Uncertainty in quantitative health impact modelling: practical exercises

Christopher Jackson, MRC Biostatistics Unit, University of Cambridge

1 What is a parameter, and why are parameters uncertain (discussion)

Suppose we want to estimate the following parameters to include in a health impact model. In an ideal world where we can measure anything, what would we measure, on what individuals? How would we summarise the individual-level measurements to estimate the parameter?

- background/baseline exposure to PM2.5
- the incidence of a disease
- the relative risk of this disease for one unit increase in exposure to PM2.5
- the mortality rate for people with a disease

Try to think of other kinds of parameters in models, and how they might be estimated in theory from summaries of individuals.

Why might the measurements we want be hard to obtain in practice? Might we measure something similar but still useful?

> Learning objectives
>
> Definition of a parameter as a summary of knowledge.
> Understanding how parameters are estimated by summarising individual observations.

2 Quantifying judgements with probability (discussion)

Suppose we have read some published data saying that there are 25.7 passengers per bus on average in a particular city. Judge 95% credible intervals for

(a) the number of passengers in a particular bus trip in this city

(b) the average number of passengers per bus in this city

(c) the average number of passengers per bus in a different city in the same country

There are no "correct answers" — you can draw on any personal experience or general knowledge you have. Think whether these intervals should be be the same or different, and record any reasoning or factors used to arrive at the judgement.

> Learning objectives
>
> Getting used to quantifying uncertainties with probabilities, based on informal judgements.

3 Obtaining full probability distributions from published estimates and uncertainties

1. Derive appropriate probability distributions for

   (a) The proportion of truck vehicle km travelled to car vehicle km travelled, given an estimate of 0.3 and a 95% credible interval of 0.15 to 0.45 (assuming car km is always greater than truck km).

   (b) The amount of time a person spends walking from home to a bus stop per day, given an estimate of 10 minutes and a standard deviation of 3 minutes.

In each case, compute the *actual* credible interval for the derived distribution, and check that it is similar to the information you supplied.

> Hint
>
> The functions `qbeta` or `qlnorm` in R can be used to compute the quantiles of the Beta or log-normal distribution, e.g. a 95% credible interval for the Beta(2, 2) is `qbeta(c(0.025, 0.975), 2, 2)`.

2. In case (1a), suppose you have used the "method of moments" technique in the slides to get a Beta distribution. This is not the only reasonable way to get an uncertainty distribution for a proportion, given an estimate and credible interval. Alternative ways are:

   - *Logit-normal distribution*: assume that the *logit* of the proportion p, that is, $log(p/(1-p))$ is normally distributed. Transform the estimate and credible limits for the proportion p to the logit scale, then convert the transformed credible interval to a standard deviation defining the normal distribution. A quantile of the distribution on the natural scale can then be obtained by using the inverse logit transform $exp(x)/(1+exp(x))$ on the logit-scale quantile.
   - *Least squares fitting*. This is a procedure implemented in the function `fitdist` from the R package `SHELF` for structured expert elicitation. For example, given an estimate of `med` (interpreted as a median) and a credible interval of `lower`,`upper`, a best-fitting Beta distribution can be derived as

     ```
     lower <- 0.15; med <- 0.3; upper <- 0.45
     SHELF::fitdist(vals = c(lower, med, upper),
                    probs = c(0.025, 0.5, 0.975),
                    lower = 0, upper = 1)$Beta
     ```

   Implement these procedures for the truck km proportion in (1a), deriving the 95% credible interval for the proportion under each of these two distributions.

3. (advanced). Suppose we have a set of 3 or more uncertain quantities which add up to 1. For example, the proportion of road transport-related PM2.5 emissions due to (a) motorcycles, (b) cars, (c) buses, (d) trucks (assuming that these together comprise all road transport-related PM2.5 emissions). How might we use Beta distributions to quantify these uncertainties?

Learning objectives

- Obtaining distributions given estimates and credible intervals (or standard errors)
- When log normal or beta distributions are appropriate
- Learning that the distributions themselves are approximations made for convenience
- Verifying that alternative approaches to getting distributions from limited information about uncertainty give qualitatively similar results
- (advanced) Expressing uncertainty around the probabilities behind a categorical variable

4 Monte Carlo simulation of health impact models

We have the following very simple health impact model (the one described in PAPER)

We want to estimate of the health impacts of a scenario where the emissions of PM2.5 air pollution from transport are D times the current amount. Consider only one population group and health outcome (incidence of stroke). The parameters include

- I_0: incidence of stroke (expected number of cases per year) in the baseline (current emissions) scenario
- μ: background concentration of PM2.5 in the baseline scenario
- π: proportion of PM2.5 due to transport…
- and the relative risk of stroke for exposure level x is described by the nonlinear function $g_2(x, \mathbf{d}) = 1 + \alpha(1 - exp(-\beta(x - \tau)^\gamma))$, with uncertain parameters $d = (\alpha, \beta, \gamma, \tau)$.

