Think Twice?

by **Tony McGrail**

Tony McGrail is Doble Engineering Company's Solutions Director for Asset Management & Monitoring Technology, providing condition, criticality and risk analysis for utility companies. Previously Tony has spent over 10 years with National Grid in the UK and the US; he has been both a substation equipment specialist and subsequently substation asset manager, identifying risks and opportunities for investment in an aged infrastructure. Tony is a Fellow of the IET, a member of the IEEE, CIGRE, ASTM, ISO and the IAM, and is currently active on the Doble Client Committee on Asset and Maintenance Management and a contributor to SFRA, Condition Monitoring and Asset Management standards. His initial degree was in Physics, supplemented by an MS and a PhD in EE followed by an MBA.

It's a matter of time and probability: flip a coin for long enough and at some point you are almost certain to get ten heads in a row, but there's no guarantee as to when it will occur, and it could take a long time.

This article looks at some examples of data analysis where making sure we have the relevant data has a significant impact on the analyses: whether it's a decision to enter an auto race, or to trust a stock market 'expert', or to assess and diagnose a critical asset, there will always be uncertainty and imprecision. Sometimes we need to look again and think twice.

Coin Trick

Derren Brown is a British illusionist who has hosted many TV specials which demonstrate his ability to apparently 'manipulate' the world around us in surprising ways.
A favorite is when, while being filmed from in front and from above, he says, "Ten heads in a row" and then proceeds to flip a coin ten times and gets ten successive 'heads', "as predicted" [1]. He's an illusionist, so, naturally, we look for the trick… but as far as we can tell there is none: it's a fair coin, we can see the flipping from two angles with no interference, and there's nothing to indicate the recording is being edited. So how does he do it?

If we only look at what's being shown, the 'data' in front of us, we may be missing a key element which enables us to reassess the data. That is certainly the case here: what are we not seeing? What information is missing? In this case it's the hours and hours spent flipping a coin where it did not result in ten heads in a row. This is a version of 'survivorship bias' where we start by looking at 'successful' efforts and overlook those that did not succeed [2].
It's a matter of time and probability:

Photo: Shutterstock

flip a coin for long enough and at some point you are almost certain to get ten heads in a row, but there's no guarantee as to when it will occur, and it could take a long time.

There's another video of the same 'trick' by a mathematician which goes a little deeper into probability theory [3]. But if all you see is the recording of the 'successful' trial, you may think there's some trick to it, some 'magic' or sleight of hand, when it's more a matter of stamina!

Carter Racing

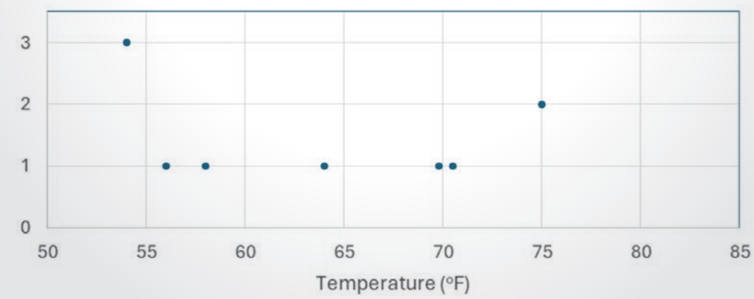
There's an interesting case presented by British Mathematician Hannah Fry involving auto racing [4]. It was dreamed up many years ago by a couple of business professors and used in risk management classes: a summary is in Box 1. What would you do: race or withdraw?

Box 1: Carter Racing Intro

John Carter has an hour to decide whether to compete tomorrow in the most important auto race of the season: success may mean sponsorship and a stable team future. But in 7 of the past 24 races the engine has blown out, and if that happens again it will put sponsorship at risk, not to mention the driver's life. But withdrawing means the team will end the season in debt and will miss out on a chance of 'glory'. And as Burns's First Law of Racing Says: "Nobody ever won a race sitting in the pits."

The team's mechanic thinks the engine's head gasket breaks in cooler weather and puts together a chart showing how many cracks were found in the gasket after each of the 7 unsuccessful races, and how they relate to ambient temperature, as shown below. Is there a link between failure and temperature? The ambient tomorrow is forecast to be about 40°F.

