Talk presented to CCT workshop 21 May 2014

Good morning everyone… Thanks to Walter Bich for the previous talk explaining the rationale for the GUM revision.

As you will know from my previous comments on the GUM revision, there is a sense in which Bayesian and frequentist statisticians have been at war for the last 100 years or so. It's not my intention to sustain that war but to take the opportunity to explain what I have learned about the differences between Bayesian and frequentist statistics and to identify a number of areas of concern that I have about the proposed Bayesian revision of the GUM.

I should add that I am not an expert in statistics, and for that reason I have found it difficult to find reliable information on the differences between the two approaches, and I'm quite sure I've more to learn. This has been made difficult, perhaps, because of the 'warfare': there is quite a bit of misinformation about both the Bayesian and frequentist approaches, including in some of the commentary published about the GUM revision. I'll do my best to explain some of these points as I progress.

Another of the reasons for my difficulty in coming to grips with the differences between the two approaches is, I think, that although they use the same terminology and share some mathematics, they are fundamentally different – as different as classical mechanics and mechanical engineering. In fact, I have come to think of Bayesian statistics as an offshoot of information or communications theory. As I hope you will see, even the basic words such as 'random' and 'probability' have different meanings. Also, it seems to me that most metrologists misunderstand key aspects of frequentist statistics, which is the foundation of the current GUM, so the starting point for understanding the differences is not good.

Before I go any further, I must acknowledge colleagues Blair Hall and Rob Willink, for their help and many useful discussions. Also I'd like to thank a recent email contact Dave LeBlond, a Bayesian statistician who has also helped explain some of the differences in the two approaches.

Outline:

- A simple statistical experiment (coin toss)
- Confidence intervals (frequentist)
- Bayes' rule and Bayesian statistics
- Credible intervals (Bayesian)
- Difference in meaning
- Differences in performance
- Other differences

2

I'll start with a simple experiment, the tossing of a coin, which I hope will give some insights into the differences in the foundations of the two approaches.

I'll then move on to explain the frequentist concept of a confidence interval. This explanation is necessary because as practising scientists most of us do not have the correct understanding.

I'll then move on to explain the Bayesian credible interval. I'll include a short review of Bayes' rule on which Bayesian statistics is founded.

Once we have the basic foundations clear, I'll spend a little bit of time explaining two of the key differences in the approaches in the context of metrology.

Spot the difference: coin toss

3

OK – let's look at a very simple statistical experiment – a coin toss.

Demonstration 1:

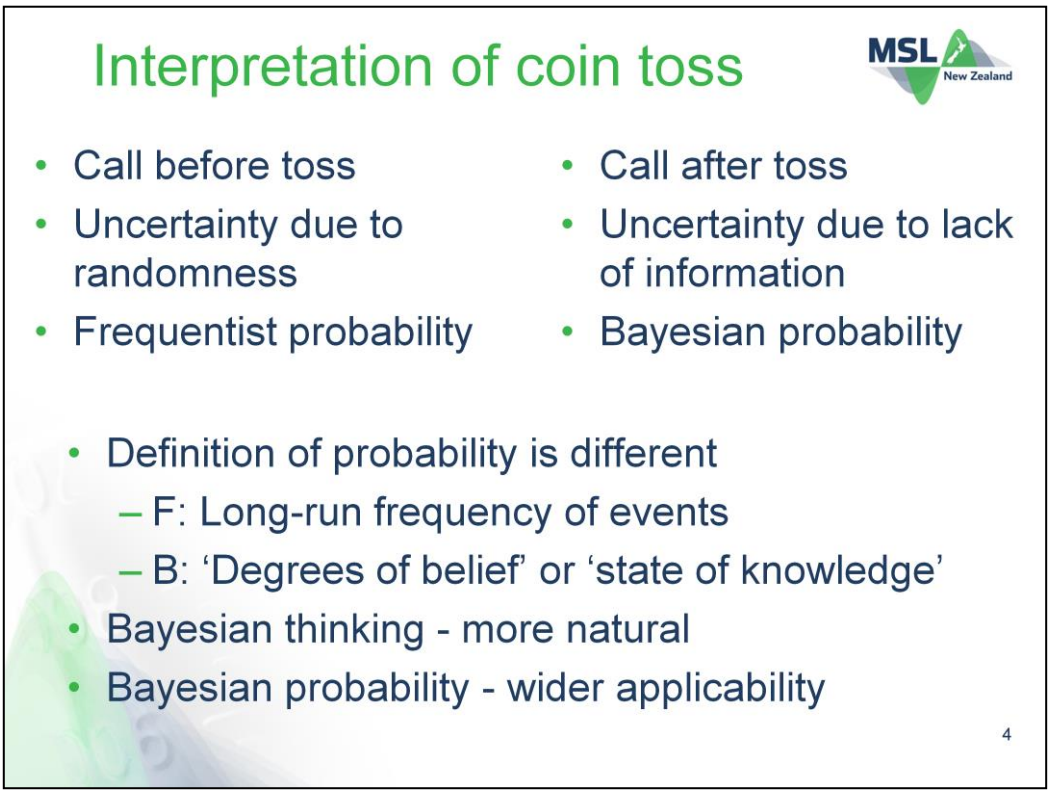
- Ask someone to call 'heads' or 'tails', then toss the coin.
- Repeat a couple of times with different people.
- Also ask, 'what is the probability that the outcome is a head'.

Most people will suggest that the probability of a head is 50%.