(a) Write down an expression for the background PM2.5 concentration in the reduced-emissions scenario, in terms of μ, π and D.

(b) Write down an expression for the *reduction* in stroke cases in the scenario where emissions are D times the current amount, in terms of $g_2()$ and your answer to (a).

(c) Implement this model as an R function which computes the reduction in stroke cases given the emissions reduction D and the parameters.

(d) Suppose I_0 is fixed at 18530, and $(\alpha = 13, \beta = 0.015, \gamma = 0.48, \tau = 4.2)$, and the remaining parameters are uncertain with the following distributions

- μ: log normal with mean 2.3 and SD 0.3 on the log scale
- π: Beta(5.7, 8.9)

Use Monte Carlo simulation to estimate a median and 95% credible interval for the expected number of stroke cases reduced when emissions are $D = 0.5$ times the current amount. Compare the results with $n = 1000$ and $n = 10000$ simulations. As a rough guess, how many significant figures do we have in the Monte Carlo estimate of the median?

(e) Compute the Monte Carlo standard error for the estimate of the median, compared between $n = 1000$ and $n = 10000$ samples. Does this confirm our guess?

(f) Compare the mean and median of the distribution of the health impacts with the health impact if the uncertain parameters were fixed at point estimates (say, at the means of their distributions: $\mu = 10.4$ and $\pi = 0.39$)

Learning objectives

- Implement a simple Monte Carlo analysis in R.
- Understanding Monte Carlo run length and standard error
- Understanding different definitions of the "best estimate".

5 Full ITHIM model

Learning objectives

- Experience an uncertainty analysis and Monte Carlo computation in a realistically complex health impact model.

6 Value of Information analysis

This is a demonstration of a Value of Information analysis in the simplified model from the Monte Carlo question above. There are no questions / exercises — just step through the demonstration, check everything makes sense, and feel free to ask if not.

We wish to estimate the amount of uncertainty in the model output that might be reduced with perfect knowledge of one of the model inputs. Or in other words, which of the uncertain inputs are most driving the uncertainty in the output. This could help us prioritise further research to get better information to improve the model.

In this simplified model, there are only two uncertain parameters: the background PM2.5 concentration μ and the proportion π due to transport. In the question above we defined uncertainty distributions for these parameters, drew Monte Carlo samples from the model inputs, and used these to get a Monte Carlo sample from the uncertainty distribution of the model output (expected stroke cases averted).

Suppose we have drawn 10000 samples from each of these quantities, and they are stored in the vectors `mu10000` (background PM2.5), `ptransp10000` (proportion from transport) and `sim10000` (model output). Look at the solution to the Monte Carlo question for code to obtain these.

The expected stroke cases averted is around 1000, with a SD of around 500 and 95% credible interval of around 400 to 2300.

```
mean(sim10000)
```

```
[1] 1099.114
```

```
sd(sim10000)
```

```
[1] 482.2935
```

```
quantile(sim10000, c(0.025, 0.975))
```

```
     2.5%     97.5%
 409.7157 2275.0427
```

