Uncertainty of Expert Judgment in Agile Software Sizing

Fearful Asymmetry

Peter J. Braxton-PBraxton@technomics.net

David H. Brown-DBrown@technomics.net

Kenneth S. Rhodes-KRhodes@technomics.net

R. Alex Wekluk-AWekluk@technomics.net

International Cost Estimating and Analysis Association (ICEAA) Conference

2/24/2022

Abstract

Agile software estimating and planning often rely on expert judgment to assess the size of the development effort at various levels of granularity and stages of maturity. Previous research by the author quantified the inherent risk and uncertainty of the self-similar scales (e.g., T-shirt sizing) commonly used in these assessments. This paper expands those a priori mathematical results and empirically tests the accuracy of experts in applying those scales. It elucidates the ideal ratio to align with the desired confidence interval, and recommends feedback mechanisms to improve consistency.

Track: Management and Risk, https://www.iceaaonline.com/pit22sessions/#MRTrack

Keywords: *Early Cost, Functional Requirements, Risk, Software, Uncertainty, Agile, Story Points*

Table of Contents

Abstract....i
Introduction.... 1
Prologue: Who Wants To Be a Millionaire?.... 3
Problem Statement.... 5
Sizing Methods.... 5
T-Shirt Sizing Scales.... 6
Thought Experiment: “Double or Half?”.... 7
Macro-Level T-Shirt Sizing.... 7
Characterizing Risk and Uncertainty.... 8
Symmetric Uncertainty on an Asymmetric Scale....10
Naïve Uncertainty: Coin Flips....11
Problem Context: Self-Similar Scales in Agile Software Development....13
Continuous Risk on a Discrete Scale....13
“Maximum” Uncertainty: The Uniform Distribution....14
“Standard” Uncertainty: The Triangular Distribution....15
“Standard” Risk: The Lognormal Distribution....16
Generalization #1: Confidence Level....19
Proportional or Symmetric Tails....23
Planning Poker, Fibonacci Numbers, and the Golden Mean....28
Micro-Level T-shirt Sizing....28
Alternate Sizing Scales....32
Generalization #2: Scale Ratio....34
Problem Context: Reliance on the Reluctant Expert....38
How Accurate Is the Expert?....38
Self-Similar Scales and the Ideal Ratio....38
Teaching Pigs to Sing: Lessons Learned for Two- and Three-Point Estimates....40
Empirical Experiment: Analogized Scales....40
SMEs in Search of a Basis....40
From Single-Point Analogy to Analogized Scales....41
Experimental Formulation....42
Experiment #1: Heights of Mountains....43
Experiment #2: Box Office Gross of Films....44
Experiment #3: Driving Distances....45

Results and Recommendations 46
Survey Results 46
SME Training and Data Accumulation 46
Conclusion and Next Steps 47
Risk and Uncertainty Benchmarks 47
Continuing Research 48
Agile Software Risk Decomposition 48
Bibliography 50
Acronyms, Initialisms, and Abbreviations 51
Appendix A: Fibonacci Sequence Convergence....52
Coda: The Proverbial Cocktail Napkin 53

Figure 1: Software Estimating Methods.... 1
Figure 1: "Double or Half" in a Game Show Context 3
Figure 2: T-Shirt Sizing Scale.... 8
Figure 3: Discrete Distribution 11
Figure 4: Uniform Distribution.... 14
Figure 5: Triangular Distribution 15
Figure 6: Lognormal Distribution 16
Figure 7: Lognormal CGF and CV as a Function of Confidence 18
Figure 8: Discrete Distribution (with Confidence) 19
Figure 9: Discrete CGF and CV as a Function of Confidence 19
Figure 10: Triangular Distribution with Confidence (Proportional Tails).... 21
Figure 11: Triangular CGF and CV as a Function of Confidence 22
Figure 12: Uniform Distribution with Confidence (Proportional Tails) 23
Figure 13: Uniform Distribution with Confidence (Equal Tails) 25
Figure 14: Uniform CGF and CV as a Function of Confidence (Proportional Tails) 26
Figure 15: Uniform CGF and CV as a Function of Confidence (Equal Tails).... 26
Figure 16: Converging Ratios of Consecutive Fibonacci Numbers 30
Figure 17: Micro-Sizing Accuracy (notional).... 31
Figure 18: Notional Sizing Model 33

Figure 19: Lognormal Risk and Uncertainty with Variations in Both Confidence and Ratio 35
Figure 20: Discrete Risk and Uncertainty with Variations in Both Confidence and Ratio 37
Figure 21: Software Estimation Decomposition, Data Needs, and Risk 48
Figure 22: Artisanal Probability Calculations 53

Table 1: Risk and Uncertainty as a Function of Confidence 27
Table 2: Fibonacci Numbers and Successive Ratios 29
Table 3: Notional Sizing Model Parameters 32
Table 4: Empirical Testing Scale (R = 2.0) 43
Table 5: Empirical Testing Scale (R = 1.5) 43
Table 6: Risk and Uncertainty Benchmarks for T-shirt Sizing 47

Introduction

As shown in the figure below, the selection of a software estimating method is highly dependent on the phase within the development and sustainment cycle, which in turn determines the data available for a software development estimate.

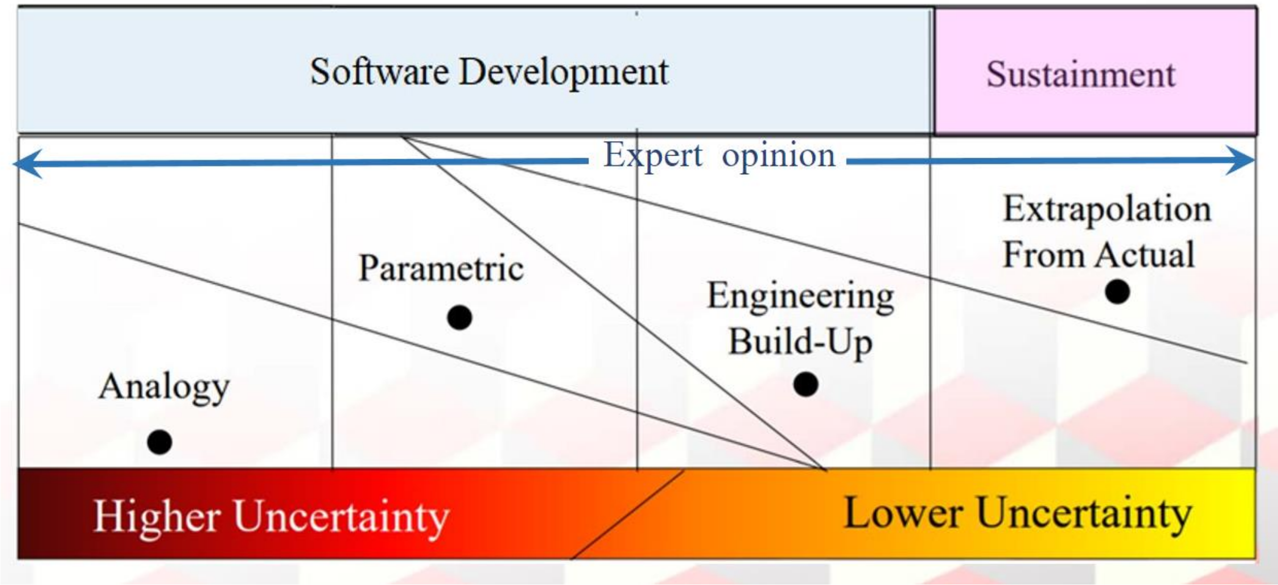


Figure 1: Software Estimating Methods

In the above graphic, the most defensible estimates are produced later in the life cycle, using an extrapolation from actuals method (Cost Estimating Body of Knowledge (CEBoK), 2013). If estimators have an actual cost history available, then this method, such as the one described in ***Are We Agile Enough to Estimate Agile Software Development Costs?*** is most appropriate. (Kosmakos & Brown, 2022) If an actual cost history is not available, and requirements are mature enough to estimate functional size, then a parametric method and tool such as SWEET is an ideal choice, as described in ***Dynamic Software Effort Estimation: How SWEET It Is!*** (Gellatly, Braxton, Brown, Jones, & Wekluk, 2022) It should be acknowledged, however, that an estimate of functional size is not always available, especially for programs/projects early in the life cycle. In these situations, an alternative method is preferred, such as the one described in this paper, ***Uncertainty of Expert Judgment in Agile Software Sizing***. (Braxton, 2022)

The genesis of this paper was to take a very straightforward question to the effect of “what if we’re wrong?” when applying the sort of scale used in T-Shirt Sizing and use it to form the basis of a series of increasingly refined thought experiments (i.e., create an excuse to have fun with probability distributions!). The goal of the research is two-fold: (1) to derive the Risk and Uncertainty implications of such scales; and (2) to carefully consider how the application of such scales can be improved.

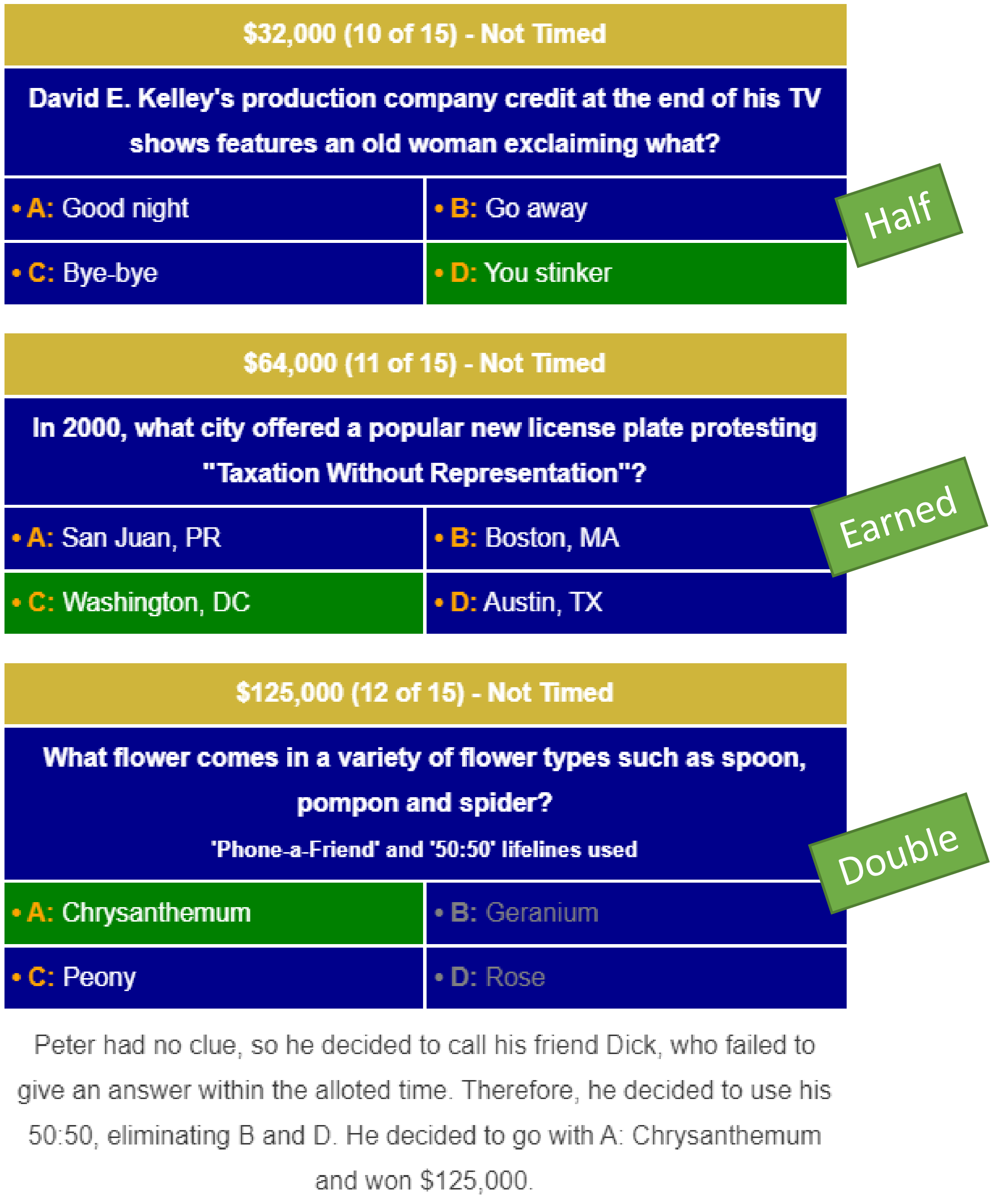
This paper builds on a 2021 IT-CAST presentation (Braxton, 2021) and addresses practical implications of low-level vs. high-level risk and uncertainty.

Prologue: Who Wants To Be a Millionaire?

> *"Baby all the lights are turned on you/*
> *Now you're in the center of the stage"*
> *- Billy Joel, "Everybody Loves You Now," Cold Spring Harbor*

While this anecdote from the author's life was *not* the genesis of this paper, it is an apt illustration of the central "double or half" conceit.

In the basic *Who Wants To Be a Millionaire* game show, the dollar value (approximately) doubles for each question, with $1,000 and $32,000 as the "safe" plateaus. Beyond $32,000, the contestant is faced with a choice: walk away with the amount already earned, or go for the next question ("double") but risk losing all but the $32,000. For the $64,000 Question (see what they did there?!) the losing side of the bet is precisely "half." The actual situation



$32,000 (10 of 15) - Not Timed

David E. Kelley's production company credit at the end of his TV shows features an old woman exclaiming what?

• A: Good night
• B: Go away
• C: Bye-bye
• D: You stinker

$64,000 (11 of 15) - Not Timed

In 2000, what city offered a popular new license plate protesting "Taxation Without Representation"?

• A: San Juan, PR
• B: Boston, MA
• C: Washington, DC
• D: Austin, TX

$125,000 (12 of 15) - Not Timed

What flower comes in a variety of flower types such as spoon, pompon and spider?

'Phone-a-Friend' and '50:50' lifelines used

• A: Chrysanthemum
• B: Geranium
• C: Peony
• D: Rose

Peter had no clue, so he decided to call his friend Dick, who failed to give an answer within the alloted time. Therefore, he decided to use his 50:50, eliminating B and D. He decided to go with A: Chrysanthemum and won $125,000.

Figure 2: "Double or Half" in a Game Show Context

faced by the author is illustrated in Figure 1.[1]

In a recent episode of *The Chase*, chaser James “The High Roller” Holzhauer offered the contestant a five times x (5x) vs. one fifth x (x/5) proposition! While the show seems to offer the chaser considerable flexibility in an effort to bait the contestant into a riskier one-on-one chase, this sort of exponential scale with R times x as the more lucrative option and $1/R$ times x as the less lucrative option, where x is the amount “earned” in the Cash Builder Round, is not uncommon. It is important to note that the ratio R is usually more on the order of two and a half (2.5) or three (3). For our T-shirt sizing, we will start off with a ratio of two (2.0) and then generalize to other possible values of R.

[1] https://millionaire.fandom.com/wiki/Peter_Braxton

Problem Statement

The sentiment "You can't cure bad cost analysis with good risk analysis" is attributed to ICEAA Lifetime Achievement Award winner Dick Coleman. Maybe not, but we have to try, and we might just improve our cost analysis in the process! This paper poses the question of how best to quantify the uncertainty (and risk) associated with agile software sizing estimates produced by subject matter experts (SMEs). Specifically, it examines self-similar scales wherein the SME provides a size assessment chosen from amongst a discrete set of size options, interchangeably referred to as "notches" or "buckets" on the scale. The simplest such scale is T-shirt size with a ratio of two (2.0), similar to the doubling dollar value in *Millionaire*. In our initial thought experiment, we ask what would happen if the SME were off by one T-shirt size in either direction. For example, what if they assessed the size as Medium, and it were really Small (half) or Large (double)?

