GUI for Bayesian sample size planning in type A uncertainty evaluation

OPEN ACCESS
IOP Publishing
Measurement Science and Technology
Meas. Sci. Technol. **32** (2021) 075005 (9pp)
https://doi.org/10.1088/1361-6501/abe2bd

GUI for Bayesian sample size planning in type A uncertainty evaluation

Jörg Martin* and **Clemens Elster**

Working group 8.42: Data Analysis and Measurement Uncertainty, Physikalisch-Technische Bundesanstalt, Abbestraße 2–12, Berlin 10587, Germany

E-mail: joerg.martin@ptb.de

Received 12 November 2020, revised 27 January 2021
Accepted for publication 3 February 2021
Published 30 April 2021

Abstract

We present a graphical user interface (GUI) for planning the sample size needed to reach a specified target uncertainty in a Bayesian type A uncertainty evaluation of normal or Poisson distributed data. To this end we build on a criterion previously introduced by Martin and Elster (2020 *Stat. Methods Appl.* 1–21) and called the variation of the posterior variance criterion. This criterion includes, and extends, standard Bayesian sample size planning procedures. Guidance is provided for the elicitation of the required prior knowledge in a way that makes the approach easily accessible for metrologists. The GUI also includes a menu that performs the Bayesian inference after the experiment has been carried out.

Keywords: sample size planning, experimental design, type a uncertainty, Bayesian statistics

1. Introduction

When planning a measurement series a natural question that arises is how many data need to be acquired. Taking too many samples can be costly or time-consuming, while taking too few can make the result of the measurement unusable. The quest for the appropriate sample size is known as sample size determination (SSD). In metrology, SSD can be relevant when planning the sample size needed to meet a specified target uncertainty in a type A (standard) uncertainty evaluation, that is an uncertainty evaluation based on observations [2].

SSD is part of experimental design, and usually performed before any measurements are taken. SSD utilizes prior knowledge about the experiment such as the expected dispersion of the data. While also frequentist SSD makes, often implicitly, use of prior knowledge [3, 4] the language of Bayesian statistics is a consistent approach for dealing with such knowledge [5–9], but diving into its depths might be deterrent for someone whose only goal is to find a sample size. To help bridging this gap this article presents an open source software package, written in Python and equipped with a graphical user interface (GUI), to do sample size planning based on a Bayesian criterion. The GUI also allows the final standard and expanded uncertainty to be evaluated once the data have been recorded. Explicit guidance is given to elicit the prior knowledge needed for applying the SSD.

The GUI is tailored for the following task: we want to determine a suitable length n of a planned measurement series $x_1, \ldots, x_n$, where we assume that the measurements satisfy the following conditions, in order to keep the framework and interface as accessible and straightforward as possible:

- The measurand is a scalar.
- The measurements are statistically independent.
- The measurements follow a normal or Poisson distribution.

The first of these assumptions is needed for the sample size criterion we introduce in (2) below, whereas the second and third assumption are taken to keep the GUI simple. To demonstrate the usage of our method we will use throughout this article a fictional experiment of determining the mean temperature in some environment, say a laboratory, within a specified target uncertainty, so that $x_1, \ldots, x_n$ represent

* Author to whom any correspondence should be addressed.

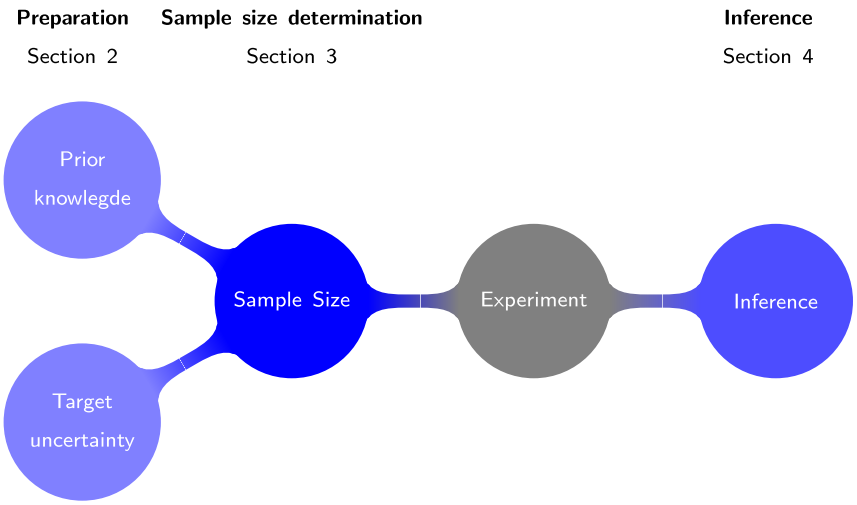

Figure 1. Overview over different phases of the sample size planning, the experiment and the subsequent inference. Stages treated in this article are blue. The actual sample size planning is treated in sections 2 and 3. Section 4 describes the evaluation of measurement results.

measurements of the temperature[1] at n different physical locations in this laboratory (measurements at the same location are typically easy to sample in large numbers and a SSD is therefore less relevant). The GUI is devoted to any measurements series of scalar quantities (e.g. mass or length) where the above applies, which is a quite common assumption (see the examples in the GUM [2]).

