Bayesian basics
let $X$ and $Y$ be discrete rvs taking values in $\mathbb{N}$. denote the joint pmf:

$$p_{XY}(x, y) = \mathbb{P}[X = x, Y = y]$$

**marginalization**: computing individual pmfs from joint pmfs as

$$p_X(x) = \sum_{y \in \mathbb{N}} p_{XY}(x, y) \quad p_Y(y) = \sum_{x \in \mathbb{N}} p_{XY}(x, y)$$

**conditioning**: pmf of $X$ given $Y = y$ (with $p_Y(y) > 0$) defined as:

$$\mathbb{P}[X = x | Y = y] \triangleq p_{X|Y}(x|y) = \frac{p_{XY}(x, y)}{p_Y(y)}$$

more generally, can define $\mathbb{P}[X \in A | Y \in B]$ for sets $A, B \in \mathbb{N}$

see also this visual demonstration
the basic ‘rules’ of Bayesian inference

let $X$ and $Y$ be discrete rvs taking values in $\mathbb{N}$, with joint pmf $p(x, y)$

<table>
<thead>
<tr>
<th><strong>Product rule</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>for $x, y \in \mathbb{N}$, we have: $p_{X,Y}(x, y) = p_X(x) p_{Y</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Sum rule</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>for $x \in \mathbb{N}$, we have: $p_X(x) = \sum_{y \in \mathbb{N}} p_{X</td>
</tr>
</tbody>
</table>

and most importantly!

<table>
<thead>
<tr>
<th><strong>Bayes rule</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>for any $x, y \in \mathbb{N}$, we have: $p_{X</td>
</tr>
</tbody>
</table>

see also [this video](#) for an intuitive take on Bayes rule
fundamental principle of Bayesian statistics

- assume the world arises via an underlying generative model $M$
- use random variables to model all unknown parameters $\theta$
- incorporate all that is known by conditioning on data $D$
- use Bayes rule to update prior beliefs into posterior beliefs

\[
p(\theta|D, M) \propto p(\theta|M)p(D|\theta, M)
\]

• Physics - Newtonian dynamics, relativity

• Note - Bayesian ML DOES NOT believe the model parameters are random
pros and cons

**in praise of Bayes**

- conceptually simple and easy to interpret
- works well with small sample sizes and overparametrized models
- can handle all questions of interest: no need for different estimators, hypothesis testing, etc.

**why isn’t everybody Bayesian**

- they need priors (subjectivity . . . )
- they may be more computationally expensive: computing normalization constant and expectations, and updating priors, may be difficult
basics of Bayesian inference
the likelihood principle

given model $\mathcal{M}$ with parameters $\Theta$, and data $D$, we define:

- the prior $p(\Theta|\mathcal{M})$: what you believe before you see data
- the posterior $p(\Theta|D,\mathcal{M})$: what you believe after you see data
- the marginal likelihood or evidence $p(D|\mathcal{M})$: how probable is the data under our prior and model

these three are probability distributions; the next is not

- the likelihood: $\mathcal{L}(\Theta) \triangleq p(D|\mathcal{M},\Theta)$: function of $\Theta$ summarizing data

the likelihood principle

given model $\mathcal{M}$, all evidence in data $D$ relevant to parameters $\Theta$ is contained in the likelihood function $\mathcal{L}(\Theta)$

this is not without controversy; see Wikipedia article
given model $\mathcal{M}$ with parameters $\Theta$, and data $D$, we define:

- the **prior** $p(\Theta|\mathcal{M})$: what you believe before you see data
- the **posterior** $p(\Theta|D, \mathcal{M})$: what you believe after you see data
- the **marginal likelihood or evidence** $p(D|\mathcal{M})$: how probable is the data under our prior and model
- the **likelihood**: $\mathcal{L}(\Theta) \triangleq p(D|\mathcal{M}, \Theta)$: function of $\Theta$ summarizing the data

the fundamental formula of Bayesian statistics

\[
\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}
\]

\[
p(\theta|D) = \frac{p(D|\theta) \ p(\theta)}{p(D)}
\]

also see: Sir David Spiegelhalter on Bayes vs. Fisher
Notes

- For discrete $\Theta$, $p(\Theta|D), p(\Theta)$ are pmfs.
  For continuous $\Theta$, $p(\Theta|D) = f(\Theta|D), p(\Theta) = f(\Theta)$ (use pdfs).

- Similarly for discrete vs continuous data (for $p(D)$).

- Likelihood $p(D|\Theta)$ is not a prob distn. It is a fn of $\Theta$ that is parameterized by the data.

