Examples of coverage intervals with very good and very bad long-run success rate

Nicola Giaquinto

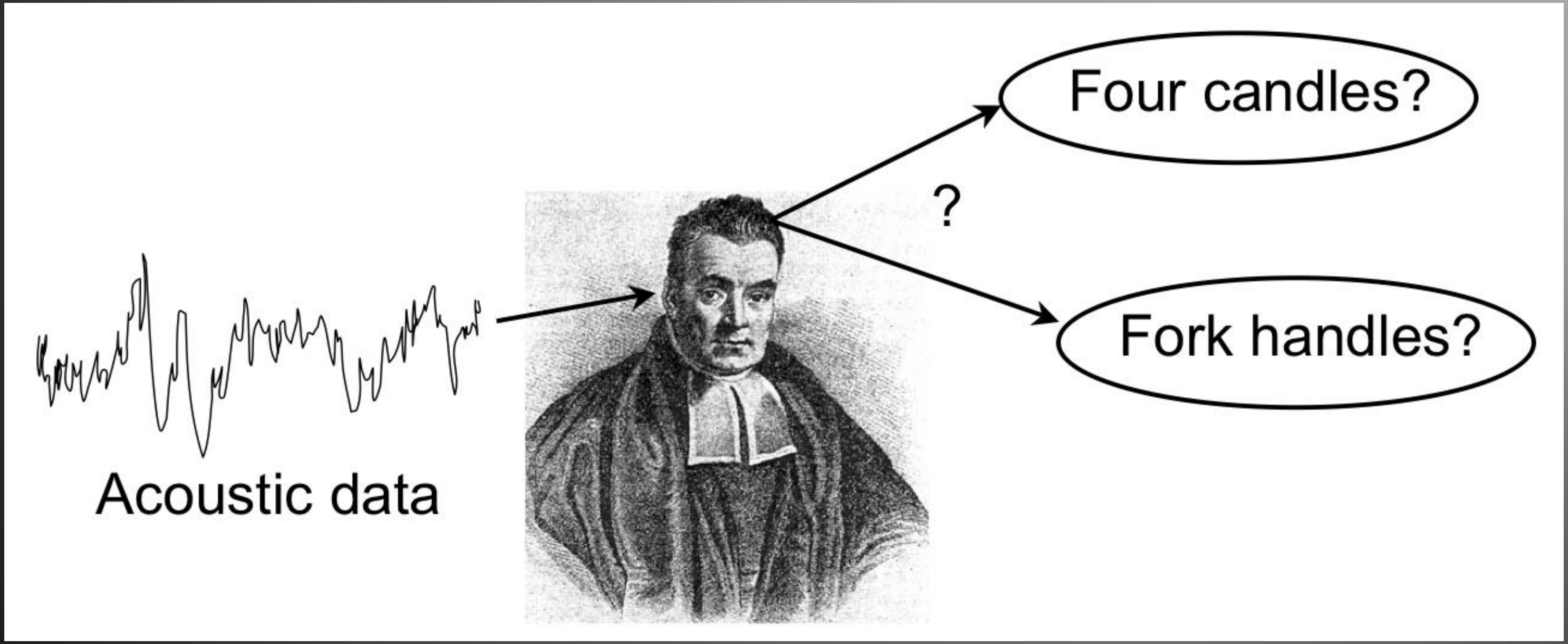
Politecnico di Bari – Italy

Outline

- Long-run success rate and Bayesian intervals
- The three examples
- Considerations
- Conclusions

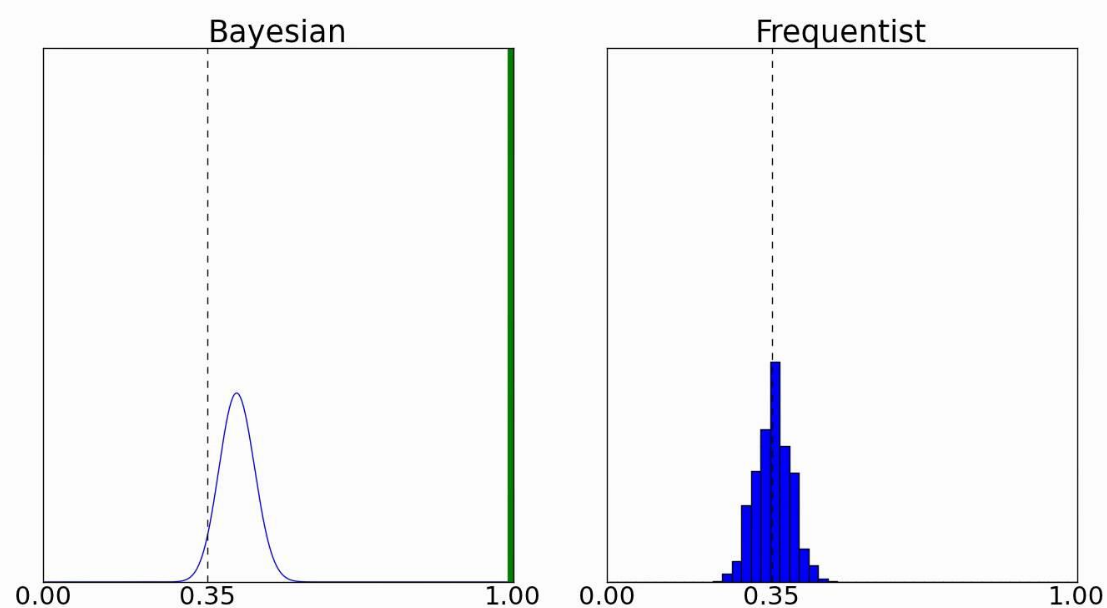
Long-run success rate and Bayesian intervals

Bayes at work



Why prefer Bayesian solutions?

	Bayesian approach	Orthodox approach
Usefulness of the resulting pdf	*Direct (probability of measurand values)*	*Indirect (probability of measured values for a supposed measurand)*
Use of the pdf	*Simple (credible intervals)*	*Convoluted (confidence intervals)*
Prior information	*Integrated in the scheme*	*Not covered*
Uncertainty	*Known quantity*	*Uncertain quantity*



When scrutinize Bayesian solutions?

- When only ***vague prior information*** is available («X is in the interval [a;b]»)
- Bayesian approach requires a clever **technical construction of the right "objective" prior pdf** (objective = non-informative = non-subjective)
- Stein's Paradox:
  different priors are needed to estimate μ and μ^2 in $N(\mu,\sigma^2)$
- Words from the (objective) Bayesian statistician J. Berger:

> Bayesian Analysis (2006) vol. 1, pp. 385–402
>
> **The Case for Objective Bayesian Analysis**
>
> J. Berger
>
> Objective priors can vary depending on the goal of the analysis for a given model.
>
> Use of constant priors, vague proper priors [...] I call such analyses pseudo-Bayes.
>
> **They do not carry with them any of the guarantees of good performance that come with (well-studied) objective Bayesian analysis**

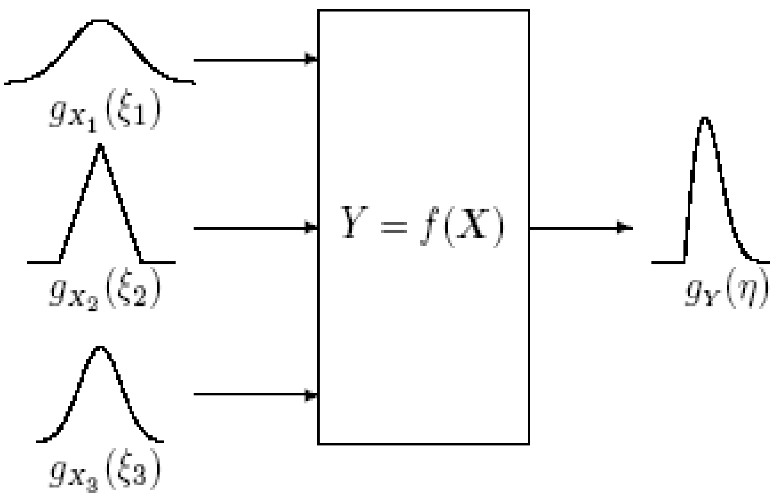
"frequentist notions are valuable in the construction of objective Bayesian methodology" (frequentist validation)

LRSR evaluation

(«Think like a Bayesian, check like a frequentist»)

