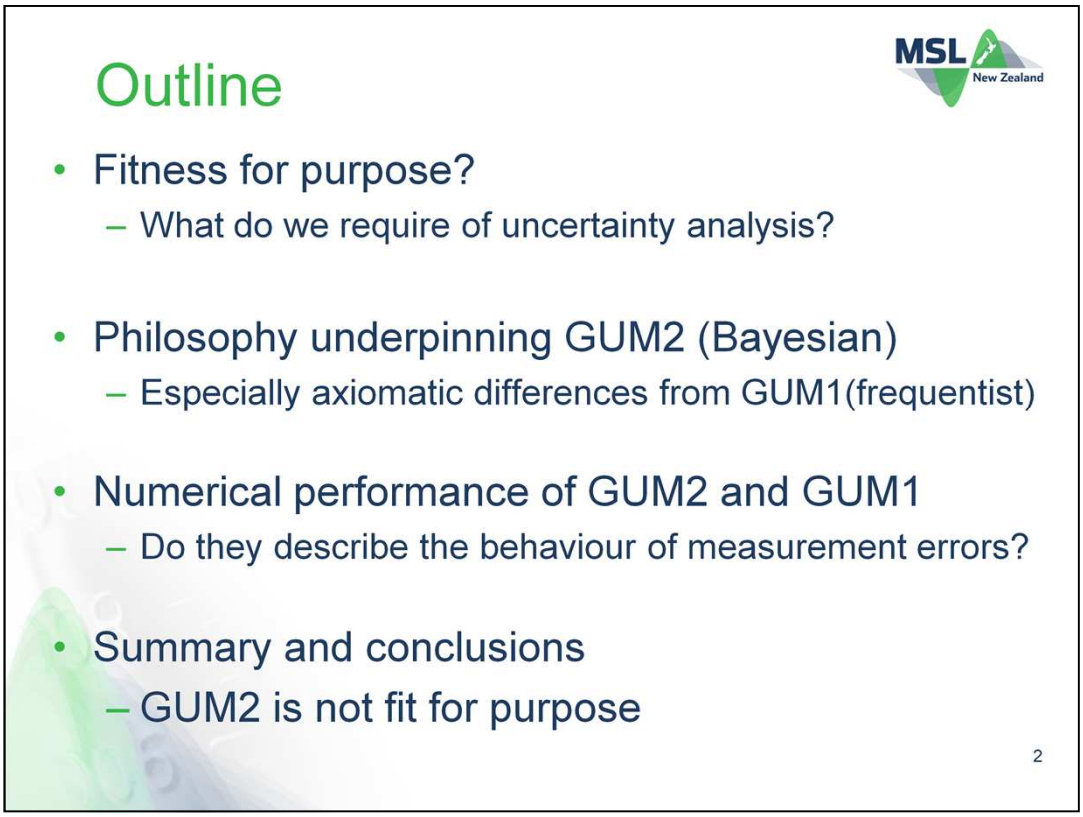
Thank you Mr Chairman… Good afternoon everyone…

Firstly, I would like to thank Walter Bich and his Working Group up for inviting me to present my concerns about aspects of the draft GUM revision. I'd also like to acknowledge the fantastic job the Working Group have done on the draft revision. I think it is exemplary in terms of its clarity and precision of its presentation. There is a lot I like about this revision.

However, I foresee a number of problems with the revision in its current form, most of which are only apparent if you know what to look for. One of the lesser aims of this presentation is to explain what lies beneath the waterline, so to speak.

Before I progress much further, I must say that I am not a statistician. I am a physicist, a metrologist. Until ten years ago, I had not heard of Bayesian statistics, and when I first saw GUM Supplement 1 (GUM-S1) and its reference to Monte Carlo, I thought it was a great idea. I expected that it would overcome all sorts of difficulties we have with the current GUM – things like non-linearity, and non-Gaussian and asymmetric distributions, and indeed GUM-S1 does allow us to address some of these problems. What I did not realise at the time was that GUM-S1 does not use Monte Carlo to simulate error processes, as I was expecting. It is purely for numerical integration, and the difference was something I did not understand. Very gradually over the last ten years, as I have come to understand something of the rationale for the GUM supplements and the draft revision, I have become increasingly uncomfortable with the application of Bayesian statistics to measurement uncertainty **as represented by GUM-S1, GUM-S2 and the draft revision** when applied to physical metrology. The differences between the approaches are subtle, but I think they have far-reaching consequences for uncertainty analysis.

I must also acknowledge colleagues Blair Hall and Rob Willink, and a bunch of others at MSL and on the CCT, for their help and many useful discussions. Also I'd like to thank a recent email contact, Dave LeBlond, a statistician who regularly uses Bayesian methods and has helped me understand some of the differences in the two approaches.



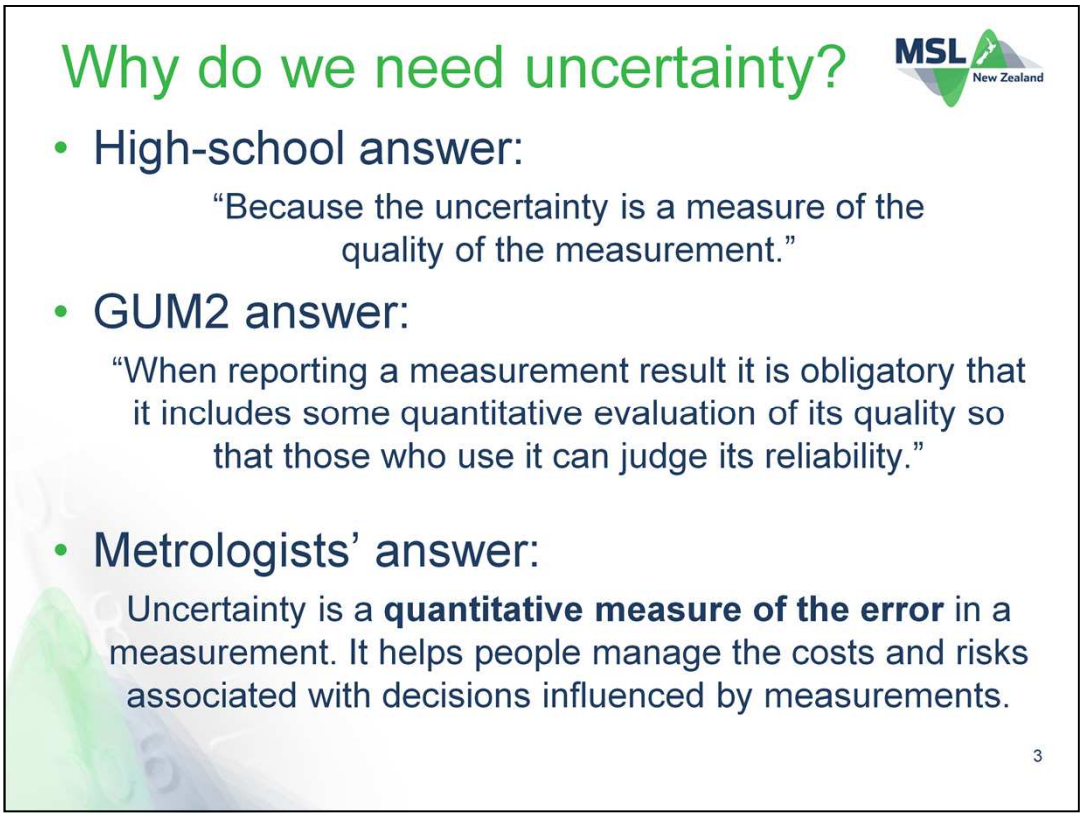
OK, so let me start with my conclusion. The draft GUM revision, which I will refer to as GUM2, is not fit for purpose.

Now, in order to explain why I think this is so, I first need to explain what I believe, from a physics perspective, makes an uncertainty analysis fit for purpose, so I will spend about 1/3 of my presentation looking at how uncertainty is used.

