Teaching Pigs to Sing: Improving the Fidelity of Assessments from Subject Matter Experts (SMEs)

Peter J. Braxton, Technomics, Inc.

Richard L. Coleman (retired)

SCEA Washington Area Chapter

Wednesday, June 13th, 2012

Adapted from "Understatement of Risk and Uncertainty by Subject Matter Experts (SMEs), SCEA/ISPA 2011

"So-crates: 'The only true wisdom consists in knowing that you know nothing.'"
"That's us, dude!"

"Never try to teach a pig to sing; it wastes your time and annoys the pig."

Technomics
The Science of Informed Decision Making

Intro

At the April 2012 SCEA Washington Area Chapter luncheon, Mr. Marc Greenberg presented a paper on the use of Subject Matter Experts (SMEs) in cost estimation, where inputs are typically provided as low, most likely, and high values and treated as the corresponding parameters of a triangular distribution. This paper focuses on different aspects of that same general issue. Whereas the previous presentation emphasized **understatement of risk**, the use of a **formal elicitation process** to improve inputs, and exploring **alternatives to the triangular** distribution, this presentation emphasizes **understatement of uncertainty**, the use of a **feedback mechanism** and a **correction factor** to improve inputs (as measured by an empirical survey), and exploring different approaches to combining or **"conflating" disparate inputs** from multiple SMEs. If you'd like to know more, read on!

Abstract

There are two critical components to improving the use of SME assessments in cost estimating and risk analysis: (1) improving the fidelity of the inputs from each individual SME; and (2) combining those individual inputs in the best way. We will treat these in turn.

In much of the literature, experiments were conducted to gauge the degree to which SMEs tend to underestimate uncertainty (the range of possible values). These experiments generally focused on knowable but uncertain quantities, such as the height of Mount Kosciuszko. The authors improved the results of these studies by conducting their own experiments to include: (1) unknown and uncertain quantities, such as the number of points scored in upcoming sporting events; (2) a direct (self-)assessment of the degree of the SME's expertise in both the subject of the assessment and the (meta-)subject of risk assessment; (3) a quantification of the understatement of risk (tendency for growth) in addition to understatement of uncertainty – cost growth factors (CGFs) in addition to coefficients of variation (CVs), if you will; and (4) cost estimators and other defense acquisition professionals as the subject of the experiments. We will present the surprising and illuminating results from these surveys and their implications for "training" SMEs by providing feedback on prior assessments and applying correction factors in addition to or in lieu of such training.

The alternative methods explored for conflating the risk distributions of multiple SMEs include averaging random draws from the expert distributions, with or without correlation, for each run of the Monte Carlo; "averaging" the distributions themselves by averaging the means (or modes) and either averaging the extrema or taking the extrema of the extrema; and sampling the expert distributions. This paper provides a mathematical foundation for the conflation of triangular distributions by showing which of these methods are equivalent and how they compare in general. It reiterates recommendations for best conflation method in the cases of "single reality" and "multiple realities."

Together, these results hold the promise to significantly increase the quality and fidelity of expert quantification of risk.

SME Risk Topics

- Understatement of Risk and Uncertainty by SMEs
- Correction of SME Assessments for:
  - Accuracy (Risk adjustment)
  - Precision (Uncertainty adjustment)
  - Logical Consistency
- Conflation of SME Assessments
- Adjustment of SME Assessments with Data
  - Bayesian Probability! [maybe next time…]

Bibliography – Predecessor Papers

- "Risk Analysis of a Major Government Information Production System, Expert-Opinion-Based Software Cost Risk Analysis Methodology," N.L. St. Louis, F.K. Blackburn, R.L. Coleman, DoDCAS, SCEA/ISPA, Journal of Parametrics, 1998. *Awarded DoDCAS Outstanding Contributed Paper and SCEA/ISPA Overall Best Paper*
- "The Manual for Intelligence Community CAIG Independent Cost Risk Estimates," R.L. Coleman, J.R. Summerville, S.S. Gupta, DoDCAS, SCEA, ASC, 2002
- "Are We at the 50th Percentile Now and Can We Estimate to the 80th?" Richard L. Coleman, Peter J. Braxton, Eric R. Druker, Bethia L. Cullis, Christina M. Kanick, SCEA/ISPA, 2009, DoDCAS, 2010
- "The Correct Use of Subject Matter Experts in Cost Risk Analysis," Richard L. Coleman, Peter J. Braxton, Eric R. Druker, Bethia L. Cullis, Naval Postgraduate School (NPS) Acquisition Research Symposium (ARS), 12 May 2010; Department of Energy Cost Analysis Symposium (ARS), 19-20 May 2010
- "Determining the Cost of the Certification and Accreditation Process using Expert Opinion and Monte Carlo Simulation," A.J. Flynn, B.J. Nethery, K. Thomas, A.E. Gerstner, B.D. Dickey, C.M. Kanick, P.J. Braxton, SCEA, 2010

Note: The first predecessor paper in green above is included in its entirety (with modest updates), since it has not been presented to a SCEA audience before. We will hit highlights as needed for context and background en route to the "main event" (the survey results). The reader is invited to peruse the remainder at his or her leisure.

Bibliography – Literature Search

- "Subjective Probability Distribution Elicitation in Cost Risk Analysis: A Review," RAND TR-410-AF, 2007
- Alpert & Raiffa (1982) *A progress report on the training of probability assessors* in Kahneman, D., Slovic, P., & Tversky A., (Eds.). *Judgment under uncertainty; Heuristics and biases*, Cambridge University Press
- Brown, T. A. (1973). An experiment in Probabilistic Forecasting, R-944-ARPA
- Lichtenstein, Fischhoff & Phillips. (1982). *Calibration of Probabilities: the state of the art to 1980* in Kahneman, D., Slovic, P., & Tversky, A. *Judgment under uncertainty; Heuristics and biases*, Cambridge University Press

Problem Statement

- Expert-based risk methodologies are a common approach to cost risk
- Expert-based risk methodologies are defined for the purposes of this paper as follows:
  - Notwithstanding that the cost estimate may be based on actuals, expert-based risk methods rely on elicitation of the parameters of the risk distribution from expert opinion
    - Typically triangles for cost risk
    - Typically Bernoullis for technical risk
    - May include normals
  - Single or multiple experts may offer estimates of a particular risk via some form of parameterization
- This paper will discuss two topics
  - The "best" approach to converting extrema and percentiles from expert opinion into risk distributions
  - The "best" approaches to conflating multiple views of the parametrization of a single risk
- For completeness, the paper will also discuss some difficult characterizations the authors have encountered and the approach they have evolved for "correcting" them
  - Inconsistent percentiles
  - Unusual characterizations
- This topic was addressed in general in a prior paper[1] under the rubric "Omission Of Elements Of Variability"
- A confession: A prior paper[2] espoused a form of combination of expert testimony that this paper now recommends against

1. "Are We at the 50th Percentile Now and Can We Estimate to the 80th?" Richard L. Coleman, Peter J. Braxton, Eric R. Druker, Bethia L. Cullis, Christina M. Kanick, SCEA/ISPA 2009, DoDCAS 2010.
2. "Risk Analysis of a Major Government Information Production System, Expert-Opinion-Based Software Cost Risk Analysis Methodology," N.L. St. Louis, F.K. Blackburn, R.L. Coleman, DoDCAS, SCEA/ISPA, Journal of Parametrics, 1998. Awarded DoDCAS Outstanding Contributed Paper and SCEA/ISPA Overall Best Paper.

The "Best" Approach To Converting Extrema and Percentiles From Expert Opinion Into Risk Distributions

Correcting Extrema and Percentiles for Truncation

- Our estimated distributions tend to be "too tight"[3,4]
- Extrema
  - Without feedback, we provide values near the 20th %-ile and 80th %-ile when we are asked Min and Max
  - This can be improved, with feedback to the 10th and 90th %-iles
  - This can be improved by asking for more-extreme values:
    - "Astonishingly-low-probability outcomes" equate to the 0.1th %-ile and 99.9th %-ile
- Quartiles
  - Without feedback, we give 25th and 75th quartiles that actually contain only 33% of the outcomes vs. the expected 50%
  - This can be improved with feedback to 43% vs. the expected 50%
- Independent investigations of this over-tightness are remarkably consistent in the degree to which it occurs[3,4]
- Our ability to probabilistically characterize the past or future or to estimate our certainty on general-knowledge facts are all about comparable[5]

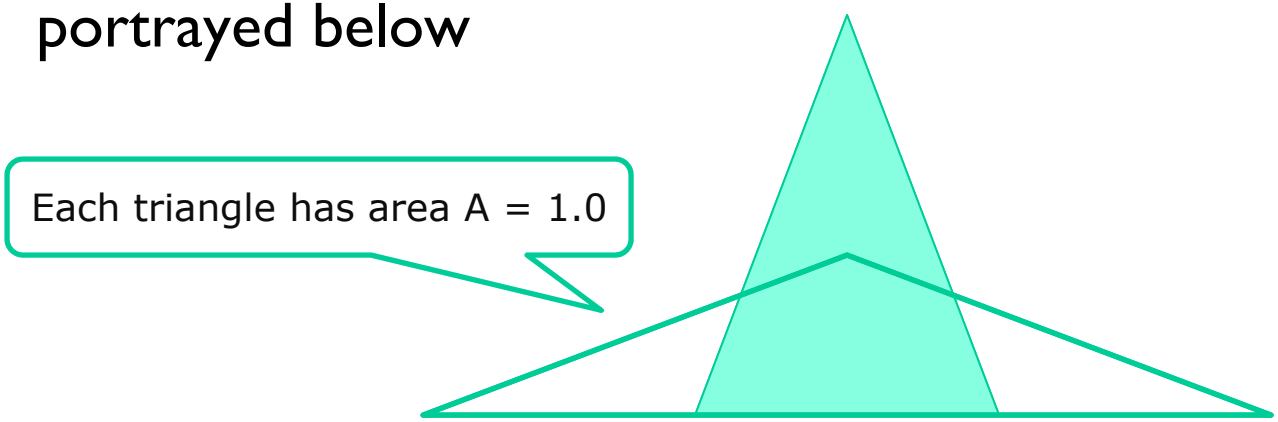
3. *Judgment under uncertainty; Heuristics and biases,* Edited by Daniel Kahneman, Paul Slovic and Amos Tversky, Cambridge University Press, 1982, Chapter 21, A progress report on the training of probability assessors, Alpert & Raiffa
4. *An experiment in Probabilistic Forecasting,* Thomas A. Brown, R-944-ARPA, July 1973
5. *Judgment under uncertainty; Heuristics and biases,* Edited by Daniel Kahneman, Paul Slovic and Amos Tversky, Cambridge University Press, 1982, Chapter 22, Calibration of Probabilities: the state of the art to 1980 Lichtenstein, Fischhoff & Phillips

Correcting Extrema / Percentiles – Two Views

- Assume that experts will return 20th and 80th percentiles when asked for the full range
  - In other words, when given "highs" and "lows," assume you are getting something more like plus and minus one standard deviation masquerading as extrema
  - This could be presumed to improve to 10th and 90th but only if the experts can be assumed to have gotten specific feedback about their accuracy at this task in the past
    - Note that this is not the same as saying they are very well qualified, it refers specifically to feedback training
    - We believe that practitioners have mistaken expertise for being trained and that this is why many practitioners believe experts provide 10th and 90th percentiles
- Although we don't typically ask for quartiles, we recommend assuming that a claimed 25-75 inter-quartile range is actually a 33-67 percentile range
  - This can be improved to a 28-72 range with specific feedback
- The two distortions above are not strictly coherent, meaning that they yield different corrections
  - The full range case is a greater understatement than the interquartile case
  - *The wider the confidence interval you ask for, the more the witness will understate it*
- When given expert testimony, therefore, it is appropriate to correct the testimony by adjusting the standard deviation or the end points using the two general results above, depending on the form given

Errors of Extrema - Pictorially

- We note that experts appear to be providing approximately the 20^{th} and 80^{th} percentiles
- We know* that the 20^{th} percentile occurs at a point that is $\sqrt{(1/10)} = 0.316$ of the base
  - The understatement of spread by experts is on the order of a factor of 2.5
- Pictorially, then, we are experiencing a reduction in distribution on the order of the teal (claimed) to the white (actual) portrayed below



* For the geometry of triangles with regard to percentiles and area, see the Appendix

The "Best" Approaches To Conflating Multiple Views Of A Distribution

Conflation of Expert Information

- Conflation refers to the combining of different (independent) views of a thing to arrive at a single (better, and more complete) view of it
- We seek to conflate expert testimony principally because we will arrive at a better estimate for the mean
  - But, what about the dispersion?
- Conflation is the most difficult problem for expert-based risk methodologies
  - This is not immediately obvious, but it is so
  - Dispersion is in turn the hard part of the conflation
- Ad hoc conflations are often used for k experts each giving estimates for the same risk or WBS element, e.g.:
  1. Use the individual expert testimonies in each run of the Monte Carlo:
     a. Make k random draws from the k different distributions and average them[6]
     b. Make k random draws from the k different distributions with correlation and average them
  2. Derive the parameters of a single distribution from the parameters of the expert testimony and then Monte Carlo
     a. Make a new distribution with i) the mean of the k expert means and ii) the mean of the standard deviations, for normals[7], or the means of the respective end points for triangles [Average the Parameters]
     b. Make a new distribution with the average mode of the k experts and the lowest low and the highest high as end points
     c. Make a new distribution with the average mean of the k experts and the lowest low and the highest high as end points
  3. Sampling has been endorsed in the literature[7]
     - For each run of the Monte Carlo, pick the answer from a randomly selected expert who provided testimony
- We will examine each of these methods
  - In backup we prove that 1b and 2a are equivalent for symmetric triangles and we speculate that for asymmetric triangles there is no significant difference, and so there is nothing to separate these beyond ease of implementation

6. "Risk Analysis of a Major Government Information Production System, Expert-Opinion-Based Software Cost Risk Analysis Methodology," N.L. St. Louis, F.K. Blackburn, R.L. Coleman, DoDCAS, SCEA/ISPA, Journal of Parametrics, 1998. Awarded DoDCAS Outstanding Contributed Paper and SCEA/ISPA Overall Best Paper.
7. *An experiment in Probabilistic Forecasting,* Thomas A. Brown, R-944-ARPA, July 1973

The First Question

- No single conflation method will work for the two possible scenarios that can confront the estimator

1. "Single Reality": There is a one (typically uni-modal) distribution, which we do not know, but which experts are presumed to know to some degree of accuracy
   - Example: What is your estimate for the GNP of Brazil for 2009?
   - Example: How big is a brown bear?
   - Example: What is the range of technical risk for the cost of the engine?
2. "Multiple Realities": There are k (typically uni-modal) distributions, we generally know neither k nor the individual distributions, but experts are presumed to know at least one each to some degree of accuracy
   - Example: How far away is your favorite planet? [there could be up to 9 answers depending on the inclusion of Pluto and Earth!]
   - Example: How big is a panda? [there is a lesser panda and a greater panda, but we don't happen to know that and fail to specify]
   - Example: What is the cost risk for the engine on the F-35? [There is a main and an alternate engine, each has a range]

- This problem may seem silly, but it is not, and our choice of conflation methods depends on the case we believe to apply
- We will recommend approaches for both, but first, decide which case applies
- The amount of spread in your expert testimony will give you an idea whether single or multiple reality is more likely
  - We recommend against feedback or "drilling down" until after initial testimony is gathered because witnesses are notoriously vulnerable to witness leading, anchoring and all other sorts of mischief … you'll never know

Desiderata for Single and Multiple Reality Cases

- Each case dictates different characteristics for the conflation technique
- Single reality:
  - Best estimate for the mean
  - Best estimate for the dispersion
  - Best estimate for the distribution
- Multiple Realities
  - Best portrayal of the multiple choices we are confronted with
- We will discuss each in turn

The Preferred Methods

- We will describe the apparent preferred solution for each method after asserting them below
- Single Reality:
  - Average the parameters and correct for the understatement of extrema (using method 1b or 2a from an earlier slide)
- Multiple Realities
  - Sample from the experts after correcting each for understatement of the extrema
- If we cannot discern whether we are in Single Reality or Multiple Realities, we recommend sampling
  - Because this is more conservative, meaning it will have wider dispersion
- We reject the use of averaging answers on each iteration despite having used the method in a Best paper Overall[8] in 1998. To see why, we will show its characteristics and indicate why it is probably unsuitable.

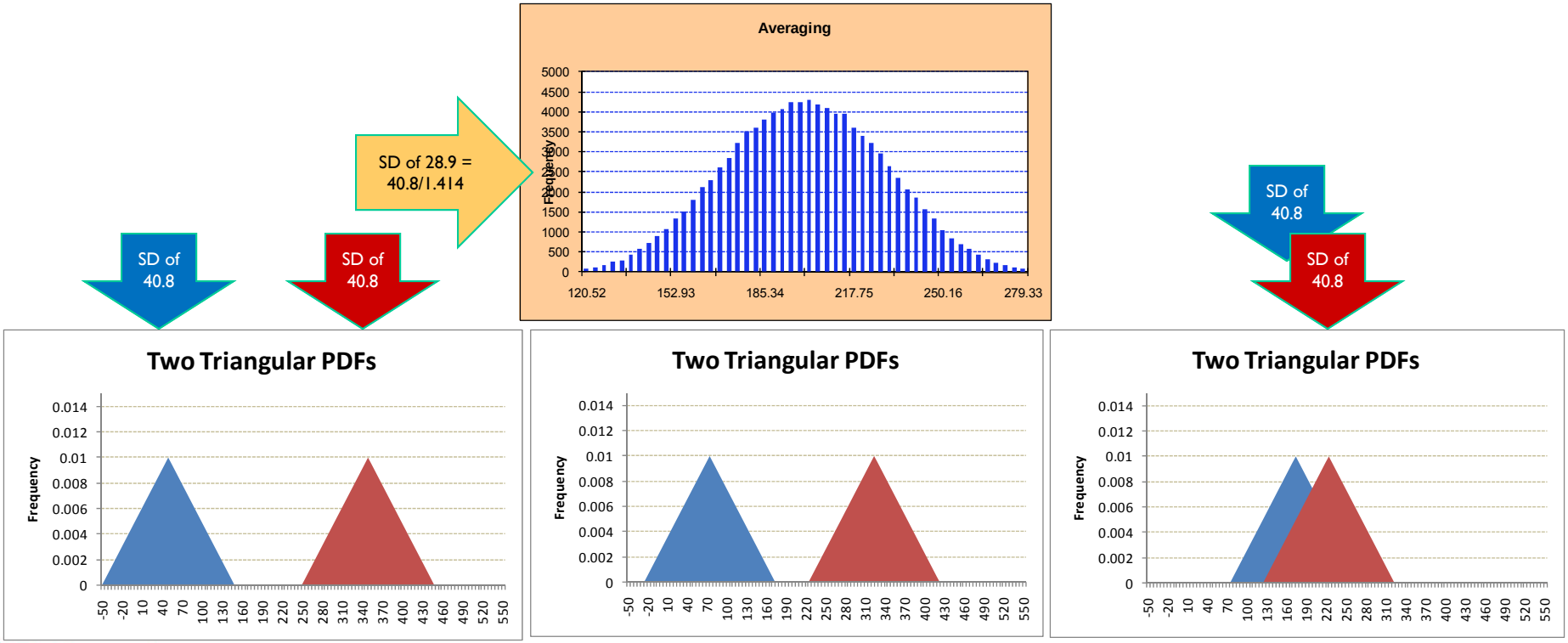
8. "Risk Analysis of a Major Government Information Production System, Expert-Opinion-Based Software Cost Risk Analysis Methodology," N.L. St. Louis, F.K. Blackburn, R.L. Coleman, DoDCAS, SCEA/ISPA, Journal of Parametrics, 1998. Awarded DoDCAS Outstanding Contributed Paper and SCEA/ISPA Overall Best Paper.

Recommendation - Single Reality

- The mean of the single reality not troublesome, almost any *reasonable* approach will yield the same mean
  - We use the word "reasonable" with trepidation
- The standard deviation presents the problem, since individuals are known to under-report, and some methods are vulnerable to distortions
- We recommend averaging parameters of the expert testimony because it is clear what is happening
- Correct each expert's testimony for truncation of the standard deviation, or correct the average, there is no obvious difference in the order of the operations
  - Techniques for correcting the standard deviation were shown in the first part of the paper

Conflation: Averaging on Each Iteration (1a)

- Averaging on each iteration can have an unexpected result: Three very different sets of testimony by two experts will produce exactly the same picture
  - This is not obvious at first, but it is so
- The standard deviation of k identical but scattered triangles, with SD = s, when iteration-averaged will produce a standard deviation s/√k
  - The SD of the conflation can be thus be arbitrarily small, if k is sufficiently large
  - This does not comport with our desire that the SD be well modeled
  - Correction for k can be achieved by a spreading with √ k but this is likely to be done wrong or omitted altogether, and at best would require row-by-row corrections
  - Correction for expert truncation can be achieved by treating the end points as if they were 20/80 points, this can be done before or after conflation

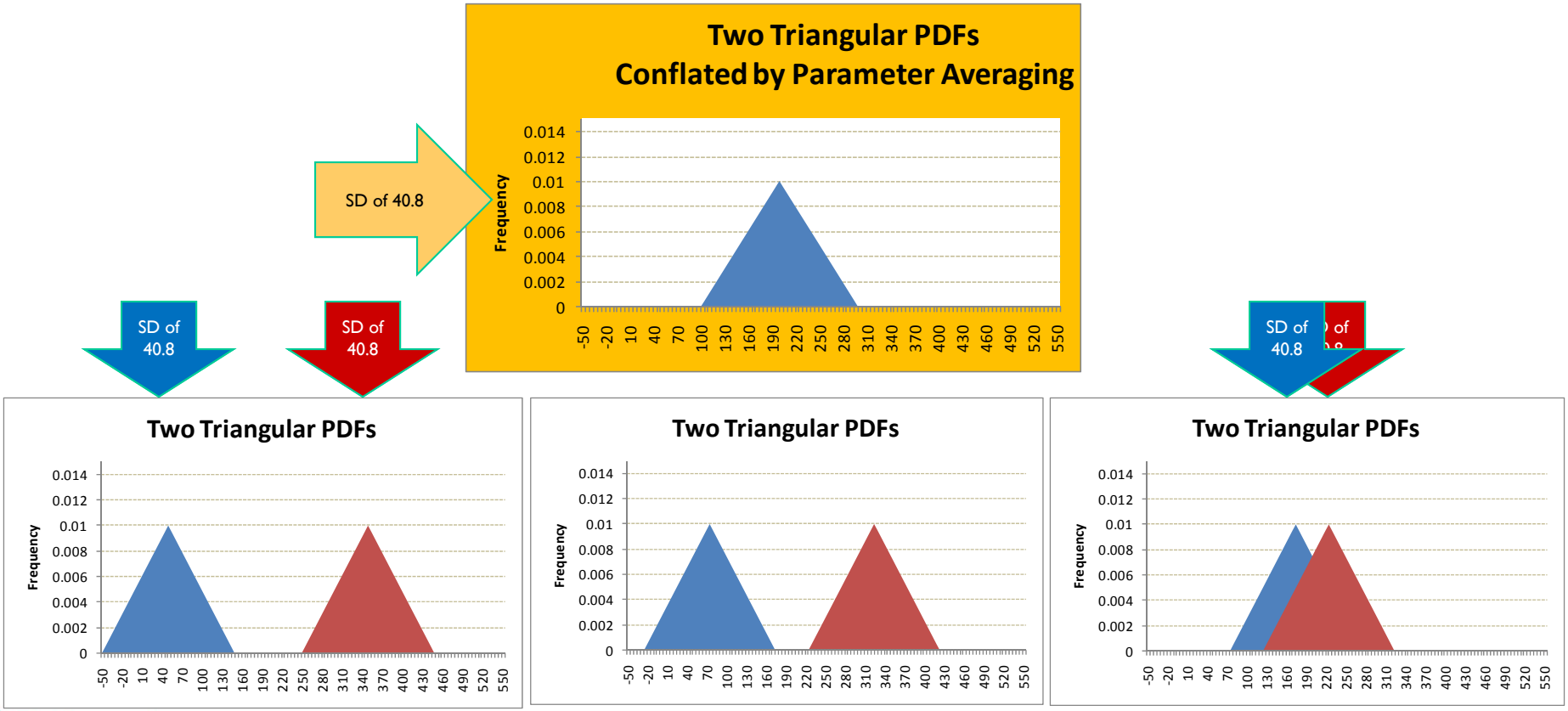


Conflation: Averaging on Each Iteration (1a)

