Deriving the Posterior Predictive Distribution (Discrete Prior Case)

Objective

In Bayesian analysis, we often want to predict future observations after observing data. Let Y_{new} be a future observation and π be an unknown parameter (e.g., a success probability). The posterior predictive distribution is:

$$P(Y_{\text{new}} = y_{\text{new}} \mid \text{data})$$

Bayesian Principle

The predictive distribution accounts for uncertainty in the parameter π. We do this by integrating over the posterior distribution of π:

$$P(Y_{\text{new}} = y_{\text{new}} \mid \text{data}) = \int P(Y_{\text{new}} = y_{\text{new}} \mid \pi) \cdot P(\pi \mid \text{data})\, d\pi$$

Intuition

The posterior predictive distribution arises from a simple and powerful idea: **we don't know the true value of π, so we average over our uncertainty**.

After observing data, we have an updated belief (posterior) about what π could be. We then imagine predicting future data under each possible π_k, but we don't commit to one value. Instead, we *blend* the predictions from all possible values of π, weighted by how plausible each value is according to our posterior.

Discrete Prior Case

If the prior on π is discrete, say:

$$\pi \in \{\pi_1, \pi_2, \ldots, \pi_K\} \quad \text{with} \quad P(\pi = \pi_k) = p_k$$

then the posterior is also discrete:

$$P(\pi = \pi_k \mid \text{data}) = \tilde{p}_k$$

So the integral becomes a sum:

$$P(Y_{\text{new}} = y_{\text{new}} \mid \text{data}) = \sum_{k=1}^{K} P(Y_{\text{new}} = y_{\text{new}} \mid \pi = \pi_k) \cdot \tilde{p}_k$$

Example: Binomial Likelihood

Assume $Y_{\text{new}} \sim \text{Binomial}(n, \pi)$, then:

$$P(Y_{\text{new}} = y_{\text{new}} \mid \pi = \pi_k) = \binom{n}{y_{\text{new}}} \pi_k^{y_{\text{new}}} (1 - \pi_k)^{n - y_{\text{new}}}$$

So the posterior predictive becomes:

$$P(Y_{\text{new}} = y_{\text{new}} \mid \text{data}) = \sum_{k=1}^{K} \tilde{p}_k \cdot \binom{n}{y_{\text{new}}} \pi_k^{y_{\text{new}}} (1 - \pi_k)^{n - y_{\text{new}}}$$

Interpretation

This is a **mixture of binomial distributions**, where each component corresponds to one possible value of π, weighted by its posterior probability.

- Each term: the probability of observing y_{new} given a fixed π_k.
- Weight: the posterior probability that $\pi = \pi_k$ after observing the data.

Key Point

The posterior predictive distribution takes into account both the randomness in the data-generating process and the uncertainty in the parameter π. It is computed using the law of total probability, averaging over all plausible values of π given the data.