

# Updating Guidelines for Evaluating and Expressing Uncertainty of NIST Measurement Results

Antonio Possolo

Statistical Engineering Division  
Information Technology Laboratory

March 6, 2014



1/63

## Steps Taken

- **May 15-16, 2012**
  - NIST's Measurement Uncertainty Policy
  - 24 presentations by NIST scientists
- **January 28, 2013**
  - Draft for revised NIST procedure on *Evaluation and Expression of Measurement Uncertainty*
  - 21 sets of written comments and suggestions for improvement
  - 59 items addressed in response to discussion
- **July 23, 2013**
  - First edition of this presentation

2/63

## Viewpoints

- *Must every new calibration setup, experiment, key comparison, proficiency test, calibration report, and publication be developed in **active consultation** with the Statistical Engineering Division?*
- *While the methods [...] can be applied to **uncertainty analysis of complex computations**, they alone are not adequate to characterize the uncertainty in this context*

3/63

## Viewpoints

- *I do not think it would be a useful document for me or my staff in estimating the uncertainty of our calibrations — the statistics jargon alone makes it nearly **impenetrable***
- *Different levels of uncertainty analysis are appropriate for different types of calculations [...] Require that uncertainty analysis be **sufficient to support conclusions** being drawn*

4/63

## Viewpoints

- You can measure a number of values of say a voltmeter, but the GUM will give you *only the uncertainty of each measurement*
- It is essential that the *data dependency* existing in the original test data be taken into account
- Currently, *none of the companies I know use uncertainty at all*. People who make airplanes, engines, computers, and other high end products seem to use Repeatability & Reproducibility studies to validate inspection methods

5/63

## CAT for Gageless Tooling

6/63

## CAT for Gageless Tooling

## F-22 Raptor

7/63

8/63

## Viewpoints

- *I'm very happy to hear that you guys are working on an update to the TN1297. I've always had difficulty using that document. It **rarely seemed to apply** to what I wanted to do*
- *One of my problems when I first came here was that 1297 is written from a metrology/statistician point of view. [...] I found 1297 almost **impenetrable***

9/63

## Viewpoints

- *Document needs to be **written in the vernacular** — I'm reminded of the effect of publishing the King James Bible in English rather than in Latin*
- *Maybe we just need a huge encyclopedia of **worked examples**, so we can see something that's closer to something we're trying to do*
- *This is truly a well written and incredibly clean and understandable uncertainty guide!*

10/63

## General Goals

- **Increase freedom of choice**  
that scientists, statisticians, and mathematicians require to address needs of rapidly evolving and expanding fields of measurement science
- **Widen class of measurement models**  
used to assign value to measurands and to evaluate measurement uncertainty
- **Facilitate critical assessment of models and assumptions**  
in particular of those that support probabilistic interpretation of uncertainty

11/63

## Grandfathering

- All **published uncertainty evaluations** associated with NIST measurement services (SRMs / calibrations) **remain valid** and will not need to be redone
- Procedures described in **NIST TN 1297** and in the **GUM** (1995, 2008) may continue to be used when the assumptions that validate them appear **plausible**

12/63

## Measurement & Measurement Result

- Experimental or computational process that produces a measurement result supporting decision-making

— cf. R. White, 2011, ACQUAL 16: 31–44

- Measurement result comprises
  - Estimate of measurand
  - Assessment of measurement uncertainty

13/63

## Message in a Bottle

### KEY FACTS

- Measurement uncertainty expressed most completely and generally by **probability distribution**
  - Characterizes state of knowledge about measurand
  - Possibly summarized to be fit-for-purpose
- Experimental data may be used alone or combined with other information to
  - Estimate measurand
  - Evaluate measurement uncertainty
- Uncertainty evaluation to be done consistently with measurement model
  - *Measurement equation* — Monte Carlo / Gauss
  - *Observation equation* — Statistical Methods