- We conclude that averaging on each distribution has some good and bad characteristics, but on the whole is not desirable
- It produces a good confidence interval for the mean of the experts, but this is not what we want
  - We already know the mean of the experts, the point estimate is the simple average of the means of each
  - What we really want is the full range of the possible outcomes, but averaging on each iteration does not do this, instead it shrinks the answer
  - By analogy, this is the same problem as the confidence interval for a CER … it bounds the line, but not the data … what we really want is the prediction interval
  - It is only a candidate (and flawed at that) for clear cases of single reality

Conflation: Averaging Parameters (2a)

- Averaging parameters provides simple results: Three very different sets of testimony by two experts produces exactly the same picture
- The standard deviation of k identical but scattered triangles, with standard deviation of s, when iteration-averaged will produce a standard deviation of s
  - The SD of the conflation will not vary with k



Conflation: Averaging Parameters (2a)

- We conclude that averaging parameters has some good and bad characteristics, but on the whole is simple and wieldy
  - It produces good estimates of the mean and the standard deviation
  - It is insensitive to scatter of expert testimony, so is only useable in clear cases of single reality
  - Correct the parameters as shown earlier because each expert is likely to truncate
    - The order of the operations does not matter

Conflation: Sampling (3)

- "Average" the probability distributions of the k experts, using one of two schemes, depending on the speed implications and the ease of implementation in your model:
  1. Put all the distributions in the mix, and scale each by 1/k, creating a (probably) multi-mode custom distribution[9]
     - We will see this pictorially on the next slide
  2. Characterize each of the k distributions and choose a first random number to select which expert distribution to use for each run of the Monte Carlo and a second random number to draw from that expert's distribution[10]
  - The two above methods are mathematically identical
- The resulting distribution will have two characteristics:
  - A better estimate of the mean and generally better predictive performance than other conflation schemes[9]
  - A wider (actually, "not narrower") standard deviation for the conflated result than those of the original individual distributions
    - We don't know the degree to which sampling will correct dispersion, although the more experts the wider the dispersion
    - We plan to attempt a study of this
  - We will give a demonstration of this effect with representative data

9. *An experiment in Probabilistic Forecasting,* Thomas A. Brown, R-944-ARPA, July 1973
10. "Determining the Cost of the Certification and Accreditation Process using Expert Opinion and Monte Carlo Simulation," A.J. Flynn, B.J. Nethery, K. Thomas, A.E. Gerstner, B.D. Dickey, C.M. Kanick, P.J. Braxton, SCEA, 2010

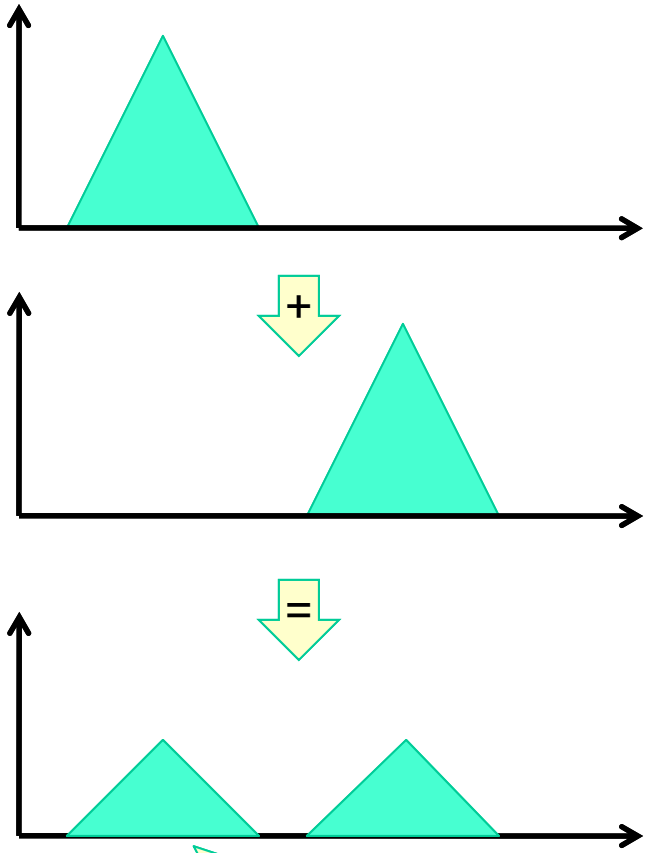
Conflation: Sampling (3)

- To conflate two triangular distributions, "average" them

The first distribution

The second distribution

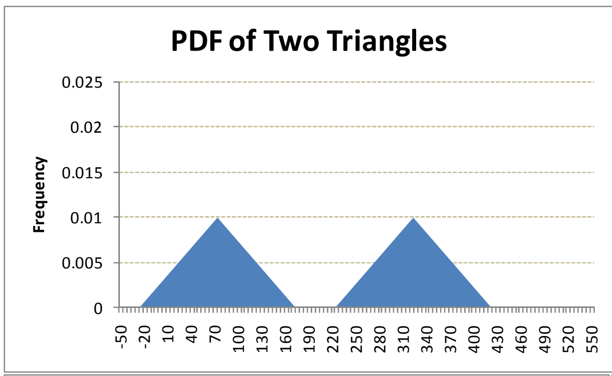
The conflated (averaged) distribution

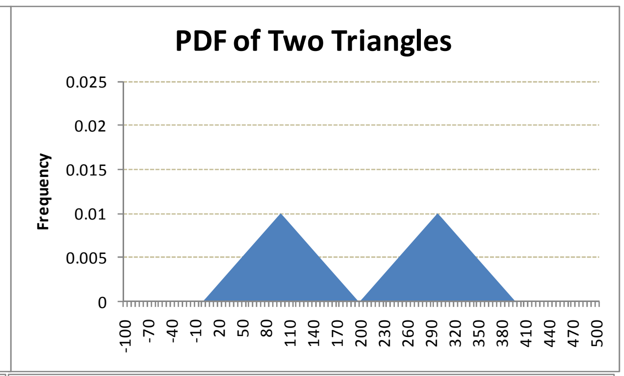


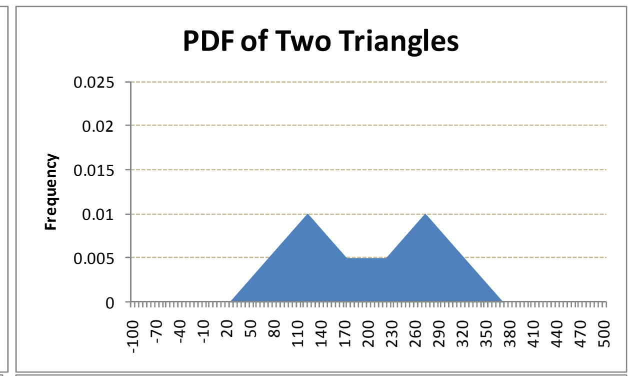
Each triangle has area A = 0.5, or more generally, A = 1/k

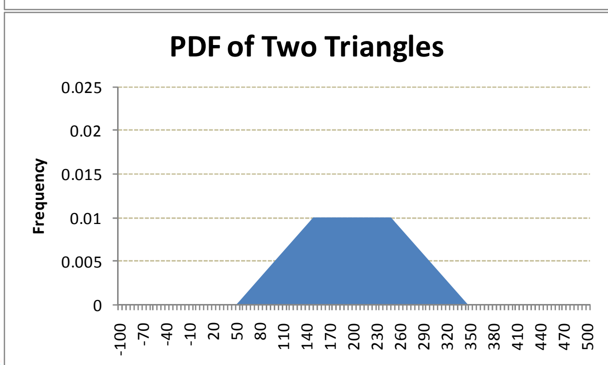
Sampling of Two Triangles - PDF

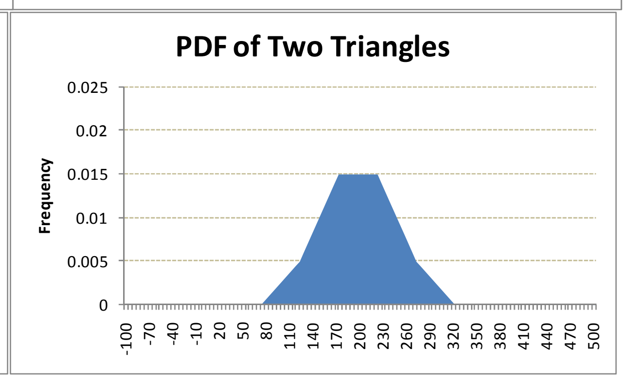
- These charts portray the conflation of two triangles as the respective experts who estimated them come into alignment
  - Each original individual triangle is symmetric, has a base length of 200, and a standard deviation of 40.8
  - Conflation is done by averaging the two PDFs (also described as sampling)
- The two triangles move closer in such a way that the conflated mean remains constant
  - We maintained the same conflated mean of 200
  - We kept the conflated mean constant to allow us to discuss the CV in a meaningful way
  - When the two triangles merge, we get a triangle that has the height and width of each individual triangle before conflation
- The standard deviation of the conflated distribution will be shown on the next graph

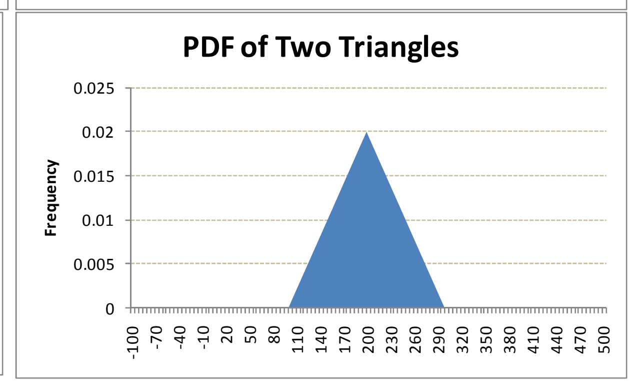








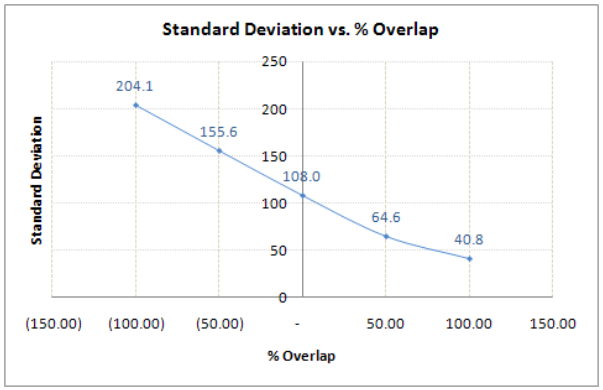


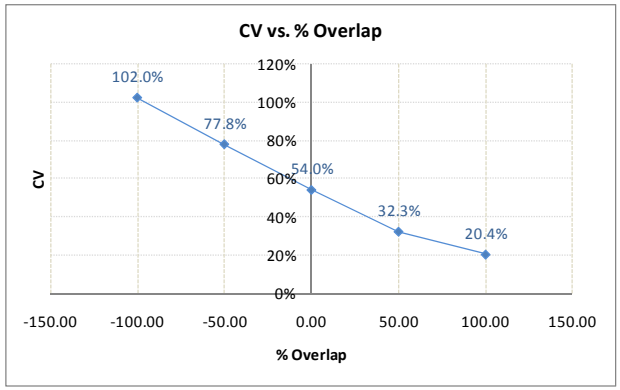


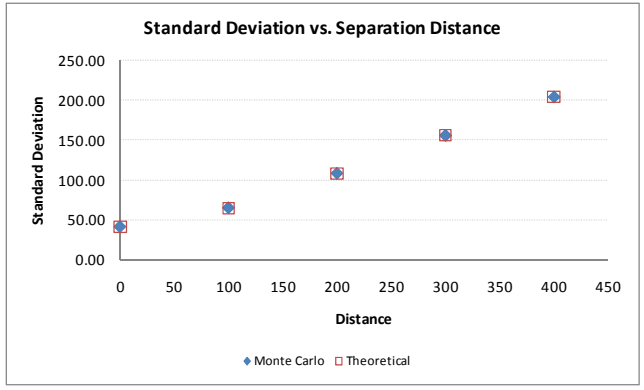
Conflation of Two Triangles - CV and SD

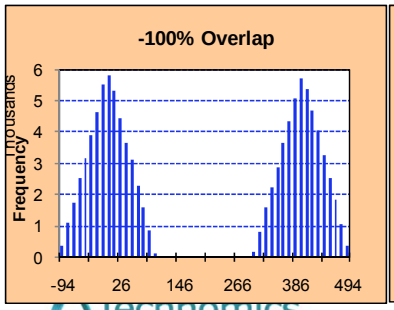
- As two triangular PDFs move closer, the conflated standard deviation and CV drop until the triangles merge and achieve the same standard deviation as that of each triangle
  - Since we chose to maintain the mean of the conflation at 200, the CV drops
- The unsettling conclusion is that the CV of conflated expert opinion can be uncontrollably large, depending on how far apart their triangles
- The standard deviation of two identical triangles separated by distance 2d can be shown* to be $\sqrt{(\sigma^2+d^2)}$

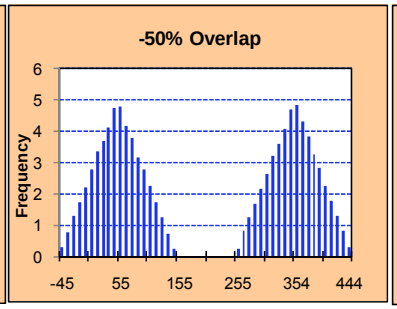
**We aren't saying it's easy … this phrase usually means the Professor is too lazy to show you or too kind to bore you, and the former is by far the more likely! We're the latter, the proof is in backup*

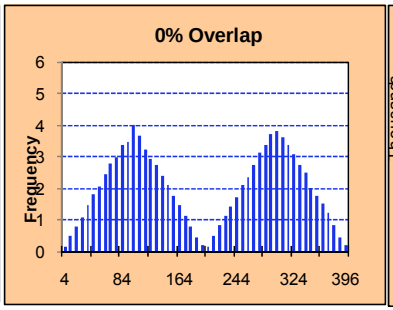


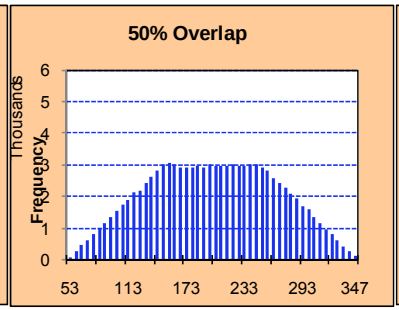


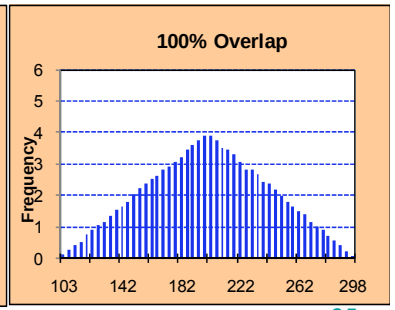












The Dispersion of Sampled Distributions

- Let:
  - σ = SD of the underlying risk
  - S_e = SD for the individual experts (we think it is about $\sigma/2.5$)
  - S_m = SD for the meta distribution of the experts opinions
  - S_c = SD of the conflation
- Then,
  - if $S_e = 0$, then $S_c = S_m$
  - if $S_m = 0$, then $S_c = S_e$
- And, further
  - $S_c \geq \max(S_e, S_m)$
  - This also implies that if S_e is corrected to σ, S_c exceeds σ
- We have shown, in backup, that once the experts have produced k triangles, then:

$$S_c = \sqrt{(S_e^2 + S_t^2)}$$

where S_t is the calculated standard deviation of the means of the k triangles from their means.
We have yet to prove that:

$$S_c = \sqrt{(S_e^2 + S_m^2)}$$

But we believe it to be true

Thoughts on the Distribution of Expert Opinion

- Assumptions:
  1. Experts will not be versed in the distribution of costs, but will be estimating the distribution based on the outcomes they have experienced and perhaps some hearsay
  2. Experts are most likely to be technical people, not cost estimators, so will have experience in a handful of projects and hearsay of somewhat larger number
- Implications
  1. Experts will perceive a mean (and perhaps the mode?) according to Chebyshev's inequality or a confidence interval bounded by $\sigma/(\sqrt{n})$, at best
     - Where n is the number they have observed
  2. Experts will perceive a standard deviation (and thus perhaps the extrema of a triangle?) as a variance σ times a chi-square (n) divided by n, at best
  3. The above do not yet consider the implications of truncation of the value of σ

Combining Corrections for Extrema and Conflation

- We have shown that individual distributions can be corrected for a consistent pattern of understatement
- We have shown that sampling of multiple experts will improve the mean and widen the spread
  - But we don't have a good sense of how much the spread will be improved
- The implication of the two above statements is that we should not endeavor to both expand and sample expert distributions
  - If we correct the individual distributions, we will have the dispersion "about right", if we then sample them, we will have a dispersion that exceeds "about right"
- So, for "multiple reality," do one or the other but not both
  - Expansion of a single distribution focuses on the dispersion
  - Sampling of diverse experts focuses on getting the mean right
  - Since we generally recommend correcting lower order moments first[11], conflation is the priority

11. "The Manual for Intelligence Community CAIG Independent Cost Risk Estimates," R.L. Coleman, J.R. Summerville, S.S. Gupta, DoDCAS, SCEA, ASC, 2002. [see tenets]

Conflation: Sampling

- Sampling of each distribution has excellent characteristics
  - It replicates what the experts told us exactly
- It has a problem in use for a single reality situation because the standard deviation is not easily correctible for scatter nor is it useable without correction
  - We can easily correct each expert's testimony for truncation
  - But we cannot undo the growth caused by expert scatter, which is theoretically unbounded … the adjustment may be a function of k, the number of experts, and has yet to be ascertained
- We conclude that, despite its popularity in the literature, the sampling technique is too tricky in a single reality case and should not be used

Recommendation - Multiple Reality

- The mean of the multiple reality is not troublesome, almost any *reasonable* approach will yield the same mean
  - Again that dangerous word "reasonable"!
- The standard deviation does not present as much of a problem in a multiple reality case because we believe each expert, like the six blind men with the elephant[12], sees a piece of the truth
- Use sampling, do not correct distributions lest you overstate variance as a result
  - If you were to correct, it would have to be before sampling, you cannot easily correct it afterwards, order matters

12. "The Blind Men and the Elephant," John Godfrey Saxe, http://en.wikipedia.org/wiki/Blind_men_and_an_elephant#John_Godfrey_Saxe.

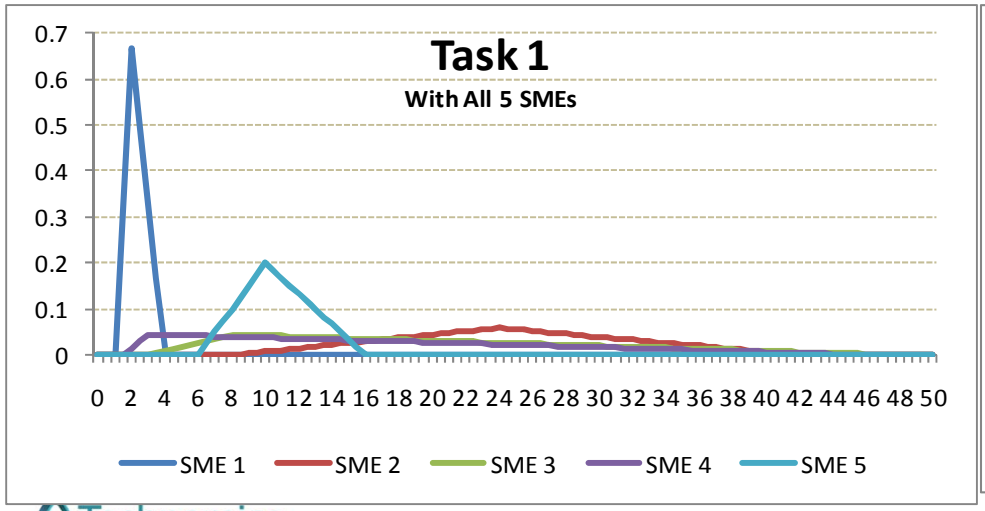
An Actual Case Study

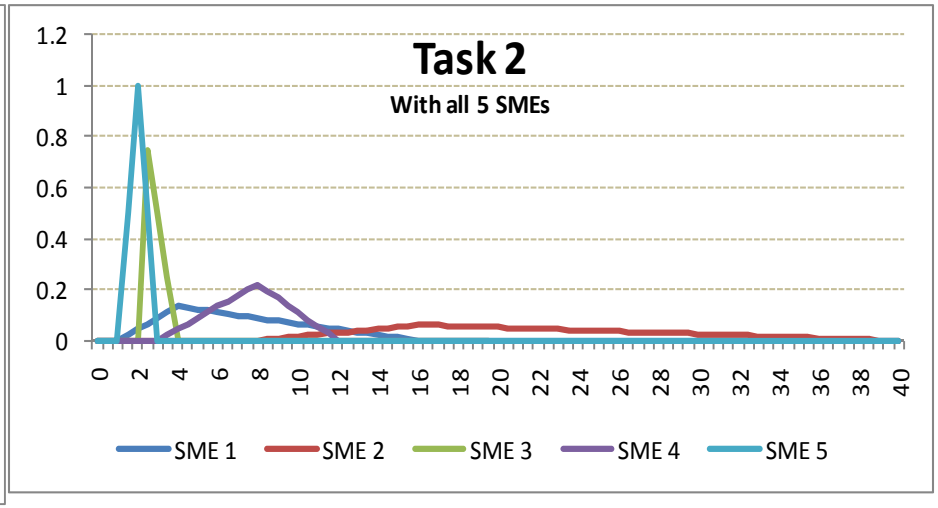
The SME Data

- Actual SME data was collected on a number of subtasks
- Each SME was providing estimates of the same tasks without collaboration
- The data, while not strictly pathological, was sufficiently different to provide a good test of our findings
- Our paper was written for this study, but our methodology development was divorced from the data until the end
- The data source is sufficiently obscured, by a single linear transformation, to prevent traceback

