The Unseen: Statistical Inference with Limited Data

Trevor L. VanAtta
US Army Tank-automotive and Armaments Command (TACOM), Warren, MI
Operations Research Analyst & Chair of the Army Cost Risk Working Group
trevor.l.vanatta@us.army.mil
March 2012

Table of Contents

ABSTRACT

"Objective measurements of probability are often unavailable, and most significant choices under risk require an intuitive evaluation of probability." -Daniel Kahneman and Amos Tversky[1]

What are the odds of rolling a sum total of seven when tossing two dice? What is the probability of red turning up after a spin of a European roulette wheel? Most analysts, given a little time and a calculator, could answer these two questions with exact precision. For both of these questions, there is only one true correct answer. Such is the nature of probability analysis for questions that are *decompositional* (all possible outcomes can be determined), *frequentistic* (the experiment can be repeated an infinite number of times), and *algorithmic* (the results can be measured with numbers). Unfortunately, as pointed out in the quote above, not all questions involving uncertainty can be measured with precise probability, and, more often than not, we must rely on our intuition to evaluate risk. For example, there is no perfect probability measure when evaluating the odds that a specific applicant will be a successful employee if hired, or when assessing the likelihood that a witness is telling the truth during testimony.

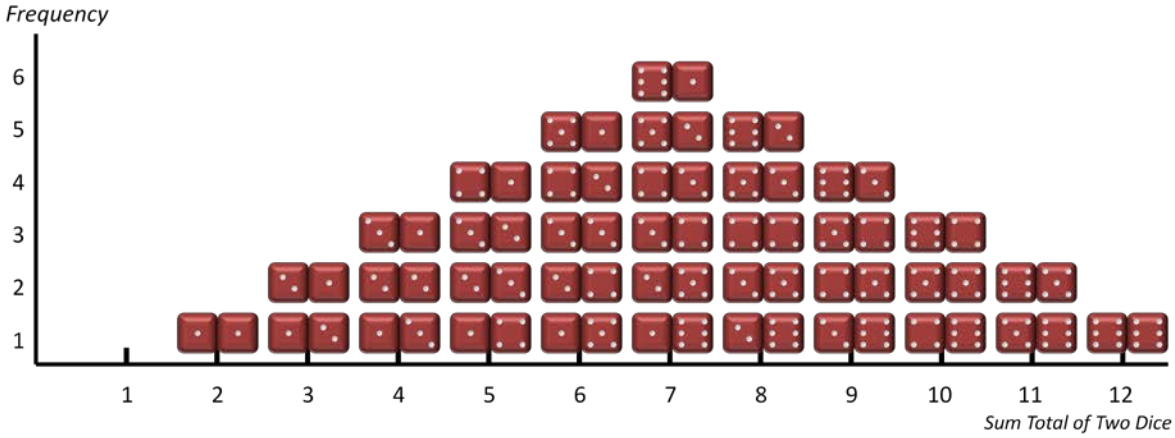
But what about cost estimates? Can we calculate the precise probability of a project overrunning one million dollars through the use of statistics? Unfortunately, no matter how much data gathering and analysis we do, we cannot place limits on the real world. We could calculate a 70% confidence interval for an estimate, but some unexpected events could occur (i.e. an earthquake, a financial collapse, etc.) that alter the odds and throw the cost of the project spinning out of control. Real world scenarios are not *decompositional* such that we can account for all possible outcomes, so an objective, perfect probability calculation of cost risk isn't possible.

So then, how do we evaluate cost risk? Must we default to our intuition and give up on statistics altogether? This question is very real in the Department of Defense, especially because we often don't have more than a few reliable data points with which to formulate our cost estimates. Nevertheless, there are practical means by which we can bridge the gap between intuition and mathematics. We can combine psychological research on intuitive judgment from recent decades with mathematics to understand how to draw distributions when the data we've gathered isn't enough to perform parametric analysis or a goodness-of-fit test. Cost risk analysis requires both mathematical and intuitive processes, so exploring and understanding the flaws and powers of our intuition is the only way to connect the real world with mathematics so that we can formulate *meaningful* statistical inferences.

The Leibniz Paradox: Intuition or Calculation?

"When we talk about analytic versus intuitive decision making, neither is good or bad. What is bad is if you use either of them in an inappropriate circumstance." – Malcolm Gladwell[2]

Imagine that you're playing a game of Monopoly with some family members, and your turn has just come up. As you survey the board, you see that your piece is approaching a cluster of an opponent's newly constructed hotels, and if you roll a 4, 6, or 7, you'll lose the game. You could roll the dice and take your chances, or you could attempt to negotiate with your opponent to somehow buy yourself one turn of immunity. In considering your options, you'd be well-advised to calculate the odds of rolling a 4, 6, or 7. As the figure below shows, there are 36 possible combinations when rolling two dice, 14 of which result in a sum total of 4, 6, or 7. This means that your odds of rolling a 4, 6, or 7 are 14 in 36, or about 38.9%. That's not good news. Consider how your odds would differ if you were a few spaces up on the board and rolling a 2, 4, or 5 would doom your chances of winning. In that scenario, there's only an 8 in 36 chance (22.2%) of the doomsday scenario occurring, and you'd be in a better bargaining position.



Calculating the odds (or probabilities) in this sort of game of chance is quite easy, and could potentially benefit any player who habitually thinks in such terms. On the other hand, imagine that you're interviewing a potential employee and trying to determine the odds that this individual will be a successful addition to your team. There are no fancy calculators or analytical tools that could tell you the exact probability of this person being a successful employee. The only way to make such a calculation is with your own subjective judgment. In doing so, you may have a plethora of numerical measurements about this individual's past, including their college GPA, length of service in previous jobs, number of felony convictions, and/or scores on specific standardized tests. These measurements are tools designed to help you formulate a *level of confidence* in the applicant, but they could not result in a perfect calculation of probability.

So why is it that in some scenarios it's easy and logical to calculate *probabilities*, but in other scenarios it would be completely absurd to waste any time attempting to do so? To answer this question, let's go all the way back to the very birth of the idea of statistical inference…

In the 17th Century, mathematicians had begun to explore the topic of *probability* in games of chance, such as the likelihood of rolling a four when tossing a die. At the dawn of the 18th Century, mathematician Jacob Bernoulli wondered if he could use these same principles in real-life scenarios, so he wrote a letter to his friend Gottfried Leibniz to ask whether it would be possible to use data gathered

from gravestones to calculate the probability of a 20 year-old male outliving a 60 year-old male. Leibniz was not so comfortable with the idea, and replied back:

> "Nature has established patterns originating in the return of events, but only for the most part. New illnesses flood the human race, so that no matter how many experiments you have done on corpses, you have not thereby imposed a limit on the nature of events so that in the future they could not vary."[3]

Leibniz was correct, and he brings up a painful reality. While we can roll a 6-sided die on a table and know that the result will be somewhere between 1 and 6, we cannot use the past to predict the probabilities of the future with absolute perfection when examining the real world. Nevertheless, Bernoulli was wise enough to know that such data could still be informative when calculating things such as life insurance rates. Sure, data from gravestones can't tell you what will happen in the future with perfect *probability*, but it can be used to understand what has happened in the past so that you can make predictions about the future with reasonable *confidence*.

That brings us to two very important terms in the world of statistics: *probability* and *confidence*. ***Probabilities*** are perfectly accurate measurements of the odds that we face in a given scenario. They can only be calculated in scenarios that are *decompositional* (all possible outcomes can be determined), *frequentistic* (the experiment can be repeated an infinite number of times), and *algorithmic* (the results can be measured with numbers)[1]. Consider the two dice. You can easily *decompose* every possible scenario into 36 combinations, repeat the rolling of the dice an infinite number of times, and record the results with numbers (2-12). On the other hand, ***confidence*** is the *perceived probability* of a given scenario occurring. When sitting across the table from a job applicant, you cannot possibly decompose all of the future scenarios of the individual's potential career, nor could you run through that career an infinite number of times, and clearly job performance cannot be rolled into a single unit of numerical measurement. That's why you express the odds in terms of ***confidence*** rather than ***probability***.

Now, consider the contrast between forecasting future costs and rolling two dice. When forecasting costs, the range of possibilities stretches outward into infinity, so a complete decomposition of all future possibilities is effectively impossible. On the other hand, as shown earlier, all of the possible combinations of two six-sided dice can be recorded into 36 possible permutations. In addition, the experiment with dice can be performed an infinite number of times, whereas the cost of a particular project or program will occur only once. While we can apply a statistical distribution in both scenarios, we can only calculate a *probability* in the scenario with two dice. When predicting future costs, statistics cannot be used to calculate probabilities, but rather can only be used to formulate *statistical inference* in the form of *confidence levels*.