Most of the problems I foresee with the draft revision (GUM2) are a consequence of the philosophical differences between GUM2 and the current GUM (GUM1). So I'll spend another 1/3 of the presentation describing what I think are the major axiomatic and philosophical differences between GUM1 and GUM2. Because of time limitations, this will be more superficial than I would like, and I will have to assume you are already familiar with GUM1.

Finally, I will look at the differences in the numerical performance of GUM1 and GUM2. Although I have serious concerns about the philosophical approach underpinning GUM2, it is primarily the poor numerical performance of GUM2 (and in the accompanying supplements GUM-S1 and GUM-S2) in some situations that lead me to conclude that GUM2 is not fit for purpose.

.



Let's start at the beginning: why do we need uncertainty?

Like most of you I expect, my teacher at school told me that we had to calculate the uncertainty in our measurements because it provided an indication of the quality of the measurements.

I never found this answer satisfactory. The statement is true enough, but it does not tell us WHY we should calculate the uncertainty.

GUM2 gives us a slightly more extensive answer, but it's no better than the high-school answer – it still does not tell us why.

My answer, and what I would hope is every metrologist's answer, is: uncertainty is a quantitative measure of the error in a measurement. It helps people manage the costs and risks associated with decisions influenced by measurements [1].

Example – regulatory compliance

- Average Quantity System (OIML R 87)
- Rule 1: (applied to 200 g packet of biscuits)
  - The average mass of the contents must not be less than 200 g.
- Rule 2:
  - Not more than 2.5% of packages may contain less than 191 g.
- Rule 3:
  - No package may contain less than 182 g.

4

Let me give an example to explain what I mean.

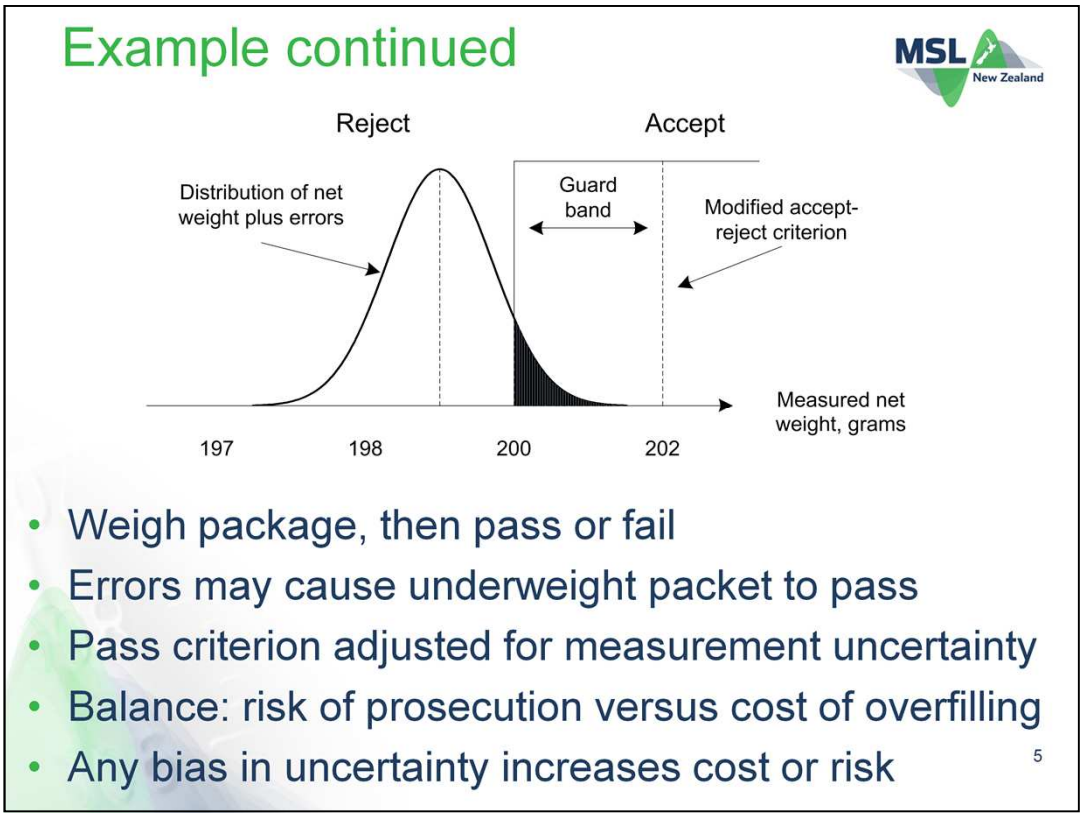
Over the road from our laboratory in Lower Hutt is a biscuit manufacturer, and this is one of the packets of biscuits he makes. Under New Zealand fair-trading laws, he must satisfy the requirements of OIML Recommendation 87, otherwise known as the Average Quantity System (AQS) [2]. I expect the same or similar legislation applies in most countries.

The AQS requires all 200 gram packets of foodstuffs to satisfy 3 requirements:

Rule 1: The net mass of the contents, averaged over many packets, must not be less than 200 g.

Rule 2: Not more than 2.5% of packets may contain less than 191 g (–4.5%)

Rule 3: No package may contain less than 182 g (–9%)



To meet these regulations, the biscuit maker weighs every packet as it leaves the production line. This weighing process is dynamic, so there are a lot of different errors that occur.

Let us suppose that a packet whose contents have a true mass of 199 g passes over the balance. Due to the various errors in the weighing process, the indicated mass may be anywhere between, say, 197 g and 201 g, as indicated by the normal probability density curve drawn on the graph.

This means that there is a possibility of the measured mass being 200 g or more (the shaded area), and the packet being approved for sale. If this happens too frequently, the biscuit maker risks prosecution, fines, and bad publicity.

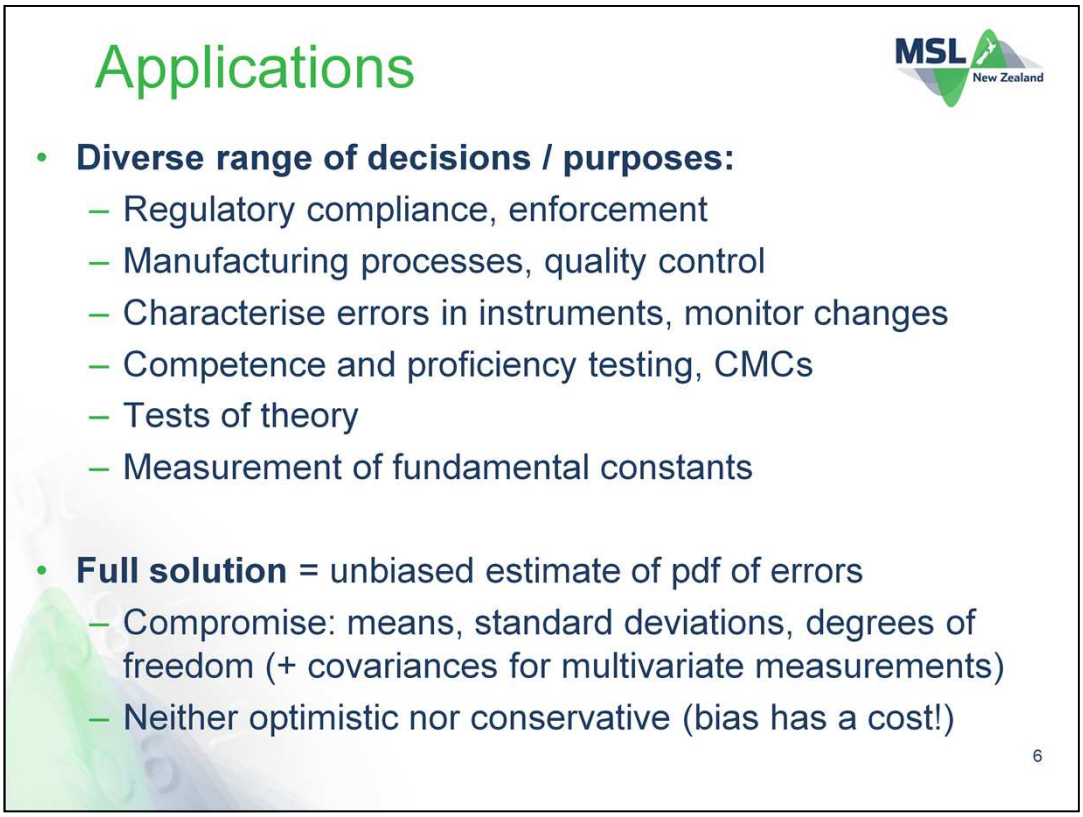
To reduce the risk of prosecution, he moves the pass/fail line to 202 g, say, to create a guard band, or tolerance band, so that on average he overfills each packet, but substantially reduces the risk of prosecution. Therefore, he trades the cost of overfilling against the risk of prosecution.