We will follow the idea [10–14] to design n such that we can specify the measurement result with a sufficient target uncertainty quoted ε in the GUI. More precisely we want to achieve that

$$u < \varepsilon, \tag{1}$$

where u is the Bayesian type A (standard) uncertainty for our measurement result[2]. A condition such as (1) could for instance become quite relevant in cases where u is the dominant source in a combined uncertainty. Unfortunately, u will depend on the measured values, whereas sample size planning is done *before the experiment is performed*. A way to overcome this cyclic dependency is to use prior knowledge, mathematically encoded in the language of Bayesian statistics [15]. A substitute for (1) in this spirit, proposed by the authors in [1], is a criterion called the variation of the posterior variance criterion (VPVC), whose formula is given by

$$\overline{u^2} + k \cdot \Delta u^2 < \varepsilon^2. \tag{2}$$

The precise definitions and formulas for the objects in (2) will not really matter for the purpose of this article and we refer to [1] for the statistical details. In a nutshell, $\overline{u^2}$ is the expected squared Bayesian type A standard uncertainty, the object Δu^2 its expected variation, both based on the prior knowledge, and the expansion factor $k \geqslant 0$ is a parameter to tune the chances that after the experiment (1) will be satisfied. For $k = 0$ the criterion in (2) reduces to the standard Bayesian SSD criterion known in the literature as the average of the posterior variance criterion (APVC) [10–12].

This article presents a software package, written in Python [16], that spares the user most of the statistical details, and thereby provides an easy access to a topic that appears to be rather seldom treated in the metrology literature. In addition, the package comes with an implementation of the VPVC sample size planning and a GUI that can be used without any foreknowledge of Python.

The structure of this article together with the typical flow of sample size planning, experiment and subsequent inference is sketched in figure 1. Section 2 explains how prior knowledge and the desired target uncertainty, namely ε in (1) and (2), can be recorded in the sample size GUI. Furthermore, explicit guidance is provided for eliciting the required prior knowledge. Section 3 then describes how the according sample size can be gained from this information. Section 4 finally explains how a Bayesian estimate of the measurand together with its uncertainty can be obtained, after the experiment has been performed, using the GUI.

A summarized Howto for using the GUI and a detailed description of the output of the included inference menu is included in appendices A.1 and A.4. Moreover, an additional example with Poisson distributed data is also included in appendix A.2. For the mathematically inclined reader some mathematical details behind the Bayesian inference are sketched in appendix A.3.

2. Preparation

Subsequent to following the instructions in the `README`, the GUI can be launched by entering `python -m vpvc_gui` in the corresponding Python or Anaconda environment. This will open a window, similar to the one displayed in figure 2. In the menu `Data distribution` in the leftmost column of the

[1] To unify the notation in this article and make it consistent with the one on the GUI, we here use x as a letter to represent the measured temperature values in contrast to the usual convention of calling the temperature 'T'.

[2] To be mathematically precise: u denotes the standard deviation of the posterior distribution of the measurand, see [1] or appendix A.3.

