

Uncertainty in quantitative health impact modelling: answers to practical exercises

Christopher Jackson MRC Biostatistics Unit, University of Cambridge

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

- A person's exposure – perhaps averaged over time and space. But this will vary according to your model. Do you want to assign exposures for a synthetic person with a particular characteristics? If so - you would want to observe many individuals with those characteristics and take an average. Or sample a value if want to include noise? Will it make a difference if average out? Is this getting too advanced.
- Does a person get the disease within one year? Proportion of people getting the disease within one year
- Proportion of people with exposure (baseline value + 1 unit) who get the disease with 1 year, divided by proportion of people with exposure (baseline value) who get the disease with 1 year. May depend on the baseline value.
- Variations between time and places, expense of carrying out study, differences between populations... and many more, depending on the example you chose.

2 Quantifying judgements with probability

- (a) To get an upper credible limit you might draw on knowledge of how many seats are in a bus, how often buses are “standing room only”, and any knowledge of this city or the country it is in. For the lower limit, how often are buses empty? Perhaps this is rare in a city. How frequent are buses?
- (b) Key issues:
- The credible interval should not have zero width. You may wonder if there should be any uncertainty about the average, since we have data stating the “average” is 25.7. If this had been obtained by accurately observing every bus trip in a particular time period, then, yes, 25.7 would be a perfect estimate of the average bus occupancy over that time period. But we want a parameter in a health impact model used to inform policy about a different set of trips in the future, so there should be some uncertainty.
 - The interval for the *average over all trips* should be narrower than the interval for a *specific* trip. The difference between the two intervals might be roughly quantified by conceiving the estimate 25.7 as an average of the observed occupancy over n trips: say $\bar{X} = (X_1 + \dots + X_n)/n$. If the occupancy in the i th trip X_i is normally distributed with standard deviation σ , then the standard deviation of the average, \bar{X} , (also called the “standard error”) is σ/\sqrt{n} .
 - Each of the intervals is then $25.7 \pm 2 \times \text{SD}$, and because the average (\bar{X}) has a smaller SD than an individual data point X_i , the interval is narrower.
 - We might also want to inflate σ a certain amount to account for any additional biases (selection bias, measurement error) underlying the estimate of 25.7. Equivalently, we could inflate the width of each credible interval by a certain percentage.
- (c) The interval obtained in (b) should be inflated to account for any difference between the two different cities.

If making judgements, the important thing is to state the judgements transparently, and any reasoning underlying them. This lets your readers scrutinise them. For example, as a credible interval for an understandable quantity, or, e.g. “10% wider interval width due to unknown biases”

3 Obtaining full probability distributions from published estimates and uncertainty

1.

(a) Proportion of trucks to cars

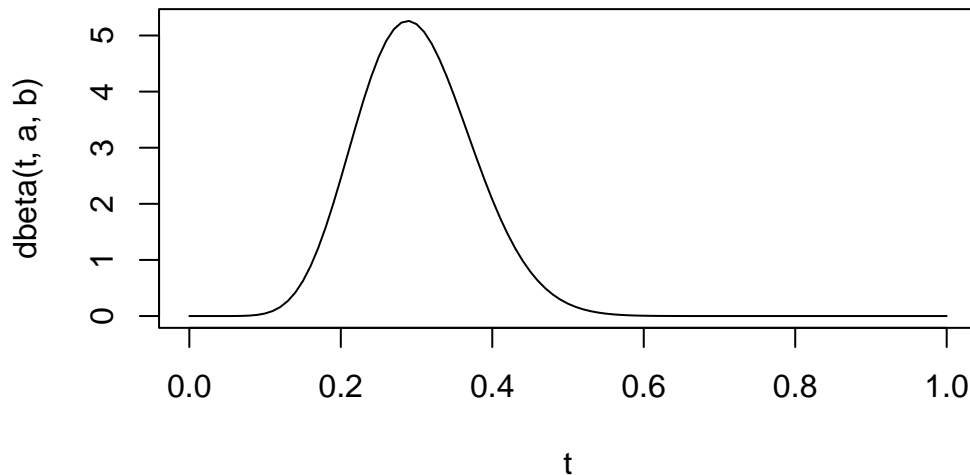
```
m <- 0.3; lower <- 0.15; upper <- 0.45  
s <- (lower - upper)/4  
a <- (m*(1-m)/s^2 - 1)*m  
b <- (m*(1-m)/s^2 - 1)*(1 - m)
```

