

Statistical Analysis For The Masses

Sam L. Savage

Several years ago I came to the conclusion that many graduates of basic Statistics courses don't understand the Central Limit Theorem.

I was wrong.

They don't even understand the concept of a probability distribution!

I will present evidence for this bad news, but also suggest that good news may be on the way.

I. The Bad News

A. The Sample Population

My evidence for bad news is based on experience with those who have attended my seminar on "Management Science in Spreadsheets" which has been offered throughout the United States since 1990. The brochure states that:

"This introductory seminar is designed for managers, accountants, engineers, and other professionals involved in forecasting, profit maximization, cost minimization, evaluation of risk, and decision-making under uncertainty. Formal training in management science or statistics is not required."

Thus this group is self selected to have a practical need for dealing with uncertainty. Few have had extensive training in probability or statistics, but by a show of hands, virtually all have had at least an introductory course in Statistics. Typical class size is between 15 and 40 with a total of over 1000 attendees.

During my lectures I use a lap-top computer and projection system to present live models in spreadsheets such as 1-2-3 or Excel. In this environment I pose questions to the audience concerning chance phenomena which, after some discussion, I answer experimentally on the spot using Monte Carlo simulation. This process has revealed some surprising misconceptions concerning uncertainty.

B. The Experiments

I have repeatedly performed three experiments within this group, at first informally, then with somewhat more care, with consistent results.

Problem 1 - Averages of Random Variables

I begin my seminar with a discussion of random variables.

Random Variables

I define a random variable as “a number you don’t know yet”. As a simple example, I introduce the spinner shown in figure 1, and point out that the result of a spin is modeled by the RAND function found in all spreadsheets. To provide motivation, I ask the participants to imagine that God will determine the profit of their company next year (in Millions) by giving such a spinner a good whack. To provide even more motivation, I tell them that if profit is less than \$200,000 they’re fired.

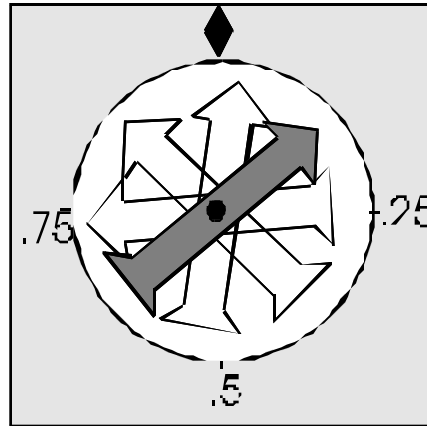
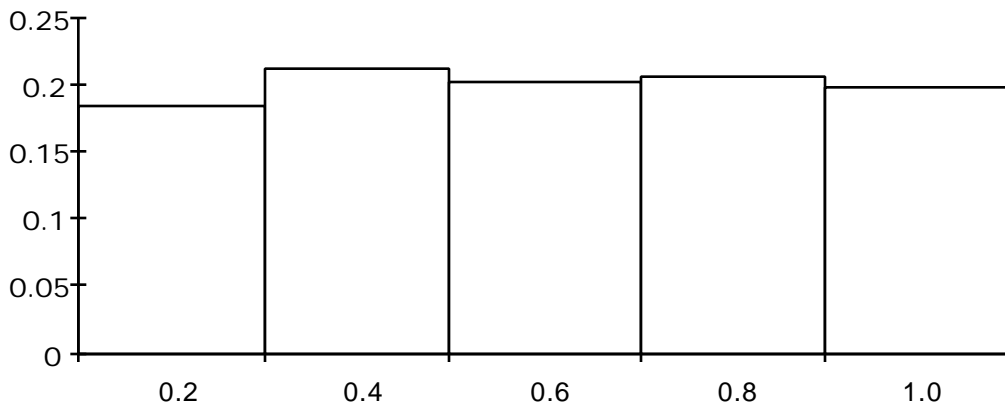


