What is Truth?
Do our SIL calculations reflect reality?

Keith Brumbaugh
aeSolutions, Houston, TX
keith.brumbaugh@aesolns.com

Abstract

Is our industry stuck in the past? The current industry trend is to only look at random hardware failures in safety integrity level (SIL) probability of failure on demand (PFD) calculations. No one would appear to be updating assumptions as operating experience is gained. Hardware failure rates are generally fixed in time, assumed to be average point values (rather than distributions), and either generic in nature or specific to a certain set of hardware and/or conditions which the vendors determine by suitable tests or failure mode analysis. But are random hardware failures the only thing that cause a safety instrumented function (SIF) to fail? What if our assumptions are wrong? What if our installations do not match vendor assumptions? What else might we be missing? How are we addressing *systematic* failures?

One obvious problem with incorporating systematic failures is their non-random nature. Many functional safety practitioners claim that systematic errors are addressed (i.e., minimized or eliminated) by following all the procedures in the ISA/IEC 61511 standard. Yet even if the standard were strictly adhered to, could anyone realistically claim a 0% chance of a SIF failing due to a human factor? Some will say that systematic errors cannot be predicted, much less modeled. But is that true?

This paper will examine factors which tend to be ignored when performing hardware-based reliability calculations. Traditional PFD calculations are merely a starting point. This paper will examine how to incorporate systematic errors into a SIF's real-world model. It will cover how to use Bayes theorem to capture data after a SIF has been installed — either through operating experience or industry incidents — and update the function's predicted performance. This methodology can also be used to justify prior use of existing and non-certified equipment.

Keywords

Bayes, Bayes' theorem, SIS (Safety Instrumented System), SIF (Safety Instrumented Function), SIL (Safety Integrity Level) calculation, PFD (Probability of Failure on Demand), RRF (Risk Reduction Factor (1/PFD)), LOPA (Layer Of Protection Analysis), systematic failures, human factors, human reliability, operations, maintenance, ISA/IEC 61511, hardware reliability, prior use, confidence interval, credible range.

Introduction

How confident might anyone be that a SIF's calculated probability of failure on demand (PFD) represents reality? Current industry practice is to only consider random hardware failures when calculating PFD. Yet does the calculation really indicate how often a function may actually fail? Specialists may be confident in the numbers used in the calculation, but might this be a false sense of security? Layer of protection analysis (LOPA) often estimates hazardous events in the range of 10^{-4} or less per year. Yet industry is still experiencing several disastrous events per year.

If one *estimates* 1,000 major operating units worldwide, and industry experiences 10 major incidents per year, the accident rate would be 10 / 1,000 per year, or 1 in 100 (0.01) per year. This is the same probability as the borderline between SIL 1 & 2. It is likely that most of the accidents so far have had nothing to do with instrumentation failure. This should give everyone pause. What might be missing in the calculations? Perhaps the key lies in the previous statement; *it is likely that many of the accidents so far have had nothing to do with instrumentation failure*. Now we might be onto something!

Current PFD models are not factoring in systematic failures. This is not entirely the fault of safety system practitioners. There is not a well-defined method or standard that discusses how to approach this topic. This author wrote a paper [1] that showed even a SIL 3 SIF spouting all the latest and greatest certified devices could be brought to its knees when poor human factors and systematic error were factored into the analysis. With just a few realistic and slightly conservative systematic issues factored in, a mighty SIL 3 SIF was degraded down lower than SIL 1! This begs the question with our current hardware-based calculations; *if they are not reflecting reality, then what good are they*?

This paper will provide an approach using Bayes theorem to include real-world factors when modelling the performance of a SIF. Such a model would include systematic errors that are unique to a facility. A statement of credibility could then be made using the Bayes approach. For example, one could say that they are certain that the real PFD value of a SIF is contained within a range with a 90% assurance (i.e., a 90% credible interval). The more data one collects, the closer one gets to the true PFD value. Bayes' theorem can even incorporate subjectivity. One can use key performance indicators, audits, and assessments to get an idea of the performance a SIF is capable of and determine if a gap in safety is going unaddressed.

Frequentist vs Bayes

It is important to define two different philosophies — Frequentist vs. Bayesian — regarding the implementation of ISA/IEC 61511 (hereafter referred to as "the standard") and PFD calculations.

Frequentist based statistics assume that the relative frequency of an event (i.e., how many times an event occurs over a previous sample) is the same as the probability of the event occurring during the next sample. Unfortunately, this assumption only applies to situations where many identical trials can be repeated ad nauseum with well-defined outcomes. Examples would be games of chance where the odds are known *in advance*, or where a "true" fixed parameter value can be assumed to exist in the population, such as the average weight of a 6-month-old male infant.

The frequentist approach is *not* good for inferring rare event probabilities. Frequency is in the past; probability is in the future. We do not — and never want to — know what the true average disaster rate is. We try to avoid ever having one! Even if one could quantify an average point value, the average would never stay "fixed" due to organizational and human changes and other systematic contributions.

Bayesian based statistics assume the parameter of interest (e.g. the PFD of a SIF) is a random variable. One assigns a prior distribution (i.e., an uncertainty distribution) that represents the belief in a certain parameter value over its range. The prior distribution is then updated with new evidence which can quantify the uncertainties of our assumptions and bring us closer to the true value. There is no fixed average in Bayesian based statistics. The average moves (i.e., the

PFD in this case) with evidence backing it up. This enables more realistic quantification of PFD before a disaster occurs.

The problem with our current approach

The current industry approach to determine SIF performance is to perform a PFD calculation. One method would be using equations published in ISA-TR84.00.02-2002 – Part 2, as well as other standards and books. The intention is that the system's failure probability is modelled, and if the function falls within the desired target SIL range, the design is acceptable. The pitfall in this approach is the numbers used in the calculations are all frequentist derived (i.e., they assume a true, fixed average value). To make a frequentist inference one needs a true, fixed average. This requires an abundance of data. Yet does industry really have such an abundance of data? Industry may have a good measure of hardware-based failures to form an accurate average, but are hardware-based failures the only thing that can make a SIF fail? What about the systematic component? The standard makes a differentiation between hardware safety integrity and systematic safety integrity, so it is no wonder that confusion exists. Many argue that PFD calculations should only be looking at hardware-based failures, and that systematic contributions have no place in PFD calculations. But is that true? Are current calculations even following this black and white approach?

PFD calculations include many different parameters such as failure rate, diagnostic coverage, proof test interval, and common cause. Practitioners take a frequentist approach and assume the values are well established, fixed constants. Most also assume that systematic error is not — and should not be — factored into these calculations. Yet are these statements true?

Common cause values are typically estimated based on a method shown in the IEC 61508 standard. The parameters for estimating the beta factor are systematic in nature. Common cause failures degrade performance by orders of magnitude. Glossing over these — as well as other systematic failures — could be viewed as unrealistic, misleading and potentially dangerous.

Proof test intervals are generally fixed and are often assumed to be the same as the plant turn around interval. Delayed testing (a systematic contributor) can have a significant negative impact on SIF performance. The impact of delayed testing can be modeled quantitatively, yet the math using traditional means can be complicated. Assumptions would be made and hopefully a management of change (MOC) study would be completed for the plant to continue to run. This would include management accepting the potential risk of delayed SIF testing.

Practitioners assume device failure rates are hardware based only and are derived from databases where a true average inference can be made. Yet this may not be the case for all data sources, of which there are many. Practitioners commonly either use publicly available generic figures which are often based on industry averages, or use data provided by manufacturers (commonly referred to as model specific data).

Do generic failure rates only contain random hardware failures (and not systematic)? Consider the following statement from Dr. Bill Goble of exida [2], "Today, a lot of very useful data is provided by the OREDA database … They published a PDF data handbook which has failure rates, failure modes, and common cause factor estimates for use in safety instrumented function verifications. … What I learned, very importantly, all realistic failures are included. That includes both failures due to the product design, manufacturing, and *failures [due] to site operational procedures… including things like maintenance errors and exceptional stress*." (Emphasis added.) Our assumption that generic data is devoid of systematic error may be an error itself.

Another question to consider is whether the data is even valid for a frequentist inference. Is there enough data for a fixed average? Generic data is based on wide sampling of industry averages. While the data may represent a true average (assuming there are enough samples to prove the very small numbers), a specific facility still may not fall within the average. Keep this thought in mind as this paper will cover confidence intervals. There may be orders of magnitude of uncertainty in assuming that industry average data applies to a specific plant. As Stephen Thomas wrote [3], "Safety Integrity Level (SIL) verifications may calculate an average Probability of Failure Demand (PFD_{avg}) to three decimal places based on data with uncertainties of two orders of magnitude." One should not assume generic data based on industry averages provides all the information needed to factor in a particular facility's systematic error contribution.

It might be tempting to believe that any systematic errors included in generic data have been applied conservatively, but that is not necessarily true. Dave Grattan wrote [4], "Hardware failure is considered random (i.e., statistical). Human error is systematic (i.e., deterministic). The two types of error differ only by the amount of information known about each… Claiming that a generic value of human error is 0.1 and that it is conservative is misleading and dangerous because of the nature of determinism. The error likelihood for an identified cause is either 1 or 0." In other words, any particular plant is not a generic plant. It will have its own deterministic systematic influence which is unique to itself. For example, if an unsafe practice becomes normalized and common practice over time, then the personnel will *always* make that error. In fact, they will not even question it or *consider* it an error. What if it were common practice to not test thermocouples because it was considered a line break and personnel do not want to suit up to run a test because it would waste time and lead to unnecessary paperwork? This would not be factored into a generic number no matter how conservative. Whatever number were assigned to proof test coverage for the thermocouples would be incorrect, unless one assumed the barest minimum.

Another source of hardware only (not systematic) failure rate data would be model specific data from a manufacturer derived from a failure modes, effects, and diagnostic analysis (FMEDA) process. Data derived by exida (a major certifier of devices) are based on FMEDA which are produced during the SIL certification process. This is likely as close as industry can get to a hardware only based failure rate where we can make a valid frequentist inference of a fixed average. Yet numbers on some certificates often appear so optimistic as to defy reality. At times one questions whether the data analyst just started making up numbers. As one example, the dangerous failure rate for one specific pneumatic spring return actuator was given from a high-profile certification company as 1.8 E^{-8}/hr, or 18 FITs (failures per billion hours). This translates to a mean time between failure (MTBF) of 6,500 years (i.e., only one device out of 6,500 fail dangerously in a year). Does this seem reasonable to anyone? If we cross reference this data with the silsafedata.com database [5] which provides industry recognized upper and lower bounds for dangerous failure rates, the lowest realistic value is 110 FITs. This is an entire order of magnitude difference! Even that number equates to a MTBF of ~1000 years, which still seems very generous.

Many functional safety practitioners question the validly of such data. One might choose to perform PFD calculations using such data and then eventually verify the number after collecting sufficient maintenance records. Unfortunately, it is generally not possible to collect enough data over a facility's lifetime to prove such low values, especially for a single device. In order to prove that an actuator has even the lower bound 110 FITs one would need a single device to

operate without failure for *1,000 years*. Decreasing the time needed would require operating more devices. To drop the sampling time to 5 years would require operating 200 devices (again, without a single failure occurring during that period). This could be justified if a facility were large enough. Some plants have taken this approach for gathering and maintaining their own site-specific data, however there is still the problem of averages. A SIF device of interest might fall outside the average, even within the same plant. What if a device were specified with an incompatible material, or was installed incorrectly, or suffered from any number of other possible systematic human errors? No average value would represent that device. And SIFs (or assemblies such as valves) are composed of more than just one device. Factoring in all the different devices, installations, and averages leads to increased uncertainty.

Why worry about systematic error?

One may be tempted to think that PFD calculations only need to factor in hardware error for contributions to overall performance. Some believe that systematic error is something to be corrected and not included in calculations. This stance is understandable given the definition of "hardware safety integrity" in the standard ("Part of the safety integrity of the SIS relating to random hardware failures in a dangerous mode of failure"). One issue industry faces is that PFD calculation tools do not incorporate a way to factor in *all* possible systematic contributions. So why even worry about it? If one ignores the systematic contributions to SIF performance, it would be neglecting what the standard requires. The standard says that safety integrity should factor in *both* hardware and systematic failures ("In determining safety integrity, all causes of random hardware and systematic failures which lead to an unsafe state can be included (e.g., hardware failures, software induced failures and failures due to electrical interferences)"). PFD calculations clearly cover the hardware failure aspect, but only partially factor in systematic failures (unintentionally or not). By not considering the "big picture", calculations could be missing any number of systematic failure contributions which could cause the predicted values to be far worse than what the calculation alone would indicate. Systematic contributions include, but are not limited to, the following.