Demonstration 2:

- Toss the coin. With the tossed coin obscured under the hand, ask someone to call 'heads' or 'tails'.
- Repeat a couple of times with different people.
- Also ask 'what is the probability that the outcome is a head'.

Again, most people suggest that the probability of a head is 50%.



Hopefully, you will have all noticed that there was a subtle difference between the two experiments. In the first experiment, I asked for the call before I tossed the coin, whereas in the second experiment, I asked for the call after I tossed the coin.

In the first experiment, the uncertainty about the state of the coin is due to the random behaviour of the coin during the toss. In the second experiment, the coin has already been tossed and is in a fixed non-random state. The uncertainty is not due to the random behaviour of the coin but due to a lack of information about the state of the coin.

The first experiment encapsulates the frequentist's view, where probability describes the random behaviour of things and is measured as the long-run frequency of random events. In the long run, 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. Strict frequentists can only ascribe a probability to the first experiment, while Bayesians are happy to ascribe a probability to both experiments.

The experiments also show how the Bayesian view of probability has much wider applicability than the frequentist view. A Bayesian 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.

Note too that all of the people involved in the coin-toss experiment were happy to ascribe a probability to both experiments - the Bayesian approach to probability is more consistent with the way we think.

The most important point of the experiment 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.

Confidence Intervals (Freq.)

- Problem:
  - Measure Boltzmann's constant, *k*
  - Error is zero-mean normal noise process
- Measure:
  - *n* samples from normal distribution
  - ➔ calculate mean, sample standard deviation
- Estimate *k*
  - The interval $\left[M - t_{0.975,n-1} S / \sqrt{n}, M + t_{0.975,n-1} S / \sqrt{n} \right]$ has a 95% probability of containing *k*
  - Note: M, S are random variables; the statement is meaningless when replaced by numerical values
  - Note: mean of measurements has normal distribution
  - Note: use *t*-distribution to calculate confidence interval

5

Now, for the purposes of demonstrating the difference in the two approaches, let's consider a simplified measurement of the Boltzmann constant, where we consider only a single measured variable subject to a zero-mean random error (i.e., random in the frequentist sense). I'll assume that the value of *k* produced by the measurement is simply the true value plus a random error drawn from a zero-mean Gaussian (normal) distribution.

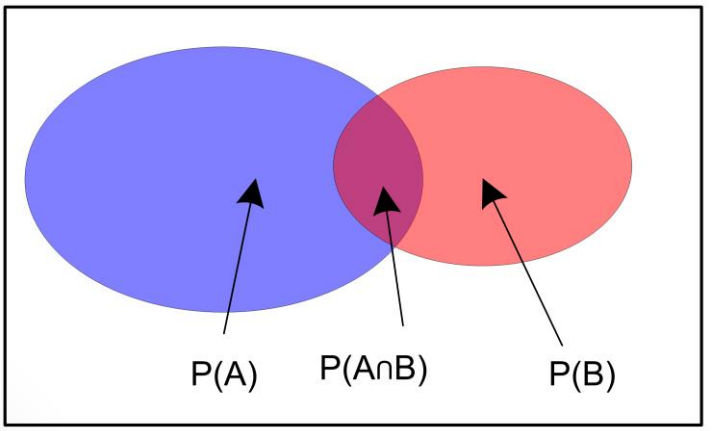
Now, if the we follow the process recommended by the current GUM, we make a bunch of measurements, and calculate the mean and the standard error of the mean (the standard uncertainty). To obtain an expanded uncertainty we generate a confidence interval using the *t*-distribution.

Note that every time we repeat the experiment, we get a different confidence interval, so that the confidence interval is itself a random variable. A frequentist says that 95% of those intervals will contain the true value of *k*.

Note that once a numeric interval is calculated, it either contains the true value or it does not contain the true value – it is no longer random. From the frequentist perspective, one cannot say that there is a 95% probability that a specific numeric interval contains the measurand (in the same way that the frequentist could not assign a probability to the state of the coin after it has been tossed). The 95% probability applies only to the corresponding fraction of a large number of such intervals. Note, again, that our natural thinking on uncertainties is closer to the Bayesian perspective than the frequentist perspective.

One of the primary aims of frequentist statistics, and a key distinguishing feature, is to develop algorithms for realising accurate confidence intervals (i.e., neither conservative nor optimistic). We'll return to this point later.

Bayes' Theorem



- Basic statistical result $\mathrm{P}(A \mid B)\,\mathrm{P}(B) = \mathrm{P}(B \mid A)\,\mathrm{P}(A)$
- Usually expressed as $\mathrm{P}(A \mid B) = \dfrac{\mathrm{P}(B \mid A)\,\mathrm{P}(A)}{\mathrm{P}(B)}$

6

Now, before explaining Bayesian statistics, let us review Bayes' Theorem.

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 total population – including the white area).

The area of overlap 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.

The mathematical result is Bayes' Theorem. It is usually written slightly differently, which is how the theorem is used for evaluating conditional probabilities.

As you can see, the explanation I have presented is in frequentist terms. The theorem is used by both Bayesian and frequentist statisticians.

Bayes' Theorem cont.

- Example: effectiveness of medical screening
  - 1% of population have disease
  - Test 80% reliable for detecting disease
  - Test 90% reliable for detecting absence of disease
- Test 10,000 people
  - Correctly identify 80% of the 100 people with disease
  - Incorrectly identify 10% of 9,900 without disease
  - Probability of having the disease having tested positive:

$$P(A \mid B) = \frac{0.80 \times 0.01}{0.80 \times 0.01 + 0.1 \times 0.99} = 0.075.$$

  - Interpreted differently by Bayesians and frequentists
  - Applies for single probability

7

This nice example shows how Bayes' theorem is used, and how it is interpreted differently by Bayesian and frequentist statisticians. It is also a good example showing how ineffective some (medical) screening processes can be.

