Number of samples required for validation

Chinmay Belthangady

June 2024

1 Introduction

In the Mosaic study, of the 194 cancers in the intent-to-test population (defined as smokers above age 50) of the validation data set, 167 were called correctly. The sensitivity is therefore $167/194 = 86\%$. To estimate the number of samples required for validation for different levels of success probability, we work in the Bayesian framework where the method of conjugate priors allows us to compute closed-form solutions for probabilities.

2 Computing posteriors

The first step in the Bayesian inference framework is to pick a prior that characterizes our belief in what the sensitivity of our classifier is before we have observed any data. Since this sensitivity σ is a number between 0 and 1, we model it as a beta distribution.

$$p(\sigma) = \frac{1}{B(\alpha, \beta)} \sigma^{\alpha-1} (1-\sigma)^{\beta-1} \tag{1}$$

where $B(\alpha, \beta)$ is the beta function. The likelihood function i.e the probability of calling k cancers correctly from a set of n cancers with a sensitivity σ is given by the binomial distribution.

$$p(k, n|\sigma) = \binom{n}{k} \sigma^k (1-\sigma)^{n-k} \tag{2}$$

Using the Bayesian formula, the posterior can be computed.

$$p(\sigma|k, n) = \frac{p(k, n|\sigma) \times p(\sigma)}{p(k, n)} \tag{3}$$

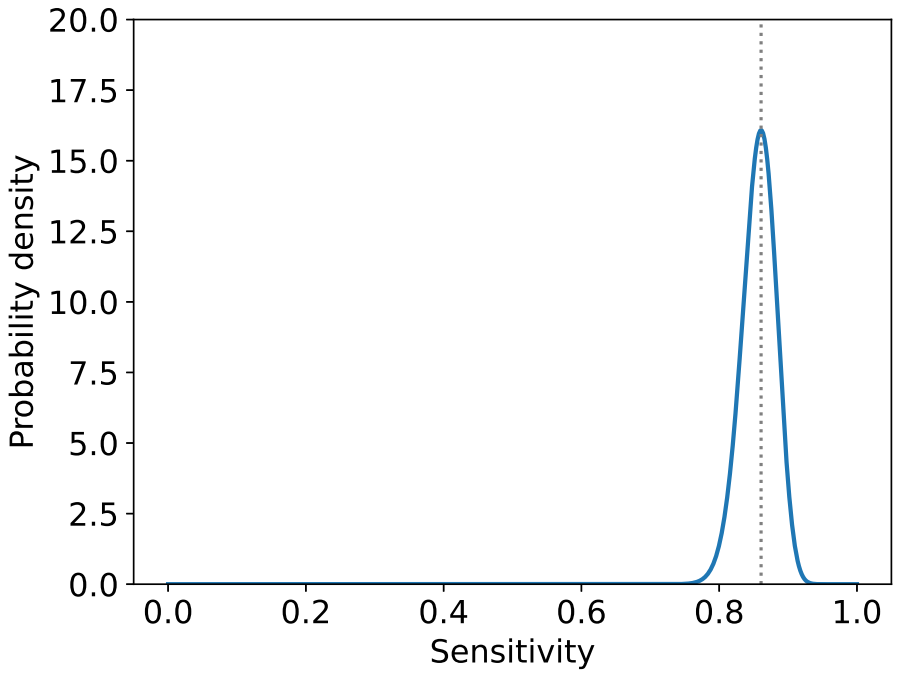
Since we have a beta-distributed prior and a binomial likelihood, the method of conjugate priors allows us to compute the posterior in a straightforward manner.

$$p(\sigma|k, n) = \frac{1}{B(\alpha + k, \beta + n - k)} \sigma^{\alpha+k-1} (1-\sigma)^{\beta+n-k-1} \tag{4}$$

To make the prior completely uninformative, we pick a uniform prior by setting $\alpha = 1, \beta = 1$. The posterior then simplifies to:

$$p(\sigma|k,n) = \frac{1}{B(k+1, n-k+1)}\sigma^k(1-\sigma)^{n-k} \tag{5}$$

The plot below shows this posterior distribution, where the gray dotted line represents the point estimate of 0.86. The 95% *credible* interval is $0.80 - 0.90$. This matches closely the Clopper-Pearson 95% *confidence* interval of $0.80 - 0.91$ that can be estimated in a Frequentist analysis. The posterior reflects our belief in the sensitivity of the classifier after observing the data.



