Using small data to support decision making when LOPA fails:

The case for incorporating site specific process safety data into our calculations, and how to do it.

Keith Brumbaugh, P.E., CFSE
aeSolutions
Keith.brumbaugh@aesolutions.com

aeSolutions Technical Team

Prepared for Presentation at
American Institute of Chemical Engineers

2023 Spring Meeting and 19th Global Congress on Process Safety
Houston, TX
March 12-16, 2023

Using small data to support decision making when LOPA fails:

The case for incorporating site specific process safety data into our calculations, and how to do it.

Keith Brumbaugh, P.E., CFSE
aeSolutions
Keith.brumbaugh@aesolutions.com

aeSolutions Technical Team

Keywords: Data, Parameters, Updating

Abstract

If we're honest with ourselves, Process Safety has a lack of data problem. Nowhere does this show up more than in the types of calculations we perform for Layer of Protection Analysis (LOPA) and Safety Integrity Level (SIL) calculations, for example. Sure, we have generic failure data. But do we have the confidence that this generic data is right for our specific application? In addition, many LOPA scenarios contain "one-off" equipment parameters (either initiating event frequency or probability of failure) for which there is no generic data, leaving teams guessing at what value to use. Worse, LOPA targets are getting smaller (i.e., 1e-5 or 1e-6 per yr) which often leaves gaps, requiring decisions to be made regarding capital spending. Sticking with generic data in these cases can leave us feeling that we are being too conservative. On the Operations and Maintenance side of the LOPA equation, we face similar problems when attempting to verify the installed performance of an IPL (Independent Protection Layer). A multitude of assumed parameters (e.g., failure rates, test and inspection intervals, time in bypass, etc.) for which we would like a method to incorporate actual site data into the values used during design. And ideally this method could optimize these parameters for potential cost savings (for example, extending maintenance intervals).

This paper will present a straightforward and easy to use method for feeding operational data back into process safety calculations, using commercial software that is already running on your

computer. The paper will explore how much data is needed to confidently claim a parameter value, starting with an assumed or generic value, and periodically updating that value with small data, as evidence (from testing, maintenance, actual demands, etc.) is collected over time. The authors have been using these methods successfully on real process safety applications for several years now, that were all triggered by difficulties and shortcomings in LOPA. These application case studies will be discussed as well.

1 Introduction

What do we mean by "when LOPA fails?" We've all been there. Participating in a LOPA when a handful of the scenarios being evaluated just don't "fit" well into the method. The reasons are varied. It could that there is no generic data available representative of the device you're evaluating. It could be a "residual gap" that exists (i.e., a gap less than 10) and you're stuck coming up with another independent safeguard. It may be that extremely ambitious (i.e., small) LOPA targets have left large risk gaps. You may be sophisticated enough to know which difficult scenarios to exclude from LOPA altogether (for example, human factors dominated scenarios), but then you're left wondering how to show that you have met the risk targets. You might be considering performing full-blown QRA (quantitative risk analysis) but then you realize that the generic data you would be using inspires no more confidence than what you use for LOPA (it may have the appearance of more precision, i.e., more decimal places, but the reality is you have no idea how well it applies in your application). All of these examples of "fails" (and there are certainly many more) can result in degraded decision-making regarding risk reduction allocation, to prevent major accident hazards. And we never have a blank check that we can write.

2 "Be smart, not conservative."

The Offensive Coordinator for the Cincinnati Bengals football team has a saying he uses with their young star quarterback, Joe Burrow. "Be smart, not conservative." A similar catch phrase is "Do you want to be correct, or conservative?" Risk-based process safety requires us to evaluate the likelihood of a potential hazard. Do we want to be correct or conservative in that evaluation? There are times for each. But this paper addresses being correct. Would you like to be able to make data driven decisions for that (**see Section 3**)? Of course, but largely we have no useful data (other than generic) that could improve our decision making. Or do we? This is where the concept of small data comes in.

3 Decisions, Decisions!

What decisions are we talking about related to LOPA (and Functional Safety in general)? Here are some examples:

1. How to compare your as-assumed to as-operating reliability parameters? This is a *required* safety lifecycle task (per **IEC 61511**).
2. How to make better decisions regarding risk reduction allocation?
3. How to extend your testing/ maintenance intervals for Functional Safety safeguards?

The methods presented in this paper can be used to help answer those questions.

4 The Big Reveal (begin with the end in mind)

Starting with generic (Big) data, we can use statistics combined with field evidence (demands, proof-tests, assessments, audits, etc.) – aka "Small Data" - to update the generic numbers, to demonstrate:

(1) increasing (or not) confidence in the assumed (generic) numbers, and

(2) eventually aid decision making regarding risk reduction allocation, extending test intervals, etc.

We equate Big Data with generic data, because there is a lot of it referenced across multiple sources (e.g., OREDA, SINTEF, RiAC, CCPS, etc.). Small Data is site data. Can we update the generic data with small data? This paper shows how to do this.

5 "An approximate answer to the right question, is better than an exact answer to the wrong question."

This is a quote from John Tukey, the legendary American mathematical statistician. How does his quote apply to us? As an example, let's ask the following "right" question, "how much data do I need to be able to make a claim to some degree of confidence?" We'll attempt to answer this question in **Section 7**, but let's first drill down into the question being asked. The "claim" we are making is on any one of a multitude of parameters we use as process safety practitioners to base decisions on, for example, risk reduction allocation to close gaps, setting maintenance intervals, etc. See **Figure 1** for examples of typical parameters that Functional Safety uses. The next part is, the parameter is claimed to "some degree of confidence," i.e., it's approximate. The parameter value chosen is **not** a point value that is exactly known (as is often assumed by LOPA Teams using generic data tables, e.g., "a risk gap of 1.001 is *significant*!" they'll claim[1]). This is an exact answer to the wrong type of question regarding risk reduction allocation (e.g., "am I on the risk target line?") This point is lost on many LOPA teams. Remember, we are trying to be correct here, not conservative.

In practice, we don't know if generic LOPA data actually represents reality (i.e., the installed and maintained (or not) performance of the equipment). And we won't ever know until we put some work into (i.e., update) those generic numbers. More to come later on that updating process.

[1] LOPA is considered an "order-of-magnitude" method, however, Teams often treat it with way more accuracy than it deserves. This paper can be used to address the issue of closing "residual gaps" (i.e., gap <10), to properly show the uncertainty has been accounted for.

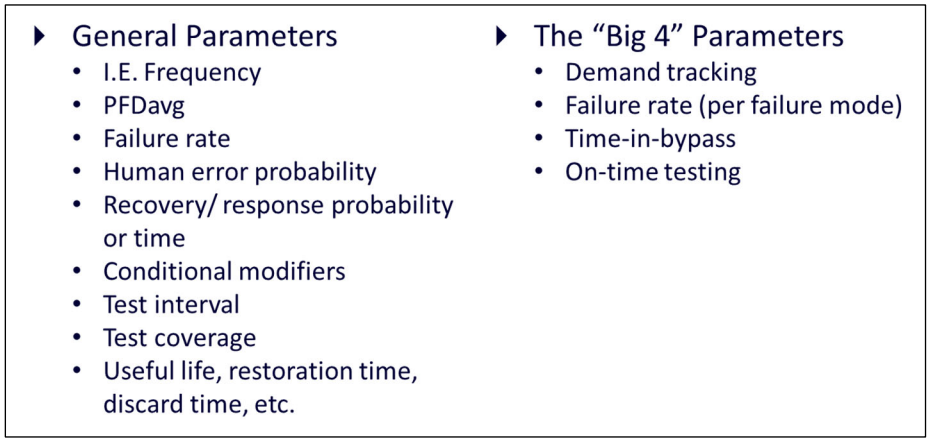

Figure 1. **Typical Process and Functional Safety Parameters**. A parameter describes a real-world manifestation of Nature we are interested in for safety and reliability. Parameters turn reality into a number. Parameters can be described with probability distributions (describing the uncertainty associated with the parameter) and can be updated with site specific evidence based on operating conditions, field testing, maintenance, etc. Accurate parameters lead to better decision making regarding, for example, risk reduction allocation, maintenance intervals, comparing as-designed vs. as-performing metrics for safeguards and IPLs (Independent Protection Layers).

