Wisdom of Crowds; Tales of the Tail

The "Wisdom of Crowds" is a powerful method to understand a likely value of a quantity. However, extreme "tails" of a distribution are more difficult to understand. A wide range of biases limit human understanding of rare events. This paper describes means to better understand rare events using models and estimates from groups; crowdsourcing distributions, not just specific forecasts. A brief survey of prior work, and some original Lone Star research is presented.

Introduction

James Surowiecki's book[1] and a Nova episode[2] popularized Francis Galton's experiments about the "wisdom of the crowd." Galton was concerned about unwashed mobs (the worst kind of groupthink) and the threat of liberal democracy. But, a crowd can be wise; wiser than an individual. Many disciplines[3] understand Galton's work to demonstrate some important principles. Most important, the median (in some cases the mean) of several estimates is more likely to be accurate than one estimate.

But humans are not very good at probabilistic thinking. Kahneman and Tversky formulated Prospect Theory to explain how our intuition breaks down. Simply stated, humans are better at pattern recognition than probability distributions. And extremely rare events are among the most challenging for human estimators. Daniel Kahneman says, "What you see is all there is" (WYSIATI)[4] meaning that people can't visualize and often resist thinking about the unfamiliar. We are surprised at both rare events, and, by things which only *seem* rare because WYSIATI.

Tales of the Tail

Lone Star's work often involves client data (whether they think of it as "big data" or not) and elicited estimates. Using this data for analysis and prediction is often difficult, particularly at the tails.

One way to think of the problems at the tails is to classify judgment and data errors about them in three types[5]: Errors, Cognitive Distortion, and Ignorance.

---

[1] The Wisdom of Crowds is the title of James Surowiecki's 2004 book, which retells the Galton story. The title is an allusion to Mackay's history of popular folly, "Extraordinary Popular Delusions and the Madness of Crowds."

[2] See the Nova episode in music and rhyme; https://youtu.be/r-FonWBEb0o

[3] Social Science, Behavior Economics, Mathematical Psychology, Intelligence Analysis, Market Research….

[4] WYSIATI is a recurring theme in Kahneman's book, Thinking, Fast and Slow

[5] No claim is made for this list, other than being a useful framework for this short paper. These are probably not three orthogonal dimensions, and are probably incomplete.

- **Errors** – People make mistakes of many different types. When they miss the "right" answer, they can seem to be predicting a rare event. This can be true in eliciting expert judgment, or in the records of data sets. Experts can give answers in the wrong units; 3600 seconds is an hour. An expert who means "one" but says "3600" made a simple mistake. For big data and metadata, it can be daunting to "clean" errors and to distinguish errors from very rare events.

- **Cognitive Distortion** – We are bad at probabilistic things, and we are often at our worst near the tails. Prospect Theory shows humans are insensitive to changes in probability of rare outcomes. We are not likely to die in our next trip whether it is in a car or an airplane. But the odds are not the same. And, there seem to be more people who irrationally fear the safer of the two options.

- **Ignorance** – People have opinions, and will offer estimates, even about topics on which they are ignorant. In some cases, ignorance can be related to the first two categories. For our purposes, it is useful to think about ignorance separately.

We often find the outliers, or tails of data distributions to be the cause of estimating errors. One client had a great deal of maintenance data. In examining the actions of a simple procedure, we found the average time to perform the task was several days.

It turned out a poorly trained mechanic had entered the serial number in the field where minutes should have gone. A six digit number (like 7,284,433) is a long time, even in minutes (almost 14 years).

A few dozen of these entries skewed the average repair time to several days instead of about an hour.

This kind of outlier is not a 'rare event.' It is a 'mythological event.' In cleaning client data, we find a surprising number of mythological events.

Errors, in judgement or in recordkeeping, if not cleaned can lead to myths about rare events.

The second category – cognitive error is also tricky. Several studies show "experts" in one field over estimate their knowledge and accuracy in other areas.

Modeling and simulation for forecasting often relies on "Subject Matter Experts" or SMEs.

Some good work has been done on the calibration of SMEs in offering estimates, including estimates of probability.[6] But SMEs are bad at conditional probability. So, eliciting to characterize complex probability distributions is very challenging.