Despite the cost community's best efforts to engender data-driven methods wherever possible, SME-driven methods persist, and it behooves us to set up both good structures (process) and good training (people) to maximize estimate fidelity. That is the ultimate goal of the disquisitions in this paper.

Sizing Methods

The following are common sizing methods used in software cost estimating, as defined by Agile practitioners.

- T-Shirt Sizing: Agile teams popularized T-shirt sizing in project estimation to track how much time or effort an initiative will take. Each project is assigned a t-shirt size—e.g., Small, Medium, Large, Extra-Large—to represent that task's relative effort.
- Planning Poker: A gamified technique that development teams use to guess the effort of a task. Since the estimates are based on the entire group's input, they can be more accurate than a single expert's judgement.
- Fibonacci Numbers: This technique is borrowed from nature: an exponentially increasing scale that allows relative sizing and a realistic way to forecast work.

The technique creates a buffer in estimating that allows for change and uncertainty.

- Story Points: A unit of measure for expressing an estimate of the effort required to write software code. Scrum teams assign story points relative to work complexity, breadth, and amount of risk or uncertainty.
- Function Points (FP): A size measure that can be used as an input to estimate cost, effort, schedule and staffing needed to develop a software project. Function point analysis identifies logical groups of data that a project or application maintains or accesses, and logical processes that query or report that data. FP analysis counts three Input/Output (I/O) processes: external inputs, external outputs, and external inquiries.
- Simple Function Points (SiFP): Developed by Meli and acquired by the International Function Point Users Group (IFPUG) in 2019, SiFP identifies only one process: Unspecified Generic Elementary Process (UGEP) corresponding to one of the three transactional processes. SiFP identifies data function and assigns a count based on how many data elements it has.
- Source Lines of Code (SLOC): A quantitative measurement in computer programming for files that contain code from a computer programming language. The number of lines indicates the size of a file and gives an indication of the amount of work involved to write the code.

Note that some of these methods may be used in combination. For example, you could use Planning Poker to assign Fibonacci numbers to story points, the second, third, and fourth choices above. There are other characterizations of software effort, such as Level of Effort (LOE), which are not generally accepted as sizing metrics.

T-Shirt Sizing Scales

Our focus in this paper is T-Shirt Sizing, which is any use of the adjectival categories commonly associated with literal t-shirts – Extra Small (XS), Small (S), Medium (M), Large (L), Extra Large (XL), etc. – to define an ordinal scale. This scale can be denominated in hours, dollars, Story Points, Function Points, or the like.

Thought Experiment: “Double or Half?”

“You may be right/ I may be crazy/
But it just might be a lunatic/ You’re looking for”
- Billy Joel, “You May Be Right,” Glass Houses

Our going-in risk position when applying a T-shirt sizing scale is that the SME performing the assessment very well may be right, but could just easily be off by one T-shirt size in either direction. If the “true” size is a “notch” higher on the scale than the SME’s assessment, that represents an underestimate (setting the stage for an overrun); and if a notch lower, an overestimate (setting the stage for an underrun).

Macro-Level T-Shirt Sizing

Figure 3 shows one particular – and perhaps extreme! – instantiation of a T-shirt sizing scale. It is baselined at a Small effort of 1,000 hours, or about half a developer-year, and each “notch” on the scale doubles from the previous. We denote this with the ratio value R = 2. We will reference this primary scale throughout the paper.

The entire scale varies more than 500-fold, from Extra Small (XS) to Six Extra Large (6XL). We use the term “self-similar” to emphasize the fact that the scale looks the same from any point of reference: the next notch is always double, the previous notch is always half. This is an exponential scale, though examples in the physical sciences, such as the Richter scale for earthquakes and the Decibel scale for sound, usually use the term logarithmic, since taking the logarithm of the phenomenon being measured actually creates a linear scale.

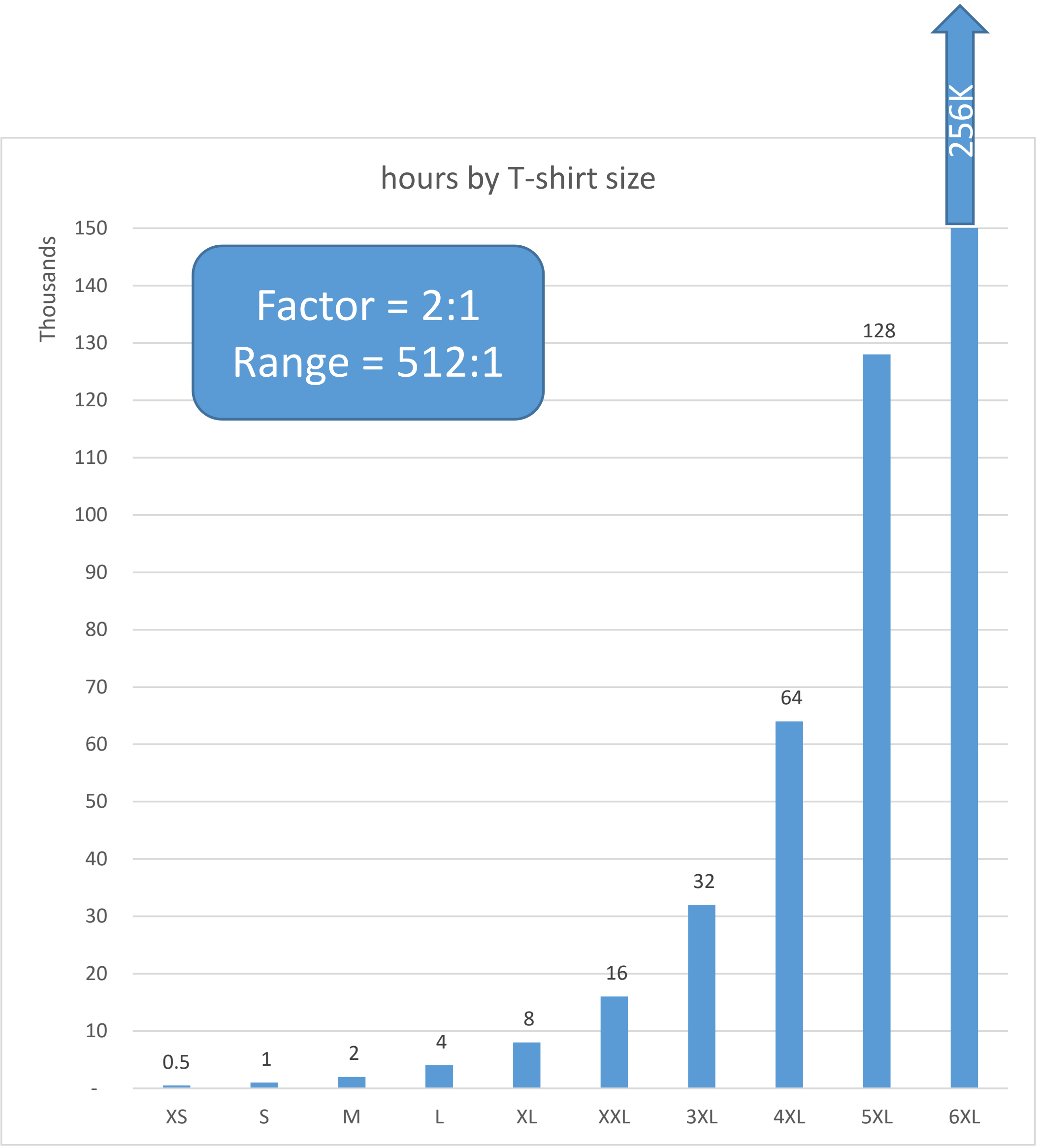


Figure 3: T-Shirt Sizing Scale

Characterizing Risk and Uncertainty

Recall that *risk* is the potential for growth of estimated quantities like size, schedule, effort, or cost, and *opportunity* is the potential for reduction. We usually use risk as an umbrella term for both risks and opportunities, since both represent deviation from a central value but in opposite directions. Our primary measure of risk in this paper is the Cost Growth Factor (CGF), which is the ratio of the expected value to the point estimate. A CGF greater than one (1.0) represents net risk; less than one, net opportunity. Throughout, we will subtract one (100%) from CGF value to isolate the growth component as a proportion of the point estimate in percentage terms.

Likewise, *uncertainty* is the variability in the estimate, reflected by properties of a probability distribution such as variance or standard deviation. Taking advantage of the self-similarity of agile scales, we generally want to express uncertainty in relative (unitless) terms. We do this by dividing the standard deviation by either the expected value (risk-adjusted estimate) or the point estimate (non-risk-adjusted). The former is the traditional coefficient of variation (CV), and we denote the latter as a pseudo-CV. We report both values throughout but treat CV as the primary metric. Note that CV is generally smaller, as it divides the same numerator by a larger denominator.

This approach is consistent with default "Parametric" cost estimating approaches like straight averages and (OLS) regressions. Those methods are generally unbiased (zero percent cost growth), and the CV is computed as the standard deviation or standard error of the estimate divided by the unbiased mean.

We will demonstrate in this paper that straightforward – if not always pretty! – math leads to growth percentages and CVs under various distributional assumptions related to our central "double or half" conceit. The roadmap is as follows.

In each instance, we'll vary the precise nature of the "double-or-half" thought experiment to establish a specific probability distribution. Next, we'll compute the mean of that probability distribution and compare it with the point estimate (H hours) to determine the resultant cost growth factor, or CGF, from which we'll subtract one (1) to express it as an expected percent growth. Then, we'll compute the variance and its square root, the standard deviation. Dividing this by the original point estimate yields a "pseudo CV," but we'll emphasize the more appropriate coefficient of variation, or CV, which necessitates dividing by the risk-adjusted mean from the previous step. We'll compile these risk and uncertainty benchmarks, both expressed as percentages.

As we proceed through the analysis, we'll move from a simple discrete case to some more realistic continuous cases. (Even though the scale itself is discrete, we believe in outcomes that are continuous for all practical purposes.) Then we'll introduce the notion of confidence to both the discrete and continuous cases. Note that the term confidence ideally reflects the true accuracy of the SMEs providing the assessment, and not their expressed confidence, which may be colored by false bravado or simply well-meaning overconfidence.

We'll also consider generalization to scale ratios other than 2.

Symmetric Uncertainty on an Asymmetric Scale

"We didn't start the fire/
It was always burning, since the world's been turning"
- Billy Joel, "We Didn't Start the Fire," Storm Front

We didn't invent geometric scales like T-shirt sizing, they were clearly found to be useful by the software development community. They are more versatile in capturing the full range of outcomes for development of simple to complex software capabilities than a linear scale would be.

You might think that plus or minus one notch would lead to uncertainty but no risk, and this would be true if the T-shirt sizing scale were linear. A constant slope as reflected in the principle of "rise over run" would mean the increase moving one notch larger would exactly cancel out the decrease moving one notch smaller, for a net expectation equal to the SME's size assessment on which the uncertainty is centered.

With the exponential T-shirt sizing scale, however, the slope is not constant but rather ever-increasing. This means that moving one notch larger incurs a greater penalty, *more than* canceling out the reduction of moving one notch smaller. With this symmetric uncertainty – equally likely to be off in either direction – applied to an asymmetric (exponential) scale, we should not be surprised that the result entails not only a significant amount of uncertainty but also some manner of risk.

Naïve Uncertainty: Coin Flips

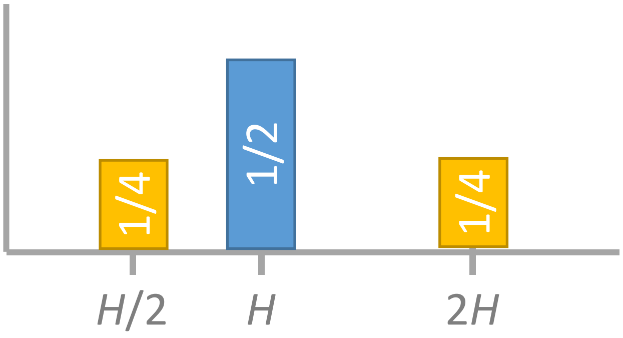


Figure 4: Discrete Distribution

Figure 3 illustrates our first and simplest scenario.

Suppose the SME's assessment is correct half the time, and that the true value is one notch higher (double) one quarter of the time and one notch lower (half) the remaining quarter of the time. We use H hours as the point estimate without loss of generality (*w.l.o.g.*). Since the scale is self-similar, the double-or-half proposition "looks" the same regardless of whether the assessed estimate is large or small.

Just as the Cincinnati Bengals won the coin toss to start Super Bowl LVI, we can model this discrete distribution as two successive flips of a fair coin: one to determine whether the SME is right or wrong; and if wrong, a second to determine whether the SME is low or high, assuming they are equally likely to under- or over-estimate.

The mean is the expected value, or the sum of products of the three possible outcome values and their respective probabilities:

$$\sum_i x_i p_i = ({}^1/_4)({}^H/_2) + ({}^1/_2)(H) + ({}^1/_4)(2H) = \frac{9H}{8} = \left(1 + \mathbf{\frac{1}{8}}\right)H$$

The CGF is 1.125, tantamount to **12.5%** growth over the point estimate, highlighted in red.

We use the shortcut formula for the variance of a probability distribution, the expected value of the square less the square of the expected value, which we just determined.

$$\sum_i x_i{}^2 p_i - \left[\sum_i x_i p_i\right]^2 = \left({}^{1}\!/_{4}\right)\left({}^{H^2}\!/_{4}\right) + \left({}^{1}\!/_{2}\right)(H^2) + \left({}^{1}\!/_{4}\right)(4H^2) - \left[\frac{9H}{8}\right]^2 = \frac{25H^2}{16} - \frac{81H^2}{64}$$

$$= \left[\boldsymbol{\frac{\sqrt{19}}{8}} H\right]^2$$

The “pseudo-CV,” i.e., the multiplier to the point estimate to get standard deviation, is highlighted in purple, approximately equal to 54.49%. The CV is the square root of 19 over nine, or about **48.43%**.

We will continue this convention throughout in the derivation of risk and uncertainty results: percentage growth over and above H hours will be highlighted in red, and pseudo-CV, as the coefficient of H before squaring in the variance calculation, will be highlighted in purple.

Problem Context: Self-Similar Scales in Agile Software Development

Continuous Risk on a Discrete Scale

As noted, T-shirt sizing takes a different approach than a traditional three-point estimate often elicited from SMEs. In that case, the SMEs are allowed to choose any range of minimum (optimistic), most likely, and maximum (pessimistic) values. This allows considerable flexibility, but SMEs generally have a poor track record accurately characterizing uncertainty in this approach. (Braxton & Coleman, Teaching Pigs to Sing: Improving Fidelity of Assessments from Subject Matter Experts (SMEs), 2012) While T-shirt sizing appears to be a one-point estimate – the SME provides one and only one size – the premise of this paper is that the risk analyst is actually getting a three-point estimate "for free," where the min and max are the adjacent sizes or "notches" on the scale. By providing the SME a "drop-down menu" with limited choices, we simplify the sizing process and enable them to focus on the basic judgment, steering clear of false precision. This also represents "sailor-proofing" of a sort, helping guard against wildly inappropriate intervals. Because we generally believe that risk is right skew, the exponential scale is appropriate. Because it is self-similar, a notion drawn from fractal geometry, the uncertainty will be proportionally identical at each step of the scale. Even though these steps are discrete, outcomes are continuous, and we should model them as such.