14/63

## Outline

- Steps Taken, Viewpoints, Goals
- Grandfathering
- Measurement & Key Facts
- Concepts, Tools, Examples
  - *NIST Uncertainty Machine*
- Measurement Quality Policy
- Why Update & Change?
- Next Steps

15/63

## Measurement Uncertainty

### DEFINITION

- Quantity that characterizes the dispersion of the values that may be attributed to the measurand and that are consistent with the experimental data and with other relevant information about the true value of the measurand
  - Probability distribution on the set of possible values of the measurand
  - Fit-for-purpose summaries
    - Standard measurement uncertainty
    - Coverage region
    - Approximation / Estimate of probability density

16/63

# Measurement Uncertainty — Evaluation

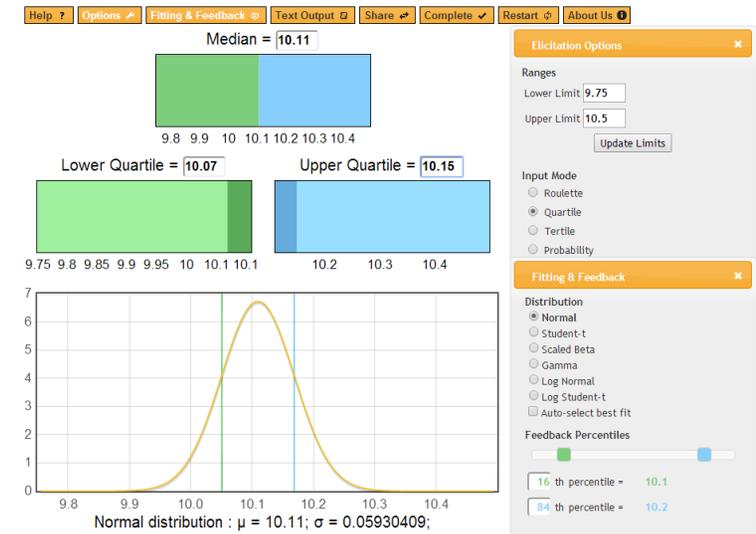
- Evaluated so as to be **fit for purpose**
- Sources of uncertainty may be evaluated based on:
  - Experimental data (**Type A evaluations**)
  - Other sources of information (**Type B evaluations**)

Elicitation of expert opinion  
— structured procedure to do Type B evaluations

17/63

# Uncertainty Elicitation Tool (MATCH)

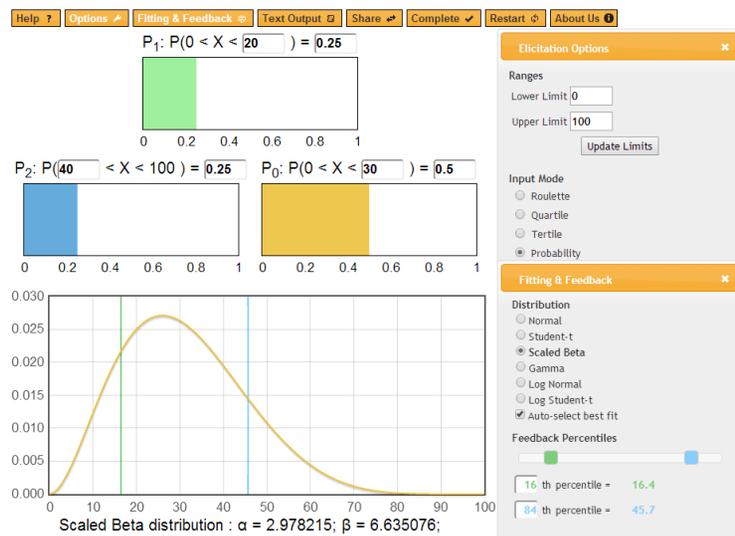
With 50% probability, length of part lies between 10.07 mm and 10.15 mm, and otherwise is as likely to be below 10.11 mm as above