6 Data Science for the Process Safety Practitioner

This paper discusses an application of statistics to Process and Operational Safety. Statistics is an important branch of Data Science (see **Figure 2**). The field of Data Science has immense potential to impact the practice of Process and Operational Safety. We'll discuss three example applications using **Figure 2**, including a discussion of how it is applicable to this paper. The three examples are:

(1) Machine Learning – If you're making "lots and lots" of decisions in a repetitive manner, you want machine learning. What might this look like in operational safety practice? Consider a data driven algorithm that is "listening" for a specific failure mode of a critical piece of equipment. One that if failed could propagate into a major accident hazard. The algorithm provides early detection of the potential unplanned event. The event signature has to be defined, the algorithm has to be trained to detect it, the system must be wired and configured to actually run, and an automatic or human response planned if the algorithm triggers. An excellent real-world application of machine learning to operation's safety is provided in reference **[1]**.

(2) Data Analytics – Following **Figure 2**, if the few decisions you're making are not critical (or you're not making decisions per se, on the data) and the uncertainty is low, you'll use data analytics. In this context, data analytics closely mirrors what is known as "descriptive statistics." If, for example, you're using Process Behavior Charts to trend

metrics **[2]**, you're using simple statistics, but you're not necessarily making inferences[2] beyond those samples (until a particularly bad or good trend develops).

(3) Statistics – This branch of **Figure 2** is applicable to what we are describing in this paper. If the few decisions you are making are important and there is large associated uncertainty, you need statistics. More precisely, inferential statistics, where you collect some data, and need to make judgements that extend beyond that data either into the future or to the part of the population that was not sampled. For example, you collect some data on equipment 'X', and want to extend the results of those statistics to the entire population of 'X'. Or, you collect some data on equipment 'Y', and you want to infer how 'Y' might perform in the future if some parameter associated with 'Y' is changed, for example, its maintenance interval.

The popular and very smart Cassie Kozyrkov, Chief Decision Scientist at Google, likes to say, "Statistics is the science of changing your mind under uncertainty." The "Null" or default position (or action) is based on your prior belief, and is the starting point from which you may change your mind based on the statistics you would perform. If you don't intend to make a decision about something and act on it (or stay with the default position), don't waste your time doing statistics. This is what analytics are for.

An example is, you collect some test data on equipment over time, and you then want to extend the test interval for that equipment based on successful tests, with high enough confidence that the equipment performance won't be degraded. In short, you want to change your mind about the test interval, whether to extend it, or stay with the default interval. You use statistics to show this. Many decisions made in Process and Operational safety would benefit from the use of statistics. And by statistics, I don't mean the kind you learned in High School or College[3]. More on that later.

---

[2] What is an inference? Inferences are all around us. A good example comes from the American football conference championship game played between the Philadelphia Eagles and the San Francisco 49ers on January 29, 2023. Near the end of the 1st quarter, a punt by the Philadelphia Eagles hit the overhead wire causing a very poor punt. Normally the down would be re-played (the Eagles would get another punt attempt) however there was no direct video evidence showing the ball hit the wire, for the game officials to overturn the play. However, based on overwhelming and independent evidence of the player's reactions to the ball hitting the wire (including pointing upward and saying the word "wire"), one could make an inference that the punted ball did, in fact, hit the wire. But there was no video evidence showing with 100% certainty that the ball did actually hit the wire. That is an inference.

[3] There are two distinct kinds of statistics: (1) Frequentist and (2) Bayesian. The two have completely opposite interpretations of statistical concepts (see **Section 7** for an example). You were likely taught "Frequentist" statistics in High School and College. But it is widely acknowledged that Process and Functional Safety should be using Bayesian statistics **[5,6,7]**. And for a fun but brilliant introduction to the Frequentist v. Bayesian "Wars" see the following link: https://www.youtube.com/watch?v=eDMGDhyDxuY "All About that Bayes: Probability, Statistics, and the Quest to Quantify Uncertainty" by Kristin Lennox.

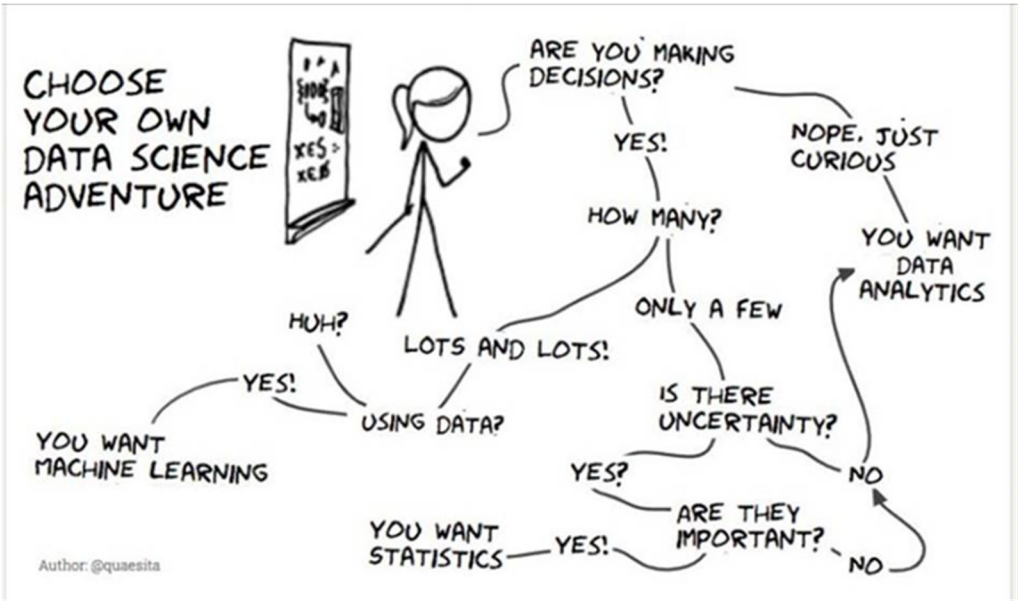

Figure 2. **Data Science Learning Paths[3]**. Each of these learning path outcomes, (1) Machine learning, (2) Statistics, and (3) Data Analytics, has applications for Process Safety. This paper is describing an application at the end branch labelled "You Want Statistics." See text above for discussion.

7 How much data do I need?

This question comes up all the time. The brilliant decision scientist Doug Hubbard likes to say, "You have more data than you think, and you need less of it than you assume **[4]**."

First we set some context. Here we're asking how much data do I need, if I collected it myself in my facility, to have a high level of confidence that my parameter of interest is correct? We're not talking about pooled generic data here. Typical generic data doesn't come from your site, which means using generic data to make inferences about equipment at your site weakens your argument. (However, below, we argue that generic data is part of the starting point for the statistical analysis described by this paper).