Let us suppose that 1% of the population have some particular disease and we have developed a test for detecting the presence of that disease. We reckon that the test is 80% reliable for detecting the presence of the disease, and 90% reliable for detecting the absence of the disease. Now, what is the probability of someone having the disease given that they have tested positive? The answer is given by Bayes' Theorem:

Consider a population of 10,000 people: on average 100 have the disease, and 9,900 do not have the disease. If we test all 10,000, then 80 of the 100 with the disease will test positive, and 990 of those without the disease will also test positive.

P(A) is the probability of having the disease = 0.01

P(B) is the probability of testing positive for the disease = 0.8 x 0.01 + 0.1 x 0.99

P(B|A) is the probability of testing positive, given that the person has the disease = 0.8 x 0.01

The fraction of those that have the disease having tested positive is therefore

$$P(A \mid B) = \frac{0.80 \times 0.01}{0.80 \times 0.01 + 0.1 \times 0.99} = 0.075.$$

To a frequentist, this means that 7.5% of those that test positive will have the disease. The probability has to do with the long-run average over the population.

To a Bayesian, there is a 7.5% chance that any individual that tests positive will have the disease. The probability has to do with the lack of complete information about the medical status of the individual.

Bayesian statistics

- Generalise Bayes' theorem from probability to the whole distribution:

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(B \mid A)}{\int \mathrm{P}(B \mid A)\,\mathrm{P}(A)dA}\,\mathrm{P}(A)$$

  - Where $\mathrm{P}(A)$ is the prior distribution
  - $\mathrm{P}(A \mid B)$ is the posterior distribution
  - And $\frac{\mathrm{P}(B \mid A)}{\int \mathrm{P}(B \mid A)\,\mathrm{P}(A)dA}$ is likelihood function
  - There are two main types of Bayesian approach
  - 'Subjective' and 'objective'

8

The medical screening example shows the application of Bayes' theorem to the calculation of a single conditional probability. Bayesian statistics extends this idea to the calculation of complete pdfs rather than single probabilities.

Suppose we start an experiment with some knowledge of the value of a quantity A, and we represent our state of knowledge of A by a pdf, P(A). We then perform a bunch of measurements, B, and then ask the question: how has our state of knowledge changed as a result of the measurements? The answer is given by the integral version of Bayes' theorem

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(B \mid A)}{\int \mathrm{P}(B \mid A)\,\mathrm{P}(A)dA}\,\mathrm{P}(A)$$

The state of knowledge before the measurements is represented by the 'prior distribution' P(A), and the state of knowledge after the measurements is represented by the 'posterior distribution' P(A|B).

One of the major problems with Bayesian statistics is that, in general, the equation often cannot be solved analytically and must be solved numerically. Unfortunately, the numerical calculations are so computationally expensive that until the last couple of decades, many Bayesian calculations were impractical. This situation has changed with the development of new integration algorithms and ready access to very fast personal computers. (It's important to note that there are also many frequentist problems where the solution is numerical or 'let's hire a statistician'.)

Although in general the equation must be solved numerically, there are particular instances where analytic solutions exist. (See for example 'conjugate prior' distributions in Wikipedia.) I'm now going to give two examples of the evaluation of the equation using two instances of these conjugate prior distributions. These two examples will also illustrate the differences in the two main branches of Bayesian statistics.

Credible Intervals (Subjective)

- Use prior distribution to represent prior knowledge of Boltzmann's constant
  - E.g. as normal distribution $N(k_{\text{prior}}, \sigma_{\text{prior}})$
- Make *n* measurements
  - Calculate mean, *m*, and standard deviation, *s*
- Calculate posterior distribution

$$k_{\text{post}} = \frac{m\sigma_{\text{prior}}^2 + k_{\text{prior}} s^2 / n}{\sigma_{\text{prior}}^2 + s^2 / n} \qquad \sigma_{\text{post}}^2 = \frac{\sigma_{\text{prior}}^2 s^2 / n}{\sigma_{\text{prior}}^2 + s^2 / n}$$

  - This is a subjective Bayesian approach
  - The 95% coverage factor is 1.96 σ_{post}

9

Let us consider first an example of the **Subjective Bayesian** approach, with application to the Boltzmann constant problem we outlined earlier. Suppose we represent our prior knowledge of the value of Boltzmann's constant by a normal distribution $N(k_{\text{prior}}, \sigma_{\text{prior}})$. We then make the measurements, with mean and standard deviation, *m* and *s*. It turns out that the posterior distribution representing the state of knowledge of *k* after the measurements is also given by a normal distribution, but with mean and standard deviation given by the two equations…

We can see that the mean of the posterior distribution is the average of the experimental and prior values weighted by the corresponding variances. Also the variance of the posterior distribution, representing the uncertainty in the posterior mean, is given by the 'parallel combination' (by analogy with resistors in parallel) of the prior and experimental variances. Thus, the standard uncertainty after the measurement is always less than both the standard uncertainty before the measurement and the experimental standard deviation, s. The fact that the posterior uncertainty is always less than the experimental standard deviation reflects the prior information included in the calculation. (Note that in the limit σ_{prior} => infinity, the Bayesian and frequentist results are the same).

