C1 Understanding and Applying "*Daubert's* Error Rate" - Part One

Mark I. Marpet, PhD, PE, 14 Cowie Road, Chester, NJ 07930-9715; Howard P. Medoff, PhD, Pennsylvania State University, 1600 Woodland Road, Abington, PA 19001; and Marcus P. Besser, PhD, Thomas Jefferson University, 130 South Ninth Street, Philadelphia, PA 19107-5233*

After attending this presentation, attendees will learn one important form that an error rate can take and how to develop an uncertainty interval, derived from the Chebychev inequality, which can serve as an error rate for a measurement process. This presentation will impact the forensic science community by providing attendees a better position to qualify or cross-examine an expert with respect to this area.

Since the *Frye* decision was superseded by Rule 702 of the Federal Rules of Evidence, as interpreted by the *Daubert*, *General Electric*, and *Kuhmo Tire* decisions (The *Daubert* Trilogy), it is no longer enough to show that basis for the admission of expert evidence in a judicial proceeding is that it is generally accepted in a relevant field of study. These Federal Rules of Evidence require that the evidence itself be reliable. In 2009, the National Academy of Sciences issued a report, entitled *Strengthening Forensic Science in the United States: A Path Forward* (the NAS Report). It looked at the state of the art of forensic science and found that, in many respects, forensic science disciplines were wanting.

One of the criteria that the court set forth in the *Daubert* Trilogy and discussed in the NAS Report is the (known or potential) error rate of the hypotheses or techniques being proffered. The concept of error rate causes consternation and confusion to both forensic practitioners and to the legal community. This presentation will impact the forensic science community by abating some of that confusion by: (a) defining in practical terms what an error rate is or can be; and (b) offering guidance to practitioners how to develop an error rate for techniques that they utilize. Thus, the impact to will be for practitioners to improve their own practice of forensic science and, by instilling this knowledge in the legal community, members of which can then ask appropriate questions, to make sure that forensic science practitioners do indeed treat error rate in a meaningful manner.

The trial judge that must act as a "gatekeeper" to determine admissibility of expert testimony to the trier of fact on the basis of both:
(a) the ability to assist the jury; and, (b) whether or not the information the expert intends to offer is indeed "scientific knowledge." *Daubert* sets forth a non-inclusive list of criteria to be considered, among them, whether there is a "known or potential rate of error", a concept from metrology: the science of measurement.

Error rate is an appropriate term when one deals with, for example, a dichotomous classification, e.g., whether or not there is or is not a match between a suspect and a piece of evidence.[1] Unfortunately, "error rate" is problematic when it comes to *measuring* things of forensic import, because it conflates *error* with *measurement uncertainty*, which is not "error" at all but, rather, a concept in metrology that reflects the fact that measurements are never without at least some variability.[2]

While expert testimony must be grounded in science (for example, mechanisms of friction). Once the science is established then the admissibility of expert testimony revolves around things more quotidian (the actual measuring of the friction between a reference material and a given floor surface). In place of hypothesis testing and its associated errors, there is accuracy, standardization, and calibration issues.

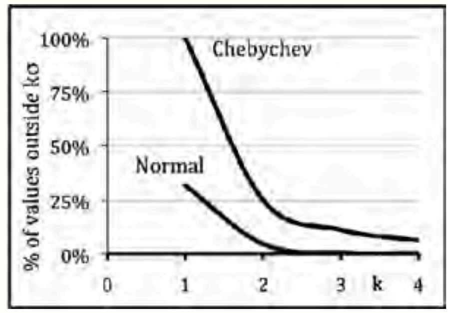
One tool that has been used for this purpose is a *Confidence Interval* (CI), which depends upon measures of the central tendency (typically, the Arithmetic Mean) and the dispersion (typically, the Standard Deviation). A CI is expressed as a range of values subject to a level of confidence in that range of values. For example, one can state that a tribometric test sequence determined that the friction between the floor and a reference surface was between 0.37 and 0.47 with 95% probability:
$P(0.37 \leq \mu \leq 0.46) = 95\%$. The production of a CI requires the collection of data in a contextually-appropriate manner, the calculation of sample statistics, typically the mean and standard deviation, and from that, a calculation of the CI using straightforward and well-known formulae. CIs for single or small samples require that the underlying probability distribution follow a Normal (Gaussian) Distribution.[3] Unfortunately, many forensically interesting phenomena are not Normally distributed.[4] Parts one and two of this paper look at alternative ways of
developing uncertainty estimates:

1. The Chebychev inequality (discussed in this part of the paper).
2. The sign-test derived confidence interval on the median (discussed in part 2).
3. The uncertainty interval based upon the cumulative function (discussed in part 2).