Some additional context is that the data is required to be what is called **homogeneous** (i.e., identical devices in the same service, installed at the same site, managed by the same people including maintenance, operations, technical services, capital projects, under the same management of change, and subject to the same economic and human constraints and pressures.) Obviously, you will want to pool together as many homogeneous devices as you can, in effect increasing the sample size[4]. Also, data should be collected by failure mode **[8]** for the like

[4] A result of Frequentist based statistics tells us that as sample size increases, the standard error decreases by the square root on 'n' (the sample size). For example, if my sample size is increased by a factor of 4x, my standard error is cut in half (relative to the expected value of the population statistic). While this is a good thing, it requires a

devices, to support RCM (Reliability Centered Maintenance) decisions. Admittedly, these are very narrow and limiting requirements, but necessary for the math to work, which will support better decision making. It's also justified for several other reasons. One is, many times you have only one kind of homogeneous device, or at best a handful of homogeneous devices. That's the reality. Second is, extending the statistics from identical to "similar" devices is, one, extremely subjective (i.e., when do devices stop being similar enough?), and two, the math gets exponentially more complicated. And there is currently no COTS (commercial-off- the-shelf) software available fit for purpose to do these extended calculations (for non-homogeneous data). In addition, the subjectivity involved in collecting like data (including by failure mode) may out-weigh "the math" no matter how sophisticated said math gets. For this reason, we argue the simpler analysis presented in this paper based on what we call "small homogeneous data" is the correct place to start, and at a minimum describes the foundation that more complex analysis would be built on, using future fit for purpose software.

Below we rank, in order going from "big data" to "small data" an attempt to answer the question "How much data do I need?"

1. As a "rule-of-thumb" (read: conservative) you need an order-of-magnitude more data than the claim you are trying to make **[9]**. For example, say you are trying to claim a 1/100 year initiating event frequency for a particular device. You would need 1,000 years-worth of essentially failure free data. For example, 100 homogeneous devices installed for 10 years-worth of collected data on the *failure effect of interest,* with very few failure modes that can't be written off (i.e., corrected) as "systematic[5]." The authors have heard of very large sites doing this exact thing. A simple MLE[6] (maximum likelihood estimator) ratio can be used (i.e., total number of failures divided total number of years) to produce the best estimate of the failure rate parameter given your data.

   This same example can be applied to probability of failure, based on field testing evidence as well. To make a claim of a 1 in 10 probability of failure, you would need 100 demands worth of failure-free homogeneous data. The MLE would take the form of total failures on demand divided by total number of demands.

2. If you don't work in a large facility such that you have access to a significant pool of homogeneous data as described in Case 1 above, you need to sharpen-the-pencil a bit. In this case we will invoke a probability distribution, (for example, the Poisson distribution

---

sampling distribution of the statistic you're interested in to be developed (the mean, variance, etc.). So for small data applications, it is limited, in making valid inferences. Bayes is a more direct and realistic method to incorporate larger sample sizes without the need to invoke hypothetical sampling distributions.

[5] "Systematic" isn't a recognized failure mode in RCM (Reliability Centered Maintenance) parlance. However, the term is so embedded at this point in Functional Safety, there is no undoing it. The good news is, the Bayesian statistics presented in this paper works with so-called "systematic" failures or faults, whether included (or not) in the data collection effort. It is up to the judgement of the practitioner to include them (or not).

[6] The MLE is a frequentist-based parameter estimator, first developed by R.A. Fisher. It is the parameter that "maximizes" the likelihood of observing your data. The main limitation of the MLE is that confidence limits do not fall out of it naturally. One must invoke an artificial distribution – such as the Chi-square – to calculate confidence limits. The reason is because the likelihood part of the MLE is not a valid probability distribution.

for failure rate or the Binomial distribution for failure on demand) to lessen the requirement for data. The assumption we make is that the probability distribution we select represents the underlying distribution found in Nature (a common assumption that is taken). An example of the reasoning used to determine how much data is needed in this case is found in **[10]**. The results can be summarized as follows (see reference **[10]**): You need three periods worth of homogenous data to "prove" a number to 90% confidence[7]. For example, if I'm trying to prove that my parameter is "1 in 10", I need 30 years or 30 demands of failure free data. For "1 in 100" you would need 300 years, etc. This is better than Case 1 above, but for smaller sites or where you have very unique installation(s), still very constraining.

3. At this point we are ready to transition from **bigger data** to **smaller data** requirements. We pause here to acknowledge that this paper will not be able to discuss relevant statistical concepts such as sampling distributions, the Central Limit Theorem, etc. which are cornerstones of big data statistics. There's an old caveat in statistics that goes as follows: "it probably doesn't mean what you think it means." We'll leave you with an example and move on. The point here is that many practitioners using even basic statistical concepts don't really understand what they mean and what the assumptions they are based on mean. Many people assume a 90% confidence interval means "there is a 90% chance that the true parameter falls within that interval." Not necessarily. It depends on what kind of statistics you are using[8]. If you don't know what kind of statistics you are using, there is good chance that a 90% confidence interval will actually mean, "If I took 1,000 samples (or, more in the limit), 90% of the confidence intervals that I would calculate from each of those 1,000 samples, would contain the true parameter." This is a completely opposite meaning (to the former), and more practically, no one ever takes 1,000 samples. The small data method we present below does in fact use the former interpretation (i.e., "the chance that the true parameter is within the interval").

4. The smaller data method presented by SINTEF **[6]**, requires at least as much total equipment time (they refer to this as an "observation period") as what you're comparing to in order to give confidence to the updated failure rate or recommended (new) test interval. We'll discuss below how they are able to lessen the data requirement.

5. And finally to small data (used in the example below in **Section 13**), requiring a sample size of only n=1 (i.e., one proof test, one failure, one demand, etc.) to begin to make updates on. How do we get away with this?!

[7] We are able to invoke the concept of a confidence level here because we are using valid probability distributions.
[8] We associate Frequentist (classical) statistics with "big data," and Bayesian statistics with "small data." Frequentist statistics works in the "long run" (the theoretical limit being infinite). That is, after many trials, draws, demands, repeats, or experiments have been performed, I can make my inference. Bayesian statistics works as soon as the data starts to trickle in.

How does small data work? It works by using something known as a "Prior." The Prior is what we bring to the table when we initially design an IPL (Independent Protection Layer)[9]. It is our beliefs, opinions, attitudes, biases, and experiences (in short, everything that makes us human) about how we think that IPL will perform. But it is also data on that IPL. Generic industry data, corporate pooled data, one-off data, certified data, best-guestimate data. All of this powerful evidence is rolled into the Prior that becomes our starting point parameter from the design phase. From there, in the Operations and Maintenance phase of the IPL lifecycle, we test, maintain, and audit. This trickle of real-world performance is then used to *update* the Prior to produce what is known as a "Posterior" (post-data). Initially, the goal is to build a case of confidence (or not) in the Prior. We are not talking about (at least initially) changing the numbers. We are arguing to first build a case (claim) in the confidence of the number (the Prior). See **Section 10** for more details about how we do this.

8 Aren't Decisions based on averages good enough?

Decisions based on single-point averages are ubiquitous in business and industry, and process safety is no exception. But critical (important) decisions that have associated uncertainty should never be based on single-point values. Why?

As Sam Savage likes to say, "Decisions based on averages, are wrong on average." This is known as the "flaw of averages" **[11]**.

Think about some hypotheticals. The drunkard that plans to walk down the centerline of a busy road (the average) to stay alive, will be dead, on average. Then there's the statistician that drowned in a stream of an average depth of 2 feet. You get the picture.

Think about this. The entire notion of Functional Safety is built around the concept of an "average." That means that our decisions based on those "averages", for example, setting maintenance intervals, allocating risk reduction, etc. will be wrong on average!