Scenario	Decompositional?	Frequentistic?	Algorithmic?
The Sum Total of Two Dice	Yes	Yes	Yes
Forecasting Future Costs	No	No	Yes
Evaluating Job Applicant	No	No	No

The difference here may seem entirely semantic, but it has widespread implications in cost risk analysis. The point is this: *No matter how much cost data you have or how much analysis you do, the formulation of a statistical distribution surrounding a cost estimate is only a hypothesis. Regardless of the width of the distribution, costs could still occur outside of the projected range.* For a particular program, cost forecasts can be expressed in a narrow range (i.e. $4-5 million) or a wide range (i.e. $2-7 million). The narrower forecast is more informative to decision makers if it turns out to be true. The wider forecast,

however, is more likely to be true, but is potentially less informative. As Nobel Laureate Daniel Kahneman put it:

> "A good forecast is a compromise between a point estimate, which is sure to be wrong, and 99.9% confidence interval, which is often too broad. The selection of hypotheses in science is subject to the same trade-off. A hypothesis must risk refutation to be valuable, but its value declines if refutation is nearly certain. *Good hypotheses balance informativeness against probable truth.*" (Emphasis added)[1]

Effectively, this means that when formulating statistical distributions in cost estimating, we are *always* (regardless of the amount of available data and analysis) infusing our intuitive judgment into our projected ranges.

The Errors of Dealing with Uncertainty without Statistics

"The Commanding General is well aware that the forecasts are no good. However, he needs them for planning purposes." – Reply to Noble Laureate Kenneth Arrow during WWII after he requested to be relieved of the responsibility to forecast weather into the distant future because the uncertainty was so great as to make the forecasts no more than a guess.[3]

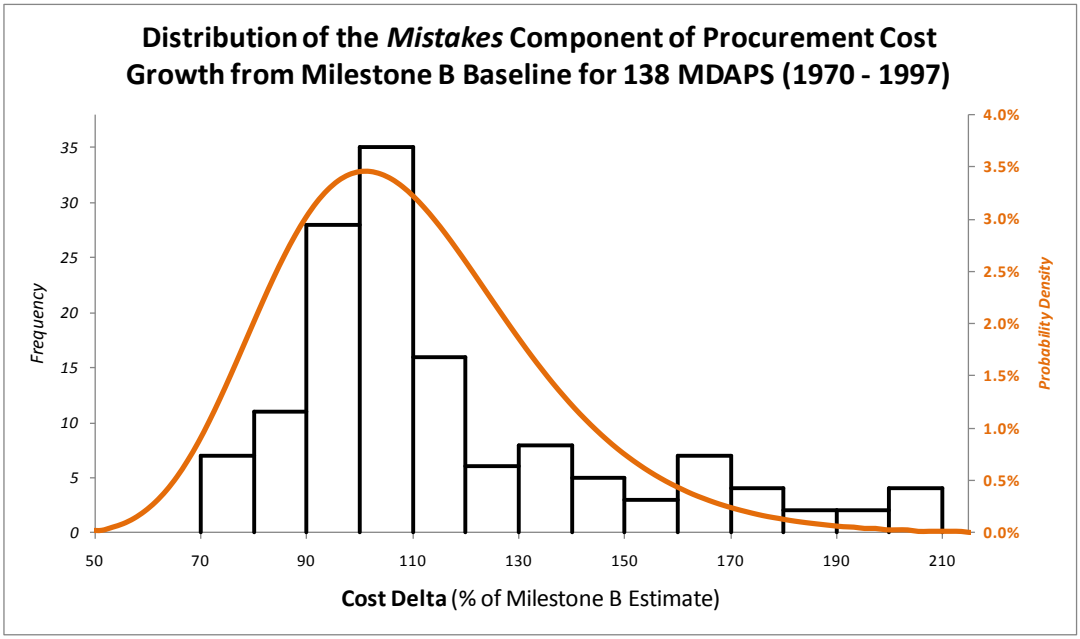
The Director of the Secretary of Defense's Cost Analysis and Program Evaluation (CAPE) office, Ms. Christine Fox, stated in testimony to Congress on March 24, 2010:

> "It is difficult to mathematically calculate the precise confidence levels associated with independent cost estimates prepared for major acquisition programs. Based on the rigor of the methods used in building CAPE estimates, the strong adherence to the collection and use of historical cost information, and the review of applied assumptions, we project that it is about equally likely that [the estimate] will prove too low or too high for execution of the restructured program as described."[4]

Ms. Fox is operating under the assumption that an estimate done with reliable historical cost data and reasoned assumptions should fall around 50% confidence. This is a common viewpoint in the Department of Defense, and it seems intuitive to believe that analysts are always trying to balance their assumptions and data to guard against the risk of being either too optimistic or too pessimistic in their forecasts. She is correct that calculating confidence levels can be difficult, and she took an important step in trying to describe the objective approach required to formulate an estimate at 50% confidence, but as the DoD moves forward, subjective inferences like this will need to be replaced with more statistical analysis. Analysts often intentionally make assumptions in their estimates that are either optimistic (perhaps to guard against the risk of a program getting cancelled) or pessimistic (to guard against the risk of underestimating), and sometimes they formulate extreme assumptions or estimates inadvertently. Assuming that an estimate is at 50% confidence could cause decision-makers to overlook the range of possible costs, a range that could show the point estimate to be an extreme scenario, depending on the assumptions used in formulating the estimate. A very small set of optimistic or conservative assumptions can *easily and unintentionally* cause an estimate to be an extreme outlier rather than the assumed 50% confidence.

We must use mathematics to explore uncertainty and risk, particularly because Congress and other decision makers are constantly bombarded with conflicting cost estimates and have no choice but to *guess* at the credibility and potential bias of various organizations performing those estimates. Not exploring risk mathematically will only give us the same results that we've been getting for decades. In a 2005 study done by the Institute for Defense Analysis (IDA)[5], Dr. David McNicol (former Chairman of the Pentagon's Cost Analysis Improvement Group, or CAIG) conducted a study of 138 Major Defense Acquisition Programs (MDAPs) that occurred between 1970 and 1997. He compared the estimated cost

of Procurement at the initiation of each program (Milestone B) to the *actual cost* of each program as reported in their final Selected Acquisition Report (SAR). Costs were adjusted to correct for major design changes and shifts in the procurement quantity, isolating what Dr. McNicol referred to as the "Mistakes Component" of cost estimating. The results are shown in the histogram below, where the x-axis represents the percent difference in cost (*actual* divided by *estimated*) and the y-axis shows the number of programs that fell into each interval. Notice that the histogram is right-skewed, meaning estimates during this period were *more likely to be too low than too high*.



One of the ways in which analysts often try to guard against the problem of underestimating is by adopting conservative / pessimistic assumptions. The problem with this approach is much the same as the problem with optimistic estimates. If you make a series of slightly conservative assumptions to guard against the possibility of a cost overrun, the compounding effect of risk could result in an extremely conservative estimate that drains resources that could be used somewhere else in the budget.

The draw of making conservative assumptions is particularly powerful in adversarial or competitive scenarios. After all, who wants to underestimate their competitors or enemies? However, *over*estimating one's adversary can have negative consequences just the same way as underestimating. Consider the nuclear arms race between the U.S. and Soviet Union during the Cold War. According to Alain Enthoven and K. Wayne Smith (two of Secretary McNamara's "whiz-kids") in their 1971 book How Much is Enough? Shaping the Defense Program 1961-1969, a small set of ill-considered assumptions made by NATO intelligence analysts in the early 1950's was a driving force that helped spark the beginning of the arms race. NATO intelligence on the Soviet military suggested that the land and air forces of the Warsaw Pact nations were far superior to those of the NATO allies, and that, if a war broke out along the Iron Curtain, the Americans and their allies had little chance of winning. In fact, not only was the Soviet military capability viewed as far superior, but it was considered to be so much more superior that the U.S. would likely go bankrupt in any attempt to match the perceived size and power of Soviet land and air forces. As Enthoven and Smith put it:

> "In the 1950s and early 1960s the standard military briefings given at NATO headquarters and by the Joint Chiefs of Staff compared the NATO and Warsaw Pact forces solely in terms of divisions. In 1961, the usual comparison was 175 well-equipped, well-trained, fully ready Soviet divisions facing about 25 ill-equipped, ill-trained, unready NATO divisions in the center region."[6]

The natural conclusion of such intelligence briefings left U.S. decision makers with the impression that the only affordable way to prevent a Soviet attack was to construct a sufficient deterrent to assure Soviet leaders that any attack would result in their total destruction. Consequently, the production of nuclear warheads by the United States ramped up into the thousands well before 1960. Several years later, the Soviet Union responded by constructing its own massive nuclear deterrent[7].