The Original Data

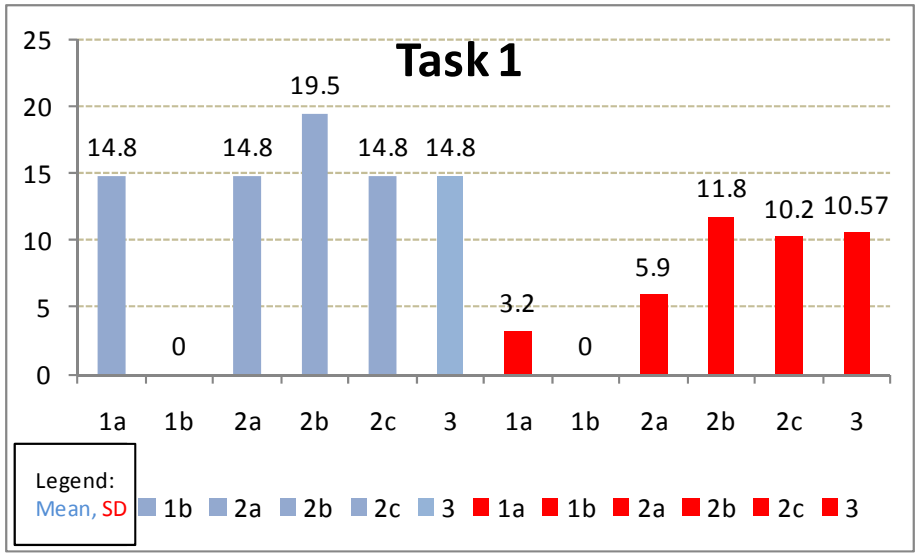
- The transformed source data shows a dispersion of opinion
- It was unclear whether this was a case of multiple reality
  - The study authors concluded that it might be, so they chose sampling
- We will compute the results from all the methods we examined and plot the results of the two methods we selected

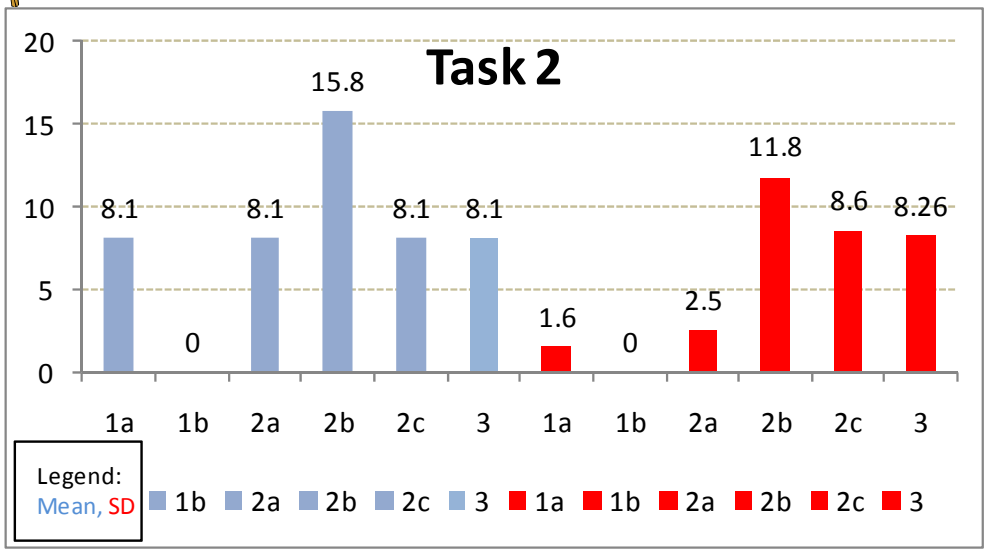




Moments of the Postulated Methods

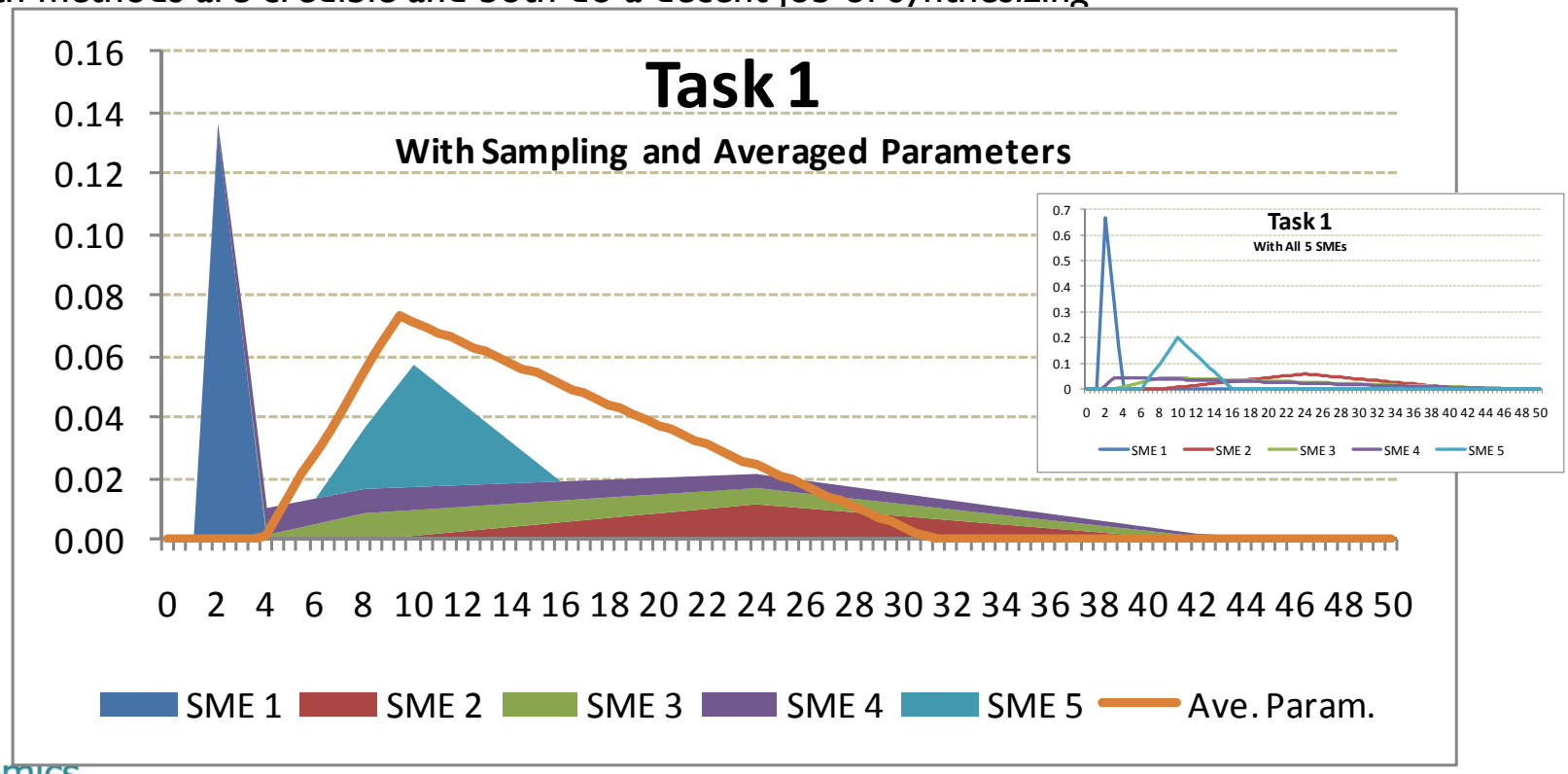
- Methods recap
  - 1a — Average the results of each SME on each run
  - 1b — Same as 1a with correlation = 1.0 (same as 2a below for symmetric, a bit different for skewed)
  - 2a — Average the parameters of the SMEs (use the average of the means *or* the average of the modes)
  - 2b — Min of the mins, average of the modes and max of the maxes
  - 2c — Min of the mins, average of the means and max of the maxes
  - 3 — Sampling (equivalent to averaging PDFs)
- As we expected, the means are all almost all the same
  - Method 2b used averaged modes, so the mean is not preserved
  - Method 2c, an attempt to salvage 2b, used average means but routinely returned modes below the min so was unusable
- As we expected, the standard deviation is the parameter that responds to our choices
  - SD of 1a was "too small"
  - SD of 2b, the rejected 2c and 3 were "too big"
  - The SD of 2a was "Goldilocks"





Graphs of the Two Recommended Methods

- The "Sand Chart" shows sampling, the preferred method for multiple realities
  - It retains all the information told to us by the SMEs equally
  - It suggests, in this example, that there may be three different modes, representing 3 different possibilities
- The "Line Chart" shows averaged parameters, the preferred method for single reality
  - It responds to all SMEs, but produces a uni-modal, less dispersed solution
  - It suggests, in this example that SME 1 was too low while SMEs 3 and 4 were a bit pessimistic on the high end
- Both methods are credible and both do a decent job of synthesizing



Conclusion for the Conflation of Experts

- As asserted, we have illustrated that the averaging of parameters for k triangles, is equivalent to averaging of draws from those k triangles with a single draw of a random number used to simulate the expert's draw, and then averaging the draws
- We have demonstrated why those two equivalent methods give the simplest and clearest result for Single Reality and seem the best representation of what the k experts seem to have meant
- We have shown why Sampling of k experts gives the best representation of what the k experts seem to have meant in the case of Multiple Realities
- We presented a case study with actuals that shows that the two recommended approaches do a decent job of synthesizing what the SMEs told us
- The issue of deciding between Single and Multiple Realities remains the most difficult issue
  - Sometimes it will be as simple as learning that each expert has in mind "a different engine"
  - Sometimes it will be a concession to the wide dispersion and the recognition that there "must be a reason"
- We will now move to a different topic, that of correcting mischaracterization of distributions, without which this paper would seem incomplete

Correcting the (Mis)characterization of Distributions

The Problem

- "Experts" who may know a lot about the technical issues, and maybe even the cost of them, will not necessarily be well versed in probability
  - Consequently, the characterizations they will produce will not be easily used and will sometimes be incoherent (meaning, internally contradictory)
- Expert testimony in risk analysis should be accorded the same respect that cost data is in cost analysis
  - Tenet 1: "Do no harm" meaning preserve as much of what the expert said as is possible in achieving coherence
  - Tenet 2: Preserve lower order moments above higher order moments
  - Tenet 3: If particular aspects are more important than others, preserve those aspects (e.g., if the variability or upper percentiles are the focus, accord those greater priority)
- It is preferable to make the corrections with direct feedback to the expert, but this feedback should be done under the same precepts as the corrections
  - Meaning, follow the tenets in your persuasions and probing

Implausible Percentiles

- "The 20/50/80 are $0.0M/$0.9M/$3.6M"
  - No triangle can fit this, and the distribution is wildly skewed, so simplifying steps were taken:
  - Assume that the stated "50%-ile" is really the mode
  - Take the 20 and 80 as "about true", and assume they are ±σ. Use the rule that the half-base lengths of a symmetric triangle are $\sqrt{6}*\sigma$. Note that these triangles are not symmetrical, but use it as a factor that probably does a decent job
  - Results:

Input	Output	
20%-ile 0	L	-1.305
50%-ile 0.9	M	0.900
80%-ile 3.6	H	7.514

  - Note that the correction may be distorting the central tendency
  - But, this distribution is clearly intended to be skewed, and the mean is therefore above the median
  - We cannot actually compute the mean with the information given
  - We also knew that in this analysis, the ROS at the 80th percentile was a particular focus, so we felt that preservation of that point should take priority (Tenet 3)

Unlikely distributions

- Risk values:
  - 20% probability of -$2M
  - 40% probability of $0
  - 20% probability of +$4M
- Suspecting that this was a just clumsy way to characterize a triangle, we asked if a triangle with the below characteristics was along the lines of what the expert meant:
  - 20%-ile -$2M
  - Mode 0M
  - 80th %-ile +$4M

  … the expert agreed readily that the precise distribution wasn't what he meant, and the triangle captured the sense of it.

Errors Of Characterization Induced by the Risk Analyst

- Categorical* risk distributions
  - Many risk models cannot easily (or rather obviously) implement a categorical random variable beyond a Bernoulli
    - Many can do it, most analysts don't realize they can
  - For a 3-value categorical, with choices of 0, a and b, many analysts implement it as two independent Bernoullis with values of 0 or a and 0 or b
  - This results in an error as the results are not the same … the two Bernoullis can turn out as a and b at the same time, but the original formulation prohibits that
  - Either implement it as a categorical or create two Bernouli's with the right characteristics
- Triangular risk distributions
  - Sometimes the end points are set at the standard deviation of the formulation
  - Sometimes triangles are used instead of normals, even when the normal was proposed, out of aversion to negative outcomes
    - In practice, negative outcomes are harmless in Monte Carlo
    - Negative outcomes ought to be fairly rare anyway
- Normals
  - Sometimes triangles are substituted incorrectly (see above)
    - If the mean and standard deviation are captured correctly there is little harm
  - Sometimes the negative portion of the normal is truncated despite that this causes a shift of the formulated mean and a reduction in the standard deviation

* Categorical risk distributions are like Bernoullis but allow 2 or more values (the Bernoulli is a member of the family)

Conclusion for Correcting Mischaracterization of Distributions

- We have presented tenets by which apparent errors of characterization may be corrected and have listed the most common Risk-Analyst-induced errors
- We finish by reiterating that the testimony of the experts we consult should be handled much as we should handle data
  - We must be careful in not ignoring the symptoms of the testimony, and avoid such elementary errors as causing anchoring and "leading the witness."
  - We should, nonetheless carefully repair any clear errors caused by the unfamiliarity with probability that can result in unlikely distributions

Conflation of SMEs – Summary Thoughts

- The conflation of expert testimony has received some attention in the literature, but little to none of the conclusions seem to have permeated the cost risk discipline
- We hope that we have provided a reasonably thorough paper by which risk analysts might be guided
- We also hope that we have provided a few good tenets for correcting mischaracterization, along with some illustrative (actual) examples.
- We hope to be able to take on the issue of what we call the meta-distribution, the likely distribution of individual expert testimony
  - Without a good model for the meta-distribution, the full demonstration of the best answers will remain incomplete, because the meta-distribution is the unseen ground truth against which these answers can be measured
  - Until we can be satisfied we have the meta-distribution, we are confined to showing the behavior of various methods and deciding if that behavior seems correct

Expert Survey Purpose

- Two surveys were conducted to explore SME risk assessment phenomena noted in the literature
  - Pilot survey conducted at author's church – laymen (literally!)
  - Full-fledged survey conducted at author's employer – analysts
- Questions were crafted so as to shed light on various (*a priori*) hypotheses, via both:
  - Descriptive statistics (summary numbers and graphics), and
  - Inferential statistics (hypothesis tests)
- Questions designed more to test risk assessment skills than subject matter knowledge
  - Mix of general topics and areas where some or most respondents might reasonably be considered "expert"
  - With rare exceptions, respondents were expected *not* to know the answer, but rather to "have an idea"

Expert Survey – The Dry Run

Survey Overview

- 25 respondents connected with the author's church and/or church musical (including the author!)
  - Ranging in age from 17 to 79, 13 male, 12 female
  - One reneged (the author's own mother!), two missed back page
- 34 questions, each asking for a low and high value
  - Confidence level not specified beyond "reasonably sure"
  - Did not ask for Most Likely
    - Assumed symmetric with mode = mean = median = average of low and high
  - Not your typical trivia quiz
    - Respondents universally reported the survey made them "feel stupid"
  - Across 6 Categories (4-7 questions per category)
    - Respondents asked to rate their expertise 1-5 in each Category
  - Including 9 pairs of similar past/future questions
    - What was Tiger Woods' golf score yesterday? and what will it be tomorrow?
- No training / guidance / coaching

Survey Results Terminology

- We shall say a respondent was "successful" or "correct" or scored a "hit" on a given question if his or her interval contained the true value
  - HL = "Hit Low" – true value contained between respondent's Most Likely and High values (i.e., Most Likely was *below* true value)
  - HH = "Hit High" – true value contained between respondent's Low and Most Likely values (i.e., Most Likely was *above* true value)
  - HE = "Hit Exact" – true value equal to respondent's Most Likely
- We shall say a respondent was "unsuccessful" or "incorrect" or "missed" on a given question if his or her interval did *not* contain the true value
  - ML = "Missed Low" – true value above respondent's High value (i.e., respondent's entire interval was *below* true value)
  - MH = "Missed High" – true value below respondent's Low value (i.e., respondent's entire interval was *above* true value)
- We shall say a respondent was "sure" if his or her Low and High values were equal
  - SR = "Sure Right" – true value equal to respondent's Low = Most Likely = High
  - SW = "Sure Wrong" – true value not equal to respondent's Low = Most Likely = High

Survey Hypotheses

- Respondents will be "correct" no more than about 2/3 of the time
  - 68.3% for plus or minus one standard deviation of normal
  - 60% for 20/80
- Respondents will do about equally well gauging past events as predicting future events
- As a group, respondents will be unbiased
  - Tend to estimate at a mean and/or median
- The distribution created by averaging the parameters of the individual distributions (Method 2a) will perform better than individual distributions
  - "Wisdom of the Crowd"
- Respondents who rate themselves expert (4 or 5) in a Category will be correct more often and/or will have narrower intervals
- Respondents will do better at Categories they "should" know better (e.g., Music, Church), independent of self-assessment
- For some questions, a combination of ignorance (don't play golf) and innumeracy (can't estimate TV viewers as a proportion of U.S. population) will cause a "discontinuity" where providing reasonable ranges is very difficult
- The spread of average responses across respondents ("peak-to-peak") will be comparable to the average (corrected or uncorrected) low-to-high spread ("tip-to-tip")
- Distribution of responses will be comparable across questions ("pseudo-iid")

Survey Normalization

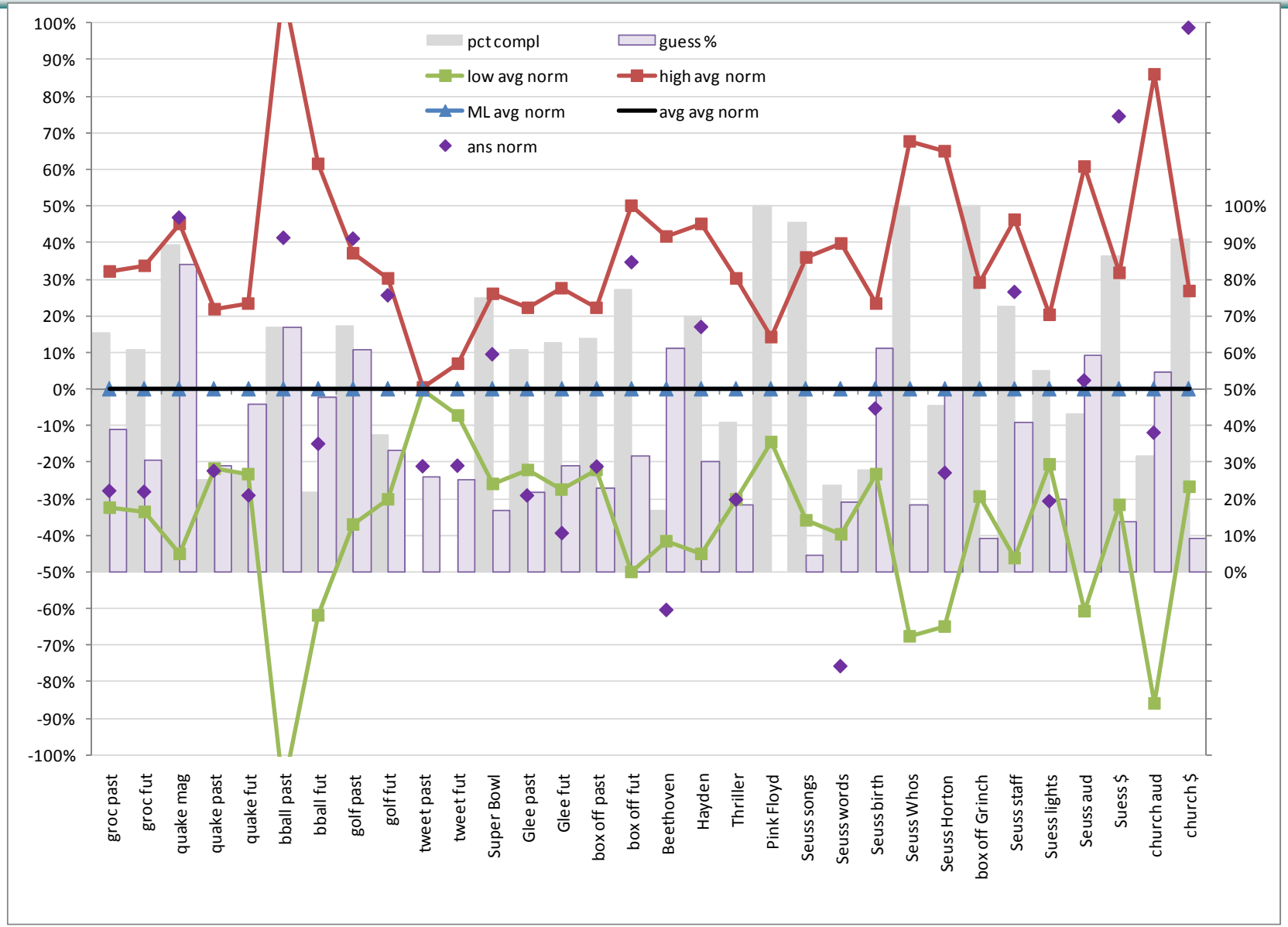
- Since the "correct answers" are point values and not distributions, the only available measure of dispersion comes from the respondents
- For each question, the standard deviation of average value across respondents was used as a proxy
  - For knowable but unknown values, this represents the dispersion of responses (the "meta-distribution" or "peak-to-peak"* measure)
  - For unknowable (future) values, it is also a proxy for the true standard deviation, though it may differ by an unknown factor
- The "baseline" was set as the mean of the average value across respondents
- A Z-score-like normalized response was then calculated by subtracting this mean and dividing by this standard deviation
  - Ex: Super bowl viewers, mean = 96.7, std dev = 150.6, so a response of 150 would have a normalized value of (150-96.7)/150.6 = 0.35
- Normalized values only used where necessitated by cross-question comparisons using spread or distance measures
  - Facilitated "apples-to-apples" comparison between questions with different scales (e.g., millions of Super Bowl viewers vs. tenths of a point on the Richter scale)

* "Peak-to-peak" evokes the modes of the triangle, and while we are using the means, the sense is the same: how spread out the triangles are from each other, as opposed to the spread of individual triangles

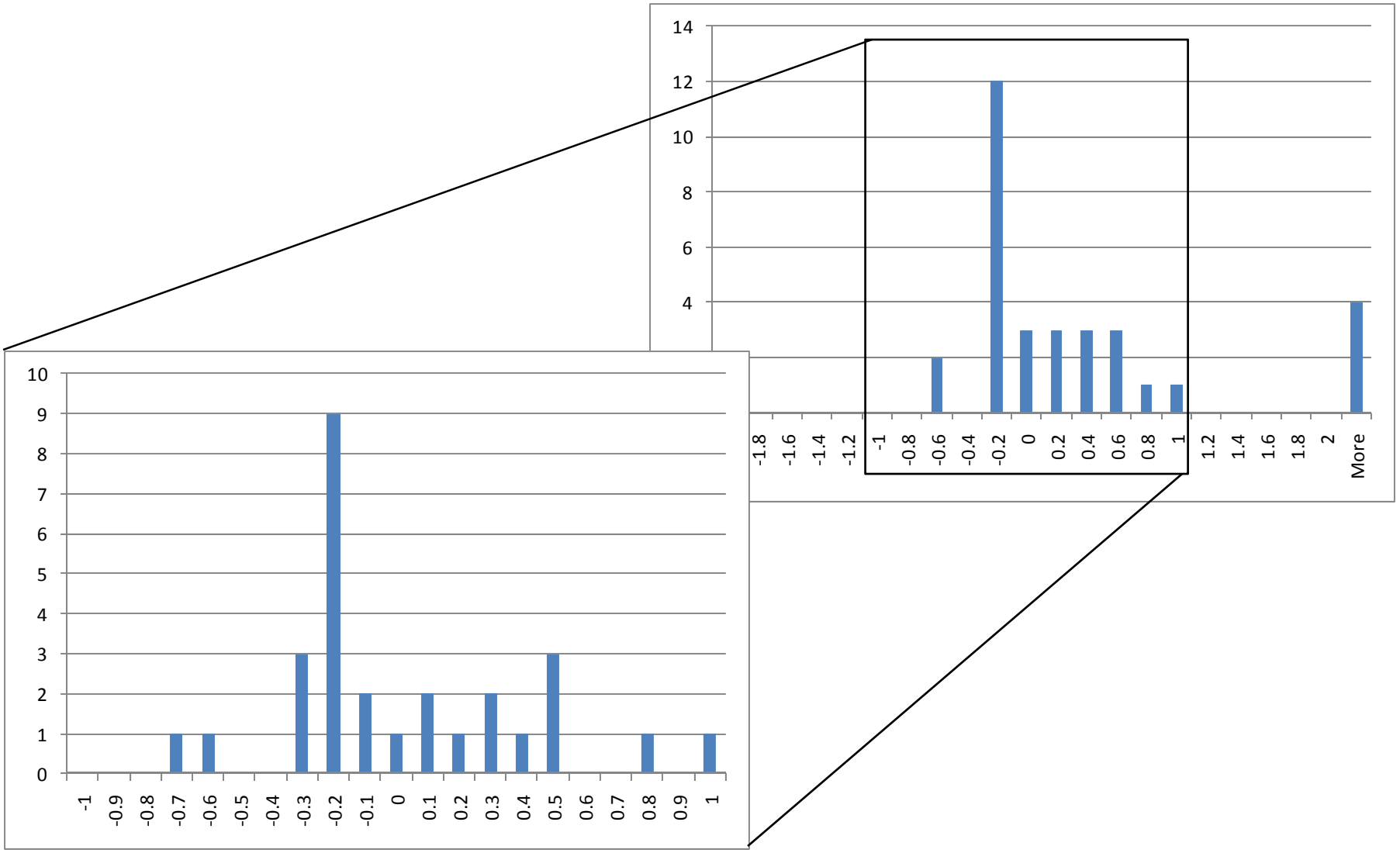
Survey Results – With Outliers

- Only 5/24 respondents had > 50% success rate (including the author)
  - Only 1/24 > 66% (71%, missed statistically-significantly-harder 2nd page, so likely biased)
  - Overall success rate was 33%
    - Consistent with 33/67 interval (when asked for quartile)
- Past (38%) vs. future (34%) not statistically significant
  - No appreciable difference in interval widths
- On average, true answer (normalized) was 0.39 higher than overall average
  - Standard deviation of correct answers (normalized) was 1.26
- Collectively, respondents were at the 53rd percentile
  - Skewed responses, underestimates were more severe than overestimates (analogous to "overruns are more severe than underruns")
- Parameter-averaged distribution was correct 47% of the time
  - Split almost evenly between Miss High, Hit High, Hit Low, and Miss Low
- No correlation between self-reported expertise and success, except for Sports and Church (see graph)
- Predictions ("tip-to-tip") were mostly in the 0.2-0.5 normalized range (see graphs)
  - Average standard deviation about 0.15
- Confidence and accuracy often inversely related (see graphs)
  - Attendance predictions in 0.6-0.9 range but accurate within 0.15
  - Donations predictions about 0.30, but off by at least 0.70!