The use of 'informative priors', as illustrated here is the most common Bayesian approach. However, it does require the experimenter to nominate a distribution to represent the prior state of knowledge. This prior distribution is considered to be personal, i.e., it varies between experimenters - hence the terminology Subjective Bayesian statistics.

We can also see from the example, that the Type B uncertainty assessment of the GUM falls very naturally within Subjective Bayesian statistics. Especially for Type B assessments of unknown systematic errors, the current GUM must stretch the frequentist meaning of random by assuming that a systematic error is a single sample drawn from a distribution. Note that there is a spectrum of frequentist approaches to statistics and this idea of using a single sample to represent a systematic error is well within the spectrum of frequentist approaches, although some frequentists might reject it.

Note too that the current GUM describes some Type B assessments in terms of degrees of belief (= state of knowledge) implying a Bayesian definition of probability; see for example GUM clause 3.3.5.

Credible Intervals (Objective)

- Assume complete ignorance - use Jeffreys' prior distribution
- Make measurements
  - Calculate mean, *m*, and standard deviation, *s*
- Calculate posterior distribution
  - *t* distribution with, *m*, *s*, $\nu = n - 1$,
  - Credible interval $\left[m - t_{0.975,n-1} s / \sqrt{n}, m + t_{0.975,n-1} s / \sqrt{n} \right]$
  - Same numerical interval as frequentist
  - This is an objective Bayesian approach

10

Now some statisticians don't like the idea of the high level of subjectivity involved with the subjective Bayesian approach. Their solution is to adopt the **Objective Bayesian** approach.

The objective Bayesian approach adopts so called 'uninformative priors' to represent complete ignorance of the value of the measured quantity. The Jeffreys prior (see Wikipedia), which is used by GUM-S1, is one of the most commonly used. One of the appeals of the Jeffreys prior is that for normally distributed measurements, the posterior distribution is a shifted and scaled t-distribution that yields 'credible intervals' identical to the corresponding frequentist 'confidence intervals'. (Remember - while the numerical values of the intervals are the same, the meanings are different.)

There are a number of problems with the objective approach. Firstly, the Jeffreys prior is an improper distribution, which means that the integral in the denominator of the integral version of Bayes' theorem does not converge. The distribution also has no mean or standard deviation. Jeffrey's prior also violates other conditions that some statisticians expect of distributions. Secondly, the computations are simple only for quantities that may take any positive or negative value. If quantities are restricted in value (e.g., temperature and mass are always positive, or an artefact value is known *a priori* to be within 20 ppm say – as in a mass comparison), a different uninformed prior should be used, and the resulting posterior distributions are not so elegant or easy to compute.

The rationale for objective Bayesian statistics and how to find uninformative priors is one of the more contentious areas of statistics, and there are several branches depending on their axiomatic foundations. One of the most popular seeks priors that maximise the Shannon entropy of information theory.

The objective approach to Bayesian statistics is also in the minority. Most Bayesian statisticians argue that there is no such thing as an 'objective' or 'uninformative' prior, and that any prior injects subjective information and requires justification based on expert knowledge of the application.

It is notable that the GUM Supplement 1, which is Bayesian, uses the Jeffreys prior for all Type A assessments. It seems very strange that GUM-S1 recommends

(i) Subjective or informed priors for Type B assessments but only the Jeffreys prior for Type A assessments.

(ii) The Jeffreys prior for Type A assessments even when it is demonstrably inappropriate.

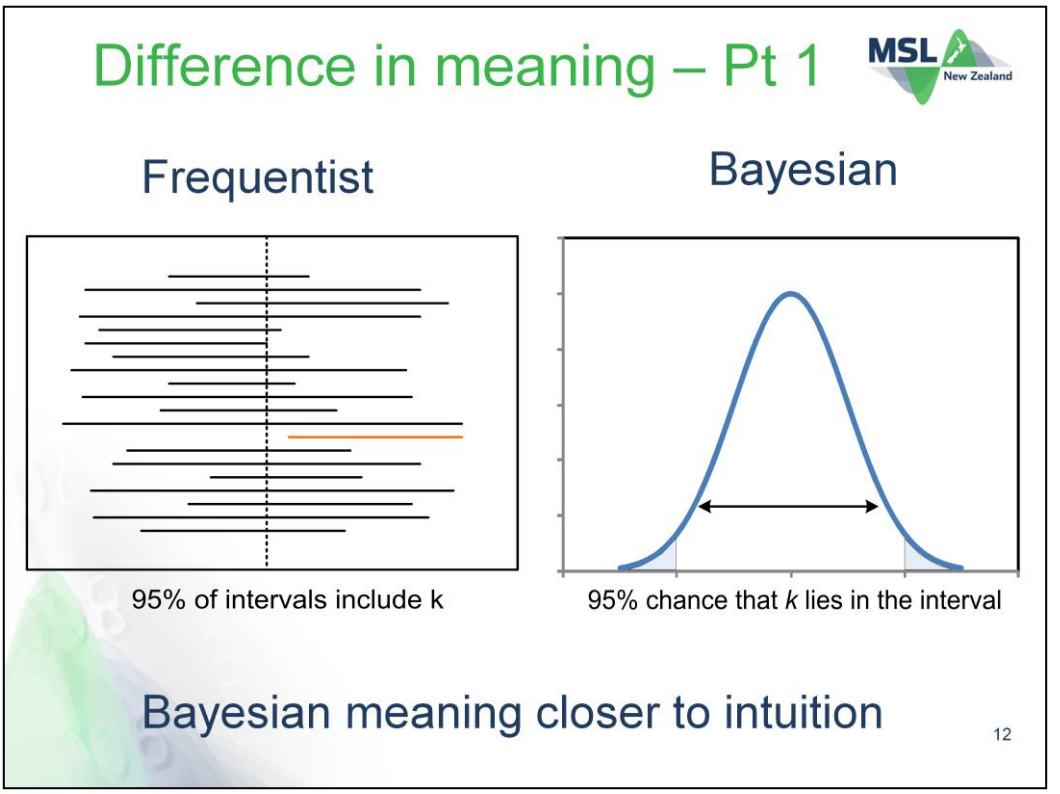
The subjective Bayesian approach particularly has had some great successes, especially where it can be applied iteratively by using the posterior distribution from one iteration as the prior distribution for the next iteration. Iterative applications have the advantage of tracking changes in measurands, and given enough time, any undesirable influence of a poor choice of first prior distribution is increasingly 'forgotten', and the posterior distribution slowly converges towards the real behaviour of the system.