- Specification
  - Incomputable materials
  - Incomplete or poor assumptions
  - Missing requirements
- Environmental factors
  - Extreme temperature
  - Humidity
  - Vibration
  - Plugging service
- Human impacts
  - Improper installation
  - Imperfect maintenance
- Poor assumptions
  - Imperfect testing
  - Incomplete testing

An entire book (or volumes) could be written on potential sources of systematic error. Quantifying the systematic failures — and then adequately representing their effects on system

performance — is a difficult prospect and not well defined. However, it is possible using Bayes' theorem. The Bayesian approach suggested in this paper is to include both quantitative and qualitative factors to give a more realistic impression of how a SIF is actually operating.

More issues with the existing approach

Additional limitations of the frequentist approach are the inability to include failures, shocks and warning signs. Suppose a SIF had a target PFD of < 0.10 and the initial calculated value was 0.083. The plant may only want to test their SIFs every 5 years to coincide with the plant turn around. Imagine performing the first proof test and finding the function in a failed state (e.g., a solenoid is stuck). What now? Most would replace the solenoid (if it were actually faulty), chalk it up to just being unlucky, continue and hope it does not happen again. The team might be very reluctant to update the PFD value of the SIF given the current tools. A frequentist approach might consider a new value of PFD as an average between 1 (1 test, 1 failure) and 0.083, so perhaps 0.54. The SIF now no longer meets the target value. It would require another 9 tests with no failures to show the SIF was back below the 0.1 target. If one were to rely on just a single SIF for samples (due to the problem with averages discussed above) and testing was only performed every five years, it would require waiting another 45 years without failures before the plant can claim there is no operational gap.

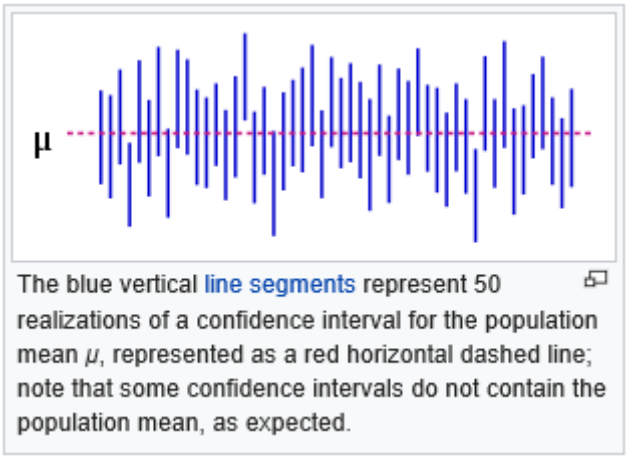
A solution to close an operational gap might be to redesign the SIF and strive for an even lower calculated PFD through the implementation of fault tolerant devices, but would this really solve the problem? What if upon further inspection it was discovered that the instrument air dryer was not functioning to specification and was allowing excess moisture (a systematic failure), which resulted in the solenoid failure. Instead of redesigning the entire SIF, one could (and should) fix the source of the problem. Perhaps equipment was failing upstream, or a different form of operations and maintenance were necessary for the faulty air system. If the instrument air system was corrected, one might now have greater confidence that the SIF would operate as originally assumed. Yet taking the frequentist approach there would be little choice but to redesign the SIF, or keep running and hope for the best. Fixing the faulty instrument air supply probably fixed the problem, but this is a qualitative, subjective judgement. How might this be quantified? Enter Bayes theorem, which allows for the application of qualitative factors into one's belief of a SIF's PFD. This approach will be discussed in more detail later in this paper.

Another issue is accounting for warning signs coming from system validation, audits, and assessments. A problem may be uncovered, but its impact on SIF performance may be unclear. There has been a push in industry to utilize key performance indicators (KPIs) as leading indications of SIF performance and overall plant safety. These are an attempt to glean systematic contributions and take pre-emptive action before there is an accident. A PFD calculation may indicate that a SIF falls within the SIL 3 range. Yet if the function is treated the same as any other control system interlock in the plant, the chances are it is not operating at a SIL 3 level. It is akin to defining a safe driver. As my colleague Paul Gruhn put it, "The definition of a safe driver is *not* one who hasn't had an accident. It's one who follows the rules, laws, doesn't text or drive under the influence, wears seat belts, has a car in good working order, etc. I'm more "confident" that such a person will not have an accident than a drunken teen who's texting and speeding but *hasn't had an accident yet*."

How do we apply these warning signs to a plant and SIF health? Audits and functional safety assessments are methods to glean this information and determine potentially negative indicators,

such as normalization of deviation, plant fragility, or near misses. The problem is that even though these indicators are all good warning signs that can lead to raising or lowering a confidence level, there is no clear method on how to factor this feedback into a PFD calculation. If it is not possible to quantify the potentially degraded performance of a system, then industry will not know what to focus on. There will be a risk of letting real issues lie dormant. This would be similar to not dealing with (i.e., just accepting) alarm floods, or the medical equivalent of "just treating the symptoms". Some SIF specific indicators *can* be quantified, albeit with some amount of difficulty. Examples using the traditional frequentist approach include excessive demands, failure rates, time in bypass, and delayed testing. Yet qualitative factors — which potentially have a more serious impact — cannot be factored in using current frequentist tools. As stated previously, a Bayesian approach can allow for qualitative updates of SIF performance.

One final note of concern with the frequentist approach is the means to quantify uncertainties and confidence intervals. The frequentist confidence interval is based on the amount of data available to make a statement of confidence that the true number lies within a particular interval X% of the time. It assumes a true value is fixed and the interval moves. If one claims a confidence interval of 95%, it is akin to saying that if one takes 100 samples, the true valve will be in 95 of them. This is not the same thing as saying there is a 95% probability that the parameter lies in our confidence interval.



The blue vertical line segments represent 50 realizations of a confidence interval for the population mean μ, represented as a red horizontal dashed line; note that some confidence intervals do not contain the population mean, as expected.

Figure 1: Example of confidence intervals

Consider the example above from Wikipedia, Figure 1 "shows 50 realizations of a confidence interval for a given population mean μ. If we randomly choose one realization, the probability is 95% we end up having chosen an interval that contains the parameter. However, we may be unlucky and have picked the wrong one. We will never know; we are stuck with our interval."

Note that practically none of the bars in Figure 1 have their actual sample average equal to the 95% confidence interval average. A value could be high or low and that would not be known until enough samples were gathered, which is often not practicable. For example, what if a plant received a batch of solenoids that operated with their sample average below the 95% confidence interval average (i.e., sample average is less than "μ" from the figure)? The sample still might contain the 95% confidence average, but if that number were used in a PFD calculation the answer would be optimistic. One might have calculated a SIF to just barely meet the

performance target, but the function would not really be operating at the target. It could be worse. It is entirely possible the sample does not even contain the 95% confidence average. The facility was just unlucky.

One might assume that there is ample data out there, so there should be enough samples to make a statement of confidence. Yet as pointed out, most of the data is averaged from the whole industry and is not representative of any particular plant. Average data will not contain a plant's systematic biases. How could it? What sort of confidence is possible if none of the data represents a specific plant? As the saying goes; garbage in, garbage out. Collecting enough data to make a statement of confidence for a plant could require more time than a plant would ever have. Knowing the confidence interval then would be too little too late. The Bayes approach has an alternative known as the credible interval, as discussed in the following sections.

What is Bayes' theorem

Much of the information in this section is directly copied and/ or paraphrased from reference #4. Reading that paper for further background is recommended.

Bayes' theorem is a statement of conditional probability. The probability of A given that B is true can be written as $P(A \mid B)$. However, this is only a small part of the overall power of Bayes' theorem. The vertical bar ($\mid$) represents the expression "given that I know." It is an epistemological statement (i.e., pertaining to knowledge) rather than a statement of frequencies and proportions. As such, it represents a subjective degree of belief.

One can use the Bayesian approach to treat a parameter of interest (e.g., PFD in the case of a SIF) as a random variable. One can assign a prior distribution, which is an uncertainty distribution that represents the belief in a certain parameter value over its range. The parameter is variable while the range is fixed. Contrast this with the frequentist interpretation above of confidence interval; the parameter is fixed while the range is variable. Using Bayes, the parameter of interest can be represented within the credible margin of distribution. For instance, the PFD of a SIF will be somewhere between 1 and 0.00001.

The initial uncertainty distribution is referred to as "the prior". When data becomes available, the prior knowledge is updated with the new evidence, such as statistics obtained from tests and other observations. The result is represented as a posterior distribution to better represent the parameter of interest. The new data can include subjective judgements. The Bayesian approach is sometime called the subjectivist approach to parameter approximation due to its ability to include subjectivism. This concept might raise some concern among frequentist practitioners. Yet knowledge is "belief justified". When we talk about belief (a.k.a. subjectivity) we are talking about a degree of knowledge. Each time an update is performed, it should follow a documented, justified approach, similar to any management of change process. This should prevent gross misuse of the tool. Even so, the Bayes methodology is somewhat immune to a "rouge" update. When the prior is updated with the new data, the prior does not vanish in the update. Its range of credibility merely shifts, as will be discussed. Besides, every process safety decision made with current tools today is with a degree of subjectivity, including what generic data to use.

Bayes' theorem is summarized in Figure 2.

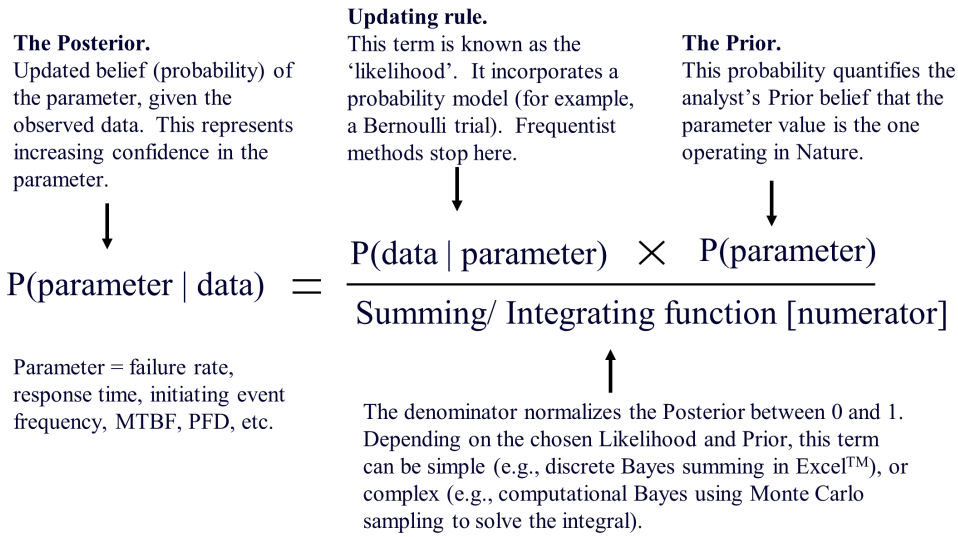


Figure 2: Bayes' Theorem

Bayes' theorem works with probabilities or probability distributions. Frequentist methods stop at the likelihood term. Calculating the value of the denominator determines the complexity of implementing Bayes, ranging from Excel™ based discrete methods, to Markov Chain Monte Carlo computational methods.

Data is represented in a distribution — also known as a Probability Density Function (p.d.f.) — such as a uniform distribution (the "bell curve") or any number of other distributions. It is up to the user to choose the distribution they feel most closely resembles reality given the data available for converting a SIF's performance (i.e., PFD) to a p.d.f. One useful distribution would be a binomial distribution, although there are numerous models available, and some may better represent a SIF's PFD compared to what this author is proposing. The goal of this paper is to discuss the concept. An example of possible distributions using a discrete (Excel based) binomial model is shown in Figure 3.