Finally, ignorance causes distortion. When we ask unqualified people we should not be surprised to get bad data. Egalitarians may be offended by this, but there is a difference between asking about simple concepts and abstract ideas. The weight of an ox or the number of jelly beans in a jar are concrete, simple concepts. Quantum probabilities of particle engagement will defy highly educated scientists. We need to bias our crowd toward people who can at least grasp the concepts we hope to explore.

Errors, cognitive distortion, and ignorance; these help explain why the tails of crowd estimates are bad.

History of Exploring the Tails

[6] For a very readable example, see Douglas Hubbard's book, "How to Measure Anything"

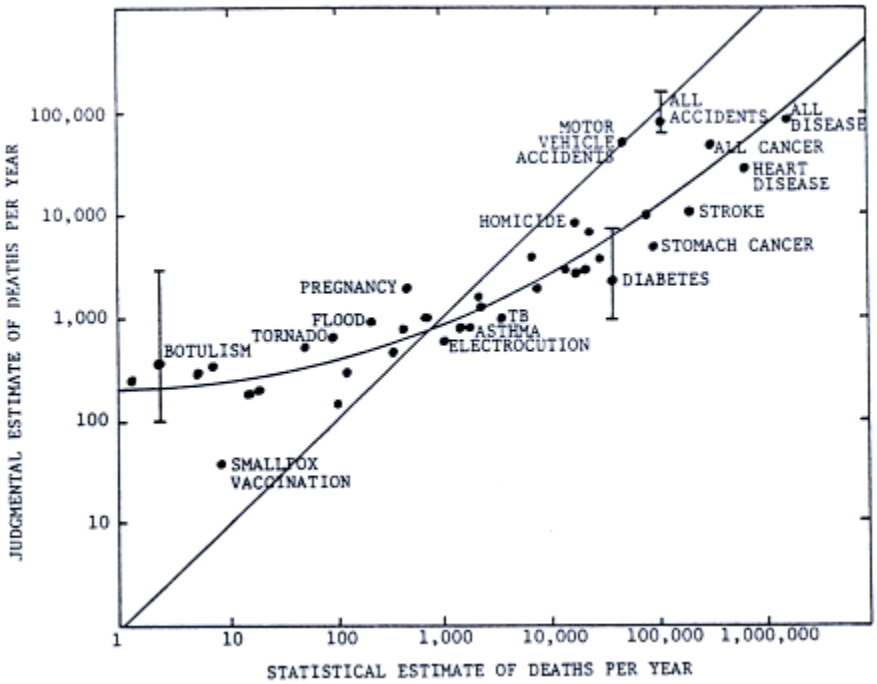
Work in the 1970's codified the combined effects of our errors, ignorance and cognitive distortions. This work, which shows our distorted perceptions across most of the range of probability. Slovic, Fischhoff, Lichtenstein, Kahneman and Tversky are among the most often cited from this era.[7]

In a number of elegant experiments, they explored the perception of uncertainty across a wide range of rarity, and across a range of risk importance.

An example is shown in the figure below taken from Slovic et al. (1979). The perceived probability of death from 41 causes is compared with statistical estimates. Each point in the figure compares the actual death frequency (horizontal axis) and the average value for the estimated mortality rates.

Correct estimates would lie on the straight diagonal line.

The estimates were obtained from 111 university students and 77 members of the League of Women Voters. Bars for selected points show the 25th and 75th percentiles of the range of perceptions.



[7]Slovic, Kunreuther and White. (1974). Decision processes, rationality and adjustment to natural hazards.

Tversky and Kahneman (1974). Judgement under uncertainty: Heuristics and biases. Science, 185, 1124-1131.

Slovic Fischhoff and Lichtenstein (1979). Rating the risks. Environment, 21(3), 14-20, 36-39.

Slovic, Fischhoff and Lichtenstein (1982). Facts versus fears: Understanding perceived risk. In Judgement Under Uncertainty: Heuristics and Biases. Cambridge University Press, New York.

For analysts interested in the tails, this data is intriguing. It shows that some familiar risks (motor vehicle accidents) are within the estimating capacity of the group. Other risks of equal probability (cancer deaths) were not accurately assessed.