One of the first successes was in the insurance industry where it is used to adjust premiums in response to a previous years' claims.

It is used extensively in cryptography and the communications industry for both code breaking and message detection.

In recent years it has also become a key technique in machine learning – we are probably most familiar with its use in spam filters.

Although many of the claims for success are justified, there is also some misinformation. Often the claims seem to be based simply on the application of Bayes' theorem, which applies equally for Bayesian and frequentist methods. It is also commonly claimed that the Bayesian approach works for drug trials where frequentist methods have failed. In fact, the problem is that the frequentist approaches are sometimes badly misused in drug trials.



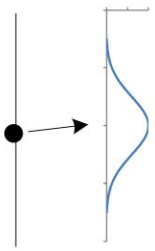
OK, we've already touched on this, but I thought it should be emphasised.

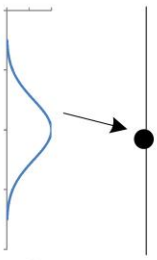
The frequentist approach of the GUM produces random confidence intervals, 95% of which will contain the measurand.

The Bayesian approach produces a posterior distribution from which a credible interval can be determined.

Note that there are many different possible Bayesian approaches each dependent on the personal assessments of the experimenter, all with different prior distributions and posterior distributions, and all of which make the claim that their various different credible intervals have the same probability of containing the measurand. This highlights again that the Bayesian probability has to do with information and strength of belief rather than characterising the real random behaviour of a system.

Difference in meaning - Pt 2





Frequentist:
Parameter fixed,
Data distributed (in error)

Bayesian:
Data fixed (known),
Parameter distributed

Bayesian model OK for distributed measurands, but incompatible with measurement, physics, chemistry,…

13

One of the most marked differences between the two approaches relates to the differences in the models of the measurement. In each of the two drawings above, the left-hand line indicates the measurand, the right-hand line indicates the measurements, and the arrow indicates the measurement process.

The frequentist approach, with which we are most familiar, assumes a fixed (but unknown) value for the measurand, and that the measurements are dispersed about this value due to randomness (errors) in the measurement process. This frequentist model of the measurement is exactly the same as our usual physical model of the measurement.

For the Bayesian, the only things known for sure are the results of the measurement, therefore they are considered to be fixed. The measurements are assumed to be drawn from a distribution representing the prior knowledge of the value of the measurand. In this case, the statistical model of the measurement is inconsistent with our usual physical model that assumes the measurand has a single, fixed, but unknown value.

This distinction between the statistical models and physical models for Bayesian statistics has an important consequence for numerical models of experiments. For example, Monte Carlo models of an experiment that assumes a fixed value of the measurand yield numerical estimates of the confidence intervals consistent with the frequentist analysis. However, in order for the numerical model to give intervals consistent with the Bayesian analysis, the numerical analysis must assume that the measurand is distributed in accordance with the prior distribution (see Mana and Palmisano, Metrologia, 51, 191).

When pushed, Bayesians will emphasise that their model of the measurement relates to the value of the measurand and not the measurand itself. However, to the Bayesian, it really does not matter whether the measurand is distributed or not - the measurand is always represented by a distribution. Bayesian statistics is so frequently used with distributed measurands that the idea of distributed measurands becomes an integral part of Bayesian thinking.

This highlights, I believe, one of the major problems with a Bayesian uncertainty analysis – the statistical model is (usually) not compatible with the physical model.

Difference in performance

- Problem: expanded uncertainty in $X+Y$
  - x_i: n_1 samples drawn from N(0, σ_1)
  - y_i: n_2 samples drawn from N(0, σ_2)
- Frequentist (GUM):
  - Use Welch-Satterthwaite approximation to give effective degrees of freedom
- Bayesian (GUM-S1):
  - Convolve two t distributions with n_1-1 and n_2-1 degrees of freedom

14

OK. I've mentioned a couple of times that the coverage for the two approaches is different. Let's look at this in a bit more detail.

Suppose we have two measurands, X and Y, and we want to estimate the sum X+Y. For simplicity, we assume that the measurements are both affected by zero-mean noise processes with standard deviations as given.