Now consider what happens when he invites his calibration consultant (me) to measure the errors in the balance and hence determine the uncertainty in the balance readings. If I give him a conservative uncertainty (too large), then he will make a very large tolerance band, and grossly overfill the packets for no good reason, essentially giving away biscuit for no return.

If I give him an optimistic uncertainty (too small), then he will make a small tolerance band. He will no longer have the gross over-filling problem, but I will be exposing him to an increased risk of prosecution.

The only way the biscuit maker can optimally adjust his quality-control processes is if I give him my best objective estimate of the uncertainty in the balance readings. Any bias in my estimate of the uncertainty will either cost him money due to overfilling or expose him to an increased risk of prosecution.

This is just one example of how measurements influence decisions.



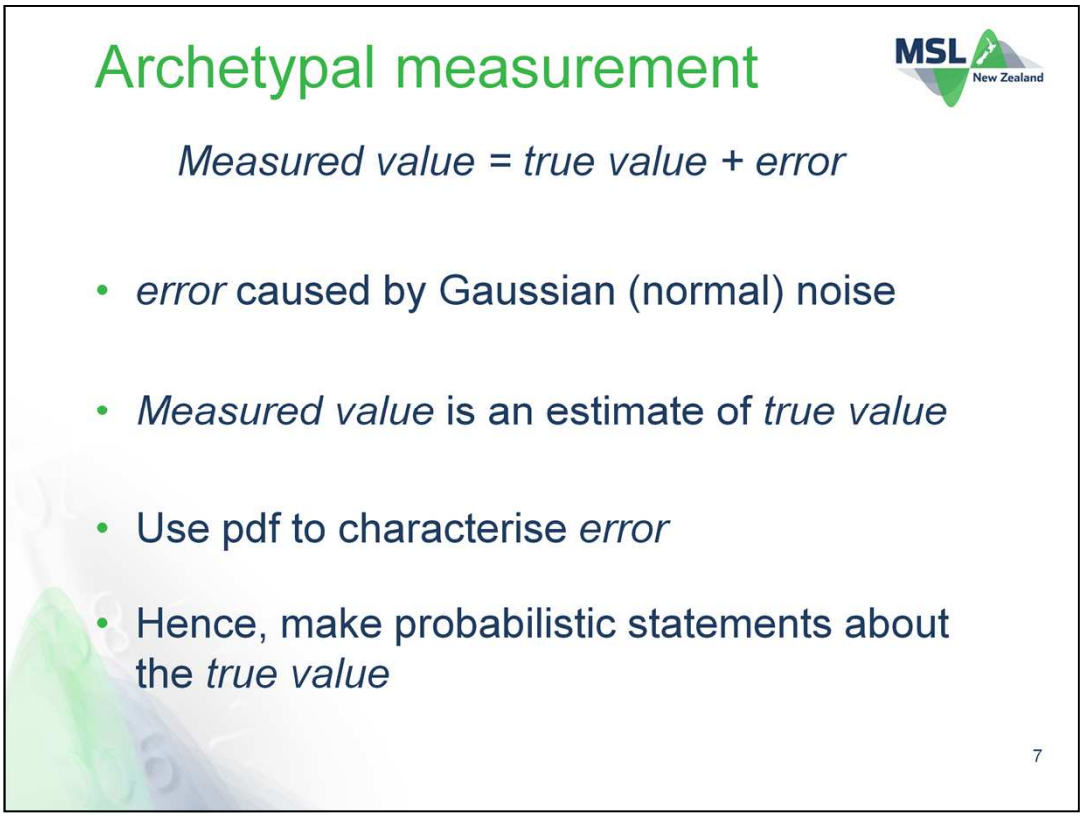
If you look at how other measurements are used you will find that the ultimate endpoint for practically all measurement chains is a decision. The decision may involve a single measurement, or many measurements, and often a lot of other information too.

The list gives examples of applications of measurements. In most cases, bias in either the measurement itself, or estimates of the uncertainty in the measurement, has a cost, much as with the biscuit maker.

These examples tell us that what is ideally required of an uncertainty statement, in most cases, is a best representation of the distribution of errors in the measurement. That is, the estimate of the uncertainty should be neither optimistic nor pessimistic/conservative. Of course, the cost of measuring errors in this much detail will be excessive in almost all cases and, therefore, we adopt a reasonable compromise…

We should provide:

- An estimate of the mean of the error distribution (so corrections can be made)
- An estimate of the standard deviation of the error pdf (standard uncertainty)
- Ideally, the number of measurements or the degrees of freedom, if the number of measurements is small
- Perhaps an expanded uncertainty (but this must be at a confidence level appropriate to the user's needs)
- Covariances, where multivariate measurements are made.



OK, before we look at statistical approaches for assessing measurement uncertainty, let's remind ourselves of what a measurement looks like.

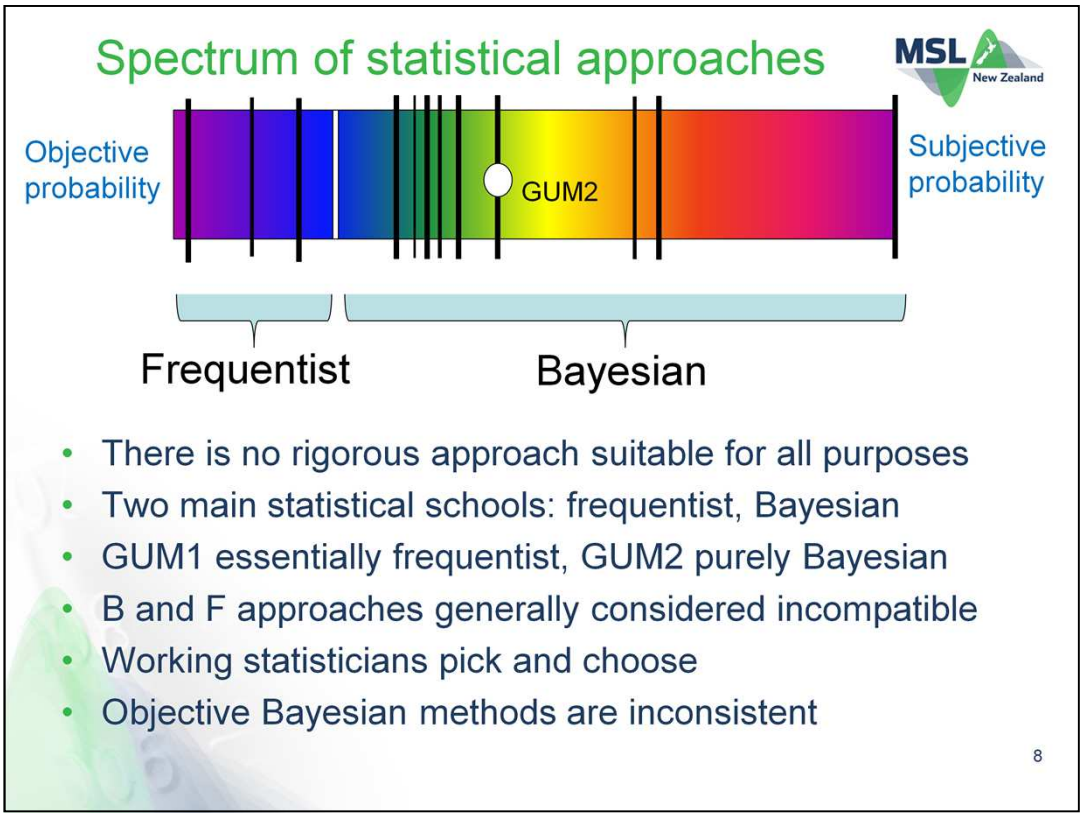
In almost all physical measurements, we have in mind (and this is practically true in almost all physical metrology) that there is a single true value for the measurand. The measurement is an estimate of the true value of the measurand, and is in error by some unknown amount.