- If $D$ is continuous, use $f(D|\Theta)$ - Note this is still some fn of $\Theta$.
• data $D = \{X_1, X_2, \ldots, X_n\} \in \{0, 1\}^n$

• model $\mathcal{M}$: $X_i$ are generated i.i.d. from a $\text{Ber}(\theta)$ distribution

**Example:** the mystery Bernoulli rv

- Fix $\theta$; what is $\mathbb{P}[X_i | \mathcal{M}]$ for any $i \in [n]$?

  - $N_1 = \# \text{ of 1s}$, $N_0 = \# \text{ of 0s}$
    - $N_0 + N_1 = n$

  $\mathbb{P}\left[ X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n | \mathcal{M}, \Theta \right] = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} = \theta^{N_1} (1-\theta)^{N_0}$
  - $x_i \in \{0, 1\}$

- Let $H = \# \text{ of '1's in } \{X_1, X_2, \ldots, X_n\}$; what is $\mathbb{P}[H | \mathcal{M}, \Theta]$?

  $\mathbb{P}[H = h | \mathcal{M}, \Theta] = \binom{n}{h} \theta^h (1-\theta)^{n-h} \sim \text{Bin}(n, \theta)$
the Bernoulli likelihood function

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0, 1\}^n$
- model $\mathcal{M}$: $X_i$ are generated i.i.d. from a $Ber(\theta)$ distribution

likelihood: $\mathcal{L}(\Theta) \doteq p(D|\mathcal{M}, \theta)$: function of $\Theta$ summarizing the data

$\mathcal{L}(\theta) = \theta^{N_1}(1-\theta)^{N_0}$

- Note: $\mathcal{L}(\theta)$ is NOT a distribution (i.e., $\int \mathcal{L}(\theta) d\theta \neq 1$)
log-likelihood, sufficient statistics, MLE

- \( l(\theta) = \log L(\theta) \)

(For Bernoulli, \( l(\theta) = \log (\theta^{N_1} (1-\theta)^{N_0}) = N_1 \log \theta + N_0 \log (1-\theta) \))

- \((N_1, N_0)\) are sufficient statistics of \( D \)

(i.e. \( L(\theta | D) = \text{parametric fn of } N_1 \text{ and } N_0 \))

- MLE - \( \arg \max_{\theta \in [0,1]} L(\theta) = \arg \max_{\theta \in [0,1]} l(\theta) = \frac{N_1}{N_1+N_0} = \frac{N_1}{n} \)
Cromwell’s Rule

How should we choose the prior?

The zeroth rule of Bayesian statistics:

Never set $p(\theta|M) = 0$ or $p(\theta|M) = 1$ for any $\theta$.

- “I beseech you, in the bowels of Christ, think it possible that you may be mistaken.” (Oliver Cromwell, 1650)

- Connected to philosophy of science (Falsifiability)

Also see: Jacob Bronowski on Cromwell’s Rule and the scientific method.
from where do we get a prior?

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0, 1\}^n$
- model $\mathcal{M}$: $X_i$ are generated i.i.d. from a $\text{Ber}(\theta)$ distribution

option 1: from the ‘problem statement’

Mackay example 2.6

- eleven urns labeled by $u \in \{0, 1, 2, \ldots, 10\}$, each containing ten balls
- urn $u$ contains $u$ red balls and $10-u$ blue balls
- select urn $u$ uniformly at random and draw $n$ balls with replacement, obtaining $n_R$ red and $n_{nR}$ blue balls

$$P(\theta) = \text{Unif}\left\{\frac{0}{10}, \frac{1}{10}, \frac{2}{10}, \ldots, \frac{10}{10}\right\}$$
from where do we get a prior

- data \( D = \{X_1, X_2, \ldots, X_n\} \in \{0, 1\}^n \)
- model \( M \): \( X_i \) are generated i.i.d. from a \( \text{Ber}(\theta) \) distribution

**option 2: the maximum entropy principle**

choose \( p(\theta|M) \) to be distribution with maximum entropy given \( M \)
we know \( \theta \in [0, 1] \)

- Maximum entropy prior on \([0,1]\) is \( U[0,1] \)
from where do we get the prior, take 2

- data \( D = \{X_1, X_2, \ldots, X_n\} \in \{0, 1\}^n \)
- model \( \mathcal{M} \): \( X_i \) are generated i.i.d. from a \( \text{Ber}(\theta) \) distribution

**option 3: easy updates via conjugate priors**

- prior \( p(\theta) \) is said to be **conjugate** to likelihood \( p(D|\theta) \) if corresponding posterior \( p(\theta|D) \) has same functional form as \( p(\theta) \)
- natural conjugate prior: \( p(\theta) \) has same functional form as \( p(D|\theta) \)
- conjugate prior family: **closed under Bayesian updating**