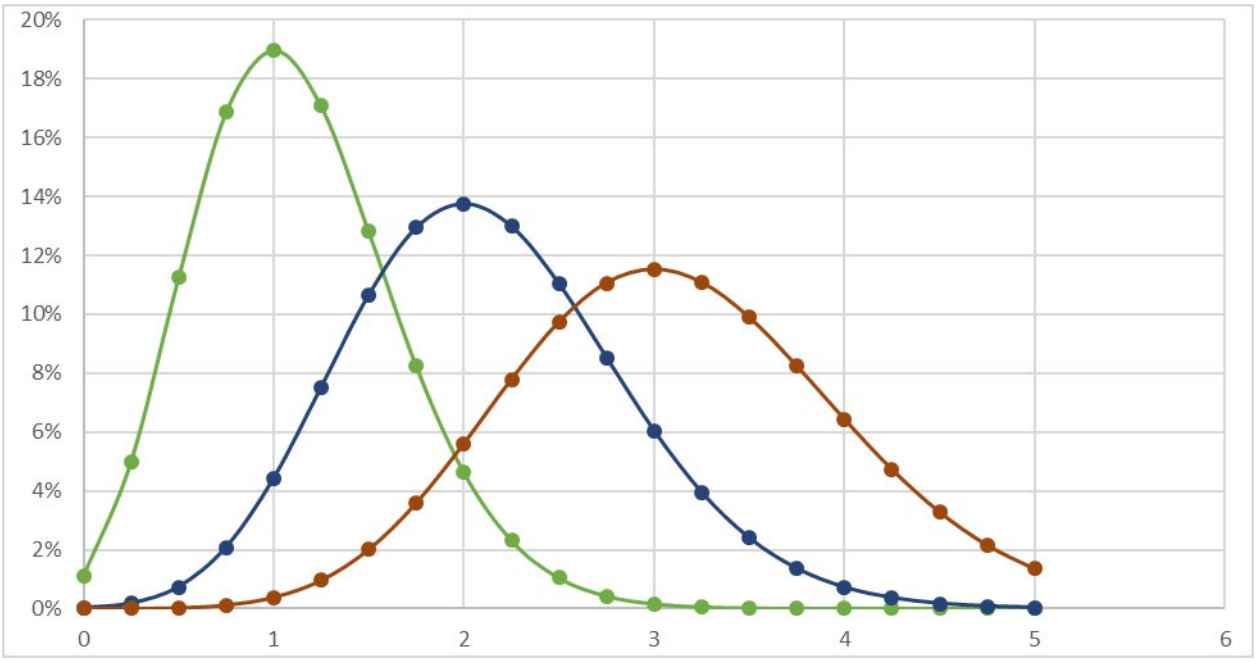


Figure 3: Sample binomial distributions for SIL 1, 2, and 3.

Figure 3 shows three different binomial distribution probability density functions (p.d.f.) for SIL 1, SIL 2, and SIL 3 SIFs with mean values corresponding to 0.1, 0.01, and 0.001. The x and y axis in Figure 3 require explanation. The x-axis essentially represents SIL range (including intermediate values). The SIL ranges are representative of PFD values corresponding to SIL. This approach was taken, since any graph of PFD that ran from 1 to 0.00001 would have values in the SIL 1, 2, 3 range compressed to a point of incomprehensibility. In order to circumvent that issue, converting PFD into SIL range allows for the graph to be divided into orders of magnitude. For example, the x-axis value of 1 equals a PFD of 0.1, the lower bound of SIL 1. An x-axis value of 2 equals a PFD of 0.01, the lower bound of SIL 2, and so on. The x-axis shown does contain values of 5 and 6. These are not real SIL numbers and are shown simply due to the author's inability to remove them from the Excel graph.

The remainder of this paper will refer to SIL numbers corresponding to the graphs. Some values will be in between whole integer numbers (e.g., SIL 1.5). It is understood that these are not real SIL numbers. They are used in this paper for simplicity sake to indicate a PFD halfway between the low bound of SIL 1 and SIL 2. For example, SIL 1.5 equals a PFD of 0.02, or RRF of 50.

The y-axis in Figure 3 represents the probability that the PFD is any particular value. 19% on the SIL 1 (green line) indicates there is a 19% chance that the true PFD value is 0.1.

The probability density function can also be represented as a cumulative density function (c.d.f). This graphically plots the summation of all p.d.f. numbers to provide a graph which shows cumulatively the probability of a value being at or below the range of possible values, as shown in Figure 4.

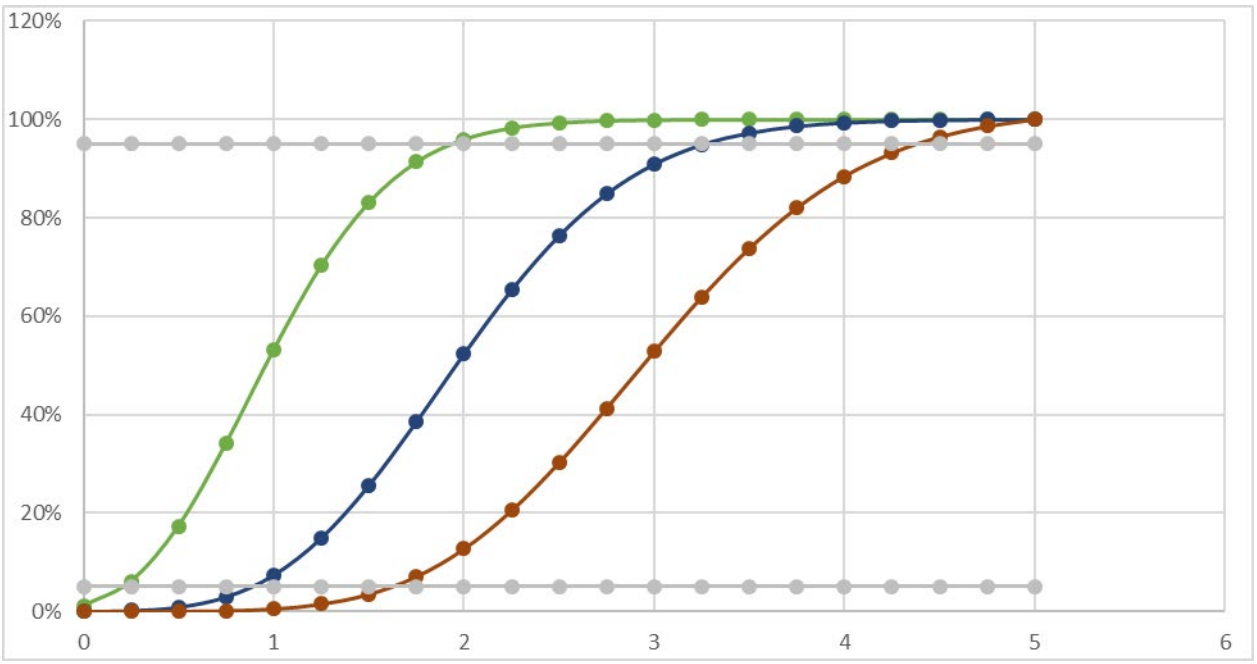


Figure 4: Sample cumulative distributions for SIL 1, 2, and 3.

Figure 4 represents the same SIL 1, 2, and 3 distributions shown in Figure 3, but depicted as cumulative distribution functions. All probabilities from the p.d.f. are summed up stepwise to 100%.

The x-axis is the same as in Figure 3. The y-axis of the c.d.f. is slightly different. The y-axis is the cumulative probability the parameter is equal to or less than the specific x-axis value. One can now make the statement that for any given x-value, there will be a y% probability that the true value is at or less than the x-axis value. For instance, the SIL 1 (green line) distribution indicates there is a 95% chance that the true value of the parameter modelled by this graph is less than SIL 2 (i.e. ~.01 *or greater* in terms of PFD).

One important use of the c.d.f is to determine the credible range. Horizontal grey lines are shown in Figure 4 at the 5% and 95% cumulative probability ranges. This allows one to make a 90% credible statement (95% - 5% = 90%) that the true value (the parameter) lies within this range. Using the SIL 1 (green line) distribution, one could say it is 90% credible that the value lies between SIL 0.25 and SIL 2 (i.e. PFD between 0.4 to 0.01). While this is a function of the binomial distribution chosen in this example, the chances are that every single SIL 1 SIF in existence has a PDF value in this range.

Credible range is similar to — but not the same as — confidence interval. Confidence interval is a statement that there is a certain chance that a sample contains the true value, which is assumed to be fixed. The Bayesian approach assumes the credible range of the parameter (via the prior) to make a statement that the true value is contained within the sample. They seem very similar and given enough data it is likely that the two would align. Yet a statement of confidence can only be made if the true average value were known, which would require ample data. Credibility does not have to know the true value of the parameter in advance, that is what we are trying to find. If one did not have enough data, yet tried to make a statement of confidence, it would be meaningless (or the confidence should be a lot lower than 95%).

Using a Bayesian methodology, a 90% credible range is stated as a low to high bound. This states that one is 90% certain that a parameter's true value lies between the low and high bounds. As more knowledge is gained, this 90% range shrinks (i.e., the range narrows) and gets closer to the actual value. As the 90% credible range shrinks, one feels more certain of where the actual parameter value lies. Consider the following depiction of a shrinking 90% credible range, as shown in Figure 5.

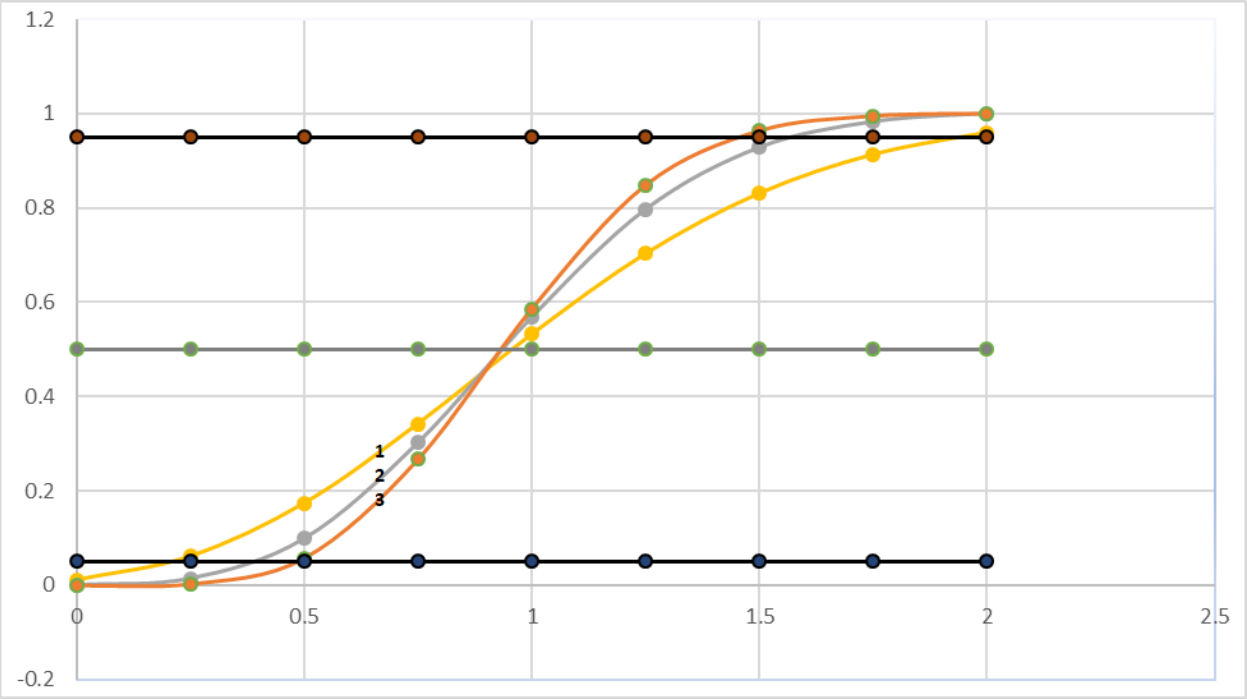


Figure 5: Example of shrinking credible ranges.

Figure 5 shows three different c.d.f. plots and two horizontal lines along the y-axis at 5% and 95%. The horizontal lines for 5% to 95% depict the 90% credible range. Line 1 in this c.d.f. is the same SIL 1 c.d.f. from Figure 4, with a cumulative range of SIL 0.25 to SIL 2. Two iterations of Bayesian updates have been applied to depict the concept of shrinking the credible range. The range for line 3 has narrowed and the parameter's 90% margin lies between SIL 0.5 and SIL 1.5. This might not look like much graphically, but the certainty has increased by almost a full order of magnitude with just two samples!

So which methodology is more appropriate for attempting to parameterize rare event occurrences such as PFD, a Bayesian credible range, or a frequentist confidence interval? Will it ever be possible to collect enough data to know what the true average is? The answer seems obvious. A good comparison between credible interval and confidence interval comes from Wikipedia [7]. "Credible intervals are analogous to confidence intervals in frequentist statistics, although they differ on a philosophical basis. Bayesian intervals treat their bounds as fixed and the estimated parameter as a random variable, whereas frequentist confidence intervals treat their bounds as random variables and the parameter as a fixed value. Also, Bayesian credible intervals use (and indeed, require) knowledge of the situation-specific prior distribution, while the frequentist confidence intervals do not."

Figure 6 was taken from the NUREG handbook [8] to compare frequentist 90% confidence vs. Bayesian 90% credible ranges.

	Frequentist	Bayesian
Interpretation of 90% interval for a parameter	If many data sets are generated, 90% of the resulting confidence intervals will contain the true parameter. We do not know if our interval is one of the unlucky ones.	We believe, and would give 9 to 1 odds in a wager, that the parameter is in the interval.