The «weakly Bayesian» GUM2/S1 approach

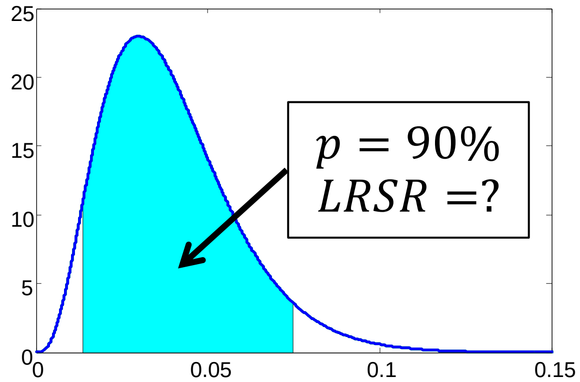
- Pdf are assigned to X_i according to rules *independent on the propagation problem*
- **Equivalent to a Bayesian analysis using *particular prefixed priors***
(Elster, Toman, Forbes, Lira, Grientschnig…)



Known problems of GUM2/S1 intervals

- In nonlinear problems the pdf of Y may have undesirable properties
  (Hall, Elster, Toman, Willink, Forbes, …)

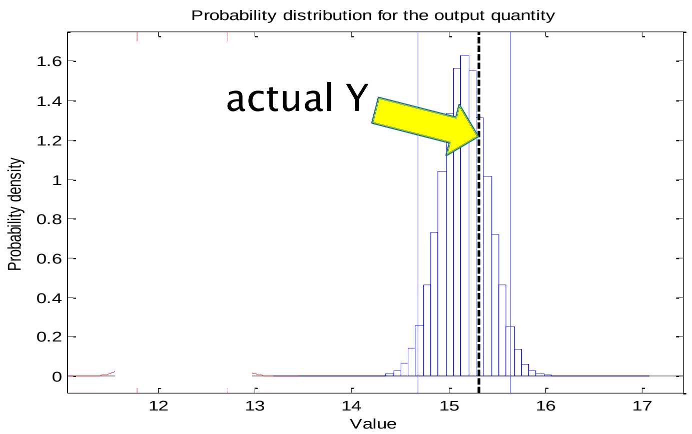
- LRSR of coverage intervals (which is NOT the coverage probability) can be low

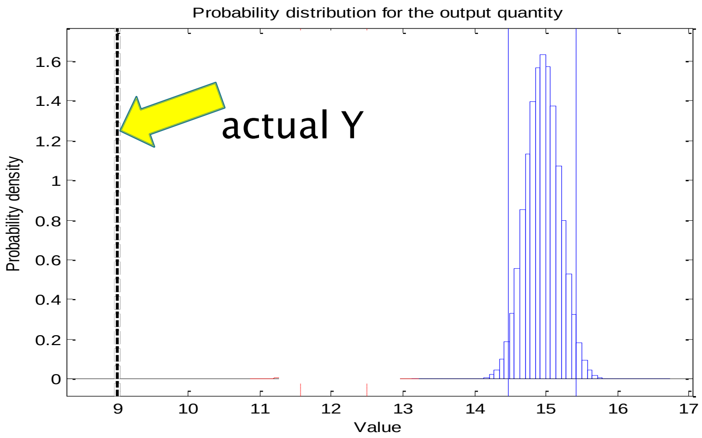


- Different positions:
  - Low LRSR demonstrates that the scheme is wrong
  - The pdf of Y can be improved with proper Bayesian techniques like MCMC
  - Low LRSR is not a problem and must be ignored

Investigating the problems by means of examples

- Three pair of examples (A, B, C)
- In each pair, input quantities has the same pdf
- In each pair:
  - n.1 yields $LRSR = p$ (satisfying)
  - n.2 yields $LRSR \cong 0$ (unsatisfying)





The examples

Effect of dice

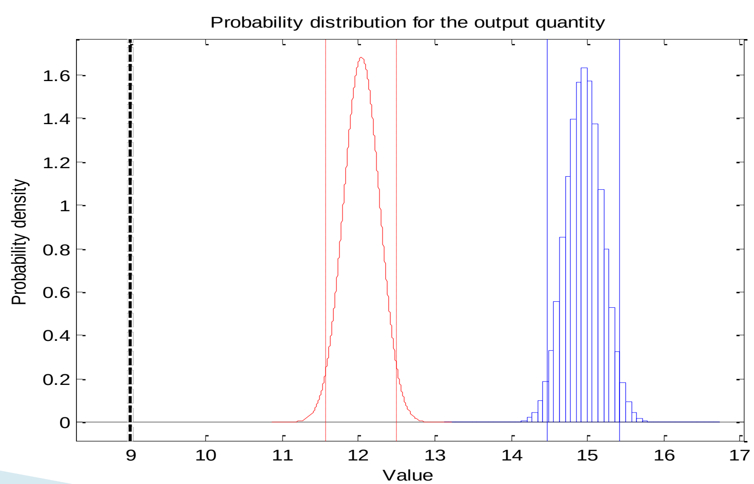
Effect of noise

Effect of quantization

Common elements in the examples

- Measurement model: **quadratic mean** $Y = f(x_1, \ldots, x_N) = \frac{1}{N}\sum_{i=1}^{N} X_i^2$
  - results can be intuitively interpreted
  - any other nonlinear model would be good
- Number of input quantities
  - $\boldsymbol{N = 2500}$, generates **BIG** differences between examples
  - $\boldsymbol{N = 1, 2}$ **generates identical (smaller) effects**
- Coverage interval
  - 95% probability symmetric interval
- Computations
  - **NPLUnc software** (exact Monte Carlo and approximate GUF solution)

    http://resource.npl.co.uk/docs/science_technology/scientific_computing/ssfm/documents/software/NPLUnc.zip
  - Analytical results are easy to obtain



A – Rolling N dice

- $X_1, \dots, X_N$ are totally unknown

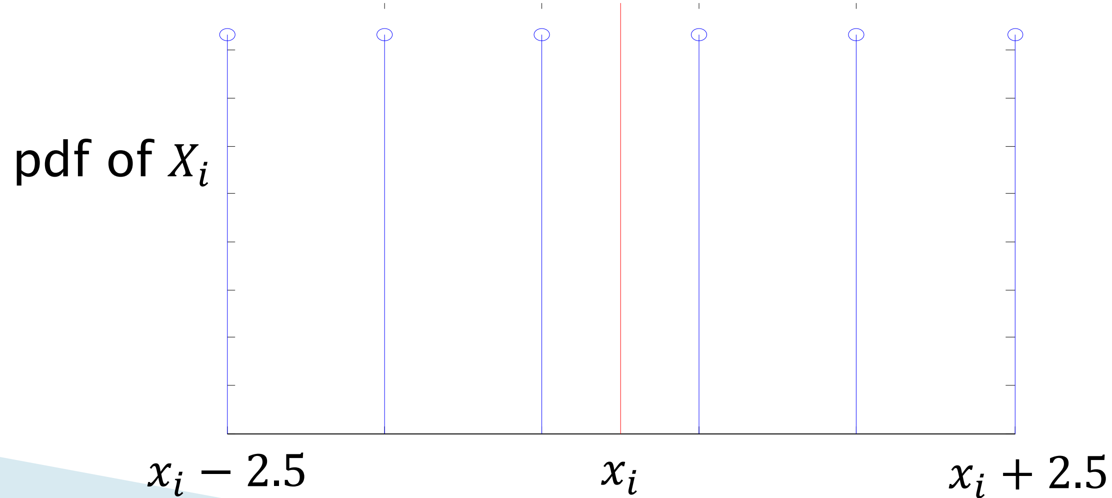
- Information:

A1

- $X_i =$ outcomes of the roll of N independent fair dice

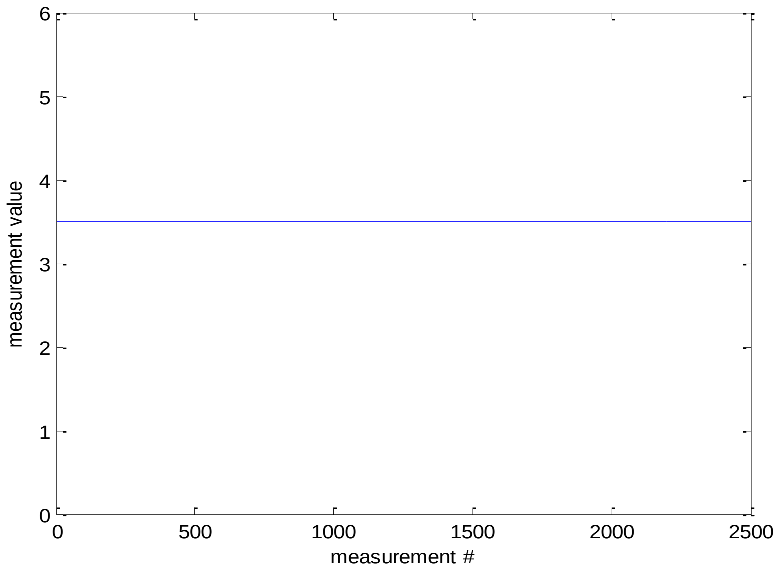
A2

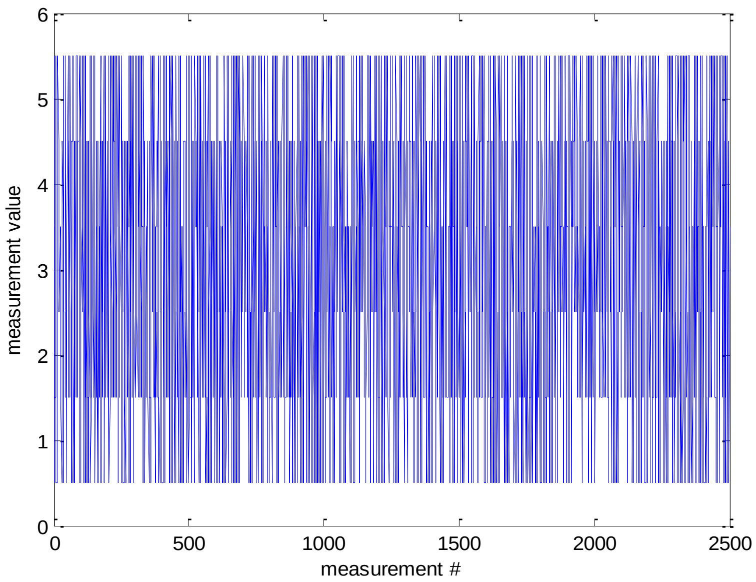
- $D_i =$ outcome of N fair dice, values $[-2.5, -1.5, -0.5, +0.5, +1.5, +2.5]$
- Values $x_i = X_i + D_i$ are given



Actual «measurements» x_i in examples A1 and A2

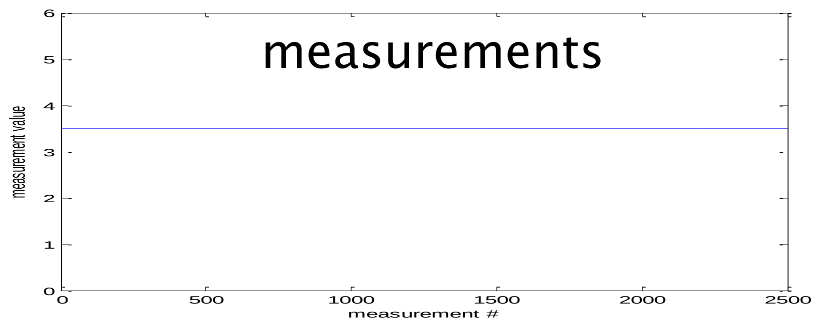
Note:
these are not repeated experiments, these are the input measurements x_i in a *single* experiment



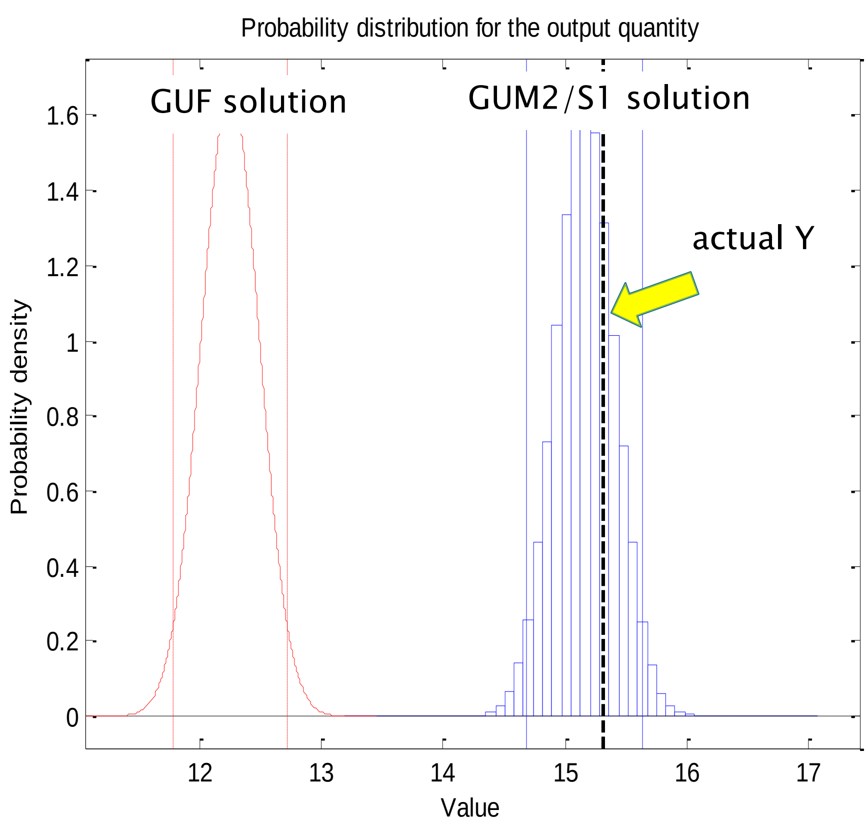


- In A1, $x_i = E[X_i] = 3.5$, $X_i = x_i + D_i$, with $D_i \in [-2.5, -1.5, -0.5, +0.5, +1.5, +2.5]$
- In A2, the situation is reversed, $x_i = X_i + D_i$ and x_i assume different («random») values

A1 – NPLUnc solution and actual Y

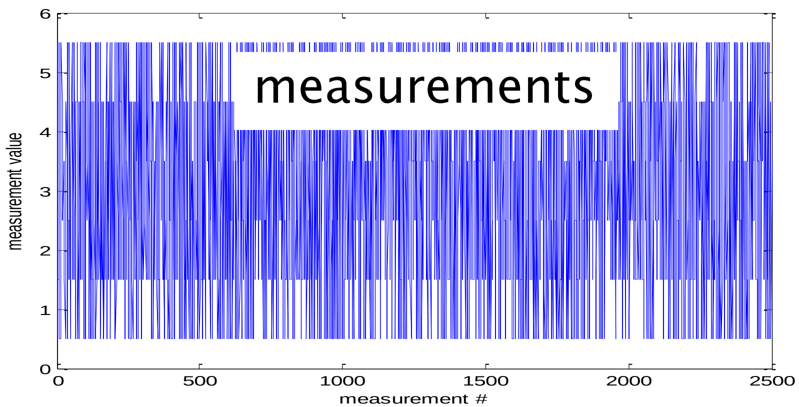


- Actual experiment: $Y = 15.3152$
- The GUM2 interval has $LRSR = p = 95\%$
- The GUM2 (bayesian) pdf is also a *frequency* pdf

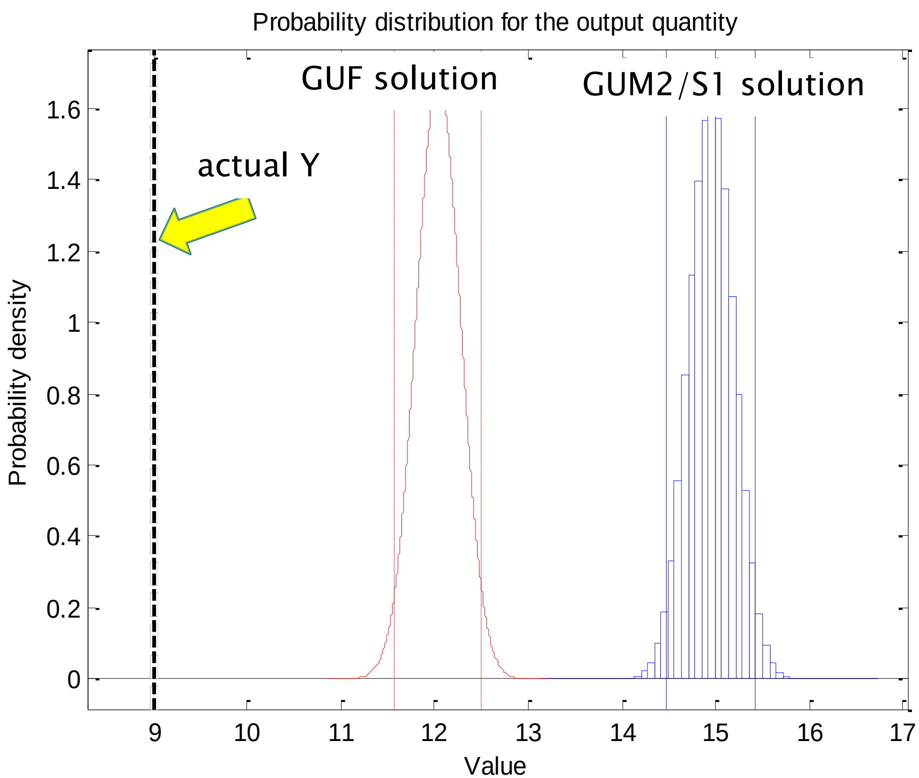



```
Results from GUM uncertainty framework:
y    = 12.25
u(y) = 0.239096
I(y) = 11.7814, 12.7186

Results from a Monte Carlo method:
y    = 15.1669
u(y) = 0.24381
I(y) = 14.6919, 15.651

MCM calculation has stabilized
GUF is NOT validated against MCM
>> quadratic_mean(X)
ans =
  15.315200000000001
>>
```

