Recently, in a document classification case a Court stated "[t]he point of random sampling is to eliminate irrelevant documents from the group identified by a computerized search and focus the parties' search on relevant documents only." In re Lithium Ion Batteries Antitrust Litigation, Dkt. 13-MD-02420, Slip Op. at 2 (DMR)(USDC Cal. Filed 2/24/15). This comment made me think about purpose of randomization while modeling for document classification; thus the purpose of this post is to consider why we use randomization and its apparent application to document classification.

For purposes of analysis, we will construct a sample response population consisting of only a response variable for document population of 1,500,000. A response of "1" would indicate that the document is responsive to a request, while "0" would indicate that the document is non-responsive or irrelevant. To simulate receipt from varying record custodians, we will assume five custodians, with submissions of varying density, i.e. the portion of the custodian's share that is responsive to a given request or prompt.