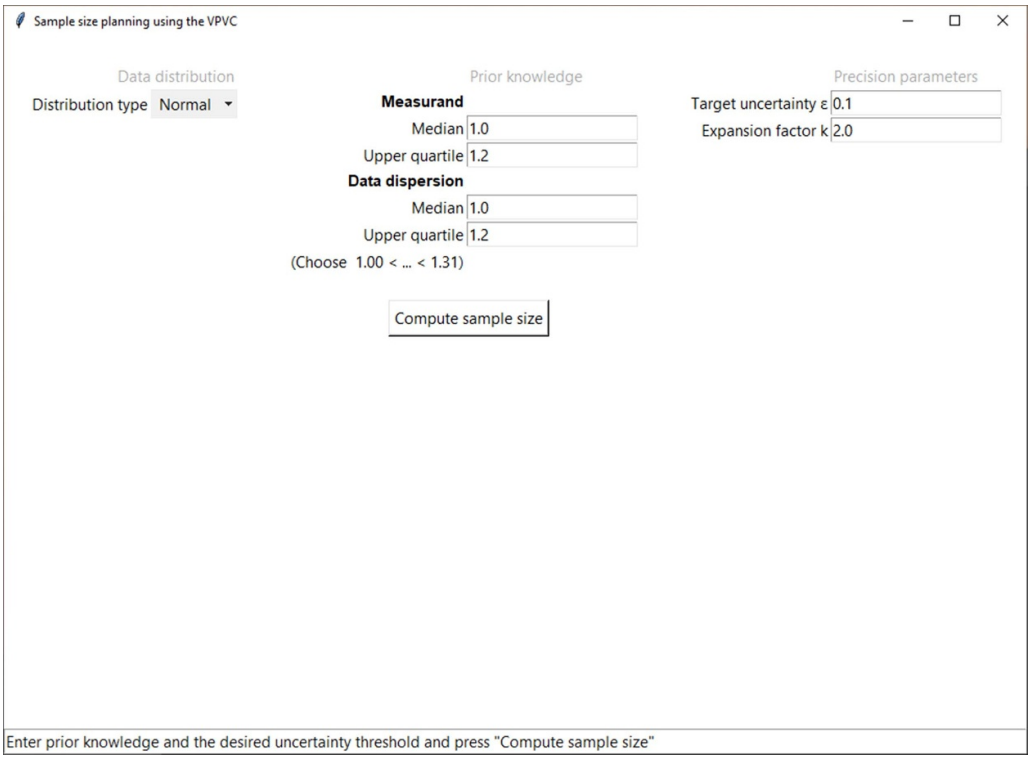

Figure 2. The GUI after starting it. On the leftmost side the distribution of the measurement data can be selected. The second and third column contain boxes where the prior knowledge (menu `Prior knowledge`, section 2.1) and parameters describing the precision of the sample size determination (menu `Precision parameters`, section 2.2) ought to be entered. Actions and error messages are shown by the status bar on the bottom.

window the distribution of the measured data can be selected to be either `Normal` or `Poisson`. For our toy example of temperature measurements we will stick to `Normal`, an example with a Poisson distribution is given in appendix A.2. The other two columns, which logically belong to the left hand side of figure 1, contain several boxes where (arbitrary) default values appear. Replacing these numbers with meaningful values is the prerequisite for determining the sample size and will be the content of this section.

Note also the status bar on the bottom of the window, where expected actions and error messages will be displayed.

2.1. Elicitation of prior knowledge

This section provides explicit guidance for the elicitation of prior knowledge that needs to be filled in the menu `Prior knowledge` in the second column of the interface displayed in figure 2. Prior knowledge is often gained from expert knowledge, or can be obtained from previous measurements. Specifically, we need prior knowledge concerning two different quantities:

- The measurand, the quantity of interest.
- The dispersion (standard deviation) of the data.

In the example above the measurand would be the mean temperature in the environment, whereas the dispersion is the fluctuation of the measurement results due to random influences such as noise in the physical environment or within the measurement device. For both quantities, the measurand and the dispersion, two numbers have to be specified that quantify what is known about them. We here follow the recommendations from [17, 18] and use the following:

- The median: estimate a value so that is equally likely that the quantity is above and below.
- The upper quartile (75% quantile): suppose the quantity is above the median, estimate a threshold which makes it equally likely that the quantity falls below or above this threshold in such a scenario.

In total we therefore need four numbers that have to be specified for the prior knowledge: the median of the measurand, the upper quartile for the measurand, the median of the dispersion and the upper quartile of the dispersion. Let us come back to the mean temperature example we introduced in section 1. Suppose that we know that the temperature in the laboratory is usually around 20 °C. In fact we would consider it equally likely that the temperature in the room is above or below this number, so that it seems natural to pick the median of the measurand as 20 °C. We further consider it equally likely that the temperature is between 20 °C and 20.5 °C or above 20.5 °C and pick therefore 20.5 °C as an upper quartile. Note that this does *not* mean that 20.5 °C − 20 °C = 0.5 °C is the uncertainty about the temperature. Figure 3 illustrates our choice of prior knowledge. The left plot shows the actual probability

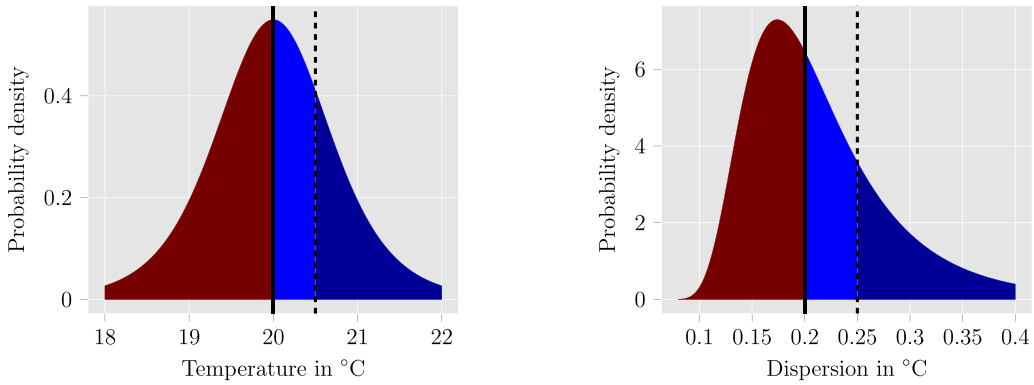

Figure 3. Probability distributions that are used by the mathematical model in the background for the prior knowledge from section 2.1. The vertical lines show the medians (solid) and upper quartiles (dashed). For mathematical details on the used distributions compare appendix A.3. Areas separated by vertical lines are marked by different colors. The blue areas are equal in size and their sum equals the red area. Note that some parts of the distribution are not shown for the sake of depiction.

distribution of the measurand (the mean temperature) that is used by the mathematical model in the background for the prior knowledge specified in this section. The median of 20 °C is marked by a solid line and the upper quartile of 20.5 °C by a dashed line. Note that the red area has the same size as the sum of the two blue ones and that, moreover, both blue areas are equal in size. The reason for using the upper quartile instead of the uncertainty is that it is often more intuitive to guess. Following [17, 18] a good approach to estimate the upper quartile is to start from a value M, larger than the median, of which one is quite confident that the true value of the quantity will be below and to reduce M until it seem equally likely that the quantity is above or between the median and M. In the GUI displayed in figure 2 we can now enter in `Prior knowledge` below `Measurand` the median and upper quartile.