Average Responses by Question

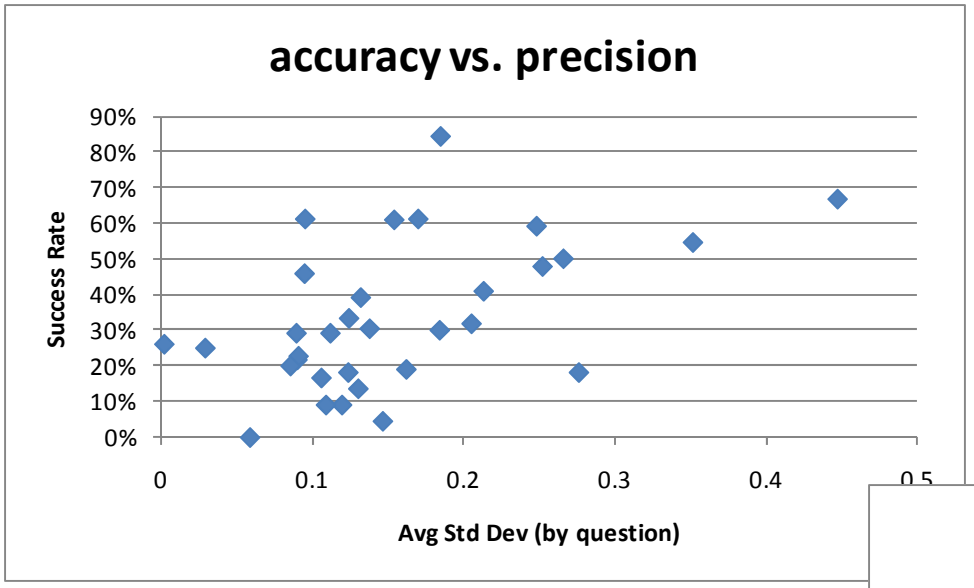


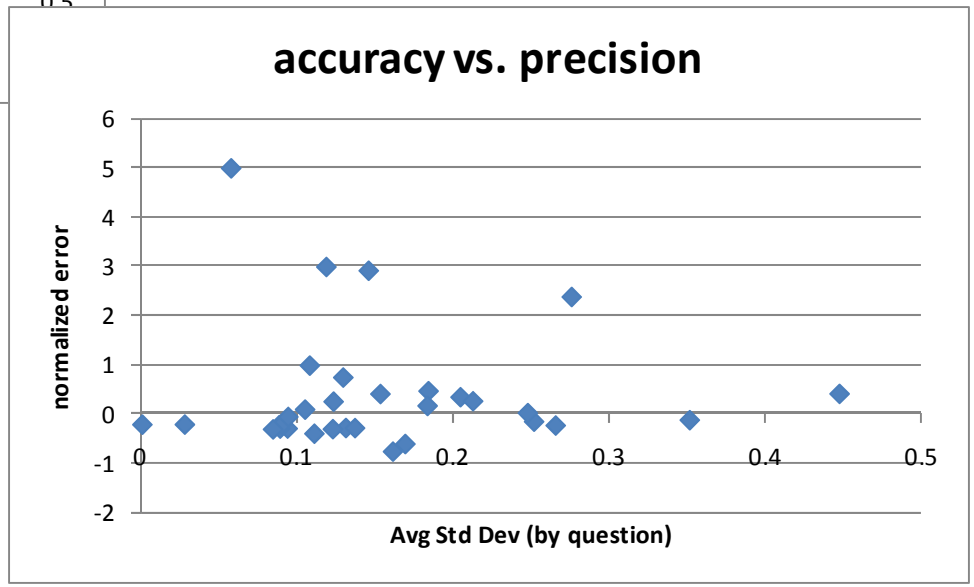
Distribution of Average Error



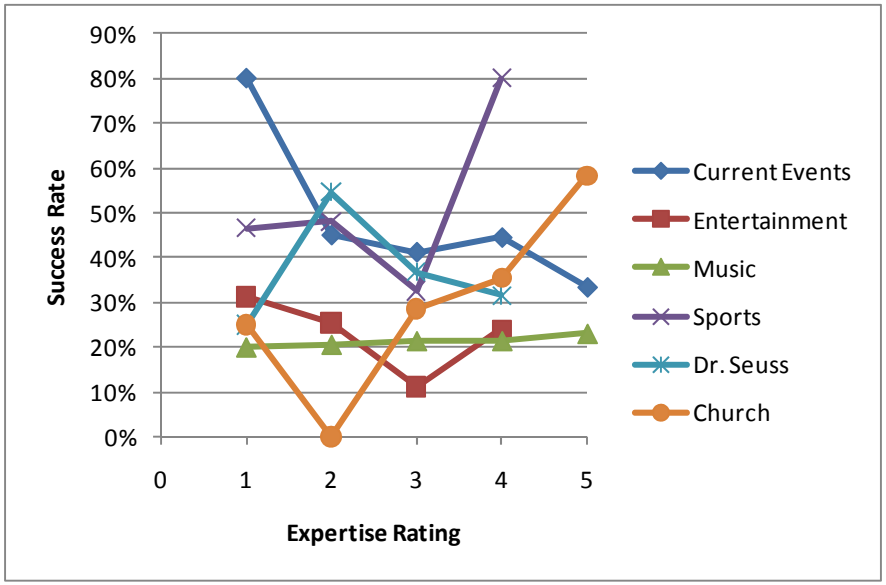
Accuracy and Precision

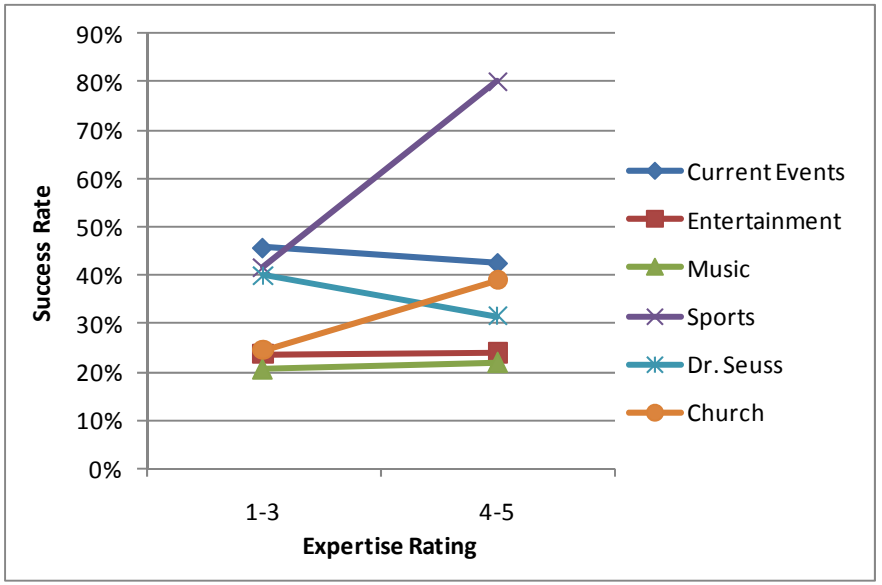
- Compare width of respondents' intervals to two measures of accuracy

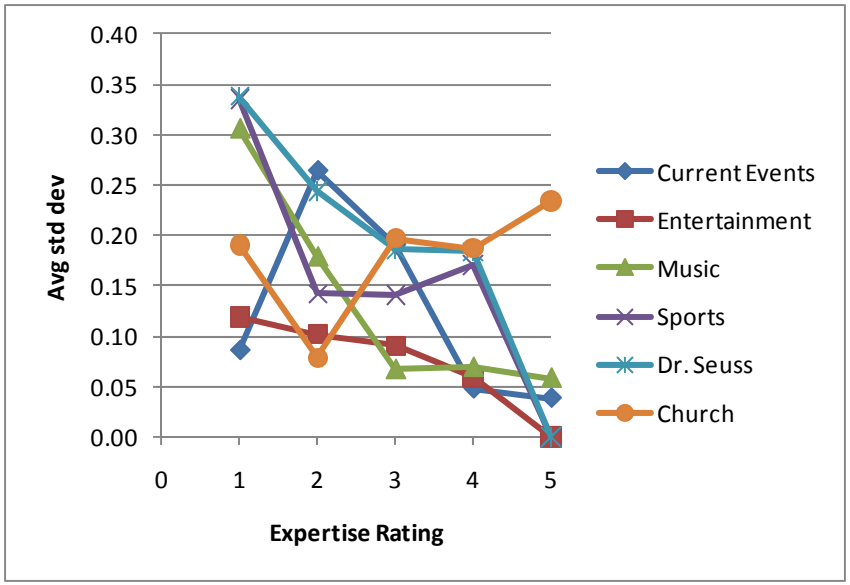


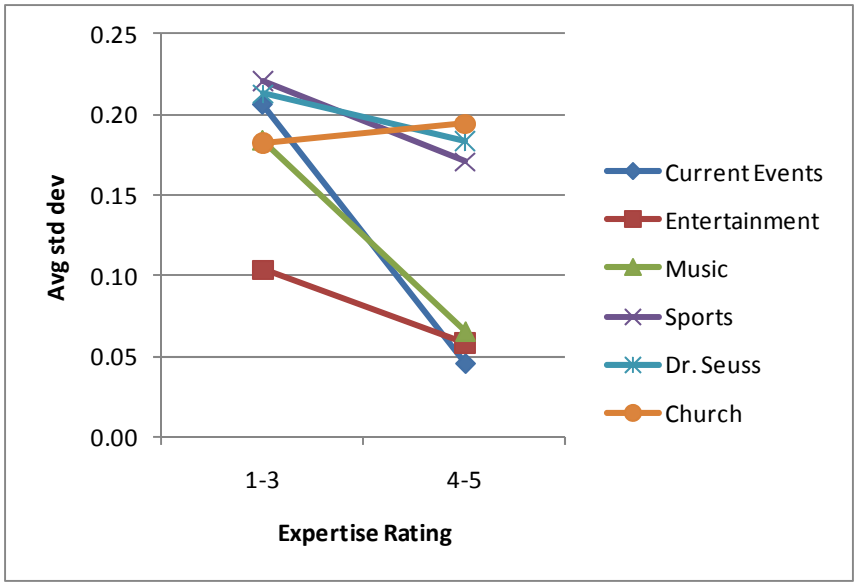


Effects of (Self-Reported) Expertise









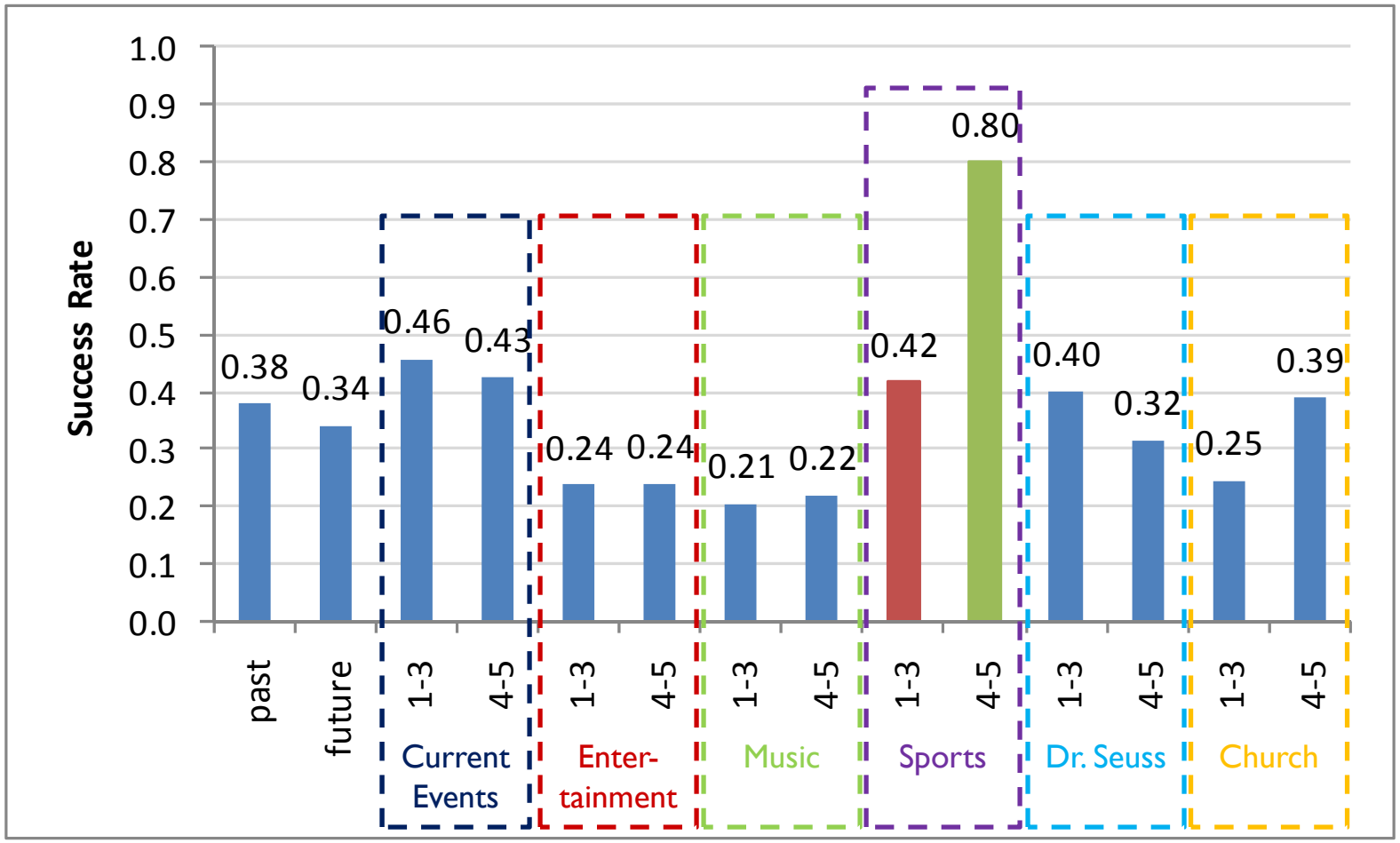
Effects of (Self-Reported) Expertise

- It turns out that most of the time "experts" are more accurate (i.e., higher success rate) or more precise (i.e., narrower intervals), but not both
- Unfortunately, no way to tell which case we have
  - Ideally, we'd like to fix accuracy (confidence level) and let the chips fall where they may on precision

	narrower interval?	more accurate?	both?
Current Events	Y	N	N
Entertainment	Y	N	N
Music	Y	N	N
Sports	N	Y	N
Dr. Seuss	N	N	N
Church	N	Y	N

Success Rates Comparison

- Binomial test conducted for statistical significance
  - Assume normal approximation, np > 5, nq > 5



Expert Survey – The Real Thing

Survey Overview

- 29 respondents, all colleagues at the author's company
  - Ranging in age from 22 to 63, 20 male, 9 female
  - 21 attended brown bag and received "training," 8 did not attend
- 51 questions, each asking for a low, most likely, and high values
  - Part 1, 17 questions
    - Confidence level not specified beyond "reasonably sure"
    - Respondents asked afterward what kind of interval they *thought* they were providing
    - Surveys immediately graded and feedback provided to brown bag attendees
      - Necessitated no "future" questions on Part 1
  - Part 2, 34 questions
  - Across 8 Categories (3-7 questions per category, except 14 for Curr Evt)
    - Respondents asked to rate their expertise 1-5 in each Category
  - Including 9 pairs of similar past/future questions
    - What was Tiger Woods' golf score yesterday? and what will it be tomorrow?

Differences from pilot survey noted in red

Survey Hypotheses

- Respondents will be "correct" no more than about 2/3 of the time
- Respondents will do about equally well gauging past events as predicting future events
- As a group, respondents will be unbiased
  - Tend to estimate at a mean and/or median

Differences from pilot survey noted in red

- The parameter-averaged distribution (Method 2a) will perform better than individual distributions
- Respondents who rate themselves expert (4 or 5) in a Category will be correct more often and/or will have narrower intervals
  - More significantly, respondents who rate themselves expert in Risk Assessment will be correct more often (may actually have *wider* intervals, depending on subject-matter knowledge)
- Respondents will do better at Categories they "should" know better (e.g., Company, Weapon Systems, SCEA), independent of self-assessment
- Since respondents are analysts who work with numbers for a living, innumeracy will be less of an issue
  - Since questions cover a broad range of subject matter, ignorance will still cause difficulty
- Respondents will have a much higher success rate after receiving training
- The spread of average responses across respondents ("peak-to-peak") will be comparable to the average (corrected or uncorrected) low-to-high spread ("tip-to-tip")
- Distribution of responses will be comparable across questions ("pseudo-iid")

Survey Design – Unanticipated Issues

- Anchoring
  - It is well-established that respondents will "anchor" to provided values
    - "Is Mt. McKinley taller or shorter than 50,000 feet?" vs.
    - "Is Mt. McKinley taller or shorter than 5,000 feet?"
  - It was not unanticipated that respondents might anchor to their own responses to previous questions
    - For example, first three questions were Mt. McKinley (20,320 ft), Mt. Kosciuszko (7,310 ft), and Pu'u O'o eruption (65 ft)
      - (*A posteriori*) hypothesis that respondents may have overstated the last two due to anchoring on their first answer
- Risk Assessment Training vs. Subject-Matter Training
  - Some past/future pairs were split across Part 1 and Part 2
  - Inadvertently but inevitably, brown bag participants received Subject-Matter Training in addition to Risk Assessment Training
    - For example, if you thought Charlie Sheen tweeted thousands of times a day but found out he tweeted zero times on March 14th, you would probably revise your prediction downward – in fact, trained average = 0, untrained average = 3,756!
    - While bad for statistical analysis, this would not be bad for practical implementation

Survey Examples – Typical Answers

- Difficulty of capturing the correct answer, even when you feel you're giving a "ridiculously" wide range
  - 25-50-100 for symphonies written, missed Beethoven (9) high, Haydn (106) low
  - 10-50-100 column-inches of Libya coverage in *Post*, missed low (112.5)
- Inability to anticipate extreme events
  - 22-51-106 (parameter averaged) for *Dark Side of the Moon*, actually 775 weeks!
  - 782-1,374-2,495 (parameter averaged) for Charlie Sheen tweets, actually zero!
- Unnecessarily wide ranges
  - Dr. Seuss born 1800-1900-2000 (really, he's written dozens of beloved children's books by age 11?!)
- Clear indications of subject matter expertise (narrower ranges)
  - Age of Jeff Beck (61-64-65 vs. 25-50-75)
  - Attendance at Jeff Beck concert (1,000-1,250-1,500 vs. 500-3,000-15,000)
- Overconfidence of SMEs
  - 1610-1668-1698 for Diet of Worms (confusion of 1600s with 16th century?)
- Inconsistency of responses
  - Donations per congregant ranging from $0.40 to $200.00

Survey Examples – Sure Answers

- Recall that a respondent is "sure" if low = most likely = high
  - Respondents on both surveys intuited this without explicit instruction
- Sure Right (SR)
  - Maryland scored 71 points in ACC quarterfinal a week before (the man's a sports genius!)
  - Japanese earthquake 8.9 on Richter scale (happened a couple days earlier)
  - Dr. Seuss born in 1904 (more impressive on company survey – she must have kids who just celebrated International Reading Day on his 107th birthday)
  - Beethoven wrote 9 symphonies (the Ninth is the famous "Ode to Joy")
- Sure Wrong (SW)
  - Japanese earthquake 9.8 on Richter scale (oops! digits reversed)
- Almost Sure
  - 1520-1521-1529 for Diet of Worms (impressive!)

Survey Examples – Extreme Answers

- Half a million members of SCEA (SCEA wishes!)
  - 10,000 Lifetime members, compared to 5,000 total! (likely misunderstanding about meaning of "Lifetime member")
- 2 billion copies of *Thriller* sold (I know Michael Jackson's popular, but one copy for every three people on earth?!)
- 300-500-700 company attendees at DoDCAS (out of a 60-person company?)
  - Must have meant total conference attendees
- Dr. Seuss born between 1287 and 1789 (Are we talking about the same Dr. Seuss here?!)
  - Or 1930-1975-1980 (so he wrote *Horton Hears a Who* almost 30 years before he was born?)
- One million tweets a day (Charlie Sheen's warlock powers aren't that strong!)
  - Must have meant followers
- Diet of Worms held 1950-1970-2010 (Vatican II, maybe!)
- $900M opening weekend for *Sucker Punch* (more than the previous top five put together!)