18/63

# Uncertainty Elicitation Tool (MATCH)

Proportion of alite in cement clinker as likely to be below 30% as above. The other quartiles are 20% and 40%



19/63

# Measurement Uncertainty — Evaluation

BOTTOM-UP / TOP-DOWN

- **Bottom-up assessments** — uncertainty budgets for individual labs or measurement methods
- **Top-down assessments** — via interlaboratory and multiple method studies
  - Often reveal unsuspected uncertainty components

*Dark uncertainty*  
— Thompson & Ellison (2011)

20/63

# Measurement Uncertainty — Evaluation

TECHNICAL DEVICES 1/2

## GAUSS'S (1823) FORMULA — GUM (10), (13)

- Estimates, standard uncertainties, and correlations of input quantities
- Measurement function  $f$  approximately linear in neighborhood of estimates of input quantities
- Uncertainty of input quantities must be small relative to size of that neighborhood
- Values of partial derivatives of  $f$
- Probabilistic interpretation involves additional assumptions

21/63

# Measurement Uncertainty — Evaluation

TECHNICAL DEVICES 2/2

## MONTE CARLO METHODS

GUM-S1/S2 Metropolis & Ulam (1949), Morgan & Henrion (1992)  
MCMC Geman & Geman (1984), Gelfand & Smith (1990)

- Detailed probabilistic modeling of contributions from all recognized sources of uncertainty
- Specialized software
- No linear approximations or derivatives needed
- Results automatically interpretable probabilistically

22/63

# Measurement Models

- Describe relationship between value of measurand and quantities used to estimate it

*Uncertainty evaluation must be consistent with measurement model*

## ▪ Measurement equation

- Monte Carlo Method
- Gauss's Formula

## ▪ Observation equation

- Statistical Methods

23/63

# Measurement Models

## MEASUREMENT EQUATION

- Measurand (*output quantity*) is known function of input quantities
- Thermal expansion coefficient  
$$\alpha = (L_1 - L_0)/(L_0(T_1 - T_0))$$

## OBSERVATION EQUATION

- Measurand is known function of parameters of probabilistic model for experimental data
- Observed lifetime of a part has Weibull probability distribution, and measurand is expected lifetime

24/63

## Example — Observation Equation

LIFETIME — F-100 *Super Sabre*

- **Measurand:** component lifetime
- **Times to failure** (hour): 0.22, 0.50, 0.88, 1.00, 1.32, 1.33, 1.54, 1.76, 2.50, 3.00, 3+, 3+, 3+

25/63

## Example — Observation Equation

LIFETIME — F-100 *Super Sabre*

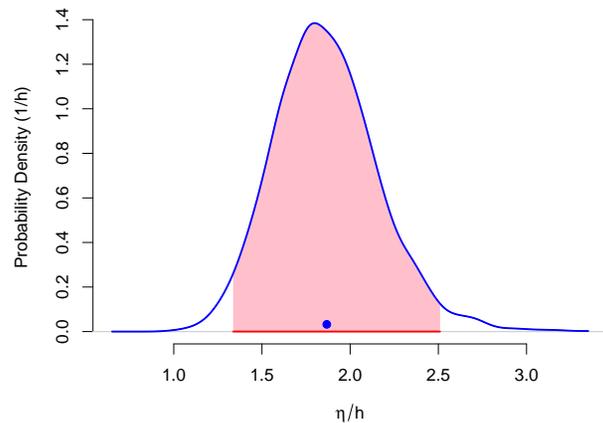
- **Observation equation** (*statistical model*)
  - Observed lifetimes are sample from Weibull distribution with shape  $\alpha$  and scale  $\beta$
  - Expected lifetime  $\eta = \beta\Gamma(1 + \frac{1}{\alpha})$
- **Value assignment**
  - *Ad hoc* methods
  - **Maximum likelihood**
  - Bayes
- **Uncertainty evaluation**
  - **Parametric statistical bootstrap**
  - Bayesian posterior distribution