In the simplest case, where the error is due to a random noise process, for example, we can make repeated measurements and build up a picture of the error distribution.

This allows us to determine the probability density function for the error, and from (i) the pdf of the error, (ii) the measurement model, and (iii) the measured (mean) value, we can then make probabilistic statements about the range of values within which the true value might lie.

The pattern set by this archetypical measurement can be extrapolated to practically all physical measurements, although the pdf of the error often cannot be determined by sampling (Type A) as here, and may require us to use a range of different methods (Type B). The development of a sufficiently detailed model of the measurements to determine the error pdf is the main job of the metrologist.

There is also a wide range of statistical methods we could use.



It turns out that there are many different statistical approaches/paradigms we could employ, including frequentist, Bayesian, fiducial, likelihood, … (+ others). For the purposes of our discussion the two most important are frequentist and Bayesian. But each of these is not just a single approach; they can both be split many ways. The graphic is my impression of these many approaches. At the LHS are approaches based on 'Objective' definitions of probability and, in particular, on the observed frequencies of random events (hence frequentist). At the RHS are approaches based on subjective probability (also known as 'degree of belief' or 'state of information'). I'll talk more about the difference in the two probability definitions shortly. This wide range of approaches has evolved because there is no single approach suitable for every application.

When you look closely at this spectrum of approaches, several problems emerge:

- Frequentist and Bayesian approaches are considered by many statisticians to be fundamentally incompatible (see [3 paragraph 1] discussion on the Likelihood Principle, for example). Because GUM1 is essentially frequentist and GUM2 is Bayesian, this means that GUM1 and GUM2 are not compatible. To my mind, backward compatibility for a revised GUM should be obligatory.
- Generally, frequentist methods are the most reliable, but are not widely applicable, while Bayesian methods are very widely applicable, but not always reliable. By adopting a Bayesian approach for ALL problems, we are exposing ourselves to problems that only an experienced statistician might recognise.
- In particular, ALL of the 'Objective Bayesian' methods crowding the middle of the spectrum above are notorious for inconsistencies, especially for multivariate and vector problems [4, 5]. Unfortunately, GUM2 is based on one of these inconsistent approaches and, as we will see shortly, there are some serious problems in GUM Supplement 2 as a result.
- Only the fully subjective Bayesian approach (extreme RHS) is mathematically sound and free of problems [5,6]. Unfortunately, most statisticians agree that a high level of statistical expertise is required for this approach.

Objective v subjective probability

Call before toss

- Coin in random state
- Uncertainty due to randomness
- **Frequentist** probability
- Objective, physical

Call after toss

- Coin in fixed but unknown state
- Uncertainty due to lack of information
- **Bayesian** probability
- Subjective, evidential, personal

9

To illustrate the difference between objective and subjective probability, consider the following experiments with coin tosses:

Experiment 1: (i) ask for the call, heads or tails, then (ii) toss the coin

Experiment 2: (i) toss the coin, then (ii) ask for the call

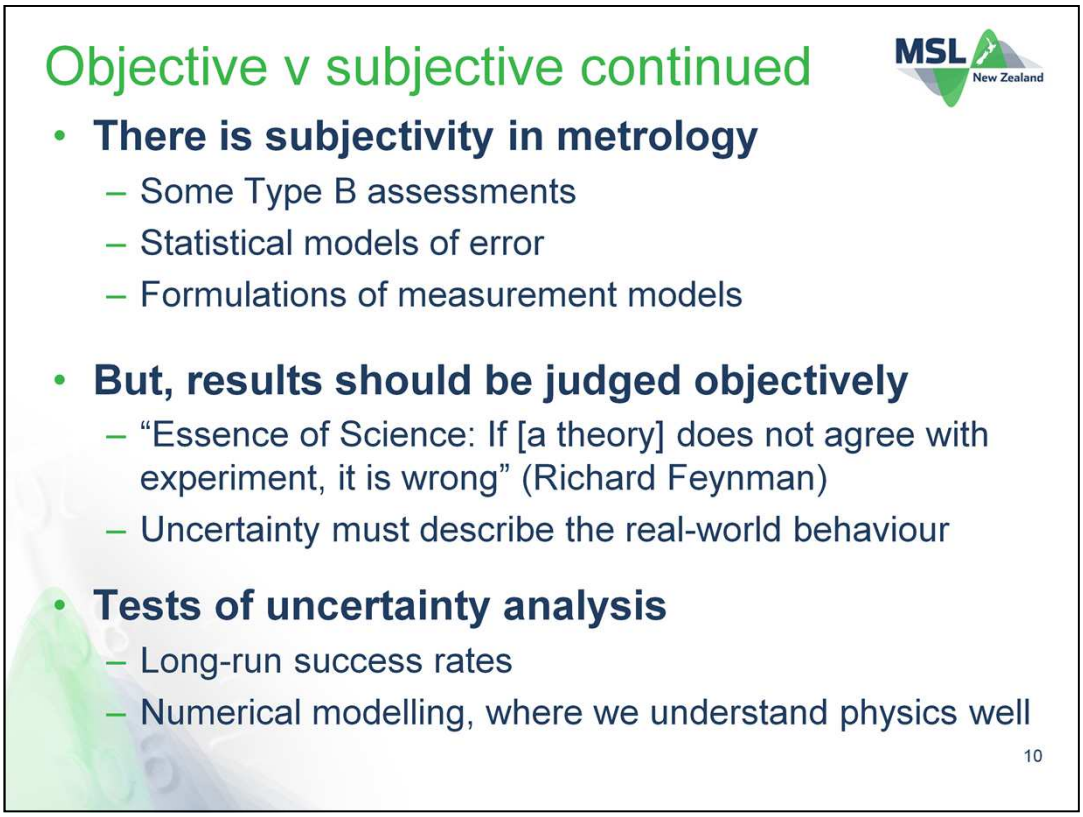
In the first experiment, the uncertainty about the state of the coin is due to the random flight of the coin during the toss. In the second experiment, at the time I ask for the call, the coin has already been tossed and is in a fixed non-random state. The uncertainty is now due to due to a lack of information about the state of the coin – there is no random behaviour.

The first experiment encapsulates the frequentist's view, where probability describes random behaviour and is measured as the long-run frequency of random events. In the long run, if we repeat step (ii) of the experiment, we expect 50% of the tosses of a fair coin to be heads. The second experiment encapsulates the Bayesian view, where probability describes the 'strength of belief' or 'state of knowledge' about a system. If we repeat step (ii) here (with different observers, and the same coin toss) we would get a 50:50 distribution of heads and tails, although the state of the coin does not change but remains unknown.

The experiments also show how the Bayesian view of probability has much wider applicability than the frequentist view. Strict frequentists can only ascribe a probability to the first experiment, while Bayesians can ascribe a probability to both experiments. A Bayesian statistician can also speak of the probability associated with one-off events, such as presidential elections, or whether someone has a specific disease, or the colour of a car responsible for a traffic accident – situations where a strict frequentist can offer no advice.

Although Bayesian methods have clear advantages, there are also worrying aspects from a scientific perspective: (i) in many cases, a Bayesian assessment cannot be verified or falsified; (ii) Bayesian methods cannot provide a credible way to summarise results independently of any previous beliefs or knowledge [6].