Survey Examples – Funniest Answers

- Re Charlie Sheen's future tweets: "could be dead" (true risk analyst speaking!)

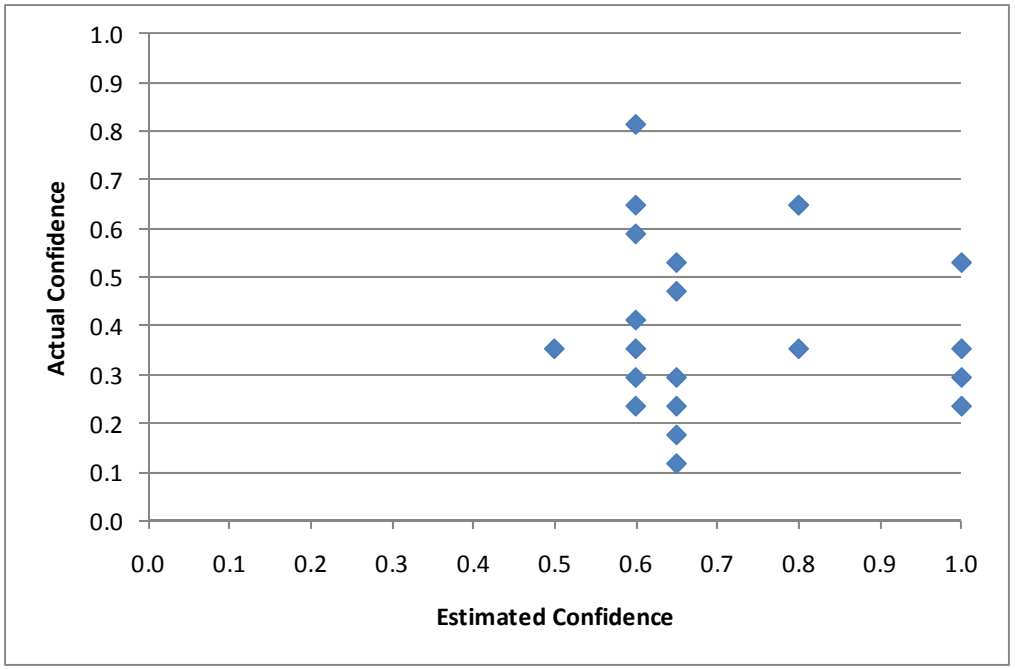
Survey Results (Part 1) – With Outliers

- 9/29 respondents had > 50% success rate
  - Only 2/29 > 66% (71% and 81%)
  - Overall success rate was 41%
    - Slightly better than 33/67 interval (when asked for quartiles)
- Collectively, respondents were at the 44th percentile
- Parameter-averaged distribution was correct 71% of the time, at 65th percentile
- 42% compared with 30% success (statistically significant) for common questions with pilot survey
- Predictions ("tip-to-tip") were mostly in the 0.3-0.8 normalized range (see graph)
  - Average standard deviation of 0.26 compared with 0.15 (pilot survey)

Differences from pilot survey noted in red

Survey Part 1 – Confidence Assessment

- The overwhelming message is that respondents didn't have a very good idea of what confidence they were actually estimating at
  - Correction factor required to get from actual confidence to estimated confidence is consistent with literature (~2.5)



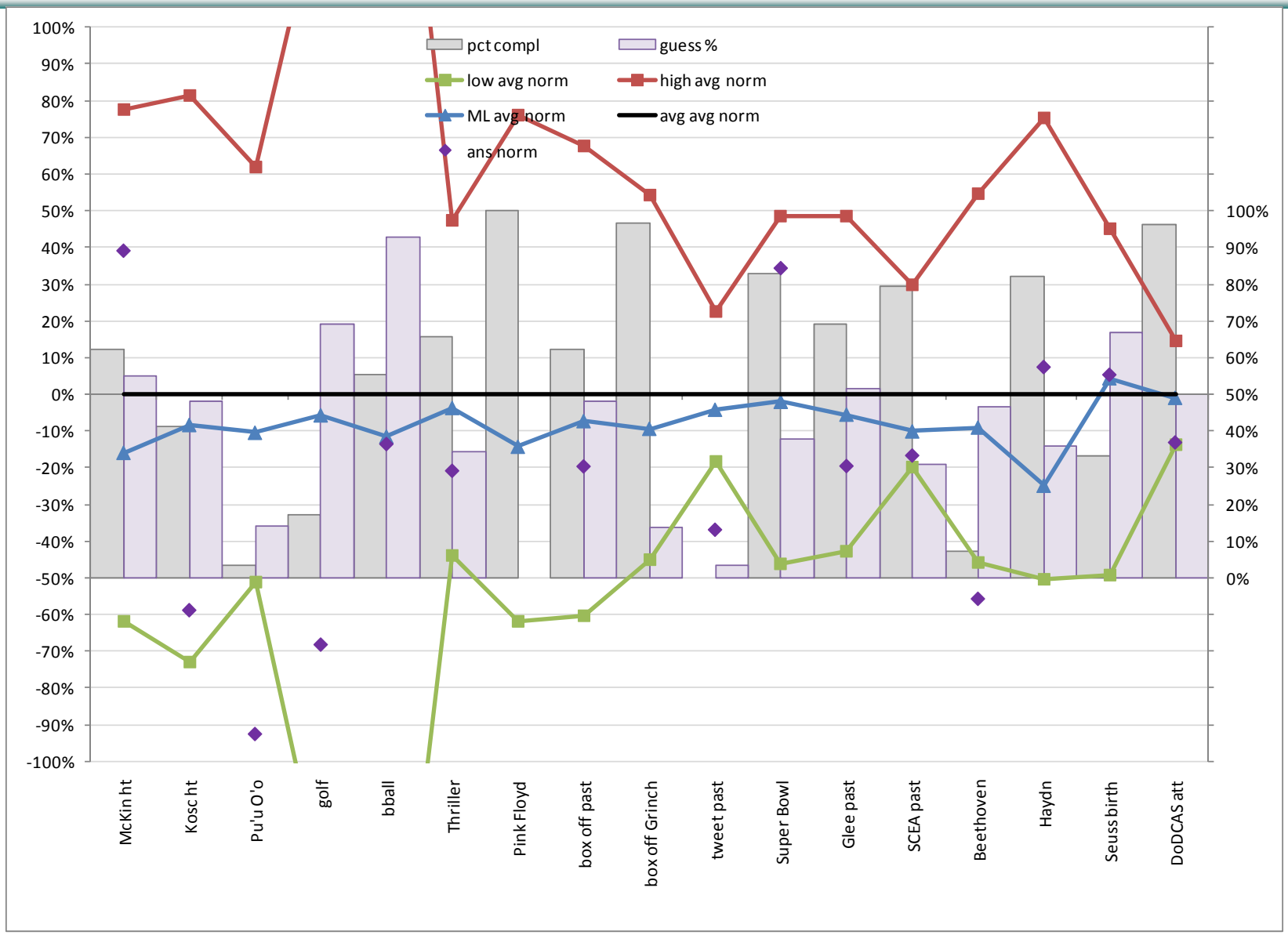
	conf est	conf act	corr est	corr act	corr fact	num
25/75	0.50	0.35	3.41	5.11	1.50	1
20/80	0.60	0.48	2.72	3.61	1.33	7
one sigma	0.65	0.30	2.45	6.04	2.46	6
10/90	0.80	0.55	1.81	3.04	1.68	3
min/max	1.00	0.35	1.00	5.11	5.11	4
				4.58	**2.43**	**21**

Survey Results (Part 2) – With Outliers

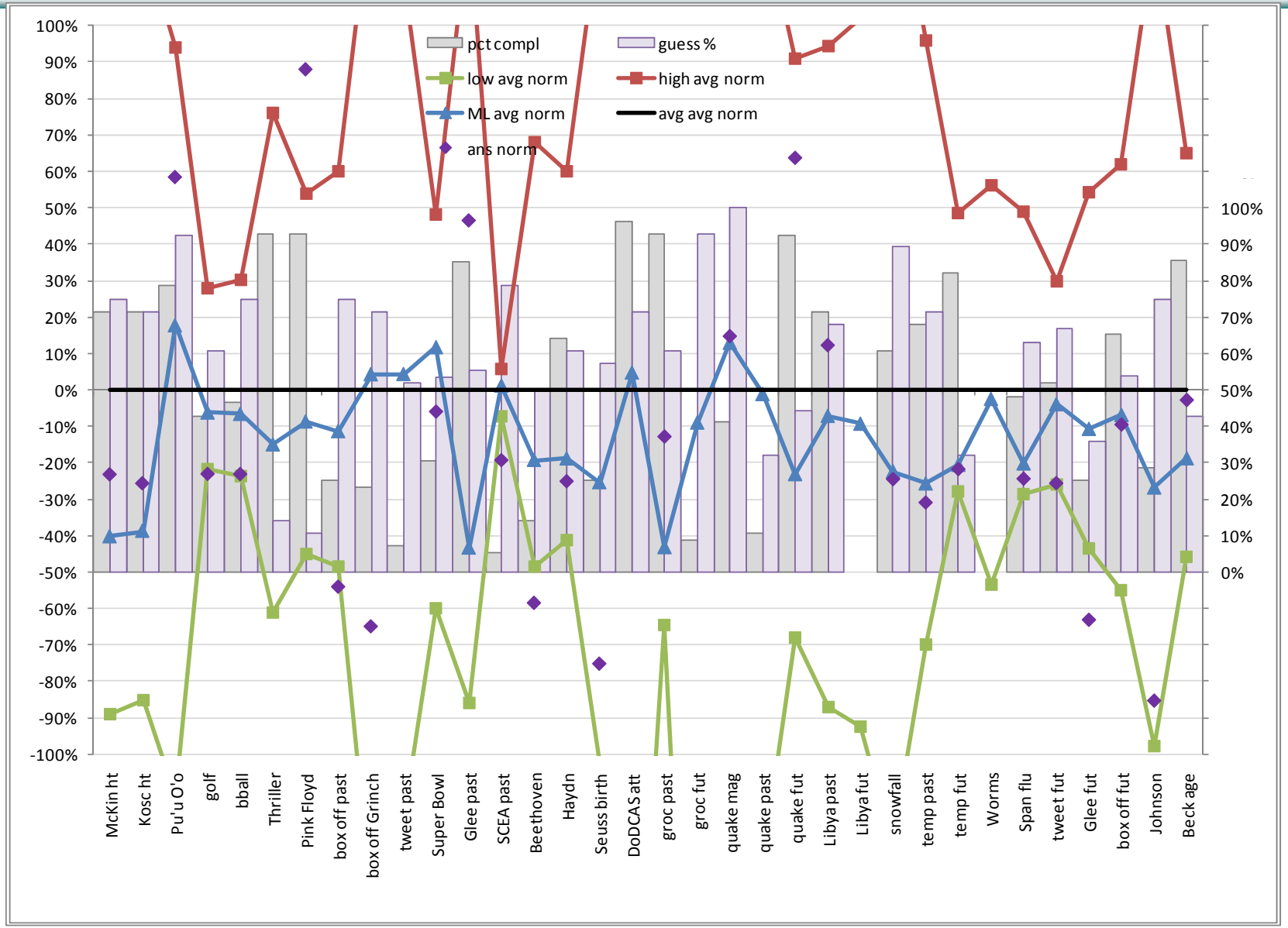
- 16/20 trained (4/8 untrained) respondents had > 50% success rate
  - 12/20 trained (0/8 untrained) > 66%
  - Overall success rate was 60%
    - Consistent with 20/80 interval
    - 65% for those who received feedback on Part 1, 45% for those who did not
- Past (62%) vs. future (60%) not statistically significant
  - No appreciable difference in interval widths
- On average, true answer (normalized) was 0.13 higher than overall average
  - Standard deviation of correct answers (normalized) was 1.78
- Collectively, respondents were at the 45th percentile
- Parameter-averaged distribution was correct 76% of the time, at 59th percentile
- Predictions ("tip-to-tip") were mostly in the 0.5-1.0 normalized range (see graph)
  - Average standard deviation increased from 0.26 (Part 1) to 0.37 (Part 2)

Differences from pilot survey noted in red

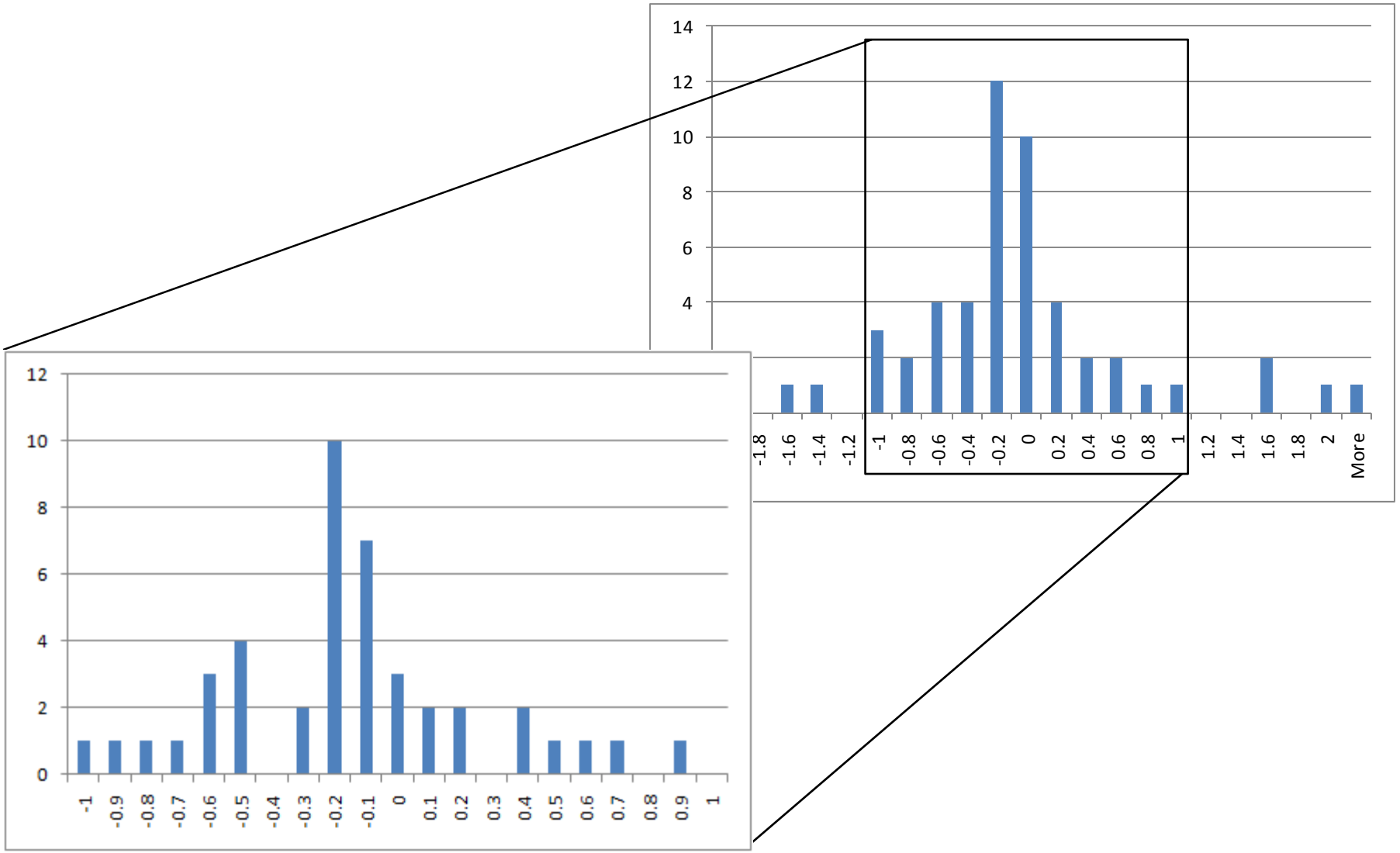
Average Responses by Question (Part 1)



Average Responses by Question (Part 2)

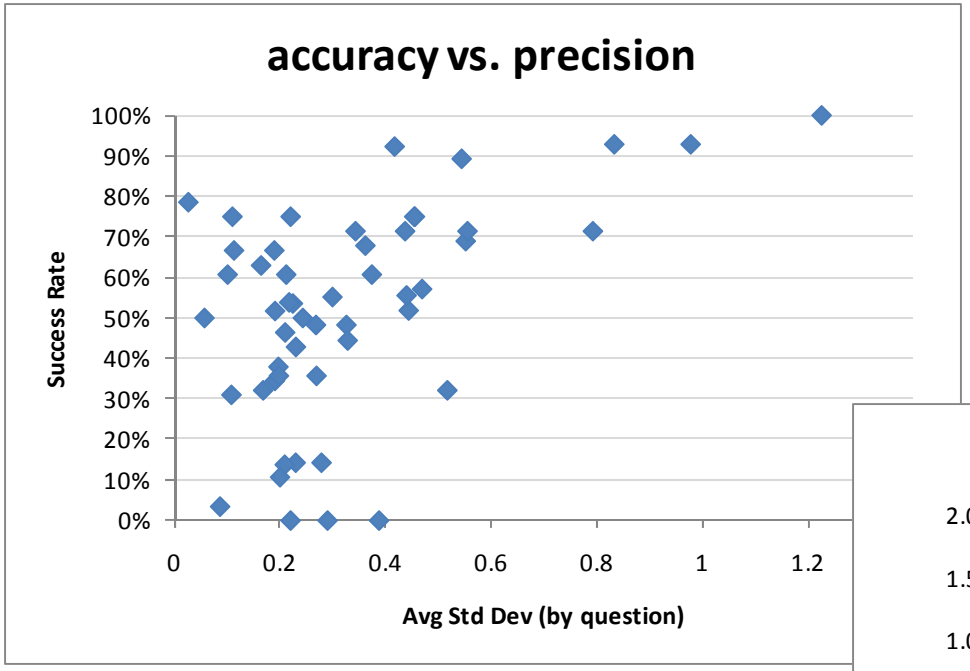


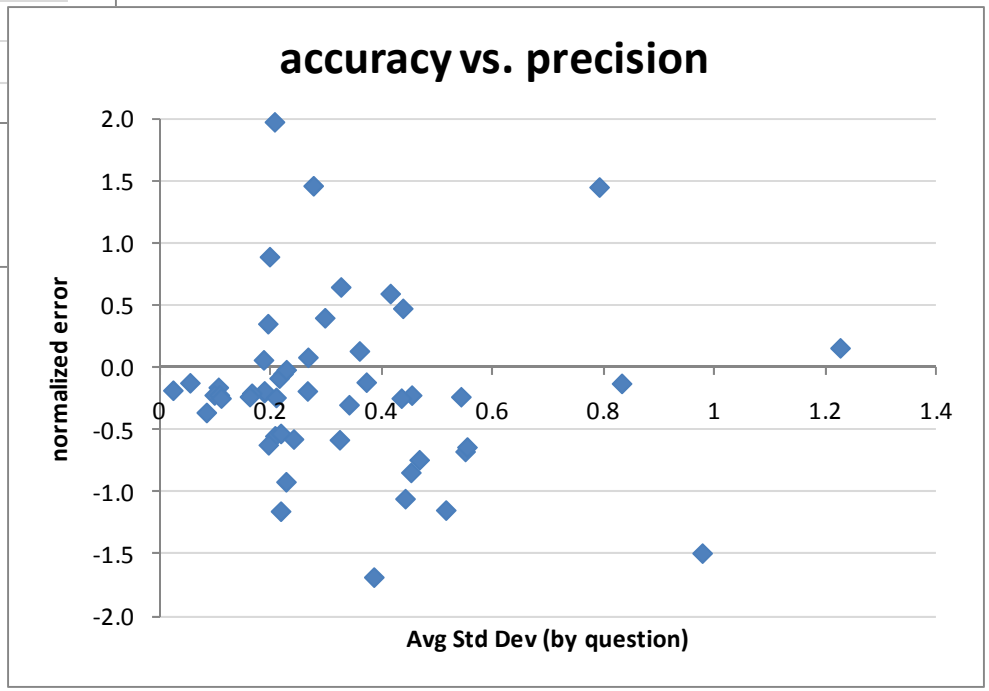
Distribution of Average Error



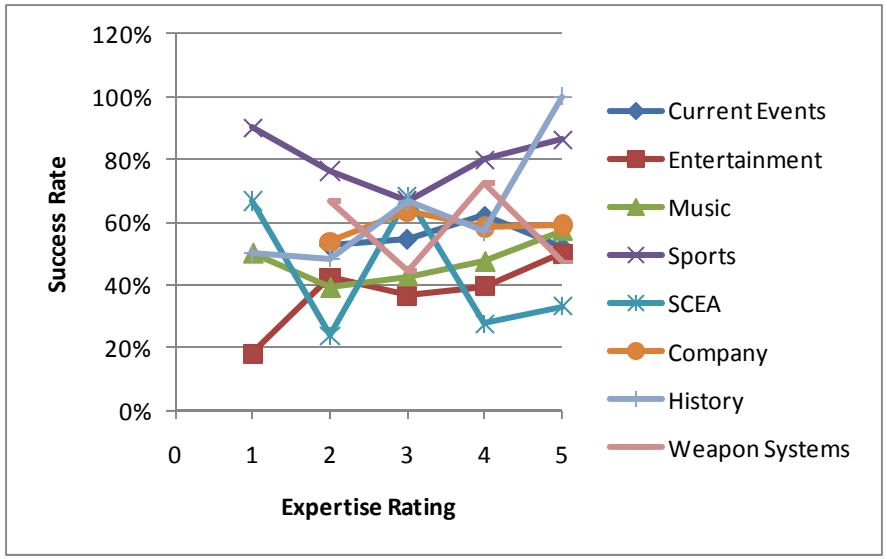
Accuracy and Precision

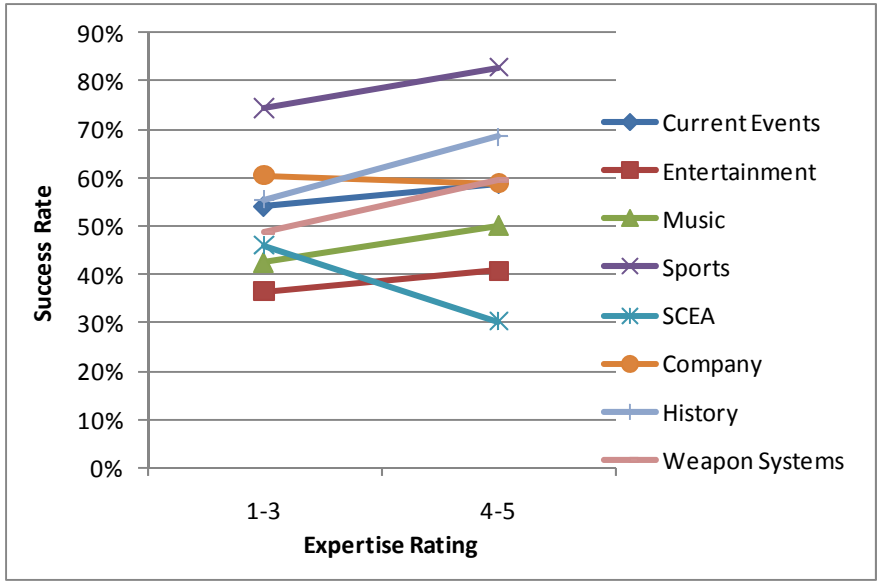
- Compare width of respondents' intervals to two measures of accuracy