Figure 6: Comparison of frequentist vs. Bayesian methodologies

An example of the concepts discussed in Figure 6 is shown in the cartoon in Figure 7 [9].

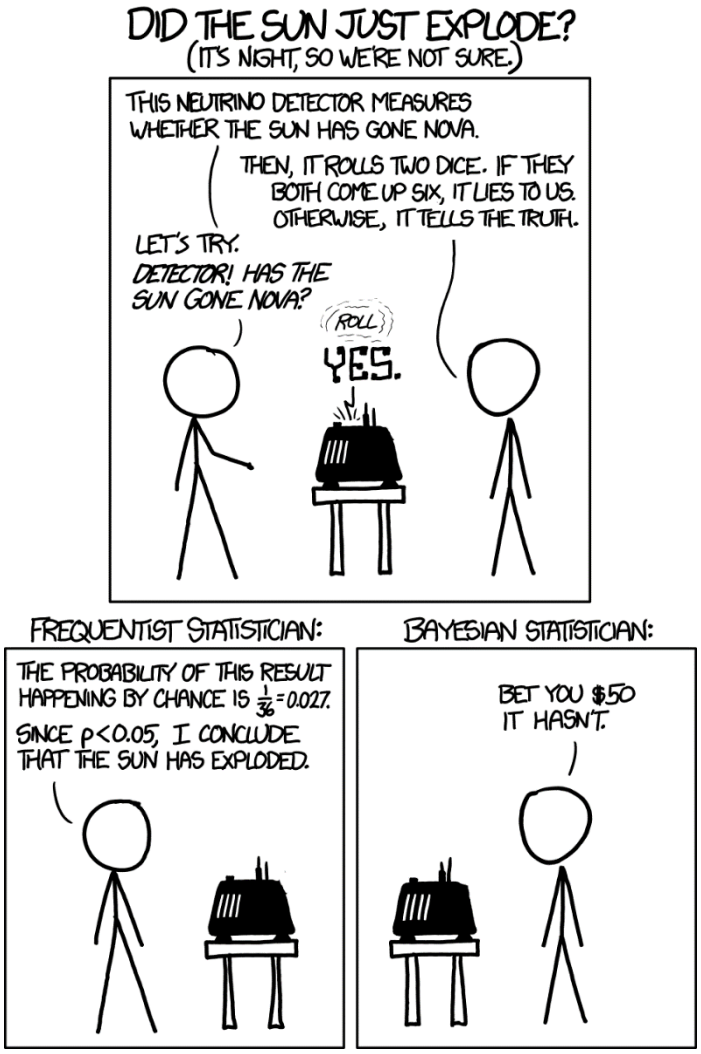


Figure 7: Example of frequentist vs. Bayesian thinking styles

The frequentist assumed that since the probability of the results being a lie was less than 5%, he could make a statement with 95% confidence that the sun has exploded. But had he collected enough data? He should have rolled the device's dice again, or maybe he had a faulty device to begin with. To paraphrase the NUREG Handbook, his sample (device) just so happened to be one of the unlucky ones (or lucky in this case). Contrast to the Bayesian statistician who has such an ample body of evidence in the prior that even given this result, it has barely shifted his belief that the sun will indeed still be there.

What can Bayes do for modelling SIFs?

When considering why one should use a Bayesian approach vs. a frequentist approach in evaluating and managing SIFs, it is useful refer to the NUREG Handbook. "In the Bayesian setting, probability is a measure of uncertainty, a quantification of degree of belief. The Bayesian methodology is used to modify uncertainty in a logically coherent way, so that "degree of belief" is rational, not merely personal opinion. In this methodology, each unknown parameter is assigned an initial prior probability distribution. This does not mean that the parameter varies randomly, but only that it is unknown, with the probability distribution modelling belief concerning the true value. […] The frequentist approach is quite different. The probability of a random event is defined as the long-term fraction of times that the event would occur, in a large number of trials. Probabilities are used only for random quantities, the possible data values. Probability distributions are never used to describe parameters, *because the parameters are [assumed] not random*. When quantifying uncertainty in an estimate, a frequentist asks questions such as, "Under similar conditions, what other data sets might have been generated? From data set to data set, how much variation would be seen in the parameter estimate? For any one data set, how far might the estimated parameter be from the true parameter?" *Any prior or external information about the parameter value is ignored.*" (Emphasis added.)

Note that the frequentist approach asks the question "under similar conditions, what other data sets might have been generated?" This is a direct parallel to the use of averages in PFD calculations. Recall the discussion from an earlier section stating that average data is not representative of any particular plant with its own systematic biases. How can one attempt to generate a statement of confidence with data representing someone else?

Further quoting from the NUREG Handbook, "Statisticians have argued vigorously over which approach is preferable. When estimating parameters for PRA (probabilistic risk assessment), the Bayesian approach clearly works better, for two reasons.

First, data from reliable equipment are typically sparse, with few or even zero observed failures. In such cases, it is reasonable to draw on other sources of information. The Bayesian approach provides a mechanism for incorporating such information as prior belief.

Second, the Bayesian framework allows straightforward propagation of basic event uncertainties through a logical model, to produce an uncertainty on the frequency of the undesirable end state. […] The frequentist approach cannot handle such complicated propagation of uncertainties except by rough approximations."

These concepts can fit well within the management of SIFs due to the (hopefully) infrequent activation and failures of these devices. Data would (hopefully) be too sparse to form a valid frequentist statement of confidence.

The Bayesian approach requires developing an initial subjective prior distribution and updating it with new evidence (as it becomes available) to generate a posterior. The posterior becomes the new prior the next time one gathers more evidence. Figure 8 is from reference #4.

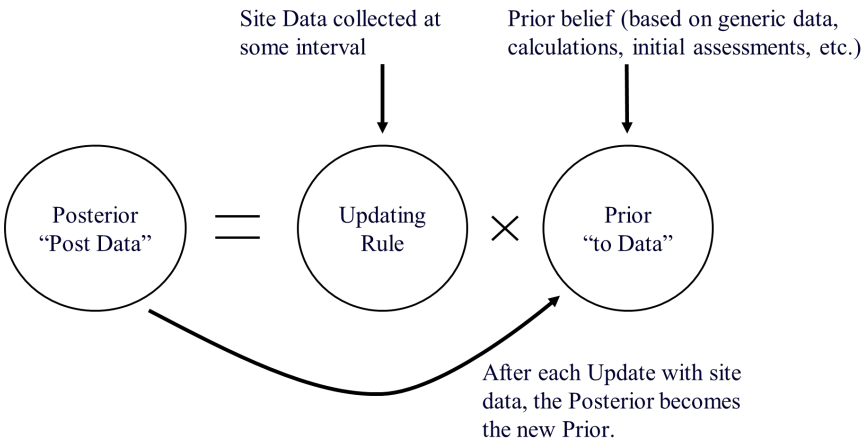


Figure 8: Simplified Bayesian engine

Figure 8 shows a Bayesian engine in simplified form. The prior belief is the starting point which is typically based on calculations using generic data and initial audits and assessments. As the Bayes engine runs, the confidence in the posterior result increases.

New data in the Bayesian approach can be anything that could validly shift the credible range. This could include data from key performance indicators (KPIs) or other red flags. For example, a plant may have had a large staff turnover due to experienced employees going into retirement. If all the employees who are left have only one year of operating experience, one should not feel as confident that the plant or SIF will be operated and maintained as well as in the past. At the very least one should be able to admit that the previous staff had a better understanding of the SIFs than the new staff. One could make a judgement call that the current credible PFD margin ought to change. It would be likely that the 90% credible range would become "wider" and perhaps even the mean value would shift downwards. Perhaps not such a drastic shift as to throw the previous credible range outside the realm of possibilities, but there should be a definite shift to allow for lower results. Bayes theorem allows the application of this new evidence. One could then generate a distribution that would revise the prior data and widen the credibility margin, as shown in Figure 9.

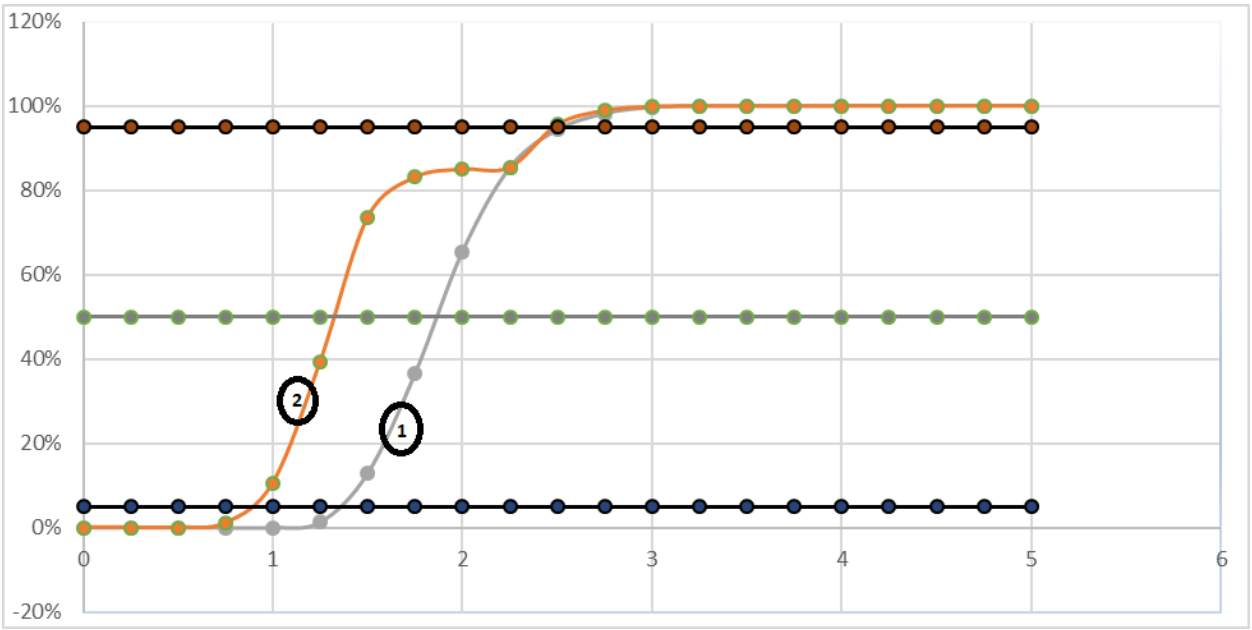


Figure 9: Updating a credibility margin

The graphical method using Excel may be viewed as clunky and there may be a better way to model the change in distribution. The author converted the earlier statement of belief (i.e., "the staff's ability to operate and maintain the SIF has degraded") into a distribution. The new evidence distribution (not depicted Figure 9) simply took a SIL 2 binomial distribution and manually shifted the lower range down closer towards a SIL 1, while maintaining the original SIL 2 upper probability ranges past SIL 2.5. This new evidence distribution was used to update the prior line 1 to the new posterior distribution shown with line 2. The posterior line's 90% credible range is still roughly at the same upper value (i.e., SIL 2.5), but the lower credible value has dropped from SIL 1.25 to SIL 0.75. This example shows the same level of performance in the upper range yet acknowledges the loss of experience might have a detrimental impact in the lower range.

Figure 9 has an additional horizontal line at the 50% y-axis cumulative probability. The mean value (average) can be approximated by the 50% cumulative value. There are mathematical methods to calculate the mean of a Bayesian distribution, but given the limitations of the discrete model in Excel, this is a good enough approximation. One might consider making the statement that the average value equals a frequentist calculated PFD. The mistake would be that a frequentist average is assumed fixed. Also, the true average per the frequentist approach could be outside the 90% credibility range. The Bayesian average shifts each time a new update is performed. One should be able to correct any out of bounds averages one may have started with as more evidence becomes available. An example would be starting with a SIL 3 design, but after collecting a few bad samples, updates show the function only meets SIL 1.

The notion that a 50% value is representative of the PFD is likely close to the truth and is an assumption of this paper. The 50% mark on a Bayesian c.d.f plot y-axis means that it is equally credible that the PFD/SIL value is higher or low than the corresponding value on the x-axis. One should not fall into the frequentist trap of saying the PFD/SIL *is* the average value. As stated previously, averages can shift using Bayesian methods. The comparison can be useful however, especially given that the frequentist PFD is a distribution itself (i.e., an average based on a range

of samples). If the task with PFD calculations is to "prove" that a SIF meets a given target, what better way than to utilize the "middle of the road" value?