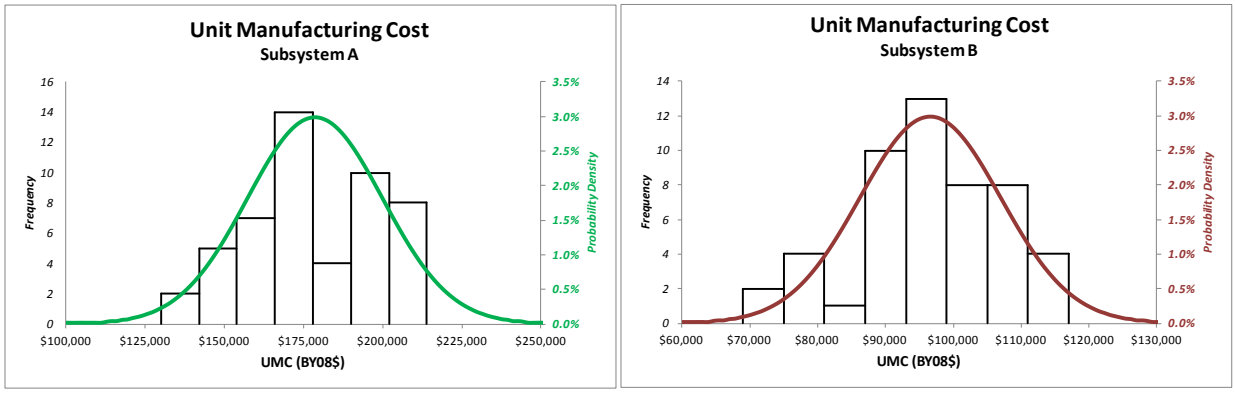
The problem is that the Soviet land and air forces really were *not* superior to the forces of the NATO nations. According to Enthoven and Smith, later analysis concluded that the Soviet forces were not organized along the same lines as the United States because one U.S. division had comparable fighting power to at least three Soviet divisions, which were organized in smaller units. Furthermore, of the 175 Soviet "divisions", at least half of them were "paper units" that were effectively void of manpower and equipment. Essentially, the nuclear arms race began because of two "conservative" assumptions made by intelligence analysts about the capabilities of our enemy in the 1950s and 1960s. Without any intelligence gathered to support their key assumptions, analysts *assumed* that the Soviet "division" was equivalent in fighting power to a U.S. division, and then *assumed* that all 175 Soviet "divisions" were well-equipped and battle-ready, all because they were attempting to avoid underestimating the capabilities of an enemy. Had the initial analysis been supplemented in any way by some sort of sensitivity or uncertainty analysis regarding these two key assumptions, more intelligence might have been gathered and the course of world history may have been altered.

In any scenario involving uncertainty, we are foolish if we rely solely upon our intuition and assumptions to generate point estimates. We must utilize uncertainty distributions to perform meaningful analysis. In a constrained budget environment where overestimating by 15% could cancel programs, and underestimating by 15% could strain future budgets, we cannot afford to lose objectivity or ignore the range of possible costs in our forecasts of the future.

Statisticians to the Rescue? The Neurosis of Numbers

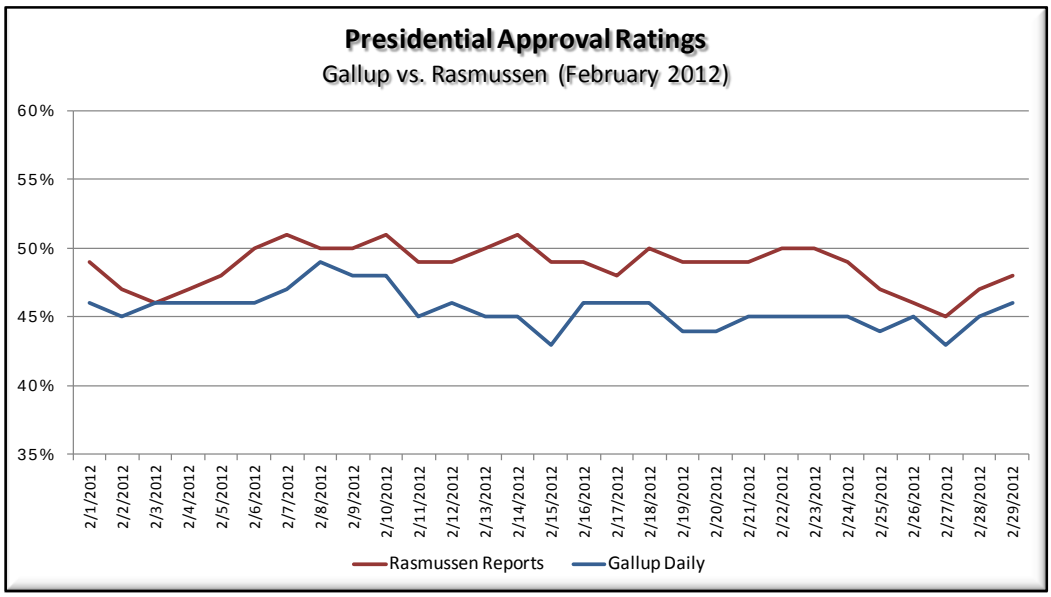
"Our lives teem with numbers, but we sometimes forget that numbers are only tools. They have no soul; they may indeed become fetishes. Many of our most critical decisions are made by computers, contraptions that devour numbers like voracious monsters and insist on being nourished with ever-greater quantities of digits to crunch, digest, and spew back." –Peter L. Bernstein[3]

To help resolve the problem of bias and risk creeping into point estimates, the concept of cost risk analysis has flooded the Department of Defense in recent decades, and was even required by law in the Weapon System Acquisition Reform Act (WSARA)[8] signed by President Obama in 2009. By placing point estimates into the confines of a statistical uncertainty distribution, analysts can establish a range of possible costs and express a level of confidence in any given projected value (i.e. 80% confidence that costs will not overrun $10 million). For example, cost data pulled from Cost and Software Data Reports (CSDRs) for subsystems on one particular weapons program (details excluded to protect proprietary information) has shown that unit manufacturing costs (recorded by production lot) varied during production according to the histograms on the following page, which show coefficients of variation of 10-12% of the average cost.



By compiling historical data like this and analyzing it with statistical regression analysis or goodness-of-fit tools, cost estimators can formulate ranges of possible costs surrounding their forecasted point estimates and begin to make statements about the likelihood of future costs falling within certain ranges. In an ideal world, ample supplies of historical cost data like this would flow readily into the hands of statistically savvy cost analysts eager to formulate ranges around their estimates. Unfortunately, *in the real world*, analysts in the Department of Defense are often dealing with only a few data points (far too few for statistical analysis) and are forecasting costs for systems that are pushing the envelope of advanced technology that falls outside the relevant range of the data gathered for existing analogous systems (if any analogous systems exist in the first place). This leaves analysts begging, "How are we supposed to formulate statistical distributions when we have so little data?"

Even when enough data is available to perform statistical analysis, it doesn't always tell a single, straightforward story, and analysts are still required to rely upon their best judgment to formulate statistical distributions. To demonstrate this concept, take a look at the following political polling data reported by Gallup[9] and Rasmussen Reports[10] on President Obama's job approval rating in the month of February 2012. Both Gallup and Rasmussen actively poll 1500 adults every three days, and both claim that the margin of error surrounding their polling data is +/-3%.



On several occasions in the month of February, Rasmussen reported President Obama's job approval rating to be 5-6% higher than Gallup, and at no point did Gallup ever report an approval rating that was higher than Rasmussen. Across the entire month of February, Rasmussen reported an average approval rating of 49% while Gallup's average was 46%. How can that be? There's clearly a statistically significant difference in these two polls, so if you look at one poll without looking at the other, you may get a false sense of confidence by actually believing the advertised margin of error. In other words, purely looking at statistics will not help you resolve this issue. ***Judgment*** regarding the *story behind the numbers* is the only way to understand why there is a divide in the polls (Note: The divide in these polls is most likely due to the fact that Gallup polls 1500 adults of voting age, whereas Rasmussen only polls "likely voters").

This same concept applies in the world of cost analysis as well. Consider a scenario in which you are formulating an estimate for the manufacturing cost of an engine. One way to estimate this cost is to use data on existing engines to formulate a regression analysis, creating what is commonly known as a cost-performance estimating relationship (CPER). Using data on existing engines, you could create multiple CPERs (i.e. one based on horsepower, one based on torque, etc), and perhaps several of these equations would exhibit favorable statistics. It isn't hard to run performance characteristics for a single engine through two different CPERs and get two different point estimates surrounded by different statistical error terms, both of which seem like entirely reasonable estimates. With two different point estimates that suggest two completely different cost ranges, how should you decide to forecast the cost of engines that will be built in the future? Mathematics can only go so far in helping you answer this question. At some point, you are still required to exercise reasoned intuitive judgment to formulate a forecast of the range of the engine's cost. Even in this scenario where substantial amounts of data can be gathered and analyzed to perform statistical analysis, any forecast based on that analysis would still require intuitive judgment.

The bottom line is that every cost forecast is a hypothesis about the future that requires the analyst to make a series of reasoned judgments, given all of the available information, about what the future will ultimately look like. Cost analysts can use statistics to help inform a forecast with a range of possible values, but the validity of that range is still determined by the credibility and realism of the underlying judgments behind the estimate. To put it another way, confidence levels are formulated based on *perceived* or *estimated* probability, not on actual probability. Numerical analysis is absolutely necessary to avoid the pitfall of erroneous assumptions and projections, but the numbers will only take you so far. Even in the realm of cost risk analysis and statistically generated confidence levels, decision makers still need to look analysts in the eye and ask hard questions.