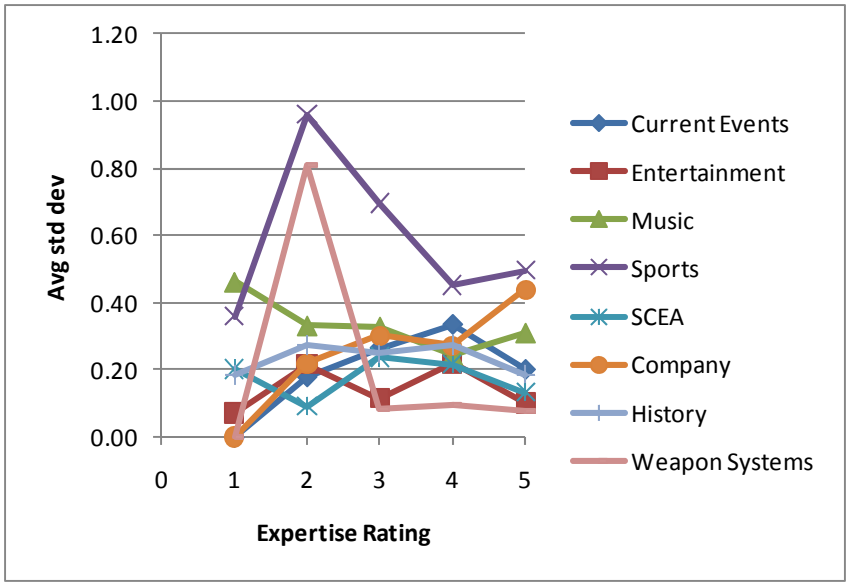


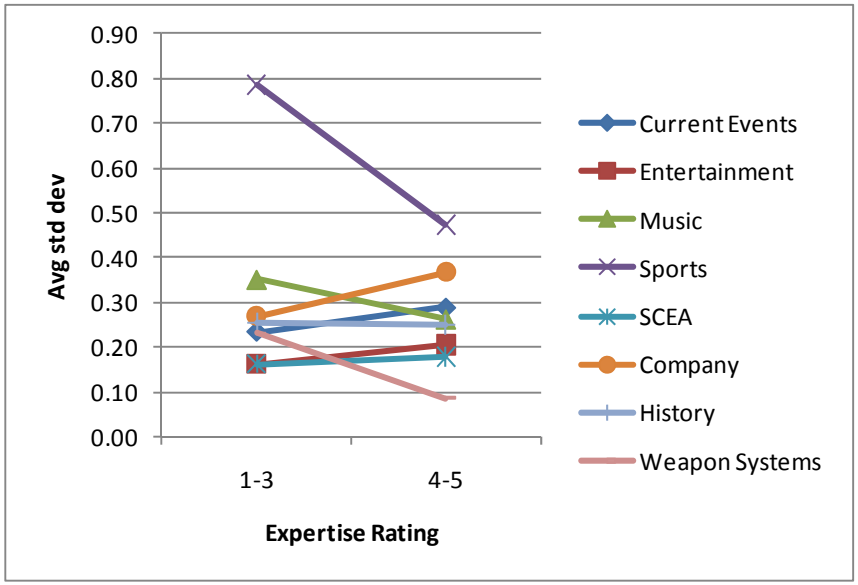


Effects of (Self-Reported) Expertise









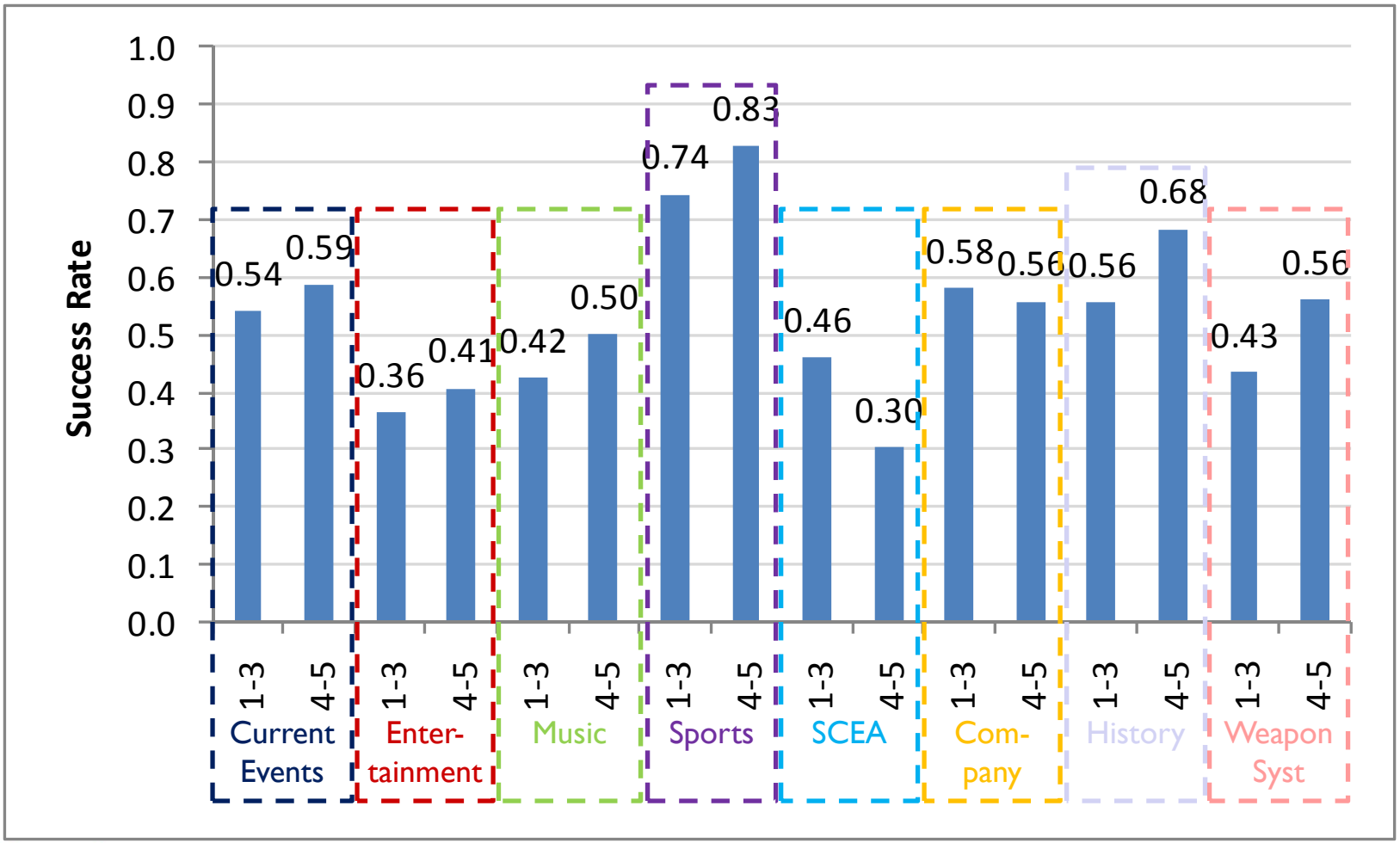
Effects of (Self-Reported) Expertise

- In this case, certain categories produced superior experts, who gave narrower intervals *and* were more accurate
  - In other categories, "expertise" seemed to make one neither more precise *nor* more accurate
- Rankings for SCEA and Company likely problematic
  - If respondents really scored relative to general population, everyone would've been a "5"
- Generalizations are dangerous given relatively small samples

	narrower interval?	more accurate?	both?
Current Events	N	N	N
Entertainment	N	N	N
Music	Y	Y	Y
Sports	Y	Y	Y
SCEA	N	N	N
Company	N	N	N
History	N	Y	N
Weapon Systems	Y	Y	Y

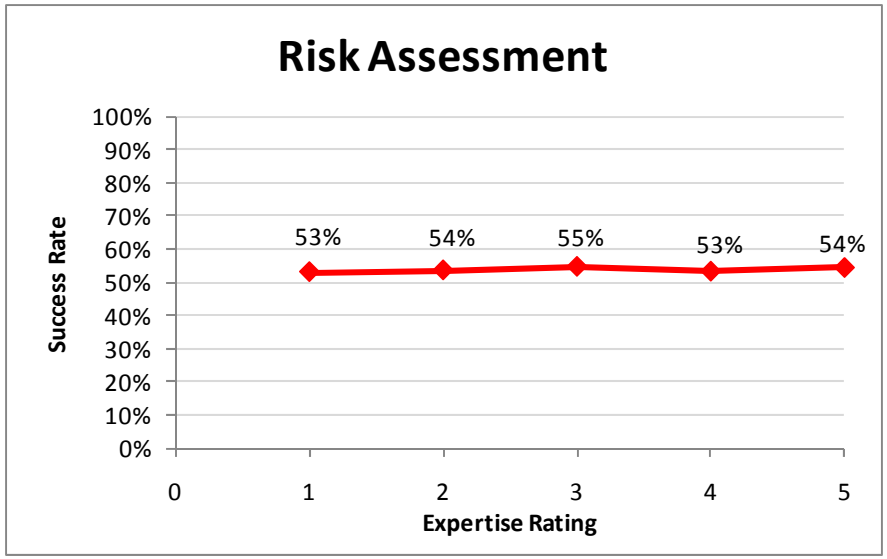
Categories Success Rates Comparison

- Binomial test conducted for statistical significance
  - Assume normal approximation, np > 5, nq > 5



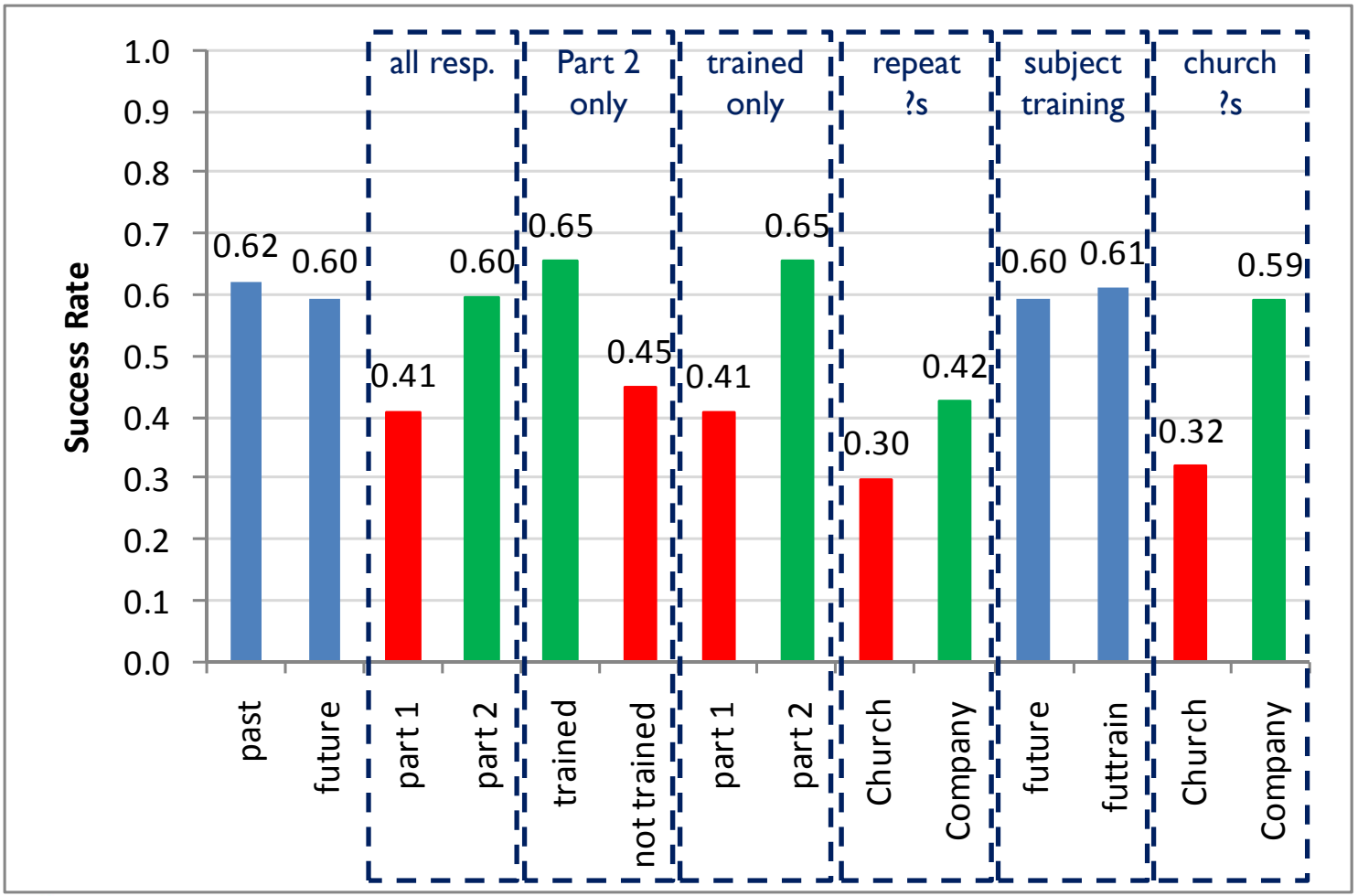
Risk Assessment Expertise

- Even more startling, expertise in Risk Assessment has absolutely no bearing on accuracy
  - All five rankings are within plus or minus 1% of grand average!
- No clear pattern of Risk Assessment expertise affecting interval width



Success Rates Comparison

- Binomial test conducted for statistical significance
  - Assume normal approximation, np > 5, nq > 5



Hypothesis Results (1 of 2)

- Hypothesis: Respondents will be "correct" no more than about 2/3 of the time
  - Result: Supported; only rare exceptions prior to "training" (feedback)
- Hypothesis: Respondents will do about equally well gauging past events as predicting future events
  - Result: Supported; percentages nearly identical
- Hypothesis: As a group, respondents will be unbiased
  - Result: Rejected; modest bias toward underestimation (percentile and mean); analysts more consistent and less extreme than laymen
- Hypothesis: The parameter-averaged distribution (Method 2a) will perform better than individual distributions
  - Result: Supported; higher success rate, higher percentile (due to right-skew meta-distribution?)
- Hypothesis: Respondents who rate themselves expert (4 or 5) in a Category will be correct more often and/or will have narrower intervals
  - Result: Rejected; very little correlation between (self-assessed) expertise and results
- Hypothesis: Respondents who rate themselves expert in Risk Assessment will be correct more often
  - Result: Strongly rejected; success rates virtually identical across Risk Assessment 1-5

Hypothesis Results (2 of 2)

- Hypothesis: Respondents will do better at Categories they "should" know better (e.g., Company, Weapon Systems, SCEA), independent of self-assessment
  - Result: Rejected; performance in these categories was no better than average
- Hypothesis: Since respondents are analysts who work with numbers for a living, innumeracy will be less of an issue
  - Since questions cover a broad range of subject matter, ignorance will still cause difficulty
  - Result: Inconclusive; still a wide range of answers, hard to tell root cause
- Hypothesis: Respondents will have a much higher success rate after receiving training
  - Result: Strongly supported; significant increase in success rate after training
- Hypothesis: The spread of average responses across respondents ("peak-to-peak") will be comparable to the average (corrected or uncorrected) low-to-high spread ("tip-to-tip")
  - Result: Rejected; variation between respondents much greater than spread "within" respondents
- Hypothesis: Distribution of responses will be comparable across questions ("pseudo-iid")
  - Result: Rejected; normalized distributions widely different across questions

Outlier Analysis

- "Bad" respondents
  - Two 17-year-olds amongst adult sample
  - Persistently disingenuous, flip, or misleading responses (quite rare)
- "Bad" questions
  - Pink Floyd question incorrectly worded as album Top 40 instead of Top 200
  - Turns out there were many Diets of Worms, but the famous one was in 1521
  - Influenza epidemic referenced as 1919 (more associated with 1918, actually extended 1917-1920)
  - DCARC question incorrectly worded as "programs" instead of Major Defense Acquisition Programs (MDAPs)
  - FY12 total in FY11 Defense Authorization Bill (outlays only, not appropriations)
- "Bad" answers
  - Blank answers to certain questions (quite rare)
  - Number of Charlie Sheen's tweets (in a day) misinterpreted as number of followers on Twitter
    - A million instead of a handful!
    - Illustrates peril of estimating "new technology"

Curiosities

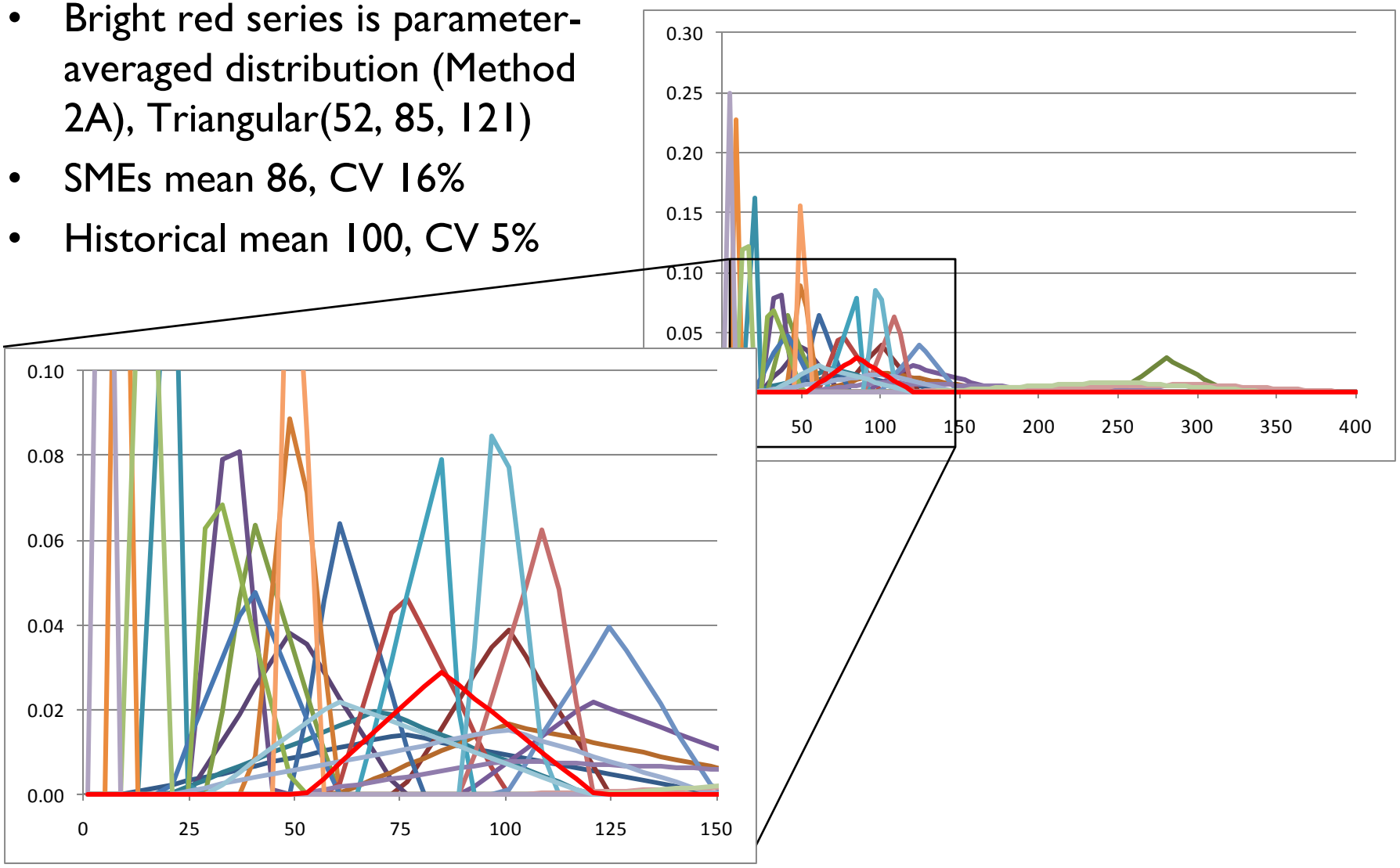
- For some questions, it was hard to find agreement on the correct answer
  - Album sales of Michael Jackson's *Thriller* given as 65-110 million
  - Deaths due to 1917-1920 "Spanish flu" epidemic given as 50-100 million
  - Those are pretty wide ranges!
- For some questions, the correct answer changed!
  - As of the church survey, the Japan earthquake was 8.9 on the Richter scale
    - A couple days later, it was reclassified as 9.0
- Lesson (to be) learned:
  - As much as we'd like to think the quantities we're estimating – cost and schedule – are objective and precise, there may be uncertainty not only in the prediction but also in the final actual value!
    - Cf. Ray Covert's Error In Variables (EIV)

Dueling Distributions

- Most of the survey "right answers" are point values (e.g., the height of Mt. McKinley in feet)
  - Even if they conceivably arose from a probability distribution (e.g., of possible high temperatures for a given day)
  - Typical problem for risk analysts: "Cost is an unrepeatable experiment"
- What if we could infer the "real" distribution and compare the experts' distribution?!
- Model this year's Super Bowl audience as a random variable based on the previous recent attendance (20 years, say)
  - Normalized for United States population to 2010
  - Approximately 100M viewers (mean) with a CV of 5%
    - Implies 2011's 111M viewers was a "two-sigma event"
  - Since this was a past characterization, respondents' means may have been influenced by reports of record viewership

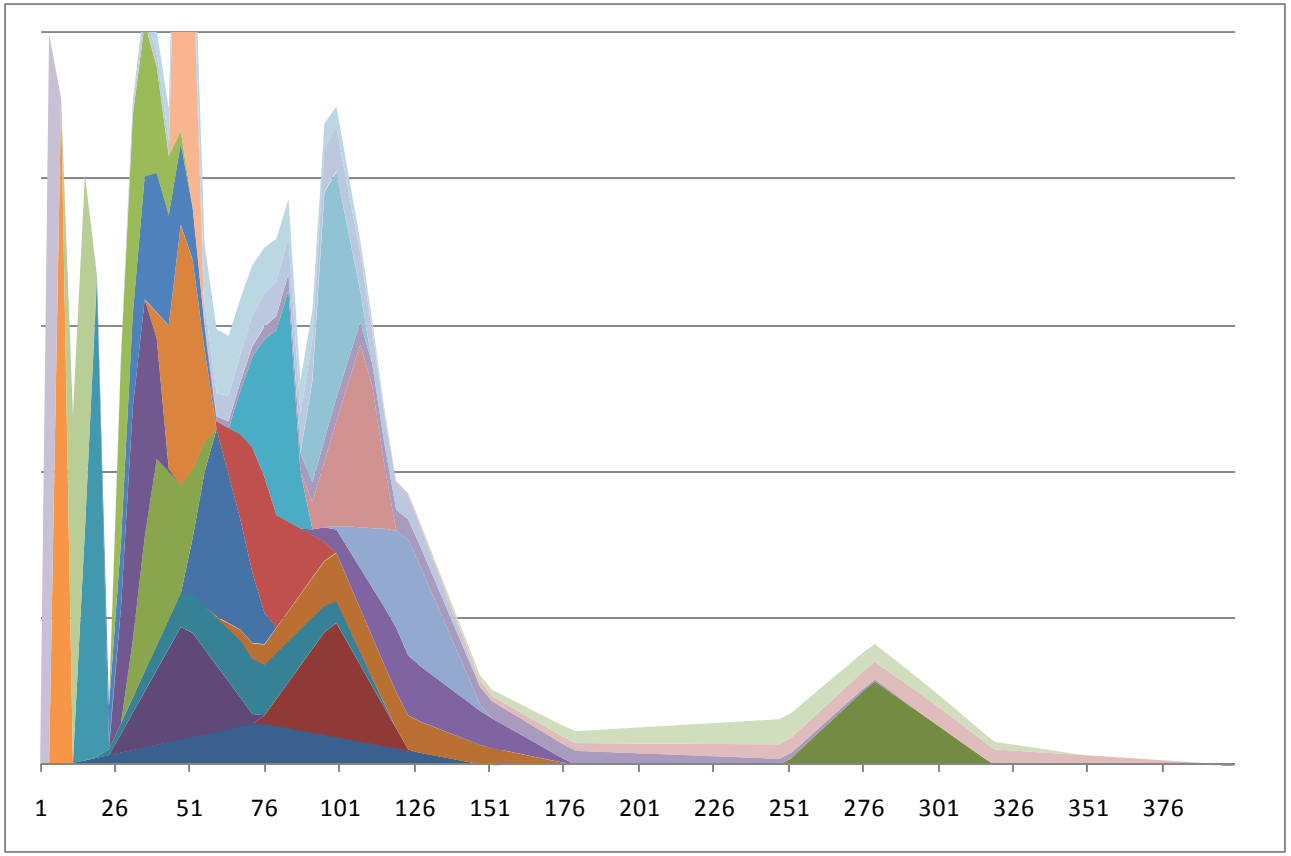
Super Bowl Audience

- Bright red series is parameter-averaged distribution (Method 2A), Triangular(52, 85, 121)
- SMEs mean 86, CV 16%
- Historical mean 100, CV 5%