A2 – NPLUnc solution and actual Y



- Actual experiment: $Y = 9$!
- The GUM2 interval has $LRSR \cong 0$
- An interval with $LRSR = 95\%$ is easily computed using a “frequentist” approach: it is the confidence interval [8.65; 9.60]




```
Results from GUM uncertainty framework:
y    = 12.0396
u(y) = 0.237033
I(y) = 11.575, 12.5042

Results from a Monte Carlo method:
y    = 14.9568
u(y) = 0.242091
I(y) = 14.4823, 15.4326

MCM calculation has stabilized
GUF is NOT validated against MCM
>> quadratic_mean(X)
ans =
     9
>>
```

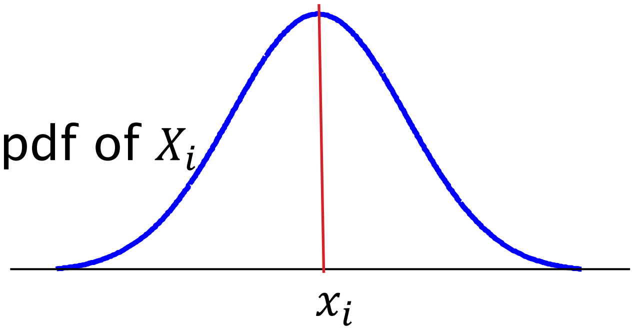
B – Gaussian noise

- $X_1, \ldots, X_N$ are totally unknown
- Information:

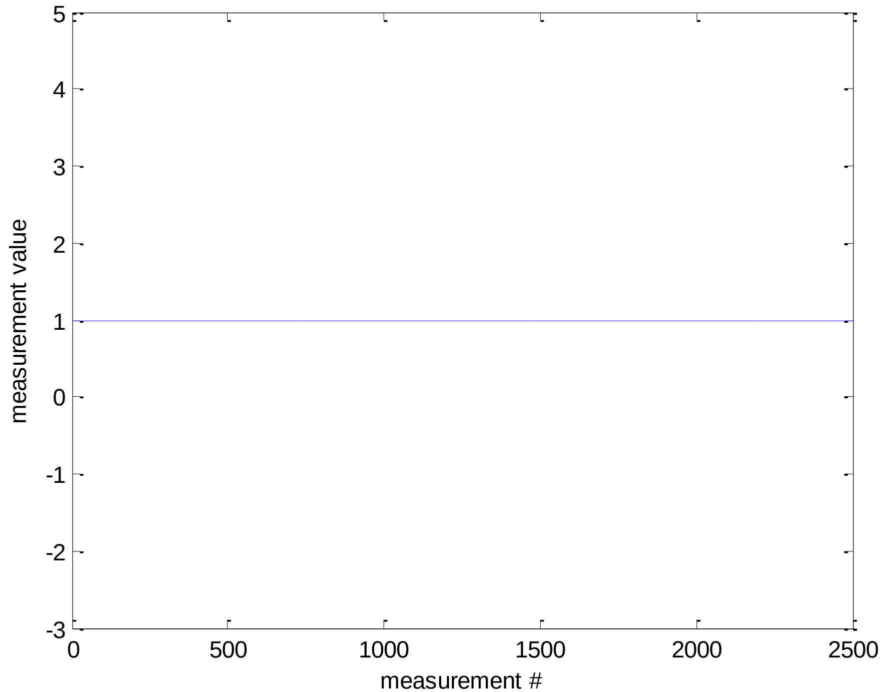
A1

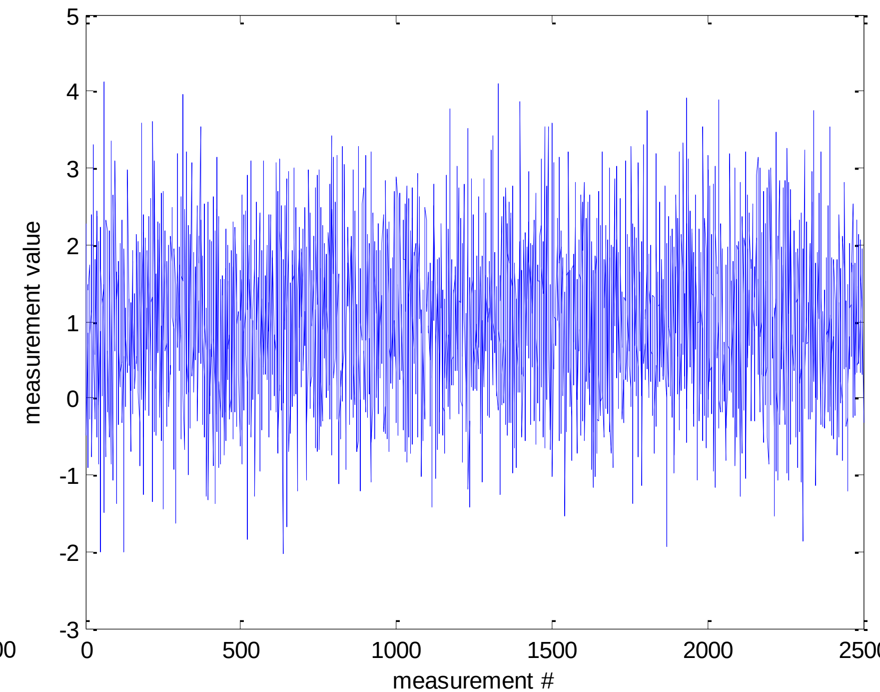
- $X_i = x_i + N_i$, with $N_i \sim N(0,1)$ i.i.d.
- measurands X_i are obtained by summing noise to measurements x_i

- A2
- $x_i = X_i + N_i$, with $N_i \sim N(0,1)$ i.i.d.
- measurements x_i are obtained by summing noise to measurands X_i



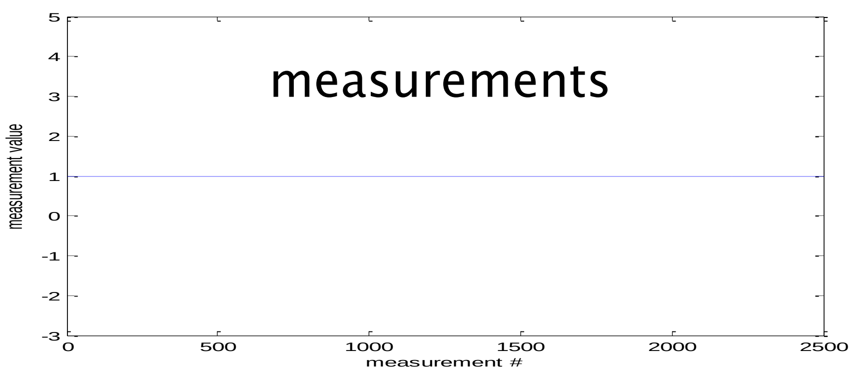
Actual measurements x_i in examples B1 and B2



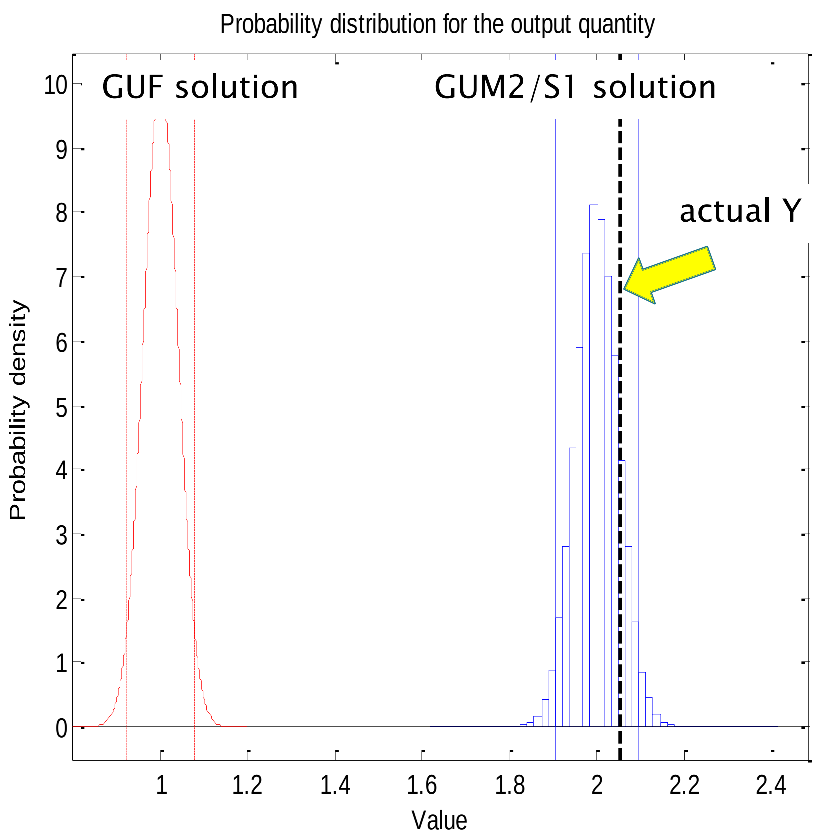


- In B1, $x_i = E[X_i] = 1$ (different x_i can be used)
- In B2, x_i assume different («random») values