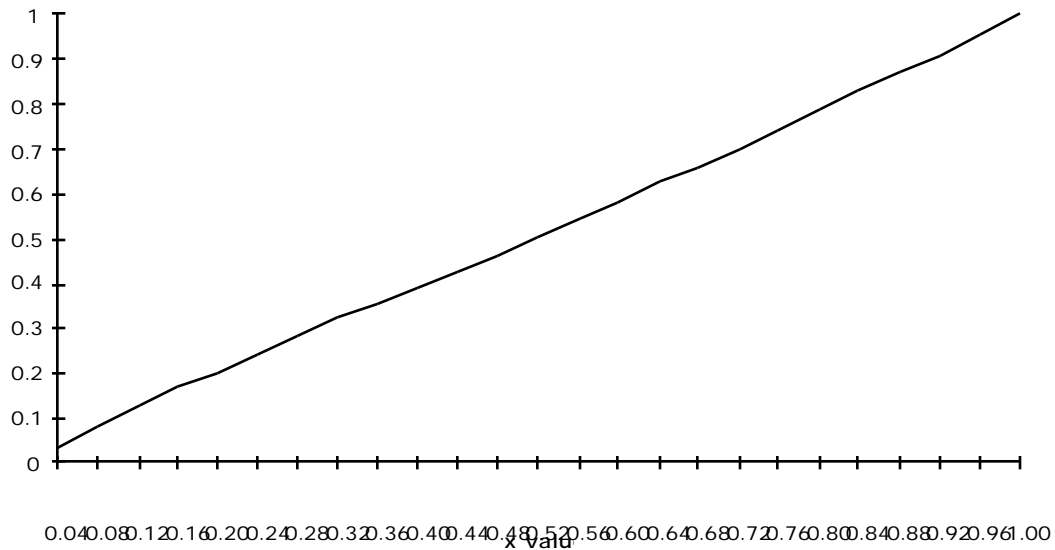
figure 1

Simulation of a Spin

- I simulate profit in Millions, by placing the RAND function in a blank spreadsheet and running a Monte Carlo Simulation of several hundred trials.
- Before showing the results, I ask what the average is likely to be. The predominant answer is \$500,000 (so far so good).
- I then display a 5 bin histogram of the outcomes of RAND as shown below.



- I also display the cumulative graph showing that the probability of getting fired ($x < .2$) is 20%.



For the first two years of giving these seminars, I would then perform the following experiment.

The Average of Two Spins

- I admit to the class that the above model of profit is naive, and that actually God whacks the spinner twice and averages the results.
- When asked what the average profit will be now, the consensus is still \$500,000.
- Next I ask the class to draw a 5 bin histogram of the results. Virtually everyone draws something like one of the two figures below. The proportion of each varies from group to group, but is roughly evenly divided.

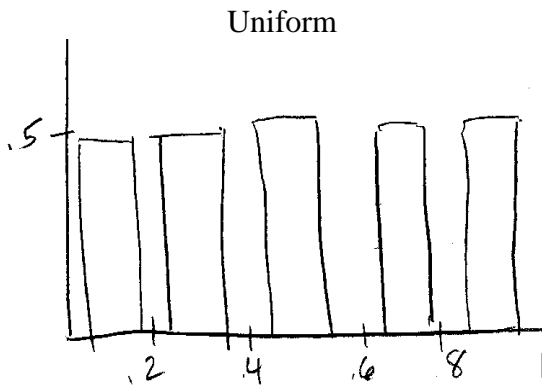


figure 2

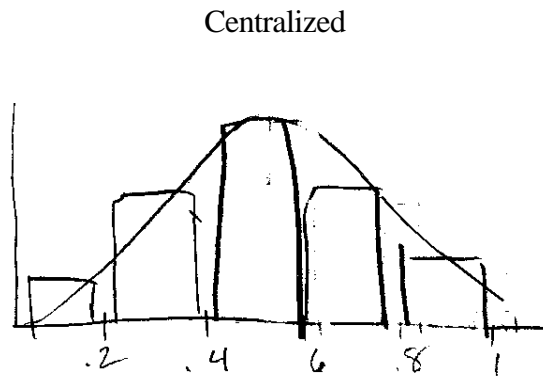
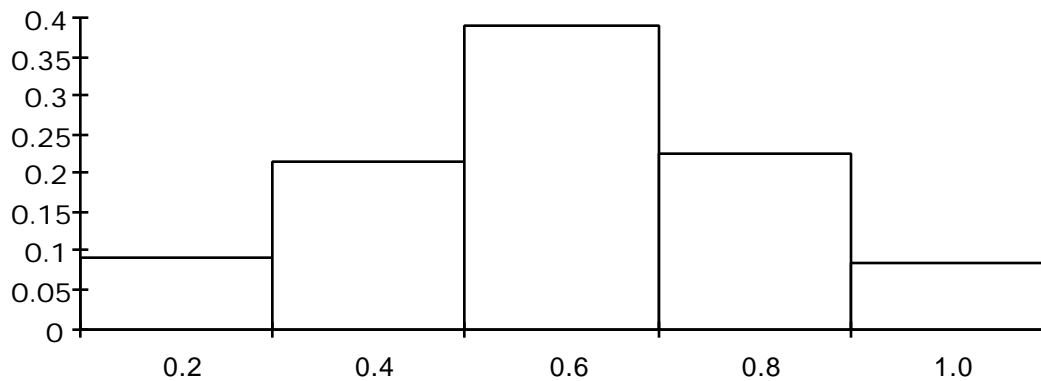