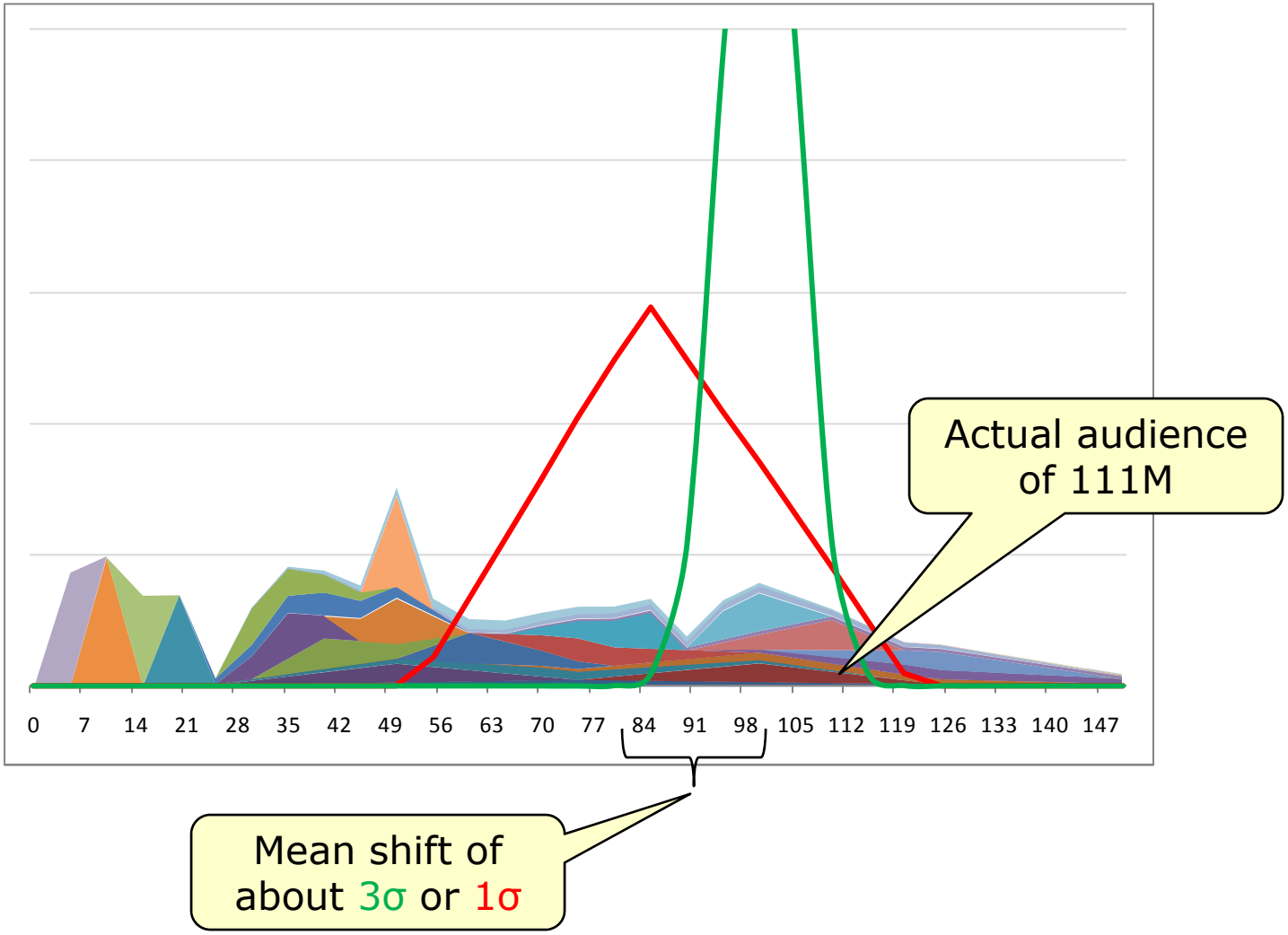
Super Bowl Audience

- Graph shows sampled distribution (Method 3)
- As noisy as this looks, I bet an S-curve would look pretty smooth!



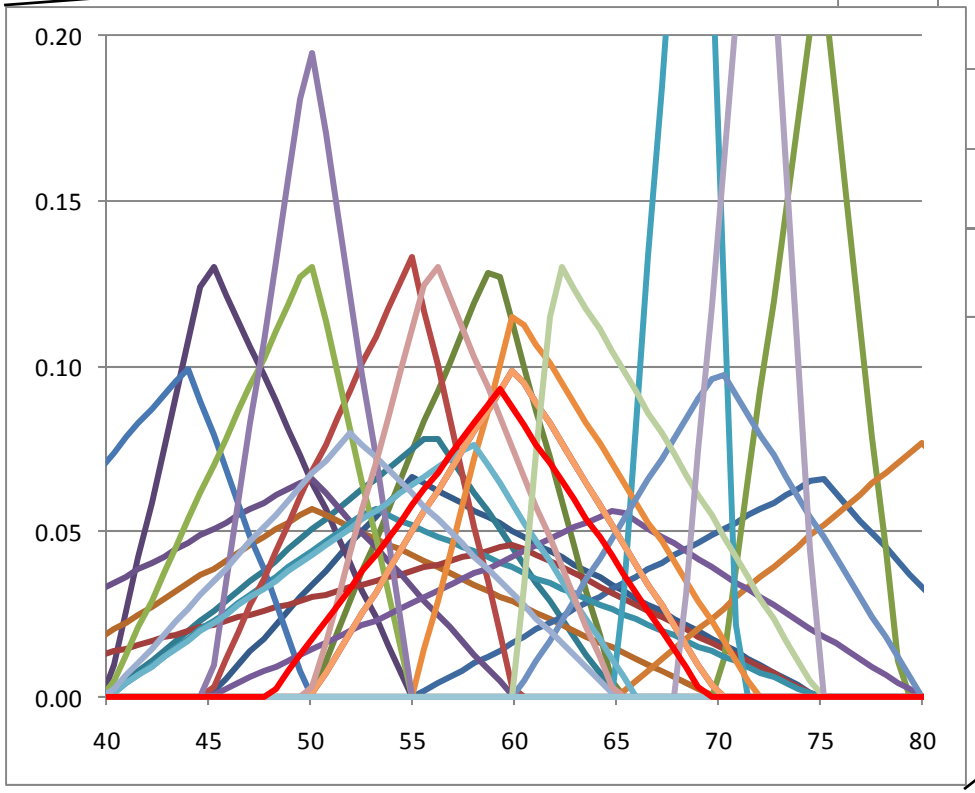
Super Bowl Audience

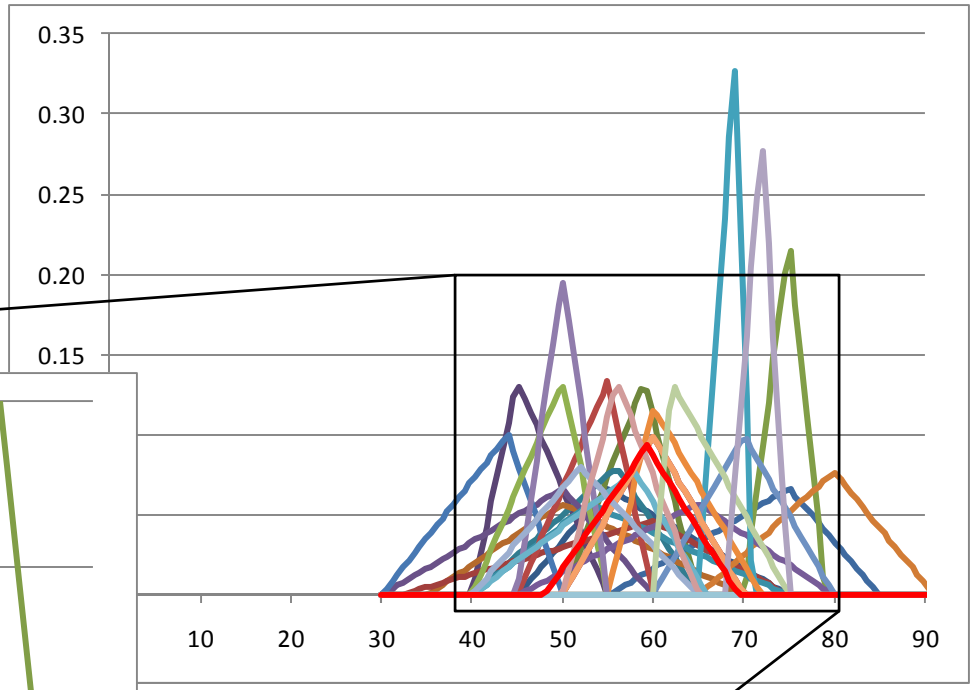
- Adds Parameter-averaged distribution (Method 2A), Historical



March 25th High Temp

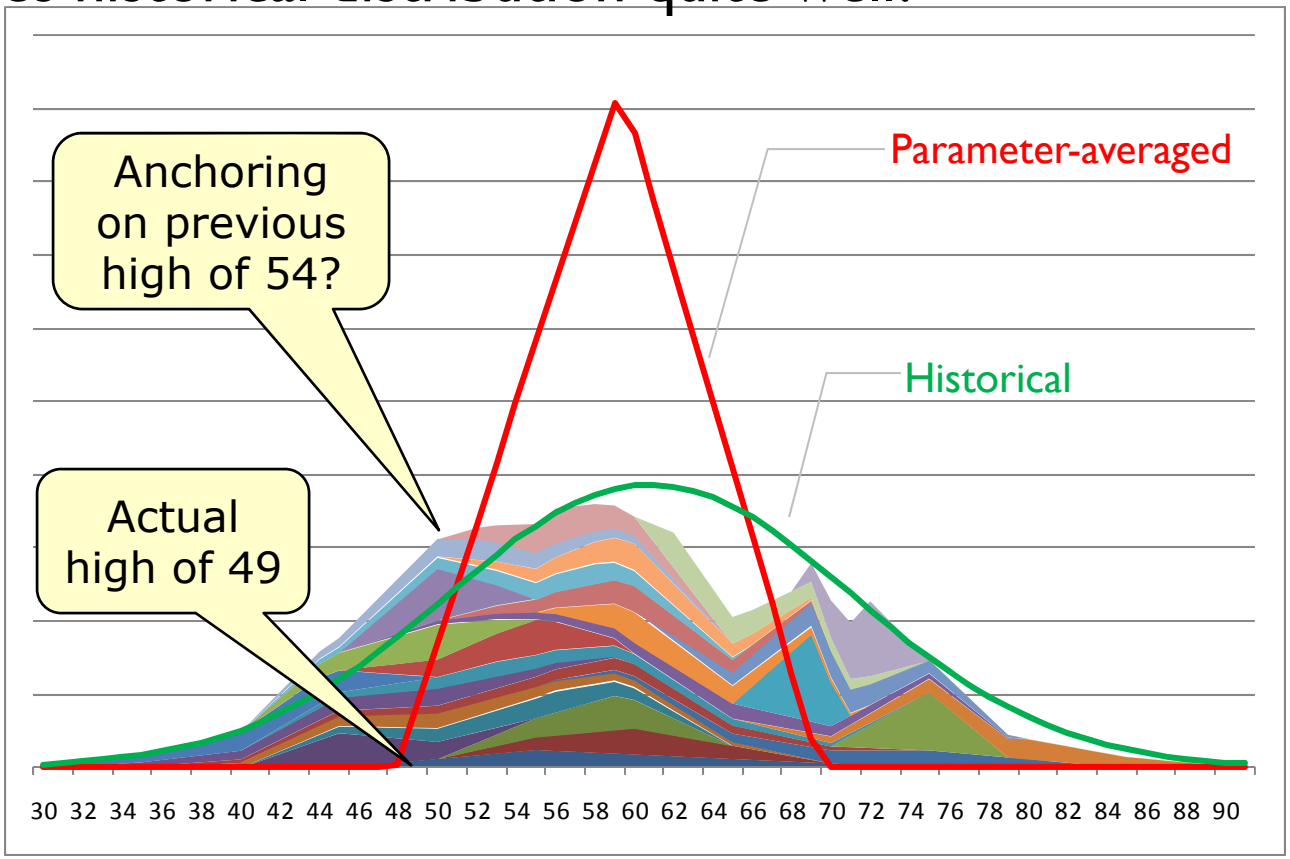
- Bright red series is parameter-averaged distribution (Method 2A), Triangular(48, 59, 69)
- SMEs mean 59, CV 7%
- Historical mean 61, CV 17%





March 25th High Temp

- Area Chart shows sampled distribution (Method 3)
- Much more coherent-looking than Super Bowl data
  - Like Sports, Weather is something people get constant feedback on
- Matches historical distribution quite well!



Two Correction Factors

- Primary correction factor is based on accuracy
  - If α = 100% - % correct, then factor = $1/(1-\sqrt{\alpha})$
- If we take the standard deviation of the average responses ("peak-to-peak") as indicative of the true underlying standard deviation, we can compare the average standard deviation of the responses ("tip-to-tip") to obtain a measure of understatement
  - Corresponding correction factor is reciprocal of average (normalized) standard deviation
  - Susceptible to outliers

Future Work

- Last year's paper focused on literature search and theoretical development
- This year's paper focused on results of surveys spanning laymen and analysts
- Next year's paper (!) might bite off:
  - Survey(s) of SMEs in their areas of expertise
    - Technical, plus Cost and Schedule implications
    - Including Risk Assessment training for improved responses
  - Survey(s) where the true underlying distribution, not just the point value, is "known"
  - Bayesian viewpoint, tempering SME input with data

SME Risk Key Messages

1. Reality: People (including SMEs) can't conceive of true **mins and maxes**
   - Recommendation: Don't ask for min and max, ask for a 20/80 or 10/90
2. Reality: People don't have a clue what kind of **interval** they're providing
   - Recommendation: Ask for a specific interval (see #1)
3. Reality: People consistently understate **uncertainty**
   - Recommendation: Multiply provided intervals by about 2.5
4. Reality: People consistently understate **risk**
   - Recommendation: Don't lose sight of #4 for #3, add appropriate risk factors
5. Reality: People have a pitiful lack of **self-awareness**
   - Recommendation: Beware of both over-confidence and self-effacement, don't take self-assessments at face value
6. Reality: People have trouble **estimating** *everything*
   - Recommendation: Don't allow cost and schedule to be portrayed as a special case, apply risk and uncertainty to technical inputs as well, for example
7. Reality: People have less trouble estimating in areas where there is constant **feedback**, like Sports (scores, statistics, odds, water-cooler conversations, talk radio, etc.)
   - Recommendation: Provide feedback on assessments, as objective and immediate as possible
8. Reality: People can be **trained**; risk assessment is a learned skill
   - Recommendation: Train your SMEs in both general and subject-specific assessment
9. Reality: People are great at **rationalization**, selective memory, and rewriting history
   - Recommendation: Don't rely on "anecdotal actuals" / "expert testimony"; develop and maintain an objective track record

Black Swans

Coleman's Law of the Restaurant Bill

Coleman's Law of Eternal Optimism

Teaching the pig to sing

"I knew LPD 17 would overrun!"

What Is Expertise?

- For SME-based risk, there are three critical skills
  1. **Technical expertise**, the ability to "difference": engineering, commodity
  2. **Cost expertise**, the ability to quantify (in unfamiliar units): what do those differences cost?
  3. **Risk expertise**, the ability to assess ranges: what is the inherent uncertainty?

Appendix: Triangular Derivations[13]

13. IN 06B "Probability Distributions for Risk Analysis," Peter Braxton, SCEA/ISPA, 2011.

The Geometry of Symmetric Triangles

- For a symmetric Triangular(L, M, H), where M-L = H-M
- Find points l and h such that l and h are the p^{th} and $1\text{-}p^{th}$ percentiles

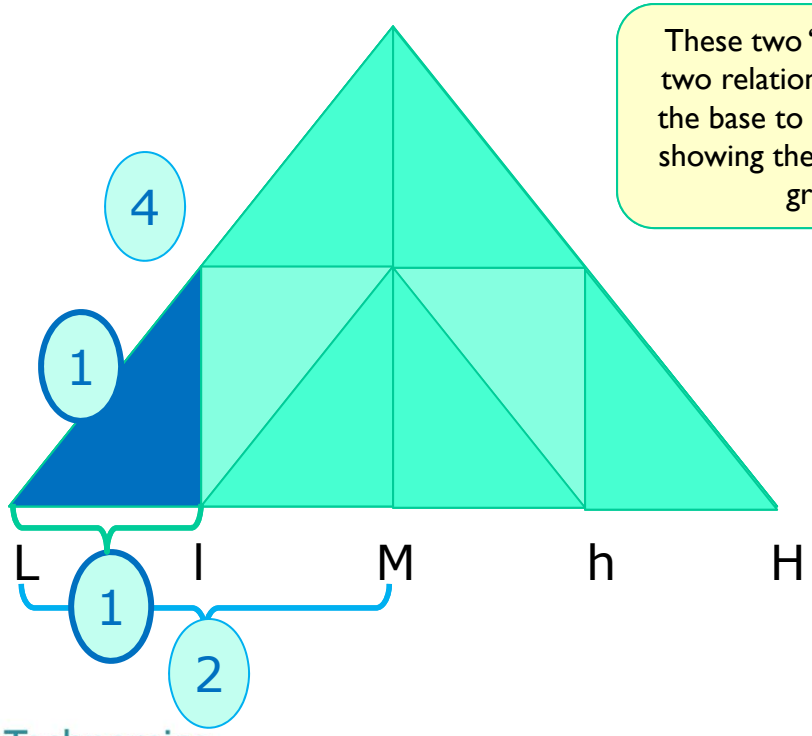
  If l-L = 1/2*(M-L), H-h = 1/2*(H-M), then p = $1/(2*2^2)$ = 1/8 = 12.5%

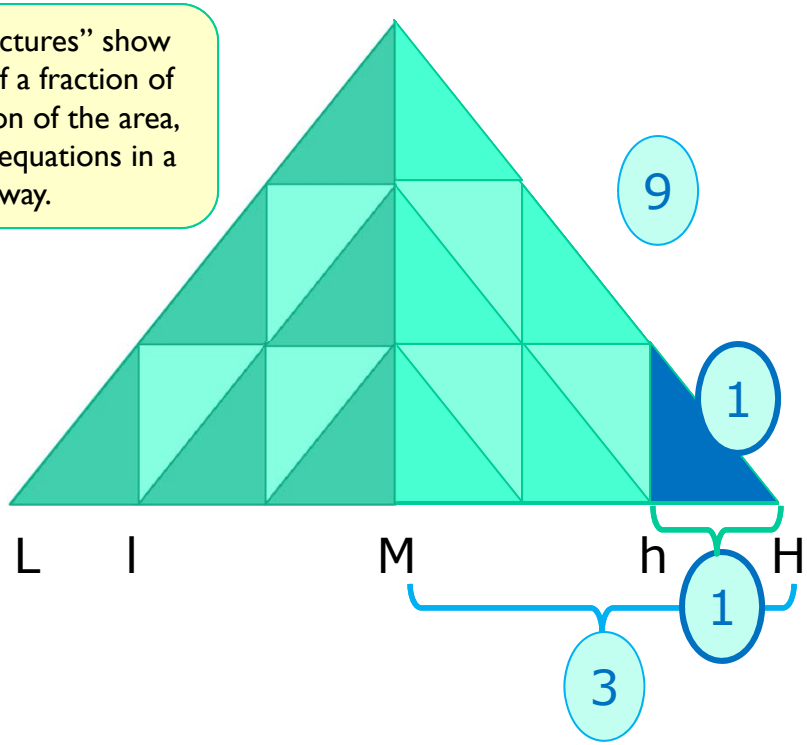
  If l-L = 1/3*(M-L), H-h = 1/3*(H-M), then p = $1/(2*3^2)$ = 1/18 = 5.6%

  p^{th} percentile -> $\sqrt{(p/2)}$ base fraction -> $\sqrt{(2p)}$ half-base fraction

  So, the 20^{th} percentile -> 1/5 occurs at point $\sqrt{(1/10)}$ = 0.3162 base fraction

These two "tiled pictures" show two relationships of a fraction of the base to a fraction of the area, showing the above equations in a graphical way.





