly, such a plot of $\bar{X}$ and s does not provide a proof of their independence for normal data. But it does illustrate that their joint distribution has no clear pattern of association. In Sections 4 and 5 you *will* see such clear patterns of association for exponential and uniform data.

Things to try on your own. Above we looked at summary statistics for 'XBAR'. If you look at the means of 'SD' in several simulation runs, you may be surprised to see that these simulated values of E(s) are around 9.40 rather than $\sigma = 10$. Actually, this is as it should be. The sample *variance* s^2 is an unbiased estimator of the population variance σ^2; in symbols, $E(s^2) = \sigma^2$. That's the reason for the $n - 1$ (instead of n) in the denominator of s^2. But s is *not* an unbiased estimator of σ: in fact, for $n = 5$, $E(s) = 0.940\sigma$. (The bias decreases as n increases.)

(a) After running the Minitab program shown in Display 1, issue the following three commands:

```
MTB > name c8 'VAR'
MTB > let 'VAR' = 'SD' * 'SD'
MTB > describe 'VAR'
```

The mean of 'VAR' simulates $E(s^2)$, which should be close to $\sigma^2 = 100$.

(b) *More advanced:* For any n, notice that $Y = \sum(X_i - \bar{X})^2 / \sigma^2$ has a chi-squared distribution with $n - 1$ degrees of freedom (df). Show that

$$E(s) = KE(\sqrt{Y}) = K\sqrt{2}\ \Gamma(n/2) / \Gamma((n-1)/2),$$

where $K = \sigma/\sqrt{(n-1)}$ and Γ is the gamma function. [Hint: The density of the chi-squared distribution with n df integrates to 1.] Show that $E(s) \to \sigma$ as $n \to \infty$. [Use Stirling's approximation.]

For $n = 5$ and $\sigma = 10$, show that $E(s) = 9.40$. Also evaluate $V(s) = E(s^2) - [E(s)]^2$. In your simulations, is the *standard deviation* of the 'SD' column close to $SD(s) = \sqrt{V(s)}$?

4. A Simulation With Exponential Data

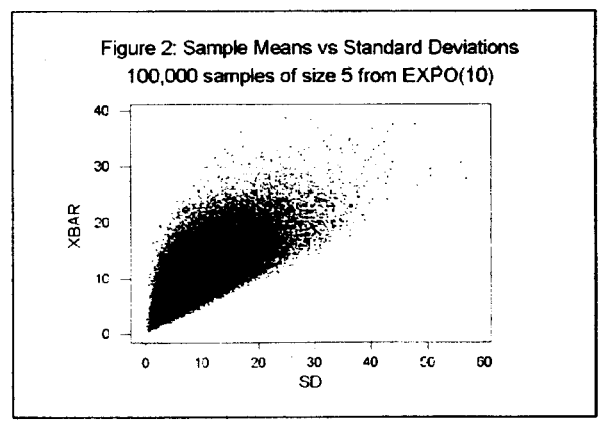
Here we will use simulation to investigate the association between the sample mean and standard deviation for random samples of size $n = 5$ from an exponential distribution with mean $\mu = 10$.

Simulation plan and Minitab program. Our simulation plan is similar to the one used in Section 3. We make two changes in the program of Display 1. Most important, we change the subcommand used with the command `random`, substituting `exponential 10` for the subcommand used above to generate normal data. In order to obtain a more revealing scatter plot, we also increase the number of iterations to $m = 100{,}000$. (If your installation of Minitab won't handle that much data, use $m = 10{,}000$ as before.)

We said earlier that $\bar{X}$ and s are independent only for normal data. We choose the exponential distribution for our second simulation because it is strongly skewed to the right, and so some samples with extremely large observations are likely. (The exponential distribution takes only positive values, so no counterbalancing extremely small values can be observed.) Because both the sample mean and the standard deviation are inflated when a sample has large observations, we suspect that large values of $\bar{X}$ will be associated with large values of s in a pattern that will make it obvious that $\bar{X}$ and s are not independent.

Simulation results. Again here, summary statistics of the column 'XBAR' show that the data are being generated as intended. For X from an exponential distribution, $E(X) = SD(X) = \mu$. Three runs of our program yield $E(\bar{X}) \approx 9.9811$, 10.013, 9.9934 (consistent with $\mu = 10$), and $SD(\bar{X}) \approx 4.4545$, 4.466, 4.4782 (consistent with $\sigma/\sqrt{n} = 10/\sqrt{5} = 4.472$). Furthermore, $\rho \approx 0.774$, 0.775, 0.775, which leaves no doubt that $\bar{X}$ and s are correlated random variables, and hence not independent.

