A One Pager on Bayesian Statistics in Baseball

Course: Bayesian Statistics
Johns Hopkins University

Shalin Shah
sshah100@jhu.edu

This short article is a summary of [1] in an expository tutorial format.

An example in this book chapter models the number of batting hits of baseball players as a Bayesian hierarchical model. The model has several layers in the hierarchy. The number of hits is modeled as a binomial variable at the bottom of the hierarchy (level 1). The probability of a hit (the parameter of a binomial) is then modeled as a beta distribution (level 2). Each player has his own binomial distribution. The beta prior is for each of the primary field positions, like pitcher, left fielder, catcher etc. There are 9 primary positions. So, there are 9 beta distributions with their own 18 parameters (9 means and 9 variances). The reason for clustering the players into primary positions is the interest in doing inference on group level means of batting efficiencies. Each of these 9 means comes from another beta distribution (level 3) and each of the 9 variances comes from a gamma distribution (level 3). These level 3 distributions represent the central tendencies and variances of all baseball players. Finally, the level 3 beta and gamma distributions have 4 parameters which have their own top-level priors (level 4). These top-level priors can be informative and can represent subjective information, like seasonal patterns, expert knowledge etc. In total there are 970 parameters in the model.

Inference on individual players and on group means is done. The key observation in the paper is that individual batting averages are highly influenced by group means. For instance, if there is not enough data for a player (less opportunities to bat), the batting efficiency of that player is highly influenced by the primary position group of the player. While doing inference, the paper mentions several interesting statistics that can be calculated from the posterior like differences in batting averages of individuals, differences of batting efficiencies between different primary positions (e.g. a catcher has a much higher batting hits as compared to a pitcher).

Also, each statistic drawn from the posterior has a distribution and a highest density interval (HDI) can be drawn across each statistic. Also, if this is done for a difference in batting hits between say a pitcher and a catcher, and the HDI around the difference does not include 0, this can be interpreted as a statistical test of difference (the null hypothesis that there is no difference is rejected).

The paper gives an example of a player for which the batting hit probability calculated by just using the data for that player is much higher than the Bayesian estimate. This is because there is not enough data for the player and so, the estimate for that player borrows information from the level up (group level means). The paper calls this “shrinkage” and notes that this is a very useful characteristic of a Bayesian hierarchical model – each player’s estimates are influenced by the group level estimates which are influenced by the global priors. This is possible as all parameters are estimated simultaneously.

References

- [1] John K Kruschke and Wolf Vanpaemel. Bayesian estimation in hierarchical models. *The Oxford handbook of computational and mathematical psychology*, pages 279–299, 2015.