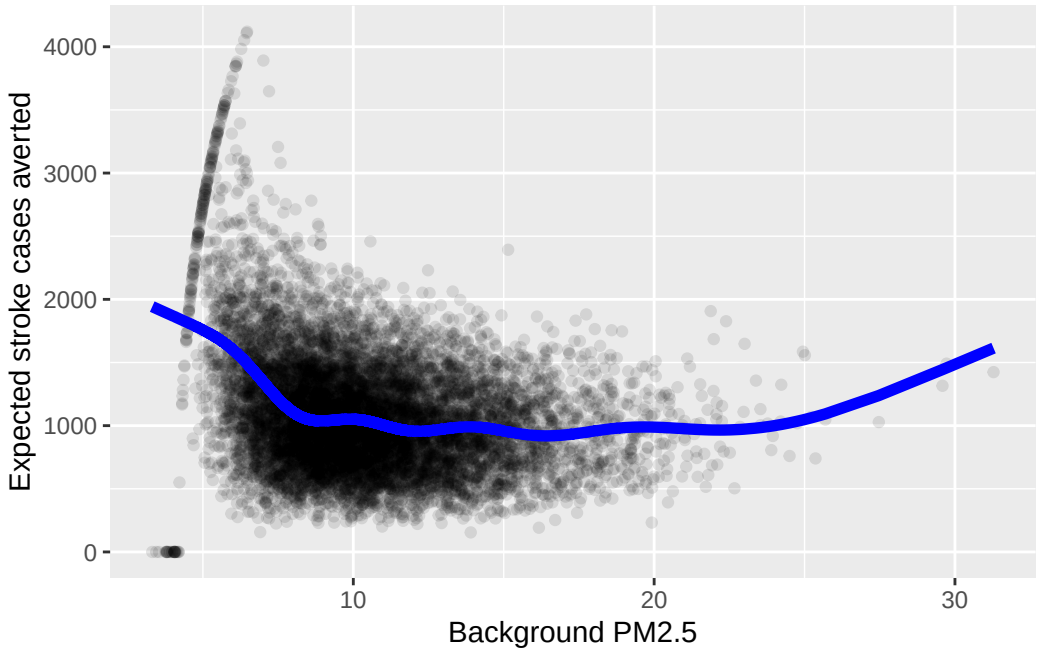
```
var(sim10000)
```

```
[1] 232607
```

Value of Information calculations for this sort of model are done in terms of the variance (the square of the standard deviation), which here is around 200,000. The expected value of partial perfect information (EVPPI) is the expected amount the variance will reduce if we learn the exact value of some parameter or parameters. As explained in the lecture slides, this is computed by fitting a flexible regression model of the simulated output in terms of the simulated inputs, and calculating the fitted values and residuals.

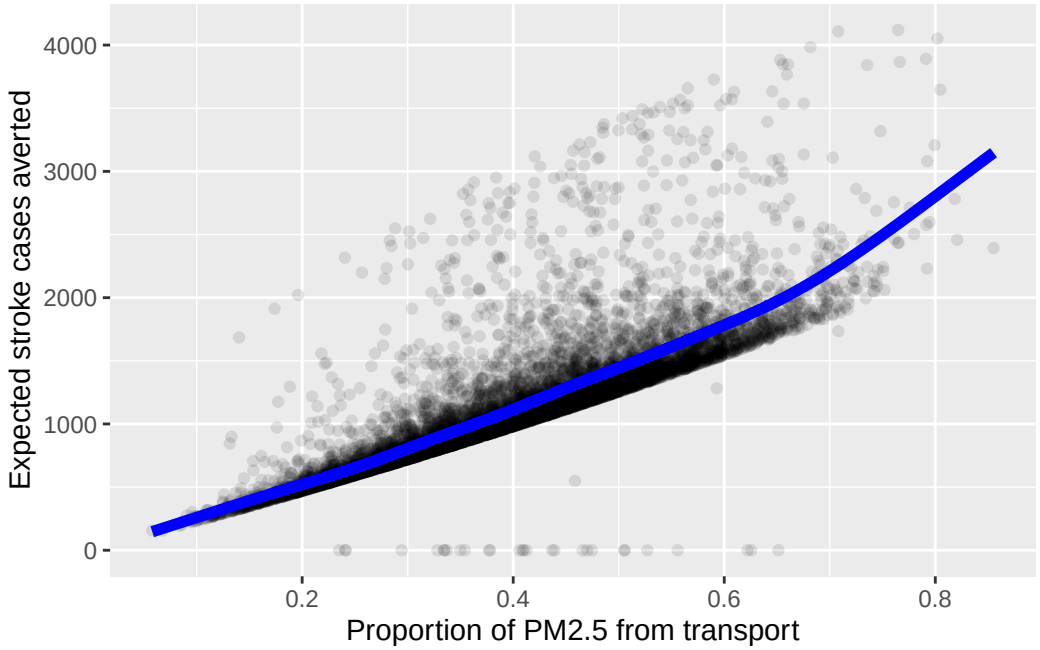
The `voi` package will take care of doing these calculations and extracting the information needed to get the EVPPI. Here we illustrate what is going on inside the package. The flexible regressions (by default) are fitted using the `gam` function from the `mgcv` package (note `s()` represents a spline function). Here are the regression models that are fitted for each of the two inputs.

```
library(voi)
library(mgcv)
library(ggplot2)
dat <- data.frame(mu=mu10000, ptransp=ptransp10000, sim=sim10000)
ggplot(dat, aes(x=mu, y=sim)) +
  geom_point(alpha=0.1) +
  geom_line(aes(y = fitted(gam(sim10000 ~ s(mu10000)))),
            col="blue", lwd=2) +
  xlab("Background PM2.5") + ylab("Expected stroke cases averted")
```



The model output appears to be weakly associated with the background PM2.5, for lower values of background PM2.5

```
ggplot(dat, aes(x=ptransp, y=sim)) +
  geom_point(alpha=0.1) +
  geom_line(aes(y = fitted(gam(sim10000 ~ s(ptransp10000)))),
            col="blue", lwd=2) +
  xlab("Proportion of PM2.5 from transport") +
  ylab("Expected stroke cases averted")
```



By contrast, the model output appears to depend strongly on the proportion of PM2.5 due to transport.