For the dispersion we proceed in a similar manner with one important exception. For mathematical reasons the approach introduced in [1] needs a prior knowledge of the dispersion that is 'precise enough', namely the upper quartile of the dispersion should be below a threshold that depends on the specified median of the dispersion[3]. Below the last box of the prior menu there are parentheses with an entry of the shape `(Choose [...] < ... > [...] )` where the precise value of the numbers [...] will depend on the value entered for the median of the dispersion. The entry for *upper quartile* for the dispersion must be between the two limits specified there (the lower limit is always simply the median). We here assume that we know from using the measurement device in the past that results typically vary by 0.2 °C which we take as median together with an upper quartile of 0.25 (below the maximal possible value of 0.26 in this case). The right plot of figure 3 shows the according probability distribution encoded by this median and upper quartile. Note that the median is not the maximum (mode) of the distribution but really marks the point that splits the area below the curve in two areas of the same size.

[3] The mathematical reason for this is that the distribution that is used to model the prior knowledge about the dispersion needs a variance in order for the VPVC to work for $k > 0$, see [1] for details.

2.2. *Specifying the target uncertainty*

This section discusses how to fill in the menu `Precision parameters` in the window displayed in figure 2. Once the prior knowledge is specified, two more ingredients are needed:

- The target uncertainty ε.
- The expansion factor k of the VPVC in (2).

The aim of the sample size determination based on the VPVC is that the final standard uncertainty is below the target uncertainty ε. However, the likeliness depends on the expansion factor k. By default the GUI will put $k = 2.0$. Loosely speaking, the object Δu^2 in (2) describes a standard deviation of u^2 (under a suitable distribution), so that the corresponding term in (2) can, in pure heuristics, be read as a '2σ' error. In [1] this was observed as a choice that performed quite good in practice. For a less stringent sample size planning one might pick $k = 1.0$. Finally, for $k = 0$ one recovers the APVC sample size planning [10].

For the temperature example we will use the default value of $k = 2.0$ and $\varepsilon = 0.1$. We can thus keep the default values chosen by the GUI.

3. Determination of the sample size

Once the prior knowledge is entered and the target uncertainty chosen, we are ready to determine the required sample size. To do this in the GUI the button saying `Compute sample size` must be pressed. If there is an error—the status bar indicating `Invalid values encountered`—this might be due to an upper quartile for the dispersion that is beyond the indicated limits, negative values for inappropriate quantities or an entry that cannot be interpreted as a float. If there are no errors the window will update as in figure 4. The sample size computed via the VPVC (2) will be printed below the button together with a plot showing the dependency of the VPVC criterion (2) on n (in blue). The plot displays in addition a line showing ε (horizontal, dashed) and the computed sample size (vertical, dotted). In the right lower corner there is a menu `Data`