3 Computing posterior predictive distributions

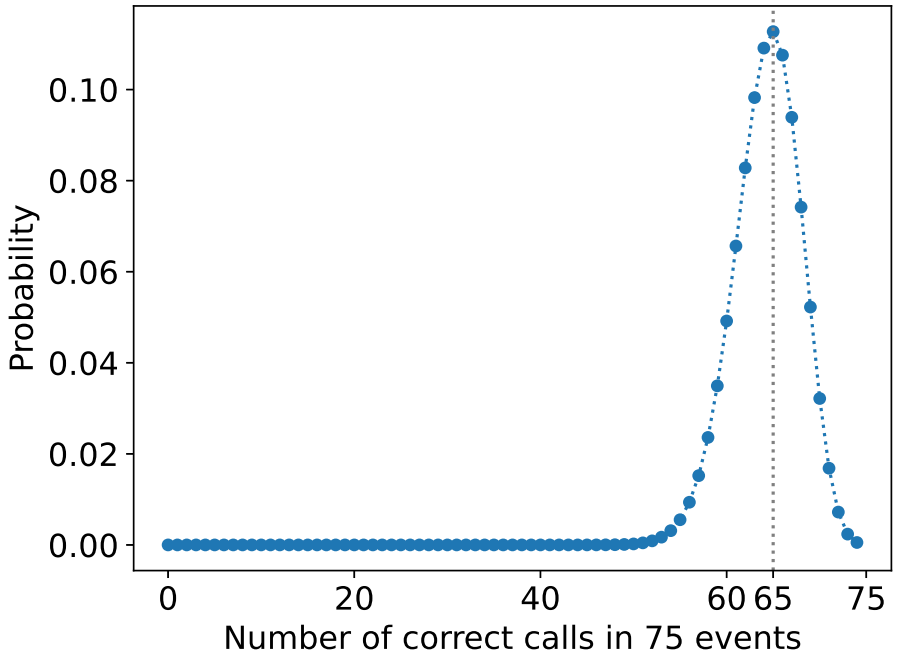
We now turn to the problem of computing a posterior predictive distribution (PPD). In our context, it answers the following question: Given that in the discovery study k of the n cancers were correctly called, what is the probability that in a future study of n' cancers, k' cancers will be correctly called. Mathematically, this is represented as $p(k', n'|k, n)$ and can be computed by marginalizing the likelihood over the posterior:

$$p(k', n'|k, n) = \int_0^1 \binom{n'}{k'} \sigma^{k'}(1-\sigma)^{n'-k'} p(\sigma|k,n) d\sigma \tag{6}$$

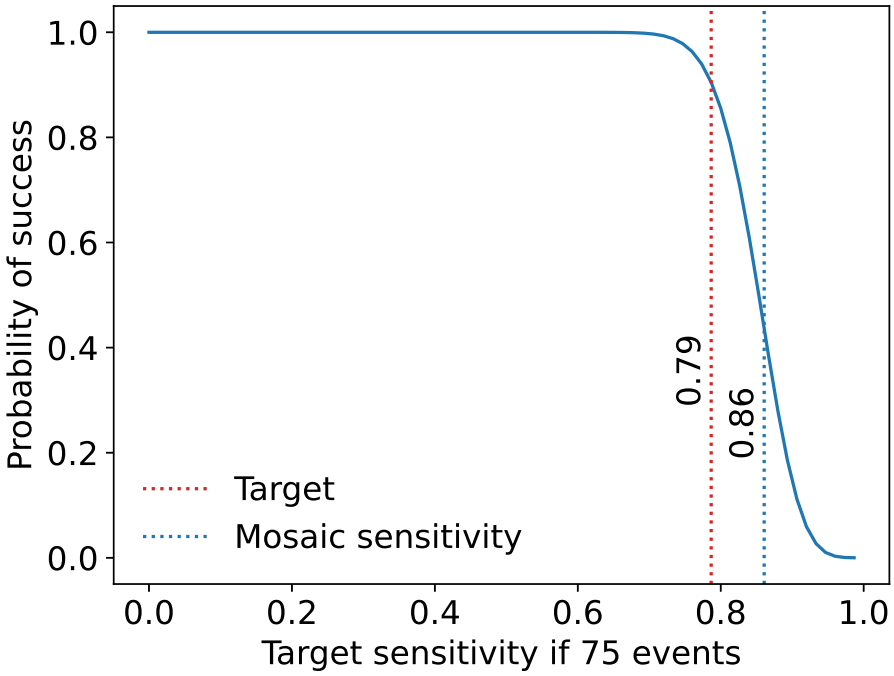
Once again, the method of conjugate priors allows us to obtain a closed form solution for this integral which is the beta binomial distribution.

$$p(k', n'|k, n) = BetaBinom(n', k+1, n-k+1) \tag{7}$$

The plot below shows the PPD for $n' = 75$ events. The vertical dashed line shows the number of cancers that are most likely to be called correctly i.e. $167/194 \times 75 = 64.56$ rounded to 65, which is also the mode of this distribution.



The probability of successful validation i.e. the probability that the point estimate of sensitivity is greater than the target, is computed as the survival function of the PPD and shown in the figure below as a function of the target sensitivity. If we want a 90% probability of successful validation, we should set our performance target at 79%.



If we want to set a higher sensitivity target for competitive reasons, we should either increase our risk tolerance or accrue more events. This information can be tabulated in a look-up table modeling the target sensitivity for given number of events and different levels of risk as shown below.

Probability of success Number of events	0.80	0.85	0.90	0.95
25	76	76	72	72
50	80	78	78	74
75	82	80	79	78
100	82	81	80	78
125	83	81	80	79
150	82	82	81	79
175	83	82	81	79
200	83	82	81	80

Table 1: Target sensitivities for different number of events and success probability.