To extract the EVPPI for each of these parameters, the `evppivar` function is used. The arguments are `outputs` (vector of samples model outputs), `inputs` (data frame including sampled model inputs), and `pars` (which columns of this data frame we want separate EVPPIs for)

```
evppivar(outputs=sim10000, inputs=dat, pars = list("mu","ptransp"))
```

```
     pars     evppi
1      mu  37071.05
2 ptransp 157732.64
```

As expected, the expected value of learning the proportion due to transport is higher. These might be presented as *proportions of variance explained*:

```
evppis <- evppivar(outputs=sim10000, inputs=dat,
                   pars = list("mu","ptransp"))$evppi
evppis / var(sim10000)
```

```
[1] 0.1593720 0.6781078
```

Or perhaps more usefully, we might predict what the remaining SD or credible interval of the output might be after learning perfect information about one of these inputs. The remaining SD is the square root of the variance remaining (original variance minus EVPPI).

```
sd_remaining <- sqrt(var(sim10000) - evppis)
sd_remaining
```

```
[1] 442.1945 273.6318
```

The SD is predicted to be of the order of 400 or 200 if learning μ and π respectively. We might convert these to a rough "predicted credible interval" for the expected stroke cases averted, by adding $\pm$ 2 SDs to the estimate of 1000. Hence if we learnt π, the CI would reduce in width, to (600,1400) from its original (400,2300).

6.1 Joint EVPPI for multiple related parameters

Note that in this model, these parameters *do not act independently* on the output. The reductions in uncertainty from learning *both parameters together* cannot be simply calculated by combining the separate EVPPIs for each parameter.

To calculate the *joint* EVPPI for more than one parameter, representing the value of learning the parameters in combination, a flexible regression model is fitted that includes all those parameters as predictors, instead of just one predictor (as we had before). To do this with the `evppivar` function, supply a *vector* of parameter names to the `pars` argument, rather than a list.

Here this shows there is expected to be a greater benefit from learning both the parameters jointly, compared to learning them separately.

```
evppivar(outputs=sim10000, inputs=dat, pars = c("mu","ptransp"))
```

```
        pars    evppi
1 mu,ptransp 201270.3
```

(In theory, since there are only two parameters in the model, the EVPPI for learning both of them should be equal to the original variance of the model output - since we reduce this variance to zero if we have perfect information on both. However it is slightly different in practice, due to the approximation made by the regression model)

6.2 Further resources

- The paper "A guide to value of information methods for prioritising research in health impact modelling" (Jackson et al, Epidemiologic Methods, 2021) (https://doi.org/10.1515/em-2021-0012).
  This goes into more detail on this particular example model.
- The web site for the `voi` package

  https://chjackson.github.io/voi

  gives resources and worked examples to learn about using VoI methods in models for making decisions between alternative policies, as well as models for estimating quantities (e.g. health impacts).

7 Technical notes about probability distributions

Many of the methods for uncertainty quantification we have discussed invoke the idea of a "best estimate" or "central point" of a distribution. There is at least three definitions of this:

Mean: average value. If $X_1, X_2, ..$ all come from this distribution, then $(X_1 + ... + X_n)/n$ (the "sample mean") approaches the mean of the distribution as n gets bigger.

Median: value for which 50% of the distribution is less than this value

Mode: commonest value, or value with highest probability density

These all agree for symmetric distributions, e.g. the normal. If you are trying to obtain an uncertainty distribution, and you have arbitrarily used one of these three as the "best estimate", but they disagree, consider whether the amount of disagreement matters. Might any results of interest be affected? If so, more information is probably needed to describe the evidence confidently.

Standard deviation: a measure of variability in a distribution

Standard error: a measure of the uncertainty in the estimate of a mean, due to having only sampled n observations $X_1, ..., X_n$. Defined as the standard deviation of $(X_1 + ...X_n)/n$. If the X_i are independent with standard deviation σ, then the standard error is $\sigma/\sqrt{n}$.