B1 – NPLUnc solution and actual Y

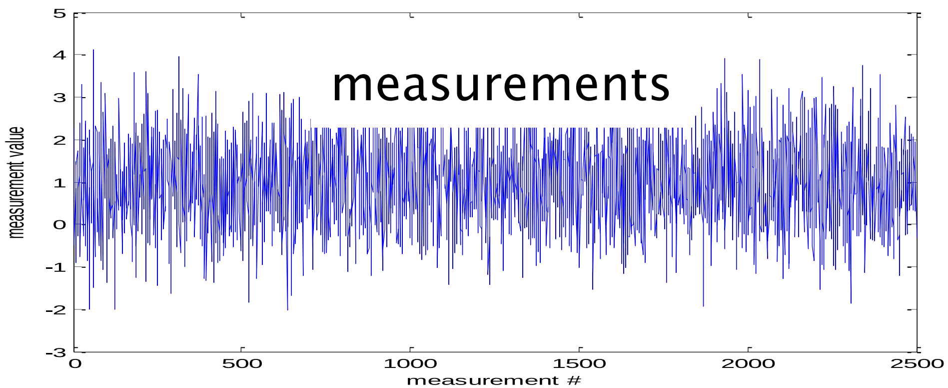


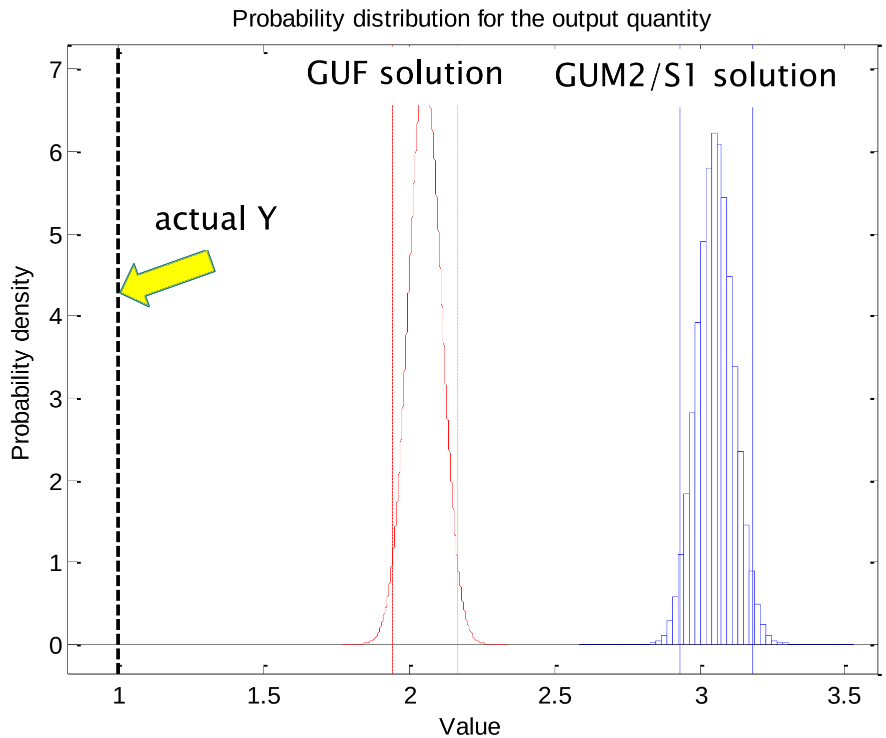
- Actual experiment: $Y = 2.0541$
- The GUM2 interval has $LRSR = p = 95\%$
- The GUM2 (bayesian) pdf is also a *frequency* pdf




```
Results from GUM uncertainty framework:
y     = 1
u(y)  = 0.04
I(y)  = 0.9216, 1.0784

Results from a Monte Carlo method:
y     = 2.00015
u(y)  = 0.0490884
I(y)  = 1.90417, 2.09711

MCM calculation has stabilized
GUF is NOT validated against MCM
>> quadratic_mean(X)
ans =
   2.05415844230092l
>> |
```

B2 – NPLUnc solution and actual Y





- Actual experiment: $Y = 1$!
- The GUM2 interval has $LRSR \cong 0$
- An interval with $LRSR = 95\%$ is easily computed using a "frequentist" approach: it is the confidence interval [0.94; 1.17]

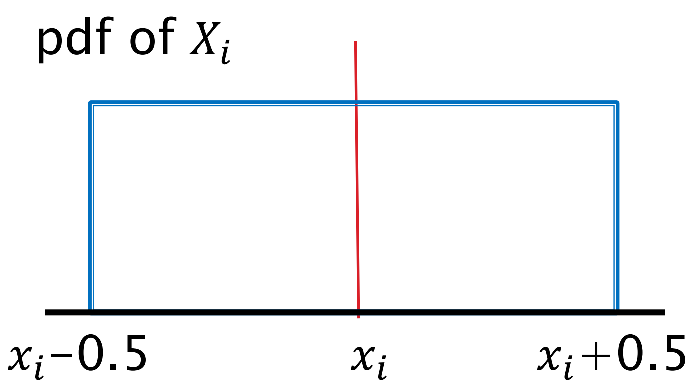
```
Results from GUM uncertainty framework:
y     = 2.05416
u(y)  = 0.0573293
I(y)  = 1.94179, 2.16652

Results from a Monte Carlo method:
y     = 3.05445
u(y)  = 0.0639412
I(y)  = 2.93006, 3.18169

MCM calculation has stabilized
GUF is NOT validated against MCM
>> quadratic_mean(X)
ans =
     1
>>
```

C – Quantization

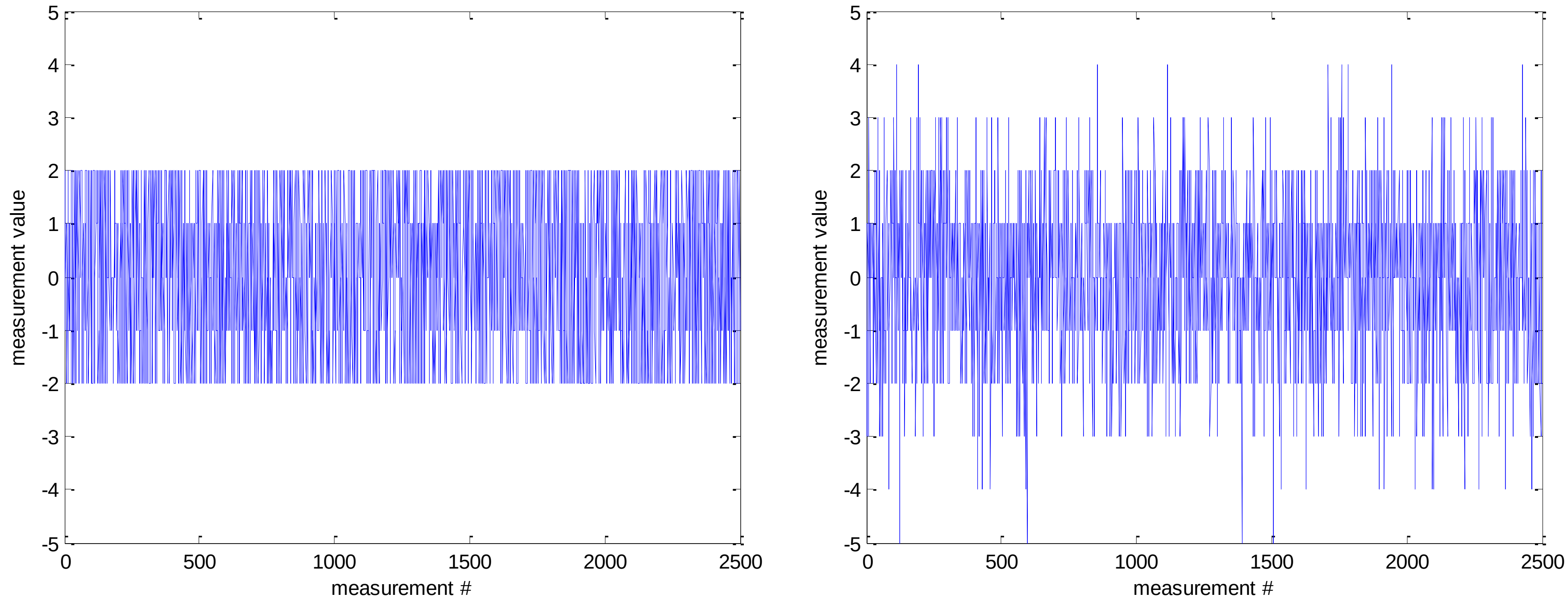
- $X_1, \ldots, X_N$ are totally unknown
- Information:

The same for C1 and C2

- $x_i = round(X_i)$ (quantization with step $Q = 1$)



Actual measurements x_i in examples C1 and C2

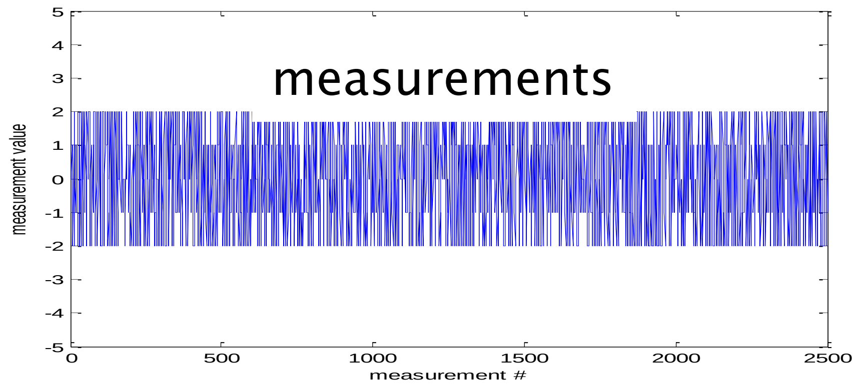


- In C1, the quantizer input is a “special” signal
  - *It spans uniformly an integer number of quantization step*

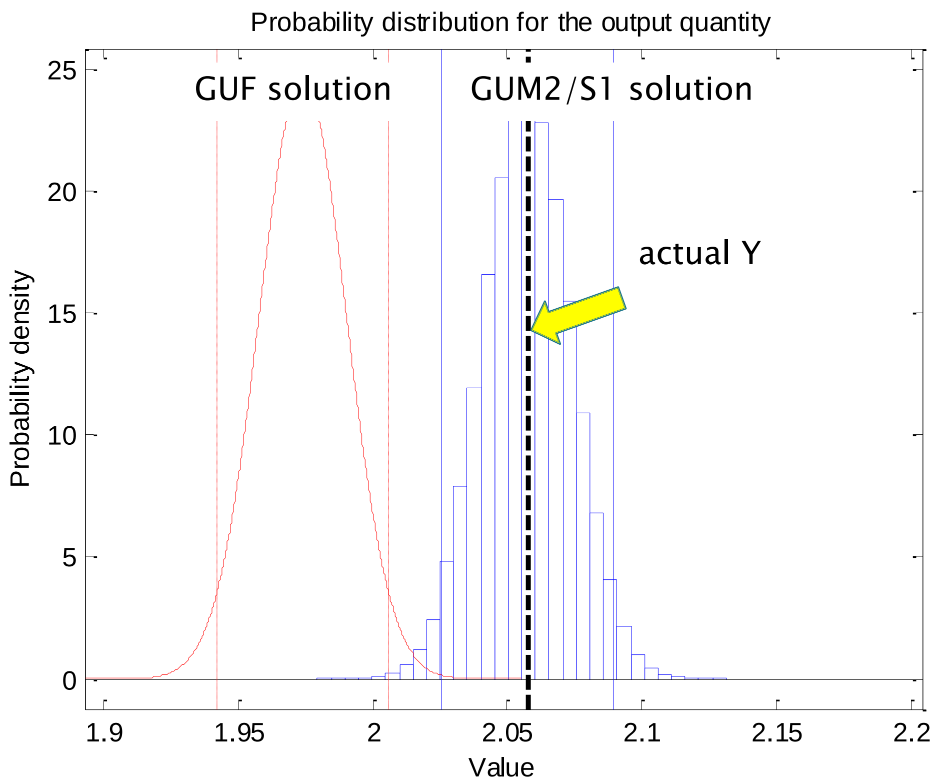
- In C2, the quantizer input is a much more generic signal
  - It has *a span that satisfies Quantization Theorem II* (bandlimited distribution)

(Widrow and Kollar, «Quantization noise», Cambridge University Press, 2008)

C1- NPLUnc solution and actual Y

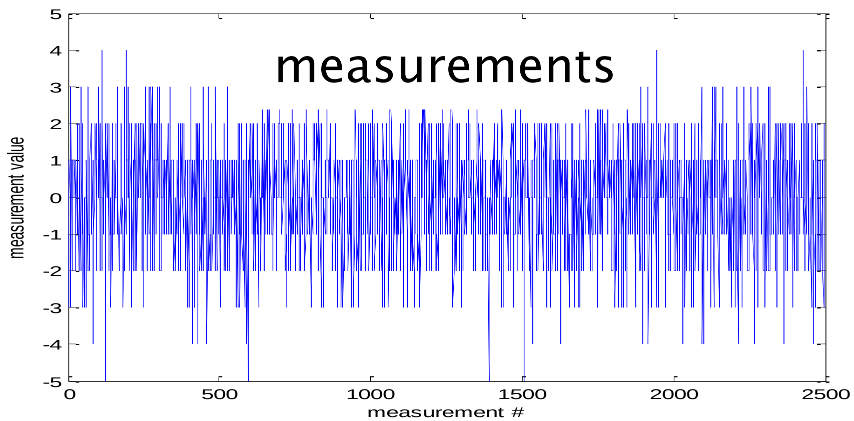


- Actual experiment: $Y = 2.0581$
- The GUM2 interval has $LRSR = p = 95\%$
- The GUM2 (bayesian) pdf is also a frequency pdf

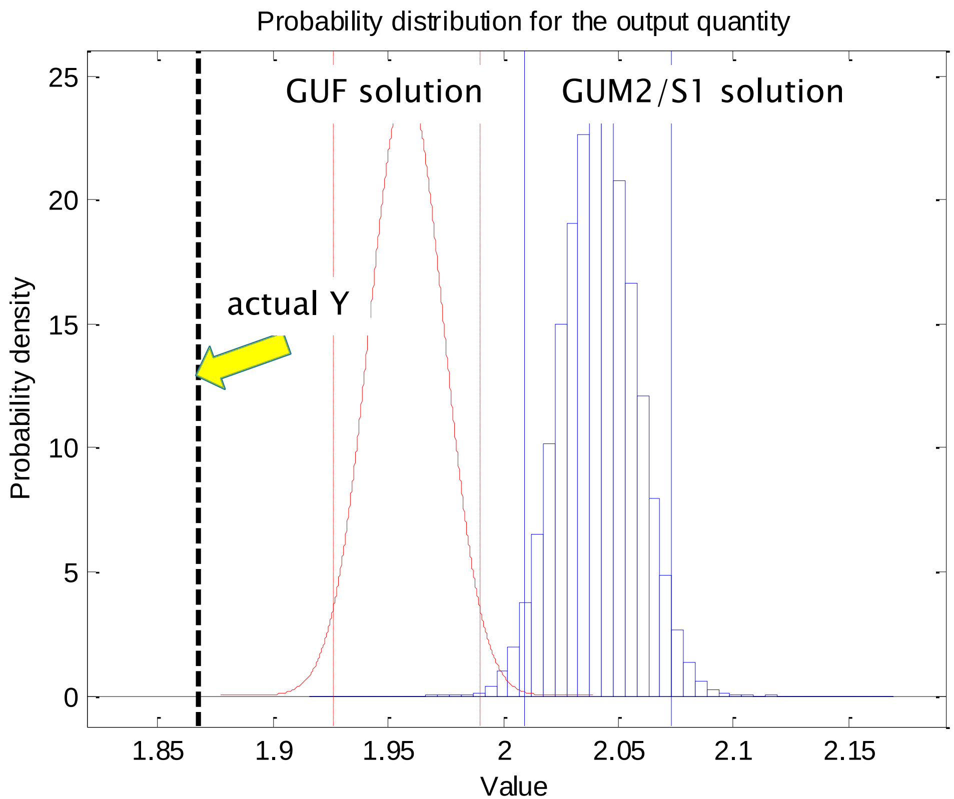



```
Results from GUM uncertainty framework:
y     = 1.974
u(y)  = 0.0162234
I(y)  = 1.9422, 2.0058

Results from a Monte Carlo method:
y     = 2.05735
u(y)  = 0.0162797
I(y)  = 2.02564, 2.08941

MCM calculation has stabilized
GUF is NOT validated against MCM
>> quadratic_mean(X)
ans =
    2.0581
>>
```

C2– NPLUnc solution and actual Y



- Actual experiment: $Y = 1.867$
- The GUM2 interval has $LRSR = 0$
- An interval with $LRSR = 95\%$ is easily computed using a "frequentist" approach: it is the confidence interval [1.843; 1.906]




```
Results from GUM uncertainty framework:
y    = 1.958
u(y) = 0.0161576
I(y) = 1.92633, 1.98967

Results from a Monte Carlo method:
y    = 2.04139
u(y) = 0.0162643
I(y) = 2.00938, 2.07331

MCM calculation has stabilized
GUF is NOT validated against MCM
>> quadratic_mean(X)
ans =
   1.867475027398909
>>
```

Considerations

Is success rate important in real life?