“Maximum” Uncertainty: The Uniform Distribution

Figure 4 represents our first potential refinement to the probability model. We are still treating “double or half” as the endpoints, but reflecting the more realistic assumption of a continuous distribution of outcomes. Though cost and risk analysts tend to leap to a three-point estimate (like Triangular), we start with Uniform as the most naïve continuous distribution.

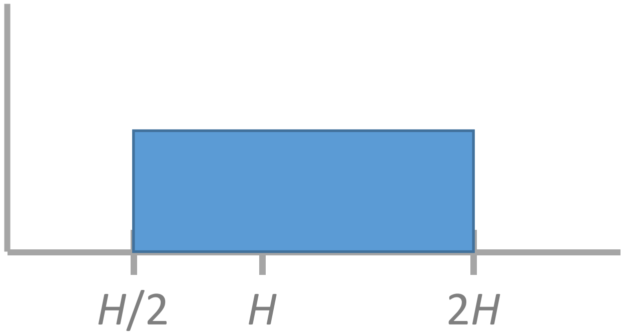


Figure 5: Uniform Distribution

The mean of a Uniform distribution is simply the average of the min and max values, which is equivalent to the middle (median) value.

$$\frac{{}^{H}\!/_{2} + 2H}{2} = \frac{5H}{4} = \left(1 + \frac{\mathbf{1}}{\mathbf{4}}\right) H$$

This represents **25.0%** growth, highlighted in red. The variance is the range squared divided by 12.

$$\frac{\left(2H - {}^{H}\!/_{2}\right)^2}{12} = \frac{9H^2}{4 \cdot 12} = \left[\sqrt{3} \cdot \frac{H}{4}\right]^2 = \left[\frac{\boldsymbol{\sqrt{3}}}{\mathbf{4}} H\right]^2$$

The pseudo-CV highlighted in purple is the square root of three divided by four, or about 43.30%. The CV is the square root of three divided by five, or about **34.64%**.

Note that by the very nature of continuous probability distributions, we are forced to abandon the second “coin flip,” wherein the SME is equally likely to over- or under-estimate. In this case, the probability of an underestimate (overrun) is twice that of an overestimate (underrun), since it is directly proportional to the ratio of the bases of the two half-rectangles: (2H-H) = H and (H – H/2) = H/2, respectively. This can be seen in the “cartoon” above.

“Standard” Uncertainty: The Triangular Distribution

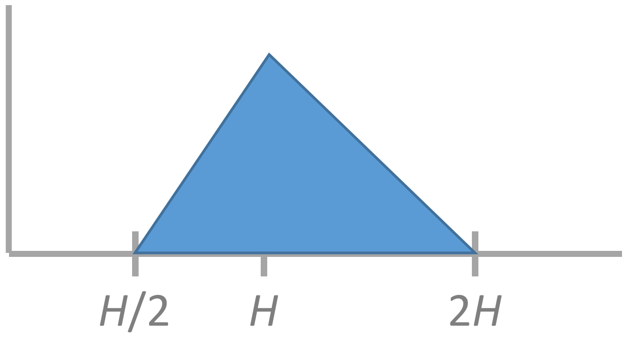


Figure 6: Triangular Distribution

Figure 5 reflects an alternative to the Uniform distribution. As with the Uniform, the Triangular assumes the “double-or-half” interval represents the entire range of possible outcomes. Whereas the Uniform arguably gives *too* much weight to the tails, the Triangular perhaps gives not enough. While this Triangular is asymmetric by design, a good rule of thumb to remember is that the half-base of a symmetric Triangular must be divided by the square root of six, or almost two and a half, to yield the standard deviation.

The mean of a Triangular is the average of the min, most likely, and max. (The intuition that the middle value should be more heavily weighted proves incorrect!)

$$\frac{H/_2 + H + 2H}{3} = \frac{7H}{6} = \left(1 + \mathbf{\frac{1}{6}}\right)H$$

This represents **16.7%** cost growth, highlighted in red.

The variance is sum of squares of the three parameters less the sum of their pairwise products, all divided by 18.

$$\frac{\left(H/_2\right)^2 + H^2 + (2H)^2 - H^2/_2 - H^2 - 2H^2}{18} = \frac{7H^2/_4}{18} = \frac{7H^2}{2\cdot 36} = \left[\sqrt{\frac{7}{2}}\cdot\frac{H}{6}\right]^2 = \left[\mathbf{\frac{\sqrt{14}}{12}}H\right]^2$$

The pseudo-CV highlighted in purple is the square root of 14 over 12, or about 31.18%. The CV is the reciprocal of the square root of 14, or about **26.73%.**

Note that both the CGF and CV are reduced from the Uniform case, since more probability is naturally concentrated near the peak of the triangle, the SME’s “most likely” assessment.

“Standard” Risk: The Lognormal Distribution

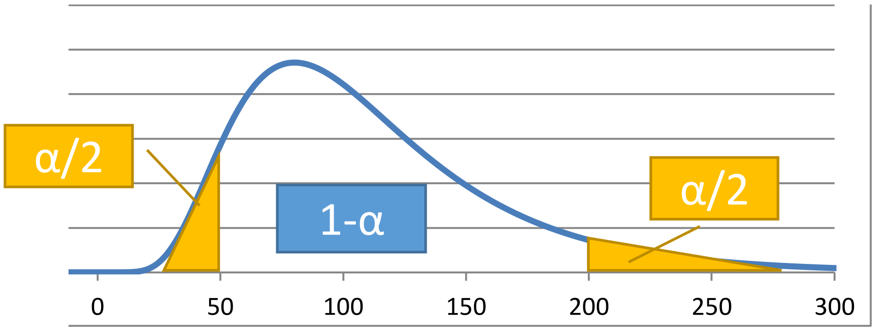


Figure 7: Lognormal Distribution

For many cost and schedule risk applications, the Lognormal is thought to be the preferred distribution. As we apply the Lognormal to our present scenario, we immediately notice the issue illustrated in Figure 6. Whereas the Uniform and Triangular are finite distributions, the Lognormal has an infinite right tail. For this reason, our 2H point (“double”) cannot be at the extreme end of the distribution. We are forced to define a right-hand tail probability which, however small, must be nonzero. For reasons of symmetry and convenience, we introduce an equal probability for the (finite) left tail, and we denote them both alpha over two. This leaves the majority of the probability, specifically one minus alpha, in the “double-or-half” range between *H/2* and *2H*. We’ll discuss this choice of notation further in the next section.

Because the log of a Lognormal is Normal, the “double-or-half” range of the self-similar scale naturally translates into a symmetric confidence interval in the transformed space (aka “related normal”). In this case, we assume that the SME estimate represents median, which for the Lognormal falls to the right of the mode (peak) but the left of the mean (balancing point). The interval is $(lnH - ln2, lnH, lnH + ln2)$. If we translate this to a standard normal by subtracting the mean (lnH) and dividing by the standard deviation, the Z-score for the top end of the interval becomes ln2 divided by sigma. Since the right-tail probability is alpha over two, we use the corresponding percentile and inverse CDF of the standard normal to solve for sigma, as shown below.

$$\Phi^{-1}\left(1 - {}^{\alpha}/_{2}\right) = \frac{ln2}{\sigma}$$

$$\sigma = \frac{ln2}{\Phi^{-1}\left(1-{}^{\alpha}/_{2}\right)} = \frac{1}{log_2 e^{\Phi^{-1}\left(1-{}^{\alpha}/_{2}\right)}}$$

From sigma, the standard deviation of the related normal, we can directly calculate the CV of the lognormal itself. (Braxton & Sayer, Probability Distributions for Risk Analysis, 2013)

$$CV = \sqrt{e^{\sigma^2}-1}$$

Since the mean of the lognormal is $e^{\mu+\frac{\sigma^2}{2}}$, the cost growth relative to the point estimate (median) is:

$$CGF = e^{\frac{\sigma^2}{2}} = \sqrt{1+CV^2}$$

Figure 8 below illustrates the increase in CGF and CV as the percent chance of being outside the “double-or-half” range (as denoted by alpha) increases. Higher alpha values results in an increasingly skewed lognormal with a heavy right tail. We truncate the graph at alpha equals one half, the proverbial “coin flip,” as CGF and CV values beyond that are impractically high.

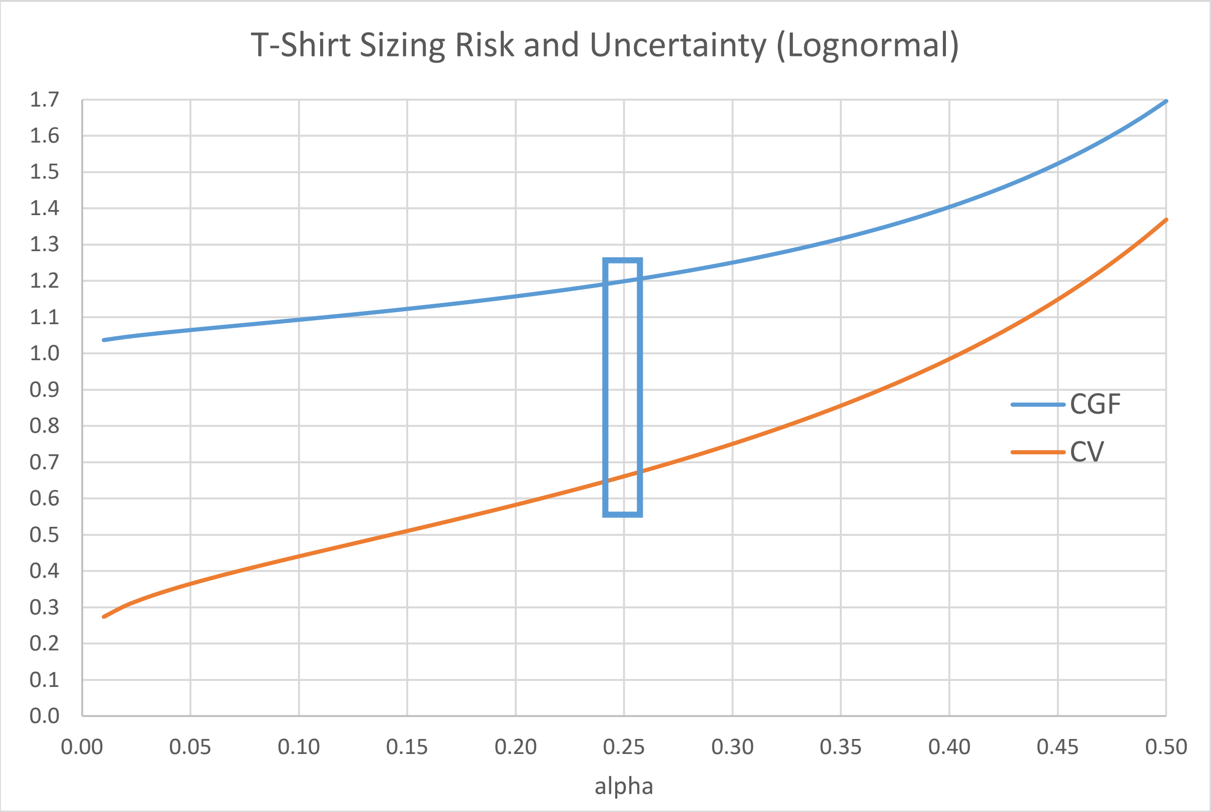


Figure 8: Lognormal CGF and CV as a Function of Confidence

Generalization #1: Confidence Level

This is a good juncture to back up and formally introduce our notion of confidence.

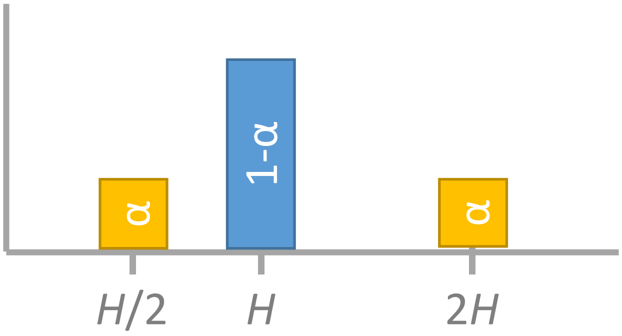


Figure 9: Discrete Distribution (with Confidence)

Let us return momentarily to the discrete case, retaining the second coin flip as fair, with an equal probability of over- or underestimate when the SME is wrong. Now, however, the first coin flip becomes unfair. We borrow the notion (and notation) of significance level from hypothesis and denote as alpha the sum of the two tail probabilities, or the chance the SME is wrong. The chance the SME is right then becomes the complement, or one minus alpha.

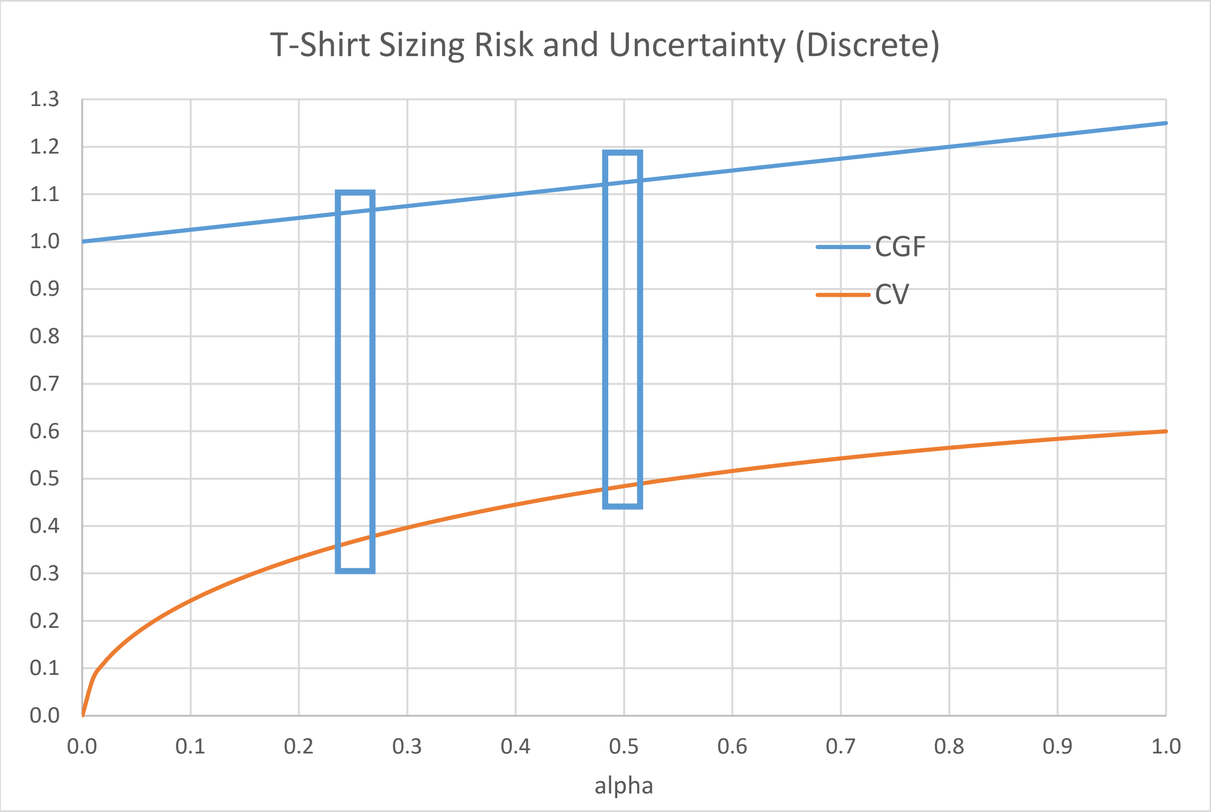


Figure 10: Discrete CGF and CV as a Function of Confidence

When alpha = 0, our estimate is certain, and both cost growth and CV shrink to zero. When alpha = 1, we are 100% wrong, and we have a “barbell” split between the two adjacent sizes, H/2 and 2H. In this case, the two “tails” in their entirety are the adjacent “buckets” or “notches” on the scale. Growth maxes out at 25%, and CV maxes at 60% (pseudo-CV of 75%), as illustrated in the graph. The equations behind these graphs are derived below.

