Multi-town sampling

28 March 2014

Ben Bolker

We have been discussing the fact that real surveys are done by sampling multiple clusters within a population. There is some vagueness in terminology here; we previously generated and sampled what we called a "population", which was a single spatially contiguous (but not necessarily homogeneous) unit. Now we want to generate and sample multiple non-contiguous places, which we will call *towns* (although the more standard survey terminology would be to call the generated units, or the samples taken from them, "clusters"); a collection of towns will be called a *population*. (Our previous simulations now represent populations that are made up of a single town.)

As has been previously defined in Roman's document, the most important difference among towns is population size; this will follow a Pareto (power-law) probability distribution, with specified lower and upper bounds and a specified exponent.

All of the other existing parameters of the models should be allowed to vary among towns. I don't think it's very important to be able to specify the parameters of individual towns explicitly – in almost all cases I can think of it's OK to draw them from a distribution. I would suggest the following rules/updates to existing specifications:

- for convenience it would be useful to change the specification of *all* town-level parameters so that they can sensibly be considered to be Normally distributed (i.e., they can take on any real/floating-point value). I haven't checked, but I think most are already OK. The only one that definitely needs to change is the average family size, which should be changed to log(family size). (In general, if a parameter needs to be positive we can log-scale it and exponentiate to get back to the original value; if it needs to be in [0,1] we can logit ($log(x/(1-x))$) scale it and logistic transform ($1/(1+\exp(-x))$) to get back to the original value.
- the default behaviour is for parameters *not* to vary across towns; any parameter that's not specified in the 'population configuration file' should just default to the value specified in the 'town configuration file'.

- we'd like to allow the parameters to vary across towns; we'd also like to allow them to be correlated with population size. We have a few decisions to make:
  - should we specify a linear regression coefficient and a residual variance, or a variance and a correlation coefficient and a variance? These are equivalent, but one way or the other might be more convenient.
  - I think it makes most sense to specify correlation/regression parameters with respect to the *log* of population size (probably easiest to understand if we use log base 10)

If we use the linear regression coefficient approach, we would say that each line of the population configuration file should specify `param_name regr_coeff resid_sd`; for example, `mean_income_b00 0 0.2` would specify that mean income was independent of population size, but varied with a standard deviation of 0.2 (we *could* specify coefficient of variation instead, but that's probably being too clever). `mean_income_b00 0.1 0` would specify that mean income increased by 0.1 per 10-fold increase in population size, but didn't otherwise vary at all (i.e. perfect correlation/$r^2 = 1$). `mean_income_b00 0.1 0.2` would combine the previous two specifications.

The only important detail I've left out here is that we should specify the regression equation so that the *mean* value is unchanged. That is, the value from the "town configuration file" should be the expected value at the (geometric) *mean* population size, not at a population size of 1 (i.e. log10(pop)=0), which would be what we'd get if we naively used the default town as the intercept.

In other words,

$$P(S) = \text{Normal}(P_0 + S(\log(N) - \overline{\log(N)}), \sigma^2)$$

where P_0 is the default value, S is the slope, N is the population size, and σ is the standard deviation.

One final question is whether $\overline{\log(N)}$ should be the mean of the *realized* log population sizes for a particular population, or whether it should be the (population) mean of the Pareto distribution from which they were chosen. I think I prefer the latter, but it's a matter of taste.