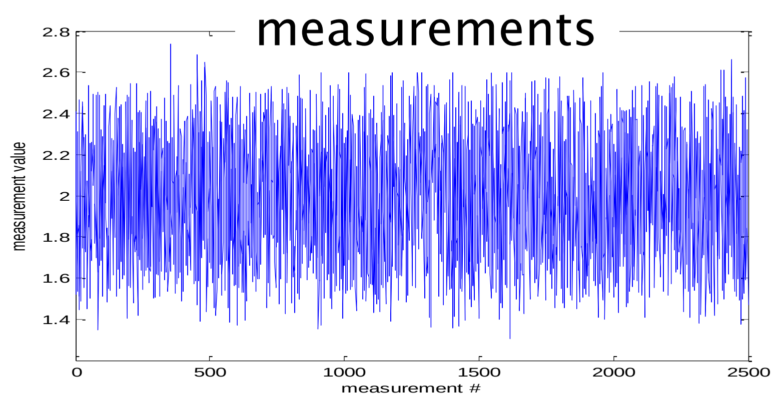
Simple interpretation of the examples

- Additive noise of known power p_n ($p_n = Q^2/12$ for quantization)
- Measured signal of (measured) power p_x
- Unknown signal of unknown power P_X

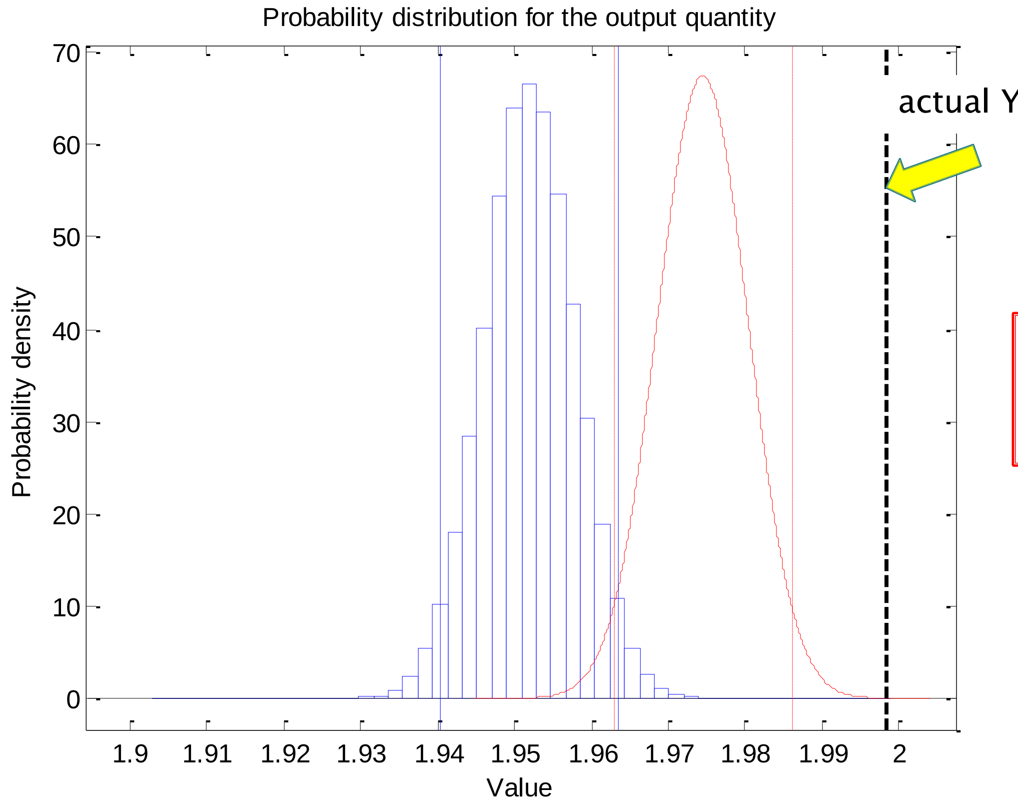
- GUM2 estimates P_X as
  - $P_X = p_x + p_N$
- This is OK in the (rare) case of noise uncorrelated *with the output* (examples A1, B1, C1)

- What usually happens is that noise is uncorrelated *with the input* (examples A2, B2, C2)
  - $p_x = P_X + p_n \Rightarrow P_X = p_x - p_n$

Different measurement model: geometric mean $Y = \left(\prod_{i=1}^{N} X_i\right)^{1/N}$



- Actual experiment: $Y = 1.998$
- The GUM2 interval has $LRSR = 0$
- An interval with $LRSR = 95\%$ is easily computed using a "frequentist" approach




```
Results from GUM uncertainty framework:
y     = 1.97447
u(y)  = 0.00591626
I(y)  = 1.96288, 1.98607

Results from a Monte Carlo method:
y     = 1.95197
u(y)  = 0.00594091
I(y)  = 1.94028, 1.9636

MCM calculation has stabilized
GUF is NOT validated against MCM
>> geometric_mean(X)
ans =
   1.99830230557388
>>
```

LRSR=0: is it a problem?

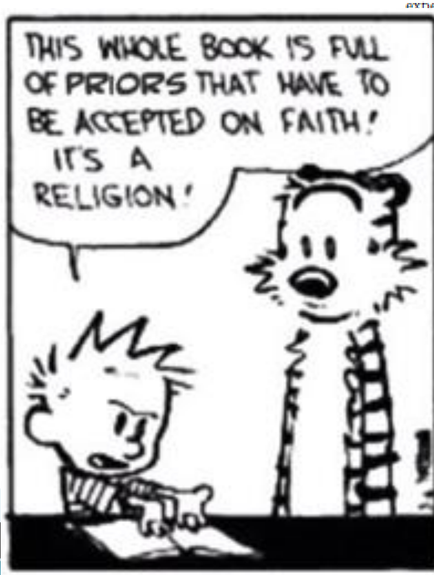
- A good Bayesian does **not** require $LRSR = p$

- But a good Bayesian **does not accept** $LRSR = 0$

  (remember the words of Berger; see also J. M. Bernardo)

- When $LRSR = 0$, the prior pdf **must be changed**, in a **problem-dependent way** («true» Bayesianism)

- Yes, it is a problem



J. Statistics Planning and Inference (1997) vol. 65

Noninformative priors do not exist: A discussion

J. M. Bernardo

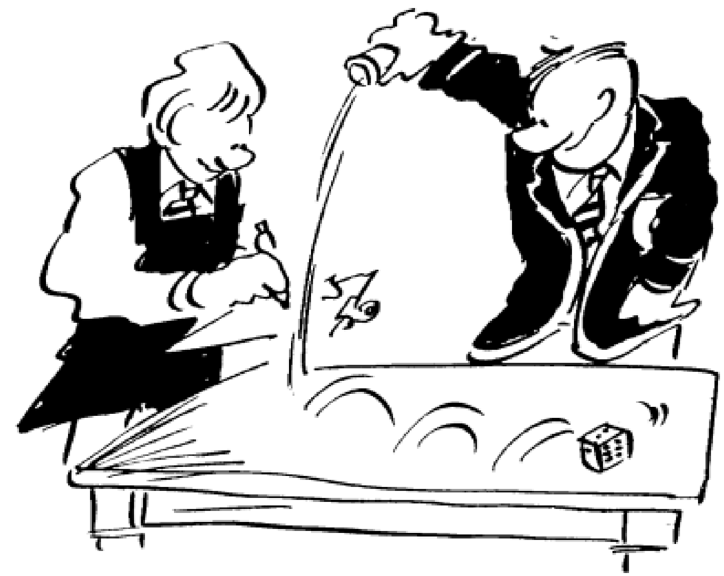
An important part of the discussion on methods for deriving non-subjective priors is based on the **analysis of the statistical properties of the posteriors they produce in specific, "test-case" examples**

Table 2 — Available knowledge of a quantity X, PDF for X on the basis of that knowledge, best estimate x of X and associated standard uncertainty $u(x)$ (7.3.2, 9.2.3 and 9.3.1)