Figure 2, the scatter plot from one of our simulation runs, illustrates this strong positive association between $\bar{X}$ and s. It also shows a specific pattern of



association that we did not expect the first time we saw it. While the points seem to scatter upward without constraint, it looks as if there may be a linear boundary below. This is not a fluke of one particular simulation run. Repeated runs show the same pattern. How does it arise?

With a little reflection it is clear that for data from any distribution restricted to positive values, the sample mean must exceed its standard error — in symbols, $\bar{X} > s/\sqrt{n}$. This follows directly from the inequality $\Sigma X_i^2 < (\Sigma X_i)^2$. With $n = 5$, the two sides approach equality when four values are very near 0 and one is relatively large. The exponential distribution yields this pattern often enough to suggest where the boundary lies. (On a scatter plot, the location of this boundary is easier to see for smaller n and harder to see for larger n).

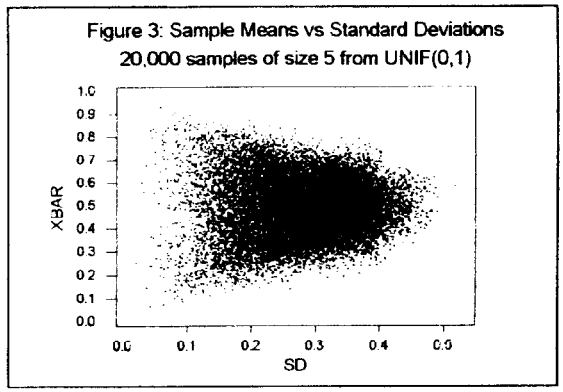
Things to try on your own. Here are some ways to explore further the linear boundary seen in Figure 2, and a suggestion to look at two sample statistics that *are* independent for exponential data.

(a) First, convince yourself that $\Sigma X_i^2 < (\Sigma X_i)^2$ when all $X_i > 0$. [Hint: What terms in the square on the right are missing from the sum of squares on the left?] Then prove that $\bar{X} > s/\sqrt{n}$. [Start with the "computational formula" for the variance s^2.] Give a numerical example to show that both inequalities could fail if negative X_i were allowed. Finally, for 5 observations from the exponential distribution with mean 10, say which of the following are positive: $P\{\bar{X} < 5\}$, $P\{s > 20\}$, $P\{\bar{X} < 5, s > 20\}$; and comment.

(b) Do several simulation runs with $m = 20{,}000$ samples of size $n = 5$ from the chi-squared distribution with 1 degree of freedom. (Use the subcommand `chisq 1`.) This nonnegative, right-skewed distribution has a heavy concentration of values very near 0. Thus the simulation should yield a scatter plot that shows the linear lower boundary quite distinctly.

(c) For a sample of size $n = 5$ from the exponential distribution with mean 10, let Y_1 be the minimum value and Y_5 be the maximum, so that $R = Y_5 - Y_1$ is the range. The "no-memory" property of the exponential distribution implies that Y_1 and R are independent. Plan, program, and carry out a simulation to illustrate this independence. (Use Minitab's row-arithmetic commands `rmax` and `rmin` and compute the range by subtracting.) Validity checks: $E(Y_1) = SD(Y_1) = 2$, $E(Y_5) = 22.833$, and $SD(Y_5) = 12.098$.

More advanced: Derive the mean and standard deviation of the maximum and minimum. It is easy to show that Y_1 is distributed exponentially with $E(Y_1) = SD(Y_1) = \mu/5 = 2$. [Find the cumulative distribution function of Y_1.] The maximum Y_5 is not exponentially distributed, but it is possible to show that $E(Y_5) = \mu(1/5 + 1/4 + 1/3 + 1/2 + 1)$ and that $V(Y_5) = \mu^2(1/25 + 1/16 + 1/9 + 1/4 + 1)$. [Look at independent increments between the order statistics $Y_1 < Y_2 < Y_3 < Y_4 < Y_5$.]