26/63

## Example — Observation Equation

LIFETIME — F-100 *Super Sabre*

- **Measurement result:**  $\hat{\eta} = 1.84 \text{ h}$ ,  $u(\hat{\eta}) = 0.30 \text{ h}$   
Lifetime shorter than 4.5 h with 99% probability



27/63

## NIST Uncertainty Machine — Example

THERMAL EXPANSION COEFFICIENT

$$\alpha = \frac{L_1 - L_0}{L_0(T_1 - T_0)}$$

	$x$	$u(x)$	$\nu$
$T_0$	288.15 K	0.02 K	3
$L_0$	1.4999 m	0.0001 m	3
$T_1$	373.10 K	0.05 K	3
$L_1$	1.5021 m	0.0002 m	3

28/63

# NIST Uncertainty Machine

USE IT TODAY!

## WEB APPLICATION

[stat.nist.gov/uncertainty](http://stat.nist.gov/uncertainty)

## DESKTOP APPLICATION

[www.nist.gov/itl/sed/gsg/uncertainty.cfm](http://www.nist.gov/itl/sed/gsg/uncertainty.cfm)

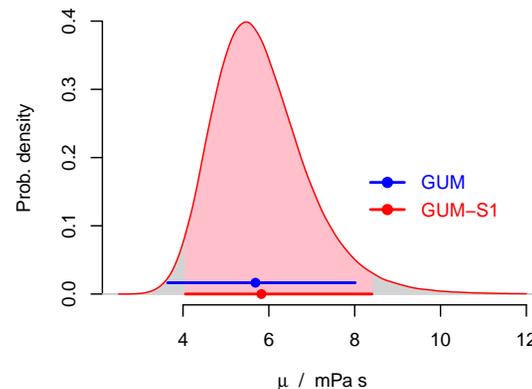
- Thomas Lafarge & Antonio Possolo
- Bug reports, corrections, suggestions, ...  
[antonio.possolo@nist.gov](mailto:antonio.possolo@nist.gov)

29/63

# Example — Type A Evaluation

FALLING BALL VISCOMETER

$$\mu_M = \mu_C \frac{\rho_B - \rho_M t_M}{\rho_B - \rho_C t_C}$$



22% solution of sodium hydroxide in water at 20°C  
HAAKE boron silica glass ball no. 2

30/63

# Outline

- Steps Taken, Viewpoints, Goals
- Grandfathering
- Measurement & Key Facts
- Concepts, Tools, Examples
  - *NIST Uncertainty Machine*
- Measurement Quality Policy
- Why Update & Change?
- Next Steps

31/63

# Measurement Quality Policy

Directive Number P 830.01

Effective Date November 20, 2012

[http://inet.nist.gov/adlp/directives/  
measurement\\_quality\\_policy.cfm](http://inet.nist.gov/adlp/directives/measurement_quality_policy.cfm)

*NIST will maintain and document the quality of NIST measurement services and of NIST measurement results by means of a quality management system described in the NIST Quality Manual*

- Sally Bruce, NIST Quality Manager

32/63

## Statements of Uncertainty

All reported NIST measurement results

- *Test or calibration reports for **calibration services***
- *Certificates and Certificates of Analysis for **reference materials***
- *Interlaboratory studies and **key comparisons***

accompanied by quantitative statements of uncertainty

— NIST QM-I 5.4.3

33/63

## Measurement Quality Assurance Program

- Provide **credibility** to measurement result
  - **Monitor performance** (stability, reproducibility, etc.) of instrument, standard, or measurement system
- **Contemporaneously measure check standards** alongside object of measurement

34/63

## QM I Appendix C

### RESPONSIBILITIES

Statistical Engineering Division responsible for providing

**technical advice and concurrence**

on statistical methods for evaluating and expressing the uncertainty of NIST measurement results,

including those that pertain to **SRMs, calibrations, interlaboratory studies, and key comparisons**

