

Bayesian alternatives to P -values and their use in Selecting and Ranking Populations

A broad area of statistical research concerns the selection of a reduced subset of $k < N$ populations or experiments, based on data X_1, \dots, X_N from N populations. For example, a common goal of *DNA*-microarray analysis is to identify a subset of genes that are differentially expressed under 2 different cell conditions. In such settings, the populations are often ranked and selected according to the p -values for the tests: $H_0 : \theta_i = 0$ vs $H_a : \theta_i > 0$, $i \leq N$ (where θ_i is the parameter of interest relating to population i). As a Bayesian alternative, one could use the posterior probability $\mathbb{P}\{\theta_i \geq c \mid X_1, \dots, X_N\}$, for a suitably chosen threshold c , to rank the different populations. Ideally, this threshold, c , should be chosen based on scientific knowledge about important values for the parameter. However, such knowledge is not always available. In this talk, we show how to use the data, X_1, \dots, X_N to construct a Loss function, \hat{L} . Using this empirical loss, \hat{L} , in selecting populations, is appropriate when we would like to rank populations based on the probabilities $\mathbb{P}\{\theta_i \geq c \mid X_1, \dots, X_N\}$ for some threshold c , but are uncertain about the correct value of c to use.

A related issue, especially when N is large, is the computational feasibility of a Bayesian analysis. For instance, consider the problem of selecting a subset of k populations, so that the proportion, O , of the selected populations that correspond to one of the k largest parameter values is as large as possible. As an alternative to a full Bayesian analysis (that selects the subset of populations maximizing $\mathbb{E}(O \mid X_1, \dots, X_N)$), we propose a computationally efficient method of selecting populations that is, in a sense, asymptotically optimal. An interesting potential use of this

methodology lies in selecting important single nucleotide polymorphisms (SNPs) in SNP microarray settings, where the number of probes is very large.

Examples are given that contrast the populations selected by both the methods described above with those selected by p -values.