The most important point of the experiments here is that, right at the foundation, the meaning of probability is different for the two approaches. In many situations, like the tosses of coins or dice, the two approaches give the same numerical answer, but we must remember that the term 'probability' is used by the two approaches to measure two quite different quantities.



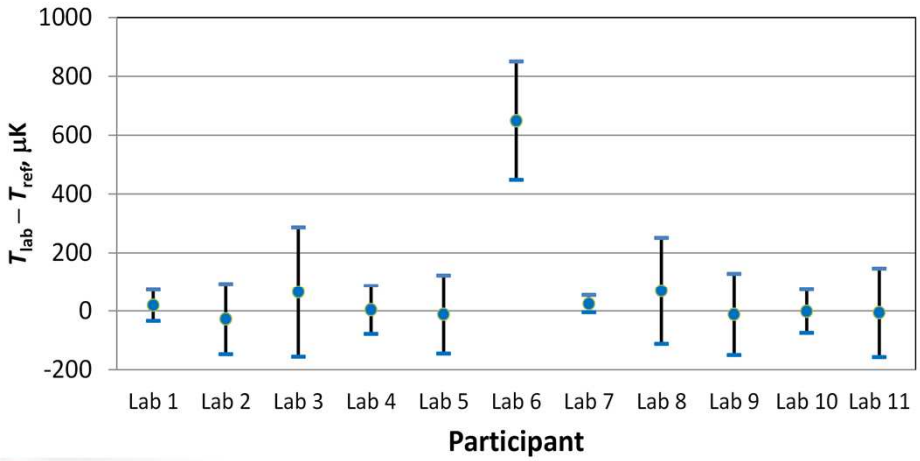
In general, scientists are very uncomfortable with all subjective aspects of science, but it is something we must accommodate. In metrology, subjectivity can enter an uncertainty analysis in several places: our choice of statistical model for the error processes (normal, rectangular, chi-square, etc), in the choice of mathematical model for the experiment (e.g., how many terms we include in a polynomial), and of course, most of our Type B uncertainty assessments have some degree of subjectivity, especially those reliant on experience and guesswork.

At the same time, we expect all of our models and theories to give a realistic description of how the experiments behave. That is, we must judge the performance of our uncertainty analyses objectively. I think this was captured very nicely by the Nobel prize winning physicist, Richard Feynman, who said "If [a theory] does not agree with experiment, it is wrong" [7]. This approach makes sure that uncertainty assessments are realistic and, for example, enable the biscuit maker to properly manage his quality control processes.

We should, therefore, be prepared to test our uncertainty analyses to confirm they give the results we expect. My colleague Blair Hall (who gives a presentation later at this meeting) is an advocate for numerical models [8]. He does not particularly care by what method an uncertainty analysis is derived, but it must work. He has built quite complex multivariate, 2-dimensional, non-linear simulations of his measurement processes to confirm that his uncertainty calculations give good values for standard uncertainties and coverage factors. He does this by repeating the simulations many times, and checking that the expanded uncertainties cover the modelled value of the measurand the requisite number of times; i.e., that the analysis has a good **long-term success rate**. In physical metrology we have the luxury of well-proven models for many measurement processes, so this is a viable option.

There are also other ways we can check our uncertainty analyses; in fact, we already do so.

Example: a thermometry comparison



- A test of analysis – have we assessed errors properly?
- Gross outliers have usually made a mistake
- Most labs are 'conservative' (uncertainty large)
- (Analysis complicated by correlation, unstable artefacts)

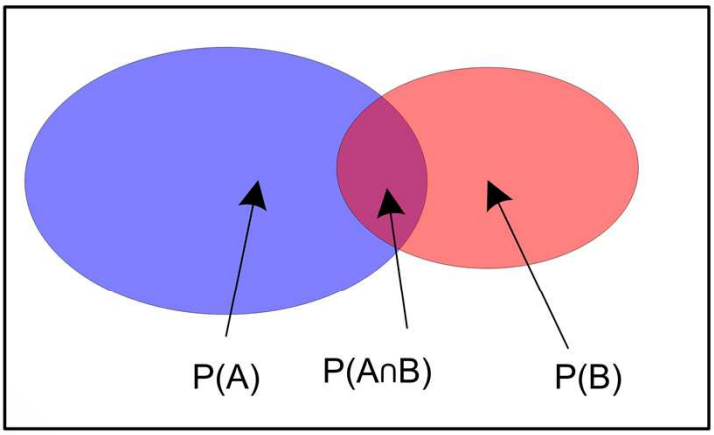
11

This graph presents the results of the most recent measurement comparison in which our lab has participated. (I've removed the laboratory labels to protect the guilty and impugn the innocent.) The uncertainty bars on the data points correspond to $k = 2$. This example illustrates several characteristics of most inter-laboratory comparisons:

- The outliers in most comparisons are very obvious. The anomalous results are usually caused by genuine experimental mistakes (often as simple as an incorrect sign in a spreadsheet) rather than extreme statistical fluctuations.
- The results of the remaining labs (non-outliers) are often, as here, clustered very close to the reference value (corresponding to 0 on this graph). In this case, the clustering is greater than the uncertainty bars would indicate. If the bars were halved in length ($k = 1$), then all but the lab with the least uncertainty would still enclose the reference value. This suggests that most uncertainty analyses here are conservative.
- Even if the uncertainty analyses are subjective and untestable in isolation, the long-term success rates of the algorithms used for standard uncertainty and coverage factors can be evaluated by comparing the results from similar experiments.

It is important to note here that comparison results are generally not simple to evaluate. The reason for clustering may be a highly-correlated uncertainty for a systematic effect shared by all participants. In contrast, some comparisons show an unexpectedly high dispersion of results due to unstable artefacts.

Bayesian paradigm, Bayes' rule



- Bayes' rule: $\mathrm{P}(A \mid B)\,\mathrm{P}(B) = \mathrm{P}(B \mid A)\,\mathrm{P}(A)$
- Used in frequentist and Bayesian statistics

12

Now, before I explain Bayesian statistics, let us review Bayes' Rule.

Suppose we have two different events, A and B. The areas of the two ellipses represent the frequencies of the events A and B (from the universe of events – the white area).

The area of overlap between the two ellipses represents the probability of both events occurring. Now consider the overlapping area in more detail – there are two ways of looking at it. We could say that the shared area represents the probability of event A given that B has already occurred, or we could say that it is the probability of event B given that A has already occurred. In either case, the resulting probability must be the same:

$$P(A|B)P(B) = P(B|A)P(A),$$

where, *P*(A) and *P*(B) are the probability of A occurring and the probability of B occurring, respectively, and *P*(A|B), means the probability of A occurring given that B has occurred. This is described as a conditional probability.

This is Bayes' Rule. The explanation I have presented is in frequentist terms; however, this is a very powerful theorem used by all statisticians.

Bayesian statistics

- Generalise Bayes' rule from single probability to the whole distribution:

$$P(\mathbf{a} \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid \mathbf{a})P(\mathbf{a})}{\int P(\mathbf{X} \mid \mathbf{a})P(\mathbf{a})\mathrm{d}\mathbf{a}}$$

where:

- $P(\mathbf{a})$ is the 'prior' distribution for the measurands $\mathbf{a}$
- $P(\mathbf{a} \mid \mathbf{X})$ is the 'posterior' distribution for $\mathbf{a}$
- $P(\mathbf{X} \mid \mathbf{a})$ is the likelihood function for the measurements $\mathbf{X}$
- Integral normalises the posterior distribution so that it has unity area

13

Bayesian statistics extends Bayes' Rule from probabilities to pdfs. You can recognise the remnants of Bayes' Rule in the LHS and the numerator of the RHS of this equation. Suppose we start an experiment with some knowledge of the value of a quantity $\mathbf{a}$, and we represent our state of knowledge of $\mathbf{a}$ by a pdf, $P(\mathbf{a})$. We then perform a bunch of measurements, $\mathbf{X}$, and then ask the question: how has our state of knowledge changed as a result of the measurements? The answer is given by this version of Bayes' Rule:

$$P(\mathbf{a} \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid \mathbf{a})P(\mathbf{a})}{\int P(\mathbf{X} \mid \mathbf{a})P(\mathbf{a})\mathrm{d}\mathbf{a}}$$

The state of knowledge before the measurements is represented by the 'prior distribution' $P(\mathbf{a})$, and the state of knowledge after the measurements is represented by the 'posterior distribution' $P(\mathbf{a}|\mathbf{X})$.

