Calibration - EVPI

Wael Mohammed

January 2024

1 Introduction

Calibration target data is one of the critical elements of any calibration process. However, calibration targets are prone to precision and/or validity issues like any data.

Commonly, we consider the data, denoted as $y = \{y_1, \ldots, y_n\}$, to be a random collection of observations drawn from a larger reference population that consists of N units, where N ¿ n. The inherent uncertainty associated with the observed variables stems from the fact that these represent just a single sample out of the many possible samples, leading to what is known as sampling variability.

Bias, on the other hand, refers to the systematic or consistent error in measurements or estimates. It can lead to inaccuracies in the results and could result from flaws in study design, data collection methods, or other systematic errors.

The Standard error (SE) measures the precision of a statistical estimate. Specifically, it quantifies the variability or imprecision of a sample statistic, such as the mean or proportion, compared to the true population parameter.

In a set of competing options, D, the optimal choice for decision-makers like NICE is that associated with the highest "utility". The net benefit (NB) is a utility function given by the following notation for each choice or option d.

$$\mathrm{NB}_d(\theta)$$

$\theta = (\theta_1, ..., \theta_2)$ is the model input parameters representing real-world quantities. Since observed data (parameters) may not be known with certainty, we are uncertain about the value(s) of θ and, therefore, uncertain about the NB. Because of the uncertainty associated with the NB, the optimal choice is determined based on maximising the *expected* NB.

$$\mathbb{E}\{\mathrm{NB}_d(\theta)\}$$

Which for the set of options, D, becomes:

$$\max_{\mathrm{d}\in\mathrm{D}}\mathbb{E}\{\mathrm{NB}_d(\theta)\}$$

We quantify the Expected Value of Perfect Information (EVPI) as the value of reducing uncertainty in decision-making. The EVPI represents the maximum amount a decision-maker would be willing to pay for perfect information that eliminates all uncertainty associated with their decision, reducing the SE to zero. The EVPI is given by:

$$EVPI = \mathbb{E}\{\max_{\mathrm{d}\in\mathrm{D}}\mathrm{NB}_d(\theta)\} - \max_{\mathrm{d}\in\mathrm{D}}\mathbb{E}\{\mathrm{NB}_d(\theta)\}$$

2 Model

We adopt a simple "toy" model with two unknown x_1 and x_2. The following equations illustrate this model.

$$\begin{aligned}\text{Mortality} &= x_1 + x_2\\ \text{iNB} &= x_1\end{aligned}$$

This model reports the incremental NB (iNB). Therefore, the most the decision maker should consider paying to reduce the uncertainty, or EVPI, is now:

$$EVPI = \mathbb{E}\{\max(\mathrm{iNB}(\theta), 0)\} - \max(\mathbb{E}\{\mathrm{iNB}(\theta)\}, 0)$$

3 Priors

The unknowns x_1 and x_2 follow a bivariate Gaussian distribution with a mean μ and variance-co-variance matrix Σ.

$$\mu = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}$$

$$\Sigma = \begin{pmatrix} v_1 & \rho \times v_1 \times v_2 \\ \rho \times v_1 \times v_2 & v_2 \end{pmatrix}$$

Where: m_1 and m_2 represent the means of x_1 and x_2, v_1 and v_2 represent the variances of x_1 and x_2, and ρ is the correlation between the two unknowns.

4 Likelihood

4.1 1. First scenario

We start with a simple assumption where calibration is not needed, x_1 is observed, and the data-generating mechanism follows a Gaussian distribution. The

likelihood function in this scenario is given by:

$$L(m_1, v_1|x_{11}, x_{12}, \ldots, x_{1n}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi v_1}} \exp\left(-\frac{(x_{1i} - m_1)^2}{2v_1}\right)$$

4.2 2. Second scenario

The second assumption is that the only observed data in this exercise is $Mortality$ data. This data is assumed to follow a Normal distribution. Hence, the data-generating function can be given by:

$$L(x, \sigma^2|y_1, y_2, \ldots, y_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - L^T x)^2}{2\sigma^2}\right)$$

where

$$y = L^T x + \varepsilon = Mortality = x_1 + x_2$$
$$L = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$
$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$
$$\varepsilon \sim N(0, \sigma^2)$$

5 Pete notes

5.1 The problem

Prior

$$x \sim N(\mu, \Sigma)$$

where

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

Measurement

$$y = L^T x + \varepsilon$$
$$\varepsilon \sim N(0, s_N^2)$$

Decision is to use intervention if $x_1 > 0$ and keep status quo otherwise.
The incremental net benefit is

$$INB(x) = x_1.\theta(x_1)$$

where

$$\theta(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

5.2 A useful integral

Using integration by parts:

$$\int_0^\infty \mathrm{d}x\ x.N(x; m, s^2) =$$

$$s^2.N(0, m, s^2) + m.\left[1 - \Phi\left(-\frac{m}{s}\right)\right]$$

$$\stackrel{\text{def}}{=} f(m, s) \tag{1}$$

where $N(x; m, s^2)$ is the density of a normal distribution and $\Phi(x)$ is the cdf of the standard normal distribution, and the last equation serves to define the function f.

TODO: check correct

5.3 Reduced normals

See the MVN wikipedia page for how to calculate the marginal of a multivariate normal.

To calculate the distribution of x conditional on observing measurement y, $P(x|y)$, note

$$P(x\&y) \propto \exp\left[-Q(x_1, x_2)/2\right]$$

where

$$Q = \frac{(L^T x - y)^2}{s_N^2} + (x - \mu)^T \Sigma^{-1}(x - \mu) = (x - a)^T M^{-1}(x - a)$$

where the last equality defines M and a. Comparing coefficients implies:

$$M^{-1} = \Sigma^{-1} + LL^T/s_N^2$$

$$a = M.\left(\Sigma^{-1}\mu + Ly/s_N^2\right)$$

So we have found

$$P(x|y) = N(x; a, M) \tag{2}$$

TODO check (e.g. probably on www somewhere)

5.4 EVSI

I think EVSI is define by the following (?):

$$EVSI(\mu, \rho, \sigma_1^2, \sigma_2^2, L, y, s_N^2) = \mathbb{E}[INB(x)|y] - \mathbb{E}[INB(x)] =$$

$$\int_{\mathbb{R}^2} \mathrm{d}^2x\ x_1\theta(x_1)\ \ (N(x; a, M) - N(x; \mu, \Sigma))$$

(using the above result for conditional normals Eqn 2)

$$= \int_0^\infty \mathrm{d}x_1 \; x_1 \;\; (N(x; a_1, M_{11}) - N(x; \mu_1, \Sigma_{11}))$$

(using the result on marginals of MVNs on wikipedia pg)

$$= f(a_1, \sqrt{M_{11}}) - f(\mu_1, \sqrt{\Sigma_{11}})$$

(using the function defined above in Eqn 1).

This should be enough to calculate answers.

5.5 Suggested next steps

1. check the places where calculations could be wrong
2. understand the behaviour of the function $f(x, y)$
3. calculate a_1 and M_{11} directly
4. see if there are limiting or special cases where this last formula simplifies and can be interpreted
5. explore more sustematically numerically
6. consider including a bias term (a parameter δ for the mean in the distribution of measurement noise ε?)