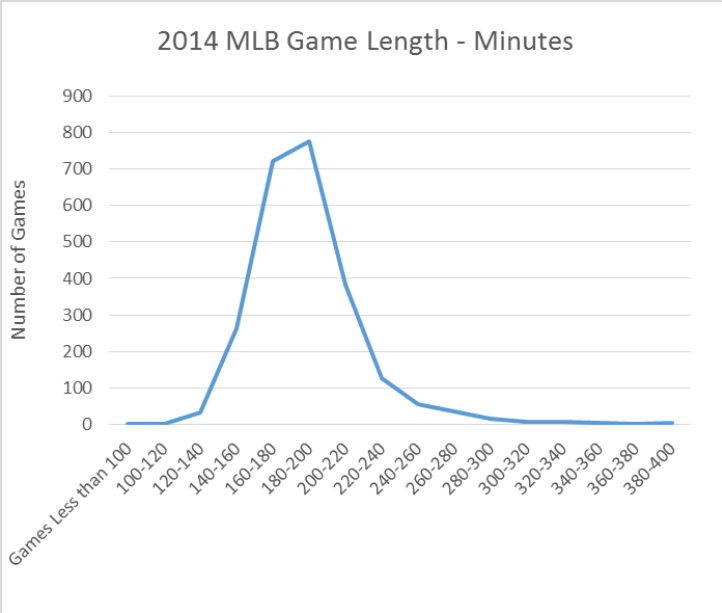
Since this figure only shows the central 50% of the estimate distributions, we are tempted to think the tails might be of more use in some cases, but this temptation is tempered; we also see some promising tails approach reality from above the diagonal line of truth, and some from below.

This data corresponds well to the work of Douglas Hubbard; estimators are often over confident in the goodness of their estimates. And, the overconfidence can either underestimate, or overestimate the probability of an event. Estimated ranges of outcomes tend to be too narrow and often exclude truth.

New Exploration of the Tails

Over the last several years, Lone Star has expanded our modeling for clients interested in rare events. These include rare catastrophes, international conflict, safety, rare accident forecasting, and competition outcomes. Our work shows a number of "proven" methods are not reliably useful.[8]

We have increasingly come to rely on formal voting among SMEs. These methods avoid many biases. In particular, secret ballots help avoid social pressure and hierarchal pressure (groupthink). Our TruCast™ electronic elicitation system is an example of a technology enabled implementation of these methods.



In 2014 Lone Star purchased several hundred survey responses to explore distribution tails, and how elicitation can reduce biases in response. We called it the "Estimating Experiment." If focused on probabilistic human judgement

Respondents chose from several topics. These included sports (baseball and football) and personal activities (commuting and grooming). The data has proven to offer rich insight about elicitation of uncertain outcomes. [9] In this paper, we present some information from respondents who chose to offer estimates about Major League Baseball

[8] For example, Scenario Based Planning (SBP) can be a useful way to "force" a group to explore rare events, but the culture, biases, and fears of a group can make provocative and well researched scenarios useless. Individuals must be able to grasp rare events, and need "permission" to be uncertain.