**Note** - The family of all distributions is trivially a conjugate prior... we want more useful families
the Beta distribution

**Beta distribution**

- \( x \in [0, 1] \), parameters: \( \Theta = (\alpha, \beta) \in \mathbb{R}^+ \) (‘# ones’+1, ‘# zeros’+1)
- pdf: \( p(x) \propto x^{\alpha-1}(1-x)^{\beta-1} \)
- normalizing constant: \( \frac{1}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \)

\[ \begin{align*}
\alpha &= \beta = 0.5 \\
\alpha &= 5, \beta = 1 \\
\alpha &= 1, \beta = 3 \\
\alpha &= 2, \beta = 2 \\
\alpha &= 2, \beta = 5
\end{align*} \]

\[ \begin{align*}
\alpha &= \beta = 0.5 \\
\alpha &= 5, \beta = 1 \\
\alpha &= 1, \beta = 3 \\
\alpha &= 2, \beta = 2 \\
\alpha &= 2, \beta = 5
\end{align*} \]
Beta-Bernoulli prior and updates

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0, 1\}^n$, contains $N_1$ ones and $N_0$ zeros
- model $\mathcal{M}$: $X_i$ are generated i.i.d. from a $Ber(\theta)$ distribution

Beta-Bernoulli model

- prior parameters: $\Theta_0 = (\alpha, \beta) \in \mathbb{R}^+$ (hyperparameters)
- Beta-Bernoulli prior: $Beta(\alpha, \beta) \sim p(\theta) \propto \theta^{\alpha-1}(1 - \theta)^{\beta-1}$
- likelihood: $p(D|\theta) = \theta^{N_1}(1 - \theta)^{N_0}$

then via Bayesian update we get

- posterior:

$$p(\theta|D) \propto \theta^{\alpha-1}(1 - \theta)^{\beta-1} \theta^{N_1}(1 - \theta)^{N_0} \sim Beta(\alpha + N_1, \beta + N_0)$$
the Beta distribution: getting familiar

**Beta**\((\alpha, \beta)\) distribution

\[ p(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1 - x)^{\beta-1} \]

properties of \(\Gamma(\alpha)\)

\[ \frac{1}{B(\alpha, \beta)} = \int_0^1 x^{\alpha-1} (1 - x)^{\beta-1} \, dx = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \]

\(\Gamma(\alpha)\) = \(\int_0^\infty e^{-y} y^{\alpha-1} \, dy\), \(\Gamma(\alpha+1) = \alpha \Gamma(\alpha)\)

- If \(\alpha\) is an integer, \(\Gamma(\alpha) = (\alpha-1)!\)
the Beta distribution: mean and mode

**Beta**\(\alpha, \beta\) **distribution**

\[
p(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1}
\]

\[
E[X] = \int_0^1 x \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1} \, dx
\]

\[
= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\beta)\Gamma(\alpha+1)}{\Gamma(\alpha+\beta+1)} = \frac{\alpha}{\alpha + \beta}
\]

Thus, mean of **Beta**\(\alpha, \beta\) dist is \(\frac{\alpha}{\alpha + \beta}\)
mode \ = \ \arg \max_{\theta \in [0,1]} \ \frac{x^{\alpha-1} \ (1-x)^{\beta-1}}{\text{B}(\alpha, \beta)}

\frac{d}{dx} \left( x^{\alpha-1} \ (1-x)^{\beta-1} \right) = (\alpha-1) \ x^{\alpha-2} \ (1-x)^{\beta-1} - (\beta-1) \ x^{\alpha-1} \ (1-x)^{\beta-1} = 0

\Rightarrow \ (\alpha-1) \ (1-x^*) = (\beta-1) \ x^*

\Rightarrow \ x^* = \frac{\alpha-1}{\alpha+\beta-2} \quad \text{(for } \alpha > 1, \ \alpha+\beta > 2)
Beta-Bernoulli model: what should we report?

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0, 1\}^n$, contains $N_1$ ones and $N_0$ zeros
- model $M$: $X_i$ are generated i.i.d. from a $Ber(\theta)$ distribution
- prior: $p(\theta) \sim Beta(\alpha, \beta)$  posterior: $p(\theta|D) \sim Beta(\alpha + N_1, \beta + N_0)$

Correct Answer - You should report Model, Prior, Posterior

Decision theoretic answer - Ask for a loss fn, report $\Theta$ which minimizes loss
• Choose ‘actions’ to minimize a loss function (stats/ML) and maximize a utility function (economic).