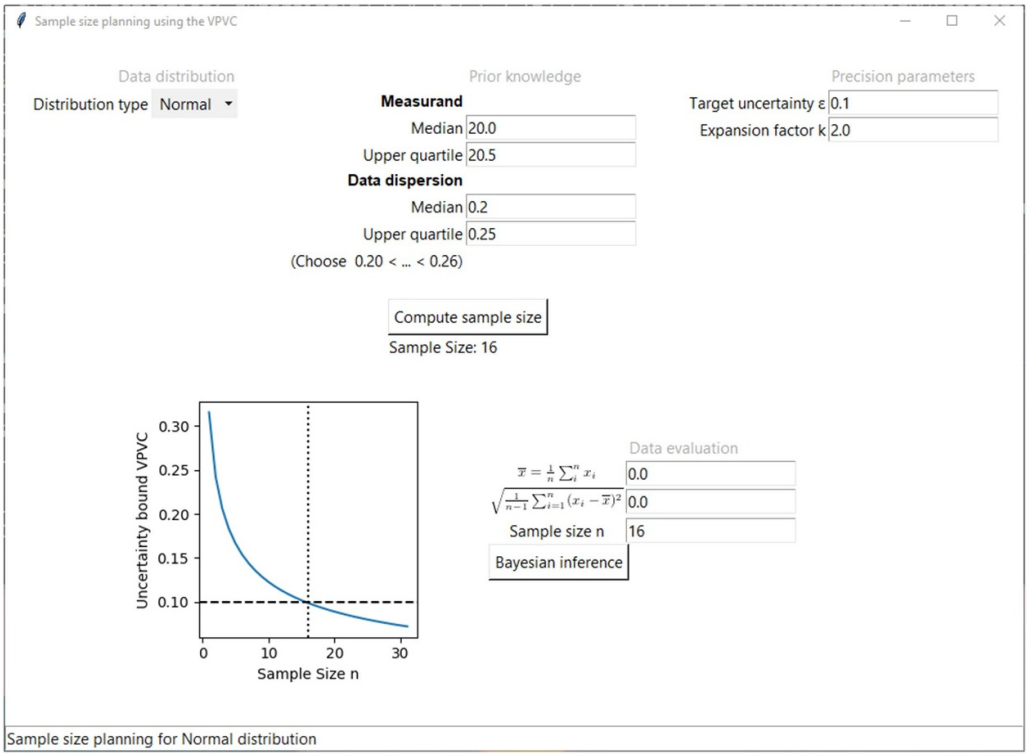

Figure 4. The graphical user interface once the prior knowledge from section 2.1 and the target uncertainty from section 2.2 were entered and the button `Compute sample size` has been pressed. Besides the sample size and a plot of the left hand side of (2) the window also contains a menu for inference in the lower right corner. Its usage will be described in section 4.

`evaluation` plotted which can be used to compute the estimate of the measurand together with its uncertainty once the experiment was performed. We will give more details on this in section 4 below.

Let us sketch the qualitative dependency of the sample size on the values entered in section 2. Understanding this might help to judge whether the computed sample size is too optimistic or restrictive and what could be done if the sample size is intolerably high:

- Median of the measurand: this quantity has *no* influence on the sample size.
- Upper quartile of the measurand: increasing this value will increase the sample size. However, as was observed in [1], the influence of this quantity on the sample size is often rather small.
- Median of the dispersion: increasing this value will increase the sample size.
- Upper quartile of the dispersion: increasing this value will increase the sample size. Due to the restriction posed by the mathematical model, see section 2.1, this number can however not be arbitrarily increased.
- Target uncertainty ε and expansion factor k: both, a smaller ε and a larger k, will lead to a larger sample size, compare also (3) below.

Concerning the prior knowledge it was demonstrated in [1] that the most conservative, that is highest, sample size is produced for a high median of the dispersion and high upper quartiles of measurand and dispersion.

The plot included in the GUI demonstrates that the left hand of side of (2) decays with an inverse power of the sample size n. To be more precise, the authors showed in [1, lemma 3.2] that for small ε (large n) the left hand side of (2) roughly scales like $\frac{a+k\cdot b}{n}$ and the computed sample size consequently as [1, lemma 3.4]:

$$n \simeq \frac{a + b \cdot k}{\varepsilon^2}, \tag{3}$$

where a and b are positive constants that depend on the specified prior knowledge.

Coming back to the temperature example we see that for the values we entered in section 2 we obtain a sample size of 16. Taking the less cautious choice $k = 1$ leads to $n = 11$ and the APVC ($k = 0$) to a sample size of 6.

4. Inference after the experiment

We now assume that the experiment has been performed and that n measurements $x_1, \ldots, x_n$ were taken. In the toy example we use throughout this article, and with the sample size from section 3, this would be a collection of $n = 16$ temperature measurements. For the sake of presentation we use some simulated data. To visualize different scenarios we consider three datasets consisting each of $n = 16$ 'measurements', all of them

Table 1. Three simulated datasets used to demonstrate the usage of the GUI in section 4. For the GUI the sufficient statistics (4) have to be computed, cf. figure 5. The computed Bayesian estimate and uncertainty are listed in the last two rows. All values are in in °C.

Dataset		1	2	3
	x_1	19.853	21.448	19.645
	x_2	19.580	20.938	19.208
	x_3	19.696	21.094	19.861
	x_4	19.948	20.744	20.063
	x_5	19.874	20.234	20.492
	x_6	19.305	21.196	20.481
	x_7	19.690	21.259	19.845
Raw	x_7	19.470	20.777	19.879
data	x_8	19.479	21.681	19.581
	x_{10}	19.582	20.564	19.432
	x_{11}	19.529	21.014	19.317
	x_{12}	19.791	20.944	20.780
	x_{13}	19.652	21.460	19.796
	x_{14}	19.524	21.441	19.825
	x_{15}	19.589	21.046	19.499
	x_{16}	19.567	21.113	20.311
Input	$\bar{x}$	19.633	21.060	19.876
to program	s	0.164	0.357	0.436
Output	Estimate	19.635	21.055	19.876
of program	uncertainty	0.044	0.083	0.098

are listed in table 1. For generation we used a normal distribution with mean $\mu = 19.5$ °C and standard deviation $\sigma = 0.2$ °C (dataset 1), $\mu = 21.0$ °C, $\sigma = 0.3$ °C (dataset 2) and $\mu = 20.0$ °C, $\sigma = 0.4$ °C (dataset 3).