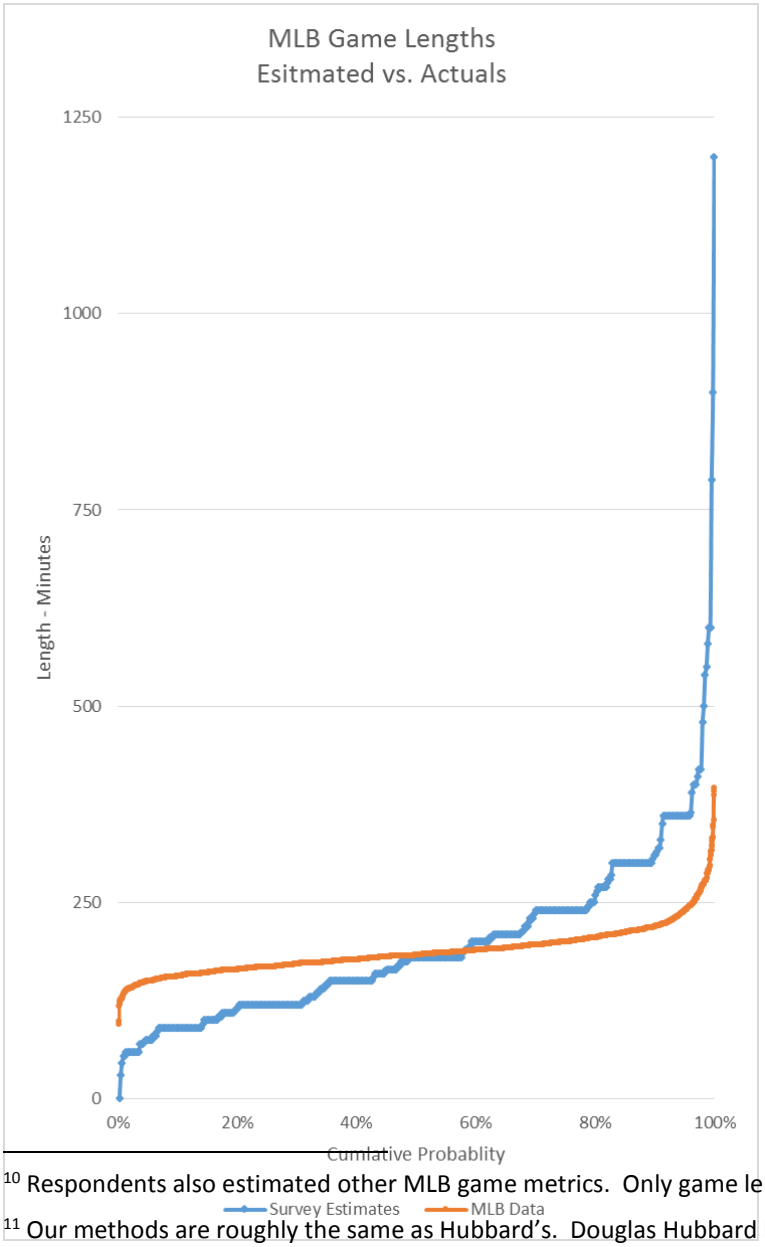
[9] Lone Star authors have previously presented some findings from the "Estimating Experiment" including findings about the relationship between overconfidence and statistical education.

(MLB).

Respondent estimates can be compared to the 2014 season. For example, respondents were asked to estimate game length in minutes[10].

The figure above (preceding page) shows the real distribution of 2014 MLB game lengths. The shortest game of the season was 96 minutes. The longest was just under 400 minutes. The distribution graph uses bin size = 20 minutes. Note how asymmetric the tails are.

The vast majority of MLB games last between 140 minutes and 360 minutes (about 98%). Television networks are said to prefer games of about 180 minutes, in order to create a three hour programming block. More than 80% of games in 2014 lasted between 160 and 220 minutes. Baseball is a very consistent sports "product" for television.



Comparing survey results with the 2014 MLB season provides some interesting observations. Shown in the chart on the left.

The blue line combines five different questions about game length. Respondents were asked about their nominal estimates, and about rare, or unlikely events.

For "real" SMEs, we prefer to provide training and calibration. TruCast™ provides a rapid means to provide this training and personal feedback to SMEs, helping them become calibrated.[11] However, in this case, uncalibrated estimators were used.

Their estimates can be compared with the actual results of the 2014 season. The figure on the left is one comparison of the human estimators (blue) compared to the actual 2014 season (gold).

[10] Respondents also estimated other MLB game metrics. Only game length is presented here.

[11] Our methods are roughly the same as Hubbard's. Douglas Hubbard has been very gracious and generous in sharing his methods with professional groups, such as Probability Management.