Number of Cracks in the Gasket

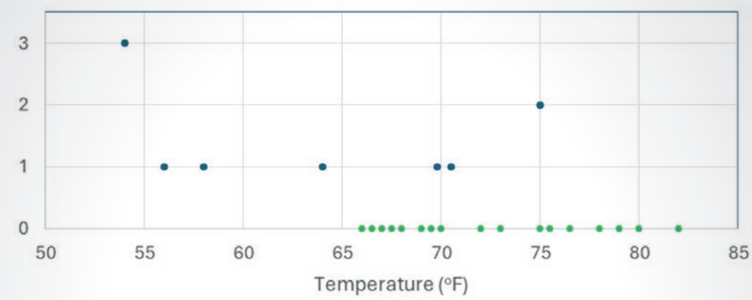


So, what do you think? Should they race or withdraw?

Box 2: Carter Racing: Decisions

How does our analysis change if we add into the chart those races which were successful? They have no cracks in the gasket and are shown in green in the chart below.

Number of Cracks in the Gasket



The additional data makes it very clear that there has never been a successful race completion below 65°F. So, with the additional data: now what would you decide? Race or withdraw?

We can, in fact, add an extra bit of spin to the analysis by noting that the charts actually have nothing to do with auto racing: the Box 1 chart is basically the data used to decide to launch the Challenger Space Shuttle, in 1986. The data was hand tabulated and faxed from the manufacturer to an emergency NASA meeting: given only the failures, most experts were not convinced there was a link between temperature and gasket failure. But nobody asked for the missing data, as per the Box 2 chart, related to successful launches, and the 'go ahead' was given, with disastrous and fatal consequences.

Ten years later Edward Tufte used this example as 'the wrong way' to display quantitative data, noting that the right chart would tell you what you need to know at a glance [5].

The case has been put to hundreds of people over the years, and most people choose to race, based on the data they have in the chart.
Very few ask for more data, and even fewer specify what that data should be – see Box 2.

Stock Buying Cows

As another example (and you're probably getting the idea now) we could look at the Norwegian documentary which looked at performance of different stock market investors: two industry 'experts', two 'influencers', an astrologer, and some cows [6].
How did the cows pick stocks?
By having their field marked into a grid with available stocks individually written on the grass in each grid cell, and wherever the cows 'relieved' themselves would be in their portfolio. At the end of a three-month period, with portfolios reviewed at the beginning of each month, the stock index had gone up by 5%.
The astrologer's portfolio rose by 4%, the industry experts by 7.28% and the cows by 7.26%.

It is often the case in generating health indices that we end up with the index being reduced to a set of categories and unless we look at the raw data which was used to put an asset into a specific category, we may be misleading ourselves.

But the influencers, who admitted to knowing nothing about the stocks and picked them 'on intuition' made 10 percentage.

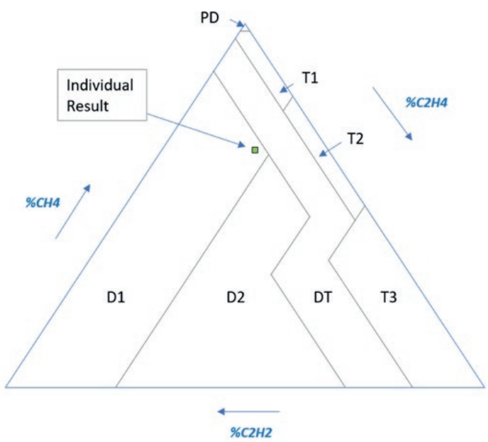
Admittedly it was only a 3-month experiment so there may have been a lot of luck involved, but the TV crew announced that they had also 'managed' a portfolio, and it had gone up 24%! Question is: how did they manage to get such high returns? The 'answer' is at the end of this article.

Dissolved Gas Analysis Accuracy

A common and popular tool for use in diagnosis of DGA results is the Duval Triangle [7] which plots three key dissolved gas concentrations as a single point on a triangular chart. The individual gas levels are summed and the percentage each one contributes to the total provides the value between 0 and 100% on each axis, as shown in an example in Figure 1. Each area of the chart is labeled with an individual diagnosis, as per the table to the right.

It would be easy to assume that the boundaries between areas are 'sharp' when, in fact, those boundaries are themselves based on historic DGA values with their own inaccuracies.