The Collision of Intuition and Statistics: Heuristics and Biases

"When faced with a difficult question, we often answer an easier one instead, usually without noticing the substitution." –Daniel Kahneman in Thinking, Fast and Slow[11]

As discussed, our intuitive judgment is constantly present in the forecasts that we make, meaning that we're constantly making probability judgments, regardless of whether or not we are performing statistical analysis. Therefore, it's important to understand how our intuitive judgment operates in scenarios of uncertainty. Over the past several decades, two psychologists, Amos Tversky and Daniel Kahneman, performed a series of experiments to study how human beings respond to uncertainty. They found that the subjects of their studies often ignored basic rules of statistics in the face of uncertainty and relied instead upon a different set of rules. Essentially, as Kahneman put it above, when attempting to resolve a question with an uncertain answer, we often take a shortcut by replacing it with

an easier question. Shortcuts such as these are referred to as heuristics, and they can lead to systemic and predictable bias in how we make probabilistic judgments. The following sections explore some of the studies that describe these heuristics and biases, and their implications for the world of cost risk analysis.

Representativeness

Consider the following description of a hypothetical woman named Linda:

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Based on the description above, which of the following statements about Linda is more probable?

(a) Linda is a bank teller.
(b) Linda is a bank teller and is active in the feminist movement.

This question was asked of 142 subjects in a landmark 1983 study done by Kahneman and Tversky[1], and surprisingly, 85% of respondents believed that statement (b) was more probable than statement (a), despite the fact that (b) is a subset of (a) because (b) represents the conjunction of "bank teller" *and* "active feminist", whereas (a) represents the full set of all "bank tellers", including those who are active feminists *and* those who are not. This question pitted the judgment of respondents against a basic rule of statistics, referred to as the *conjunction rule*, and for a large majority of respondents, the conjunction rule lost.

The study also included experiments done in real world scenarios regarding the healthcare and sports gambling industries, and consistently subjects violated the conjunction rule. Why? Consider the Linda problem above. Given her description, it seems more likely that she would be active in the feminist movement than a bank teller because the stereotypical, or *representative*, characteristics of an active feminist seem to line up well with Linda's description (31, single, outspoken, concerned about discrimination and social justice, etc), whereas the stereotypical characteristics of a bank teller don't seem like a good match. When answering this difficult question, subjects more often replace it with a question about which statement is a more *representative story* than a question regarding one statement being a subset of the other. This is one mental shortcut that we often take when faced with uncertain scenarios, and it is referred to as the *representativeness heuristic*.

Why is this important in the world of forecasting costs? To put it simply, the representativeness heuristic is powerful because we tend to view scenarios that make *good stories* as being more likely than generalized or broad scenarios. In criminal trials, juries tend to favor stories with more details to fill in the unknowns[12], even though increasing the level of detail actually decreases the likelihood of the story being true. For example, stating, "He was murdered in town with a weapon of some kind" is more likely to be true than, "He was murdered behind the saloon with a double-barreled shotgun by his disgruntled business partner," but juries might incorrectly believe the later statement to be more probable because the additional detail helps them piece together a more complete story in their minds. The easier it is to imagine a plausible scenario in our minds, the more likely we perceive it to be. In cost estimating, this same principle applies. We favor estimates that make analogies to specific systems over estimates that draw wide ranges between two less representative systems. We look for the point estimate that makes the best story and struggle to convince ourselves that other slightly less representative systems are still within the range of possibility. Consider a 1990 study done by psychologists Ilan Yaniv and Dean Foster[13]. Subjects were asked "What amount of money was spent on education by the U.S. federal government in 1987?" After being told that the correct answer was $22.5 billion, subjects were asked to choose which of the following options represented the best estimate:

(a) $18 to $20 billion
(b) $20 to $40 billion

In this study, 80% of the subjects chose answer (a), $18 to 20 billion, over answer (b), even though answer (b) actually contained the correct value, a fact which subjects were well aware of when selecting their response. They selected answer (a) because the narrow range located near the correct answer made a better story than the wider range that contained the correct answer.

Availability

Another mental shortcut that often prevails in the face of uncertainty is the *availability heuristic*, which is the process by which we mentally examine the likelihood of a scenario based upon how easily instances come to mind. As psychologists Norbert Schwarz and Leigh Ann Vaughn put it:

> "When asked to form a judgment, people rarely retrieve all information that may bear on the task, but truncate the search process as soon as enough information has come to mind to form a judgment with sufficient subjective certainty. Accordingly, the judgment is based on the information most accessible at the time."[14]

Kahneman and Tversky have demonstrated this heuristic in multiple studies. In a 1973 experiment[15], they gave subjects 60 seconds to list seven-letter words of a specified form, and found that subjects were able to list 6.4 words of the form "_ _ _ _ i n g" and 2.9 words of the form "_ _ _ _ _ n _" (on average), meaning that the first form lent itself to a greater ease of availability than the second, despite the first form being a subset of the second (all seven-letter words ending in *-ing* also end in *-_n_*). Subjects were also asked to estimate how many times they would expect to find words of each form in a four-page novel (which they were told was about 2000 words long). As expected, words of the first form (*-ing*) were estimated to be far more likely to occur than words of the second form (*-_n_*). The median estimates were 13.4 and 4.7, respectively. Similar to the Linda problem, this violates the conjunction rule, but in this case, the violation is caused by the ease with which words of each form can be imagined (*availability*).

Word Form	Average # of Words Listed in 60 Seconds	Estimated Appearances in a 2000 Word, Four Page Novel
____ing	6.4	13.4
_____n_	2.9	4.7

Consider what this means in terms of cost analysis. Over time, cost analysts tend to gather experience with particular types of systems or programs by slowly familiarizing themselves with the technical characteristics, acquisition histories, data gathering efforts, and the personnel of a particular type of system or program. For example, cost analysts in the Department of Defense might specialize in tracked combat vehicles, missiles, fighter jets, satellites, or submarines. Over time, an analyst who gains great familiarity with the cost data and technical characteristics of combat vehicles is potentially very unfamiliar with data on tactical vehicles, and vice versa. An analyst who works for fifteen years on fighter jets will overload their data files with information on fighter jets, but might struggle to find reliable information on UAVs, refueling tankers, or commercial airliners. In the search for a valid estimate, this forces analysts to reach for the most available data and study its representativeness, meaning that analysts often truncate their search with whatever fills the folders on their desktops, rather than exploring a wider range of possibilities. Similar to the representativeness heuristic, the availability heuristic causes analysts to favor narrow distributions. It also feeds other mental heuristics. When searching for the best estimate, whatever data is most available tends to become what analysts believe is the most representative, and consequently, they tend to anchor to it (the *anchoring* heuristic

is discussed in the next section). Essentially, the stories that we already know overpopulate the futures that we forecast.

Anchoring and Adjustment

Perhaps the most important mental heuristic in the field of cost analysis is known as the *anchoring and adjustment heuristic*. When responding to a question with an unknown answer, we often grab for values that we know are wrong and attempt to adjust to formulate the correct answer. For example, if you were asked to identify the year in which George Washington was elected President, you could quickly recall that the Declaration of Independence was signed in 1776 and that Washington was elected President at some point well after that date. In this case, you *anchor* to the value *1776* (which you know to be wrong) and use judgment to *adjust* to a more likely estimate of the correct value.

Studies of anchoring have even shown that subjects will anchor to completely irrelevant values. In a 1974 study[16], Tversky and Kahneman asked subjects to spin a "wheel of fortune" that would stop at a specific number, say 52. Subjects were then asked to estimate the percentage of African countries in the United Nations. The study found that the numerical value that emerged on the wheel of fortune had a significant impact on the estimated percentage provided by the subjects, meaning that subjects incorporated an irrelevant number as an anchor in their judgment.

The heuristic of anchoring and adjustment applies in real world environments as well. In determining our own willingness to pay for a certain good or service, we anchor to prices listed by sellers. For example, in a 1987 study done by two professors at the University of Arizona[17], subjects visited a piece of property for sale in Tucson, Arizona and were given 20 minutes to examine the property. All subjects were asked to estimate a reasonable price to pay for the house. The only difference in their experience was that subjects were shown different listing prices for the house, $65.9K, $71.9K, $77.9K, or $83.9K. The results showed that subjects were significantly influenced by the listing price in determining a reasonable purchase price, suggesting average reasonable prices of $63.6K, $67.6K, $70.1K, and $69.5K, respectively. Even a group of real estate experts, who would presumably have some experience estimating the value of properties, were significantly influenced by the listing price in the study.

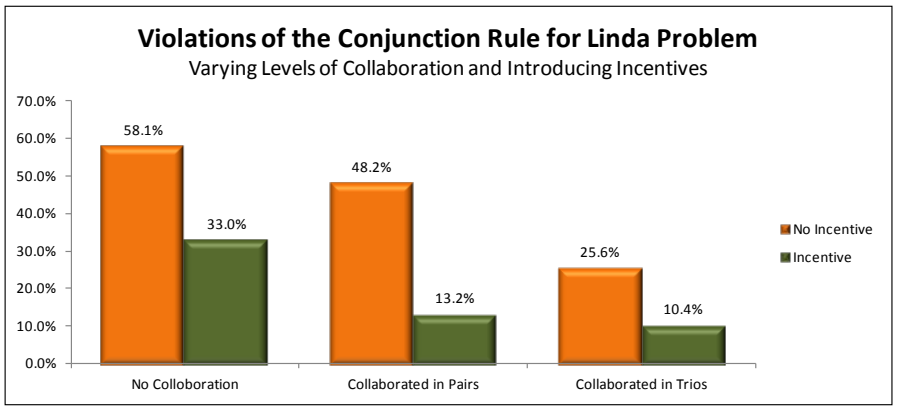
University of Arizona Study (1987) *Participants viewed a Tucson, AZ Home for 20 Minutes*				
List Price	$65.9K	$71.9K	$77.9K	$83.9K
Estimated Reasonable Price	$63.6K	$67.6K	$70.1K	$69.5K

The anchoring and adjustment heuristic often emerges in cost estimating when we draw analogies to whatever data points are most representative or available, but are then unwilling to challenge the stories that support those analogies in a sufficient way so as to allow ourselves to explore the broader range of informative data. When we anchor to a particular estimating methodology to formulate a point estimate, we tend to focus on information that lends credibility to that value rather than information that leads us away from it. Questioning a credible story or estimate is often challenging. As historian Barbara Tuchman once noted when describing the beginning of World War I, "The impetus of existing plans is always stronger than the impulse to change."[18] Similarly, the impetus of existing estimating methodologies is always stronger than the impulse to explore other options and data.