The function $P(\mathbf{X}|\mathbf{a})$ is called the likelihood function, and I'll talk more about it shortly. As the number of data points increases, this function tends to get very narrow and very small. The integral in the denominator normalises the area of the posterior pdf so that it has unity area.

One of the major problems with Bayesian statistics is that, in general, the integral in the denominator cannot be solved analytically and must be solved numerically. Unfortunately, the numerical calculations can be so computationally expensive that until the last couple of decades, many Bayesian calculations were impracticable. This situation has changed with the development of new integration algorithms and ready access to fast computers. Although in general the equation must be solved numerically, there are particular instances where analytic solutions exist [9]. (See also 'conjugate prior' distributions in Wikipedia. **[9] is a also good reference for anyone wanting to know more about Bayesian methods**.)

Likelihood function…

$$P(\mathbf{X} \mid \mathbf{a}) = L(x_1, x_2, \ldots : \mu, \sigma) = \prod_{i=1}^{n} p(x_i)$$

- Product of pdfs, e.g., for x_i drawn from $N(\mu, \sigma)$

$$L(x_1, x_2, \ldots : \mu, \sigma) = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \prod_{i=1}^{n} \exp\left(\frac{1}{2\sigma^2} (x_i - \mu)^2 \right)$$

- Maximum Likelihood Method:
  ➔ unbiased estimates of μ and σ

$$\hat{\mu} = \frac{1}{n} \sum x_i, \quad \hat{\sigma}^2 = \frac{1}{n-1} \sum (x_i - \hat{\mu})^2$$

14

OK, now let's just flashback for a minute to remind ourselves of the likelihood function (often just called the likelihood). For our purposes, it is simply the product of the pdfs for the data, based on an assumed statistical model of the data. For example, if we assume that our measurements are random and drawn from a normal distribution with mean μ and standard deviation σ, then the likelihood is simply the product of the pdfs with the different values for the measurement results substituted into the equations. Note, in this metrological problem, μ is the unknown quantity we are trying to measure, and σ is the unknown standard deviation of the errors in the measurement.

It's worthwhile taking a short detour here, and looking at the maximum likelihood method. Having written down the expression for the likelihood, one of the things we can do is differentiate it with respect to both μ and σ, and find the values of μ and σ that maximise the likelihood. If we do that, we obtain the classical estimates:

$$\hat{\mu} = \frac{1}{n} \sum x_i, \quad \hat{\sigma}^2 = \frac{1}{n-1} \sum (x_i - \hat{u})^2$$

Maximum likelihood continued

- Approach requires a change in perspective:
  - The data x_i are fixed
  - The measurands μ, σ are variable
  - These are features of all Bayesian statistics
  - Bayesian models incompatible with physics models
- Example: rearrange likelihood function…

$$L(x_1, x_2, \ldots : \mu, \sigma) = F(x_i) \times \exp\left(\frac{n}{2\sigma^2} (\hat{\mu} - \mu)^2 \right)$$

- "μ is distributed about $\hat{\mu}$ with variance σ^2/n"
- Uncertainty in μ given by likelihood function

15

However, by analysing the likelihood in this way, we've done something odd. We have assumed the data is fixed and the measurands μ and σ are variables. These are characteristics of Bayesian analysis. In fact, it is often said that the maximum likelihood method is really a Bayesian method.

There is a second thing we can do with the likelihood function that also gives an impression of the Bayesian thought processes. It turns out that the likelihood can be rearranged so that σ occurs only within a Gaussian function, and all of the other parts of the function (i.e., the data points) form multiplying constants, so that

$$L(x_1, x_2, \ldots : \mu, \sigma) = F(x_i) \times \exp\left(\frac{n}{2\sigma^2} (\hat{\mu} - \mu)^2 \right)$$

The Bayesian interpretation of this function is that μ is distributed about $\hat{\mu}$ (which is fixed because the data is fixed) with variance σ^2/n [9].

This difference in perspective is fundamentally incompatible with the physics model (slide 6) of a measurement that assumes the measurand is fixed. This difference is also reflected in the different meanings assigned to frequentist and Bayesian confidence and credible intervals (expanded uncertainties):

Frequentist/physical: long-term success rate determined from different interval estimations (i.e., repeated experiments, different data), given the same measurand.

Bayesian: long-term success rate determined from different measurands drawn from the prior distribution, given the same interval (i.e., same data).

These differences in interpretation are sometimes used to explain the differences in the numerical values of confidence and credible intervals [10]. However, in both types of analysis we should expect the long-term success rate averaged over many uncertainty assessments to be the same and equal to 95% (or whatever). Often, the priors are chosen so that this is the case, but a different prior is required for every class of problem.

Bayesian statistics continued

- Posterior distribution given by

$$P(\mu,\sigma : x_1, x_2, ...) = \text{const} \times L(x_1, x_2, ... : \mu,\sigma) P(\mu,\sigma)$$

- Uncertainty is described by posterior distribution

- Many Bayesian approaches – use different priors
- Objective Bayesian approaches use priors to represent zero information (ignorance)
  - Many variations
  - E,g., flat prior (Laplace)
  - **E.g., Jeffreys priors (1 instance used by GUM2)**

16

OK, so let us return to Bayes' Rule. We have:

posterior = likelihood x prior

The distinction between the many different varieties of Bayesian statistics arises from the different choices for the prior distribution. In particular, Objective Bayesian analyses use priors that attempt to characterise or approximate zero prior information [9, 11]. For example, for continuous variables, the so-called flat prior (an approach attributed to Laplace) gives the same result as the maximum-likelihood approach. However, there is a source of confusion: with non-linear systems (e.g., temperature measurement with a thermistor) should the prior be flat with respect to current through the thermistor, the resistance, or temperature? – it's not clear.

Jeffreys developed a method for determining priors that have an invariance property (see [9]) with respect to these types of transformations, and these Jeffreys priors are very commonly used in Objective Bayesian analysis.

However, although there are many different Jeffreys priors, GUM2 (and GUM-S1) are based on the use of one specific Jeffreys prior for all applications.

Jeffreys prior: univariate measurement

- Assume: *n* measurements from normal distribution
- $-\infty < \mu < +\infty, \quad 0 < \sigma < +\infty$
- ➔ Posterior distribution for μ is a *t*-distribution:

$$p(\mu \mid \mathbf{X}) \propto \left[1 + \frac{n(m-\mu)^2}{(n-1)s^2}\right]^{-N/2}, \quad \mathrm{Var}(\mu) = \frac{n-1}{n-3}s^2$$

- Use area under distribution to find 95% credible interval:

$$\left[m - t_{0.975,n-1}\, s/\sqrt{n}, \quad m + t_{0.975,n-1}\, s/\sqrt{n}\right]$$

- GUM2 gives same interval as for GUM1!
- But: variance is not an estimate of the variance of the error!
- Also, not appropriate for $0 \le \mu < +\infty$ (e.g., temperature, mass, impedance….)

17

The Jeffreys prior used by GUM-S1 and GUM2 is one of the most commonly used, and has a very appealing property. When the data are assumed to be dispersed according to a normal distribution (as we assumed for the likelihood a few slides ago), the posterior distribution is a shifted and scaled *t*-distribution that yields credible intervals identical to the corresponding frequentist 'confidence intervals'.

The Jeffreys prior used in GUM2 is strictly only appropriate for quantities that may take any positive or negative value. If quantities are restricted in value (e.g., temperature, mass, and impedance are always positive) then a different Jeffreys prior should be used, and the resulting posterior distributions are not so convenient.