5. A Simulation With Uniform Data

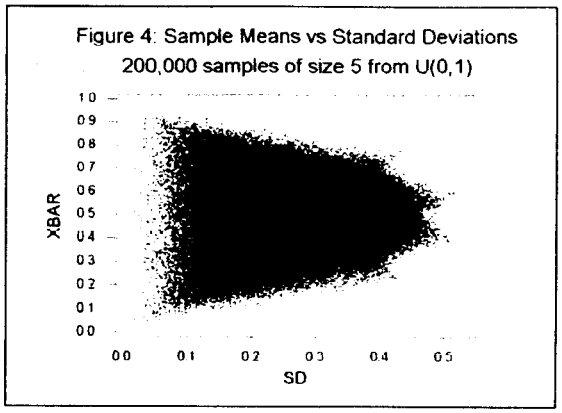
Because of its symmetry, it seems reasonable that the uniform distribution will yield samples for which the mean and standard deviation are uncorrelated random variables. Because of the theoretical result that $\bar{X}$ and s are independent only for normal data, we wondered whether simulation methods would reveal a noticeable pattern of association for uniform data in spite of the lack of correlation.

Simulation plan and Minitab program. We choose to sample from the uniform distribution on the interval (0, 1). The simulation for uniform data is similar to those in previous sections. In Minitab, the relevant distributional subcommand for generating random data is `uniform 0 1`. In our first simulation the number of iterations is $m = 20{,}000$.

Simulation results. This uniform distribution has $\mu = 1/2$ and $\sigma^2 = 1/12$. As expected, means of 'XBAR' in several simulations were close to 1/2, standard deviations of 'XBAR' were close to $1/\sqrt{60} = 0.129$, and correlations of 'XBAR' and 'SD' were close to 0.

Figure 3 does indeed show a pattern of association. As the standard deviation increases, the mean is restricted to the vicinity of 1/2. But there are some curious "lumps" and straggling clusters of points at the right side of the scatter plot. One of the challenges in dealing with simulations is to separate the signal from the noise. If we see something surprising, is it a random anomaly of one run or something worth looking into? In several more runs of the simulation we saw similar irregularities, always in about the same places. What is going on here?

We increased the number of iterations to $m = 200{,}000$ and tried again. One of the scatter plots from the larger simulation is shown in Figure 4. Here the random-looking straggles and lumps of Figure 3 have been resolved into clearly-defined "horns" that cannot be dismissed as accidental. The pattern of association between $\bar{X}$ and s is more intricate than we first supposed.



The explanation of the horns in Figure 4 is that they are images under transformation of the corners of the 5-dimensional hypercube in which 5 independent observations from the unit interval lie.

Things to try on your own

(a) A 5-dimensional hypercube has 32 corners. They correspond to data consisting entirely of 0s or 1s. One of the two horns at the far right in Figure 4 corresponds to samples that have three observations close to 1 and two close to 0, so that $\bar{X}$ is near 0.6 and s is near 0.5477. Such data points are found near 10 of the 32 corners of the hypercube. (Why 10?) Some combinations of 0s and 1s match fewer corners (have smaller "multiplicities") and so make less-distinct horns. Including the two at the top and bottom corners on the left side of the scatter plot, there are 6 horns in the scatter plot. Can you find the multiplicity of each, accounting for all 32 corners?

(b) Try simulations using the uniform distribution with m = 20,000 and n = 2, 3, and 4. How many horns do you see (on the right) in each?

(c) A beta distribution with parameters $\alpha = \beta = 1/2$ has relatively more values near 0 and 1 than does the uniform distribution on (0, 1). Try a simulation with m = 10,000 and n = 5 (subcommand `beta .5 .5`). Compare with Figure 3.

(d) Plan, program, and execute a simulation to explore whether the range and the midrange (average of the maximum and minimum) are correlated for uniform data. Are they independent? Use m = 20,000 and n = 5.

6. Nuts and Bolts of Simulation

This section gives some information to assist you in doing simulations of the kind shown above. The speed and memory capacity of computers in general use is increasing rapidly. A few years ago simulations of the scope suggested here would have put unrealistic demands on the computers available to most undergraduate students, and a few years from now they will probably seem tiny. We used Minitab 13 and S-Plus 2000 (professional) on a 700 MHz computer with 128 MB of RAM.[3] (Student versions of Minitab are of limited use for simulation because of their restricted worksheet capacity.) For your simulations, first try to use the values of m suggested here, adjusting downward or upward as necessary or desirable.