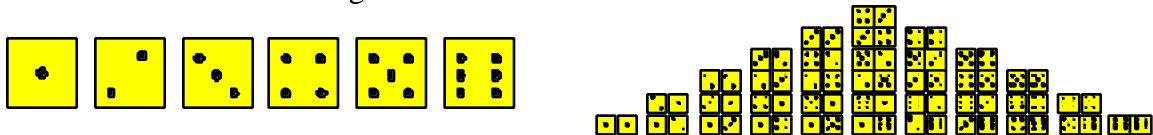
figure 3

Note that in this example probability sums to 2.5. Bins, continuous lines, or as in this case both, are sometimes drawn.

- Next I simulate $(\text{RAND}+\text{RAND})/2$ and display the resulting histogram below.



Theoretically this is a triangular distribution as can be appreciated intuitively by thinking about the outcome of rolling one and two dice.

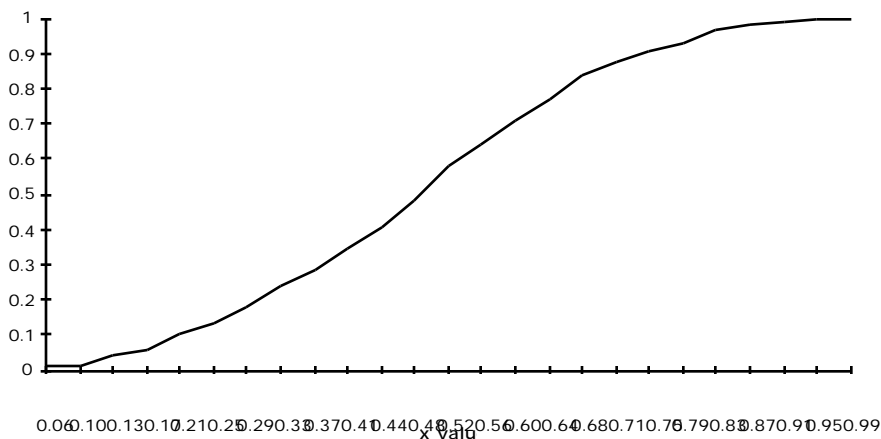


Only a handful out of 1000 subjects have drawn a clearly triangular distribution, and such people are probably better suited to being statisticians than managers. In short, **it's dumb to be too smart**. As far as I am concerned the correct practical answer to the question:

What does the distribution of the average of two spins look like?
is
It goes up in the center and down on the ends

It is far more important to be aware of the economic implications of diversification than to know the exact shape of any particular distribution.

- With this in mind I continue my demonstration by displaying the cumulative distribution which shows that the probability of getting fired is now under 10%, less than half of what it was with one spin.



- After the correct answer has been presented, I ask how many have heard of the Central Limit Theorem. By a show of hands, virtually **all** have, although apparently few have developed a mature relationship with the subject. I also ask for the term by which their stock broker refers to this concept. The word I am looking for and occasionally get is **Diversification**. I also suggest that whenever they are faced with an important “number they don’t know yet” they should try to visualize its **shape** as approximated by a histogram

The Unjustified Conclusion

For two years I concluded that since roughly half the subjects drew something like figure 3, that they understood that the distribution would centralize in going from one to two spins. The following experiment indicated otherwise.

Problem 2 - Probability Distributions

The Distribution of One Spin

In about the third year of giving these seminars, I asked the participants in one session to draw the histogram of the outcome of the **single** spin before showing them the results of the Monte Carlo experiment.

I was surprised to find the same results I had been getting for the average of two spins; that is roughly 50% each of figures 2 and 3. I went on with the lecture and as usual asked for the average of two spins, again finding similar results. Since then, I have always asked them to draw the histograms for both 1 and 2 spins with consistent results.

I Have Never Let My Education Interfere With My Learning: Mark Twain



In a classic case of the phenomenon referred to by Mark Twain, these graduates of a basic Statistics course have learned that the maximum likelihood answer to a statistics “picture” question is a “bell shaped curve”. I suspect that if the 1 spin test were administered to those with no formal training in probability or statistics, that the percentage of correct answers would go up.

The Joint Distribution

Once people have seen the correct results for a single spin their opinions have no doubt been biased concerning the result for the average of two spins. Nonetheless it is interesting to observe the joint distribution of the two responses over the two cases (1 spin, 2 spins).