Once again, the expected value is the sum of the products of the possible outcomes and their associated probabilities. The former remain unchanged, but the latter now reflect the “unfair” first coin flip. Cost growth as a function of alpha is highlighted in red.

$$\sum_i x_i p_i = ({}^{\alpha}/_{2})({}^{H}/_{2}) + (1-\alpha)(H) + ({}^{\alpha}/_{2})(2H) = \left(1+\boldsymbol{\frac{\alpha}{4}}\right)H$$

As always, we use the variance shortcut, the expected value of the square minus the square of the expected value.

$$\sum_i {x_i}^2 p_i - \left[\sum_i x_i p_i\right]^2 = ({}^{\alpha}/_{2})\left({}^{H^2}/_{4}\right) + (1-\alpha)(H^2) + ({}^{\alpha}/_{2})(4H^2) - \left[\left(1+\frac{\alpha}{4}\right)H\right]^2 =$$

$$\left(1+\frac{9\alpha}{8}\right)H^2 - \left(1+\frac{\alpha}{2}+\frac{\alpha^2}{16}\right)H^2 = \frac{10\alpha-\alpha^2}{16}H^2 = \left[\frac{\boldsymbol{\sqrt{10\alpha-\alpha^2}}}{\mathbf{4}}H\right]^2$$

The pseudo-CV is highlighted in purple above, and the CV is below. Both are now expressed as a function of alpha.

$$CV = \frac{\sqrt{10\alpha-\alpha^2}}{4+\alpha}$$

The previous specific case of alpha equals one half is called out on the graph, with a CGF of 1.125 (12.5% growth, linearly interpolated between 0% and 25%) and CV of 48.43%. Also, the perhaps more realistic case of alpha equals one fourth. Because CGF is linear with alpha, this equates to half as much cost growth, or 6.25%. However, since CV is quadratic, it is only reduced to 36.75%, or by about a quarter.

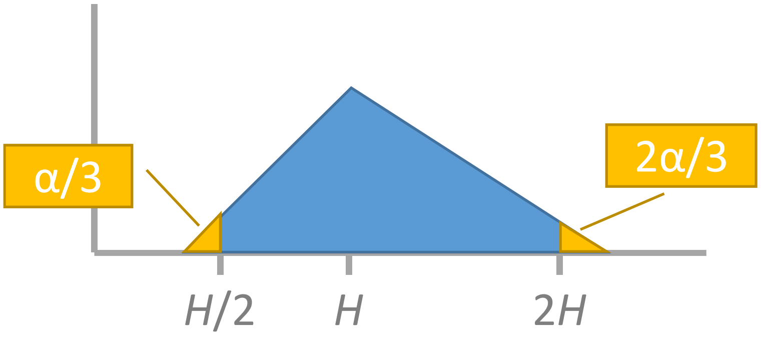


Figure 11: Triangular Distribution with Confidence (Proportional Tails)

We can now extend the traditional three-point estimate Triangular distribution to incorporate our confidence construct. Again, for reasons of convenience and convention (not to mention aesthetics!), we extend the two tails proportionally, maintaining the one-to-two ratio of the left and right half bases. With the tail probability split by the same ratio, a common scale factor can be determined. Because the small orange right triangles are similar to the large blue right triangles they overlap, the respective ratios of their areas (orange to blue) is equal to the *square* of the ratio of a corresponding pair of sides. Thus, the scale factor is seen to be a function of the *square root* of alpha. Note that this formulation is called the Trigen distribution in some Monte Carlo simulation tools.

One might consider a practical limit on alpha in this case to be when the left tail extends all the way to zero (0). When this happens, both half-bases have doubled, so that the expanded Triangular has become T(0, H, 3H), maintaining the two-to-one ratio of half-bases. The two tail triangles are both half as wide and half as tall as the half-triangles in which they are embedded, and therefore have one fourth the area. Thus, this corresponds to alpha equals one fourth, highlighted on the graph below. For any values of alpha greater than one fourth, the left tail extends beyond zero and into the negatives, which is often considered untenable for cost estimating. (The developer will never pay *us* for the privilege of developing the software!) As long as the occasional negative values are not too extreme, however, this will generally “come out in the wash” of a Monte Carlo simulation.

The expected value is again the average of the min, most likely, and max.

$$\mu = \left[\left(1-\frac{\sqrt{\alpha}}{1-\sqrt{\alpha}}\right)\frac{H}{2}+H+\left(2+\frac{\sqrt{\alpha}}{1-\sqrt{\alpha}}\right)H\right]\Big/_{3} = \left(1+\frac{\mathbf{1}}{\mathbf{6-6\sqrt{\alpha}}}\right)H$$

The cost growth as a function of alpha is highlighted in red above. The variance is again the sum of the squares of min, most likely, and max, less the sum of their pairwise products, all divided by 18. A derivation is given in the appendix.

$$\sigma^2 = \left[\frac{\sqrt{\frac{7-4\sqrt{\alpha}}{2}}}{6-6\sqrt{\alpha}} H\right]^2$$

The pseudo-CV is highlighted in purple above, and the CV is given below. Substituting alpha equals zero, we can check that the CGF and CV reduce to the previous Triangular values.

$$CV = \frac{\sqrt{\frac{7-4\sqrt{\alpha}}{2}}}{7-6\sqrt{\alpha}}$$

Alpha = 0.25 is highlighted on the graph, with **33.3%** cost growth and CV = **39.53%.**

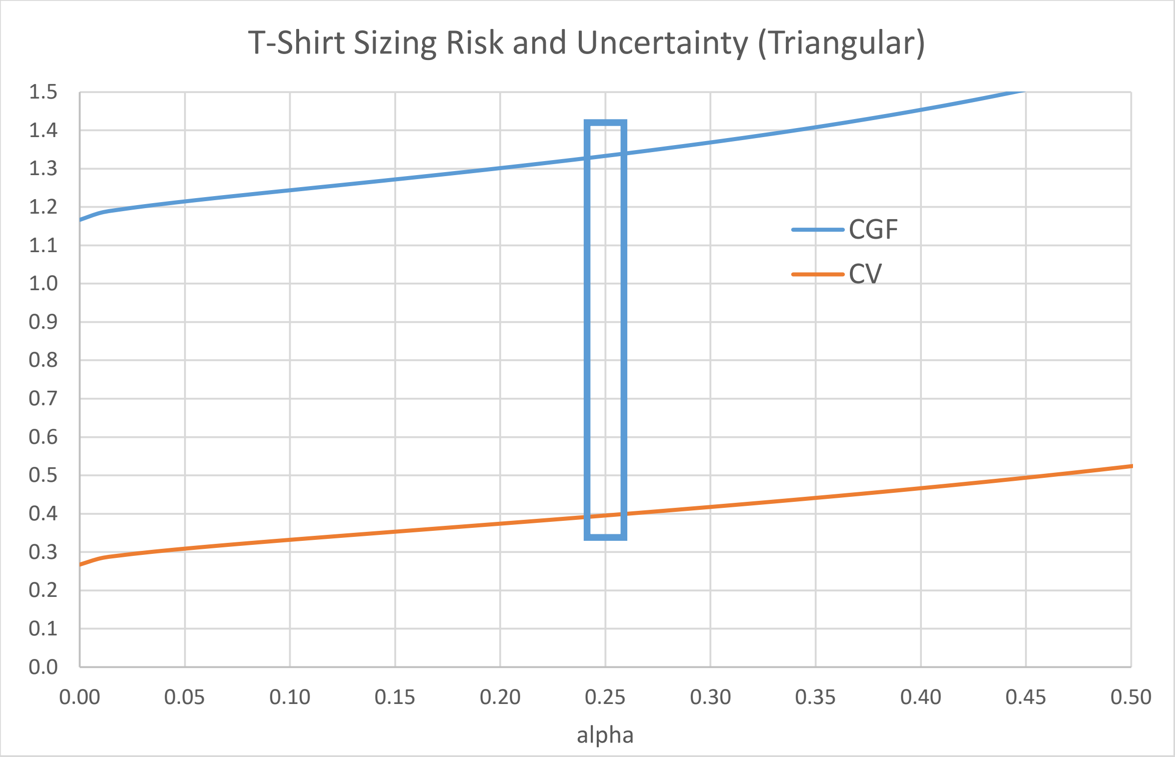


Figure 12: Triangular CGF and CV as a Function of Confidence

Proportional or Symmetric Tails

As the Lognormal, Discrete, and Triangular examples are introduced, our new question is, what if things get “worse than double” or “better than half”?! When the author faced the famous Canadian provinces question for $250,000, it was now a “double or quarter” proposition[2]. (Note that for project execution, the left tail is good, but for game show winnings it is bad!)

When extending the tails, the question arises as to whether the tail probability should be split proportionally or symmetrically. Because of the Lognormal’s relationship with the underlying normal, it can kind of have its cake and eat it, too. In our formulation, the tail probability was split evenly, which creates literal symmetry of tails for the related normal in transformed space. In the original space, the infinite right tail has more influence on growth and CV than the finite left tail.

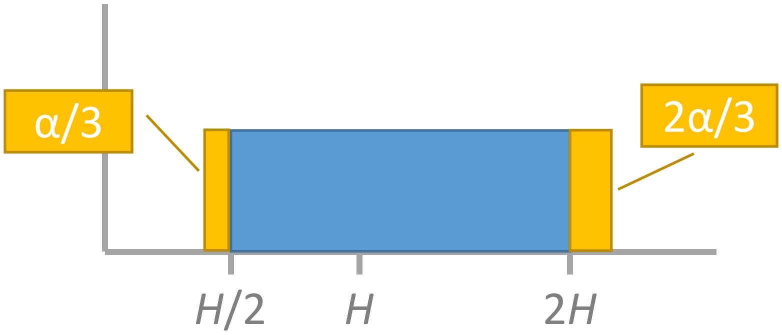


Figure 13: Uniform Distribution with Confidence (Proportional Tails)

Figure 12 shows the natural extension of the Uniform distribution to go beyond “double or half” where the two tails are split in proportion to the half-bases, i.e., one-third/two-third, just as previously demonstrated for the Triangular. While the height of the rectangle in the tailless case of the Uniform was 2/(3H), to ensure the total probability under the “curve” was one, the height is now reduced by a factor of one minus alpha to accommodate the tail probabilities.

[2] The author rattled off all ten provinces, but shamefully omitted Nunavut amongst the territories. Had he been wrong, he would’ve dropped down to $32,000, or roughly one-fourth of the $125,000 previously earned.

Once again, the mean is the average of the min and max:

$$\mu = \left[\frac{(1-2\alpha)}{(1-\alpha)}\frac{H}{2} + \frac{(2-\alpha)}{(1-\alpha)}H\right]\Big/_{2} = \frac{5-4\alpha}{4-4\alpha}H = \left(1 + \boldsymbol{\frac{1}{4-4\alpha}}\right)H$$

Cost growth as a function of alpha is highlighted in red.

The variance is the range squared divided by 12:

$$\sigma^2 = \frac{(3H)^2}{12[2(1-\alpha)]^2} = \left[\boldsymbol{\frac{\sqrt{3}}{4-4\alpha}}H\right]^2$$

The pseudo-CV as a function of alpha is highlighted in purple. The CV is given below.

$$CV = \frac{\sqrt{3}}{5-4\alpha}$$

If instead the tails are split equally, we have the slightly different scenario illustrated in Figure 13. The dimensions of the rectangle are the same, but it is shifted so that the tails contain equal probability, namely alpha over two.

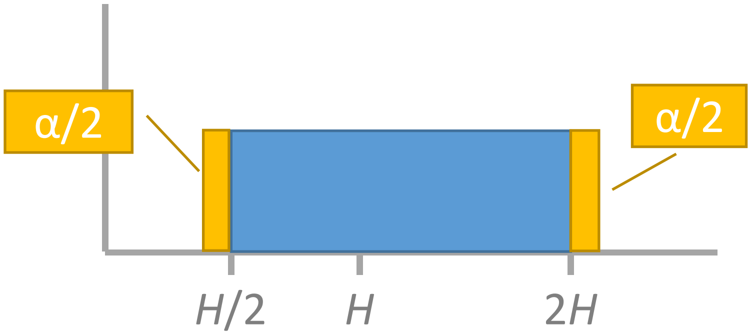


Figure 14: Uniform Distribution with Confidence (Equal Tails)

Notice that in this case the cost growth highlighted in red is *not* a function of alpha, because the distribution balances on the same point (25% growth) regardless of how long or short the symmetric tails are.

$$\mu = \left[\frac{(2-5\alpha)}{(4-4\alpha)}H + \frac{(8-5\alpha)}{(4-4\alpha)}H\right]\Big/_{2} = \frac{5}{4}H = \left(1 + \frac{\mathbf{1}}{\mathbf{4}}\right)H$$

The variance, and with it the pseudo-CV highlighted in purple, are the same as the Proportional Tails case, because the spread of the distribution is the same, but the CV is slightly different.

$$\sigma^2 = \frac{(6H)^2}{12[4(1-\alpha)]^2} = \left[\frac{\boldsymbol{\sqrt{3}}}{\boldsymbol{4-4\alpha}}H\right]^2$$

$$CV = \frac{\sqrt{3}}{5-5\alpha}$$

The case of Triangular with Symmetric Tails is deferred to the appendix.

