2.2 Simulated Data

We replicate the simulation design used by Strobl et al. (2007a) where a binary response variable Y is predicted from a set of 5 predictor variables that vary in their scale of measurement and number of categories. The first predictor variable X_1 is continuous, while the other predictor variables $X_2, \ldots, X_5$ are multinomial with $2, 4, 10, 20$ categories, respectively. The sample size for all simulation studies was set to n = 120. In the first *null case* all predictor variables and the response are sampled independently. We would hope that a reasonable variable importance measure would not prefer any one predictor variable over any other. In the second simulation study, the so-called *power case*, the distribution of the response is a binomial process with probabilities that depend on the value of x_2, namely $P(y = 1|X_2 = 1) = 0.35, P(y = 1|X_2 = 2) = 0.65$.