The median and credible interval of the derived Beta are close to the assumed mean and credible interval:

```
qbeta(c(0.025, 0.5, 0.975), a, b)
```

```
[1] 0.1645796 0.2962953 0.4562605
```

```
t <- seq(0, 1, by=0.01)  
plot(t, dbeta(t, a, b), type="l")
```



(b) Time spent walking

```
m <- 10; s <- 3  
mu <- log(m/sqrt(s^2/m^2 + 1))  
sigma <- sqrt(log(s^2/m^2 + 1))
```

A credible interval on the natural scale can be produced as

```
qlnorm(c(0.025, 0.5, 0.975), mu, sigma)
```

```
[1] 5.387753 9.578263 17.028085
```

which is roughly consistent with a standard deviation of 3 (the heuristic that the 95%CI equals the mean ± 2 standard deviations will not work exactly with skewed non-normal distributions like this one).

If doing this in practice, you should check that the credible intervals match your judgement, and revise accordingly if not.

2. (Alternative ways to elicit a proportion)

Logit-normal distribution

```
logit_mean <- log(0.3 / (1 - 0.3)) # qlogis(0.3) also does this in R
logit_lower <- log(0.15 / (1 - 0.15))
logit_upper <- log(0.45 / (1 - 0.45))
sd_logit <- (logit_upper - logit_lower) / 4
lower_fitted_logit <- qnorm(0.025, logit_mean, sd_logit)
upper_fitted_logit <- qnorm(0.975, logit_mean, sd_logit)
lower_fitted_natural <- exp(lower_fitted_logit) / (1 + exp(lower_fitted_logit))
upper_fitted_natural <- exp(upper_fitted_logit) / (1 + exp(upper_fitted_logit))
```

The fitted credible limits from the logit-normal distribution roughly match the limits of (0.15, 0.45) that we originally judged.

```
lower_fitted_natural
```

```
[1] 0.168134
```

```
upper_fitted_natural
```

```
[1] 0.4760968
```

Least squares fitting. We use `fitdist` from the `SHELF` package to find the Beta distribution which “best fits” an assumed median of 0.3 and 95% credible interval of 0.15 to 0.45.

```
shelf_beta <- SHELF::fitdist(vals = c(0.15, 0.3, 0.45),
                             probs = c(0.025, 0.5, 0.975),
                             lower=0, upper=1)$Beta
a <- shelf_beta$shape1
b <- shelf_beta$shape2
```

The credible interval of the fitted Beta agrees with the original (0.15, 0.45), within reasonable range. Note a “perfect fit” to any given quantiles is not generally possible. Simple two-parameter distributions such as the Beta and Normal can only express a limited range of probability judgements, but it is generally good enough in practice.

```
qbeta(c(0.025, 0.975), a, b)
```

```
[1] 0.1676177 0.4590843
```

3. (advanced) Suppose we want to obtain $p_{motorcycle}, p_{car}, p_{bus}, p_{truck}$, the proportions of all emissions due to each source, with $p_{motorcycle} + p_{car} + p_{bus} + p_{truck} = 1$.
- (a) Define a Beta distribution for $p_{motorcycle}$ the proportion of all emissions due to motorcycles, based on a guess with a credible interval.
 - (b) Define a Beta distribution for q_{car} , the proportion of car, bus and truck emissions due to cars, likewise. Then the proportion of all emissions due to cars is $p_{car} = (1 - p_{motorcycle}) * q_{car}$
 - (c) Define a Beta distribution for q_{bus} , the proportion of bus and truck emissions due to buses. Then the proportion of all emissions due to buses is $p_{bus} = (1 - p_{motorcycle} - p_{car}) * q_{bus}$
 - (d) p_{truck} is defined as $1 - p_{motorcycle} - p_{car} - p_{bus}$.