Available knowledge	PDF and illustration (not to scale)	x and $u(x)$	Subclause
n indication values having average $\bar{\xi}$ and standard deviation s drawn independently from Gaussian PDF with unknown expectation and variance	Scaled and shifted t: $t_{n-1}(\bar{\xi}, s^2/n)$	$x = \bar{\xi}$, $u(x) = \left(\frac{n-1}{n-3}\right)^{1/2}\frac{s}{\sqrt{n}}$	9.2.2
...nts having average $\bar{q}$ ...i independently from ...on distribution with ...own expectation	Gamma: $G(n\bar{q}+1/2, 1/n)$	$x = \bar{q} + \frac{1}{2n}$, $u(x) = \left(\frac{\bar{q}}{n} + \frac{1}{4n^2}\right)^{1/2}$	9.2.8
...r and upper limits a, b	Rectangular: $R(a, b)$	$x = \frac{a+b}{2}$, $u(x) = \frac{b-a}{\sqrt{12}}$	9.3.2
...estimate x and standard ...tainty $u(x)$	Gaussian: $N(x, u^2(x))$	x, $u(x)$	9.3.3
...estimate x, expanded ...tainty U_p, coverage ... k_p and effective degrees ...edom ν_{eff} (> 2) obtained ...plying JCGM 100:2008	Scaled and shifted t: $t_{\nu_{\mathrm{eff}}}(x, \sigma^2)$, $\sigma^2 = \left(\frac{U_p}{k_p}\right)^2$	x, $u(x) = \left(\frac{\nu_{\mathrm{eff}}}{\nu_{\mathrm{eff}}-2}\right)^{1/2}\sigma$	9.3.4
...oidal cycling between ... a, b	Arc sine (U-shaped): $U(a, b)$	$x = \frac{a+b}{2}$, $u(x) = \frac{b-a}{\sqrt{8}}$	9.3.6
...estimate x of non-negative ...ity	Exponential: $\mathrm{Ex}(1/x)$	x, $u(x) = x$	9.3.7

Important note

- If the dice are not fair, and we get LRSR = 0, the problem is **not** in the methodology
- But if the dice are fair and we get LRSR = 0, the problem **is** in the methodology



Frequentist properties of pdf in GUM2: examples A, B

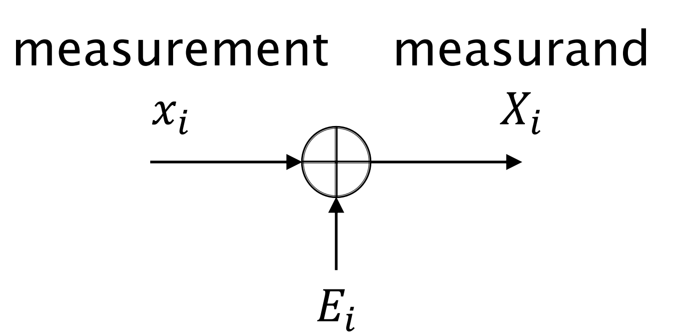
Examples A (dice), B (noise) show that:

- When measurands X_i are obtained summing random numbers E_i to fixed measurements x_i

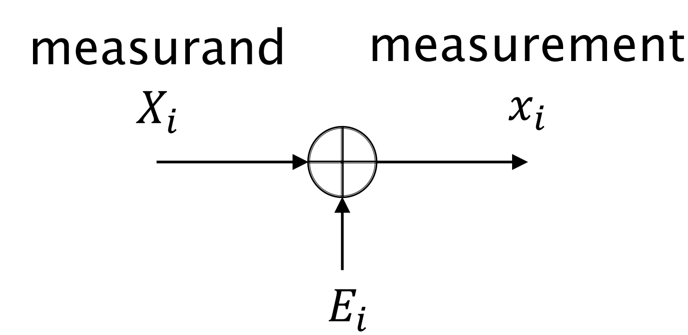
⇒ **the Bayesian pdf of Y is good** (btw it is also the frequency pdf of Y in repeated experiments)

- When measurements x_i are obtained summing random numbers E_i to fixed measurands X_i

⇒ **the Bayesian pdf of Y is bad**



GUM2 OK ☺

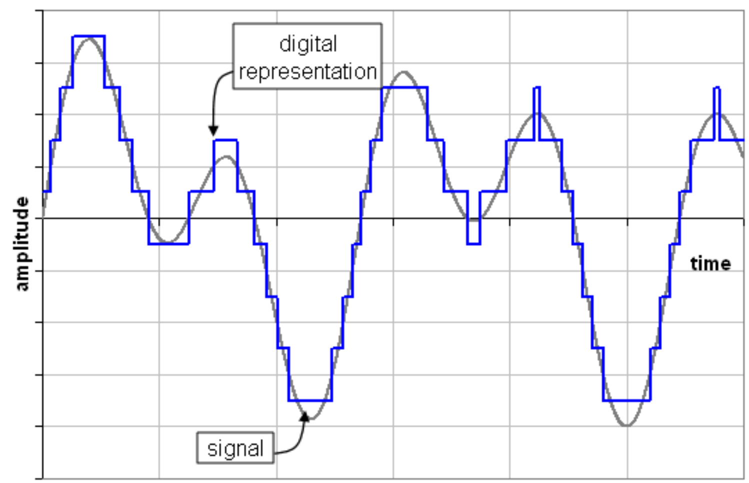


GUM2 KO ☹

Frequentist properties of pdf in GUM2 : examples C

In examples C1, C2 we do not have a «summation» mechanism

- We have different statistics for the quantization error
$$E_i = round(x_i) - X_i = x_i - X_i$$
- E_i, x_i are independent $\Rightarrow$ **the Bayesian pdf of Y is good** (C1, «special» input) and is also the frequency pdf of Y in repeated experiments
- E_i, x_i are not independent $\Rightarrow$ **the Bayesian pdf of Y is bad** (C1, «generic» input: E_i, X_i independent)



Conclusions

Possible immediate action

- Add to GUM2/S1 a statement

« »

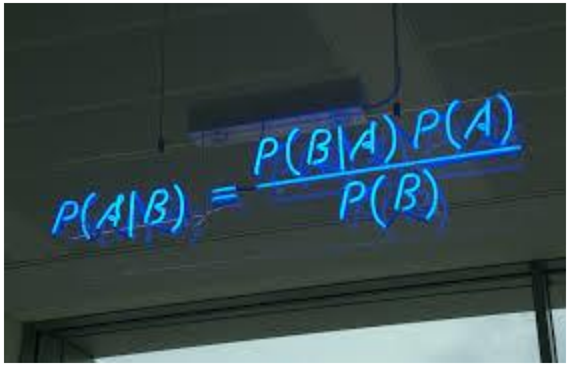
Long-run success rate of coverage intervals in repeated measurements

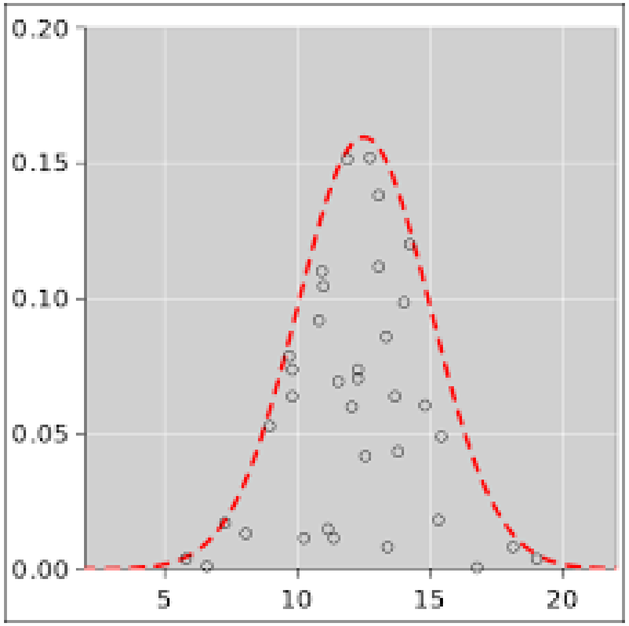
The long-run success rate of a coverage interval in repeated measurements **do not coincide**, in general, with the coverage probability p.

(LRSR may become arbitrarily low. Details may be added…)

Possible long-term actions

- Guidance to implement «true» Bayesian analysis
  - JCGM 108
- Guidance to assess LRSR of intervals / determine intervals with given LRSR
  - It can be done easily by Monte Carlo simulations
  - LRSR can be assessed for any interval, however obtained (also frequentist methods)
  - The statistics of $E_i = x_i - X_i$ is of essential importance here





One final suggestion

- Success rate is of practical interest, and intervals with assured LRSR have a market
- People use frequentist, Bayesian, and mixed approaches – all sort of tools
- Providing guidance to everybody is a lot of work, but it can be worth it

Thank you!