Correcting Errors of Inference

While these mental heuristics have a profound impact on our ability to perform accurate cost risk analysis, psychologists have found ways that we can actively reduce these heuristics and think more statistically.

In a May 2009 study[19], several psychologists repeated the Linda problem (described earlier regarding the representativeness heuristic), but added some twists in an attempt to reduce violations of the conjunction rule. By allowing minimal collaboration among study participants, and then adding financial incentives, the researchers were able to substantially reduce violations of the conjunction rule. Some students were forced to answer the question alone while others were allowed to collaborate in pairs or trios. At each of the three levels of collaboration, some subjects were offered a $4 reward for providing the correct answer, while others were offered no incentive. As the level of collaboration increased, violations of the conjunction rule dropped. Furthermore, in all three categories of collaboration, the error rate dropped significantly when subjects were offered a $4 reward. In fact, subjects that were offered the $4 incentive and were allowed to collaborate in groups of three only violated the conjunction rule 10.4% of the time, compared to a 58.1% violation rate for those who were not allowed to collaborate and were not offered incentives (Note: The 58.1% violation rate is lower than the 85% rate found in the 1983 Kahneman and Tversky study, perhaps due to changes in perceptions and stereotypes regarding the phrase "active in the feminist movement").



Other studies suggest that the effect of anchoring can be reduced by allowing subjects more time to adjust away from the anchor value, or by forcing them to shake their heads side to side when answering questions with well known anchors. For example, a study[20] done by Nicholas Epley and Thomas Gilovich asked subjects a series of such questions and subjects were asked to either nod, remain still, or shake their heads side to side while answering. When asked the year in which George Washington was elected President (anchor value = 1776), subjects who were nodding replied with an average answer of 1777.6, while subjects who remained still answered an average 1779.1, and finally subjects who were shaking their heads side to side gave an average answer of 1788.1 (Note: The correct answer is 1789). Similar results occurred for every other question presented in the study. The results suggest that a healthy skepticism, combined with sufficient time to act out that skepticism, is an effective means to objectively reduce the impact of anchoring.

Adding to these findings, a 1991 study[21] suggested that the representativeness heuristic was reduced when subjects were asked to "think as statisticians" rather than clinical psychologists. Other studies suggest that the draw of the representativeness heuristic is reduced in subjects well-educated in the field of statistics[1], and that subjects who practice logic questions prior to testing reduced their violation

of the conjunction rule[22]. All of these studies lay out a fairly clear formula for reducing the draw of heuristics. By encouraging collaboration, incentivizing long-term accuracy, maintaining an atmosphere of reasoned skepticism, and pushing the study of logic and statistics, leaders in the cost analysis community can improve the nature of probabilistic judgments across the board. Now that we've explored the errors present in our capacity to make sound probabilistic judgments and discussed some ways to reduce those errors, it's time to discuss how to translate these findings into a practical means for creating statistical distributions when we don't have reams of data to analyze.

Overcoming Heuristics and Formulating Distributions with Limited Data

"I am sure that no significant military problem will ever be *wholly* susceptible to purely quantitative analysis. But every piece of the total problem that can be quantitatively analyzed removes one more piece of uncertainty from our process of making a choice." –Robert S. McNamara, Secretary of Defense 1961-1968[23]

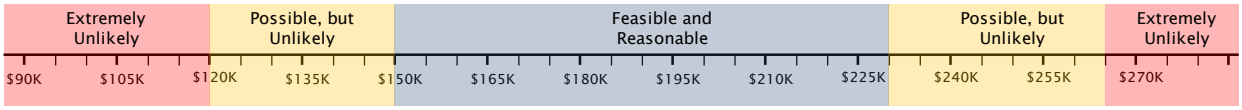
The less we are able to use reliable data and statistical analysis in our estimates, the more we are forced to rely upon our own intuitive judgments, but as we've seen, our judgment can be clouded by unintended flaws and biases. While statistical risk analysis in the form of regression and goodness-of-fit tests can be informative when ample data is available, the vast majority of cost analysis in the Department of Defense is done when little data is available. Risk analysis experts in recent years have begun to suggest that statistical distributions around cost estimates are often too narrow[24], and have described detailed mathematical solutions to this problem, such as the application of correlation[25] or the lack of sufficient use of "fat-tailed" distributions[26]. While these suggestions are useful and merit attention, they won't resolve the problem by themselves because most probabilistic distributions are formulated around cost estimates utilizing subjective judgment as the primary tool of choice, and unfortunately little has been made available to the average analyst to help them infuse sound statistical thinking into their judgments. Thus, the powers of heuristics and biases prevail, and the potential informativeness of narrow distributions is more powerful than the potential accuracy of wider distributions. To curtail the power of heuristics and biases in formulating distributions, analysts could consider some of the following techniques.

Infinity Cropping / Focusing on the Extremes: Starting with the Uniform Distribution

When attempting to estimate how much something will cost, a search begins for a target value somewhere in the realm of infinity. Rather than bounding the problem by focusing on finding a practical low and high, we are drawn by mental heuristics and the potential informative power of an accurate point estimate, and we begin where we shouldn't, by initiating a search for the most accurate single value. As Ilan Yaniv and Dean Foster mentioned in a 1997 study that will be described later:

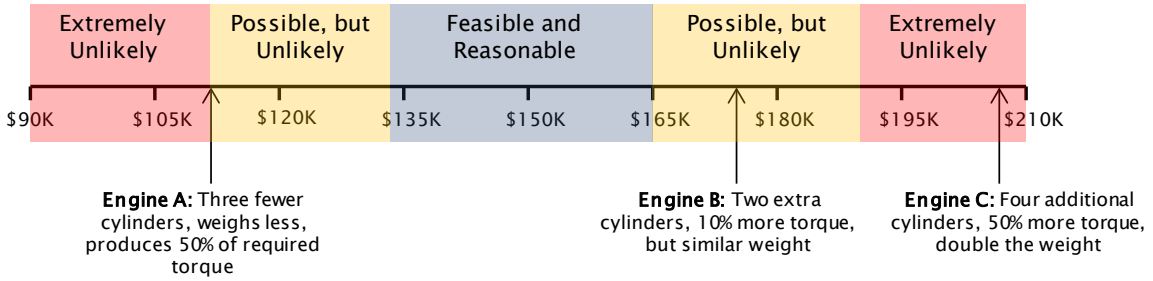
> "Rewards for being informative are immediate, as recipients evaluate the informativeness of a forecast upon hearing it. Rewards for being accurate are typically delayed to a later point in time when the relevant feedback becomes available and the forecast's accuracy can be assessed. This timing difference may further induce judges to provide highly informative [a.k.a. *narrow*] estimates." (Brackets added)[27]

In searching for the right analogy, cost estimating relationship (CER), or subject matter expert, we get sucked into the *point estimate mentality* and anchor to the most representative and available data. Consequently, the first task of any good risk analysis should be to "crop" out of infinity those costs that are judged to be extremely unlikely because beginning with an infinite range and narrowing/cropping down to a more reasonable range prevents the analyst from resolving the search with heuristics based on one or two data points.



Once this is done, practical endpoints can be determined to formulate a uniform distribution that accurately depicts what little the analysts knows. This is the equivalent of saying, "Given the little evidence that is currently available, all I can say is that costs will fall somewhere in the range of X to Y." Regardless of the amount of available data, this can be a valuable place to begin any cost estimate. The uniform distribution is, after all, the most basic probability distribution and is intended entirely for this purpose. As more data is gathered and analysis is done, uncertainty near the extreme endpoints of the uniform distribution can shift toward the best possible (most likely) point estimate. This is done by the process of eliminating or discounting possibilities near the extremes and simultaneously uncovering defendable evidence pointing towards more likely portions within the range. Until a practical minimum and maximum are determined through a process of "infinity cropping", any distribution formulated surrounding a single, most-likely point is extremely vulnerable to the "distribution narrowing" effects of heuristics.