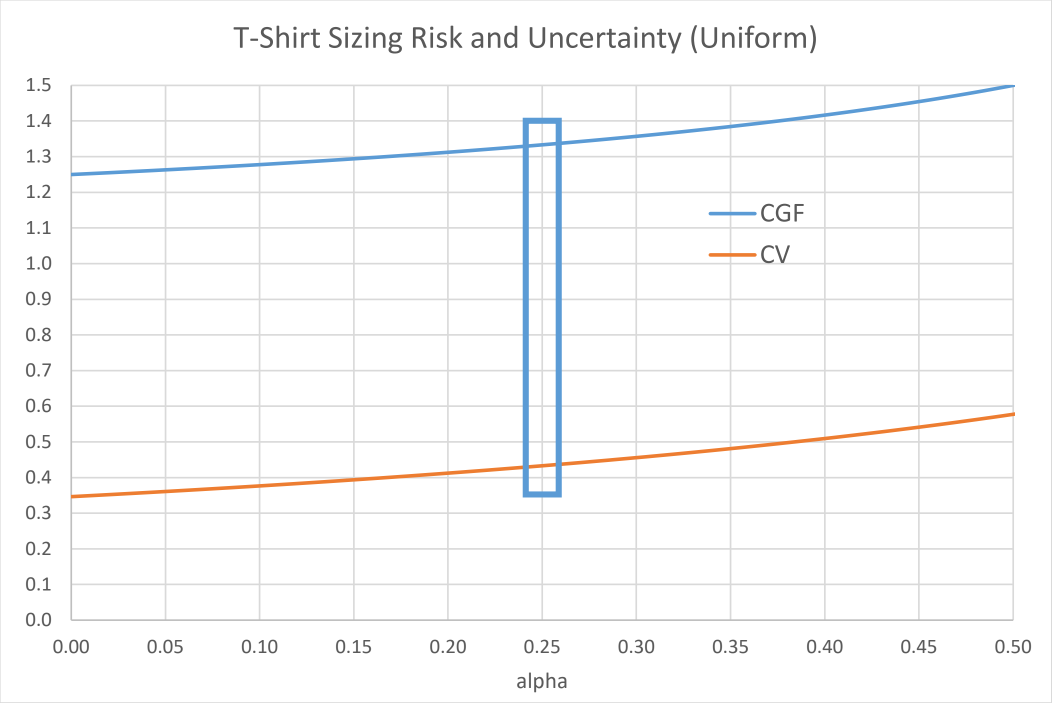


Figure 15: Uniform CGF and CV as a Function of Confidence (Proportional Tails)

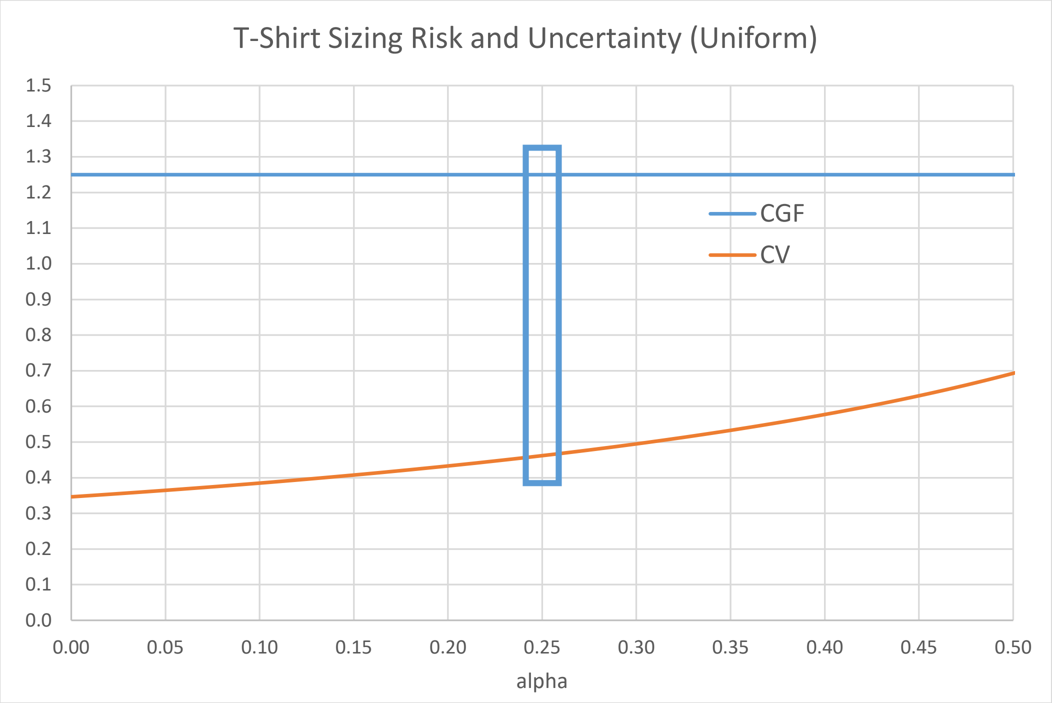


Figure 16: Uniform CGF and CV as a Function of Confidence (Equal Tails)

Here is a handy summary of all the previous results generalized with the confidence parameter, alpha. The first pair of columns shows the Growth and CV percentages as a function of alpha, and the second pair of columns provides specific values for the case where alpha equals one fourth. It is left as an exercise for the student to determine appropriate upper bounds on alpha for each scenario.

Table 1: Risk and Uncertainty as a Function of Confidence

	Growth %	CV	Growth % (α = 0.25)	CV (α = 0.25)
Discrete (Generalized)	$\frac{\alpha}{4}$	$\frac{\sqrt{10\alpha - \alpha^2}}{4 + \alpha}$	6.2%	36.74%
Lognormal	$\sqrt{1 + CV^2} - 1$	$\sqrt{e^{\sigma^2} - 1}$	19.9%	66.16%
Uniform (Proportional)	$\frac{1}{4 - 4\alpha}$	$\frac{\sqrt{3}}{5 - 4\alpha}$	33.3%	43.30%
Uniform (Equal)	$\frac{1}{4}$	$\frac{\sqrt{3}}{5 - 5\alpha}$	25.0%	46.19%
Triangular (Proportional)	$\frac{1}{6 - 6\sqrt{\alpha}}$	$\frac{\sqrt{\frac{7 - 4\sqrt{\alpha}}{2}}}{7 - 6\sqrt{\alpha}}$	33.3%	39.53%

For the Lognormal case, it is not immediately apparent how Growth and CV are functions of alpha, but recall that $\sigma = \frac{ln2}{\Phi^{-1}(1 - \alpha/2)}$. The general logical flow is to calculate variance of the related normal (sigma squared) first, then the CV of the lognormal as a function of sigma squared, and finally the Growth as a function of CV (right to left).

As is often the case in risk, some distributions – in this case, the Lognormal – indicate lower growth but higher CV, whereas others – in this case, Uniform and Triangular – indicate higher growth but lower CV. Fortuitously, these tend to have similar impacts at high percentiles of interest, such as the 70th and 80th.

Thus far, we have generalized our distributions with the confidence of the SME but maintaining the scale ratio of two ($R = 2.0$). Now we turn to generalizing the scale ratio and the context in which it is used for assessment.

Planning Poker, Fibonacci Numbers, and the Golden Mean

Micro-Level T-shirt Sizing

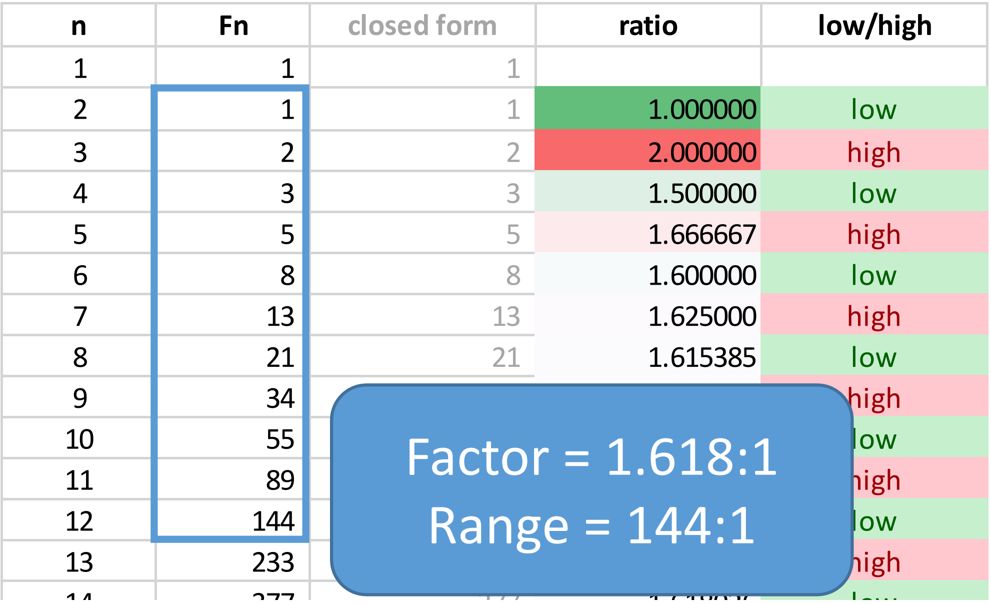
T-shirt sizing as described is a “macro-level” method, assessing Program Epics (PEs) very early in development when very little detail is available. By contrast, in a typical agile implementation, scrum teams conduct planning for three-month program increments and three-week sprints. This involves “micro-level” assessments of work to be completed in the immediate future. We posit that *a priori* SME assessments at these two levels are of comparable accuracy and precision. If Risk is indeed fractal, then these assessments should “look” the same whether they are on a large scale or a small scale.

One possible counterargument is that the micro-level assessments should be more accurate because the assessor has more direct control over the execution of the work, which will be conducted more or less right away. At this detailed planning stage, we would hope that there are relatively few unknowns. A possible counterargument in the opposite direction invokes the “size effect” in risk analysis, wherein smaller efforts are often “allowed” to overrun by a greater percentage because they have a smaller absolute impact. If a two-story-point task doubles in size, it’s no big deal, but if a 16,000-hour Epic doubles in size, that’s a significant impact to the program.

Before we return to these questions of accuracy, we introduce an alternate sizing method commonly used in Planning Poker, wherein individual development tasks are sized in Story Point using Fibonacci numbers. Recall that Fibonacci numbers, the eponym of Italian mathematician Leonardo da Pisa (literally “son of Bonacci”), are the sequence starting with 1 and 1, and whose subsequent entries are the sum of the two previous numbers:

$$1+1=2, 1+2=3, 2+3=5, 3+5=8, 5+8=13, 8+13=21, 13+21=34, \cdots$$

Table 2: Fibonacci Numbers and Successive Ratios



n	Fn	closed form	ratio	low/high
1	1	1		
2	1	1	1.000000	low
3	2	2	2.000000	high
4	3	3	1.500000	low
5	5	5	1.666667	high
6	8	8	1.600000	low
7	13	13	1.625000	high
8	21	21	1.615385	low
9	34			high
10	55			low
11	89			high
12	144			low
13	233			high
14	377			low

Factor = 1.618:1

Range = 144:1

In some alternative formulations, larger sizes are replaced with “rounder” numbers, such as 50 and 100 instead of 55 and 89. At some point, an “infinity” (∞) designation is used for a task that is too large to effectively size individually and which should be broken down into smaller tasks prior to sizing. The scale is often visualized using fruits! (This was perhaps the subliminal inspiration for the later idea of an analogized scale.)

One appeal of using Fibonacci numbers is that they combine “additive” and “multiplicative” properties. The additive aspect is more apparent, since the sum of any two consecutive sizes is equal to the next largest size. The SME may think, consciously or subconsciously, “This work is about equivalent to a 2 and a 3, therefore it must be a 5.” The multiplicative aspect emerges when we examine the ratio of consecutive terms and notice that it appears to converge quickly to a constant. It turns out that this constant is a very special number in geometry in particular and mathematics in general, the Golden Ratio or Golden Mean. (Livio, 2008)

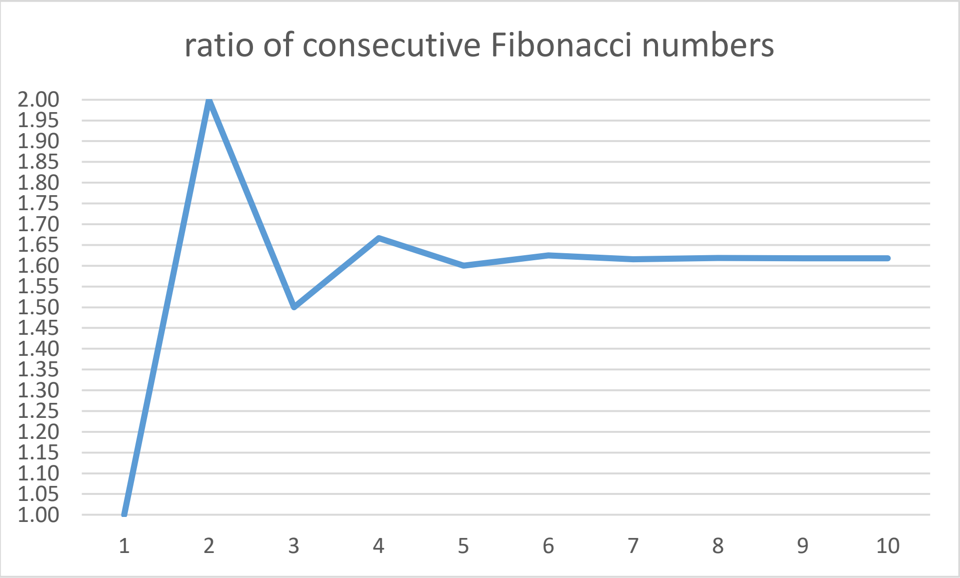


Figure 17: Converging Ratios of Consecutive Fibonacci Numbers

$$F_n = \frac{1}{\sqrt{5}}[\phi^n - (1-\phi)^n]$$

See Appendix A: Fibonacci Sequence Convergence for a derivation of the above closed-form formula for the nth Fibonacci number, which implies convergence of the ratio of success terms to the Golden Mean, shown below in exact and approximate terms.

$$\phi = \frac{1+\sqrt{5}}{2} = 1.618\ldots$$

Preliminary results show that at a micro level, a good rule of thumb is that SMEs will be correct about a third of the time, will underestimate about a third of the time, and will overestimate about a third of the time. In fact, analysis of thousands of line items shows that a binomial distribution with probability 0.5 is a pretty good approximation of micro-level sizing accuracy. The notional graph is shown in Figure 18 below. The horizontal axis label indicates the number of notches on the scale the SME was off. A zero represents the correct notch; a negative two (-2) represents an *over*-estimate by two notches; and a positive three (+3) represents an *under*-estimate by three notches. The empirical data shows spikes in the zero and negative one (-1) bins of the histogram.

This indicates that SMEs are accurate or minimally overestimate more often than predicted by the notional distribution. Further research and data aggregation may enable a more detailed releasable version of these findings in the near future.

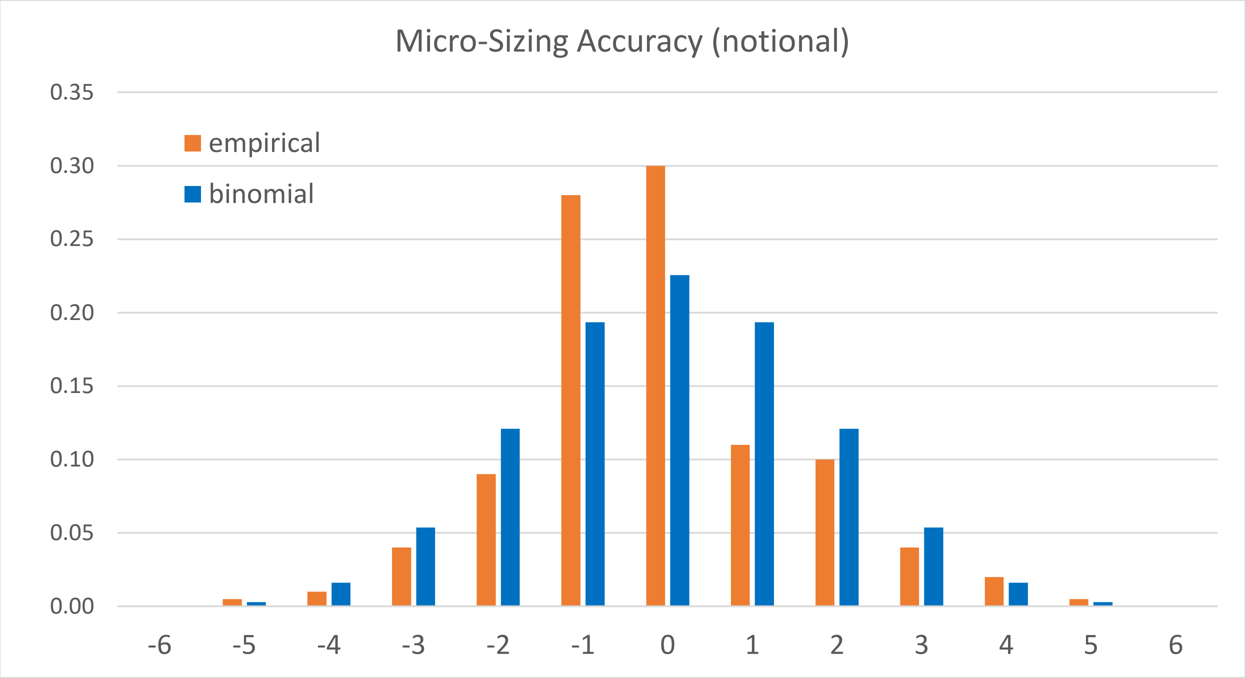


Figure 18: Micro-Sizing Accuracy (notional)

The empirical distributions in this case do fail the chi square test with an assumed distribution of binomial with probability equals one half, but the maximum likelihood estimator (MLE) for binomial probability does come out eerily close to 0.50. This distribution has a certain appeal as an extension to our original thought experiment. Instead of one coin flip, there are five or six coin flips, and with each heads the SME moved one notch to the right, and with each tail one to the left. The sum of those individual (iid) Bernoulli trials is a Binomial, which by the Central Limit Theorem (CLT) converges fairly rapidly to a Normal. This can be visualized with a *quincunx*, or returning to our game show theme, the Plinko board on *The Price Is Right*. This distribution is symmetric, but since it is imposed of the asymmetric self-similar scale, the result is expected growth and greater uncertainty.