To do inference on these datasets, that is to determine an estimate of the measurand together with an uncertainty, the menu `Data evaluation` can be used. A cutout is shown in figure 5. To use this menu the full dataset $x_1, \ldots, x_n$ is actually not needed, but only two scalar quantities, the sufficient statistics for the normal distribution:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,, \qquad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\,, \tag{4}$$

which are the mean and standard deviation of the measured values $x_1, \ldots, x_n$. For the minimal sample size $n = 1$ the standard deviation s is ill-defined but is not used in the computation, so that an arbitrary value such as 0 can be taken. In table 1 we included the sufficient statistics for each of the datasets. In the GUI both quantities can be entered in according boxes. The program assumes by default that the computed sample size was used. If the sufficient statistics were determined from a set of measurements of a different sample size, the box `Sample size n` can be adjusted. Once the sufficient statistics in (4) are entered, the button `Bayesian inference` can be pressed and should return the (Bayesian) estimate and uncertainty using the prior knowledge entered in the program unless invalid values were entered (such as a negative s)[4]. The results include

[4] To be precise, the mean and standard deviation of the posterior distribution of the measurand are computed, see [1, 15].

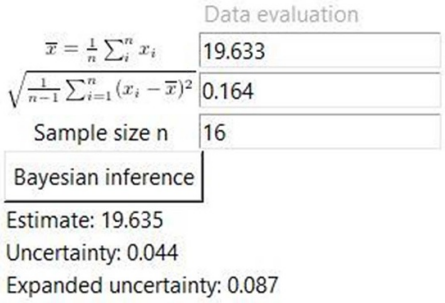

Figure 5. Cutout of the GUI window (figure 4) after the sufficient statistics from (4) have been entered and the button `Compute sample size` has been pressed. Below the button the Bayesian estimate and standard uncertainty of the measurand are displayed. In addition, the GUI also displays the expanded 95% uncertainty, compare appendix A.4 for details.

the Bayesian estimate and (standard) uncertainty of the measurand and, in addition, the expanded 95% uncertainty. All three output values are rounded to two significant digits of the standard uncertainty. A more detailed description of the output quantities is included in appendix A.4 of this article. Moreover, the precise formulas for the Bayesian estimate and uncertainty can be found appendix A.3, which illustrate in particular that the Bayesian inference combines $\bar{x}$ (for the estimate) and $\frac{s}{\sqrt{n}}$ (for the uncertainty) with the prior knowledge in a sort of weighted average.

In table 1 we included the computed estimate and standard uncertainty for each of the three datasets. Recall that the actual dispersion σ of the data increases from the left to the right. If we see (1), that is $u < \varepsilon$ with $\varepsilon = 0.1$, as a measure of success for the sample size planning then we can observe that succeeding essentially depends on whether σ is markedly larger than the value of the dispersion we specified by the median and upper quartile in the prior knowledge (recall that we chose 0.2 as median and 0.25 as upper quartile). The same observation was also found in [1]. We observe that only for the dataset 3, which has a substantially larger dispersion of 0.4, the standard uncertainty starts to scratch the threshold $\varepsilon = 0.1$.

5. Conclusion

Guidance for the usage of the variation of the posterior variance criterion (VPVC) for sample size planning was given. This criterion uses Bayesian inference to incorporate prior knowledge. To make the method accessible for users unfamiliar with Bayesian statistics a Python package with an easy-to-use graphical user interface (GUI) is offered. Detailed instructions for the usage and behavior of this GUI were provided as well as background information for the underlying criterion. To perform Bayesian inference after the measurements have been taken a small menu is included on the GUI where the sufficient statistics can be entered. The GUI can help metrologists in sample size planning and provides at the same time a simple-to-use software for the incorporation of prior knowledge into a Bayesian type A uncertainty evaluation.

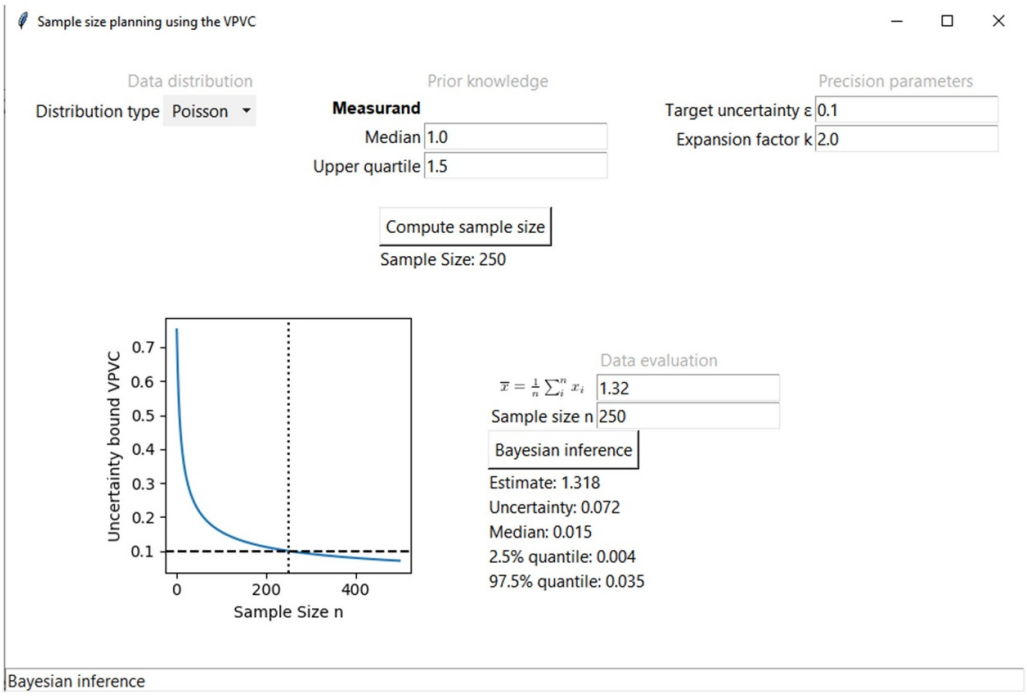

Figure 6. The GUI for sample size planning for Poisson distributed data, with values as in section A.2.