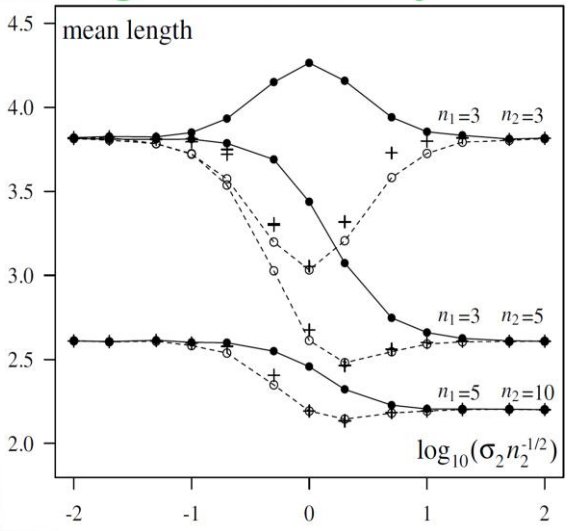
Although this is the simplest possible problem in combining uncertainty, it is very closely related to the famous Behrens–Fisher problem (see Wikipedia) for which there is no analytic solution.

The standard frequentist approach is to use the Welch-Satterthwaite approximation. This approximation reflects the way random process combine. For example, in the case where the two standard deviations are equal, and we have, say, 10 samples of each (9 degrees of freedom for each measurement), then when we add the estimates of X and Y, the measurement of the sum exhibits the same random behaviour as a measurement of a single variable with 18 degrees of freedom.

The Bayesian analysis ignores the random behaviour of the two variables and instead simply represents the state of knowledge of each variable by the corresponding distribution (which are t-distributions with 8 degrees of freedom in the case of the GUM-S1 approach). The distribution representing the combination of uncertainties is then obtained as the convolution of the two distributions. This is done numerically using Monte Carlo and adding pairs of samples drawn from each of the distributions.

Note: It has been suggested by some of the advocates of the GUM revision that the GUM is a fundamentally Bayesian document. However, the Welch-Satterthwaite approximation has no place in Bayesian statistics. The construction of confidence intervals, the Welch-Satterthwaite formula for combining uncertainties, and the use of effective degrees of freedom for Type B assessments are all signatures of frequentist statistics.

Combining uncertainty



- Coverage factor vs $\log_{10}\left[\left(\sigma_1/\sqrt{n_1}\right)/\left(\sigma_2/\sqrt{n_2}\right)\right]$
- GUM (dotted curve) < GUM-S1 (solid curve)

15

First let me thank Rob Willink for the plot. This was prepared by performing a Monte Carlo simulation of the physical model of the measurement of X+Y, as described in the previous slide. The numerical experiment covers a range of values of the standard deviations, and several different numbers of measurements of each variable.

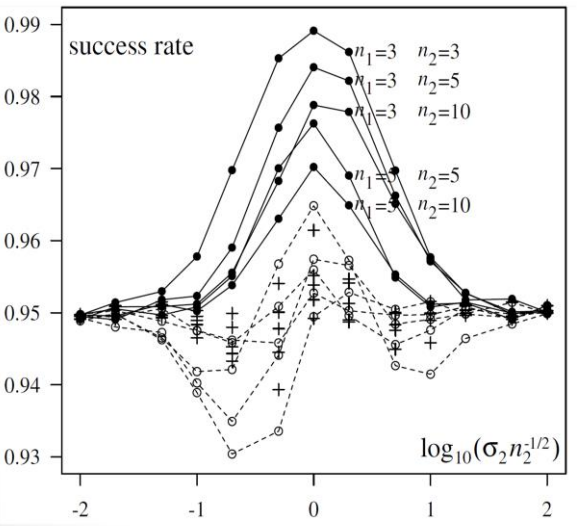
The plot compares the 95% coverage factors computed for X+Y for the frequentist approach (GUM, dotted lines) and Objective Bayesian approached (GUM-S1, solid lines). Both sets of results are plotted against the log of the ratio of the standard errors so that the different curves are readily visible on the same scale.

Several observations can be made:

- The objective Bayesian approach yields credible intervals that are always larger than the corresponding confidence intervals (in contrast to the subjective Bayesian approach).
- Where the uncertainty in one of the variables dominates the other, the system behaves in the same way as a single-variable measurement, and the two approaches give the same result.
- When the uncertainties are similar (and we are in the greatest need of an accurate approximation) the differences between the two approaches are greatest.
- Although not immediately apparent from the figure, when the number of measurements (n_1, n_2) are large, both coverage factors converge on $k = 1.96$ and the differences are negligible.

But there remains the question of which, if either, gives the correct coverage factors.

Combining uncertainty cont.



- GUM is nearly always closer than GUM-S1

16

This chart, again from Rob Willink, plots the actual success rates for the two approaches compared to the ideal of 95%. Again, the solid line reports the percentage of credible intervals (GUM-S1) that enclosed the measurand, while the dotted line reports the same success rates for the frequentist (GUM) approach. The simulation is the same as for the previous plot, so is based on numerical simulations.