While these findings may not be immediately applicable to macro-level sizing, there are two important things to note. First, there is quite a range of under- and overestimates at the micro level. Whereas our original thought experiment only allowed the SME to be off by one notch in either direction, the evidence here is that they can easily be off by two or three notches or more in either direction. (This is a compilation across a wide range of initial size estimates. Extreme underruns are only possible for large initial estimates, whereas for small initial estimates, there is “nowhere to go but up.”) Second, these initial results corroboratehigh the principle of (near) symmetry on an asymmetry scale.

Alternate Sizing Scales

Thus far, we have examined a T-shirt sizing scale with an explicit ratio of two (2.0) between sizes, and Fibonacci numbers with an implicit ratio that approaches the golden mean (1.618…). These may both seem a little abstract and arbitrary, but we now introduce a third notional sizing model that was constructed more as a build-up but behaves in very much the same way.

Table 3: Notional Sizing Model Parameters

Sked (mo)	S	M	L
E	12	15	18
M	15	18	21
C	18	21	24

LOE (FTE)	S	M	L
E	2.5	3	3.5
M	3	4.5	6
C	3.5	6	8.5

effort (PM)	S	M	L
E	30	45	63
M	45	81	126
C	63	126	204

effort (relative)	S	M	L
E	37.0%	55.6%	77.8%
M	55.6%	100.0%	155.6%
C	77.8%	155.6%	251.9%

This Notional Sizing Model actually purports to incorporate both Size and Complexity. Each is assessed on a three-point scale: Small (*S*), Medium (*M*), or Large (*L*); and Easy (*E*), Moderate (*M*), or Complex (*C*), respectively. This pair of assessments places the effort in one particular square of a series of 3 x 3 tables, which simultaneously determines schedule duration (first sub-table, between 12 and 24 months); level of effort (second sub-table, between 2.5 and 8.5 full-time equivalents); and total effort (person-months, between 30 and 204). The third is of course just the product of the first two. The fourth sub-table expresses this effort calibrated to a percentage of the “middle” square, representing size Medium and complexity Moderate.

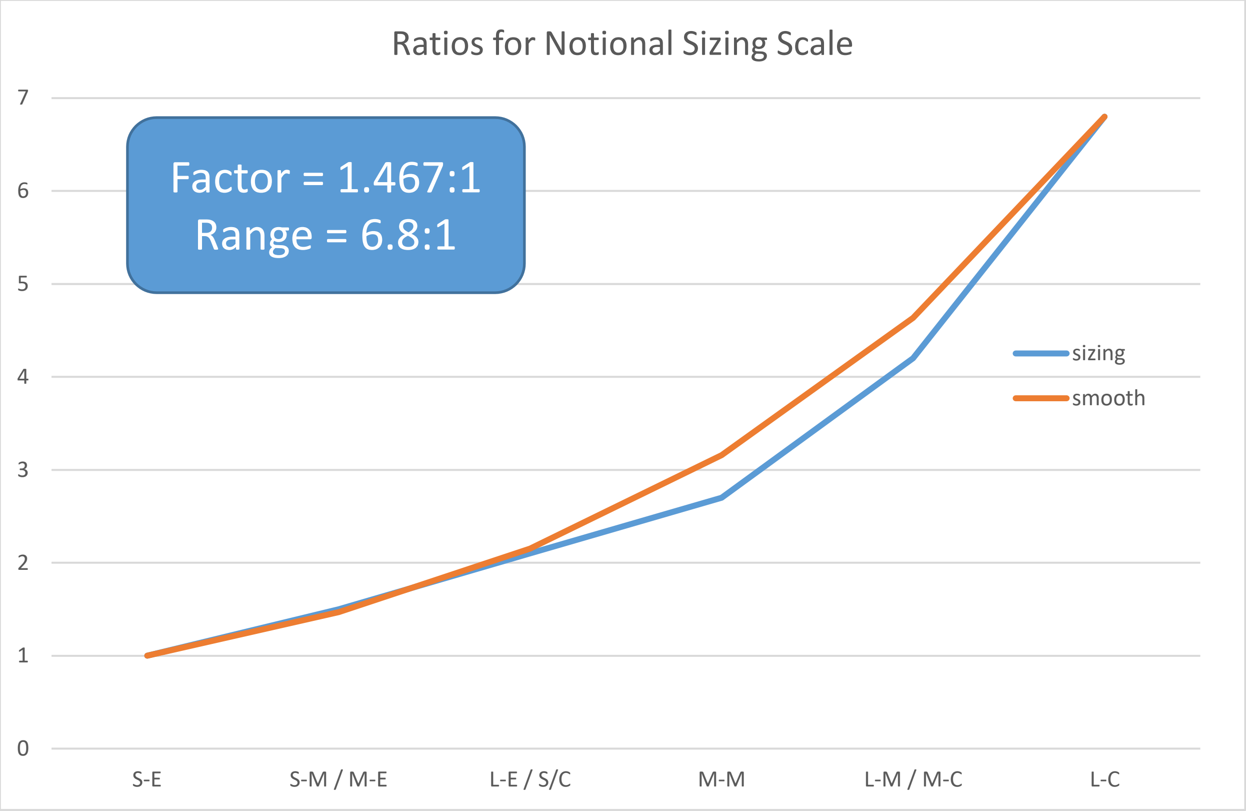


Figure 19: Notional Sizing Model

At this point, it becomes apparent that the additional (implicit) assumption of symmetry maps this 3 x 3 model to 6-point scale, wherein there are three pairs of equivalent assessments as shown on the horizontal axis of Figure 19 above: Small-Moderate and Medium-Easy; Small-Complex and Large-Easy; and Medium-Complex and Large-Moderate. The largest defined notch on the scale is 6.8 times as large as the smallest. This allows us to calculate the (geometric) average ratio between notches, which is illustrated by the orange "smooth" series on the graph.

$$6.8^{1/5} = 1.467\cdots$$

Note that "sizing" is generally intended to be an objective assessment, i.e., a property of the code itself. However, these sorts of assessments that generate size estimates in hours or dollars may implicitly take into account productivity and even labor rates as well. We will revisit the question of decomposition of risk and uncertainty at the end of the paper.

Generalization #2: Scale Ratio

We have heretofore introduced three concrete examples of SME-driven sizing, each with its own scale ratio. There is no reason we cannot repeat our previous derivations with a generalized scale ratio, *R*. Our erstwhile "double or half" values now become *RH* and *H/R*, respectively.

Though the Lognormal equations involve the most specialized functions, they are actually easiest to generalize! The confidence interval in log space now becomes:

$$(lnH - lnR, lnH, lnH + lnR)$$

The derivation of the standard deviation of the related normal is the same, with *R* replacing *2* in the logarithms.

$$\Phi^{-1}\left(1 - {}^{\alpha}\!/_{2}\right) = \frac{lnR}{\sigma}$$

$$\sigma = \frac{lnR}{\Phi^{-1}\left(1 - {}^{\alpha}\!/_{2}\right)} = \frac{1}{log_R e^{\Phi^{-1}\left(1 - {}^{\alpha}\!/_{2}\right)}}$$

The equations for CV (of the lognormal) in terms of variance (of the related normal) and for CGF in terms of CV remain the same.

We can now begin to understand the interplay between ratio and confidence. Figure 20 shows Lognormal CGF and CV as a function of alpha for the three different R values presented above. For example, a CV of 50% is achieved for an alpha of about 15% for the T-shirt scale with ratio 2.0, but to achieve that same CV takes an alpha of about 30% for the Fibonacci scale with ratio 1.618 and more than 40% for the Notional model with ratio 1.467. Intuitively, SMEs should have about the same inherent accuracy, so that the reported confidence should be dependent on the ratio of the scale being used. As a loose analogy from cost estimating, the same nominal learning curve slope (LCS) means different things depending on which learning curve theory is being applied.

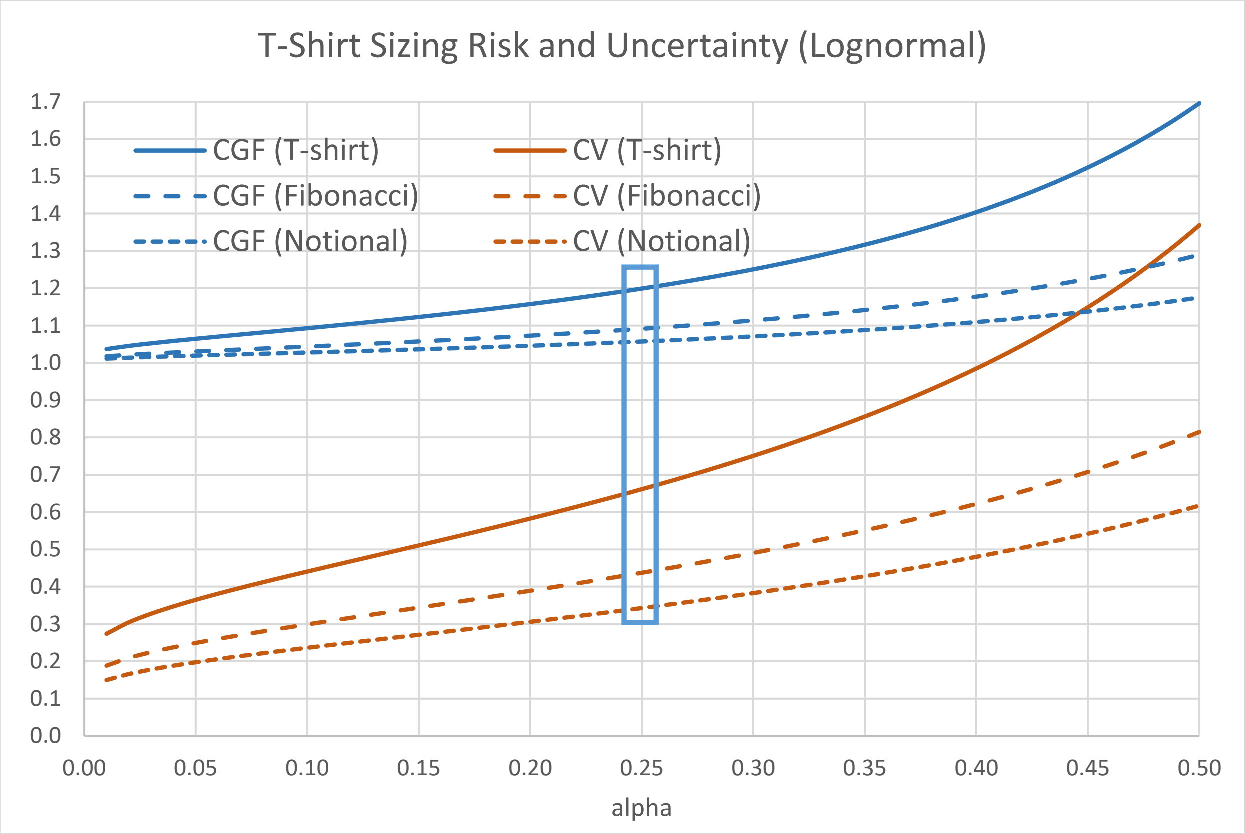


Figure 20: Lognormal Risk and Uncertainty with Variations in Both Confidence and Ratio

We can also generalize the original discrete distribution, in two steps. First the fair coin flips. The expected value is calculated below, with cost growth highlighted in red. Note the natural appearance of the $(R-1)$ term. By the nature of the self-similar scale, we must require R>1, and in practice R>>1, or the scale will be too “bunched up” and create the illusion of precision, offering the SME way too many choices.

$$\sum_i x_i p_i = \left({}^{1}\!/_{4}\right)\left({}^{H}\!/_{R}\right) + \left({}^{1}\!/_{2}\right)(H) + \left({}^{1}\!/_{4}\right)(RH) = \frac{1}{R}\left(\frac{R+1}{2}\right)^2 H = \left[1 + \boldsymbol{\frac{1}{R}\left(\frac{R-1}{2}\right)^2}\right]H$$

The variance calculation is below, with the pseudo-CV highlighted in purple.

$$\sum_i {x_i}^2 p_i - \left[\sum_i x_i p_i\right]^2 = \frac{1}{4}\left(\frac{H}{R}\right)^2 + \frac{1}{2}H^2 + \frac{1}{4}(HR)^2 - \frac{1}{R^2}\left(\frac{R+1}{2}\right)^4 H^2$$

$$= \left[\frac{3R^4 - 4R^3 + 2R^2 - 4R + 3}{(4R)^2}\right]H^2 = \left[\boldsymbol{\frac{R-1}{4R}\sqrt{3R^2 + 2R + 3}}\right]^2 H^2$$

Dividing the pseudo-CV by the CGF yields the CV.

$$CV = \frac{R-1}{(R+1)^2}\sqrt{3R^2+2R+3}$$

As a sanity check, substituting $R = 2$ produces the same result as before ($\sqrt{19}/_9$).