Simple averages may not reflect the underlying distribution in Nature that we are trying to make an inference on. The other problem is using an average value doesn't inspire much confidence. An average is approximately a 50% confidence level (depending again on the underlying distribution). But 50% is pretty low confidence. Sure, you could artificially improve that 50% value to a higher confidence – say, at 90% - but then you're sacrificing by being conservative. Remember, the point of this paper is about being correct, not conservative.

How do we be more correct? First, we don't have to throw the proverbial baby out with the bath water. We retain the generic 50% point-value as the *starting point* of our effort (this is the Prior

[9] It is important to note that Frequentist statistics does not allow the use of a Prior. To be a Frequentist means you must check your beliefs and opinions at the door (and supposedly let the data "speak" for itself – no matter how long it takes!). The problem is, the data can't speak for itself (only humans can speak), and if data could speak, it couldn't tell the whole story anyway!

as discussed in **Section 7**). From there we gather actual field evidence (data) to update that generic 50% point-value to show increasing confidence (or not) in that value.

9 Subjective v. Objective -what's the difference?

Subjectivity is everywhere in process and operational safety. It shows up in every important decision we have to make that is under uncertainty, including from ranking hazards to choosing how much credit to give a particular safeguard. We have seen some subjective choices already in this paper that we will face, such as, deciding what to include as homogeneous data, classifying failure modes (causes) for field data, determining what distribution matches your underlying process, and even just deciding if some statistics are needed to help improve decision making. And we have seen that being subjective is a big part of what makes us human (i.e., our beliefs, opinions, past experiences, etc.). Being subjective isn't a bad thing, it's what makes us human!

Still, there can be a lot of resistance from purely objective and rational engineers to using, for example, a subjective Prior (i.e., based purely on beliefs). Even though as we've discussed our Prior will be largely based on generic data, however, there is nothing preventing us from modifying the generic value based solely on subjective belief. It's up to the person (or team) setting up the Prior. But even this hint of subjectivity is enough to close their minds to the method. To that end, let's seek a middle ground between objective and subjective.

First, objective doesn't mean "quantitative[10]." A better definition for objective is to "see beyond our personal biases and prejudices **[12]**." But since we are human, we can never make perfectly objective decisions and choices. But we can be aware of them (our biases). And that is the power of subjective probability. We're not asking you to ignore being human, we are asking you to be aware of it!

10 Updating Rule – To be less and less wrong[11]

We've discussed starting with a Prior for our parameter of interest, and updating our Prior based on actual field evidence. What does this look like? Before we get into that we briefly discuss the tools used to perform the updates. The Microsoft Excel™ spreadsheet application has improved its statistical capability especially within the past 5 years or so. The number of included probability distributions, along with the ability to do, for example, Monte Carlo

[10] There exists an entire literature discussing how the mathematics of subjective probability (based solely on beliefs) is as valid as the mathematics of probability based on hard data **[13,14]**. If you're still not convinced subjective probability is a thing? Check out DraftKings or BetMGM websites as examples.

[11] The complete quote is "to err and err and err again, but less and less and less," from "The Road to Wisdom" by Danish mathematician Piet Hein. We are not trying to make a claim of 100% certainty (that is impossible with any important decision under uncertainty) however, with each update we run incorporating site data, the goal is to become less and less wrong.

simulation, allows one to perform Bayesian updating on the Prior directly in Excel, to produce Posterior distributions. There are some limitations encountered by staying in Excel, versus coding the updating process using, for example, the R programming language. However, we want to start simple, to encourage the uptake of these concepts by Process and Functional Safety practitioners. The other reason is, no matter how sophisticated the math becomes, the decision-making process as we've been discussing, can never "boil down" to a number (or even a distribution of numbers). The subjectivity involved is too great, for example, even the data collection process touched on in **Section 7** is dominated by subjectivity. We keep the tools and methods simple for this reason also – yes we advocate better decision making, but if the tools are so complex that no one will use them, it defeats the purpose.

Figures 3, 4, 5, and 6 [15] describe the updating process. Refer to the caption in each figure for discussion.

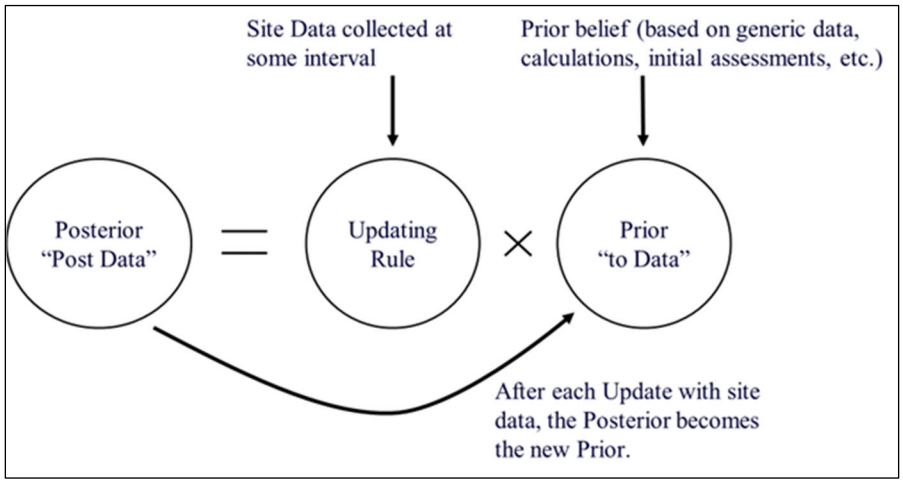

Figure 3. The updating process. The Prior is established first (before data is collected) and combined with an updating rule to produce the Posterior distribution. The Posterior provides an inference on the future interval between the new Prior and the next update.

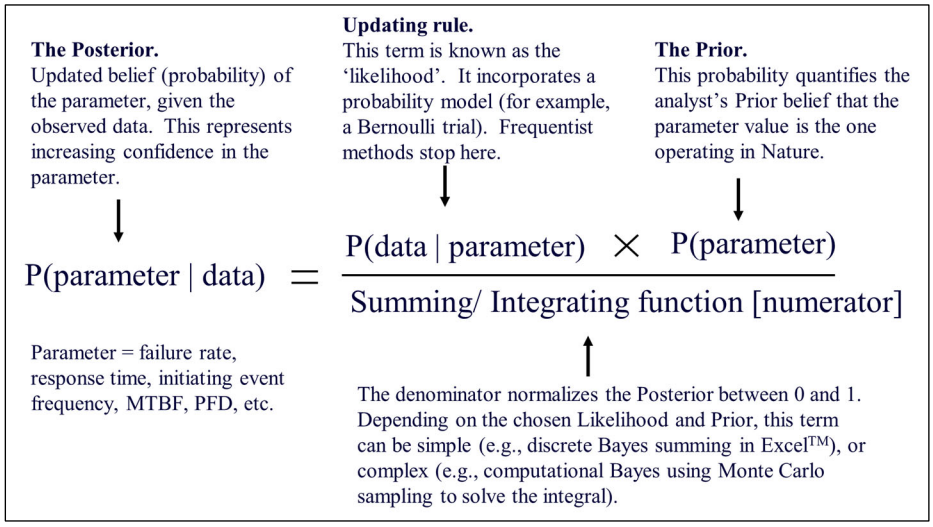

Figure 4. This is known as Bayes' Rule. Bayes' Rule is a theorem of conditional probability, it is exact. However, when used to perform inferences on parameters (as we are doing in this paper), Bayes' Rule becomes "Bayesian", i.e., it is no longer "exact." This form of Bayes' Rule can easily be created in standard Excel™ using its extensive list of probability distributions that are included.