Appendix A

A.1. Howto for using the GUI

Step 1 Select distribution (Normal/Poisson);
Step 2 Specify prior knowledge, i.e.
- Median and upper quartile of measurand[5];
- Median and upper quartile of data dispersion (only for normal case);

Step 3 Enter target uncertainty ε;
Step 4 Press button `Compute sample size`;
Step 5 Carry out experiment and collect data;
Step 6 Enter mean (and standard deviation) of data;
Step 7 Press button `Bayesian inference`.

A.2. A Poisson example

Consider the number x of decays in a probe of radioactive material in a time interval of length τ. It is common to assume that x is governed by a Poisson distribution [19]:

$$x \sim \mathrm{Poi}(\lambda),$$

where $\mathrm{Poi}(\lambda)$ denotes the Poisson distribution with parameter λ. We want to determine n so that n measured samples yield λ with an uncertainty below a pre-specified target uncertainty ε.

[5] Median and upper quartile refer to two parameters of a distribution that encodes the *a priori* degree of belief about the value of the measurand (or the dispersion of the data, respectively). The *a priori* probabilities for the measurand to be greater or less than the median are both 50%. Similarly, the upper quartile splits the domain of values for the measurand above the median into two equally probable subregions, see figure 3.

We then proceed similar as described for the normal example in the article or in the Howto in appendix A.1.

A.2.1. Specify prior knowledge and desired precision From experience or rough guessing we know that probes of the considered kind yield roughly around one decay per time interval τ and enter this as median of the prior knowledge in the GUI, see figure 6. We judge it to be equally likely that there are between 1.0 and 1.5 counts per time τ and more than 1.5, so that we take 1.5 as an upper quartile. We want to determine λ with an uncertainty below 0.1 and set the target uncertainty ε accordingly. Finally, we leave the default expansion factor $k = 2.0$ for the VPVC as is.

A.2.2. Sample size planning and inference Pressing `Compute sample size` reveals a n of 250.

Suppose that after performing the experiment and observing 250 intervals of time τ we get an average of $\overline{x} = \frac{1}{250}\sum_{i=1}^{250} x_i = 1.32$ counts. Pressing then `Bayesian inference` yields the estimate 1.318 together with an uncertainty of 0.072, which is safely below the target uncertainty of 0.1. Note that in contrast to the normal case the GUI also returns the median and 2.5% and 97.5% quantiles to give a more detailed description of the skew posterior distribution, see A.4 and [1] for details.

A.3. Mathematical background (Bayesian inference—normal distribution case)

This appendix sketches the mathematics behind the Bayesian inference for the case of normal distributed data and is *not*

needed for the usage of the GUI. For the Poisson case, we refer to [1, 20].

Assuming a data distribution $p(x_1,\ldots,x_n|\mu,\sigma^2) = \prod_{i=1}^{n}\mathcal{N}(x_i|\mu,\sigma^2)$, with the parameter of interest μ and the nuisance parameter σ^2, as well as a normal inverse Gamma prior $\pi(\mu,\sigma^2) = \mathcal{N}(\mu|\mu_0,\lambda\sigma^2)\cdot \mathrm{IG}(\sigma^2|\alpha,\beta)$ with hyperparameters μ_0 and $\lambda,\beta>0,\alpha>2$ the posterior distribution for μ is given by

$$\pi(\mu|x_1,\ldots x_n) \propto \int \pi(\mu,\sigma^2)p(x|\mu,\sigma^2)\mathrm{d}\sigma^2,$$

due to Bayes' theorem and marginalization, which can be explicitly specified as a shifted and scaled t-distribution [21]:

$$\pi(\mu|x_1,\ldots,x_n) = t_{2\alpha'}\left(\mu_0', \frac{\beta'}{n_\lambda\alpha'}\right),$$

where $n_\lambda = n+\lambda^{-1}$, $\mu_0' = \frac{1}{\lambda n_\lambda}\mu_0 + \frac{n}{n_\lambda}\bar{x}$, $\alpha' = \alpha + \frac{n}{2}$ and $\beta' = \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2 + \frac{n}{2\lambda n_\lambda}(\bar{x}-\mu_0)^2$ and $\bar{x}$ is as in (4). The posterior mean and variance are then

$$E_{\mu\sim\pi(\mu|x_1,\ldots,x_n)}[\mu] = \frac{n}{n_\lambda}\bar{x} + \frac{1}{\lambda n_\lambda}\mu_0, \tag{5}$$

$$u^2 = \mathrm{Var}_{\mu\sim\pi(\mu|x_1,\ldots,x_n)}(\mu) = \frac{n(n-1)}{n_\lambda(n+2\alpha-2)}\frac{s^2}{n} + \frac{2\beta+\frac{n}{\lambda n_\lambda}}{n_\lambda(n+2\alpha-2)}(\bar{x}-\mu_0)^2, \tag{6}$$

with s^2 as in (4).

