**John’s Sample Size Calculator**

Aim

To develop a sample size power calculator using preliminary data.

Language

R

Context

A group of workers’ exposures to a toxic chemical are sampled. There is a legislated exposure limit, called ‘OEL’. If the sampled data’s 70% upper credible level of the 95th percentile is less than the OEL, then the workers are ok. If this value is larger than the OEL, action is needed.

The population and data are assumed to be lognormal.

Small (3-10) sample sizes are often collected – due to cost and lack of statistical understanding.

Process

The process is based on “Doing Bayesian Data Analysis” by J. Kruschke. Refer to diagram below.

**Step 1: Create Hypotheses**

* Preliminary data, OEL, and calculator variables are entered
* A JAGS MCMC program is used to calculate the posterior distribution of the data set and of the lognormal distribution parameter values (µ and σ)
  + This program was developed by experts and will not need to be reviewed
* The goal is decided – either:
  + 70% UCL of P95 > OEL, or
  + 70% UCL of P95 < OEL

**Step 2: Generate simulated data sets**

* 1000 sets of parameters are randomly sampled from the µ and σ distributions in step 1
  + 1000 is arbitrary
* 6 ‘simulated’ samples are generated using each of the 1000 pairs of µ and σ values

**Step 3: Analyse simulated data sets**

* The 70% UCL of P95 for each data set is estimated
* Power =
* If the power isn’t high enough (0.8 – arbitrary), then repeat step 2 & 3 with +1 simulated samples per data set…

Issues / Questions

* Is the process meaningful / valid?
* The70% UCL of P95 of simulated data sets varies wildly set to set – results vary from 20 to 2000!
  + Why might this be?
  + Can I / Should I try to reduce this variability? (such as truncate the µ and σ distributions sampled from)
* The power / sample size calculations vary wildly each time the program is run. 1 run the calculator will say the power of 6 samples = 0.75, the next run it may say = 0.34. Sometimes it will say power is reached at 15 samples… then next time 85!
  + Why might this be?
  + Can I / Should I try to reduce this variability to make it (mostly) consistent each run?
* Is the upper tolerance level UTL70,90 the same as 70% UCL of P95?
* Low priority: is there obvious ways to make the program more efficient – run time.

Note

* Step 3 is not done using Bayes’ rule. I tried the JAGS program from step 1 but its too slow to run 1000s of times
* This is my first programming project. Apologies in advance if it is not pretty to read.

Diagram

Description automatically generated