— NIST QM-I Appendix C3

35/63

## QM I Appendix C

### EXCEPTIONS

**Any statistical methods that the Statistical Engineering Division determines to be valid [...] may be employed**

uncertainty report must document what was done, and why

— NIST QM-I Appendix C4

36/63

## Outline

- Steps Taken, Viewpoints, Goals
- Grandfathering
- Measurement & Key Facts
- Concepts, Tools, Examples
  - *NIST Uncertainty Machine*
- Measurement Quality Policy
- Why Update & Change?
- Next Steps

37/63

## Why Update and Change?

- Many measurands are **neither** quantitative **nor** scalar
- $Y = f(X_1, \dots, X_n)$  may not be best measurement model or even applicable
- Inadequate guidance for **Type B evaluations**
- No means to incorporate relevant **external information** about measurand or measurement method

38/63

## Why Update and Change?

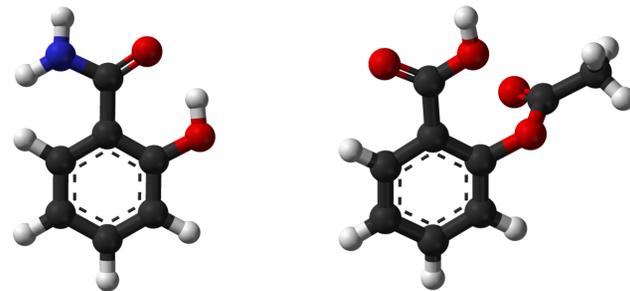
- How to reduce data from **interlaboratory studies**?
- How to **pool independent measurement results**?
- How to validate { **approximations / assumptions** } necessary for probabilistic interpretation of coverage regions?

39/63

## General Measurands

NOMINAL / CATEGORICAL

- **Identity of substance**
  - **Salicylamide** or **Aspirin**?



40/63

# General Measurands

NOMINAL / CATEGORICAL

- Sequence of nucleotides in DNA

## NIST SRMs

- 2374 DNA Sequence Library for External RNA Controls
- 2392 Mitochondrial DNA Sequencing
- 2391c PCR-Based DNA Profiling
- 2393 Huntington's Disease CAG Repeats
- 2394 Heteroplasmic mtDNA Mutation Detection
- 2395 Human Y-Chromosome DNA Profiling
- 2399 Fragile X Human DNA Triplet Repeat

41/63

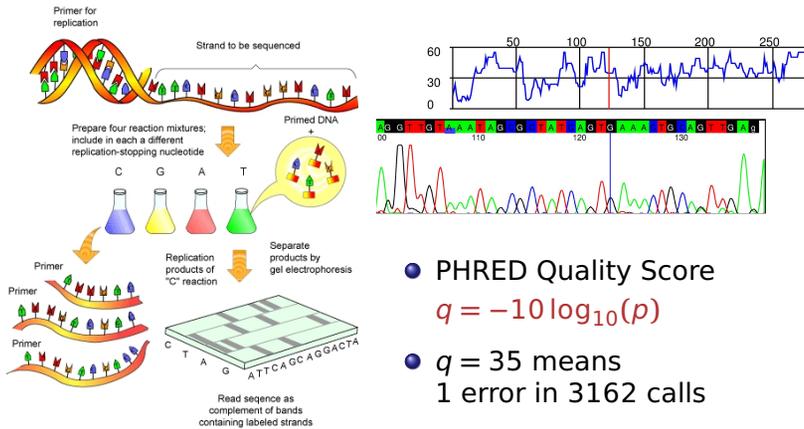
# General Measurands — Nominal/Categorical

## ISO GUIDE 35

- Properties can be quantitative or qualitative
- Concept of value includes qualitative attributes such as identity or sequence
- Uncertainties for such attributes may be expressed as probabilities