Seminar Participants

The results for 43 participants at a recent seminar are shown below.

		Average of 2	
		Correct*	Incorrect
Spins 1 Spin	Correct 	11	12
	Incorrect 	9	11





The Joint Distribution for 43 Seminar Attendees

* I count as correct any centralized distribution.

Note the roughly uniform distribution over the four possibilities. They are, clockwise from the upper left, correct 1 and 2 spins, correct only for 1 spin, incorrect for both, and correct only for 2 spins.

A More Sophisticated Population

Admittedly, the backgrounds of those attending seminars vary widely, and most have been out of school for decades. Therefore I was particularly surprised when I tried this with 8 masters level students in Operations Research at Stanford University. These students had all recently taken courses beyond basic Statistics, and some were taking graduate level Statistics concurrently. The results below show that 5 out of the 8 thought that the distribution of a single spin was non-uniform!

		Average of 2		Correct*	Incorrect
		Spins			
1 Spin	Correct 	3			
	Incorrect 	3		2	

The Joint Distribution for 8 Masters Students in Operations Research

Problem 3 - Functions of a Random Variable

Although the concepts of the central limit theorem and the diversification of risk are important, there is an even more important lesson that seems to have slipped by most graduates of a statistics course. It concerns functions of a random variable.

A Stochastic Cost Model

A supply firm carries a piece of equipment for which demand is random with a monthly average of 5 units. The firm stocks 5 units of inventory to satisfy this demand.

The cost of maintaining the inventory has two components:

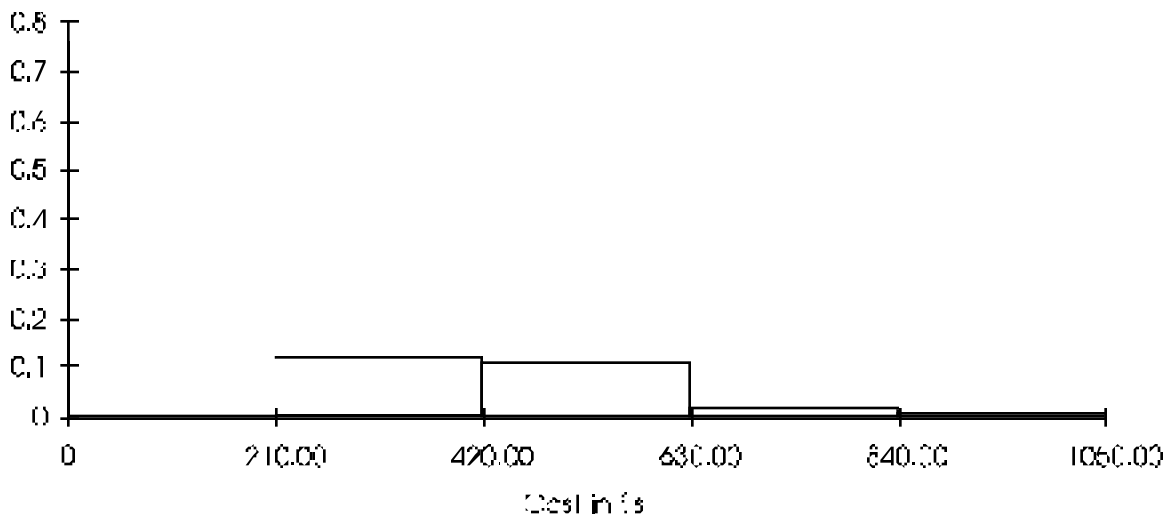
- a. If the demand is less than the number stocked, a \$50 storage cost is incurred for every unit stocked in excess of demand.
- b. If the demand is greater than the number stocked, an air freight cost of \$150 per unit is incurred for the shortfall.

This situation is reflected in the spreadsheet model shown below (the formulas are shown in cells C5,C6 and C8).

	A	B	C	D	E	F
1	Demand		Amt Stocked			
2		5		5		
3						
4	Costs		Per Unit	Total		
5	Storage Cost		\$50.00	IF(Amt_Stocked>Demand,+(Amt_Stocked-Demand)*S_COST,0)		
6	Air Freight		\$150.00	IF(Amt_Stocked<Demand,+(Demand-Amt_Stocked)*F_COST,0)		
7						
8	Overall Cost		C5+C6			

The Average Overall Cost

- I plug the *average* demand of 5 into cell A2, and neither storage nor freight costs are incurred, yielding an overall cost in C8 of zero. Therefore I say: “the *average* overall cost is zero, right?”
At this point many eyes have glazed over, but those who understand the question generally nod their heads approvingly. Rarely does someone object or give any indication that this is in error.
- Simulation of overall cost quickly shows what’s going on. I replace demand (A2) by a Poisson random variable with mean of 5, and run 50 or so iterations, yielding an average cost of around \$200, not 0. Of course, I explain, since this experiment was run with random numbers, one would not expect exactly 0 even if this were the theoretical result.
- By inspecting the histogram of outcomes, it is clear that since there are no negative costs to cancel out the positive ones that a mean of 0 is out of the question.