The suggested approach using Bayes' theorem is to begin with some basis for the initial value, most likely from a traditional PFD calculation. One could then convert the PFD value into a distribution. Each time an update is desired, rather than re-run a PFD calculation, one could revise the previous PFD distribution. *In other words, we are only concerned about failure rate data for the initial calculation. Once the prior has been established, the focus is on the overall PFD distribution (i.e., the credibility range)*. Any changes to the system could be factored in as a Bayesian update. This would provide more meaning results than re-running a frequentist calculation with numbers that often cannot be trusted. (Refer to the Annex for more details.)

What are the drawbacks to the Bayesian approach?

One potential drawback of a Bayesian approach is the subjectivity. An unrealistic update could potentially skew the results in an unlikely fashion. However, the updating rule does have a measure of protection against shock changes since the prior data is never thrown out. Also, any changes should have some form of documented justification to protect against rouge data.

Another apparent drawback of the Bayesian approach is that if one relies only on quantifiable data (e.g., SIF demand rate and actual performance), once the initial noise of the system settling into place has been established (if the first prior did not closely match true operations), the 90% credible interval will only shift very slowly without many samples (which is typically the case for a low demand SIF). For example, consider Figure 10.

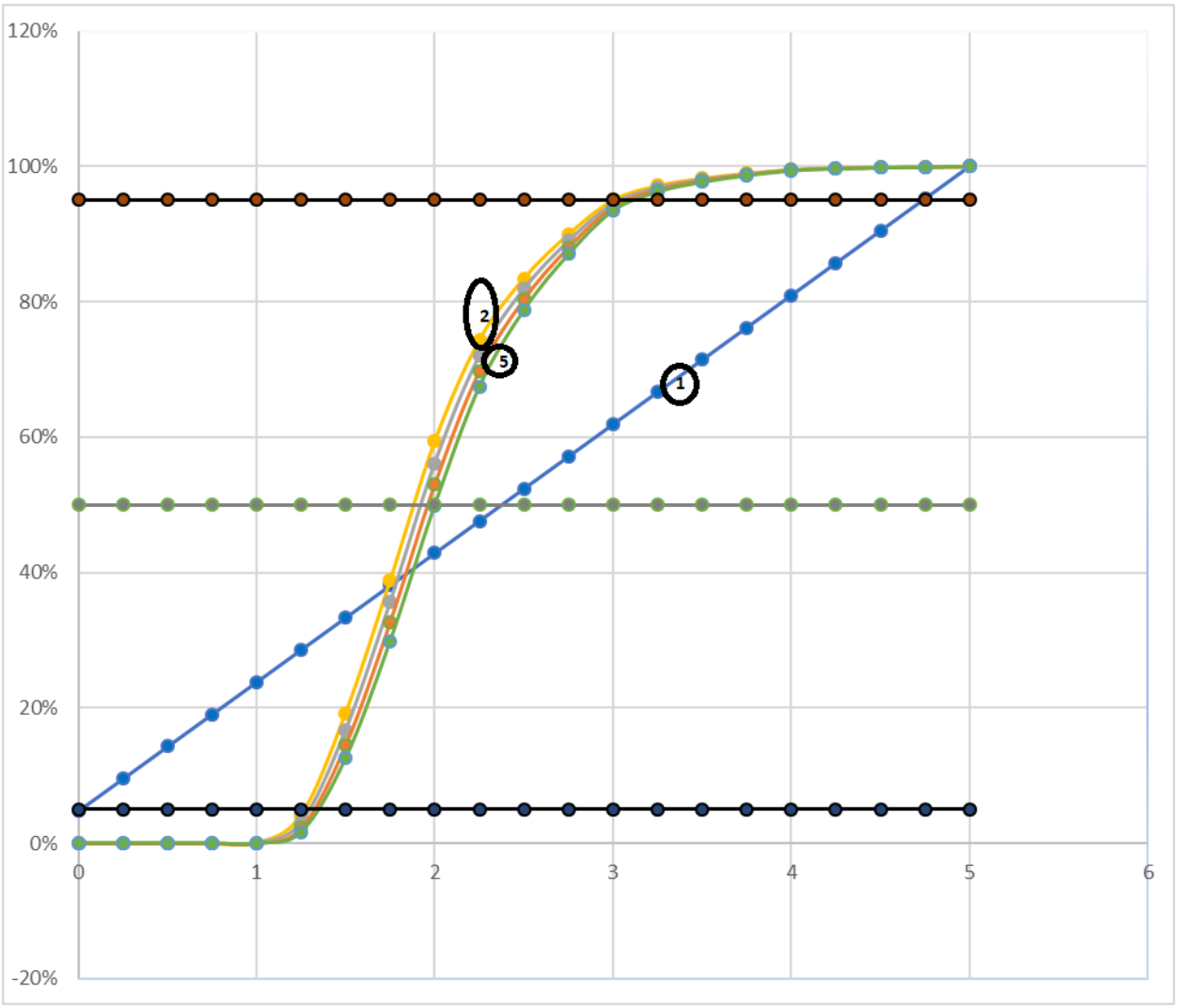


Figure 10: A uniform prior distribution (line 1) updated to binomial distributions (lines 2–5)

Line 1 if Figure 10 represents an initial prior as a uniform distribution (i.e., a flat distribution where every possible value of our parameter is equally likely). The first update for this example has applied the first PFD calculation result (0.01) to a binomial distribution and updated the uniform prior, shown as line 2 on the far left. Imagine operating the SIF over the next 15 years and challenging the function 30 times over that period (e.g., a small batch process that actuates the SIF twice a year). This is represented with lines 2 through 5. The credible range has shrunk (i.e., the final line is steeper) and the average has improved, but the change is hardly significant. It is likely that frequentist methods would have come up with a similar result.

Further problems with relying on quantitative data only (based solely on samples and results) is that if there are relatively few samples, the 90% credible region will not shift much until there is either an abundance of samples (e.g., unwanted spurious trips), or a failure. Once a single failure occurs, if the system experiences no further failures, the credible interval will initially shift leftwards (towards SIL 1 or less), but with a lack of samples afterwards, the recovery will be a long time coming. This is shown in Figure 11.

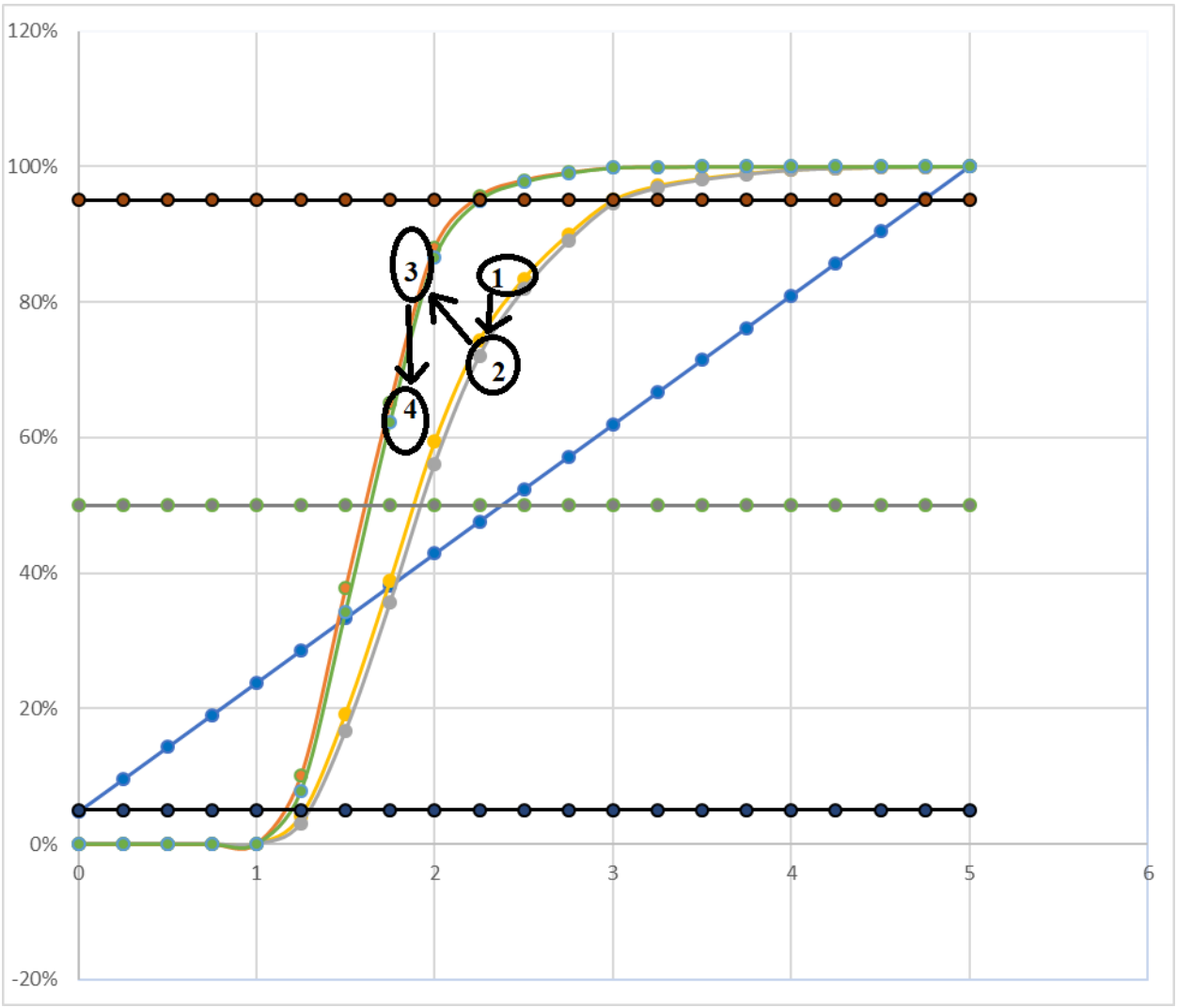


Figure 11: Incorporating the impact of a failure and recovery

Figure 11 is similar to Figure 10 with the same number of samples (30 over a 15-year period). However, it is based on a failure occurring between the times represented by lines 2 and 3. This would be sometime between the 5- and 10-year marks when testing twice a year (i.e., every batch rotation). Line 3 shifted the 90% credible range leftwards. Between years 10 and 15 (line 4) there were no further failures for another 10 samples. This results in the credibility range shifting back rightward towards SIL 2, although though not very quickly. A low SIL 2 (a PFD between 0.01 and 0.004) is still within the 90% credibility range even with one failure. Quantifying this in frequentist terms with an original calculated PFD value of 0.01, and assuming 30 samples with 1 failure, would result in a PFD point value of 0.015 (a RRF of 65). Maintaining SIL 2 performance would no longer be possible using the frequentist approach.

If one only updates PFD distributions with quantifiable data (i.e., samples and failures), and samples are taken infrequently (e.g., once every turn around when proof tests are performed), with a proof test every 5 years the Bayesian approach is not going to show much difference from a frequentist approach unless there is a failure. Once the failure occurs the shock to the system

will not be as bad as indicated using the frequentist approach, but the recovery will be slow without an abundance of samples.

Updating prior distributions with qualitative data

The apparent drawback of the Bayesian approach being slow to update when relying on quantifiable data need not be. Bayesian updating allows subjective judgements to be factored into the distribution. One way to utilize qualitative data would be to incorporate the results of a stage 4 functional safety assessment (FSA). An FSA is an assessment from an independent team that assesses whether the documentation and historical record of the operations and maintenance of a SIF conforms to the original design assumptions. The assessment makes a judgement call as to whether there are gaps. The assessment could be based upon a repeatable and well documented checklist to maintain consistency. The results of such an FSA could be qualitatively assigned to a Bayesian engine as new evidence to shape the belief of a new update distribution that may widen, narrow, or shift the prior's 90% credible interval.

For an example, assume an assessment is performed on a SIL 2 SIF with an original calculated PFD of 0.01. Incorporating an FSA update may indicate the SIF is operating at, below, or above the original calculated SIL 2 level. This could be interpreted as a vote of certainty that the SIF is actually operating closer to SIL 1.5 (a PFD of 0.05), SIL 2 (a PFD of 0.01), or SIL 2.5 (a PFD of 0.005). These point values could be shown as binomial distributions, as shown in Figure 12.