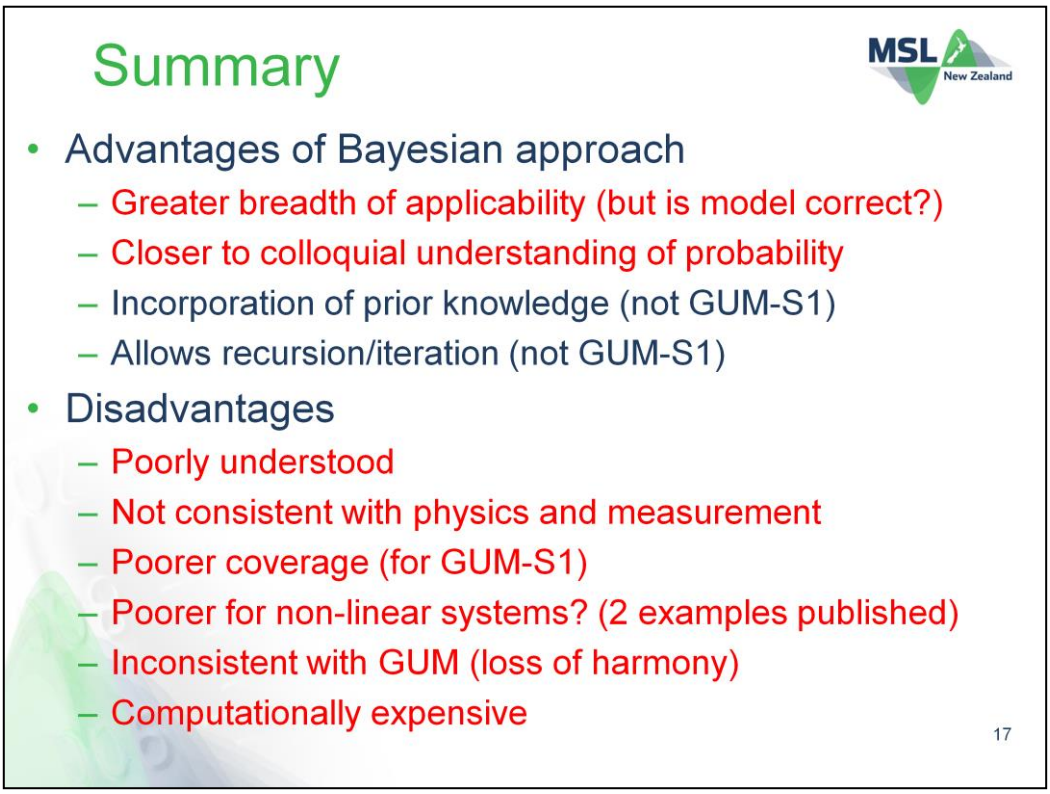
The plots shows rather nicely that the GUM approach, employing the Welch-Satterthwaite approximation, is always closer to the correct coverage factor than the GUM-S1 approach. Also, as expected both approaches give better estimates of the coverage factors as the number of measurements increases, or if the uncertainties in one measurement dominate the other.

We must be very careful about the interpretation of these results. Firstly, these experiments simply illustrate the consequence of the difference between the frequentist and Bayesian models of the experiment - this is a frequentist experiment (fixed measurand), so we should expect the frequentist results to be close the expected 95%, and the Bayesian (GUM-S1) intervals to be too large.

Secondly, the Bayesian results presented here apply strictly to the GUM-S1 approach, for which combined uncertainties (credible intervals) are always larger than the corresponding frequentist (GUM) intervals. In contrast, and as we saw earlier, for single-variable measurements, the subjective Bayesian approach always leads to credible intervals smaller than the frequentist intervals. There are too many possible subjective Bayesian experiments to make general statements about combined uncertainties for the subjective case.

Thirdly, Bayesian statisticians make no claims about coverage for these types of numerical experiments. Their claim is that a subjective Bayesian uncertainty estimate is always better because it includes prior information. A physicist's counter to this claim might be that the only prior knowledge in physics is experimental – if we wanted to include the prior information we would (e.g., as a weighted average, or by weighted least squares, as with the CODATA adjustment).

This numerical experiment highlights a very serious concern. Around the world, regulations (e.g., the Average Quantity System – OIML R87) and quality-control processes in commercial contracts are often governed by tests based on sampling. If measurement uncertainties are to be meaningful and to allow manufacturers to make high-quality decisions about manufacturing tolerances and avoid prosecution, measurement uncertainties must accurately reflect real-world frequencies.



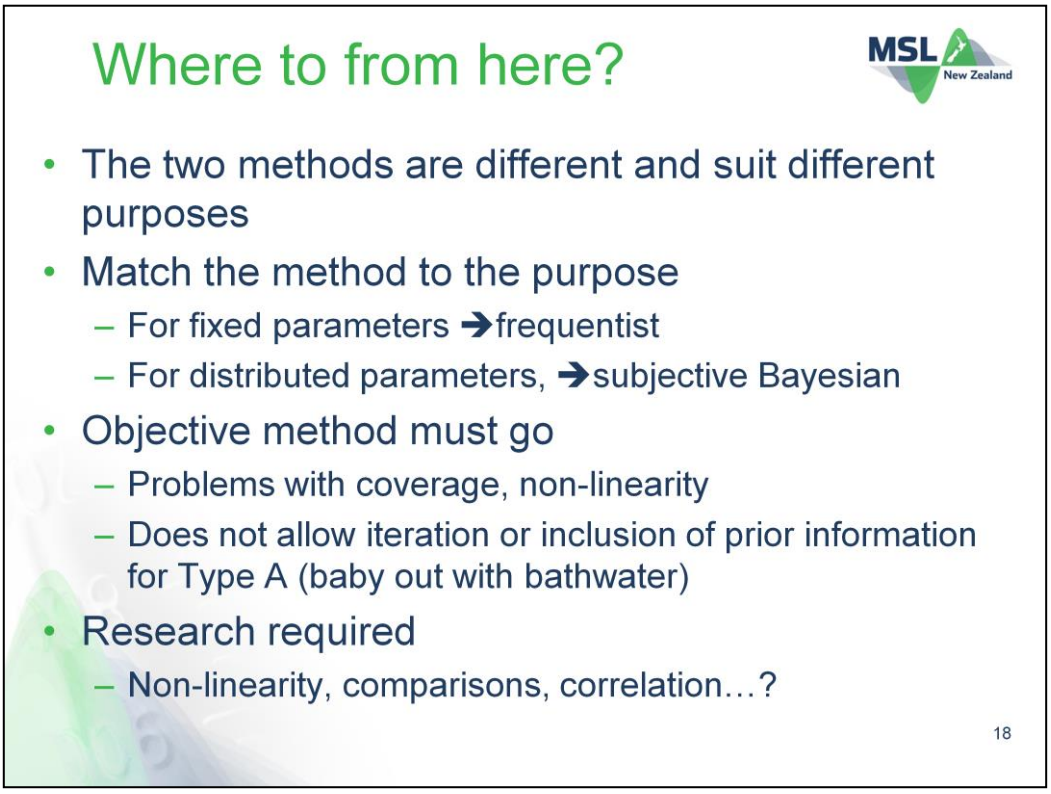
OK, we've covered a couple of the issues in detail, and since I don't have time deal with all the issues I'm aware of, let me just post a summary. Firstly, it's clear that the Bayesian approach has advantages:

