# Bayesian Statistics Notes

## Overview of Bayesian Statistics

- **Bayesian Statistics** utilizes probability distributions to model uncertainty about parameters and employs the posterior distribution to represent this uncertainty.
- **Bayes' Rule:**

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)}$$

  - $p(\theta)$: Prior distribution (knowledge before data).
  - $p(D|\theta)$: Likelihood function (probability of the data given the parameters).
  - $p(D)$: Marginal likelihood (normalizing constant).

## Posterior Predictive Distribution

- **Posterior Predictive:** Used for making predictions after observing data.

$$p(y|x, D) = \int p(y|x, \theta)p(\theta|D)d\theta$$

- **Bayesian Model Averaging (BMA)**: Reduces overfitting by averaging over models.

## Conjugate Priors

- A prior $p(\theta)$ is conjugate if the posterior $p(\theta|D)$ belongs to the same family as the prior.
- **Exponential Families**: Closed-form solutions are possible when conjugate priors are used.

## Beta-Binomial Model Example

In the Beta-Binomial model, we infer the probability of heads in $N$ coin tosses. Let $y_n$ denote the outcome of the $n$-th toss, where $y_n = 1$ represents heads and $y_n = 0$ represents tails. The data is represented as $D = \{y_n : n = 1, \ldots, N\}$, and we assume $y_n \sim \mathrm{Ber}(\theta)$, where $\theta$ is the probability of heads.

### Bernoulli Likelihood

Assuming the data are independent and identically distributed (iid), the likelihood function is given by:

$$p(D|\theta) = \prod_{n=1}^{N} \theta^{y_n}(1-\theta)^{1-y_n} = \theta^{N_1}(1-\theta)^{N_0}$$

where $N_1$ is the number of heads and $N_0$ is the number of tails (sufficient statistics).

### Binomial Likelihood

Alternatively, we can consider the binomial likelihood:

$$p(D|\theta) = \mathrm{Bin}(y|N, \theta) = \binom{N}{y}\theta^y(1-\theta)^{N-y}$$

where $\binom{N}{y}$ is the binomial coefficient, and $y$ is the number of heads observed in $N$ trials.

### Prior

We assume a Beta prior distribution:

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} = \text{Beta}(\theta|\alpha, \beta)$$

where $\alpha$ and $\beta$ are hyperparameters.

### Posterior Distribution

Given the Beta prior, the posterior distribution is:

$$p(\theta|D) \propto \text{Beta}(\theta|\alpha + N_1, \beta + N_0)$$

where $N_1$ and $N_0$ are the counts of heads and tails, respectively.

### MAP Estimate (Posterior Mode)

The Maximum A Posteriori (MAP) estimate is the value of $\theta$ that maximizes the posterior:

$$\theta_{\text{map}} = \frac{\alpha + N_1 - 1}{\alpha + N_1 - 1 + \beta + N_0 - 1}$$

For a Beta(2,2) prior:

$$\theta_{\text{map}} = \frac{N_1 + 1}{N + 2}$$

For a uniform prior ($\text{Beta}(1, 1)$), the Maximum Likelihood Estimate (MLE) is:

$$\theta_{\text{mle}} = \frac{N_1}{N}$$

### Posterior Mean

The posterior mean is given by:

$$E[\theta|D] = \frac{\alpha + N_1}{\alpha + \beta + N}$$

### Posterior Variance

The variance of the posterior is:

$$V[\theta|D] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

If $N$ is large relative to $\alpha + \beta$, the variance approximates:

$$V[\theta|D] \approx \frac{\theta_{\text{hat}}(1 - \theta_{\text{hat}})}{N}$$

Uncertainty decreases at a rate of $1/\sqrt{N}$.

## Bias-Variance Tradeoff

- **Bias:** The difference between the expected value of an estimator and the true parameter value. High bias leads to systematic errors.
- **Variance:** The variability of the estimator across different datasets. High variance means the estimator is sensitive to small changes in the data.
- **Tradeoff:** Reducing bias increases variance, and vice versa. The goal is to minimize the Mean Squared Error (MSE):

$$\text{MSE} = \text{Bias}^2 + \text{Variance}$$