• E.g., let $\Theta$ be a sample from posterior. Output $\hat{\Theta}$ to minimize:
  1) $L(\Theta, \hat{\Theta}) = \mathbb{1}[\Theta \neq \hat{\Theta}]$ (L_0 loss) - $\hat{\Theta}_{L_0}$ = mode of posterior distn
  2) $L(\Theta, \hat{\Theta}) = |\Theta - \hat{\Theta}|$ (L_1 loss) - $\hat{\Theta}_{L_1}$ = median of posterior distn
  3) $L(\Theta, \hat{\Theta}) = (\Theta - \hat{\Theta})^2$ (L_2 loss) - $\hat{\Theta}_{L_2}$ = mean of posterior distn

In general, return $\arg \min_{\hat{\Theta}} \mathbb{E}_{\Theta \sim \text{posterior}} [L(\Theta, \hat{\Theta})]$.
Beta-Bernoulli model: posterior mean

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0, 1\}^n$, contains $N_1$ ones and $N_0$ zeros
- model $M$: $X_i$ are generated i.i.d. from a $Ber(\theta)$ distribution
- prior: $p(\theta) \sim Beta(\alpha, \beta)$  
  posterior: $p(\theta|D) \sim Beta(\alpha + N_1, \beta + N_0)$

Posterior mean:

$$\mathbb{E}[\theta|\alpha, \beta, N_0, N_1] = \mathbb{E}\left[\frac{1}{\alpha + \beta + N_1 + N_0} \right]$$

Define $m = \alpha + \beta$ 
$n = N_1 + N_0$ 
$m \equiv \text{number of prior samples}$
$\alpha \equiv \text{prior mean}$
$\beta \equiv \text{prior mean}$
$N_1 \equiv \text{data mean (also, MLE)}$
$n \equiv \text{data mean (also, MLE)}$
$w = \frac{m}{m+n} \equiv \text{strength of prior relative to data}$

$$= \frac{\alpha + N_1}{\alpha + \beta + N_1 + N_0} = \frac{\alpha + N_1}{m + n}$$

$$= \frac{\alpha}{m} \cdot \frac{m}{m+n} + \frac{N_1}{n} \cdot \frac{n}{m+n}$$

$$= w \cdot \frac{\alpha}{m} + (1-w) \cdot \frac{N_1}{m} \quad \text{regularization, 'shrinkage' of MLE}$$
Beta-Bernoulli model: posterior mode (MAP estimation)

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0, 1\}^n$, contains $N_1$ ones and $N_0$ zeros
- model $M$: $X_i$ are generated i.i.d. from a $Ber(\theta)$ distribution
- prior: $p(\theta) \sim Beta(\alpha, \beta)$  posterior: $p(\theta|D) \sim Beta(\alpha + N_1, \beta + N_2)$

posterior mode: $\max_{\theta \in [0,1]} p(\theta|\alpha, \beta, N_0, N_1) = \frac{\alpha + N_1 - 1}{\alpha + \beta + N_1 + N_2 - 2}$

- If $\alpha = \beta = 1$ (i.e., uniform prior), then $\theta_{MAP} = \frac{N_1}{N_1 + N_2} = \theta_{MLE}$

In general, if prior is uniform, then $\theta_{MLE} = \theta_{MAP}$
Beta-Bernoulli model: posterior prediction (marginalization)

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0, 1\}^n$, contains $N_1$ ones and $N_0$ zeros
- model $M$: $X_i$ are generated i.i.d. from a $Ber(\theta)$ distribution
- prior: $p(\theta) \sim Beta(\alpha, \beta)$  \quad \text{posterior: } p(\theta|D) \sim Beta(\alpha + N_1, \beta + N_2)$

posterior prediction: $P[X = 1|D] = \int_0^1 p(\theta) \cdot \theta \cdot d\theta = E[\theta] = \frac{\alpha + N_1}{\alpha + \beta + N_1 + N_2}$

If $\alpha = \beta = 1$, $P[X = 1|D] = \frac{N_1 + 1}{N_1 + N_2 + 2}$ (Laplace Estimator or 'add-one' smoothing)
If we observe $N_0 = n$, then what is $P[X_{n+1} = 1]$?

- MLE: $P_{\text{MLE}}[X_{n+1} = 1] = 0$, $P_{\text{MLE}}[X_{n+1} = 0] = 1$

- Laplace (i.e., Bayesian update with Beta(1,1) prior)

$$P_{\text{lap}}[X_{n+1} = 1] = \frac{1}{n+2}, P_{\text{lap}}[X_{n+1} = 0] = \frac{n+1}{n+2}$$

More ‘0’s that we see, less unlikely the arrival of a ‘1’ however, not impossible! Remember Cromwell’s law