- The less restrained definition of probability means the Bayesian approach is more widely applicable,
- The Bayesian definition of probability is closer to our colloquial usage of the term
- A major advantage is that it provides a formal way of including prior information, and
- Of allowing the use of iteration as more data comes available.

Unfortunately GUM-S1 does not allow us to benefit from these last two advantages.

There are also some serious disadvantages:

- Bayesian statistics is poorly understood by non-statisticians (frequentist statistics is also poorly understood, but less so).
- Statistical models and physical models are different: the concept of a measurand having a single true value, physicists' concept of randomness, and the idea of measurements being in error seem to be foreign to Bayesian statistics. This contradiction with the most basic foundations of physics and metrology can only cause trouble in the long term.
- The prime motivation of (most branches of) Bayesian statistics is not to provide accurate coverage with credible intervals. Where uncertainties are used to characterise dispersion of measurements (at least 99% of metrology), this is a problem. A bigger problem, perhaps, is that some of our commercial clients use uncertainty to manage compliance with regulations based on the Average Quantity System, which is inherently frequentist.
- The Bayesian approach is inconsistent with the current GUM – a Bayesian revision will cause a loss of harmony. Actually, what is really peculiar is that GUM-S1 mixes both the subjective and objective Bayesian approaches, but does not use the subjective approach where it would give the greatest benefit (incorporation of prior knowledge and iteration for Type A analyses).
- In contrast to claims commonly made for the Bayesian approach, GUM-S1 has poor coverage for some non-linear systems. Hall (Metrologia 45, L5-L8, 2008) cites an example where GUM-S1 yields coverage intervals that NEVER enclose the measurand. (See also Wilrich, AQUAL, 19, 81, 2014).
- The computational complexity requires moderately powerful computers to carry out numerical analysis. The lack of an algebraic formulation for many problems also obscures insights otherwise obtainable with a simpler algebraic frequentist analysis. It will make the treatment of uncertainties even more obscure for most non-expert users.



As a metrologist used to questioning the purpose and definitions of measured quantities, it seems that Bayesian and frequentist probability measure different quantities. Despite the two approaches using similar terminology, similar mathematics, and yielding similar numbers, the basic meaning of probability is different. As in other areas of metrology where different measurement techniques realise different definitions of measured quantities, we should recognise that the Bayesian and frequentist approaches realise different definitions of uncertainty, and may be suited to different applications.

If, as Walter indicates, other sciences routinely require a statistical treatment that accommodates distributed measurands then a Bayesian approach might be well suited. This also suggests the measurement community has a need for both uncertainty tools. In the cases where Bayesian statistics are used, the subjective approach should be used to allow the full benefits of a Bayesian approach; i.e., the incorporation of prior information and use of iteration.

A major concern is that Bayesian statistical models are fundamentally incompatible with the models employed in most of physics (including physical metrology). The use of both models together can only lead to confusion, and eventually, probably, an undermining of current measurement and physics philosophy.

If a Bayesian approach is adopted, there are also outstanding problems with GUM-S1 and GUM-S2 that need to be solved for any GUM revision:

- How is correlation included? It particular, if all uncertainty analyses include prior information, how do we unravel the correlations caused by inclusion of common prior information (e.g., in time-series, comparisons, least squares). For these measurements, at least, there is a good reason to report only the novel information, as is conventional.
- How can the approach be applied to comparisons? All three published frequentist approaches to comparison analysis are now known to be identical (Koo and Clare, Metrologia, 49, 340) and experience shows they are functional (also, contrary to Walter's suggestion that uncertainties are generally too small, comparison results suggest otherwise). At the same time, there have been two instances where a Bayesian analysis has been dropped due to apparently silly results. (One may have been based on Elster and Toman Metrologia 50, 549 ?? Need to check). Also, given Bayesian credible intervals for comparison results, how should these be interpreted in the assessment of CMC claims.
- GUM-S2 is still not a satisfactory treatment for multivariate problems; coverage seems to get exponentially worse as the number of dimensions increases, while alternative treatments don't suffer the same problems (Hall, private communication). More research is required here for both the Bayesian and frequentist approaches.
- Is a Bayesian approach appropriate for managing compliance with fair-trading regulations using the Average Quantity System, which is now employed in most countries and based on frequentist statistics? A similar concern applies to uncertainty estimates in readings of calibrated instruments where we are genuinely interested in accurately measuring the dispersion of measured values, and coverage is important.

Finally, I (and others at MSL) believe we should treat all our algorithmic tools (not just uncertainty analysis) in the same way we treat measuring instruments. It is our duty to understand them, to carry out experiments and evaluate their performance, and to make sure they are fit for purpose.