42/63

# Sanger Sequencing of DNA



- PHRED Quality Score

$$q = -10 \log_{10}(p)$$

- $q = 35$  means 1 error in 3162 calls

SOURCES Carolyn Elya, *SciFly*, [www.eisenlab.org/FunFly](http://www.eisenlab.org/FunFly)  
CodonCode Aligner, [www.codoncode.com](http://www.codoncode.com)

43/63



SOURCE: Wellcome Trust Sanger Institute

Frederick Sanger  
1918 – 2013

44/63

# Nominal Properties

## FORENSIC STUDIES

### INNOCENCE PROJECT

- [www.innocenceproject.org](http://www.innocenceproject.org)
- More than 300 people in the US have been exonerated by DNA testing

### WRONGFUL CONVICTIONS

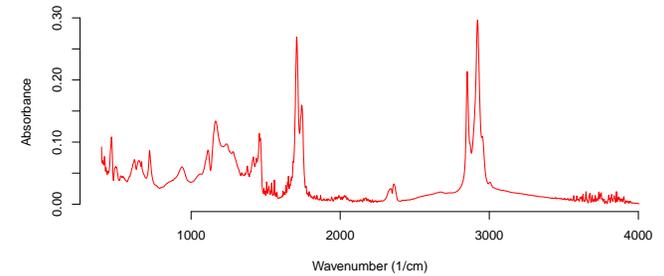
- 50 % attributable to
  - Unvalidated or improper applications of forensics
  - Misleading expressions of examination uncertainty
    - Hair / fiber matches
    - Bite-mark / shoe-print comparisons
    - Firearm tool-mark examinations
    - Serological studies

45 / 63

# General Measurands

## ORDINAL / MULTIVARIATE / FUNCTIONAL

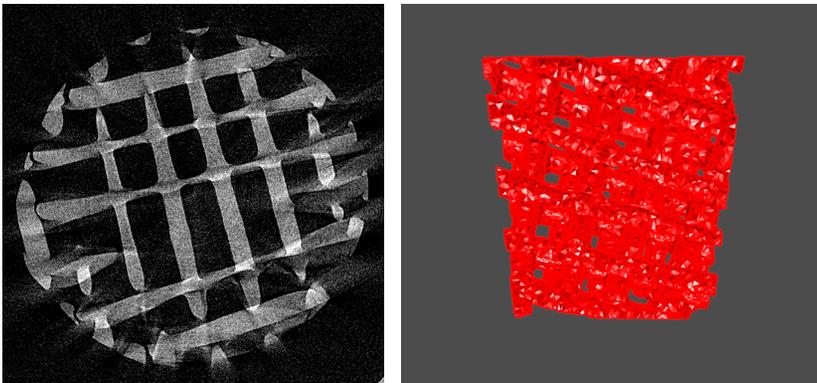
- **Ordinal:** Mohs hardness  
Mercalli earthquake intensity
- **Multivariate:** Mineralogical composition of cement clinker
- **Functional:** Spectra (Mass, IR, NMR, etc.)



46 / 63

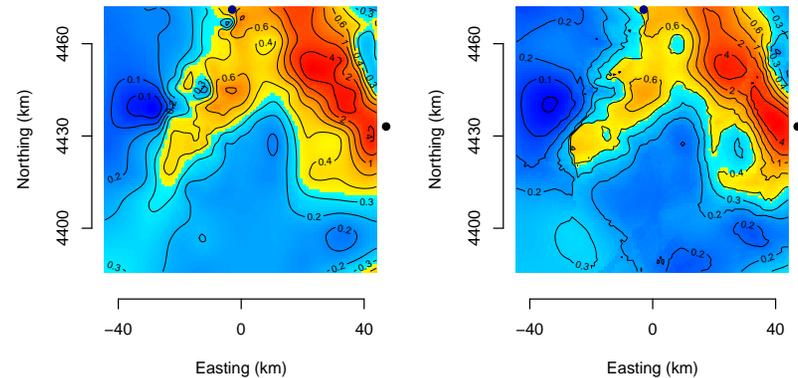
# General Measurands

## SHAPES



47 / 63

# General Measurands — Maps



- **Fukushima, Japan (September, 2011)**  
Radioactivity values (mR/hour) measured by Safecast  
Robust Local Regression and Ordinary Kriging

$$1 \text{ mR} = 2.58 \times 10^{-7} \text{ C/kg}$$

48 / 63

## General Measurands — Movies

- Flux of anthropogenic CO<sub>2</sub>, Indianapolis, IN  
**Hestia Project** — Kevin Gurney  
Arizona State Univ.