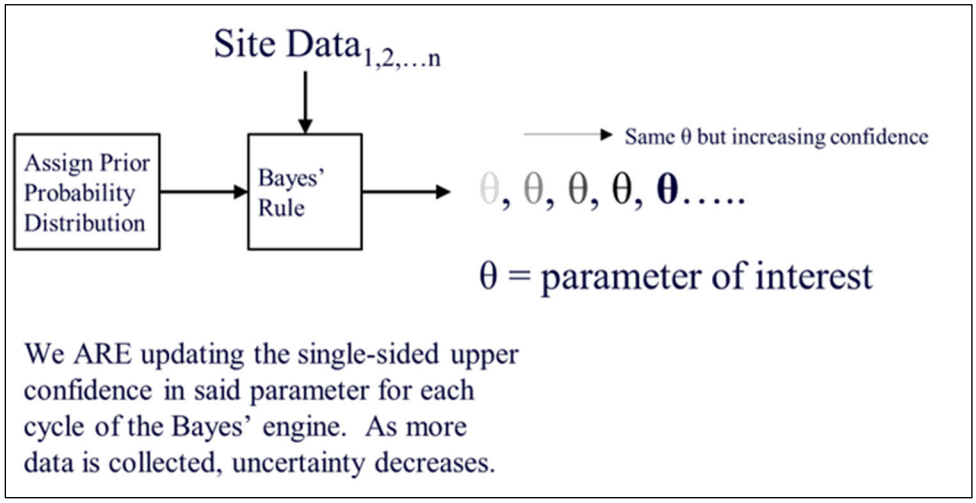

Figure 5. Bayesian updating brings the as-designed (Prior) parameter into focus as data is collected over time and updates are made. An update can be made whenever new data is collected. A direct result of this is you are able to evaluate how well your assumed parameters from the design phase are actually performing in the field, as required by **ANSI/ ISA 61511-2018 clause 5.2.5.3**.

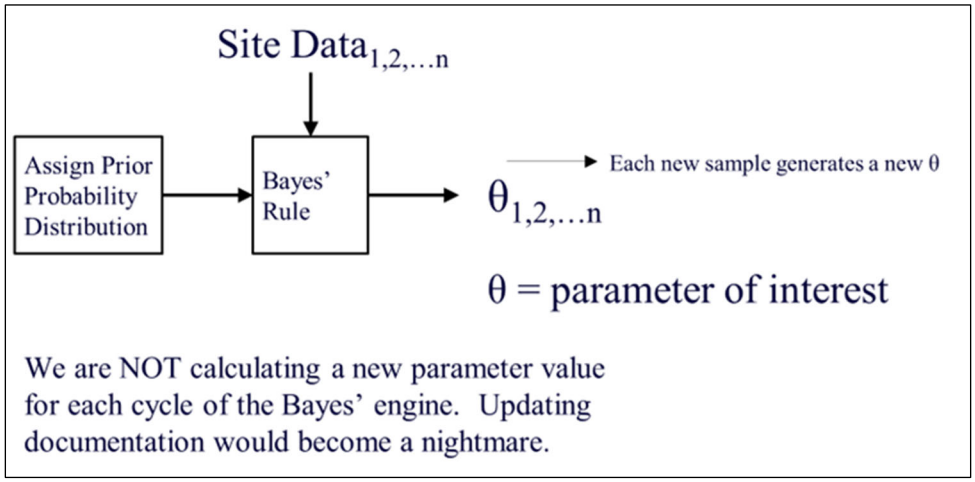

Figure 6. Over time, if a claim of high confidence can be made on the parameter, it is possible then to adjust the design value of the parameter (i.e., redo the calcs and update documentation) to modify, for example, initiating event frequency, proof-test interval, scheduled restoration or discard intervals, etc. all of which would have potential cost savings without reducing safety.

11 Bayesian Schmayesian

The intellectual giant, Nassim Nicholas Taleb **(Figure 7)**, whose books have been called among the most important to be written since WWII, uses the phrase "Bayesian Schmayesian" in his Technical Incerto **[16]**. The technical Incerto is a wide ranging treatise describing the math behind such concepts as Fat Tails, Black Swans, anti-fragility, resilience, and being fooled by randomness (in general). Any practitioner of Bayesian statistics would be wise to take notice of that tongue-in-cheek jab to see what he is talking about.

Taleb is poking fun at attempts to use something "vaguely" Bayesian (called "Schmayesian") to solve problems about the unknowns under thick tails. Luckily that is NOT what this paper is about. However, Taleb goes on to write, "One of course can use Bayesian methods (under adequate priors) for the estimation of parameters…" **[16]**. Full stop. That is fundamentally what we are describing in the paper. And who develops the best possible Priors if not Functional Safety people? If you've ever been involved with (even peripherally) the SIS Engineering effort on a Capital Project, you know first-hand the effort and brain power that goes into selecting model parameters! These are the Priors! And they are very adequate!

The fact is Process and especially Functional Safety is missing the proverbial boat on implementation of Bayesian inference. The majority of effort comes from developing good Priors, which Functional Safety is already doing. Many of current references, all related to Functional Safety, support the use of Priors as the starting point for incorporating Operational and Maintenance data into the calculations **[5,6,7]**.

Figure 7. **Nassim Nicholas Taleb**, wrote the 5-book series Incerto (Latin for "Uncertainty"), discussing (in general) the ways in which humans get fooled by randomness.

12 Small Data – putting it all together

Before jumping into some real-world examples in **Sections 13-15**, let's summarize what small data means. Small data, as discussed in this paper, works on the premise of a robust Prior, the kind that already comes out of detailed SIS Engineering efforts, or the ubiquitous data tables used for LOPA. After data is collected (testing, maintenance, audits, etc.) the Prior is then combined with an updating rule to produce a distribution of the parameter and its uncertainty reflecting site specific influences. An immediate benefit is that compliance can be demonstrated (to **Clause 5.2.5.3 of ANSI/ ISA 61511-2018**, for example) that you are assuring the parameters assumed during design phase are in fact reasonable for the site operating conditions. As additional data is collected over time, potential cost savings can be realized by, for example, extending proof-test intervals, extending useful life, lowering initiating event frequency, etc. without compromising safety. The tools for performing this analysis are accessible to the practitioner and we argue that because of the large amount of subjectivity involved in these decisions, the use of more sophisticated mathematical techniques is not justified, at least initially.

The alternative, known in this paper as big data, is to wait literally for decades until enough data has been collected to attempt the same analysis. In practice, where significant changes in process operations can occur year to year (to process, hardware, people and budgets for example), any benefit gained by having larger data is completely eliminated by the gradual introduction of non-homogeneity into the data itself over time (i.e., the inferences you would be making on that data are greatly weakened).

13 Example #1 - Justifying deviations from corporate standard IPL credit values - DCS 1 in 100

Our first example comes from a case study "A Tale of Two BPCS Credits, A Bayesian Case Study" **[17]**. In this case study the LOPA team wanted to challenge the corporate standard initiating event cause frequency for a robust Basic Process Control System (BPCS). Typically initiating cause frequency values are mandated by client standards. These static numbers are normally not an issue as it keeps LOPA scenarios consistent. The drawback of a static number is it can be restrictive, especially when LOPA severities targets run extremely low (i.e., many zeros after the decimal).

13.1 What was the problem?

For this example, the team had reached a crossroads. The LOPA had a high severity scenario and had applied all existing Independent Protect Layers (IPLs) but still had a gap of one risk reduction credit. The options to close the gap were to install a new Safety Instrumented System (SIS), or "sharpen the pencil" and take a closer look at the existing IPL credits and the initiating cause frequency (a BPCS initiator). Typically BPCS initiating causes frequencies are limited to one failure in 10 years, but could the team justify 1 failure in 100 years?