Note that the posterior *t*-distribution is not an estimate of the pdf of error in the measurement (which we assume to be normal). It represents the uncertainty in the state of knowledge about μ. For the same reason, the variance of the posterior distribution relates to the uncertainty in μ, and is not an estimate of the variance of the error, s^2, as was the case with a GUM1 analysis.

This observation highlights the second axiomatic change between GUM1 and GUM2: the uncertainty is no longer a quantitative measure of the error, but is instead a measure of the 'state of information' about the measurand.

Like all objective Bayesian approaches, there are problems [4, 5, 6,12]. The Jeffreys prior is an improper distribution, which leads to undesirable properties in the posterior. The posteriors for GUM2 with small numbers of measurements have the peculiar property that the standard uncertainties are infinite or undefined, while the expanded uncertainties are finite and well behaved. The problem with the standard uncertainties is evident from the $n - 3$ factor in the equation for the variance above.

As indicated already, this specific Jeffreys prior has been chosen for GUM2 purely because it gives the same answer as the GUM1 analysis for single-variable problems. However, for other problems, the result is only an approximation to the behaviour that would be observed in the lab.

Jeffreys prior: multivariate measurement

- **Uncertainty in $X + Y$?**
  - Measurements x_i and y_j normally distributed
- **GUM1 (frequentist) approach:**

$$u^2_{X+Y} = s^2_x / n_x + s^2_y / n_y$$

  - Unbiased estimate of the variance of the error in $X + Y$
- **GUM2 (Objective Bayesian) approach:**
  - Convolution of the posteriors for X and Y, hence

$$u^2_{X+Y} = \left(\frac{n_x - 1}{n_x - 3}\right)\frac{s^2_x}{n_x} + \left(\frac{n_y - 1}{n_y - 3}\right)\frac{s^2_y}{n_y}$$

  - Does not describe the error in $X + Y$!
  - Uncertainties always 'conservative' (biased)

18

Let us consider a very simple multivariate problem: calculating the uncertainty in $X + Y$ where small numbers of measurements of X and Y have been recorded. This is an intriguing problem closely related to the famous Behrens–Fisher problem [13, see also Wikipedia]. Although the problem seems very simple, there is no simple algebraic expression for the confidence intervals.

First, let's remind ourselves of the standard uncertainty obtained with GUM1. The variance in the estimate of $X + Y$ is simply the sum of the variance in the estimate of X and the variance in the estimate of Y. This variance is an unbiased estimate of the variance of the error in the measurement of $X + Y$.

The GUM2 approach assumes that the states of information are additive, which means that the state of information for $X + Y$ is obtained as the convolution of the two t-distributions for X and Y, which usually have different degrees of freedom. When distributions are convolved, the variances add linearly, and hence the GUM2 variance in $X + Y$ is simply the sum of the variances as calculated on the previous slide. (GUM-S1 computes the convolution numerically by Monte Carlo, by drawing samples from the two t-distributions and adding.)

There are no real surprises here; once again, the GUM2 uncertainty describes the uncertainty in state of knowledge of $X + Y$ rather than the error in the measurement of $X + Y$.

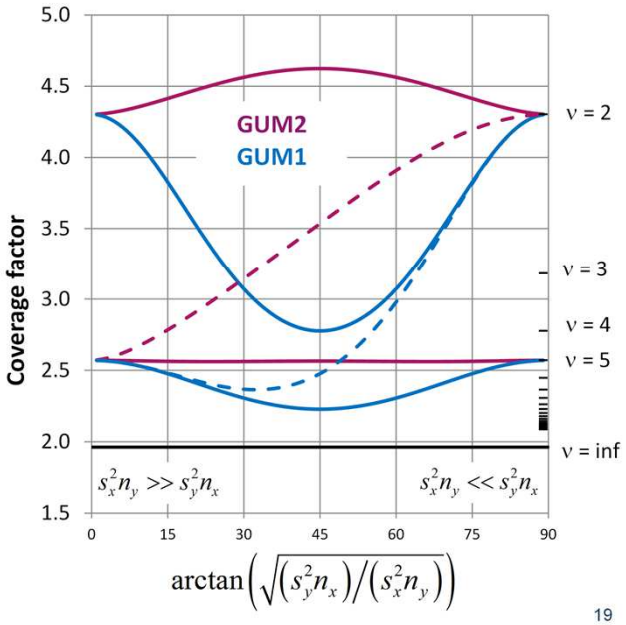
But there is a problem when we look at expanded uncertainties in $X + Y$…

Jeffreys prior: multivariate coverage

- *k* for 95% coverage for expanded uncertainty

$$U_{X+Y} = k\sqrt{s_x^2/n_x + s_y^2/n_y}$$

- Both GUM1 and GUM2 are approximations
- GUM2 always large
- GUM1 always closer to physics – mid points on GUM1 curves are exact
- Differences greatest for small *n*



19

This graph plots the coverage factor *k* for the two approaches, expressed as a multiple of

$$u_{X+Y} = \sqrt{s_x^2/n_x + s_y^2/n_y}$$

Because the basic problem has no simple solution, both of the GUM1 and GUM2 approaches are approximations. GUM1 uses the Welch-Satterthwaite approximation to derive a *t*-distribution from which a confidence interval can be derived. GUM2 convolves the pdfs for the states of information to obtain a single posterior from which credible intervals can be determined.

The horizontal axis plots the arctan of the ratio of the contributing uncertainties, $(s_y^2/n_y)/(s_x^2/n_x)$ The arctan function is simply a means to compress the horizontal scale from 0 to ∞. Thus at the LH end of the axis the uncertainty due to errors in the *Y* measurements is negligible, and at the RH axis the uncertainty due to errors in the *X* measurements is negligible. At the centre (arctan = 45) the two contributions are the same. The RH vertical axis plots the equivalent degrees of freedom according to the Welch-Satterthwaite approximation.

This graph plots coverage factors for three different experiments, where different numbers of measurements are made. Note that the points where arctan = 45 are known to be exact for GUM1. This shows that the GUM2 solution is a poorer approximation to the correct solution than the GUM1 solution, and always conservative (biased), especially where we need an approximation to work well. The poorer performance of GUM2 is expected as GUM2 is a first-order approximation and GUM1 a second-order approximation [13]. It follows that GUM-S1 should not be advanced as a reference method (as it is in GUM2), to which other methods of calculating uncertainty must be compared in order to claim satisfactory performance.

Note that two of the sets of curves include experiments where $n_x = 3$ or $n_y = 3$, for which the standard uncertainties for GUM2 are not defined, yet the expanded uncertainties can be calculated as shown. Experiments involving just 2 measurements ($\nu = 1$) are very common where the presence or absence of an effect is determined.

Jeffreys prior: vector measurement

Input vector → Output *m*-vector

$$\begin{pmatrix} X \\ Y \\ t \\ \vdots \end{pmatrix} \longrightarrow \begin{pmatrix} X+Y \\ X-Y \\ t \\ \vdots \end{pmatrix}$$

Same problem

$$u^2_{X+Y} = \left(\frac{n_x - m}{n_x - m - 2}\right)\frac{s_x^2}{n_x} + \left(\frac{n_y - m}{n_y - m - 2}\right)\frac{s_y^2}{n_y}$$

Different result

- Uncertainty affected by number of output functions!
- See GUM1 (H.2) and GUM2 example (Sec. 5)
- Such inconsistencies appear well known

20

The mathematical problems with the improper priors used in Objective Bayesian analyses get very much worse with multi-dimensional measurements [12]. GUM Supplement 2 uses a prior that yields a multidimensional *t*-distribution posterior with a variance that depends on *m*, the number of elements in the output vector. If we suppose that the first element in the output vector is $X + Y$, the same function as in the previous example, then the variance in $X + Y$ is

$$u^2_{X+Y} = \left(\frac{n_x - m}{n_x - m - 2}\right)\frac{s_x^2}{n_x} + \left(\frac{n_y - m}{n_y - m - 2}\right)\frac{s_y^2}{n_y}$$

So now the uncertainty in $X + Y$ can depend on factors that are unrelated to the calculation of $X + Y$. For example, if we chose to calculate $X – Y$, the scale factor in the equations changes from $(n-1)/(n-3)$ to $(n-2)/(n-4)$ And if I chose to record the time I make the measurement, then the uncertainty increases further to $(n-3)/(n-5)$, and now I must make 6 measurements for the standard uncertainty to be calculable. These bizarre problems make GUM Supplement 2 inappropriate for any uncertainty calculation.