 ** Presenting Author*

1-**The Chebychev Inequality:** The Chebychev inequality, which requires no information about the underlying probability distribution other than its mean (µ) and a standard deviation (s), states that the probability of being farther than *k* standard deviations from the population mean is less than $^1/_k$:

$$P(|x-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$



For example, the probability of being greater than two standard deviations from the mean is less than 25%.

$$P(|x-\mu| \geq 2\sigma) \leq \frac{1}{2^2} = \frac{1}{4} = 25\%$$

This result is weak compared to the narrower confidence interval that can be arrived at given information about the underlying probability distribution. If the underlying probability distribution was normal, for example, the probability of being greater than two standard deviations from the mean is about 5%, one-fifth of the Chebychev-determined value. (See the graph above.)

Example 1: The Chebychev Uncertainty Interval

Col. (1)	Col. (2)	Col. (1)	Col. (2)	Col. (1)	Col. (2)	Col. (1)	Col. (2)
0	0	7	1	14	6	21	0
1	0	8	2	15	3	22	1
2	2	9	4	16	1	23	0
3	0	10	3	17	1	24	0
4	1	11	7	18	1	25	1
5	0	12	11	19	0	26	0
6	0	13	8	20	2	27	0

The following data represents the time, in milliseconds, that it takes a laser rangefinder to stabilize its reading. Column one represents the stabilization time, rounded to the nearest millisecond, that the rangefinder takes to settle; column two represents the number of times that that particular reading is seen out of 55 readings. For example, the device took 2 ms to settle on two occasions and on eight occasions, it took 13 ms to settle. The mean and standard deviation can easily be calculated in Excel to be 12.31 and 4.09 ms respectively.[5]

Thus, 25% of the readings are within 12.31 ± 2x4.09 = 12.31 ± 8.18 = (4.13, 20.49).

If one is working with single samples or very small samples, as is often the case with forensic samples, and if there is historical instrument data from which to build a standard deviation, the Chebychev Inequality will enable one to develop an Uncertainty Interval (UI), albeit a rather wide one.

Discussion: An instrument or process, absent an externally imposed rule or standard, cannot "pass" or "fail" an UI or a CI. Rather, and this is important, the utility or lack of utility is situational. A hypothetical floor-friction example illustrates this. In civil litigation, a plaintiff must prove that a floor is "slippery." If the tribometer gave us a 95% CI or UI of the available friction as ($0.35 \leq \mu_a \leq 0.43$) on the true coefficient of friction, i.e., it is 95% certain that the true value of floor friction is between 0.37 and 0.43, that in and of itself cannot determine whether the floor is or is not slippery. There must also be an external reference, a threshold: if the pedestrian-required friction is, say, μ_r=0.45, with values at or above 0.45 being slip resistant, and values below that slippery. If the high side of the CI or UI is, say, 0.43, the floor would be proven, within reasonable certainty, to be slippery.

One might argue that a one-sided CI is more appropriate because we are only interested in one side of the error rate CI (or UI). In this friction- testing example, where the movant wants to prove that the friction is,

 *Presenting Author

within reasonable certainty, below slipperiness threshold, one could argue that a one-sided, "high side" CI, better suits the task (if the high side of a 2-sided CI—or a 1-sided, high-side CI—is below the friction threshold, this will prove the slipperiness. Conversely, if the high side of the available-friction CI is above friction required by the pedestrian, then the floor cannot be proven to be slippery). A one-sided, upper CI would give a lower high side, making it easier to show that the floor was slippery. The more conventional two-sided CI is preferable because (a) the two-sided interval gives a larger interval on the side of interest (whether that's the low or the high end), reflecting a larger uncertainty, making it *harder* for the movant to prove the slipperiness hypothesis, and is thus more conservative and, (b) the uncertainty of most processes is in fact better reflected by a two-sided UI.

References:

1. Dichotomous classifications, like this 'match' example, actually have two error rates: the probability of a false positive (that a match is mistaken) and the probability of a false negative (that a match will be missed). *Daubert* concerns itself with the former.
2. Measurement Uncertainty is the parameter, associated with the result of a measurement (the measurand) that characterizes the dispersion of the values that could reasonably be attributed to that measured value. (Guide to the Expression of Uncertainty in Measurement, ISO, Geneva, 1993.)
3. For the sample mean (the arithmetic average), the Central Limit Theorem states that as the sample size increases, the distribution of that mean rapidly converges to the Normal Distribution. Practically, if the underlying distribution of the values has a unimodal (one-hump) shape, the average of a sample of 10 or more will converge to a normal distribution; Regardless of the distribution's shape, the average of a sample of 30 or more will onverge to a Normally-distributed statistic. Unfortunately, many forensic analyses use single or very small samples, making the Central Limit Theorem inapplicable.
4. The distribution of Perception-Reaction Time is log-normal; the distribution of the probability of an indicated slip on a Brungraber Mark II walkway friction tester is binomial, and the CF can be determined using Logistic Regression. ("The Relationship Between the Measured Friction Coefficient and the Safety of a Walkway Surface," 2010 Proceedings of the American Academy of Forensic Sciences, Vol. 16, p. 159-160, Seattle, WA, and part 2 of this paper).
5. In Excel, the formulae are =average(range) and =stdev(range) respectively, where range is the selection of the individual values.

***Daubert* Error Rate, Forensic Science, Uncertainty Interval**