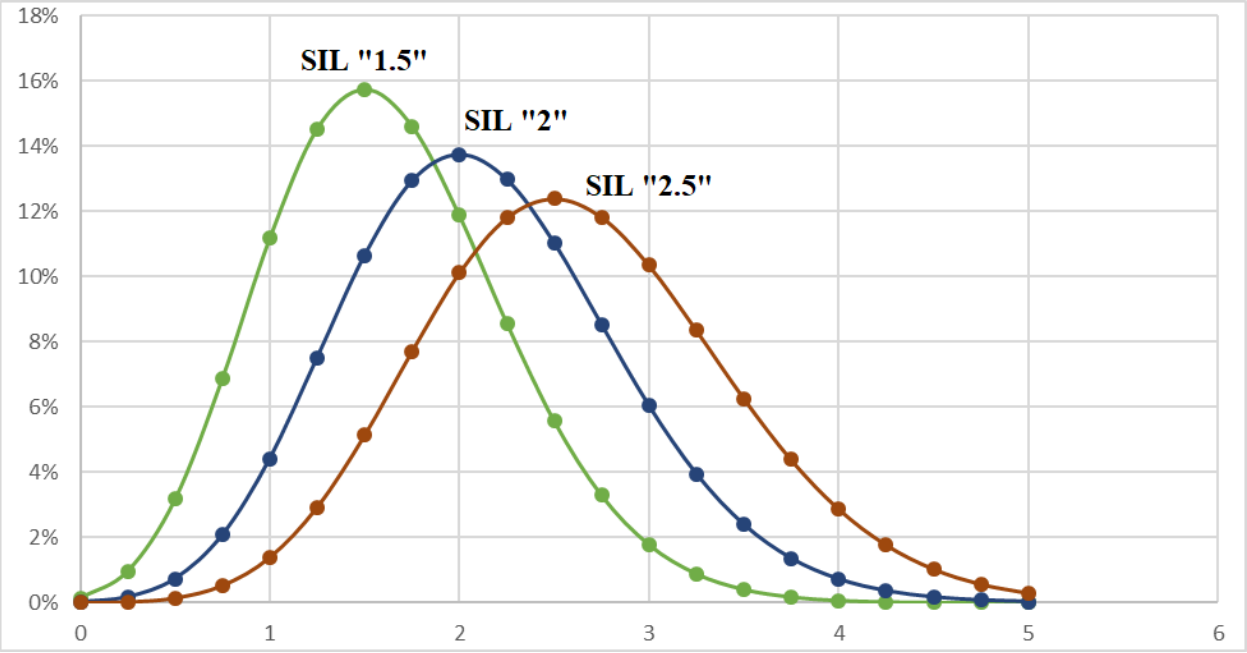


Figure 12: Incorporating the impact of operating below, at, or above expectations

Utilizing this assessment "judgement call" as an updating factor to a Bayesian prior gives one a way to recover more quickly from a single bad datapoint (i.e. a single failure with slow recovery). As a side note, the qualitative updating distribution could use a more suitable distribution such as a gamma or beta distribution with a shape parameter applied. However, the author has chosen to stick with the already familiar binomial distribution for simplicity.

As an example of qualitative updating, consider an earlier example of faulty instrument air to the solenoid. The originally calculated PFD of the SIF was 0.04 (via frequentist means), or a RRF of 25. A proof test was performed every 5 years, and the first test showed the SIF was in the failed state. Plugging this example into a Bayes engine assuming no failures for the next 15 years, and continued sampling every 5 years, would show an initial shock to the system that would shift the 90% credible margin leftwards. The performance recovery back towards the original calculated value would be slow, as shown in Figure 13.

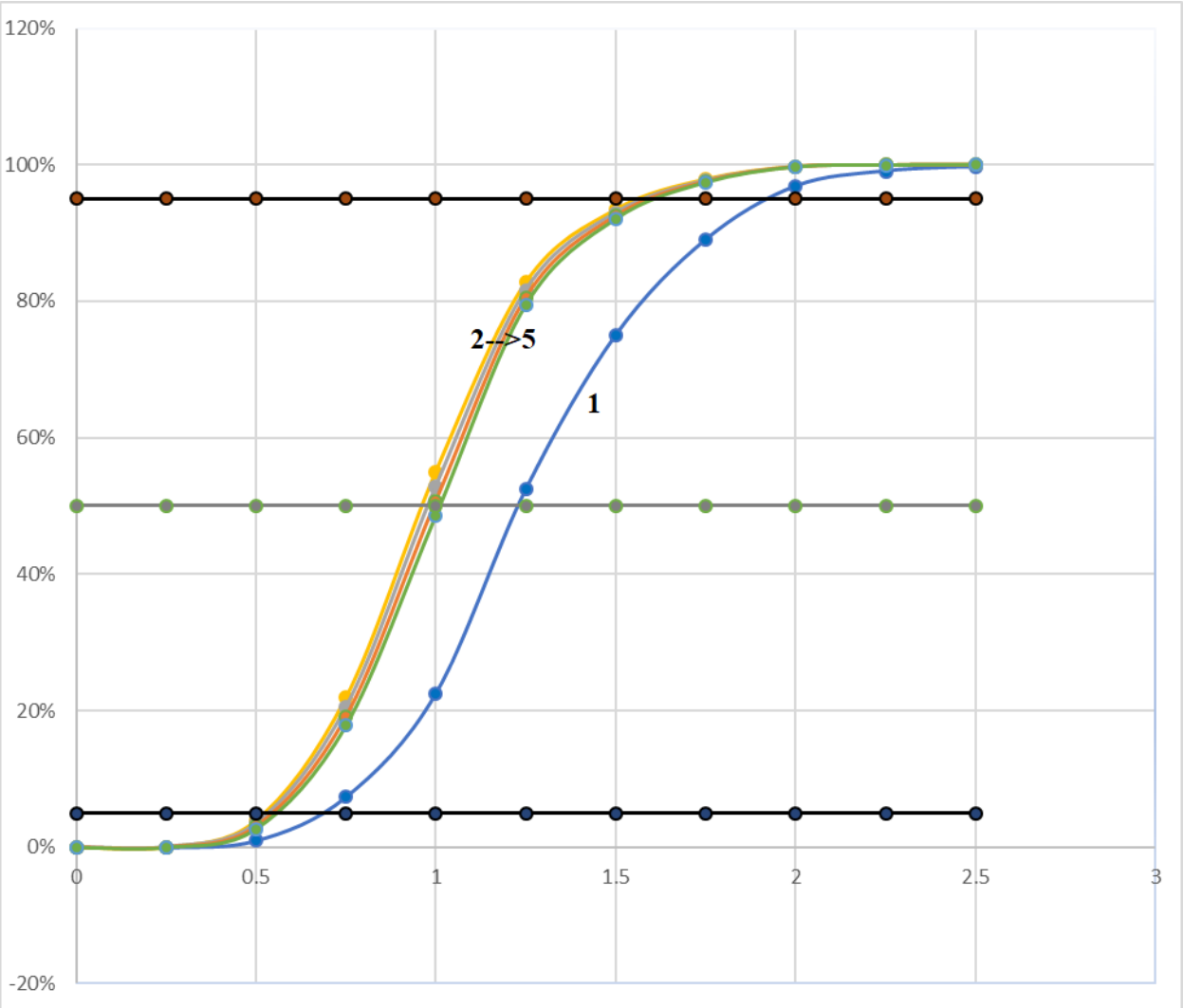


Figure 13: The impact of a failure (line 2) and steady recovery (through line 5)

Line 1 represents the original failure distribution with 90% credibility range from approximately SIL 0.75 to almost SIL 2. Lines 2 through 5 show the progression from a single failure through three more updates every five years. The initial drop in credibility shown by line 2 reduces the margin range to between SIL 0.5 and SIL 1.5. Note that the PFD of 0.04 (i.e., SIL 1.25 per the graph) is still within the 90% credible range, but it is on the outskirts. If one wishes to base performance off the 50% credibility value (i.e., the approximate mean value), the mean is around 0.1, which is greater than the PFD target (if we can call it that).

Suppose evidence that the instrument air system was faulty is discovered after performing the first test and that was determined to be the cause of the solenoid failure. Imagine the problem being corrected, leading staff to feel confident that the SIF's failure mechanism has been resolved. Staff believe the function should now be back to operating at the originally assumed average PFD of 0.04. One could update the previous graph with a binomial distribution centred at 0.04 and come up with Figure 14. (Note: Only steps 1, 2, and 3 are shown.)

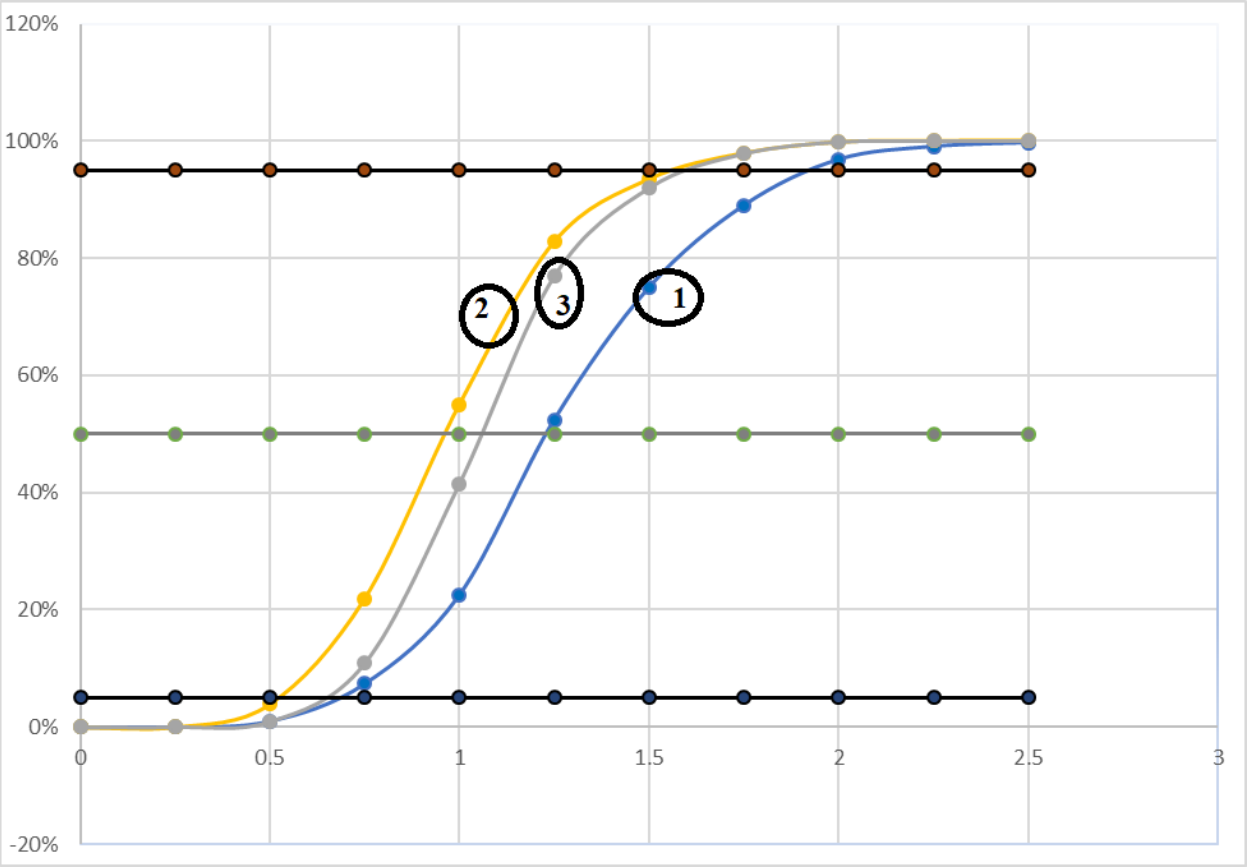


Figure 14: An update (line 3) based on subjective input

Figure 14 shows the original line 1, the failure represented by line 2, and the judgement that the problem was fixed represented by line 3. Line 3 has not jumped back to the original curve (i.e., line 1), but the performance is closer to line 1 than assumed and shown in Figure 13 compared to waiting for the next 15 years to indicate anything different.

Imagine now that after fixing the faulty air system, the facility decides to perform a stage 4 FSA per the requirements in the 2018 edition of the standard. The results of the assessment paint a good picture and the team believes the SIF is operating better than originally assumed, perhaps a PFD of 0.02 rather than 0.04. The update might look like that shown in Figure 15.