We may conclude from the point plotted in Figure 1 that we definitely have a D1, discharge of low energy. However, DGA comes with some inaccuracy, and a reasonable laboratory measurement is expected to be within 15% of the 'true' value. If the oil has exactly what we've measured in a sample, then we have just one 'successful' data point, but there are many more possible sets of values for the actual dissolved gases which could also yield the single result we got. On the left in Figure 2 the cloud of green points represents possible 'true' values of the DGA levels for the three gases which could yield the measurement we have made with



Label	Diagnosis
PD	Partial Discharges
D1	Discharges of Low Energy
D2	Discharges of High Energy
DT	Mixture of Thermal and Electrical Faults
T1	Thermal Faults of temperature < 300°C
T2	Thermal Faults of temperature 300°C < T2 > 700°C
T3	Thermal Faults of temperature > 700°C

Figure 1: Plotting a Single DGA result on a Duval Triangle

Photo: Shutterstock

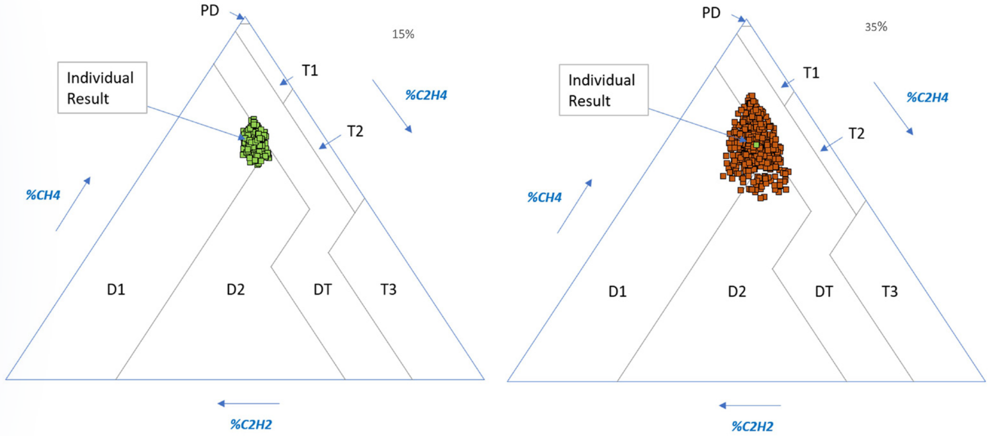


Figure 2: Plotting Possible True DGA Values at 15% and 35% accuracy on a Duval Triangle

an accuracy of 15%, these are the 'missing data' and most are in D1. But to the right in Figure 2 the cloud of dark red points all represents possible 'true' values with an accuracy of 35%, which is what online DGA systems may have in practice [8].

The result of the monitor inaccuracy is that more possible values for the actual DGA in the oil, the red dots in Figure 2, lie outside of the D1 diagnostic area than inside. Without knowing the accuracy of the laboratory or the monitor used we may have a very poor diagnostic result and an inappropriate prognosis.

Similar diagnostic inaccuracy can be found in other test/assessment measurements, such as power factor, winding resistance and so on, where the imprecision may have significant impact on interpretation and diagnosis.

Health Index Boundaries

When we make a measurement of some value, in whatever measurement unit we are interested in, our measurement technique and measurement system will provide sources of both systematic and random errors. The result of our measurement is thus only an 'estimate' of the true value. Numerous measurements of the same value will provide a 'distribution' around the actual value, often in the form of a Normal (a.k.a. Gaussian) distribution, symmetrical about a 'true' value. The spread of the distribution is characterized by the standard deviation, the confidence interval and the error [9].

In Figure 3, a measurement of 25 'units' (whether feet or degrees or picofarads or whatever…) is characterized by an error of +/-10% with a confidence interval of 90%, and a resulting standard deviation of 1.52 units. The vertical axis indicates the probability, on a scale of 0-1.0 (representing 0-100%), that the result is at a particular x-axis value.