Three Perspectives of What Went Wrong

I consider this to be such an important lesson that I repeat what went wrong in three ways, one of which hopefully hits home.

1. Average Inputs (demand) Don't Necessarily Yield Average Outputs (cost)

I explain that if the model happens to be linear, then average inputs **do** yield average outputs. But when I ask by a show of hands, how many know what linear means, only a small percent respond. Therefore it is safer to assume that you **don't** get average outputs.

2. $E(F(X)) \neq F(E(X))$ unless F is linear

I explain that this is the way it is expressed (and for the most part left to languish) in text books.

3. A Sobering Example

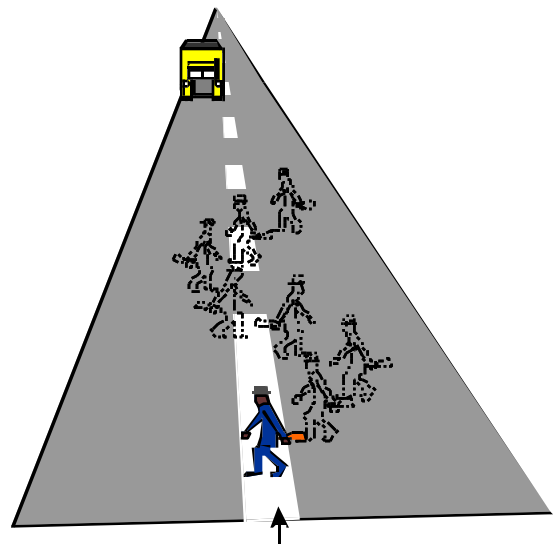
Finally I impress on the class, that this can have serious consequences.

Consider a drunk wandering back and forth across a busy highway whose position is denoted by X . Assume that his average position, $E(X)$, is on the center line, and let $F(X)$ represent the state of the drunk. Then

$$F(E(X)) = \text{ALIVE}$$

whereas

$$E(F(X)) = \text{DEAD}$$



Average Position of Drunk, $E(X)$
is Center Line

C. The Tip of The Iceberg

And what have others learned about errors in judgment under uncertainty? I recommend *Judgment under uncertainty: Heuristics and biases* edited by Kahneman, Slovic, and Tversky (1982). This contains a wide variety of well documented and consistent errors in judgment, even among “experts” who have had substantial statistical training.

If the Statistics courses are not providing basic statistical intuition, it is not necessarily because they failed to cover the material, but perhaps just the opposite. That is, by covering so **many** topics, they may yield a poor signal to noise ratio when it comes to the basics.

And if the graduates of Statistics courses don't know what a probability distribution is, how about the tens of millions of users of electronic spreadsheets who have had no Statistics whatever. After all, 100% of them have numbers in their spreadsheet models about which they are uncertain. One can only guess at the economic impact of tens of millions of average values being plugged into potentially nonlinear models upon which business and government decisions are being based.

II. The Good News

A. A Goal

In the preface to *Judgment under uncertainty: Heuristics and biases* it states that two earlier investigators, J. Bruner and H. Simon, were “concerned with strategies of simplification that reduced the complexity of judgment tasks, to make them tractable for the kind of mind that people happen to have.”

I believe a similarly worthy goal is to “design strategies of simplification that reduce the complexity of judgment tasks, to make them tractable for the kind of mind *and computer* that people happen to have.” And the good news is that the kind of computer people have today is the size of a book and contains more raw computing power than the entire United States at the time Bruner and Simon initiated their work in the mid 1950’s.

B. Consumer Stochastics

I refer to the fulfillment of this goal as Consumer Stochastics (part of a larger trend that I call Consumer Analytics). This differs from the overall field of statistics, only in its target audience. Whereas statistics has aided scientists and engineers in almost every discipline since long before electronic computing, Consumer Stochastics is aimed at non-scientists and non-engineers, who face uncertainty on a regular basis, that is, most professionals.

This will not be the first time that a technology has moved from a highly trained audience to a more general one.