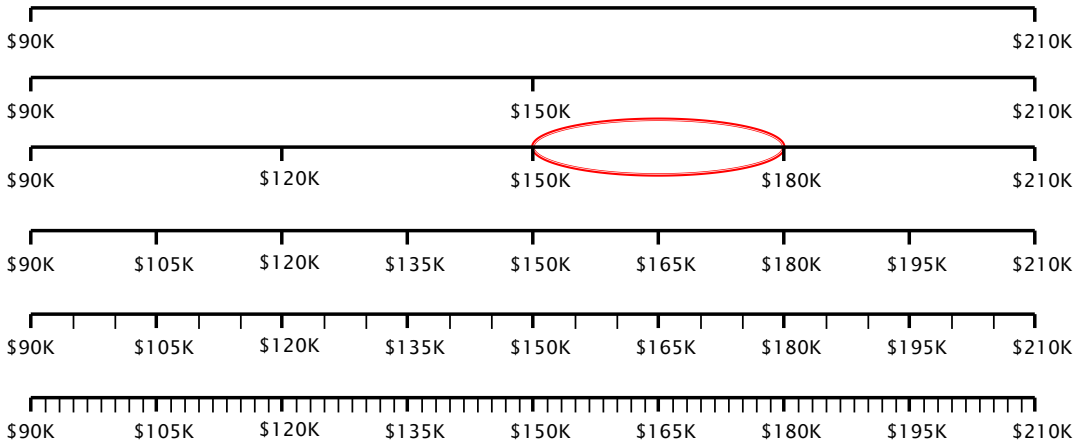
One of the most effective ways to search for practical minimums and maximums is to explore the systems / programs that we view to be poor analogies to our target system / program. Rather than initiating a search for the best possible analogy and getting caught up in the point estimate mentality, beginning with data points that we know are extreme can help bound the problem. While the anchoring heuristic suggests that we may be tempted to anchor to these extreme non-analogous data points, beginning with the extremes actually minimizes the effect of anchoring. Remember that anchoring is caused when we tend to focus on information that lends credibility to a value rather than information that leads us away from it. By initiating the search for a value that we believe to be an extreme, we begin with something that we already believe is not a credible data point, rather than beginning with the most analogous system and struggling to challenge its merits. Furthermore, beginning with a uniform distribution defined by practical low and high values puts us in a position where we begin with statistics (i.e. a probability distribution), rather than sticking a dart in the single point that we view as being most likely and then having to invent ways to create a low or high.



Grain Scales / Fidelity Intervals

A 1997 study published in the Journal of Behavioral Decision Making described a concept referred to as "grain scales". In the study, titled *Precision and Accuracy of Judgmental Estimation*[27], Yaniv and Foster (mentioned earlier) asked subjects to answer difficult questions (i.e. In what year did the first trans-Atlantic flight occur?), and then explored the merits of three methods of eliciting range estimates from subjects in the study. The first method explored the concept of grain scales, in which subjects were shown multiple scales with intervals of varying fidelity, and were asked to choose a scale where they felt

comfortable identifying one full interval as a best estimate. Subjects were asked to circle that entire interval from one tick mark to the next. An example of this is shown in the figure below:



The second method elicited range estimates from subjects by asking them to identify a range by providing low and high values so as to include the correct answer 95% of the time (i.e. a 95% confidence interval). The third method asked subjects to make a best guess at the correct value and then provide an error term (e.g. 1930 ±15 years). For all three methods, the same 42 questions were asked of all study participants. When providing estimates based on grain scales, the estimated intervals contained the correct answer 55% of the time. In contrast, the ranges elicited using the 95% confidence interval and the plus/minus error term methods contained the correct answer only 43% and 45% of the time, respectively. Despite being asked for a range that would be correct 95% of the time, subjects were only right 43% of the time! In contrast, subjects in the grain-scales method were asked to respond as if providing answers to a close friend, a person with whom they presumably would not feel a burden to supply the correct answer 95% of the time. Not only that, but when supplying answers with the grain-scale technique, respondents adjusted their chosen scales so that their answer was correct roughly 40-60% of the time for every scale, meaning that subjects intelligently balanced informativeness and probable truth in selecting an appropriate interval (Note: Subjects were always provided 6 scales, the 1st being the coarsest and the 6th being the most precise, similar to the figure shown above):

Hit Rate (%)	
Scale	Observed
1st	100%
2nd	51%
3rd	37%
4th	46%
5th	55%
6th	56%

Not only are grain-scales or fidelity intervals a practical means by which cost analysts can solicit estimates from subject matter experts (SMEs) when no data or information is available, they're also an effective means for utilizing subject matter experts to turn a uniform distribution into a triangular, beta, or Pert distribution. If an analyst uses the Infinity Cropping methodology to establish a practical low and high (i.e. a uniform distribution), in the absence of additional data to establish a most likely point within

that range, analysts can ask SMEs to provide estimates using the grain-scale technique to identify a narrower range of more likely values. To do this, they would draw the grain-scales with the practical low and high that were established via *infinity cropping* as the endpoints of each scale.

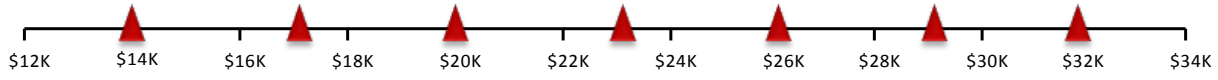
Eliminating the Average

Often times, we analysts have more data than we realize, but the draw of the *point estimate mentality* causes us to ignore the range of that data and simply calculate an average value. By doing this, we are effectively *throwing away* reams of valuable data. Recall the example mentioned earlier utilizing CSDR data. A weapon system that is produced over a span of several years might result in useful cost data reported on every production lot. The temptation is to simply average this data to establish a single point estimate. The problem is that this exact "average" value may never have been observed in any of the production lots delivered to date, and it may never be observed in any future production lots, even if it still ends up being the average value. Stanford professor Sam Savage wrote an entire book on this topic titled The Flaw of Averages: Why We Underestimate Risk in the Face of Uncertainty[28], a book in which he laid out the dangers of utilizing averages and ignoring dispersion. As Dr. Savage discussed in the book, consider a drunk stammering down the middle of the street. Over time, the drunk may be able to maintain an average path that stays on the center line where no cars are traveling, but he may be spending the majority of his time in harm's way. When you compute averages, you give a false sense of security that the future will occur near that average, and furthermore you establish that average as the anchor-point so that even when you look at the total dispersion of the data, you are tempted to explain away the endpoints, even though you can see as plain as day that you observed those very data points in the real world.

Cost analysts can resolve this problem simply by refusing to average data and to focus instead on its dispersion, a decision that opens up new avenues of possibility for forming distributions. Imagine that a particular program has completed seven lots of production, and the manufacturer has reported costs for each lot of a particular component, let's say the engine. You compile the data into the following table:

Production Lot #	Reported Unit Cost of Engine
1	$29,000
2	$23,000
3	$26,000
4	$14,000
5	$20,000
6	$32,000
7	$17,000

Before you throw these numbers into a spreadsheet and compute the average, remember that you currently have no information that would lead you to believe that any one of these seven data points is any more correct than any other. Nonetheless, a cursory look at the data will cause you to see that the low value is $14,000 and the high is $32,000. Is the temptation to compute the average getting to you yet? If you're predicting the engine's unit cost for Production Lot #8, you currently have no reason *whatsoever* to believe that $23,000 (the average) is any more likely to occur than $14,000 or $32,000. After all, each of these observations occurred exactly once. Instead of computing an average, let's try looking at these points plotted on a number line.



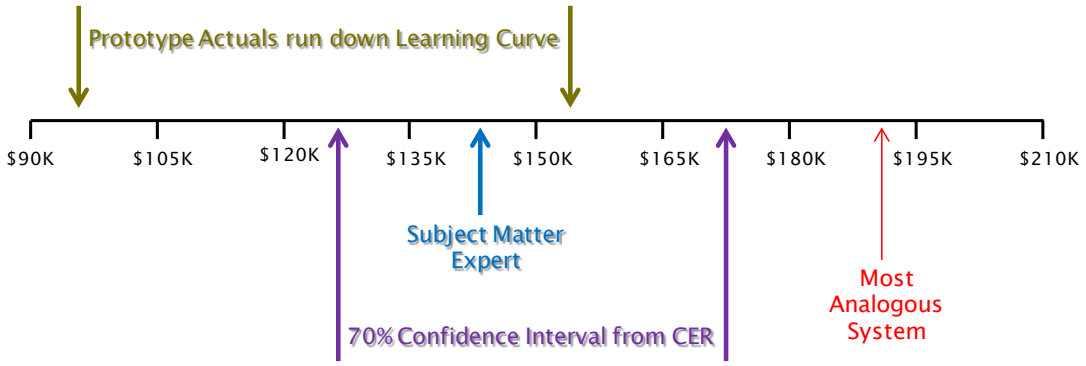
In this example, you have seven data points evenly dispersed over a range beginning at $14K and ending at $32K. In the point estimate mentality, you'd compute the average and report an estimate of $23K to decision makers. Clearly Production Lot #8 could come in higher or lower than this value, and probably will, so it's worth it to put a range around that average. How should you draw this range using these seven data points? The dispersion of the data suggests a uniform distribution, but the powerful draw of the average ($23K) leads us to believe that somehow it's more likely to occur than the endpoints. In other words, *the average acts as the anchor*, and our inner psychology causes us to look for merit in that number and to discount the endpoints, even though statistically there is absolutely no reason to do so. When you observe data near endpoints, don't average it away. Give it the credibility and attention it deserves because it could be tomorrow's fate.