49/63

## External Information & Type B Evaluations

- Neither GUM nor current policy afford means to incorporate relevant **external information** about:
  - Value of measurand
  - { Within / Between } { Lab / Method } dispersion of values

50/63

## Interlaboratory Studies

- Neither GUM nor NIST Policy provide guidance for:
  - Computing **consensus values**
  - Evaluating associated uncertainty
  - Characterizing differences between participants

*agreement between laboratories is, in general, much less satisfactory than can be achieved within a single laboratory, even in the case of well-established methods of measurement*

*the introduction of statistics in this field has not been so successful as might be expected*

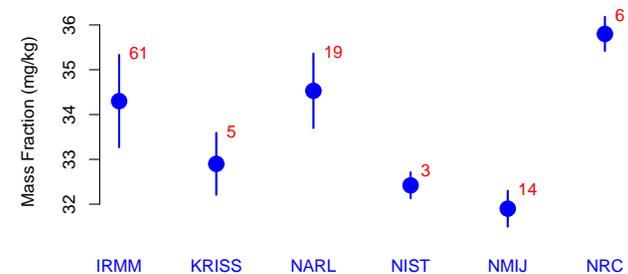
— **John Mandel** (1959, Page 251)

51/63

## Interlaboratory Studies

### CCQM-K25: PCB CONGENERS IN SEDIMENT

- 6 NMIs measured mass fraction of PCB 28
- **DATA:** Measured value, standard measurement uncertainty, number of determinations



52/63

# Interlaboratory Studies

PCB 28, CCQM-K25

- **Random effects models** evaluate and propagate **dark uncertainty**

Laboratories participating in an interlaboratory study are never a “random drawing” from *all* laboratories

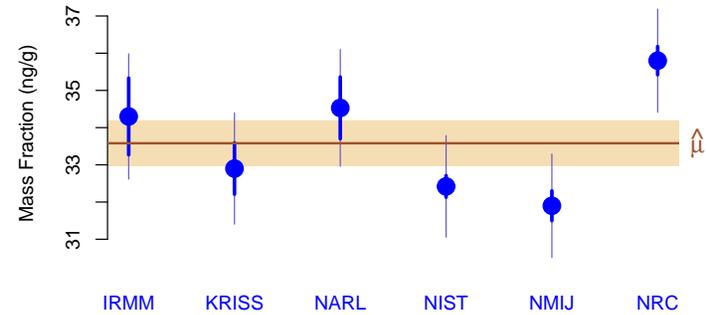
Inferences drawn [ . . . ] will be valid for laboratories *similar* to those who participated in the study

— **John Mandel** (1991, Page 64)

53/63

# Interlaboratory Studies

PCB 28 — CCQM-K25



- Consensus value  $\hat{\mu}$
- Standard measurement uncertainty  $u(\hat{\mu})$
- **Dark uncertainty: 2.5 times larger** than within-lab variability