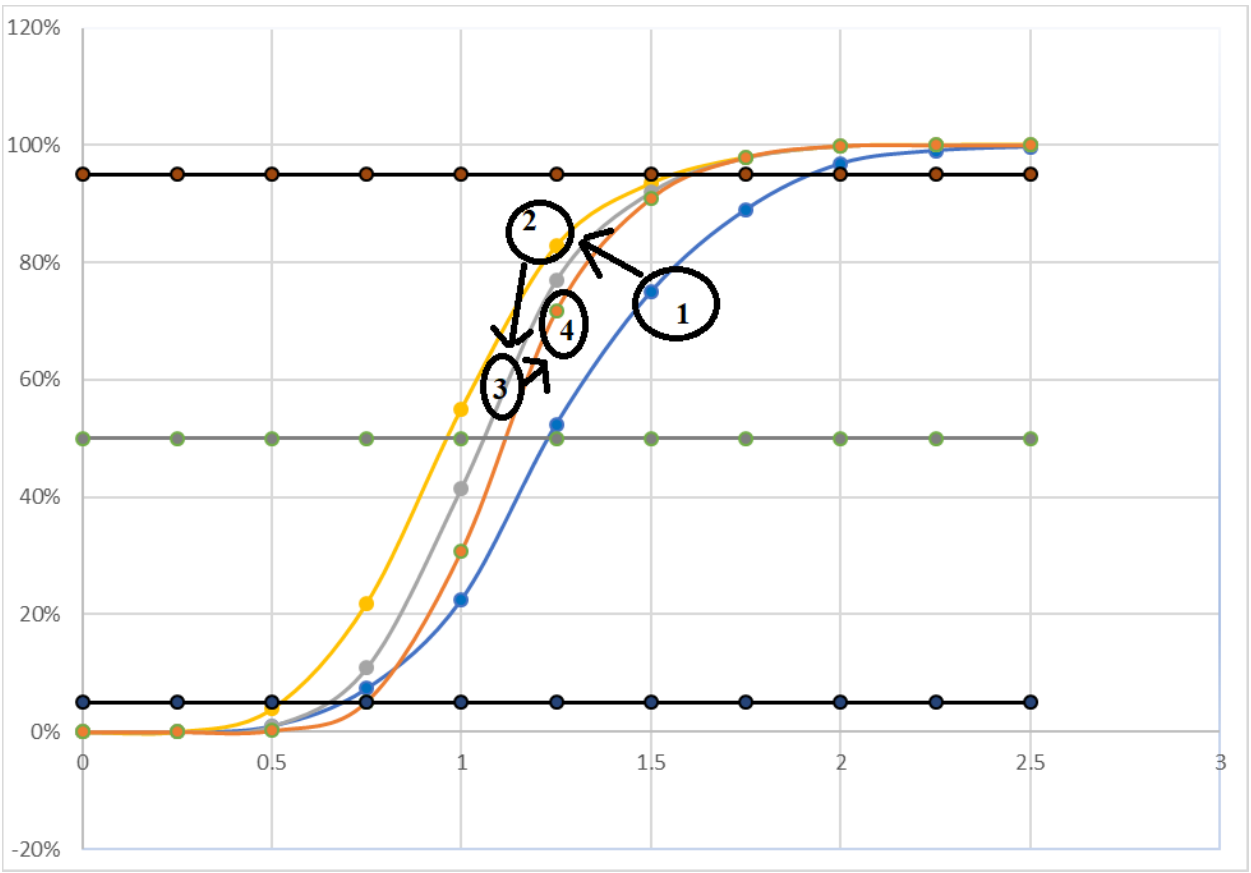


Figure 15: An update (line 4) based on further subjective input

This figure shows the original line 1 line, the first failure indicated with line 2, the judgement call of fixing the air problem indicated with line 3, and finally that statement of confidence based on the favourable FSA indicated with line 4. Note that line 4 has almost returned the 90% credible region to the original prior. In fact, the lower bound is now better than the original assumption, yet the upper bound is taking longer to recover. This makes sense given the history described. The knowledge gleaned from this level of detail is the point of applying Bayes theorem.

Many other opportunities exist to factor in qualitative parameters besides stage 4 FSAs. Updates can come from officially documented belief influencers such as audits, human reliability analysis (HRA), and assessments, to more subtle influencers such as staff turnover, new projects, lagging maintenance, delayed testing, excess failures, outside influences, and policy changes. Every one of these factors could be positive or negative and could be factored into evaluating a SIF's credible PFD range.

What should we do with our frequentist methods?

Previous sections might give the impression that a Bayesian approach is the only approach, one that advocates for throwing out the entire body of knowledge contained in existing PFD calculation methods. This is *not* the proposal! The frequentist approach is actually a very *good* method for generating the *initial* belief in what a SIF should be capable of. That information could serve as the starting point for a Bayesian approach which could be updated incorporating

subjective modifiers based on operating at, below, or above targets. The prior distribution could then be carried forward into a Bayesian updating system while new evidence comes in.

Figure 16 is a graphical representation of converting frequentist based SIL values based on PFDs of 0.1, 0.01, and 0.001, to binomial distributions for use as Bayesian priors ranging from SIL 0 to the upper limits of SIL 4. (Note: SIL 5 and 6 on the x-axis are not real SILs, they are simply unavoidable values due to the charting tool in Excel).

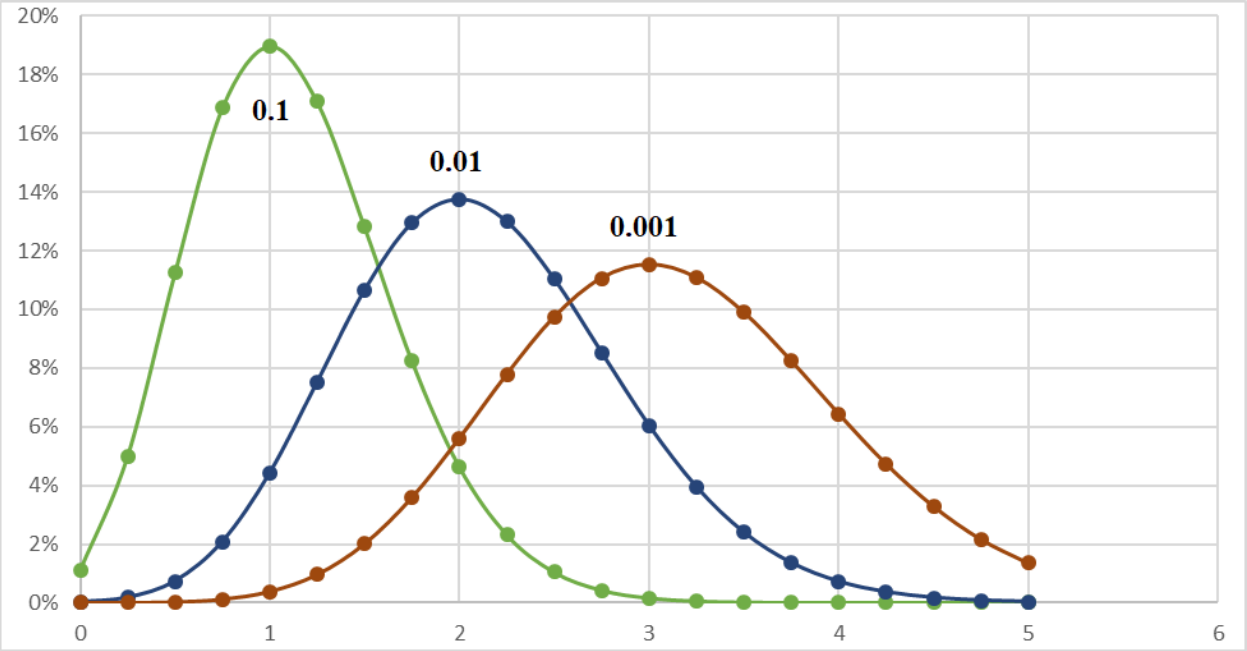


Figure 16: An example of converting frequentist based SIL values to binomial distribution priors

Frequentist PFD calculations are relatively simple to perform and provide a good starting point. No major subjective judgements are really needed. Considering the large number of greenfield projects, it may be better to start with industry averages until more is learned about how a plant actually operates.

One could potentially start with a fully Bayesian approach and subjectively come up with an initial prior distribution. However, this runs the risk of the design team being over-confident in their methodology (i.e., they may be overly optimistic with initial subjective judgements and believe they have designed a perfect plant.) Of course, the perfect plant is impossible and sooner or later the team will be proven wrong. Yet the team might have to wait a significant amount of time to collect enough evidence to justify updating the Bayesian prior. Hopefully by then it will not be too late.

To again quote from the NUREG handbook, "The primary use of the frequentist approach is in preliminary examination of the data, to check the correctness of model assumptions, and to decide which model to use… Then Bayesian methods are used for estimating the parameters. In addition, frequentist estimates are often simpler to calculate than Bayesian estimates, and therefore are useful for rough approximate calculations."

Conclusion

The topics covered in this paper might seem a little overwhelming at first, but given time and study, they will become clearer. (Details of the method used are covered in the following annex.) There may be a fear of moving PFD calculations away from "real data" by throwing subjectivity into the mix. Yet practitioners need to seriously consider whether the data industry has been using is "real." Might current approaches already include subjective, systematic factors in the failure rate data? Even the choice of what failure rate data and factors to use in a frequentist calculation is subjective. Ignoring systematic factors in the analysis of protection layer performance needs to stop. If the current industry methods do not even *attempt* to address this issue, yet standards mandate that it must be accounted for, what are practitioners to do?

Bayes' theorem provides a method to realistically model SIF performance (and that of other safety layers as well). It allows for the collection of data which is used to update performance models, rather than throwing the data out and chalking it up to bad sampling based on frequentist methods. Prior use justification of failure rate data is about as good as the industry can do with current approaches. But this can require mountains of evidence and extensive time to make even the slightest change in calculation results. If practitioners were to update their failure rate data, they would have to redo the calculations anyway. Why not redo it the right way using Bayes' theorem? Even if the data collected were from one particular plant, and there was enough to determine true average values, they would still be average values. The plant average data would not include systematic biases that might be possible even between SIFs in the same operating unit. The concept of prior use data is also ripe for the application of a Bayesian updating engine based on generic data and the qualitative evaluation of potential systematic failures on SIF performance, as covered by Steven Thomas [3].

Top 5 reasons why process safety needs Reverend Bayes [3]

1. Rare event frequencies (targets) cannot be validated experimentally. It is neither feasible nor ethical.
2. Rare events are one-off events. Long-run frequentist-based probability has no meaning in this context.
3. Rare event frequency as a statistical parameter is better interpreted as a single-sided confidence limit, not a point value.
4. Generic data is only the starting point (i.e., the prior).
5. The PFD calculation is our best leading indicator, and some effort should be made to make it trustworthy.

References

1. Can we achieve Safety Integrity Level 3 (SIL 3) without analysing Human Factors, Keith Brumbaugh
2. https://www.exida.com/Blog/the-exida-fmeda-process-accurate-failure-data-for-the-process-industries
3. A Hierarchical Bayesian Approach to IEC 61511 Prior Use, Stephen Thomas
4. Reverend Bayes, meet Process Safety. Use Bayes' Theorem to establish site specific confidence in your LOPA calculation, Dave Grattan
5. http://silsafedata.com/
6. Improving Human Factors Review in PHA and LOPA, Dave Grattan
7. https://en.wikipedia.org/wiki/Credible_interval

8. NUREG CR-6823
9. https://www.xkcd.com/1132/
10. Reliability Engineering and Risk Analysis A Practical guide, second edition, by Mohammad Modarres, Vasiliy Krivtsov, Mark Kaminskiy
11. https://en.wikipedia.org/wiki/Confidence_interval
12. Introduction to probability, statistics, and random processes, H. Pishro-Nik
13. https://www.probabilitycourse.com, Kappa Research LLC, 2014

Annex

OK great, but how does one actually *apply* all this to SIFs?

Many examples have been covered in this paper, but little information on the "how to" has been provided. Many assumptions have also been made in applying Bayes' theorem. Incorporating Bayes theorem in SIF performance evaluations *is* possible, yet there are no standards on just *how* to do this. The author's method may not be the *best* approach, but it is *one* approach.

To fully explain the techniques used is beyond the scope of this paper and would require some interface with the author. However, the key concepts used will be provided here. The author became aware of this approach through discussions with colleagues, the application of Microsoft Excel and the NUREG CR-6823 handbook (see section 6.2.2.3), additional research on probability distributions, and more.

The approach advocated is:

1. Begin with an initial PFD calculation of a SIF using the traditional frequentist approach.
   - Try to use failure rate data and other assumptions that match a plant's own systematic biases (although this is not required).
2. Convert the PFD point value to a Bayesian prior.
   - The distribution assigned will ideally be centred on the calculated PFD value and be within a credible range (i.e., a PFD between 1 and 0.00001).
   - There are many different tools available to convert the PFD point value into either a continuous or discrete distribution. Even manual setting of likelihoods is permitted. (Note: No value within the distribution should be set to 0, otherwise the value will never move from the 0% likelihood.)
   - The author has chosen to use a discrete based binomial distribution in Microsoft Excel. Any distribution is valid, even a uniform distribution. A discrete distribution using whole integer values greatly simplifies the approach. A 21-point distribution was chosen centred on the PFD value. The 21 points correspond to 4 step changes per SIL from 0 to 4, with one extra for when the PFD equals 1. More samples can be used to smooth the graph, but 21 was enough for the examples used in this paper. Figure A.1 is the same distribution used in previous figures of binomial distributions for PFD values of 0.1, 0.01, and 0.001, corresponding to SIL 1, SIL 2, and SIL 3.