An alternative approach would be to use the Dirichlet distribution, which generalises the Beta distribution to deal with sets of 3 or more proportions that add up to 1. However it is tricky to define a Dirichlet in terms of means and credible intervals for the uncertain quantities - expressing as a sequence of “conditional” Betas as above is easier.

4 Monte Carlo simulation

(a) $g_1(\mu, \pi, D) = \mu(\pi D + 1 - \pi)$

(b) $Y = f(\theta) = I_0 - I_0 g_2(g_1(\mu, \pi, D), \mathbf{d}) / g_2(\mu, \mathbf{d})$

(c)

```
dose_response <- function(pm,alpha,beta,gamma,tau){
  1 + alpha * ( 1 - exp(- beta * ( pmax(pm - tau, 0) )^gamma ) )
}

model <- function(D, mu, ptransp, dr) {
  scenario_pm <- mu*(ptransp*D + 1 - ptransp)
  RR_base <- dose_response(mu, dr$alpha, dr$beta, dr$gamma, dr$tau)
  RR_scenario <- dose_response(scenario_pm, dr$alpha, dr$beta,
                              dr$gamma, dr$tau)
  inc_diff <- inc_base - inc_base*RR_scenario/RR_base
  inc_diff
}
```

This is just one of many, many ways to write the code for a model like this.

This code works in a “vectorised” way - given a vector of different parameters, we can obtain the model output for each parameter as a vector. Not all models will be able to be vectorised in this way, but if this is possible, it often makes R code cleaner and faster.

(d)

```
inc_base <- 18530
mu1000 <- rlnorm(1000, 2.3, 0.3)
mu10000 <- rlnorm(10000, 2.3, 0.3)
ptransp1000 <- rbeta(1000, 5.7, 8.9)
ptransp10000 <- rbeta(10000, 5.7, 8.9)
dr <- list(alpha=13, beta=0.015, gamma=0.48, tau=4.2)
sim1000 <- model(D=0.5, mu=mu1000, ptransp=ptransp1000, dr=dr)
sim10000 <- model(D=0.5, mu=mu10000, ptransp=ptransp10000, dr=dr)
quantile(sim1000, c(0.025, 0.5, 0.975))
```

2.5%	50%	97.5%
432.2303	1029.6825	2182.3901

```
quantile(sim10000, c(0.025, 0.5, 0.975))
```

2.5%	50%	97.5%
413.4384	1026.2016	2233.5287

It looks like we can safely report the median of the uncertainty distribution (a reasonable point estimate) as 1000 (to the nearest multiple of 100), but the third significant figure (e.g. is it 1020 or 1030?) depends on which random samples are drawn.

- (e) Monte Carlo standard error

```
sd(sim1000) / sqrt(1000)
```

```
[1] 14.51028
```

```
sd(sim10000) / sqrt(10000)
```

```
[1] 4.735773
```

This represents uncertainty in the median due to having drawn a limited number of Monte Carlo samples. It can be reduced by drawing more samples. The Monte Carlo standard error will reduce to 0 as the n increases to ∞ .

The Monte Carlo standard error (MCSE) can be used as a way to guide how many significant figures to present in the estimates. Our Monte Carlo estimate might be judged accurate to within ± 2 MCSE.

The Monte Carlo error is a different concept from *uncertainty in knowledge* due to having imperfect knowledge of the model parameters (μ and π in this case) — which can only be reduced with more/better observed data on these parameters. This kind of uncertainty is reflected by the credible interval (estimated by the 2.5% and 97.5% quantiles of the Monte Carlo sample).

(advanced)

Note the upper and lower credible limits also each have Monte Carlo error — so we can also use a similar technique to judge the appropriate number of significant figures to present in these. Tools for computing MCSEs for *sample quantiles* (as opposed to the *sample mean*) are available in the R **posterior** package.

- (f) The model evaluated at the means of the parameters is not the same as the mean (or median, computed above) of the uncertainty distribution of the model output. The difference may or may not be practically important, but if we are able to quantify uncertainty and do the Monte Carlo computation, then the estimate that accounts for uncertainty should be preferred.

```
model(D=0.5, mu=10.4, ptransp=0.39, dr=dr)
```

```
[1] 995.7316
```

```
mean(sim10000)
```

```
[1] 1094.24
```