beta mapper

Mohsen Sadatsafavi

2023-12-28

## Estimating the distribution of NBs from summary statistics of a model performance

Our task is to generate many samples from the joint probability distribution P(prev, se, sp) from distributions on the performance of the model in a population.

We assume we have specified probability distribution for the following parameters:

* Expected prevalence of the outcome in the target population
* c-statistic of the model in the target population
* Calibration slope and intercept for the model in the target population
  + Calibration slope and intercept refer to A and B in the following equation
  + (the calibration function)

## Lemma:

Let the predicted risk and the corresponding binary outcome. If then and .

Proof: trivial by using Bayes’ theorem and the Beta-Bernoulli cojugacy.

We will use the above finding twice: one to simplify the estimation of c-stat given to the probability of one Beta RV being greater than another Beta RV, and again when estimating sensitivity and specificity of a (potentially miscalibrated) model if the calibrated risk follows a Beta distribution.

## Approach

1. We assume the distribution of CALIBRATED RISKS (ie., follows a Beta distribution in the population. This is much more tractable than making assumptions about the distribution of . We note that because the logit calibration connecting and is monotonic on , our knowledge of the c-statistic for predicted risks carries to the same distribution for .
2. Because we assume we know the mean and c-statistic, we should be able to derive the parameters of this Beta. Given the mean is known, this becomes a matter of finding the alpha parameters of the Beta distribution that gives rise to a given c-statistics.

Given the above lemma, we will be dealing with Beta distribution all along. TODO: show that a=f(c-stat) is monotonical for a given mu?

1. We estimate the sensitivity and specificity of the model at a given threshold given the above-mentioned Beta, as well as calibration intercept and slope

,

.

Again, given the lemma, these probabilities follow a Beta distribution, so this is easy to calculate. For example, (assuming B>0 which is safe).

And we do know given the Lemma.

END OF DOCUMENT