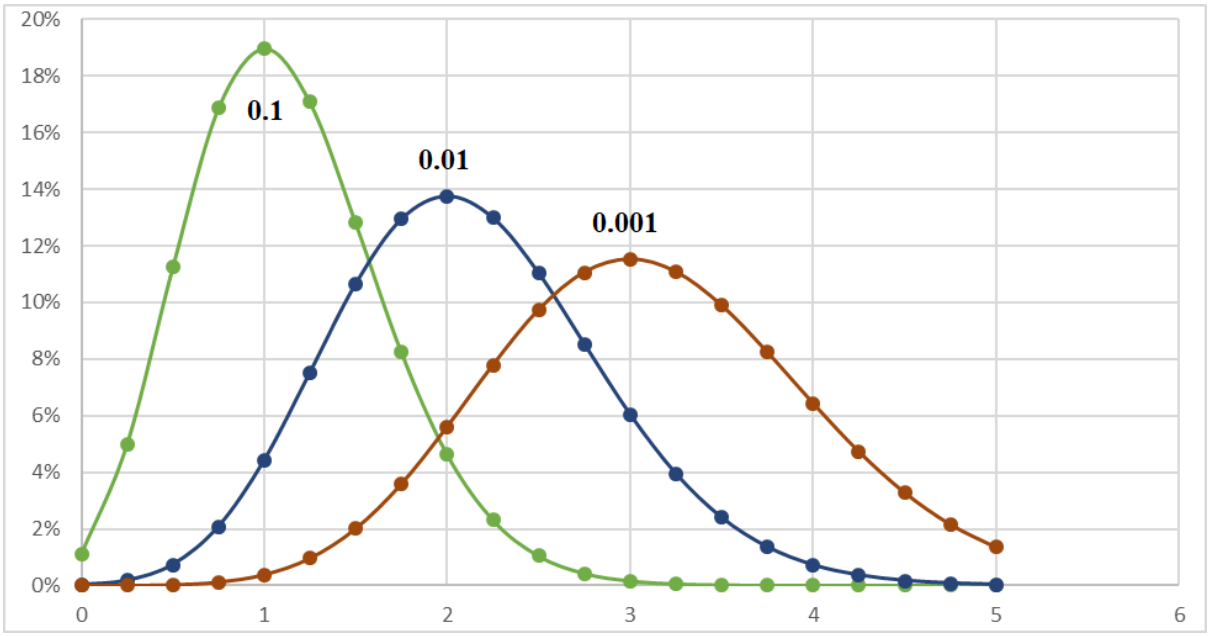


Figure A.1: Three sample binomial distributions

3. Gather and convert evidence into a usable metric.
   - Two different forms of evidence explored in this paper have been:
     i. Quantitative: based on actual demands and their results (e.g., failure or no failure).
     ii. Qualitative: based on judgement calls (e.g. FSA results, audits, assessments).
   - Quantitative evidence may be captured at any time desired with any form of distribution. The author used a binomial distribution per recommendations in the NUREG handbook. The formula for a binomial distribution is:
     i. $$P_X(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{for } k = 0, 1, 2, \cdots, n \\ 0 & \text{otherwise} \end{cases}$$
     ii. In this case, k = number of failures, n = number of demands.
     iii. p = the PFD value being modelled (i.e., the range of 21 points along the x-axis)
   - Qualitative evidence a bit more nebulous. The author has chosen to apply another 21-step binomial distribution representative of a SIL at, below, or above the target. A judgement call can be made whether one thinks a SIF is operating the same, better, or worse than expected.
     i. Figure A.2 shows the application of operating a SIF at less than, equal to, or better than targeted SIL 2 performance (with a PFD of 0.01).

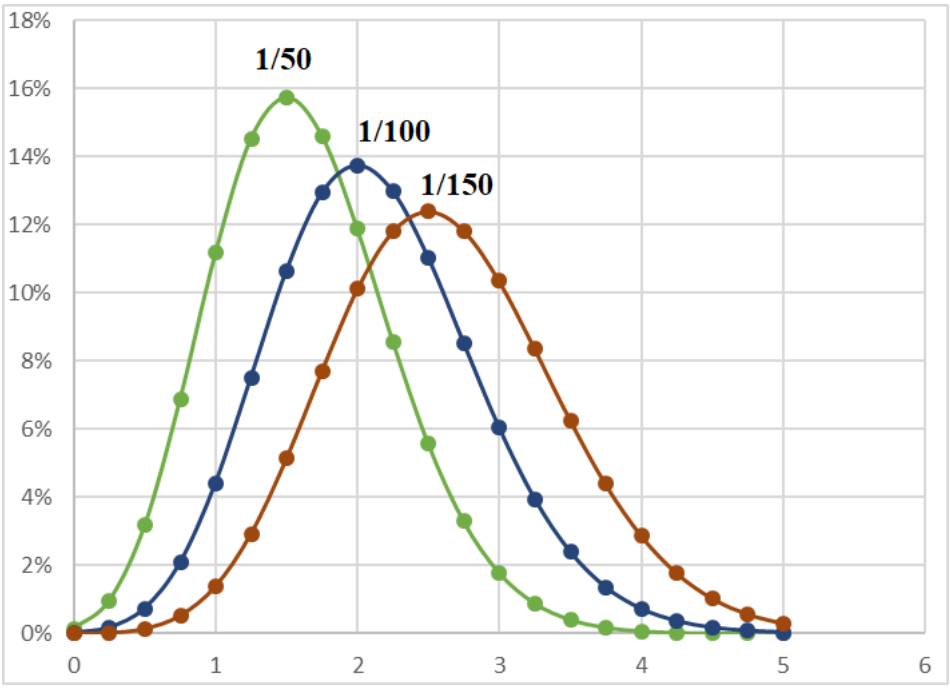


Figure A.2: Operating a SIL 2 SIF at less than, equal to, or better than expectations

4. Once the evidence distribution is known and generated, update the prior with the evidence into a posterior.
   - The steps to do this are well documented in the CR-6823 Handbook, section 6.2.2.3. The author was able to do this in Excel.
   - Figure A.3 shows probability density function for an initial prior (line 1), an update with a quantitative single failure with single sample (line 2), and another update with a qualitative assessment that the SIF has improved in performance due to operations/ maintenance improvements (line 3).

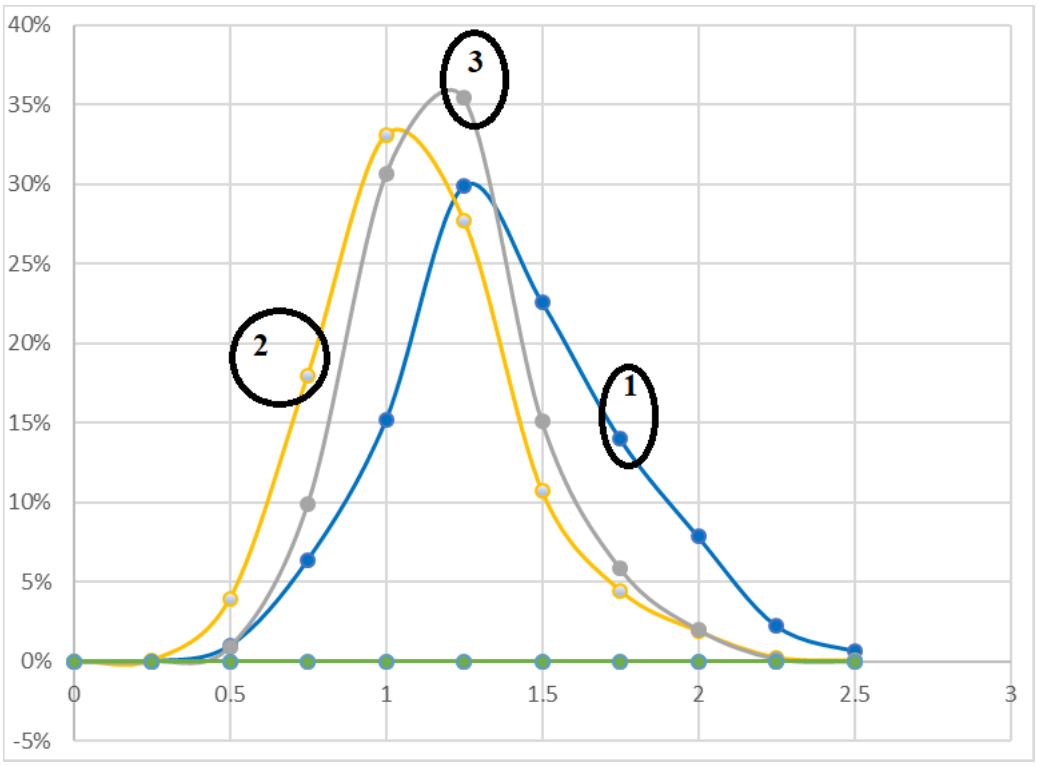


Figure A.3: An initial distribution incorporating quantitative and qualitative updates

5. Finally, one can plot a cumulative density function simply by adding the tabularized probabilities together. All probabilities should add up to 100%.
   - The values between 5% to 95% indicate the 90% credible range, meaning one can put 9 to 1 odds that the parameter of interest lies within those bounds.
   - Figure A.4 takes the p.d.f. distributions from Figure A.3, and converts them to c.d.f distributions. 5% and 95% are represented via black horizonal bars near the top and bottom of the range.

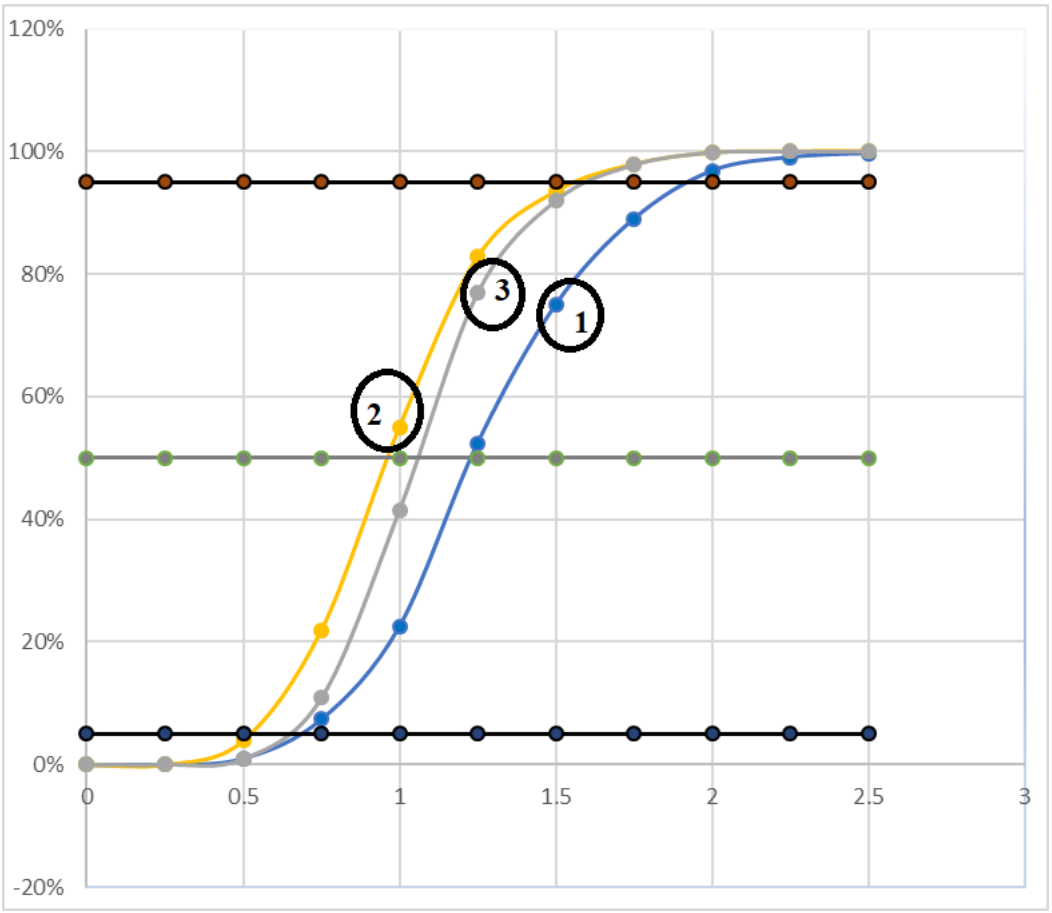


Figure A.4: The p.d.f. distributions of Figure A.3 shown as c.d.f. distributions.