54/63

# Interlaboratory Studies

PCB 28 — CCQM-K25

- Effect of prior information about relative size of **dark uncertainty**

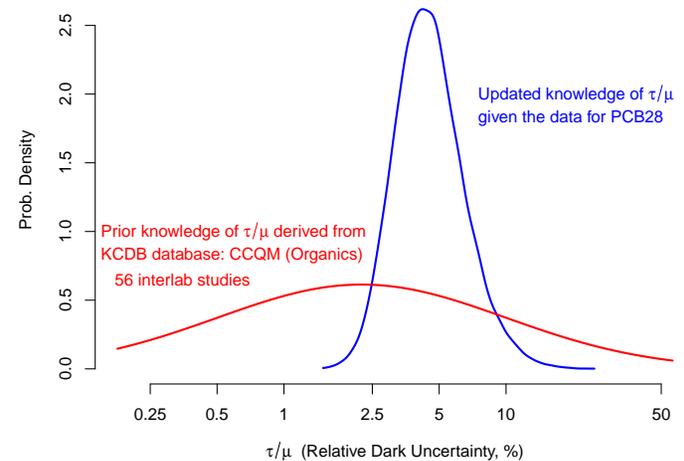
		ng/g	
		NONE	KCDB
Measurand	$\mu$	33.6	33.6
Std. uncertainty	$u(\mu)$	0.877	0.760
Dark uncertainty	$\tau$	1.60	1.52

**KCDB:** Distribution of  $\tau/\mu$  in 56 CCQM (Organics) interlab studies (Andrew Rukhin, 2013)

55/63

# Interlaboratory Studies

PCB 28 — CCQM-K25



56/63

## Probabilistic Interpretation

### COVERAGE INTERVALS

- Probabilistic interpretation of coverage intervals relies on **myths** about practical relevance of **Central Limit Theorem**

For many practical measurements in a broad range of fields [...] **probability distribution** characterized by the measurement result and its combined standard uncertainty **can be assumed to be normal because of the Central Limit Theorem**

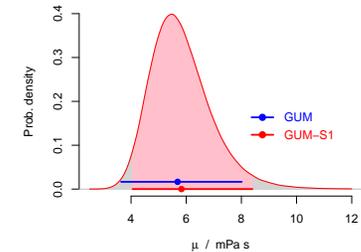
— GUM, G.6.6.

57/63

## Probabilistic Interpretation

### MONTE CARLO METHODS (1/2)

- Monte Carlo methods characterize distribution of output quantity generally more reliably than Gauss's formula and the CLT
  - Even under "default" assumptions (input quantities independent and Gaussian), output quantity need not be approximately Gaussian



58/63

## Probabilistic Interpretation

### MONTE CARLO METHODS (2/2)

- Monte Carlo methods automatically provide probabilistic interpretation for coverage regions
  - because they **produce samples** from probability distribution of measurand
  - *Other statistical techniques, including Monte Carlo or bootstrap methods, may be used to determine the uncertainty associated with the property value of a CRM* — 6.1, ISO Guide 35 (2006, 3rd ed)

59/63

60/63

## Next Steps

Circulate for comments:

- Updated **draft procedure**
  - *Please read and comment on 2-pager handed-out today*
- Updated **collection of examples**
  - ◆ Stefan-Boltzmann
  - ◆ Titration
  - ◆ Load cell calibration
  - ◆ Ionization energies
  - ◆ PCBs
  - ◆ Refractive index
  - ◆ Thermal bath
  - ◆ Cement clinker
  - ◆ Fukushima

61/63

## Acknowledgments

### PRODUCTION

- Mary Jo DiBernardo — **Producer**
- Patrice Boulanger — **Production Coordination**
- Joseph Hynes & Crew — **Audio-Visual**

### REVIEWERS

- Participants — **May 2012 Workshop**
- Steve Lund, Adam Pintar — **Workshop scribes**
- Authors (21) of written comments — **1st Draft**
- Will Guthrie, Alan Heckert, Jack Wang — **WERB**
- Jolene Splett — **WERB & WWW**

62/63

## Acknowledgments

### EXAMPLES

- Tom Bartel, Ricky Seifarth
- Simon Kaplan
- Mark Stiles
- Paul Stutzman
- Tom Vetter

### SPECIAL EFFECTS

- Thomas Lafarge — **NIST Uncertainty Machine CAVOSS3D, ImageJ Plugin**

63/63