1. *Establishing a "baseline" error rate.* Suppose that we have a classification problem with K classes, and suppose that the proportions of observations in each class are $p_1, p_2, \ldots, p_K$. Suppose that class Q has the largest proportion, so that $p_Q > p_m$ for all other m 6= Q.

If you had no explanatory variables and still had to do prediction, you would use a naive classifier that always assigns most common class to all predictions. In our problem, **what would be the misclassification rate for the naive classifier?**

This is sometimes called the baseline error rate for the problem, and represents a guess at the worst error rate you expect and "real" classifier to have, assuming that future samples have the same distribution of classes as this one.

- ➔ If I always predict the predicted class would be Q, which has the largest proportion, than the error rate = number of variables in class which is not Q / number of all variables.

2. *Difficulties with classifying unbalanced responses*. Suppose you have a classification problem with K = 2, and that 95% of the responses are class 1. **What is the baseline error rate for this problem?**

It is often the case that the baseline error rate is hard to beat with a "real" classifier, because correctly classifying a portion of the class-2 data often causes an even larger number of class-1 data to be misclassified. For example, if the ratio of class 1 to class 2 is 95:5, then correctly classifying even one or two class-2 observations may cause 5 or 10 class-1 responses to be misclassified. For this reason, we may choose to use other measures besides total misclassifications to judge a classifier. We will talk about these more later.

- ➔ Number of variables in class which is not 2 is 5%. So if I always predict the variables' class would be K=2, 5% of them would be incorrect. Therefore, the baseline error rate is 5%.