Minitab. It is not necessary to retype a sequence of commands each time it is used. You can highlight commands in the Session window, and cut (CTRL-C) and paste (CTRL-V) them to the *active* `MTB >` prompt at the *end* of the Session Window. You can also cut and paste commands recorded in the History window or composed in a text editor. (When you paste, the copied *prompts* will disappear temporarily.) If desired, you can modify pasted commands before pressing `ENTER`. When you are ready to run the simulation, move the cursor to the very end of the block of pasted/ edited text and press `ENTER`.

Display 3: S-Plus Code For the Simulation of Section 1

```
m <- 10000          # iterations (samples)
n <- 5              # sample size
k <- m*n            # total random data values
X <- matrix(rnorm(k, 100, 10), m, n)
 # simulated normal data into an m by n matrix
XBAR <- apply(X, 1, mean)
VAR <- apply(X, 1, var)
 # arg. 1 indicates a matrix row operation.
SD <- sqrt(VAR)
summary(XBAR)
sqrt(var(XBAR))
 # some versions accept: colStdevs(XBAR)
summary(SD)
cor(XBAR, SD) # correlation
plot(SD, XBAR, pch=".")
 # dot instead of default circle in plots
```

A more sophisticated approach would be to write a Minitab macro.[3] For example, after a test run, cut and paste the commands of Display 1 into Notepad; add two lines to the beginning: `gmacro` and `template`; one line to the end: `endmacro`; and save as `a:\normsim.mac` (no .txt extension). Then you can run the macro from a `MTB >` prompt by typing `%a:\normsim.mac`. (If you are not familiar with macros, you may need to read Minitab, or even Windows, documentation.)

S-Plus. S-Plus is a widely used statistical software package. If you are not yet familiar with it, the simple simulations of this article are good first projects for becoming acquainted.[3, 4] For us, some of the simulations ran slowly in S-Plus, the one shown in Figure 4 at "coffee-break" speed.

Names of functions and objects are case sensitive in S-Plus: `SD` and `sd` would be different variables, `sqrt` is a built-in function and `Sqrt` is not. The `summary` function gives the "5-number summary" plus the mean, but not the standard deviation. If needed, request variances or standard deviations separately.

Display 3 shows an S-Plus program for the simulation in Display 1. This program should be typed into a new Script window for the first simulation, and later saved or modified as desired. (Comments, indicated by #-symbols, need not be typed.) Press F10 to run the program.

For suggested modifications of this program, the following functions simulate the obvious distributions: `rexp()`,[4] `rchisq()`, `runif()`, `rbeta()`. As shown in Display 3 for `rnorm()`, the first argument in parentheses is the number of observations to be generated and subsequent arguments, separated by commas, provide the parameters. (In S-Plus, the parameter of the exponential distribution is its *rate* $1/\mu$.) Functions such as `max` and `min` work with `apply` as shown in Display 3 for `mean` and `var`.

Reference Notes and Acknowledgments

1. For an introduction to Gibbs sampling, see our earlier article, "Elementary Uses of the Gibbs Sampler: Applications to Medical Screening Tests," *STATS* #27 (Winter 2000).

2. Proofs are available in standard mathematical statistics texts, including R. V. Hogg and A. T. Craig: *Introduction to Mathematical Statistics,* 5th ed., Prentice Hall (1995), page 214; and A. M. Mood, F. A. Graybill, and D. C. Boes: *Introduction to the Theory of Statistics,* 3rd ed., McGraw Hill (1963), page 243. Also, the suggested activity (b) in Section 3 is similar to problem VI.17 in Mood, et al. (1963).

3. Annotated Minitab macros and S-Plus code, along with other instructional material related to this article, are available at: *www.telecom.csuhayward.edu/~stat/Simul*.

4. The free software R can be obtained from *www.r-project.org*. R accepts the same code we used here for S-Plus. (Exception: in R, the *mean* μ is the parameter of the exponential distribution, in S-Plus, the *rate* $1/\mu$.)

Acknowledgments. Parts of this article were included in Bruce E. Trumbo, Eric A. Suess, and Christopher M. Fraser: "Contemporary simulation methodology for undergraduates," *ASA 2000 Proceedings of the Section on Statistics Education*, American Statistical Association (2001), pp. 241–245. Thanks to John Sayer for useful conversations on section 5.