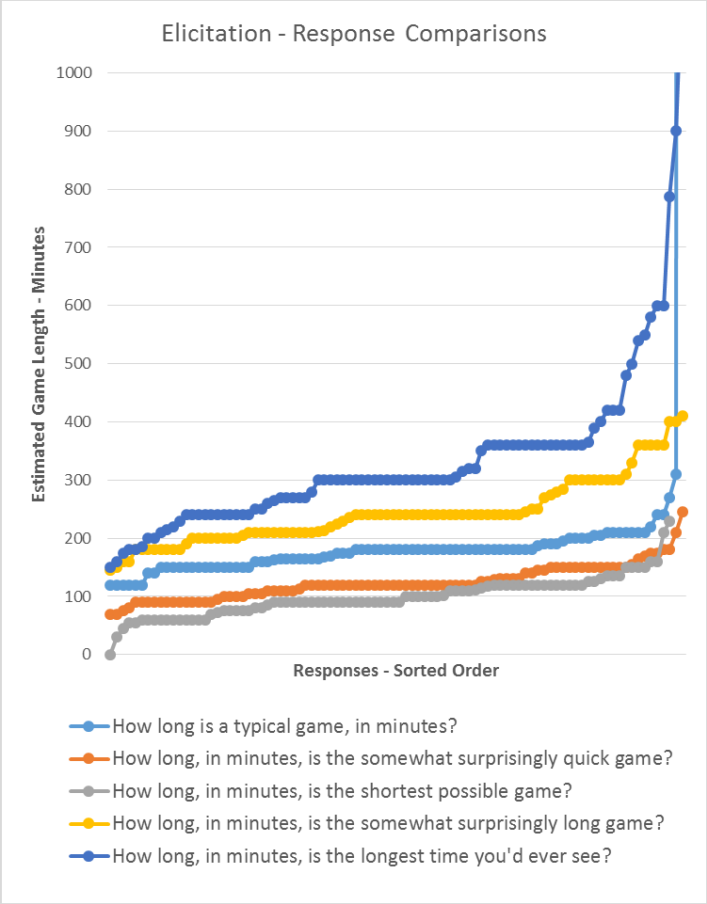
At the 50% probable (median) we see good agreement. The wisdom of crowds successfully estimated the nominal game length.

One effect is lumpy human estimates. Responses based on judgement did not create the smooth distribution of a real baseball season.[12]

Some other aspects of this comparison is notable. Both distributions are asymmetric on the long side. While the estimators could imagine more variability than the 2014 season generated, the character of the variability is roughly right.

This is an important finding. Understanding asymmetric tails is an important element of rare event modeling. This is also consistent with prior studies at Lone Star.

The respondents were asked to estimate game length in five ways. The responses to all five methods of elicitation are shown in this figure.



The five questions were:

1. How long is a typical game, in minutes?
2. How long, in minutes, is the somewhat surprisingly quick game?
3. How long, in minutes, is the shortest possible game?
4. How long, in minutes, is the somewhat surprisingly long game?
5. How long, in minutes, is the longest time you'd ever see?

From examining these responses, several questions and hypothesis come to mind.

1. Is it possible to have a game of zero length, though we didn't see one in 2014?

2. Is a game of 1200 minutes (20 hours) possible? Could the respondents be predicting things we can't see in a data set of about 2400 games?

3. Given that four of the five questions explored results in the tails of the distribution, what is the most appropriate way to blend the elicited responses (i.e., how to deal with oversampled tails)?