Triangles With Related Areas

- We wish to know how to draw triangular distributions that are related to one another
- Constant area:

$$A = \frac{1}{2}bh = \frac{1}{2}(bk)\left(\frac{h}{k}\right)$$

  - Used in **expansion of experts** (correcting understated variance)
  - For area to remain constant, in this case A = 1, as the base increases by a factor, the height must be multiplied by the reciprocal of that factor
- Reduction in area:
  - For area to be reduced by a factor, the dimensions of a <u>similar</u> triangle must be reduced by the square root of that factor

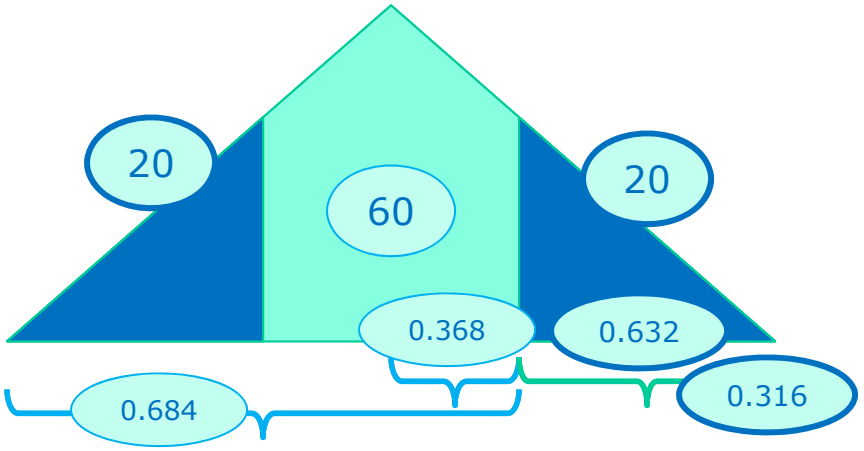
$$A_2 = \frac{1}{k}A_1 = \frac{1}{2k}b_1h_1 = \frac{1}{2}\left(\frac{b_1}{\sqrt{k}}\right)\left(\frac{h_1}{\sqrt{k}}\right)$$

  - For area to be reduced by a factor, the height must be reduced by that factor if the base is to remain constant
    - Used in **sampling of experts**

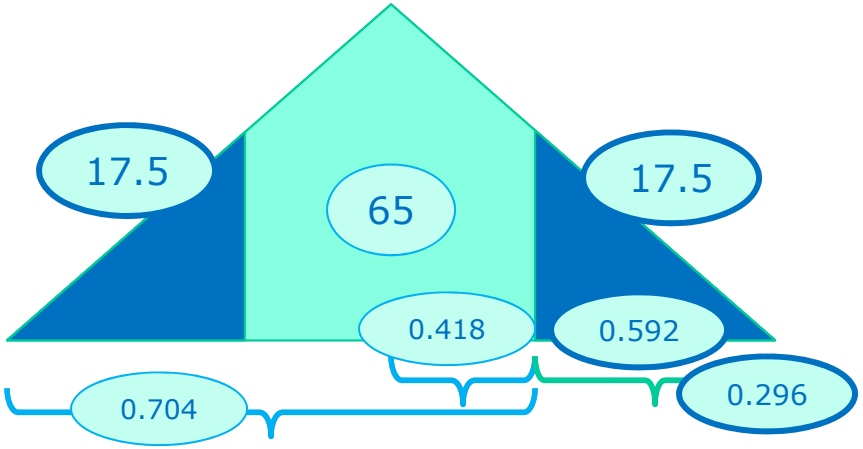
$$A_2 = \frac{1}{k}A_1 = \frac{1}{2k}b_1h_1 = \frac{1}{2}(b_1)\left(\frac{h_1}{k}\right)$$

Correction of Understated Variance for Triangles

- For symmetric triangles
  - To expand from 20-80 to Min-Max, multiply by 2.72 = 1/0.368
  - $\sqrt{(1/10)}$ = 0.3162 base fraction
  - $\sqrt{(2/5)}$ = 0.6325 half-base fraction

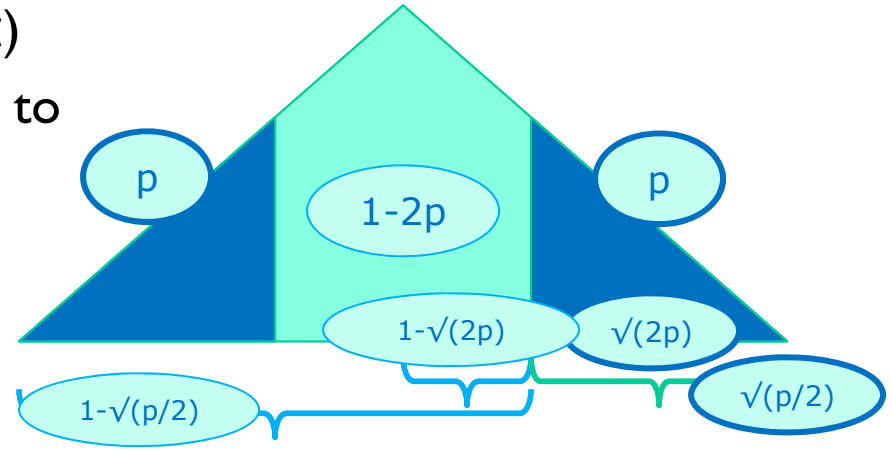


  - To expand from plus-or-minus-one-sigma to Min-Max, multiply by 2.45 ($\sqrt{6}$)
  - $(\sqrt{6}-1)/2\sqrt{6}$ = 0.2959 base fraction
  - $(\sqrt{6}-1)/\sqrt{6}$ = 0.5918 half-base fraction
  - Compare with 68.3% within one sigma rule of thumb for Normal distribution



Correction of Understated Variance for Triangles

- For symmetric triangles
  - General case
  - To expand from p^{th}-$(1-p)^{th}$ to Min-Max, multiply by $1/(1-\sqrt{(2p)})$
  - If $2p = \alpha$, then multiply by $1/(1-\sqrt{\alpha})$
  - To expand from $(\alpha_1/2)^{th}$-$(1-\alpha_1/2)^{th}$ to $(\alpha_2/2)^{th}$-$(1-\alpha_2/2)^{th}$ $[\alpha_1 > \alpha_2]$, multiply by $(1-\sqrt{\alpha_2})/(1-\sqrt{\alpha_1})$
  - For example, to expand from 33-67 to 20-80, multiply by $(1-\sqrt{(2/5)})/(1-\sqrt{(2/3)}) \approx 2.0$



- This approach can easily be extended to asymmetric triangles, with assumption of proportionality[14]:
  - We seek T(a*, c, b*) such that the interval (a, b) contains $1-\alpha$ probability
  - Assume (c-a*) : (c-a) :: (b*-c) : (b-c), or (b*-c) : (c-a*) :: (b-c) : (c-a)

$$a^* = c - \frac{c-a}{1-\sqrt{\alpha}} \qquad b^* = c + \frac{b-c}{1-\sqrt{\alpha}}$$

14. Probability Methods for Cost Uncertainty Analysis: A Systems Engineering Perspective, Paul R. Garvey, Marcel Dekker, 2000..

Triangular Distribution – PDF and Mean

- For Triangular(L,M,H) , denote L=a, H=b, ML=c by T(a,c,b)
- Since the area of the triangle must be 1 (100%), the height is twice the reciprocal of the base
  - We can then derive the PDF by using similar triangles

$$p(x)=\begin{cases}\dfrac{2}{b-a}\dfrac{x-a}{c-a} & a\le x\le c\\ \dfrac{2}{b-a}\dfrac{b-x}{b-c} & c\le x\le b\end{cases}$$

$$\mu=E[X]=\int_a^b xp(x)dx=\int_a^c \frac{2x}{b-a}\frac{x-a}{c-a}dx+\int_c^b \frac{2x}{b-a}\frac{b-x}{b-c}dx$$

$$=\frac{1}{b-a}\left[\left.\frac{\frac{2}{3}x^3-x^2a}{c-a}\right|_a^c+\left.\frac{x^2b-\frac{2}{3}x^3}{b-c}\right|_c^b\right]=\frac{1}{b-a}\left[\frac{2}{3}c^2+\frac{2}{3}ac+\frac{2}{3}a^2-ac-a^2+b^2+bc-\frac{2}{3}b^2-\frac{2}{3}bc-\frac{2}{3}c^2\right]$$

$$=\frac{1}{b-a}\left[\frac{bc-ac}{3}+\frac{b^2-a^2}{3}\right]=\boxed{\frac{a+b+c}{3}}$$

Triangular Distribution – Variance

$$\sigma^2 = E\left((X-\mu)^2\right) = E\left(X^2\right) - \mu^2$$

$$E\left(X^2\right) = \int_a^b x^2 p(x)dx = \int_a^c \frac{2x^2}{b-a}\frac{x-a}{c-a}dx + \int_c^b \frac{2x^2}{b-a}\frac{b-x}{b-c}dx = \frac{1}{b-a}\left[\left.\frac{\frac{1}{2}x^4 - \frac{2}{3}x^3 a}{c-a}\right|_a^c + \left.\frac{\frac{2}{3}x^3 b - \frac{1}{2}x^4}{b-c}\right|_c^b\right]$$

$$= \frac{1}{b-a}\left[\frac{1}{2}\left(c^3+ac^2+a^2c+a^3\right) - \frac{2}{3}\left(c^2a+a^2c+a^3\right) + \frac{2}{3}\left(b^3+b^2c+bc^2\right) - \frac{1}{2}\left(b^3+b^2c+bc^2+c^3\right)\right]$$

$$= \frac{2}{3}\left(c^2+bc+ac+b^2+ab+a^2\right) - \frac{1}{2}\left(c^2+bc+ac+b^2+ab+a^2\right) = \frac{a^2+b^2+c^2+ab+ac+bc}{6}$$

$$\mu^2 = \left(\frac{a+b+c}{3}\right)^2 = \frac{a^2+b^2+c^2+2ab+2ac+2bc}{9}$$

Square of the base minus product of the half-bases!

$$E\left(X^2\right) - \mu^2 = \frac{3a^2+3b^2+3c^2+3ab+3ac+3bc}{18} - \frac{2a^2+2b^2+2c^2+4ab+4ac+4bc}{18}$$

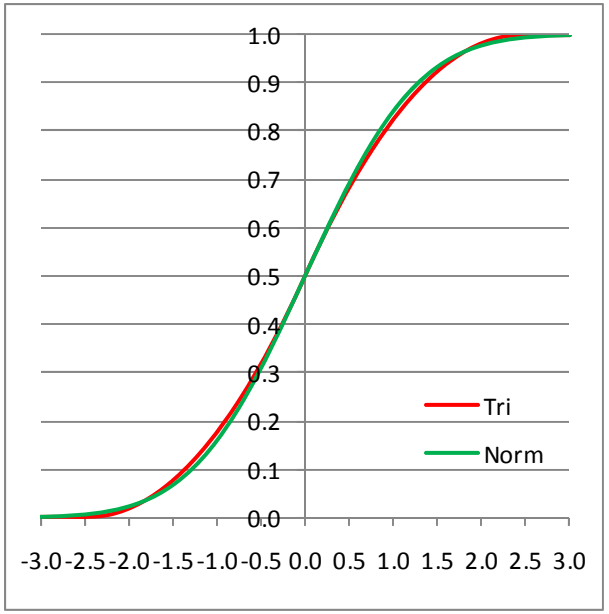
$$= \frac{a^2+b^2+c^2-ab-ac-bc}{18} = \frac{b^2-2ab+a^2+c^2+ab-ac-bc}{18} = \frac{(b-a)^2-(c-a)(b-c)}{18}$$

$$= \frac{c^2-2ac+a^2+b^2-2bc+c^2+bc-ab-c^2+ac}{18} = \frac{(c-a)^2+(b-c)^2+(c-a)(b-c)}{18}$$

Sum of squares of half-bases and product of half-bases!

Substituting a Triangular for a Normal: The √6 Factor

- For a symmetric triangle, let M = m, L = m-w, H = m+w, where w is the half-base
  - Then the mean is m, and the variance is $w^2/6$
- It follows that the half-base is greater than the standard deviation by a factor of √6
- To approximate a normal, N(μ, σ) the factor of √6 is multiplied by the standard deviation of the normal to be emulated to produce the half-base
  - By this means, end points are found that will produce a triangular distribution that emulates the underlying normal in mean and standard deviation
  - This triangular distribution, Tri(μ -√6σ, μ, μ +√6σ) differs from the underlying normal in all other moments, and at all percentiles other than the median and two "cross-over" points, but the difference is minor



Variance of Hybrid Distributions – A Pythagorean Relationship

- Suppose k distributions with pdf $p_i(x_i)$, mean μ_i, and standard deviation σ_i are sampled
- Then the pdf of the hybrid distribution is the "average" of the pdfs

$$p(x) = \frac{1}{k}\sum_{i=1}^{k} p_i(x_i)$$

- The mean of the hybrid distribution is the average of the means

$$\mu = E(X) = \frac{1}{k}\sum_{i=1}^{k}\int x_i\, p_i(x_i)\,dx_i = \frac{\sum_{i=1}^{k}\mu_i}{k}$$

- The variance of the hybrid distribution is the average of the variances plus the variance of the means taken as a discrete probability distribution!
  - See next slide for derivation

Variance of Hybrid Distributions – A Pythagorean Relationship

$$E\left(X^2\right) = \frac{1}{k}\sum_{i=1}^{k}\int x_i^2\, p_i\left(x_i\right)dx_i = \frac{\sum_{i=1}^{k}\left(\sigma_i^2 + \mu_i^2\right)}{k}$$

$$\sigma^2 = E\left(X^2\right) - \mu^2 = \frac{\sum_{i=1}^{k}\left(\sigma_i^2 + \mu_i^2\right)}{k} - \left(\frac{1}{k}\sum_{i=1}^{k}\mu_i\right)^2$$

$$= \frac{\sum_{i=1}^{k}\sigma_i^2}{k} + \left[\frac{\sum_{i=1}^{k}\mu_i^2}{k} - \left(\frac{1}{k}\sum_{i=1}^{k}\mu_i\right)^2\right]$$

- In the special case of two congruent distributions with centers at m-d and m+d, the variance is

$$= \sigma^2 + \left[\frac{(m-d)^2 + (m+d)^2}{2} - m^2\right] = \sigma^2 + d^2$$

Equivalence of Averaging Distributions and Averaging Parameters for Symmetric Triangles

- In the case of symmetric triangles, averaging the individual triangles (with perfect rank correlation) – method **1b** – can be shown to be equivalent to averaging the parameters – method **2a**
  - We will prove it in the case of two triangles, but the proof can easily be extended to more
- As previously shown, the p^{th} percentile ($p<0.5$) for a symmetric triangle is at the $\sqrt{(2p)}$ half-base fraction
  - So the p^{th} percentiles of the two triangles and their average are:

$$a_1 + \sqrt{2p}(c_1 - a_1) \quad a_2 + \sqrt{2p}(c_2 - a_2) \Rightarrow \frac{a_1 + a_2}{2} + \sqrt{2p}\,\frac{(c_1 - a_1) + (c_2 - a_2)}{2}$$

  - But this is clearly just the p^{th} percentile of the average distribution

$$\left(\frac{a_1 + a_2}{2}\right) + \sqrt{2p}\left[\left(\frac{c_1 + c_2}{2}\right) - \left(\frac{a_1 + a_2}{2}\right)\right]$$

  - A similar proof works for $p>0.5$
  - Since all percentiles are equal, the resulting distributions are identical
- Monte Carlo simulation could be used to explore the difference between the two methods for asymmetric triangles, but it is not expected to be large

Equivalence of Averaging Means and Averaging Modes for Triangles

- If we average parameters – method **2a** – as long as we average mins and maxes, it doesn't matter whether we average means or modes
  - Algebraically equivalent
  - Any number of triangles, symmetry not required
- Let the kth triangle be T(a_i, c_i, b_i), and parameter-averaged triangle be T(A, C, B), where

$$A = \frac{\sum_{i=1}^{k} a_i}{k} \quad C = \frac{\sum_{i=1}^{k} c_i}{k} \quad B = \frac{\sum_{i=1}^{k} b_i}{k}$$

- This is averaging the modes; the resulting mean is

$$\frac{A+B+C}{3} = \frac{\sum_{i=1}^{k} a_i + \sum_{i=1}^{k} b_i + \sum_{i=1}^{k} c_i}{3k} = \frac{\sum_{i=1}^{k} \left(\frac{a_i + b_i + c_i}{3} \right)}{k}$$

  which is just the average of the means!
- Reversing the flow, averaging the means can be shown to produce a mode which is the average of the modes

Appendix: Survey Questions

Church Survey (1 of 2)

quantity	units	correct
Total unit price of food items on the front page of the Giant grocery sales flyer last Wednesday (March 9th)	dollars ($)	60.39
Total unit price of food items on the front page of the Giant grocery sales flyer *next* Wednesday (March 16th)	dollars ($)	65.61
Measurement of this week's earthquake in Japan	Richter scale (one decimal place)	8.9
Deaths in Japan due to the earthquake and tsunami, as reported in today's Washington Post (Saturday, March 12th)	# of people	413
Deaths in Japan due to the earthquake and tsunami, as reported in *Monday's* Washington Post (March 14th)	# of people	1000
Points scored by Maryland in last night's Atlantic Coast Conference (ACC) men's basketball tournament game against Duke (Friday, March 11th)	# of points	71
Points scored by the *losing* team in tomorrow's ACC men's basketball championship game (Sunday, March 13th)	# of points	58
Tiger Woods' score for yesterday's round at the Cadillac Championship in Doral, FL (Friday, March 11th)	# of strokes	74
Tiger Woods' score for *tomorrow's* round at the Cadillac Championship in Doral, FL (Sunday, March 13th)	# of strokes	66
Tweets by @charliesheen on Twitter yesterday (Friday, March 11th), not counting retweets	# of Tweets	4
Tweets by @charliesheen on Twitter on Monday (March 14th), not counting retweets	# of Tweets	0
American television audience for this year's Super Bowl (Pittsburgh Steelers vs. Green Bay Packers)	millions of viewers	111
American television audience for the episode of *Glee* that ran immediately after this year's Super Bowl	millions of viewers	26.8
American television audience for this coming week's episode of *Glee* (Tuesday, March 15th)	millions of viewers	10.8
Domestic box-office gross for the opening weekend of *Justin Bieber: Never Say Never*	millions of dollars ($M)	29.5
Domestic box-office gross for the opening weekend of *Battle: Los Angeles* (through Sunday, March 13th)	millions of dollars ($M)	35.6
Symphonies written by Ludwig van Beethoven	# of symphonies	9
Symphonies written by Franz Josef Haydn, "The Father of the Symphony"	# of symphonies	106
Worldwide album sales to date of Michael Jackson's *Thriller*	millions of copies	65
Consecutive weeks in the Billboard 200 for Pink Floyd's *Dark Side of the Moon*	# of weeks	775

Church Survey (2 of 2)

quantity	units	correct
Songs in the *Seussical* score (counting "Part 1" and "Part 2" or 5A and 5B as separate songs), including those cut from the BPC production	# of songs	86
Appearances of the word "fish" in *One Fish, Two Fish, Red Fish, Blue Fish* (not counting the cover and title pages)	# of appearances	11
Birth year of Theodore Seuss Geisel aka Dr. Seuss	year A.D.	1904
Number of Whos depicted on the two facing pages in *Horton Hears a Who* when they first cry out "We are here! We are here! …"	# of Whos	65
Number of drawings of Horton the Elephant in *Horton Hatches the Egg* (not counting the cover and title pages)	# of drawings	24
Domestic box-office gross of *The Grinch Who Stole Christmas*, starring Jim Carrey	millions of dollars ($M)	260.0
Production Leadership and Support personnel listed on p. 21 of the program for BPC's production of *Seussical*	# of adults	22
Cost of light rentals from Atmospheres, Inc., for BPC's production of *Seussical*	dollars ($)	1300
Audience on Friday night (March 4th) for BPC's production of *Seussical*	# of attendees	200
Total donations on Friday night (March 4th) for BPC's production of *Seussical*	dollars ($)	xxxx
Congregation at the second service at BPC tomorrow (Sunday, March 13th)	# of attendees	137
Total donations at the second service at BPC tomorrow (Sunday, March 13th)	dollars ($)	xxxx

Company Survey Part 1

quantity	units	correct
Height of Mt. McKinley, the tallest mountain in North America	feet (ft)	20,320
Height of Mt. Kosciuszko, the tallest mountain in Australia	feet (ft)	7,310
Height of the recent Pu'u O'o eruption on Mauna Loa (March 5th, 2011)	feet (ft)	65
Tiger Woods' score for the final round at the Cadillac Championship in Doral, FL (Sunday, March 13th)	# of strokes	66
Points scored by Maryland in their Atlantic Coast Conference (ACC) men's basketball tournament game against Duke (Friday, March 11th)	# of points	71
Worldwide album sales to date of Michael Jackson's *Thriller*	millions of copies	65
Consecutive weeks in the Billboard 200 for Pink Floyd's *Dark Side of the Moon*	# of weeks	775
Domestic box-office gross for the opening weekend of *Justin Bieber: Never Say Never*	millions of dollars ($M)	29.5
Domestic box-office gross of *The Grinch Who Stole Christmas*, starring Jim Carrey	millions of dollars ($M)	260
Tweets by @charliesheen on Twitter last Monday (March 14th), not counting retweets	# of Tweets	0
American television audience for this year's Super Bowl (Pittsburgh Steelers vs. Green Bay Packers)	millions of viewers	111
American television audience for the episode of *Glee* that ran immediately after this year's Super Bowl	millions of viewers	26.8
SCEA members as of today (Monday, March 21st)	# of members	2,092
Symphonies written by Ludwig van Beethoven	# of symphonies	9
Symphonies written by Franz Josef Haydn, "The Father of the Symphony"	# of symphonies	106
Birth year of Theodore Seuss Geisel aka Dr. Seuss	year A.D.	1904
Technomicians who attended DoDCAS (February 16th-18th, 2011)	# of attendees	23

Company Survey Part 2 (1 of 2)

quantity	units	correct
Total unit price of food items on the front page of the Giant grocery sales flyer last Wednesday (March 16th)	dollars ($)	69.46
Total unit price of food items on the front page of the Giant grocery sales flyer *next* Wednesday (March 23rd)	dollars ($)	65.61
Measurement of the recent major earthquake in Japan	Richter scale (one decimal place)	9
Deaths in Japan due to the earthquake and tsunami, as reported in today's Washington Post (Monday, March 21st)	1000s of people	7.197
Deaths in Japan due to the earthquake and tsunami, as reported in Friday's Washington Post (March 25th)	1000s of people	10
Total coverage of Libya and Muammar Gaddafi in the A section of the Washington Post, Saturday, March 19th (not counting headlines, photos, and op-ed page)	column-inches	112.5
Total coverage of Libya and Muammar Gaddafi in the A section of the Washington Post, Friday, March 25th (not counting headlines, photos, and op-ed page)	column-inches	121.75
Total winter snowfall at Washington National Airport, 2010-2011	inches (in)	10.1
High temperature at Washington National Airport yesterday (Sunday, March 20th)	degrees Fahrenheit	54
High temperature at Washington National Airport on Friday (March 25th)	degrees Fahrenheit	49
Year in which the Diet of Worms was held	year A.D.	1521
Number of deaths worldwide due to 1919 influenza epidemic	millions of people	50
Tweets by @charliesheen on Twitter on Friday (March 25th), not counting retweets	# of Tweets	1
American television audience for this coming week's episode of *Glee* (Tuesday, March 22nd)	millions of viewers	5.267
Domestic box-office gross for the opening weekend of *Sucker Punch* (through Sunday, March 27th)	millions of dollars ($M)	19.1
Number of distinct songs (not counting alternate recordings) on the Robert Johnson boxed set	# of songs	29
Age of guitarist Jeff Beck, whom Rick and Bill are going to see in concert Thursday night (March 24th)	# of years	66

Company Survey Part 2 (2 of 2)

quantity	units	correct
Paid attendance at Jeff Beck concert on Thursday night (March 24th)	# of attendees	1,200
Points scored by Northern Colorado in their opening round game against San Diego State in the NCCA men's basketball tournament on Thursday, March 17th	# of points	50
Points scored by Connecticut in their Sweet Sixteen game against San Diego State in the NCCA men's basketball tournament on Thursday, March 24th	# of points	74
Technomics FY10 total sales, per FY11Q3 CAM, January 19th, 2011	millions of dollars ($M) to the nearest tenth	xx
Recipients of Making Technomics Great Award at FY11Q3 CAM in January	# of recipients	7
Backlog reported at FY11Q3 CAM in January (including > 90% Probability but not Options)	# of months	xx
Backlog reported at upcoming FY11Q4 CAM in April (including > 90% Probability but not Options)	# of months	xx
Size of "more mature" team at 3rd Annual Poohbah Pong Tournament	# of players	15
Number of distinct attendees for Technomics CEBoK training at Naval Center for Cost Analysis (NCCA)	# of attendees	28
F/A-18E/F Super-Hornet total quantity in December 2009 Selected Acquisition Report (SAR)	# of planes	515
FY12 total in 2011 Defense Authorization Bill	billions of dollars ($B)	
U.S. troops killed in Afghanistan in calendar year 2010	# of people	499
Number of distinct defense programs in the Defense Cost Analysis Resource Center (DCARC) database	# of programs	172
Lifetime members of SCEA as of today (Monday, March 21st)	# of members	49
SCEA members as of Friday (March 25th)	# of members	2,092
Congregation at the second service at Burke Presbyterian Church on Sunday, March 13th	# of attendees	137
Total donations at the second service at Burke Presbyterian Church on Sunday, March 13th	dollars ($)	xxxx