Picking up Groceries with an Internal Combustion Engine

The first internal combustion engines, for example, were employed solely by “engineers” for turning drive belts in factories and other industrial settings. No one would have thought of using one of these noisy, dangerous and tremendously heavy pieces of equipment for picking up groceries. Today of course, the internal combustion engine is the method of choice for picking up groceries even among non-engineers, thanks to power steering, automatic transmissions, and oil pressure warning lights.

Similarly, the benefits of statistical analysis are now spreading beyond the realm of the statistically trained. Come to think of it, the oil pressure warning light embodies a statistical test for predicting engine failure with clear penalties for both type 1 and type 2 errors.

C. Examples

Things I would classify as consumer stochastics fall into three categories which I will call Sublimated Statistics, Institutionalized Statistics, and General Tools for Thought.

1. Sublimated Statistics

Today we are surrounded by mechanical and electronic devices like the warning light above, which continually perform statistical tests. Some of these are quite complex. For example, headphones which generate anti-noise, and video cameras which electronically correct images for vibrations. The key to Sublimated Statistics is that if it performs correctly we don’t know it’s there.

2. Institutionalized Statistics

There are also narrow contexts in which traditional statistical measures have been adopted by non-statisticians. For example, in the 1950’s, Harry Markowitz (1959) pioneered the use of variance as a risk measure for stock portfolios. Derivations of this approach, which assumes a multivariate normal distribution of returns, is alive and well today in the finance community. In fact, in an attempt to establish industry benchmarks (and perhaps preempt additional banking regulation), J.P. Morgan & Co., now publishes its in house risk measurement techniques which are based on the Markowitz approach, see the Wall Street Journal (Oct, 11, 1994). This includes daily estimates of the covariance of numerous market factors.

Although an institutionalized measure, as in this case, can have a tremendous social impact, it may nonetheless be difficult to generalize. For example, when the underlying distributions are highly asymmetric, as with certain derivative financial instruments, the Markowitz risk measure is no longer appropriate. This may partially explain the reluctance of some investors to use derivatives: it is not that they are necessarily riskier, but merely that a method for measuring their risk has not yet been institutionalized.

Risk Detectors

Given the increased use of derivative investment instruments, and a series of recent fiascoes such as Orange County and Barings Bank, the financial community is currently aware that it needs good standardized indicators of risk beyond the original Markowitz model. A measure known as Value at Risk (VAR) has now been proposed which is applicable in situations with asymmetric risk, Beckstrom, Lewis, Don, and Roberts (1994). In addition, the SEC, for example, is trying to devise “Yardsticks” for mutual fund risk, the Wall Street Journal, (Mar, 29, 1995). I call such institutionalized constructs ***risk detectors***, and believe they will become increasingly important in the area of finance.

3. General Tools for Thought

Perhaps the most remarkable aspect of consumer stochastics is the recent proliferation of general analytical tools to tackle problems of uncertainty. Many of these are available as enhancements to electronic spreadsheets.

Monte Carlo Simulation: Killing Cockroaches With Steamrollers

The most widespread among these is the brute force application of Monte Carlo Simulation. Although suggested as an analytical tool for business as early as thirty years ago, Hertz (1979), one often heard the sentiment that this technique was akin to killing cockroaches with a steamroller. I like that analogy, because had steamrollers kept technological pace with computers, they would cost pennies today, fit in your wallet, travel hundreds of miles per hour, and be the weapon of choice against the cockroach.

Currently there are several widely available commercial packages which bring Monte Carlo capability to spreadsheets: @RISK, Crystal Ball, Savage (1993) and Savage (1994). If these products do nothing else, they provide first hand intuition into the concept of probability distributions, which may put them slightly ahead of most Statistics courses.

Ham vs. Spam

But the users of Monte Carlo simulation won't know what distributions to use, you say. I used to feel that way too until I fully comprehended the gravity of approximating random values by their averages (remember what happened to the drunk in the highway). I am now convinced that modeling every distribution in the world as triangular, specified by a minimum, maximum, and most likely value, would be a significant improvement over the status quo.

But we can often easily do much better than the triangular distribution. I take the simplistic view (and it better be simplistic if the masses are going to use it) that every random process has its own distribution which you can get a picture of by executing the histogram function on the historical data in your spreadsheet. The theory of probability has isolated certain idealized classes of these which have been given names (Normal, Binomial, Poisson etc.). But the relationship between these idealized distributions and the historical data is analogous to that of Spam and the ham from which it was derived: something is usually lost in the process.