Further generalizing, we revert to the unfair second coin flip. Here is the expected value with cost growth highlighted in red.

$$\sum_i x_i p_i = \left({}^{\alpha}/_2\right)\left({}^{H}/_R\right) + (1-\alpha)H + \left({}^{\alpha}/_2\right)(RH) = \frac{\alpha - 2(1-\alpha)R + \alpha R^2}{2R}H$$

$$= \left[1 + \boldsymbol{\alpha}\frac{\boldsymbol{(R-1)^2}}{\boldsymbol{2R}}\right]H$$

The variance calculation, with pseudo-CV highlighted in purple:

$$\sum_i x_i{}^2 p_i - \left[\sum_i x_i p_i\right]^2 = \frac{\alpha}{2}\left(\frac{H}{R}\right)^2 + (1-\alpha)H^2 + \frac{\alpha}{2}(HR)^2 - \left(\frac{\alpha - 2(1-\alpha)R + \alpha R^2}{2R}\right)^2 H^2 =$$

$$\left[\left(\frac{\boldsymbol{R-1}}{\boldsymbol{2R}}\right)\sqrt{\boldsymbol{\alpha[(2-\alpha)R^2 + 2\alpha R + (2-\alpha)]}}\right]^2 H^2$$

The CV is given below.

$$CV = \frac{R-1}{2R + \alpha(R-1)^2}\sqrt{\alpha[(2-\alpha)R^2 + 2\alpha R + (2-\alpha)]}$$

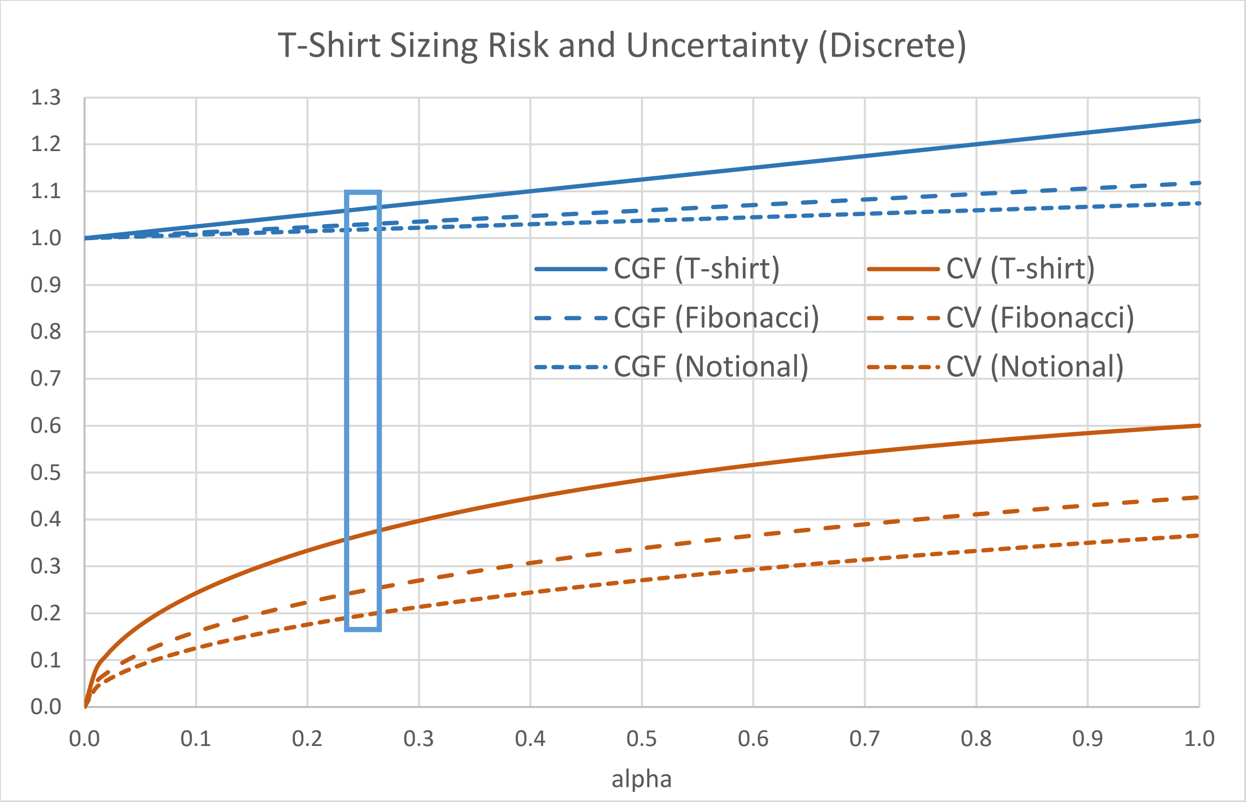


Figure 21: Discrete Risk and Uncertainty with Variations in Both Confidence and Ratio

Substituting one half for alpha (fair coin flip) reduces to the generalized ratio results just derived. Substituting two for R reduces to the earlier confidence-based result for T-shirt sizing.

Generalizations of the Uniform and Triangular cases are deferred to the appendix.

Problem Context: Reliance on the Reluctant Expert

"Honesty is such a lonely word/ Everyone is so untrue
Honesty is hardly ever heard/ But mostly what I need from you"
- Billy Joel, "Honesty," 52nd Street

How Accurate Is the Expert?

This brings us back to the central question of how accurate the SME really is. Daniel Kahneman and Amos Tversky won a Nobel Prize in Economics for cataloguing the foible of human assessments and decisions. Andy Prince and Christian Smart have cleaned up in the ICEAA Best Paper department applying these findings to the business of cost estimating.

The default answer to the question "How accurate is the expert?" is "Not very!" Not only do expert assessments tend to be biased (risk), they tend to understate the true underlying variation of the phenomenon being estimated (uncertainty). The typical expert is overly optimistic *and* overconfident. We purport to translate SME inputs to be more realistic through appropriate use of a self-similar scale. Once this is done, SME training may hold promise for further improvement of such estimates.

Self-Similar Scales and the Ideal Ratio

Self-similar scales are fractal in that mis-estimation will result in growth (or reduction) by the same ratio regardless of position on the scale. This enables the convenient application of a single risk factor to the entire software development effort, or at least large chunks thereof. It also holds promise for linking the macro- and micro-level assessments in a consistent manner.

In these self-similar scales, there is an evitable trade-off between accuracy and precision (or granularity). We can make a *reductio ad absurdum* argument and imagine the number of notches or the ratio growing without bound (approaching infinity). In the former case, there is excessive granularity, with the notches losing all practical meeting. The scale collapses to a continuous assessment by the SME with no guardrails, and the derivation of risk and uncertainty from adjacent notches is no longer possible. In the latter case, the gaps between notches become too large, and the SME is frustrated by the inability to find a choice that closely represents the estimated size. While CGF and CV can still be calculated, they may be prohibitively large, as it is no longer reasonable to think that the SME will be off by even one notch that large an appreciable proportion of the time.

Intuition tells us there's a happy medium, and based on our three examples, practical experience seems to indicate that that happy medium occurs for a scale ratio somewhere in the range of about one and a half (1.5) to two (2.0). Another candidate on the upper end is the ubiquitous number $e = 2.71828$ ..., which leaps to mind because it is the base of the exponential function that is its own derivative! The relevance of this fact to our estimation problem remains unclear, though it arguably makes it a better candidate than, say, its even more famous transcendental counterpart $\pi = 3.14159$ ....

Any ratio of about three (3.0) or greater makes the "jump" between notches on the scale too great, increasing the penalty for being wrong. Conversely, any ratio of about one and a half (1.5) or less makes the gradation of the scale too fine, increasing the probability that the assessment will be off by *several* notches and not just one.

Ultimately, the scale should be chosen to reflect the actual accuracy of the SMEs as captured by a given alpha value. The literature seems to establish alpha in the neighborhood of a fourth to a third. That is, SME ranges tend to capture somewhere between 66% and 75% of the true range of outcomes. (Not coincidentally, plus or minus one standard deviation of the normal distribution encompasses about 68% of the probability.) What remains to be seen, and we propose to test empirically, is what scale ratio corresponds to this alpha when it comes to software sizing.

It may be premature to try to use the generalized derivations above to "convert" between ratio (R) and confidence (one minus alpha). The math is messy, and we do not yet have enough empirical evidence to know which distribution(s) may be most appropriate.

Teaching Pigs to Sing: Lessons Learned for Two- and Three-Point Estimates

As move toward empirical testing of scales, we leverage an approach used in a previous paper on the use of SMEs in Cost and Risk. (Braxton & Coleman, Teaching Pigs to Sing: Improving Fidelity of Assessments from Subject Matter Experts (SMEs), 2012)

We propose to follow the same approach as in the cited paper, where polling can be conducted live, either via pencil and paper or a survey platform such as Qualtrics (https://www.qualtrics.com/). Because we cannot afford to wait three months, 12 months, or five years for the test results to come to fruition, we generally have to ask about either knowable but unknown quantities such as the height of Mount Pinatubo or the box office gross for *Avengers: Endgame*, or unknowable (uncertain) near-future events such as the high temperature in Washington, D.C., next Friday or the box office gross for *Thor: Love and Thunder*. (Note that the global pandemic has wreaked havoc with using box office gross as a potential test case! For this reason, we would eschew any films released in 2020 or later.) The previous paper tentatively established that SMEs have comparable accuracy for unknown but knowable and unknowable (uncertain) quantities.

First and foremost, we wish to assess the prevailing accuracy and precision of our pool of SMEs and reflect them in our estimates. Second, we wish to consider whether that accuracy and precision can be improved by how we devise our scale and how we train SMEs to use it.

Empirical Experiment: Analogized Scales

SMEs in Search of a Basis

CEBoK goes out of its way to discount Expert Opinion as an acceptable Cost Estimating Technique. This is less about cautioning against over-reliance on SMEs, and more about making a philosophical and conceptual distinction between expert-driven and data-driven methods. In Expert Opinion, the one and only expert-driven method, the estimate is presented as a direct assessment by the SME with no *apparent* basis, like

Athena springing fully formed from the mind of Zeus. Our experience working with Basis Of Estimate (BOE) authors and other technical contributors to cost estimates, individuals often labeled “experts,” shows that there usually is a concrete basis underlying their opinions. Consciously or subconsciously[3], SMEs leverage their experience with relevant (and perhaps not-so-relevant) programs in interpolating or extrapolating to the current assessment. As professional estimators and elicitors, our job is to push them toward a data-driven method.

This is where Expert Judgment comes in. The four accepted Cost Estimating Techniques (Analogy, Parametric, Engineering Build-Up, and Extrapolation from Actuals) are all data-driven, but they also rely on the SME to help interpret and contextualize the data so that they are appropriately normalized and analyzed. The data themselves are still the basis of the estimate, but Expert Judgment has played a key role. At worst, estimators rely on SMEs to make a direct assessment as to the scope on which the estimate is based (e.g., software sizing!). How then can we make this Expert Judgment more effective by borrowing best practices from the realm of data-driven assessments?

From Single-Point Analogy to Analogized Scales

In cost estimating, it is never our intent to cut out the Expert (SME) but rather to force Experts to couch their assessments in concrete terms that have an explicit basis and rationale, which can be *independently verified* before the fact and *empirically measured* after the fact. To this end, we propose to “Analogize” the self-similar scale and augment or replace numerical values with historical examples that (approximately) correspond to those sizes. This approach is not entirely novel. The Mohs scale of mineral hardness is not just 1 to 10, but talc to diamond. The Beaufort scale is not just categories or wind velocities, but “light breeze” to “hurricane.” One of the challenges is to find a sufficient range of analogy examples to populate the scale, but the prevalence of such Digital Engineering (DE) tools as Atlassian’s Confluence and Jira should help in this endeavor.

[3] What Daniel Kahneman refers to as System 1 and System 2, irrespectively.

As previously discussed, the analogized scale is still just asking for a one-point estimate from the SME – essentially a drop-down menu choice most closely matching their single best guess – but it provides a “stealth” three-point estimate in the form of the adjacent notches on the self-similar scale. This approach also transcends Expert Opinion with a sort of a “stealth” Analogy. It is hypothesized that sizing and similar assessments can be improved by labeling each notch on the scale with an actual example reflecting that approximate size.

Based on previous research, there is some evidence that expertise in uncertainty assessments is equally or more important than expertise in subject area in itself. We will assume at least a modicum of the latter and propose to develop the former through training. The precise training approach is outside the scope of this paper.

Experimental Formulation

The first empirical research question is whether the accuracy of judgments on the same scale are affected by how the scale is labeled. Three possible labeling schemes are numbers only, analogies only, or both. Our hypothesis is that providing analogies will improve accuracy over the numbers only scale, and that analogies only might be the best of all. Since we cannot ask the same individuals to repeat their assessments with a different scale (pending mind erasing technology!), we propose to randomize and present each respondent with one of the three possible scales. We would then present the same set of ten (10) assessments, again in a random order.

The second empirical research question is whether (absolute) accuracy or precision are dependent upon the scale ratio. Our hypothesis is that SME accuracy and precisions are largely invariant to the chosen scale, as long as it is within a “reasonable” range. That being said, risk ranges would still need to be calibrated to the actual scale ratio used. To avoid combinatorial explosion, we propose to test ratios of 1.5 and 2.0 only, again randomizing between the subjects. The former is between the Notional and Fibonacci models, and the latter is the T-shirt model. We are now up six different randomized treatments: {numbers, analogies, both} x {1.5, 2.0}.

The third and final empirical research question is whether SMEs do better at judging large things, small things, or about the same. Our hypothesis is that their accuracy and precision are roughly invariant to size. This would fit into our fractal concept of risk and enable us to link otherwise disparate micro- and micro-level sizing models.

Experiment #1: Heights of Mountains

The first experiment falls into the "unknown but knowable" category. While there is a certain appeal to using an objective logarithmic scale for physical phenomena, such as sound intensity (decibels) or earthquakes (Richter scales), these offer more limited analogy possibilities. Instead, we chose heights of mountains, which most analysts have a general conception of (e.g., "Denver is the mile-high city") but not an expertise in. This scenario also offers sufficient variety of analogy possibilities. Doubling between 1,000 feet and 32,000 feet results in a six-notch scale, whereas a similar range – pardon the pun! – can be covered by an eight-notch scale with ratio 1.5. Examples of these scales are shown in Table 4 and Table 5.

Table 4: Empirical Testing Scale (R = 2.0)

scale (ft)	mountain	location	elevation (ft)
500	Driskill Mountain	Louisiana	535
1,000	Woodall Mountain	Mississippi	807
2,000	Mount Arvon	Michigan	1,979
4,000	Black Mountain	Kentucky	4,145
8,000	Guadelupe Peak	Texas	8,751
16,000	Mont Blanc	France	15,774
32,000	Mount Everest	Nepal	29,031

Table 5: Empirical Testing Scale (R = 1.5)

scale (ft)	mountain	location	elevation (ft)
1,000	Woodall Mountain	Mississippi	807
1,500	Crown Mountain	St. Thomas, USVI	1,555
2,250	Eagle Mountain	Minnesota	2,302
3,375	Mount Davis	Pennsylvania	3,213
5,063	Black Mesa	Oklahoma	4,975
7,594	Black Elk Peak	South Dakota	7,244
11,391	Mount Hood	Oregon	11,249
17,086	Pico Pan de Azucar	Colombia	17,060
25,629	Nanda Devi	India	25,643

Devising these scales produced several lessons learned. First, we tried to use more familiar mountains as the reference analogies, but familiarity had to take a back seat to a relatively consistent scale. Second, we say “relatively consistent” because it is nigh impossible to hit each notch on the scale exactly with a useful example. In the tables above, the desired elevation in feet for the chosen scale ratio is in the leftmost column, and the actual elevation in feet of the analogy mountain is in the rightmost column. Third, this is a helpful reminder that every analogy data point represents either a “lucky” or “unlucky” program, which is to say it came in with a lower or higher cost than it “should have” *on average*. (Some mountains used to be taller but had their tops blown off in volcanic explosions!) Finally, while the world population of mountains is not necessarily right skew, the readily documented ones certainly are. Lists of peaks, as the name implies, are invariably the tallest ones in a range – those are the ones people are excited to climb. There may be a similar challenge when we try to analogize a software development scale. The memorable projects are big, expensive ones. We need to do the work to identify smaller more mundane efforts, as that end of the scale is equally important.

Experiment #2: Box Office Gross of Films

While heights of mountains may or may not be skew right, it is well established that the popularity of books, songs, and movies have long, fat right tails. Thus, for the second part of the survey, we use popular films from 1990-2019 and ask respondents to assess their domestic box office as reported by Box Office Mojo. By focusing on recent films, we minimize the impacts of inflation. (For cost estimating purposes, we would obviously want to normalize for inflation. For risk assessment purposes, we instead stick with nominal dollars because that’s the form in which amounts are almost always reported in the press, meaning that’s what respondents are mostly likely to inherently calibrate to.) As previously noted, we steer clear of the last two years so as to avoid any confounding pandemic effects. (Box office returns were severely depressed or even non-existent for most of the pandemic, though they have recently rebounded, with *Spider-Man: No Way Home* recently surpassing *Avatar* to edge into third place all time – again, in non-inflation-adjusted dollars.) While Experiment #1 may be more typical of micro-level sizing, Experiment #2 may be more representative of macro-level sizing. To extend from one million dollars ($1M) to one billion dollars ($1B) – or $1,024M to be more precise – generates an 11-point scale with ratio 2.0, or about a 17-point scale with ratio 1.5.