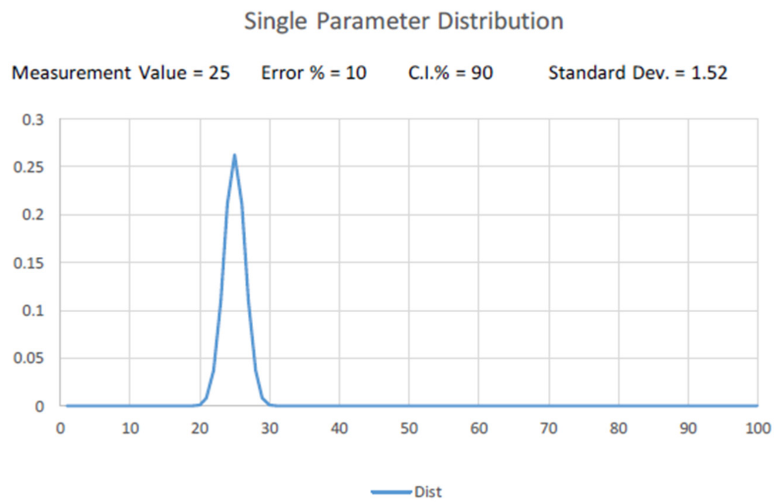


Figure 3: Parameter Measurement with Normal Distribution Around a True Value

The data in Figure 3 shows that there is uncertainty built into our measurements. Any system which attempts to encode the data into a scale – say 1-5 for condition with 1 is 'good' and 5 is 'bad', will be subject to the precision of the original measurement. As an example, Table 1 shows category, or coding, boundaries which may be applied to the measurement in Figure 3.

Overlaying the categories on to the data gives an indication of where the true value might lie, as shown in Figure 4. Note that in this case the likelihood that the true value is in the 20-40 range is 99.95%. A video on the pitfalls of thinking in categories, by Prof. Sapolsky of Stanford University, USA, is worth watching – it's a very human thing to assume that everything within a category is similar to everything else within that category, and everything in one category is very different to anything in another category [10].

Category/Code	Lower Limit	Upper Limit
1	0	20
2	20	40
3	40	60
4	60	80
5	80	100

Table 1: Category or Coding Limits for a 0-100 Measurement

If we now make a new set of measurements which happen to lie on or near a boundary, as shown in Figure 5, we can see that the likelihood of being in a particular category is 65.59% and 34.41% of being in an adjacent category. We have a lot less certainty in deciding the 'true' category.

The result of the inherent error in our measurement, as shown in Figure 5, is that we have an imprecision built into subsequent diagnoses and in anything which uses the result in calculations. What may be missing from the data we use is an estimate of the imprecision and thus the likelihood of being in a particular category.

It is often the case in generating health indices that we end up with the index being reduced to a set of categories and unless we look at the raw data which was used to put an asset into a specific category, we may be misleading ourselves.

This is especially true with risk matrices, where the health index is often used as a proxy for a probability of failure: we need to see the raw data with its imprecision, the 'audit trail' from that data to a diagnosis and thence to a probability of failure range which can be justified in terms of timescale: say 'the probability of failure in the next 12 months is between 0.5% and 1.2% with 90% confidence [11]. In fact, it would be more sensible to start with the data trail and then develop a health index based on the resulting probability ranges and their urgency. This does, of course, require that our health index addresses a well stated 'question' in a clear and auditable manner, and it may be that we need