The LOPA scenario delt with a batch reactor which was sequenced by the BPCS. There was a failure mode where if a particular block valve were to open during a certain phase of reaction, a high severity thermal decomposition runaway reaction could potentially occur. The mitigation target was going to be difficult to achieve while the amount of protections available were not enough, yet the team thought their current safety design was adequate (a reasonable subjective judgement as a starting point). The problem was their safety system was also the initiating cause (the same batch processor). The team was limited by their corporate standard which only allowed them to apply an initiating cause frequency of 1 in 10 years while not being able to apply the same BPCS valves as an IPL. This limitation is typical in industry due LOPA requiring IPLs to be Independent (per their definition).

The team believed that their BPCS was a highly robust, dangerous failure resistant system. They were sure that nearly all dangerous failure modes had been eliminated by design. One such example was the initiating event valves were motor-operated valves which were "fail last" (the safe state). The dangerous mode could not occur as long as the valves held their position and motive force was removed by the BPCS sequence. Furthermore, ever since a redesign from 25 years earlier, there had never been a failure of the system.

The system sounded good; however, as any process safety practitioner knows, taking a 1 in 100 initiating event frequency for BPCS systems is not typically allowed from industry or corporate standards. This client was no different. The corporate LOPA standard only allowed a 1 in 10 initiating event frequency for BPCS. The other issue was **ANSI/ISA 61511-2018 clause 8.2.2** which stated "BPCS as an initiating source shall not be assumed to be <10-5 per hr. (~1 in 10yr)." Many LOPA teams would recommend a safety instrumented system when faced with such restrictions, however this client did not want to incur the increased complexity and cost of ownership that comes with one. Also, there was concern that adding another SIS valve in the mix could actually result in new hazards (certain batch modes would become more dangerous with

additional valves). Given the restrictions, how could the client hope to achieve a 1 in 100 initiating event frequency for their BPCS interlock?

The client did have ideas in mind to satisfy the requirements. First the corporate LOPA standard limitation. The client had an approved generic data deviation policy and procedure (for all generic data used in their LOPA's). This deviation would be documented in a form which the team would need to fill out, that also needed to be accompanied by supporting evidence. The project scope of work would serve as this evidence.

Other limitations are the various standards, a major one including **ANSI/ISA 61511-2018** limitation, as well as CCPS guidance that a BPCS IPL shall not be over 1 in 10 initiating event frequency. They also intended to stick with their design and consider the system risk reduction to be As Low as Reasonably Practical so long as they performed their due diligence in justifying their position. Further examination of the standards, problems, and solutions can be found in the whitepaper reference **[17]**. The conclusion from the paper is if a deviation is needed from any standard, it is not outlawed outright, but there had better be a justification documented. One of the proposed methods was a Bayesian analysis with a confidence assessment. This is a small data method.

13.2 Why was this small Data?

In order to justify this system was operating at a 1 in 100 initiating cause frequency, the team would need to provide a justification. The team was confident in their system; however the problem was they did not have a lot of data to back up their subjective beliefs. If the team was limited to big data methods, such as a Frequentist interpretation, up to a 1,000 years' worth of data would be required to justify the frequency of 1 in 100 years (refer back to **Section 7** for this discussion). Unfortunately, there was no way to get 1,000 years' worth of data as this is the only such system in existence known to the client. The client only had 32 years' worth of operation history. Rather than "big data," the team was dealing with "small data."

How would the team justify their 1 in 100 failure rate with only 32 years' worth of data? The answer was to use the data they had, their expert opinion, and a Bayesian analysis with a confidence assessment. The data used was the actual operating history (32 years of history), and expert/ team opinion in the form of a Failure Modes and Effects Analysis (FMEA) where the team analyzed all known failure modes and effects of the system. These numbers were used in the analysis to generate a confidence distribution which answered the question "what is the confidence our system is operating at a frequency of 1 in 100 years?" Detailed discussion of the method is explained in the paper reference **[17]**.

13.3 What were the results of small Data?

The team developed a confidence level in their system being 1 in 100. The confidence was good to start with, 0 failures in 32 years, with a favorable FMEA to back up the numbers; the team had a >80% confidence their system was operating at least as good as 1 in 100 (see **Figure 8**). All seemed good, but unfortunately for the team the system *had a failure during the middle of the*

study! This unfortunate event needed to be factored into the analysis. The confidence assessment was updated (using the technique shown in **Figures 3 and 4**) and it was discovered the confidence of 1 in 100 frequency had dropped to <35% (see **Figure 9**). This was deemed as unacceptable from a risk standpoint[12].

The team unfortunately decided that their 1 in 100 frequency should not be used in their LOPA. In this case the standard, even though it is conservative, was proven to be "correct… for now." Small data does have a drawback that it can be contradicted by failures fairly easily, but the good thing is it will happen quickly. This is the right type of conservative. The system frequency could still truly be 1 in 100, maybe the first failure in 32 years was unlucky. If the system operated for a long time without a failure it would be a good bet that the system could be operating at the 1 in 100 frequency, but now the conservativeness has a basis, rather than just being blindly trusted.

The result of the recent failure gave the team immediate feedback that action needed to occur on their system. Based upon the unfavorable result, the team was able to generate different recommendations based on knowledge gleaned from the study, including performing a full incident investigation. The root cause analysis ended up determining the system failed at the valve level, namely the limit switches gave false indication on a critical step. Based on this finding, and the knowledge that the BPCS should not be credited with a frequency of 1 in 100, the team decided to revisit the SIS idea.

The team ultimately decided they could close their LOPA gap by adding in an SIS controlled limit switch permissive between the hazardous batch steps. This also avoided introducing a new hazard from unwanted new valves. These decisions were supported with the data, whereas if not for the data and the study and the team just went with their feelings, they could have been operating with a significant safety gap. Rather than thinking their system was good and did not need any improvement, the data supported the safer decisions of taking the more costly approach of installing a new SIS.

[12] The question often comes up in this kind of analysis, what confidence level is acceptable/ unacceptable? We recommend as a guideline, that 50% confidence (loosely representing the average) is the lowest confidence level you would want to claim, without dropping below this (if you did have a parameter calculated below 50% confidence this would be proverbial “red flag” that you may be under-protected). Notably, 70% confidence is recommended in **ANSI/ISA 61511-2018**. Traditional QRA (Quantitative Risk Analysis) uses 90% to 95% confidence levels. That being said, more important than the static number, is how you are trending over time (good or bad). And as noted previously, it is possible to artificially inflate the confidence by choosing a more (generic) conservative number, but you actually haven’t learned anything in the process, and at best you’re just being more conservative (versus being correct).