Experiment #3: Driving Distances

While indie films and Marvel blockbusters may pretty much be different species, we further test the fractal risk hypothesis by asking respondents to assess the driving distance (as specified by Google Maps) to both local and interstate destinations from the Technomics headquarters in Arlington, VA, The Artist Formerly Known As Crystal City (now “National Landing,” thanks to the arrival of Amazon’s HQ2).

Results and Recommendations

Survey Results

The empirical survey has not been completed as of this writing. An initial survey is planned prior to the conference itself, and the author may leverage the conference app, if permitted, to solicit additional responses before, during, and after the presentation!

SME Training and Data Accumulation

The analogizing of the sizing scale forces research into past actuals, which is a significant side benefit, especially for those organizations that have neglected the “blocking and tackling” of cost and software data collection for far too long. Where recourse to true actuals fails, we may have to resort to “anecdotal actuals” aka “expert testimony,” but by using it for scaling, at least we get it on the record! This also serves as a cautionary tale, inspiring us to do better going forward. Keep in mind that we want both Cost and Risk data (snapshots in time, at a minimum Initial and Final).

Even at a more detailed (scrum team) level, an accumulation of examples can help bolster the memory and the judgment of the SME. There we have an embarrassment of riches, but it may still take significantly resources to sift through the data representing various time intervals and scopes (sprint, PI, release; task, story, feature, PE, SE).

Conclusion and Next Steps

Risk and Uncertainty Benchmarks

The bottom line is that significant risk and uncertainty are inherent in these self-similar sizing scales *even if we are off by no more than one size in either direction*. (Or in some cases, a little more than one size.) While this may be little surprise to a cost and risk analyst, we have found this thought experiment to be an effective communication tool for conveying to program teams the inherent risk and uncertainty of their expert-driven sizing.

We have indulged in a great deal of algebra, for purposes of both research and enjoyment, but let's not lose sight of the forest for the trees. Table 6 reflects the fruits of our algebraic labor. While the precise calculations have been shown to depend on underlying distributional assumption, good rules of thumb are that we should expect **growth** on the order of **10% to 30%** and **CVs** on the order of **30% to 50%** for macro-level sizing. We recommend applying these ranges to expert-based methods such as T-shirt sizing. They can also be used as benchmarks for other early-phase software estimates.

Table 6: Risk and Uncertainty Benchmarks for T-shirt Sizing

	Confidence	Growth %	CV
Discrete	$\alpha = 0.50$	12.5%	48.43%
Uniform	$\alpha = 0.00$	25.0%	34.64%
Triangular	$\alpha = 0.00$	16.7%	26.73%
Discrete	$\alpha = 0.25$	6.2%	36.74%
Lognormal	$\alpha = 0.25$	19.9%	66.16%
Uniform (Proportional)	$\alpha = 0.25$	33.3%	43.30%
Uniform (Equal)	$\alpha = 0.25$	25.0%	46.19%
Triangular (Proportional)	$\alpha = 0.25$	33.3%	39.53%

These benchmarks all reflect a ratio of $R = 2.0$.

Continuing Research

The next step in this line of inquiry is to conduct the survey experiments described above and compile the results. We are aiming to have that done in time to present at the conference itself, but it may also be expanded in a follow-on paper. We seek to better understand the interplay between R, the ratio of an analogized scale, and α, the true confidence of SMEs in applying that scale. Those empirical results should illuminate this question considerably, and the above derivations demonstrate how that interplay is dependent on distributional assumptions.

Agile Software Risk Decomposition

The rubric in Figure 22 below has been invaluable in discussing the need for robust data collection to improve upon expert-driven methods. It can also serve as a basis for discussing alternative approaches to risk and uncertainty. As previously pointed out, macro-level methods such as T-shirt sizing tend to span all the way from Requirements to Effort or Cost along the top tier of the diagram. So, this approach may be appropriate early in the life cycle when little detail is available.

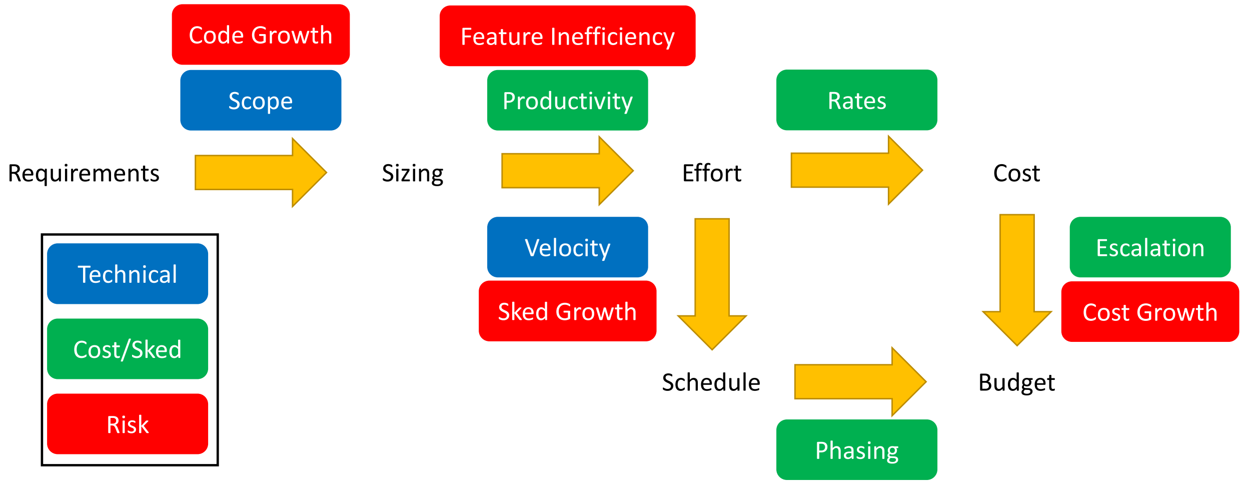


Figure 22: Software Estimation Decomposition, Data Needs, and Risk

The more traditional Inputs Risk approach, feasible with both more historical data and more detail on the present program being estimated, would compound the uncertainty associated with Sizing, Productivity, Rates, and possibly Escalation as separate factors, as implied by the arrows in the diagram. As we work to improve expert-driven methods, we are continually relating them to data-driven methods and exploring how the two can work most effectively in concert.

Bibliography

Braxton, P. J. (2021). Inherent Risk and Uncertainty of Self-Similar Sizing Scales in Agile Software Development, or "Does This T-Shirt Make My Estimate Look Big?". *IT-CAST Proceedings.* Washington, DC: DHS.

Braxton, P. J. (2022). Uncertainty of Expert Judgment in Agile Software Sizing. *ICEAA Conference Proceedings.* Pittsburgh, PA: ICEAA.

Braxton, P. J., & Coleman, R. L. (2012). Teaching Pigs to Sing: Improving Fidelity of Assessments from Subject Matter Experts (SMEs). *ICEAA Chapter Luncheon Proceedings.* Washington, DC: ICEAA Washington Chapter.

Braxton, P. J., & Sayer, L. H. (2013). Probability Distributions for Risk Analysis. *ICEAA Conference Proceedings.* New Orleans, LA: ICEAA.

Cost Estimating Body of Knowledge (CEBoK). (2013). Annandale, VA: ICEAA.

Gellatly, W., Braxton, P. J., Brown, D. H., Jones, L. F., & Wekluk, R. A. (2022). Dynamic Software Effort Estimation: How SWEET It Is! *ICEAA Conference Proceedings.* Pittsburgh, PA: ICEAA.

Kahneman, D. (2011). *Thinking, Fast and Slow.* New York, NY: Farrar, Straus and Giroux.

Kosmakos, C., & Brown, D. H. (2022). Are We Agile Enough to Estimate Agile Software Development Costs? *ICEAA Conference Proceedings.* Pittsburgh, PA: ICEAA.

Livio, M. (2008). *The Golden Ratio: The Story of PHI, the World's Most Astonishing Number.* Crown.

Radigan, D. (2022, February 24). *Story points and estimation*. Retrieved from Atlassian Agile Coach: https://www.atlassian.com/agile/project-management/estimation

Acronyms, Initialisms, and Abbreviations

Abbreviation	Expansion	Notes
BOE	Basis Of Estimate	
C	Complex	Notional Sizing Model
DE	Digital Engineering	
E	Easy	Notional Sizing Model
FTE	Full-Time Equivalent	Notional Sizing Model
L	Large	T-Shirt and Notional
LCS	Learning Curve Slope	
M	Medium	T-Shirt and Notional
M	Moderate	Notional Sizing Model
PE	Program Epic	Agile
PI	Program Increment	Agile
PM	Person-Months	Notional Sizing Model
S	Small	T-Shirt and Notional
SE	Solution Epic	Agile
SME	Subject Matter Expert	
XL	Extra Large	T-Shirt Sizing Model
XS	Extra Small	T-Shirt Sizing Model
XXL	Double Extra Large	T-Shirt Sizing Model

Appendix A: Fibonacci Sequence Convergence

Derivation of close-form formula for Fn and demonstration that ratio converges to Golden Mean

Suppose the closed-form formula is of the form:

$$F_n = c \cdot a^n + d \cdot b^n$$

Must satisfy basic Fibonacci relationship:

$$\begin{aligned} &F_n + F_{n+1} = c \cdot a^n + d \cdot b^n + c \cdot a^{n+1} + d \cdot b^{n+1} \\ &= c(a^n + a^{n+1}) + d(b^n + b^{n+1}) = c \cdot a^{n+2} + d \cdot b^{n+2} = F_{n+2} \end{aligned}$$

This will be true if both a and b are roots of the quadratic:

$$x^2 = x + 1 \;\rightarrow\; x^2 - x - 1 = 0 \;\rightarrow a = \frac{1+\sqrt{5}}{2} = \phi \,, b = \frac{1-\sqrt{5}}{2} = 1 - \phi$$

Now we solve for the coefficients c and d:

$$\begin{aligned} F_1 &= 1 = \phi c + (1-\phi)d \\ F_2 &= 1 = \phi^2 c + (1-\phi)^2 d \end{aligned}$$

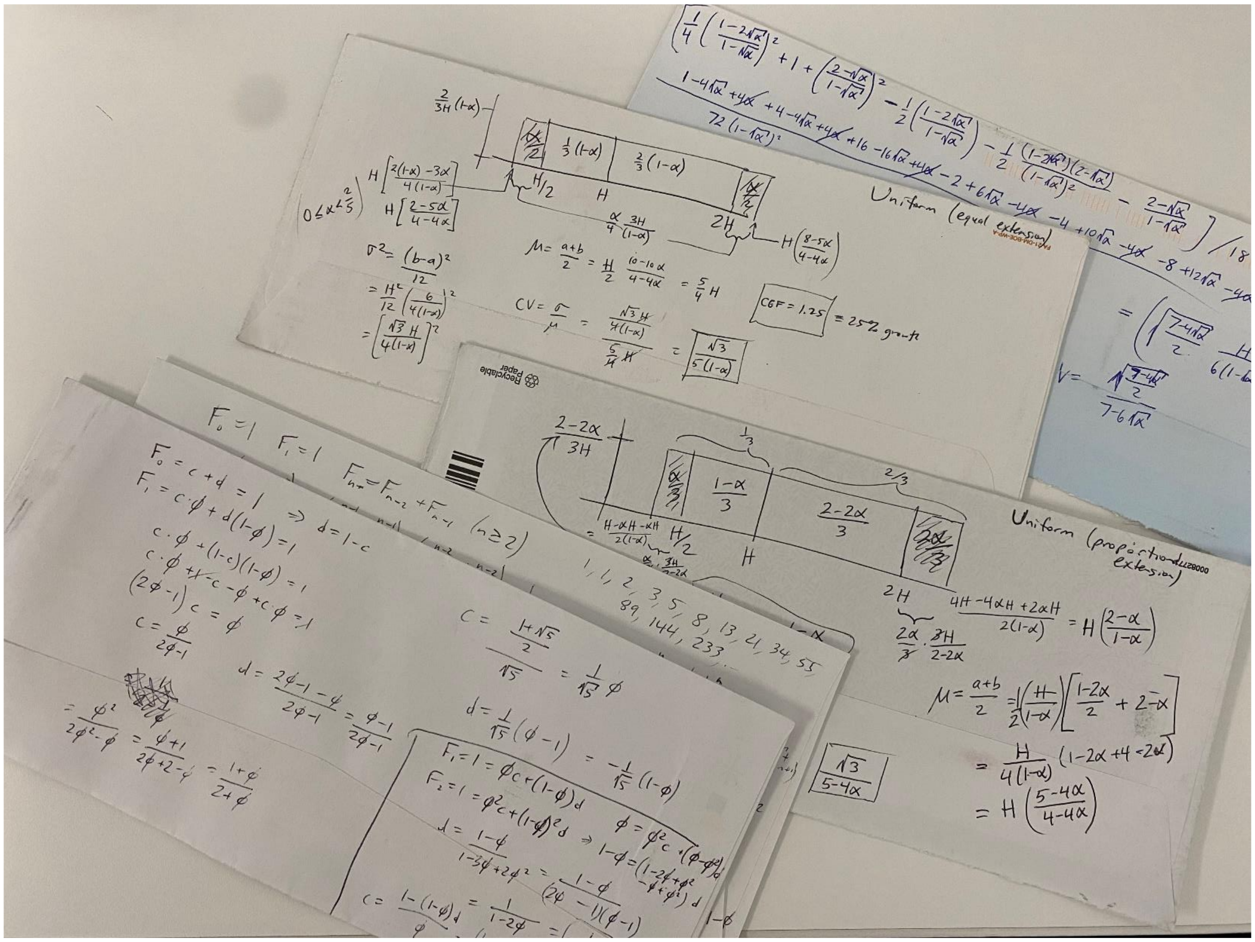
$$c = \frac{1}{2\phi - 1} = \frac{1}{\sqrt{5}} \,, d = \frac{1}{1 - 2\phi} = -\frac{1}{\sqrt{5}} \;\rightarrow\; F_n = \frac{1}{\sqrt{5}}[\phi^n - (1-\phi)^n]$$

Since the second term vanishes as n increases without bound, the ratio of consecutive terms approaches *a*. Also alternating with odd and even n, fluctuations seen earlier in graph.

Coda: The Proverbial Cocktail Napkin

"It's a pretty good crowd for Saturday
And the manager gives me a smile
'Cause he knows that it's me they've been comin' to see
To forget about life for a while"
- Billy Joel, "Piano Man," Piano Man

Once again proving that I do all my best work on the back of a white envelope. I wish I'd saved some more of these over the years, how about you?



Uniform (equal extension)
Uniform (proportional extension)
CGF = 1.25 = 25% growth
$\mu = \frac{a+b}{2}$
$\sigma^2 = \frac{(b-a)^2}{12}$
$CV = \frac{\sigma}{\mu}$
$F_0 = 1 \quad F_1 = 1 \quad F_n = F_{n-2} + F_{n-1} \quad (n \geq 2)$
1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233
$= H\left(\frac{5-4\alpha}{4-4\alpha}\right)$
$\frac{\sqrt{3}}{5-4\alpha}$

Figure 23: Artisanal Probability Calculations