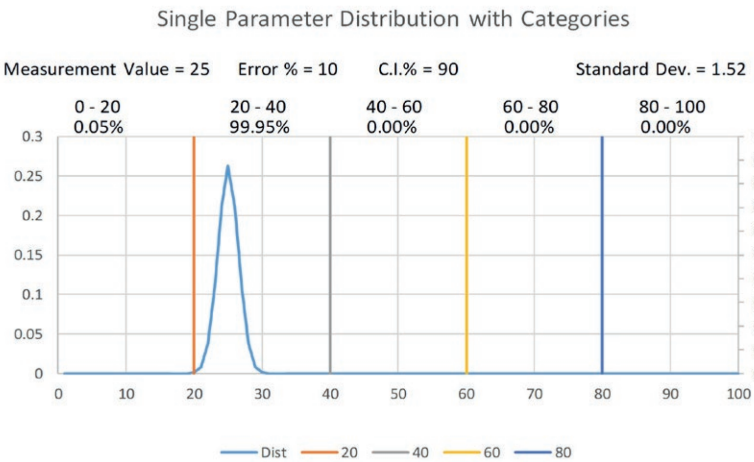


Figure 4: Parameter Measurement with Normal Distribution and Category Codes

Photo: Shutterstock

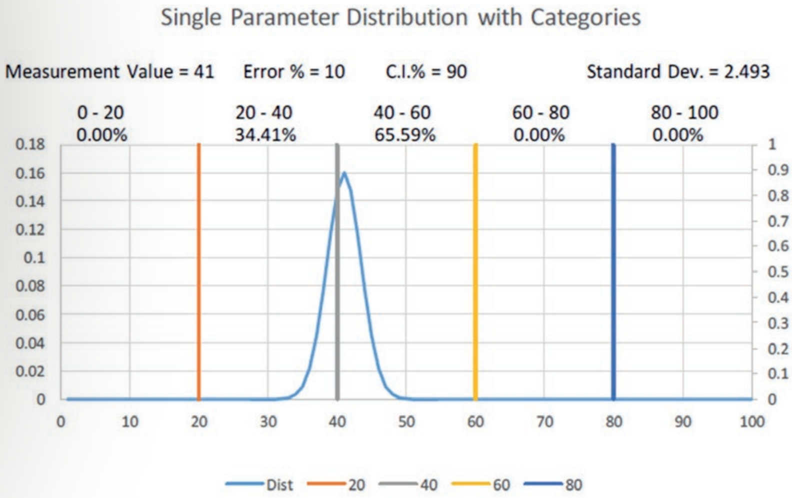


Figure 5: Parameter Measurement near a Category Boundary

multiple indices to address multiple questions.

Discussion

“What’s missing?” is a useful question to ask and helps remind us to look for more than just the data we may have available and avoid ‘survivorship bias’. It may be that we can never have a complete set of data, but at least we should be aware of that, and do what we can to counter the shortfall.

For the coin flip and the stock picking TV crew, we see survivorship bias in action: adding the unsuccessful efforts shows a very different story in each case. For the stock picking, there’s a famous quote form Burton Malkiel: “A blindfolded monkey throwing darts at a newspaper's financial pages could select a portfolio that would do just as well as one carefully selected by experts [12].”

For the ‘Carter Racing’ we have a couple of different effects, first being that we start with the failure data, rather than successful data, but the result is similar – knowledge of what’s missing can be crucial. The second effect is when it is revealed that the initial data was used to justify the space shuttle launch: the consequence of making the wrong decision becomes a lot greater and usually has a lot more influence on the decision-making process; it becomes a lot more about the risk involved and not just the ‘go/no-go’ decision.

> **...by reducing a lot of data, including data about the precision of that data, to a single value or category we may make things easier to understand but lose the information about the precision and its consequences for categorization.**

Duval’s triangle is a useful diagnostic tool, but interpretation of an individual result, or even a series of results, requires an understanding of how the chart works and what the implications of imprecision are.

Similarly, an asset health index can be a great tool for communicating technical information in a condensed form to less technically adept colleagues, often for asset ranking or prioritization purposes. The problem is that by reducing a lot of data, including data about the precision of that data, to a single value or category we may make things easier to understand but lose the information about the precision and its consequences for categorization.

As a colleague once asked: “I can see a few power transformers in the ‘Category 4’ box, which means do something within 12-24 months, but which one do I do first?” To answer that, we need to look at the individual units and their individual data and their individual urgencies: someone has to know what’s going on.

Stock buying cows: how did the TV crew do so well? It’s the ‘missing data’ scenario again: unlike the other participants, where each ‘team’ had just one portfolio, the TV crew secretly set up 20 different portfolios, and one of them did very well, and that was the one they shared to show their ‘success’. The other portfolios didn’t ‘survive’ and if we don’t ask for the ‘silent evidence’ of history’s losers we may well deceive ourselves [2].

Acknowledgment: thanks to Rick Aguilar for his comments and feedback.

References

[1] Derren Brown: https://www.youtube.com/watch?v=XzYLHOX50Bc
[2] Survivorship bias: https://en.wikipedia.org/wiki/Survivorship_bias
[3] Flipping ten heads in a row: https://www.youtube.com/watch?v=rwvIGNXY21Y
[4] Hannah Fry: https://www.newyorker.com/magazine/2021/06/21/when-graphs-are-a-matter-of-life-and-death
[5] Edward Tufte: https://www.amazon.com/Visual-Explanations-Quantities-Evidence-Narrative/dp/1930824157
[6] Cows: https://www.fynsa.com/en/newsletter/vacas-vs-expertos-el-secreto-esta-en-el-pasto/
[7] “IEEE Guide for the Interpretation of Gases Generated in Mineral Oil-Immersed Transformers”, C57.104-2019
[8] “DGA Monitoring Systems”, CIGRE Technical Brochure 783, 2019
[9] “Elementary Statistics” M. Triola, ISBN-13: 978-0-321-50024-3
[10] “Introduction to Human Behavioral Biology”, Prof. R. Sapolsky, Stanford, https://www.youtube.com/watch?v=NNnIGh9g6fA
[11] “The Risk of Using Risk Matrices” Thomas, Bratvold & Bickel, Soc. Petroleum Engineers, Annual Tech Conf., New Orleans, USA, 2013
[12] https://www.azquotes.com/quote/894760