A.4. More detailed description of the inference results

All quantities displayed below the `Bayesian inference` button relate to the marginal posterior distribution of the measurand. For details on the used priors we refer to appendix A.3 and [1].

We will list the meaning of the displayed values in dependency of the used data distribution.

A.4.1. Normal

- `Estimate`: The *mean* of the posterior distribution of the measurand. As the latter is a symmetric t-distribution, this number is equal to the median.
- `Uncertainty`: This is the Bayesian type A standard uncertainty, i.e. the *standard deviation* of the posterior distribution of the measurand.
- `Expanded uncertainty`: The *distance of the 97.5% quantile to the median*. As the posterior distribution is symmetric, this is identical to the distance of the 2.5% quantile to the median. Note that the coverage factor, that is the quotient between `Expanded uncertainty` and `Uncertainty`, will depend on the prior knowledge.

A.4.2. Poisson distribution

- `Estimate`: The *mean* of the posterior distribution. As the latter is not symmetric, this number will in general be different from the `median` value.
- `Uncertainty`: The *standard deviation* of the posterior distribution of the measurand.
- `Median`: The *median* of the posterior distribution of the measurand.
- `2.5% quantile` and `97.5% quantile`: In contrast to the normal distribution case, the distance between these quantiles to the median is in general not equal so that no simple expanded uncertainty can be specified.

Code details

The code for the GUI including installation instructions can be found in the repository: https://gitlab1.ptb.de/JoergMartin/vpvc_sample_size.git.

ORCID iD

Jörg Martin https://orcid.org/0000-0001-5066-7661

References

[1] Martin J and Elster C 2020 The variation of the posterior variance and Bayesian sample size determination *Stat. Methods Appl.* 1–21
[2] BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, OIML 1995 Guide to the expression of uncertainty in measurement (GUM) (International Organization for Standardization)
[3] Desu M 2012 *Sample Size Methodology* (Amsterdam: Elsevier)
[4] Adcock C 1997 Sample size determination: a review *J. R. Stat. Soc.* D **46** 261–83
[5] Toman B 2007 Bayesian approaches to calculating a reference value in key comparison experiments *Technometrics* **49** 81–7
[6] van der Veen A M H 2018 Bayesian methods for type A evaluation of standard uncertainty *Metrologia* **55** 670
[7] Elster C 2014 Bayesian uncertainty analysis compared with the application of the GUM and its supplements *Metrologia* **51** S159
[8] Wübbeler G, Schmähling F, Beyer J, Engert J and Elster C 2012 Analysis of magnetic field fluctuation thermometry using Bayesian inference *Meas. Sci. Technol.* **23** 125004
[9] Martin J, Bartl G and Elster C 2019 Application of Bayesian model averaging to the determination of thermal expansion of single-crystal silicon *Meas. Sci. Technol.* **30** 045012
[10] Wang F and Gelfand A E *et al* 2002 A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models *Stat. Sci.* **17** 193–208
[11] Pham-Gia T and Turkkan N 1992 Sample size determination in Bayesian analysis *J. R. Stat. Soc.* D **41** 389–97
[12] De Santis F 2007 Using historical data for Bayesian sample size determination *J. R. Stat. Soc.* D **170** 95–113
[13] M'lan C E *et al* 2008 Bayesian sample size determination for binomial proportions *Bayesian Anal.* **3** 269–96

[14] Joseph L and Bélisle P 2019 Bayesian consensus-based sample size criteria for binomial proportions *Stat. Med.* **38** 4566–73
[15] Robert C 2007 *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation* (Springer Science & Business Media)
[16] Van Rossum G and Drake Jr F L 1995 *Python Reference Manual* (Amsterdam: Centrum voor Wiskunde en Informatica)
[17] Oakley J E and O'Hagan A 2016 SHELF: the Sheffield elicitation framework (version 4.0) (UK: School of Mathematics and Statistics, University of Sheffield) (available at: http://tonyohagan.co.uk/shelf) (Accessed: 1 October 2020)
[18] Gosling J P 2018 Shelf: the Sheffield elicitation framework *Elicitation* (Berlin: Springer) pp 61–93
[19] Arfken G B, Weber H J and Harris F E 2012 *Mathematical Methods for Physicists* 7 (Amsterdam: Academic)
[20] Fink D 1997 *A Compendium of Conjugate Priors* vol 46 (available at: www.people.cornell.edu/pages/df36/CONJINTRnew%20TEX.pdf)
[21] Murphy K P 2007 Conjugate Bayesian analysis of the Gaussian distribution *Def* **1** 16