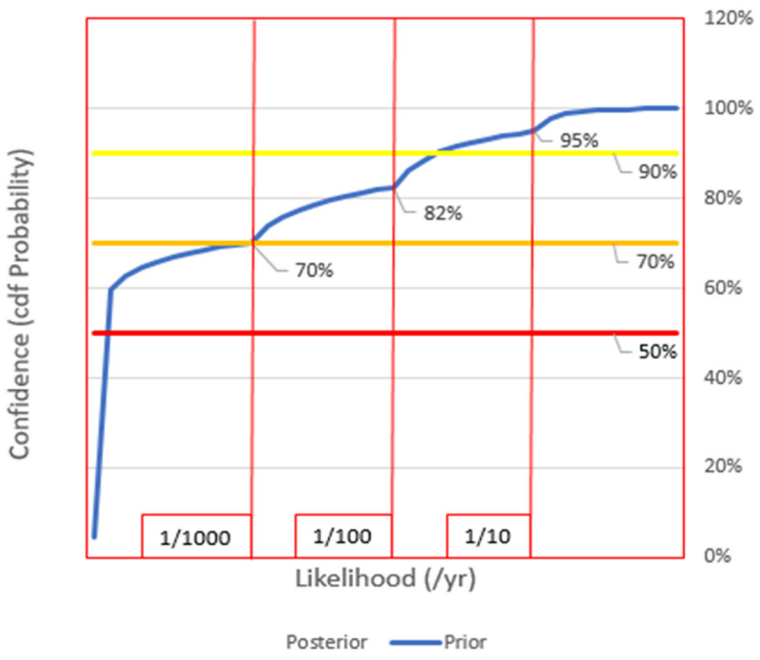

Figure 8. **Prior distribution for Example #1**. This distribution was developed by the team for a specific BPCS function where it was desired to take a 1 in 100 initiating event frequency in LOPA. To develop the Prior, the team used site historical records as well as an FMEA performed for the BPCS function (that considered potential human impact as well). The plot shows that the Prior confidence that the system was operating at 1 in 100 is 82%. This was in line with what was seeded, i.e., 95% confidence for 1 in 10, and 70% confidence for the FMEA result (1064 MTBF). Some additional features of the plot are (1) the c.d.f. is used (cumulative distribution function) because it allows one to read the probability (confidence level) directly from the vertical axis (versus a p.d.f density function which require one to integrate to get the same information) and (2) using a c.d.f. gives the confidence level as a "single-sided upper" confidence (i.e., the probability the parameter is at this value *or less*), which for most parameters used in process safety, e.g., failure rate, PFD, response times, maintenance intervals, etc. the desire is to be confident you're below some upper limit.

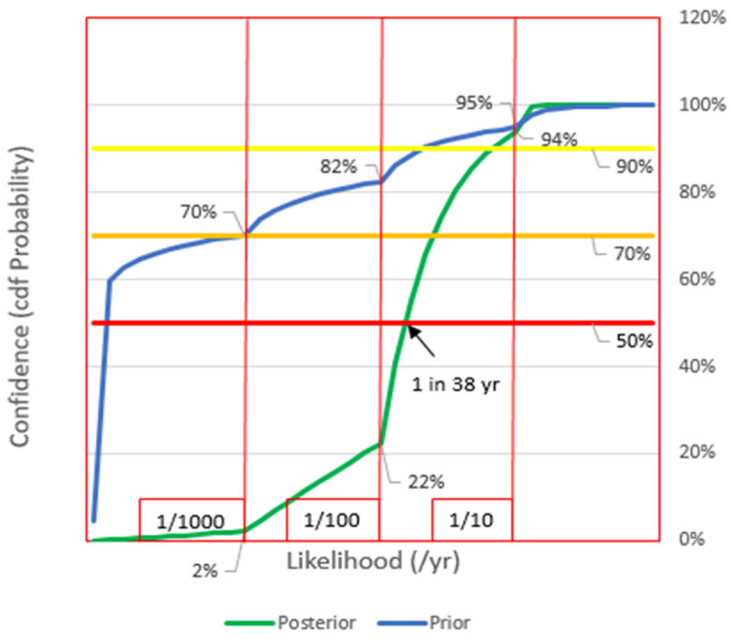

Figure 9. **Updated Prior to produce the Posterior based on new field evidence**. A failure of the BPCS function occurred that required updating the Prior with this new data. You can see that the single-sided upper confidence level took a big hit. In general, when trying to make a claim on a small number (the kind we encounter in Process and Functional Safety) just one failure will cause the statistical confidence to nose-dive. See the text for a discussion of how to handle this.

14 Example #2 - Instrument Failure Rate Data and Prior Use Justification (Failure Rate Data)

Our next example explores seemingly big data, but it is actually small data in disguise. This example deals with plant-based instrument failure rate data. This data is directly used in the calculation of a SIF's risk reduction performance. While this is not exactly a "failure" of the LOPA method, these numbers do directly influence the over optimistic trust we sometimes have in the LOPA method's ability to take us to the negligible risk zone. Many times a team places extremely considerable risk gap closure targets on SIFs. Many time these numbers are completely unrealistic. For example, when is the last time you saw a SIL 3 system that was actually operating with less than 1 failure in 1,000 - 10,000 years (or demands)? How would you ever hope to have any sort of confidence in that number, or prove it? If you are ever late on testing, is the system still SIL 3?

Why do these considerable risk reduction target SIFs get recommended? What if the team had no other option? The total target mitigated event likelihood was simply too low, and the team applied everything else they could think of, but they were left with a seemingly insurmountable gap to close. The LOPA team would likely take the easy out and apply a SIF and move on with life.

Another unfortunate reality of high integrity target SIFs is that chances are also high that the process is unique, and Safety Certified devices likely aren't available or rated for the particular service. Some form of analysis should be undertaken on the data used to perform these SIL calculations in order to better inform the LOPA team of what is really possible so everyone can make better decisions.

14.1 Why is it small Data?

All facilities that follow the **ANSI/ISA 61511-2018** lifecycle are supposed to be proving their assumed failure rate numbers with actual field data, yet hardly anyone does. Some will have an industry generic database, some a corporate database, and the most advanced sites have their own site database. This would all be based on big data, but all of these miss the mark! Per **ANSI/ISA 61511-2018**, reliability data SHALL BE based on field feedback and also uncertainties assessed per the operating environment (see for example **Clause 11.9.3 and 4**). Since most high integrity SIFs operate in their own unique environment, the only data to pull information from is a small handful of devices, or worst case just the single SIF's information. This is small homogeneous data.

The abundance of various data sources can be mis-leading as potential inputs that can be used for this data, including data historian outputs (DCS and SIS), Preventative Maintenance reports, proof test reports, and work orders. This might seem like a lot of data, but keep in mind the relevant data is a small subset of all these larger data sources, and it is very subjective when deciding what to leave in and what to leave out. If your device is a differential pressure transmitter in an oxygen mixing system, you might only have a few other similar transmitters with which to compare it to that are also in the oxygen service. And you should not pull data from other operating facilities because your plant is the only plant with your procedures and particular maintenance staff (refer back to **Section 7** for additional discussion of **homogeneous data** requirements). In other words applying the data from industry might inflate or deflate your systematic biases, only your own data will factor in the systematic bias. It is possible that the only data available is from the device itself. It could even be possible that there is no data (such as could be the case for a brand new “green field” installation).

One might ask how many samples would be required to form a confidence assessment on a failure rate number. Interestingly with a Bayesian engine, the minimum required samples are actually zero! As long as a Prior distribution can be generated (based on expert opinion often using generic data sources), there can be a confidence assessment. Qualitative data that can be used in the analysis includes interviews with personnel with repeatable check lists, or expert review of the instrument installation compared to the vendor literature, or even feedback from peers at other facilities. This feedback can be used to “hedge your bets” when it comes to generating the Prior distribution (see more details about Prior distributions at paper reference **[17]**). That said, it would be recommended to have some amount of (typically generic) quantitative data for generating the Prior distribution. If the amount of quantitative data is low, the best thing to rely on is more qualitative data. With a Bayesian engine you are not limited to quantitative vs qualitative.

All of this is not to say the data used in a failure rate assessment has to be small data. The “best fit” qualitative data used for this form of analysis would primarily be specific to the instrument

(i.e. small data), however the information could come from the entire plant. There should be a tweaking of the confidence assessment when using plant wide data. For example if you know your installation is more hazardous than the typical installation in the plant, then your confidence in the plant wide number should be lower.