Multiple Methodologies

When tests are given to measure a person's intelligence quotient (IQ), the questions in those tests are what are referred to as questions of *convergence*, meaning that the goal is for the test subject to "converge" onto the *one correct answer*. Similarly, psychologists have been trying in recent years to formulate tests of a person's creativity through *divergence* tests[29]. To provide an example, a divergent question might require you to write down as many different uses for a brick as you can possibly come up with in 60 seconds. There is no single correct answer, and a person who is skilled at answering questions of convergence might lack the creativity and quick wits to generate answers when asked questions of divergence.

This concept applies to cost estimating because ultimately *the formulation of a forecast is an exercise in divergence, not an exercise in convergence*. This is counterintuitive for most cost analysts because they have built their careers in the mathematical sciences and found success ultimately due to their skills in convergent thinking. Even though the future cost of a particular program will result in one number, we do not currently know what that number is, and in reality, we may never know. Thus, exploring that unknown requires divergent, or creative, thinking to visualize and investigate as many scenarios as possible. Exploring multiple scenarios is another way in which analysts can begin to formulate distributions.

Consider a scenario in which you are working on a project that has just completed the production of prototypes in the design and development phase of the program. Actual costs are available for the prototype build, and they show that the prototypes were roughly $1.5 million per unit. What will the system cost in production? There are multiple ways that you might estimate this cost. You could talk to experts and solicit their opinions, you could draw analogies to other similar systems in production, or you could adjust the prototype costs for production and run them down a learning curve. Different people might select different answers to this question, but why not try to do all of them and capture the results in one comprehensive location? If running the prototype cost down a learning curve leads to a production cost of $500-600K, does that mean that costs will absolutely fall inside that range? Perhaps another methodology will suggest that costs will be closer to $800K.

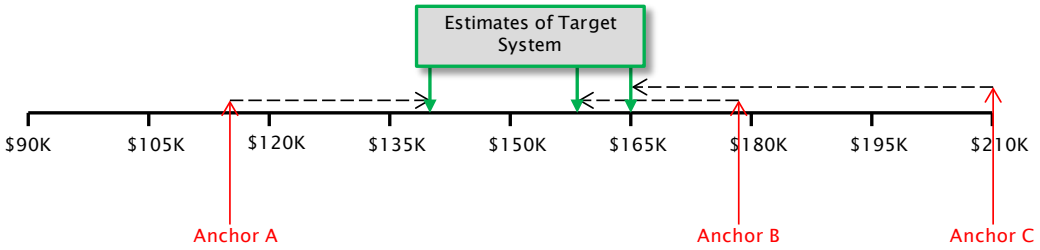


Carefully planning multiple approaches to estimating a cost and discussing the merits of each *before ever calculating a number* can prevent analysts from *anchoring* to a particular methodology. Using historical actual costs for prototypes can be very informative, but unknowns about how costs will adjust for production (i.e. the learning curve rate) can leave a lot of uncertainty and risk that might be mitigated by another approach to the estimate. Similar to the way that averaging historical data can disguise or throw away credible projections, ignoring alternate estimating methodologies can do the same thing. Every estimating methodology is a possibility worth at least some consideration. Plotting the results from multiple estimating methodologies in a histogram or on a number line can help inform the formulation of a distribution.

Forced Anchoring

If multiple data points are available for other systems or programs that have different attributes than the target system in question, cost analysts often times try to formulate cost estimating relationships (CERs) via statistical analysis, but what if only a few data points are available? Rather than trying to draw an analogy to the most representative system and fall into the traps of the representativeness and availability heuristics, analysts can utilize these data points in a different manner.

Let's say you have costs for three different radios. Radio A is smaller, lighter, and slightly less capable than your target system. Radio B is a little heavier, offering slightly more capability. Radio C is made of exotic lightweight materials and is packed with substantially more capability. Rather than choosing the best analogy from these three platforms yourself, you might decide to take all three data points to an engineer for feedback. The problem with this approach is that the engineer will be susceptible to anchoring to the most representative system just as much as the cost analyst. Radios A and C may seem too different to be useful data points, so the engineer could throw them out and ignore them, anchoring to Radio B, possibly without you even being aware of it. Instead, the analyst can get a sense of a reasonable range by separating these systems into a more divergent process. The cost analyst can provide the cost of Radio A to an engineer, and ask them to formulate an estimate of the target system from it. A similar process can be done with different engineers for Radios B and C. The results can be compiled in one spreadsheet or number line, similar to the following example:



This approach forces the engineer to anchor to particular values and then adjust. While Radio A might represent a practical minimum and Radio C a practical maximum in the formulation of a uniform distribution, that uniform distribution can be shifted towards a triangular, beta, Pert, or normal distribution by forced anchoring. SMEs will adjust upwards from practical lows and downwards from practical highs, and in so doing will reveal a more likely range. To force the issue of anchoring, analysts may have to "sell" the anchors by persuading SMEs to focus on the attributes that make them informative data points for projecting costs of the target system, but will need to be careful to keep in mind that multiple studies[20, 30] have shown that subjects need time to "detach" themselves from an anchor in order to provide meaningful estimates. Overselling an anchor value as a highly representative system can lead to an unintended lack of adjustment away from the anchor value.

Accuracy Levels: How Accurate Could You Possibly Be?

In the November-December 2008 issue of *Defense AT&L* magazine, COL Brian Shimel, USAF published an article titled *Risk, Uncertainty, and Trouble: Escaping the RUT of Program Instability* in which he stated as follows:

> "When predicting the price of a commodity as simple as a carton of eggs five years into the future, there is a standard error of 15%... Now imagine how much larger the standard error is for our sophisticated, state-of-the-art weapon systems that will take more than a decade to develop and procure."[31]

COL Shimel was on to something. If you can't predict gas prices 5 weeks from now within a range of ±1%, you almost certainly can't predict the unit cost of the latest developmental fighter jet within that range. In that sense, it's worth it to dissect our most reliable data sources, pour through information on our most continuous and homogenous production runs, and compile historical data for our most important variables to answer questions like the following:

- How much variability is there in our most reliable data sources?
- How much does the variability increase for slightly less optimal data?
- How well can we predict variables common to all estimates, like inflation, labor rates, fee/overhead, learning, etc?

Uncertainty levels may differ for different commodities and types of systems, and the removal of different controlling factors may impact distributions differently. For example, when looking at cost data reported for a production run of vehicles with a homogenous set of technical characteristics, normalized for inflation and learning, the reported data should vary less than data for another production run in which the manufacturer switched key materials multiple times during the process.

Armed with data analysis such as this, we can pull what we confidently feel is a single highly representative data point from past history and expect the variability around that point to be similar to other data resources of the same type. In other words, you may not have cost data on numerous production lots of a particular gun system, for example, but you may have a reliable contract cost from the latest production lot of that gun system and a sense of how much variability is typically present in

contract cost data. A similar methodology can be applied to labor rates, inflation rates, commonly used factors, productivity rates, fee rates, etc.

Consider an article published on CNN.com in August 2011[32] prior to Hurricane Irene striking the eastern seaboard of the United States. The article described how meteorologists project the disaster path of a hurricane and how they surround it with a "cone of uncertainty", such as is shown in the images below.

The article described a process by which forecasters utilize extremely fast computers that are ingesting data from satellites, ships, radar stations, weather balloons, and aircraft to perform billions of calculations using complex equations in various models to predict the path of the hurricane. With all that modeling and calculation, you might imagine that the "cone of uncertainty" is formulated with complex statistics, but it's not. The National Hurricane Center uses records of predicted paths on past hurricanes and "figures out what its average error is at various forecast horizons". The article pointed out that the "12-hour forecasted position is, on average, 36 miles off", and "at 48 hours, it is around 100 miles off." To put it simply, the National Hurricane Center is not using complex statistical models to determine the uncertainty of its forecasts. They are simply using their past history to determine how confident they are in their data and analysis. Cost analysts can, and should, utilize this same concept.

Conclusion: The Unseen

"Rather than conceal uncertainties, a good analysis will bring them out and clarify them… It is desirable to examine the available evidence and determine the bounds of uncertainty."
- Alain C. Enthoven, Assistant Secretary of Defense for Systems Analysis, 1965-1969[6]

In an article published in the New York Times on May 21, 2009, psychologist and author Daniel Gilbert discussed the power of uncertainty to generate fear. In the article[33], he described an experiment done at Maastricht University in the Netherlands in which subjects were given a series of 20 electrical shocks. As Gilbert described it:

> "Some subjects knew they would receive an intense shock on every trial. Others knew they would receive 17 mild shocks and 3 intense shocks, but they didn't know on which of the 20 trials the intense shocks would come."

The results of the experiment showed that subjects who knew for sure that they would receive 20 intense shocks sweated less profusely and maintained lower heart rates than subjects who were uncertain as to when the three intense shocks would come. This is what our program managers are going through when it comes to utilizing our cost estimates. If our estimates were consistently bogus, they would lose faith in us completely and quit listening to our forecasts because they would know that

we were going to shock them. Unfortunately, they're never quite sure whether they'll be shocked or not, so the fear of cost overruns is crippling. As Gilbert put it, "An uncertain future leaves us stranded in an unhappy present with nothing to do but wait."