Note: this example may seem contrived, but it parallels very closely the example in Sec. 5 of the GUM2 examples document. The results should also be compared to the original example in GUM 1 Sec. H.2, which were based on 5 measurements and yielded a realistic uncertainty more than a factor of 2 smaller than the GUM2 example.

Summary

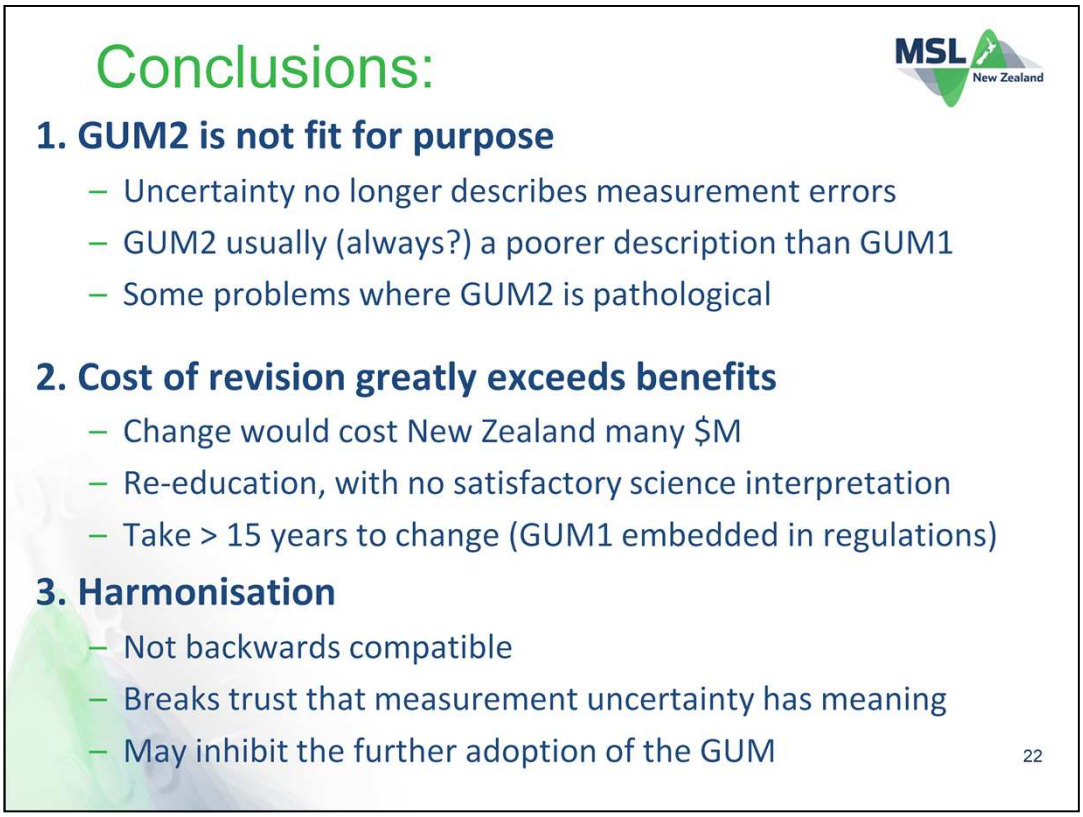
- GUM2 changes the definition of probability
  - OK, but we do expect it to describe real measurements
- GUM2 changes the definition of uncertainty
  - No longer a quantitative estimate of errors
  - Now 'state of information' (not observable) ???
- GUM2 performance
  - Same as GUM1 for univariate, and *n* very large
  - Not as good as GUM1 for multivariate and *n* small
  - Often biased (always conservative?)
  - Several pathological behaviours
  - No example where GUM2 is better than GUM1 (?)
    (better = closer to the real-world behaviour)

21

(Most of this slide is self explanatory.)

Early on this presentation, when I was talking about fitness for purpose, I concluded that uncertainty should be an unbiased measure of the errors in a measurement. A few of us at MSL have been comparing GUM1 and GUM2 approaches to calculations of measurement uncertainty. What we have found is that for many applications there is practically little difference.

However, there are a few instances where they give very different results. Where they differ we have performed numerical experiments to determine which is closest to describing real world behaviour. So far, we have found that GUM1 always gives better results than GUM2. See also [8] and [14].



(This slide is self explanatory.)

A way forward:

Considering:

(i) that the vast majority of GUM users have a minimal mathematical education;

(ii) that there is a large number of measurement processes for which a conventional (GUM1) analysis is perfectly satisfactory and accurate;

(iii) that the primary purpose of the guide is harmonisation of uncertainty practice outside the research lab (it should not, for example, constrain research into statistical methods);

(iv) that Bayesian methods cannot be fashioned into a cookbook approach (most statisticians agree that a high level of experience and expertise is required to avoid problems with Bayesian analysis);

(v) that most scientific results are published as summaries, separately from prior beliefs and knowledge;

(vi) that the decisions, cost functions, and prior knowledge is often made/held by the client and is inaccessible to the metrologist;

then:

- the GUM should be redrafted, in a simpler, clearer, and more accessible form based on the current GUM1 methodology; advice can be included, e.g., when uncertainty due to model is required, or prior information must be included, for the need for more sophisticated (Bayesian) analysis;
- Supplements can be used to guide on more difficult topics, as proposed;
- GUM-S1 should not be advanced as a reference method, but an approximation to use when GUM1 is inappropriate (e.g., dominant error distributions not Gaussian);
- GUM-S2 should be scrapped.

References

[1] D R White, “The meaning of Measurement in Metrology”, *Accred. Qual. Assur*., **16**, 31–41, 2011.

[2] ORGANISATION INTERNATIONALE DE MÉTROLOGIE LÉGALE, International Recommendation R 87, “Quantity of product in prepackages”, 2004.

[3] J O Berger and R L Wolpert, Lecture Notes-Monograph Series Vol 6: “The Likelihood Principle”, 2nd Ed. (Institute of Mathematical Statistics, Hayward, California), 1988.

[4] F James, *Statistical Methods in Experimental Physics*: 2nd Ed. (World Scientific, Toh Tuck Link, Singapore), 2006.

[5] R E Kass and L Wasserman, “The selection of prior distributions by formal rules”, *J. Am. Stat. Ass*., **91**, 1343–70, 1996.

[6] F James, “Principles and Problems in Statistics” (Training presentation, private communication), 2000.

[7] See YouTube video (several), search “The essence of science” Richard Feynman, e.g., https://www.youtube.com/watch?v=5v8habYTfHU.

[8] B D Hall, “Evaluating methods of calculating measurement uncertainty”, *Metrologia*, **45**, L5–L8, 2008.

[9] W M Bolstad, *Introduction to Bayesian Statistics*, 2nd Edition (John Wiley & Sons, Hoboken New Jersey), 2007.

[10] G Mana and C Palmisano, “Interval estimations in metrology”, *Metrologia*, **51**, 191–196, 2014.

[11] J O Berger, “The Case for Objective Bayesian Analysis”, *Bayesian Analysis*, **1**, 385–402, 2006.

[12] A R Syversveen, *Noninformative Bayesian priors: Interpretation and problems with construction and applications*, Preprint Statistics 3, Department of Mathematical Sciences, NTNU, Trondheim,1998.

[13] M Ghosh and Y Kim, “The Behrens–Fisher problem revisited: A Bayes–frequentist synthesis”, *Canad. J. Statist*., **29**, 5–17, 2001.

[14] F Attivissimo, N Giaquinto, and M Savino, “A Bayesian paradox and its impact on the GUM approach to uncertainty”, *Measurement*, **45**, 2194–2202, 2012.