Simulation can be run by directly sampling the historical data itself (the ham), as championed as early as 1969 by economist Julian Simon, see Simon, Atkinson, and Shevokas, (1976). This resampling process is simple, direct, and doesn't require a statistician as an intermediary to estimate distributions. This technique is the basis for the bootstrap method pioneered by Bradley Efron, Efron and Tibshirani (1993) which is now popular even among some professional statisticians. Furthermore, multivariate data can also be used in this manner without losing the statistical dependence between variables. As both the power of computers and the quantity of digitized information continue to grow, this approach can only become more relevant.

There is a catch to this approach, however. It is only accurate if the future of the random process being modeled behaves like its past, that is, the process is *stationary*. This problem will be addressed below.

Time Series Analysis: Forecasting the Future

Time series analysis is comprised of statistical techniques to forecast series of random variables such as sales, prices or interest rates which are measured at regular time intervals. Traditionally these routines have been contained in serious statistical analysis software packages like SAS and SPSS, but these hardly qualify as consumer stochastics. Recently, with the aid of expert systems, time series analysis has been automated in moderately priced commercial software packages, see for example Forecast Pro. Automated forecasting packages are in use in a wide variety of industries.

As an example, past monthly sales for some piece of equipment might be fed into a forecasting package, whereupon it yields the expected demand, 5 for example, for the next month. Then typically this number is plugged into a spreadsheet model to predict profit or cost wait a minute. How many times do I have to go through this? Remember, plugging in averages is foolhardy.

The “Non-Stationarity” Warning Light

Instead of just spitting out an expected or average demand, the forecasting software can also provide an estimate of the distribution of demand. This *can* often be sensibly plugged into a Monte Carlo simulation such as the spreadsheet inventory model of section I. Perhaps this process of channeling the output of time series analysis directly into Monte Carlo simulation can be automated to include something like an oil pressure light that warns of non-stationarity.

Decision Trees: Forcing us to Assess Probabilities

There are also currently several commercial packages for formulating decision trees either on their own or coupled to spreadsheets: Data, DPL, and Savage (1993). Of course decision trees don't yield correct decisions unless we input correct probabilities for uncertain events, so in most cases what good are they? What they are good for is helping us focus on those issues most relevant to our decision making. And when these issues involve probabilities that are hard to estimate, the software provides graphical sensitivity analysis showing managers the consequences of various levels of uncertainty. All this leads towards improved intuition in decision making under uncertainty, and away from what Efron refers to in jest, the “time honored tradition” of using averages.

But That's Not All

Other consumer stochastics tools are also quickly falling into the hands of the masses.

Regression

Linear regression is included with all spreadsheets. In fact, a recent seminar attendee from the legal analysis department of a large accounting firm told me that some courts now require a regression to be run of total cost against quantity for settling disputes involving fixed and variable costs.

Markov Chains

Markov chains Savage (1993) and (1994) are a compelling paradigm in the spreadsheet, in which it is easy to graph the populations in various states over time.

Stochastic Optimization

Stochastic optimization, considered computationally prohibitive only a few years ago, is now becoming popular. Small problems may be solved using built-in spreadsheet solvers, or in auxiliary software coupled to spreadsheets, for Example *What'sBest!*.

The Lunatic Fringe

In less chartered waters, products based on fuzzy logic and neural networks are also gaining acceptance. Seminar attendees from a major bank are using neural nets to predict credit-worthiness in loan applications, for example.

D. Where Do Concepts Leave Off And Tools Begin?

Some academicians aware of my approach concede that the spreadsheet is a powerful tool, but claim to prefer teaching concepts instead of tools. My response is that one can teach the concept of transferring momentum from a large mass swung on a lever arm to thin iron spikes, forcing them through pieces of wood which are thereby fastened together. I myself prefer a hammer and nails.

Can some of the concepts of statistics be similarly codified into tools as successfully as concepts of physics have been codified into the carpenters tool box? What would these tools look like? These questions will ultimately be settled through a process of both intellectual and market evolution which has just begun. But one thing is for sure: they will not be widely adopted unless they are extremely easy to use.

The Keystroke Metric






I measure the complexity of an analytical technique in terms of the number of keystrokes required to apply it. I pick the electronic spreadsheet as my origin (I don't count the number of keystrokes required for using the spreadsheet itself). For example, to run the Monte Carlo simulation of the inventory model of section one required 6 keystrokes as follows:

- 1) Paste Poisson(Mean) function into Demand cell
- 2) Replace Mean with 5
- 3) Click Simulate Run command
- 4) Type in number of iterations
- 5) Click on output cell (total cost)
- 6) Click OK

Icons of Uncertainty

To get the ball rolling I have listed several potential tools for people with spreadsheet models (worksheets) containing formulas and numbers, some of which they don't know yet. This represents a market of tens of millions. I will assume further that the users have access to historical data on their uncertain numbers (down to a market of merely millions). These tools would augment the current spreadsheet commands and function sets, and be accessed by clicking icons. Undoubtedly, once these were tried in practice they would

prove to be flawed, but my hope is that they would ultimately evolve into a set of tools which improved the users ability to deal with uncertainty.

