Design-based vs Model-based framework

2020-07-20

> The postulate of randomness thus resolves itself into the question, "Of what population is this a random sample?" which must frequently be asked by every practical statistician. (Fisher 1922)

When we say our sample set is a sample from a population, what population are we talking about?

In Fisher's model-based framework, you begin with specifying a statistical model which specifies a data generating process for your y_i's. Now imagine we generate and store y_i's ad infinitum. The resulting data set forms our hypothetical infinite population and our sample is considered a simple random sample (SRS) from this infinite population. If your sample is not a SRS, then your statistical model specifying the infinite population is incorrect. Note, the model-based framework does not require an actual random sample; it only requires that the distributional assumptions of the infinite population (and thus the statistical model) match that of the sample.

In contrast, Neyman's design-based framework begins with specifying a finite population and not a statistical model (and thus no hypothetical infinite population). In the design-based framework the targets of inference is not the statistical model, but the parameters of the fixed, finite population e.g. the average age of all New Zealand residents. Then the sample $\{y_i\}$ is not assumed to be intrinsically random like in the model-based framework; instead the randomness of the $\{y_i\}$'s are inherited from the randomness of how each individual is sampled. Note, in the design-based framework requires a random sample but not a SIMPLE random sample; that's required is knowing each π_i.

Adjusting for complex sampling schemes

The practicalities of collecting data often requires complex sampling schemes: stratified sampling, cluster sampling, multi-stage sampling etc. The design-based framework allows for conceptually straightforward way to take this into account by calculating π_i for each individual according to the sampling scheme. However a pure design-based framework only allows for estimation of quantities like means, totals and ratios. There are very good scientific reasons for wanting to use a statistical model.

One way of getting around this limitation is by using pure model-based framework and the conditioning on all sampling variables (e.g. stratification, clusters, etc.) and allowing for interactions. However adjusting the model under the conditionality princple both complicates the model interpretation and may require many additional degrees of freedom.

Model-based vs Design Based

Model-based

1. Define a statistical model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

2. Make the y_i's random by imposing a distributional assumption

$$\epsilon_i \sim N(0, \sigma^2)$$

3. "Conditionality Principle"

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 I_{strata1} + \beta_3 I_{strata2} \cdots + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2)$$

Design-based

1. Define the sampling frame, sample design, sampling scheme

- sample frame: list of primary sampling units $\{1, \ldots, N\}$
- sample design: probability (π_i) of selection assigned to each unit i
- sampling scheme: a draw-by-draw mechanism to sample units according to the probabilities assigned to each unit

2. Construct estimators for the finite population parameters using inverse probability weighting

$$\widehat{y_{total}} = \sum_{j \in S} \frac{1}{\pi_i} y_j$$

where $y_{total} = \sum^N y_i$