14.2 What is the Desired Result of the analysis?

The desired results of all of this analysis would primarily be a better failure rate number for the instrumentation, but greater than that is the achieved SIF performance with a confidence in that number. This assists the team in knowing they have work to do, either in taking some burden off of the SIF and making the target smaller, or potentially working to reduce or eliminate the hazard altogether. LOPA recommends at least an (approximately) 50% confidence in IPLs. **IEC 61511** requires a 70% confidence in any failure rate numbers used. If the failure rate with confidence assessment is using any numbers lower that those targets then the actual gap closure of the LOPA scenario should be suspect. If the team has done everything that they can, the big decision that must be made is “is it time to eliminate the hazard by process redesign?” This is the safest decision that can be made, but you would never know it needs to be made if you do not analyze the confidence in the numbers.

Ancillary outputs from the analysis also include generic plant failure rate numbers, either per process, per unit, per line, or even per device. The failure rate data and confidence assessment shows compliance with the **IEC 61511** standard (used in case of audit, both are required by the standard). The updated failure rate info can be applied to SIL Calcs (either new or revised). This data feeds back to the LOPA to confirm risk gaps have been closed.

Another useful output would be device “Bad Actors” (including spurious trips, or dormant failures). This would come from reviewing the plant historian and examining each device for comparable failure rate numbers. The Bad Actor assessment could identify instruments that need immediate action and tell the team exactly where to focus their efforts. The data can answer the questions such as: Is the device suited for the process? Is maintenance doing a bad job? Does the device need tested more frequently? Is this device a lemon? Why does the unit keep going offline?

A final note is this form of analysis provides useful feedback. The data could tell an owner-operator if their systematic errors are greater than, or less than industry generic. If they have installed a SIF device whose failure rate is worse than industry average, or their confidence is lower than average, then there is something going on at site leading to a higher systematic contribution. This would lead to another set of decisions which hopefully lead to an improved site safety culture.

15 Other uses of small data in support of LOPA

A few other honorable mentions where "large data" is not the answer for assisting in decision making when LOPA falls short.

15.1 Multiple human credits in LOPA

What do we do when a LOPA tried to stack multiple human credits on a high severity scenario? There likely were no other options, or the failure was due to a manual procedure that cannot be fully automated (for example, a loading operation involving a highly hazardous process to a rail car). The small data would come from qualitative plant experience, events per total operational history, and any failure data if it exists. All of the data is likely less than 100 years' worth on a scenario with a target of one failure in 100,000 years. The methods discussed in this paper can be used to show risk targets are being met.

15.2 Changing LOPA targets or severities during PHA revalidation or redo cycles

What do we do if LOPA teams keep changing a scenario's risk ranking each Revalidation cycle (or the Corporate risk matrix has changed) especially following a large investment in risk reduction allocation? Perhaps the scenario is not well understood. What if we examine other similar facilities or do some research of other similar facilities accident reports from the Chemical Safety Board. This qualitative analysis could shape a confidence assessment in a Prior, which could then be used to set the most likely severity and likelihood of the hazard (to some confidence level). We could then see how the confidence level has changed at the new risk ranking or LOPA target. We should not just blindly take the worst credible outcome in the name of being conservative. We need to analyze what makes our plant different, and account for these differences. Once the numbers are justified then the corporate standard should specify the values to use for future revalidation teams to utilize. As always document the justification and re-examine it periodically.

15.3 Using Two-BPCS credits in LOPA

Can we take two BPCS IPL credits in one LOPA scenario? This is a similar question to our first example discussed in **Section 13**. Typically the standards say you cannot, but what if our plant experience says the standard is being too conservative and our plant DCS is being operated and maintained better than average? Let us review our data and make a confidence assessment to justify our decision.

15.4 Independence assumption in LOPA

What do we do if our IPLs are not independent? A lot of times we might need to jump to a qualitative risk assessment and just abandon LOPA, but is this always necessary? What if we have data and expert opinion that tells us we are okay, or this is not a typical installation? Let us use that data (qualitative or quantitative), make a confidence assessment, document our justification, and move forward without having to go through "paralysis by analysis."

16 Conclusion

This paper has made the case for doing what is not being done yet in Process and Functional Safety. Incorporating site specific data back into the generic parameter values we use in LOPA, SIS Engineering, and QRA, etc. In a more general case, this supports the effort of closing the safety lifecycle by feeding back data and other evidence into the as-designed parameters to ensure these assumed parameters and metrics are in fact valid. We've also shown the method to do all this.

We've discussed that although generic data can be very trustworthy (especially as a starting point for the Prior), it is not appropriate for making inference-based decisions under uncertainty, for your specific applications, regarding risk reduction allocation, maintenance intervals, and knowing how well you are doing compared to what you've assumed during design. In short, you're not going to be using the best inputs to base decisions on, if you don't account for site specific factors in your evaluation.

We recommend to start by "test-driving" these methods on a few difficult decisions you're facing regarding LOPA. We've kept the tools required in our examples simple Excel™-based calculations. There is a lot of traction that can be made without getting more complex with the math.

17 References

[1] "Data-driven approach for labelling process plant event data." International Journal of Prognostics and Health Management, ISSN2153-2648, 2022 000. Corrˆea1, D., Polpo, A., Small, M., Srikanth, S., Hollins, K., and Hodkiewicz, M.

[2] "Measure of Success," Graban, Mark. Constancy Inc., 2019.

[3] Kozyrkov, Cassie, Chief Decision Scientist, https://www.linkedin.com/in/kozyrkov/

[4] Hubbard, Doug, et. al., "How to Measure Anything in Cybersecurity Risk," Wiley, 2016.

[5] "Probabilistic Risk Assessment Procedures Guide for Offshore Applications (Draft)" JSC-BSEE-NA-24402-02

[6] "Guideline for Follow-up of Safety Instrumented Systems (SIS) in the Operating Phase," Edition 2, 2021. SINTEF.

[7] "A Heirarchical Bayesian Approach to IEC 61511 Prior Use." Thomas, Stephen L., 14th GCPS, Orlando, FL, 2018.

[8] Moubray, John. "Reliability Centered Maintenance." Industrial Press, Inc. 2nd Ed., 1997.

[9] "Guidelines for Improving Plant Reliability Through Data Collection and Analysis," CCPS, Wiley, 1998.

[10] Kletz, Trevor, "Dispelling Chemical Engineering Myths," 3rd Edition, CRC Press, 1996.

[11] Savage, Sam L., "Chancification – How to Fix the Flaw of Averages," Coppell, TX. 2022.

[12] Silver, Nate. "The Signal and the Noise – Why so many predictions fail, but some don't" Penguin Press, 2015.

[13] Diaconis, P., Skyrms, B., "Ten Great Ideas about Chance," Princeton Press, 2017.

[14] Bertsch-McGrayne, S., "The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy." Yale Press, 2011.

[15] Grattan, D., Brumbaugh, K., "Reverend Bayes, meet Process Safety: Use Bayes' Theorem to establish site specific confidence in your LOPA calculation" https://www.aesolutions.com/post/reverend-bayes-meet-process-safety-use-bayes-theorem-to-establish-site-specific-confidence-in-lopa Accessed January 17, 2023.

[16] Taleb, N., "Statistical Consequences of Fat Tails: Real World Pre-asymptotics, Epistemology, and Applications (Technical Incerto)" STEM Press, 2020.

[17] Brumbaugh KA. "A Tale of Two BPCS Credits, A Bayesian Case Study." https://www.aesolutions.com/post/a-tale-of-two-bpcs-credits-a-bayesian-case-study , Accessed January 1, 2023.