Icon	Command or Function	Action	Syntax
	<i>Histogram</i>	Display histogram of selected data along with mean and standard deviation. Warning message for obvious non-stationarity	Select data. Click icon or perform menu command. Specify number of bins in dialog box.
	<i>Input</i>	Specify data to be resampled for a given cell in the worksheet. This results in the following formula being placed in the cell: =INPUT(data_range) where data_range is the range containing the data to be resampled during simulation.	Select cell containing uncertain number. Click on icon or perform menu command. Specify range containing data to be resampled.
	<i>Output</i>	Specify that the output distribution of a cell containing a formula depending on input cells be estimated through simulation.	Select cell. Click icon or perform menu command. Specify in dialog box, whether to add cell to or remove cell from list of output cells.
	<i>Forecast</i>	Perform simple time series forecast. Place random number formulas representing future periods in specified cells. For example, if the forecast were performed through linear regression, then the formula in the cell representing future time t would be: $b + a * t + \text{INPUT}(\mathbf{resid})$ where a and b are the slope and intercept of the regression line, and resid is the range in which the residuals of the regression were stored.	Select data. Click icon or perform menu command. Specify locations of residuals and future time periods in dialog box.
	<i>Run</i>	Run a simulation creating histograms (including cumulative) of output cells.	Click Run icon or perform menu command.

E. Skepticism, Pessimism and Hope

Skeptics who think that none of this will work run the risk of those who thought consumers would never figure out how to use internal combustion engines, cathode ray tubes, or computers.

There may be pessimists who are afraid that it just might work, and in the process degrade the traditional fields of probability and statistics. However, as statistical measures gain further acceptance in business, law, medicine and government it should only increase the demand for high quality statistical analysis.

Also, as with nearly all technological advances, there will be those who warn that encouraging those beyond the priesthood to try things on their own will lead to costly mistakes. Of course people will make mistakes with these tools, but none so costly as depriving the masses from more convenient and intuitive ways of dealing with uncertainty.

I prefer to hope that the statistical techniques launched in the 20th century will lead to widespread benefits in the 21st century, in the way that the thermo, fluid and electro-dynamics of the 19th century led to the transportation systems, electronics, and telecommunication of today.

[1] @RISK, Palisade Corp., Newfield NY

[2] Crystal Ball, Decisioneering Inc., Boulder, CO

[3] DATA, Treeage Inc., Boston

[4] DPL, Applied Decision Analysis, Menlo Park, CA

[5] Efron, Bradley and Tibshirani, Robert J. 1993. *An Introduction to the Bootstrap*, Chapman & Hall, New York

[6] Forecast Pro, Business Forecast Systems, Belmont, MA

[7] Hertz, David B. 1979. *Risk analysis in capital investment*, Harvard Business Review, Vol 57, No. 5

[8] Kahneman, D., Slovic, P. & Tversky, A. 1982. *Judgement under Uncertainty: Heuristics and biases*, Cambridge University Press, Cambridge

[9] Markowitz, H. M. (1959) *Portfolio Selection, Efficient Diversification of Investments* John Wiley & Sons, Inc.

[10] Savage, Sam L. 1993. *Fundamental Analytic Spreadsheet Tools for Quantitative Management*, McGraw-Hill, New York

[11] Savage, Sam L. 1994. *Fundamental Analytic Spreadsheet Tools for Production Operations Management*, McGraw-Hill, New York

[12] Simon, J.L., Atkinson, D.T., and Shevokas, C. 1976, *Probability and statistics: Experimental results of a radically different teaching method*, American Mathematical Monthly, 83 (9)

[13] Beckstrom, Rod A., Lewis, Don, and Roberts Chris, 1994, *VAR: pushing risk management to the statistical limit*, Capital Market Strategies No. 3, Nov. IFR Publishing, London

[14] Wall Street Journal, Oct, 11, 1994, *Morgan Unveils the Way It Measures Market Risk*

[15] Wall Street Journal, Mar, 29, 1995, *SEC Seeks Aids in Devising Yardstick for Funds' Risk*

[16] *What's Best!*, Lindo Systems Inc., Chicago

Note: the drawing of the drunk is taken from [10] and is used by permission.