The influence of heuristics and biases, the tendency towards convergent thinking, and the temptation to be informative today rather than accurate tomorrow all combine together with a synergistic effect that leads to the *point estimate mentality*. Breaking the point estimate mentality is difficult, especially when the proponents of cost risk analysis flood analysts with complex theoretical mathematics and the kind of convergent thinking that causes the point estimate mentality in the first place. Effectively, without good judgment and a willful battle against the powers of heuristics, the ever-touted 80% confidence level[8] is just as meaningless and useless as a point estimate. In forecasting cost, fancy mathematics alone will not steer decision makers clear from disaster, because the very heart of our projected confidence levels still lies within the subjective judgments that we make in formulating our estimates. The point of uncertainty and risk analysis is to help decision makers get a sense of what we know, what we don't know, and what we can project with reasonable confidence. Thus, cost risk analysis is not about trying to forecast a single number or range as close to accurate as possible, but about trying to look into the eyes of the infinite unknown, eliminate unlikely scenarios, and illuminate real *unseen* possibilities that merit more attention.

References

1. Tversky, Amos and Kahneman, Daniel; "Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment"; *Psychological Review*, Volume 90, Number 4, 293-315 (October 1983)
2. Gladwell, Malcolm; Blink: The Power of Thinking without Thinking; Published by Back Bay Books (April 3, 2007); ISBN 978-0316010665
3. Bernstein, Peter L.; Against the Gods: The Remarkable Story of Risk; Published by Wiley, August 31, 1998; ISBN 978-0471295631
4. Testimony of Christine H. Fox, Director of Cost Assessment and Program Evaluation (CAPE), Office of the Secretary of Defense (OSD); Before the United States House Committee on Armed Services Air & Land Forces Subcommittee and Seapower & Expeditionary Forces Subcommittee, March 24, 2010
5. McNicol, David J.; "Cost Growth in Major Weapon Procurement Programs", Second Edition; Institute for Defense Analysis (IDA); 2005; ISBN 0-9762550-1-4
6. Enthoven, Alain C. and Smith, K. Wayne; How Much is Enough? Shaping the Defense Program 1961-1969; Initially Published in 1971, Republished by The Rand Corporation in 2005; ISBN 0-8330-3826-5
7. Kristensen, Hans M. and Norris, Robert S.; "Global nuclear stockpiles, 1945-2006," *Bulletin of the Atomic Scientists*¸ Volume 62, No. 4 (July/August 2006), 64-66
8. Public Law 111-23; Weapon Systems Acquisition Reform Act of 2009; Signed by President Obama May 22, 2009; Title I, Sec. 101, §2334, (d) Disclosure of Confidence Levels for Baseline Estimates of Major Defense Acquisition Programs
9. "Gallup Daily: Obama Job Approval"; http://www.gallup.com/poll/113980/gallup-daily-obama-job-approval.aspx; Gallup, Inc. February 2012
10. "Obama Approval Index History"; http://www.rasmussenreports.com/public_content/politics/obama_administration/obama_approval_index_history; Rasmussen Reports, LLC; February 2012
11. Kahneman, Daniel; Thinking, Fast and Slow; Published by Farrar, Straus and Giroux (October 25, 2011); ISBN 978-0374275631
12. Rubenstein, Ariel; "False Probabilistic Arguments v. Faulty Intuition"; *Israel Law Review*, Volume 14, Number 2, 247-254 (1979)
13. Foster, Dean P. and Yaniv, Ilan; "Graininess of Judgment: An Accuracy-Informativeness Tradeoff"; *Journal of Experimental Psychology*: General, 21, 1509-1521 (1990)
14. Schwarz, Norbert and Vaughn, Leigh Ann; "The Availability Heuristic Revisited: Ease of Recall and Content of Recall as Distinct Sources of Information"; Heuristics and Biases: The Psychology of Intuitive Judgment; Cambridge University Press, 2002; ISBN 978-0-521-79679-8
15. Tversky, Amos and Kahneman, Daniel; "Availability: A heuristic for judging frequency and probability"; *Cognitive Psychology*, Volume 5, 207-232 (1973)
16. Tversky, Amos and Kahneman, Daniel; "Judgment Under Uncertainty: Heuristics and Biases"; *Science*, Volume 185, 1124-1131 (1974)
17. Neale, Margaret A. and Northcraft, Gregory B.; "Experts, Amateurs, and Real Estate: An Anchoring-and-Adjustment Perspective on Property Pricing Decisions"; University of Arizona; Published in *Journal of Organizational Behavior and Human Decision Processes* 39, 84-97 (1987)
18. Tuchman, Barbara W.; The Guns of August; The Random House Publishing Group (1962); ISBN 0-345-47609-3
19. Charness, Gary et al.; "On the Conjunction Fallacy in Probability Judgment: New Experimental Evidence Regarding Linda"; 19 May 2009

20. Epley, Nicolas and Gilovich, Thomas; "Putting Adjustment Back in the Anchoring and Adjustment Heuristic: Differential Processing of Self-Generated and Experimenter-Provided Anchors"; *Psychological Science*, Volume 12, Number 5, 391-396 (September 2001)
21. Schwarz, N., Strack, F., Hilton, D., and Naderer, G.; "Base rates, representativeness, and the logic of conversation: The contextual relevance of *irrelevant* information."; *Social Cognition*, Volume 9, 67-84 (March 1991)
22. Agnoli, Franca; Krantz, David; "Suppressing Natural Heuristics by Formal Instruction: The Case of the Conjunction Fallacy"; *Cognitive Psychology*, Volume 21, Number 4, 515-550 (October 1989)
23. McNamara, Robert; "Managing the Department of Defense"; *Civil Service Journal*, Vol. 4, No. 4 (1964), p. 13
24. Book, Stephen A.; "Cost S-Curves Through Project Phases"; NASA Independent Project Assessment Office technical report, September 2007
25. Book, Stephen A.; "Why Correlation Matters in Cost Estimating" presented at the *32nd Annual DoD Cost Analysis Symposium*, February 1999
26. Smart, Christian; "Covered with Oil: Incorporating Realism in Cost Risk Analysis" presented at the *Joint ISPA/SCEA Conference*, 2011
27. Foster, Dean P. and Yaniv, Ilan; "Precision and Accuracy of Judgmental Estimation"; *Journal of Behavioral Decision Making*, Vol. 10, 21-32 (1997)
28. Savage, Sam; The Flaw of Averages: Why We Underestimate Risk in the Face of Uncertainty; Published by Wiley (2009); ISBN 978-1118073759
29. Gladwell, Malcolm; Outliers: The Story of Success; Published by Back Bay Books (June 7, 2011); ISBN 978-0316017930
30. Barr, Dale J; Keysar, Boaz; "Self-Anchoring in Conversation: Why Language Users Do Not Do What They Should"; Heuristics and Biases: The Psychology of Intuitive Judgment; Cambridge University Press, 2002; ISBN 978-0-521-79679-8
31. Shimel, Brian COL, USAF; "Risk, Uncertainty, and Trouble: Escaping the RUT of Program Instability"; *Defense AT&L Magazine*, November-December 2008
32. "How Forecasters Develop Hurricanes' Cone of Uncertainty"; http://news.blogs.cnn.com/2011/08/24/how-forecasters-develop-hurricanes-cone-of-uncertainty/; 24 August 2011
33. Gilbert, Daniel; "What You Don't Know Makes You Nervous"; *The New York Times*; 21 May 2009

Author's Biography

Trevor L. VanAtta is a graduate of the University of Michigan, where he obtained a B.A. in Economics in 2006. He is currently an Operations Research Analyst at the U.S. Army Tank-automotive and Armaments Command (TACOM), where he specializes in cost and risk analysis for tracked combat vehicles. He has performed, analyzed, and validated cost estimates across the entire life-cycle of numerous systems, including tracked combat vehicles, route clearance vehicles, unmanned helicopters, autonomous and non-autonomous robotic vehicles, unmanned aerial vehicles, ground sensors, wheeled tactical vehicles, and combat support systems. Mr. VanAtta also worked in the Office of the Secretary of Defense's Cost Analysis and Program Evaluation (OSD CAPE) office in 2009, where he helped perform Independent Cost Estimates (ICEs) for ground robots, unmanned aerial vehicles, ground sensors systems, and the Navy's EA-18G Growler aircraft.

Mr. VanAtta has also served as Chair of the U.S. Army Cost Risk Working Group since 2008, leading the effort to write the Army's *Cost Uncertainty and Risk Analysis Guidance* in 2009. He also published the Army's January 2011 *Cost Uncertainty & Risk Analysis Quick Reference Guide.* He has performed cost uncertainty and risk analysis for the Army's Ground Combat Vehicle (GCV) program, consulted with numerous tactical and combat vehicle program offices at TACOM in their development of cost risk analysis, and provided detailed instruction on the topic of cost risk analysis across the Army. In addition to performing cost risk analysis for Program Office Estimates, Mr. VanAtta has also successfully applied the concepts of cost risk analysis in support of Army program affordability assessments.