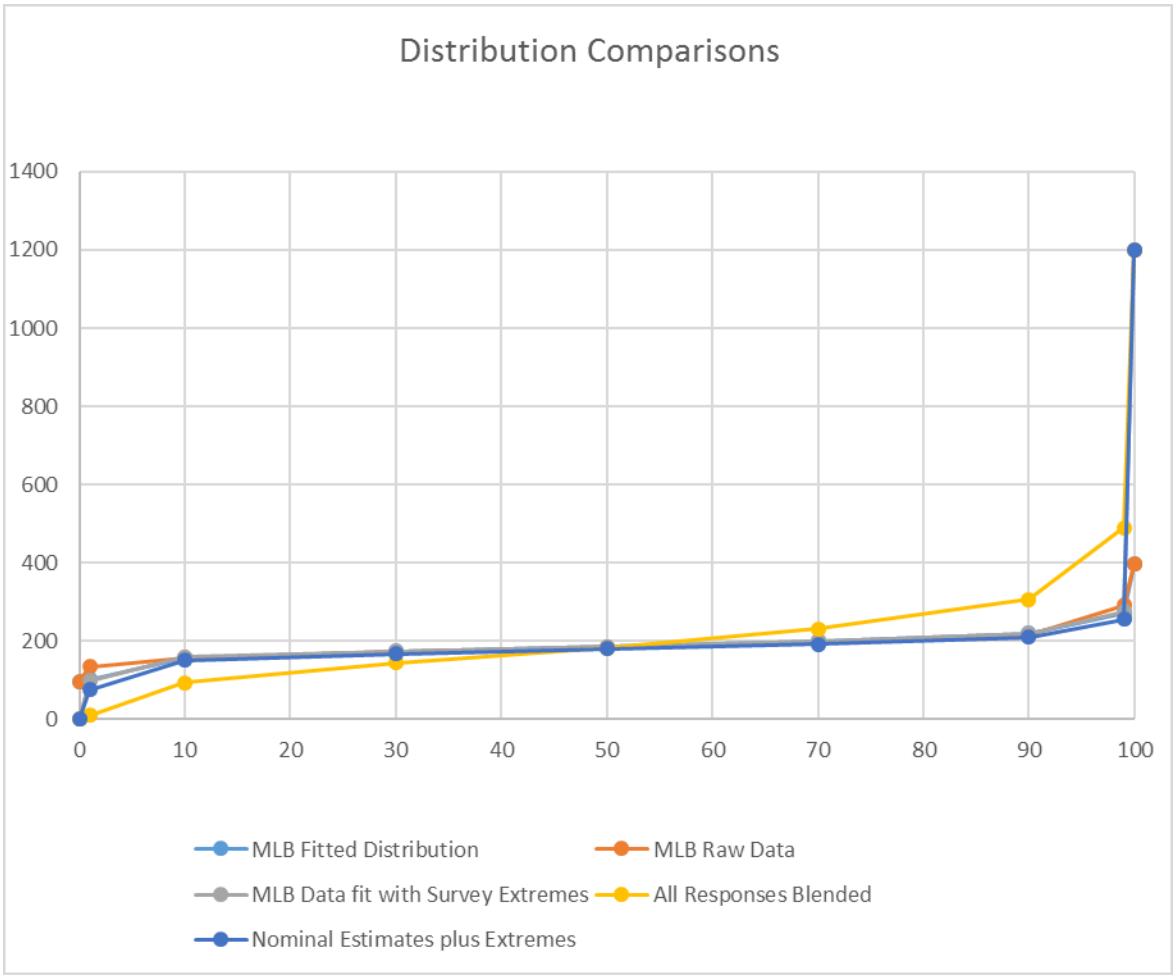
[12] For use of real data, Lone Star endorses the SIPMath™ open interface standard for sharing probabilistic data.

The first question, of zero length (presumably a forfeit) is in part based on the definition of "a game." Whatever the "shortest game" is, the 2014 season did not produce it. In 1919, New York beat Philadelphia 6 to 1 in a full 9 inning game lasting only 51 minutes. An MLB game must complete 4 ½ innings to be an official game. By that logic, it seems possible an official game could last less than 25 minutes (early innings are faster), even if we reject forfeits as official games. It seems the respondent who estimated zero is closer to the true end of the tail than the 96 minute shortest game of 2014.

The answer to the second question is "yes" it is theoretically possible for games to last for any length of time. Games over 400 minutes are rare. There was one in 2013, and there has been one as of this writing in the 2015 season. The 2015 game was between the Red Sox and the Yankees. When the game started, the Yankee first baseman was 34. He was 35 when it ended.

The longest game on the MLB record books was played in 1984, between the Brewers and the White Sox. It went 25 innings and lasted 8 hours and 6 minutes (486 minutes). If we sold rare event insurance we'd feel the actuarial risk of a 1200 minute game is very low, but not zero.

In attempting to build distributions from the survey data, a wide variety of methods can be compared. The following figures compares both raw MLB 2014 data, and fits of various combinations of data.



Except for the Raw MLB 2014 season data, we used representation compression. Distributions were represented by a method which sampled only 5 points. This is a method Lone Star frequently uses as a first order representation[13]. Even with compressed representations, the distributions match closely. The main differences come from questions of data cleaning (should we throw out the zero estimate?).

The yellow curve (all five responses to the elicitations) is the worst match, but this is hardly a surprise. Four of the five questions asked about extremes, so we oversampled the tails.

Perhaps the most intriguing representation is the comparison of distributions using the extremes (zero and 1200 minutes). Even with graphical rescaling, the three distributions which use the extremes are too close to distinguish by graphical means.

The following table summarizes a comparison of the five distributions.

	1	2	3	4	5
Cumulative Percentage	TruNav Fit of MLB Real Data	MLB Season Raw Data	Fit Raw Data with Tails from Survey Only	All Responses Blended and Fit	Fit Nominal Estimates with Limits from Extremes
0	96	96	0	0	0
1	101	135	99	10	76
10	157	157	158	93	150
30	173	172	173	144	167
50	185	184	185	181	180
70	199	196	199	231	191
90	219	214	220	307	210
99	270	292	275	490	255
100	397	397	1200	1200	1200

While these five distributions do not represent all of the potential methods for fitting or for loading data into distributions, they show some interesting diversity.[14]

Columns 1 and 2 demonstrate that a sparse representation creates a credible fit. The 1 & 99% points show some potential distortion, but the data here is so sparse goodness of fit is hard to estimate, even with several thousand points.

This is a good illustration of the problems with truly rare events.

[13] A full description is outside the scope of this paper. These were generated with Lone Star's TruNavigator™ or TruNav™ modeling tools.

[14] In a related, unpublished study, we explored a number of goodness of fit measures. It is fairly simple to apply these tests to data sets of all sizes, whether the distributions are fitted or raw data.

Comparing columns 2 and 3, we explore how elicited responses on rare events can be blended with the real data from the season. If we assume we need more than 2400 games to know what might transpire over a century of baseball, adding the zero and 1200 extremes to the real data from the season results in the fitted distribution shown in column 3. Again we see some distortion at 1 & 99% but otherwise the fit produces the same result as column 1 across the center of the distribution.

Column 4 is the same data shown on the yellow curve. As in the graph, we can see the results of oversampling the tails. The interesting thing here is how easy it is to escape the curse of overconfidence. In many elicitation methods, groupthink or other biases yield estimates over too narrow a range. Here we see under confidence; we asked respondents to think about rare events.

Column 5 is perhaps most intriguing. This used the elicited nominal responses, and added the extremes (zero and 1200 minutes). This very crude blending method of the elicited estimates is a good match to the real 2014 season in the middle of the distribution, and accounts for the extremes we might not see in decades of baseball. Said another way, if we used the distribution described in column five as the basis for random draws of 2400 games, that distribution would be much like a real season in most cases.

Implications for Lone Star's Preferred Methods

Lone Star typically decomposes problems into Little Questions™. Because of bias (WYSIATI) asking SMEs about "big questions" is often futile. And as this paper demonstrates, it can be hard to test the accuracy of predictions at the tails of distributions.

By decomposing problems into smaller components (Little Questions™) we become less dependent on the input tails as shown in this table.

Number of Input Factors Modeled	Odds All Factors Are in the 0 - 10% Tail Region	Odds All Factors Are in the 0-1% Tail Region
1	1.00E-01	1.00E-02
2	1.00E-02	1.00E-04
4	1.00E-04	1.00E-08
8	1.00E-08	1.00E-16
16	1.00E-16	1.00E-32
32	1.00E-32	1.00E-64

Even a fairly simple model, with only 8 independent sub-factors does not rely on the tails of input distributions to create the tails of output distributions. The odds that all eight sub-factors will be in their 0 -10% tails at the same time is 0.000001% (one in ten to the minus eight).

Several Lone Star models for our clients depend on accurate depiction of the tails. The MLB data shows it is practical to elicit useful data in the 10% to 90% portion of the cumulative probability distribution.

Decomposing into Little Questions™ provides a means to ensure model inputs depend primarily on this reliable portion of elicited distributions, whether "reliable" is from 1% to 99%, or from 15% to 85%. The resulting outputs are far more accurate at the tails than the inputs were.

Conclusions

Exploring this fairly simple data set shows a number of promising findings for several fields, including big data and problems related to the tails of distributions.

For very large data sets (e.g., big data) the use of extreme compression to represent some data is practical and useful. But humans are "lumpy" in their responses and no smooth distribution function is likely to mimic this behavior. For some problems, real data is critical.

By specifically eliciting responses for the tails of the distribution we found:

- Respondents as a group correctly suggested the asymmetry of a lop sided distribution

- Casual, non-mathematical questions, without respondent calibration provided good results over most of the distribution. Most methods to build the center of the distributions were roughly equivalent.

- Only the tails were hard to elicit and fit; the wisdom of crowds is weakest at the tails but even here, a very simple method using data from uncalibrated "man on the street" respondents provided a very respectable distribution.

- All of the methods we tested were more than adequate for the Lone Star Little Questions™ method