Example 0.1. Let $x_1, \ldots, x_6 \overset{iid}{\sim} \text{Bern}(\theta)$ be the data set $\langle 0,0,1,0,1,0 \rangle$. Then:

$$\begin{aligned}
l(\theta; x) &= \ln(\prod_{i=1}^{6} \theta^{x_i}(1-\theta)^{1-x_i}) \\
&= \sum_{i=1}^{6} \ln(\theta^{x_i}(1-\theta)^{1-x_i}) \\
&= \sum_{i=1}^{6} x_i \ln(\theta) + (1-x_i)\ln(1-\theta) \\
&= \ln(\theta)\sum_{i=1}^{6} x_i + (6 - \sum_{i=1}^{6} x_i)\ln(1-\theta) \\
&= \ln(\theta)6\bar{x} + (6-6\bar{x})\ln(1-\theta) \\
&= 6(\ln(\theta) + (1-\bar{x})\ln(1-\theta))
\end{aligned}$$

Now let's differentiate this to maximize it:

$$\frac{d}{dt}6(\ln(\theta) + (1-\theta)\ln(1-\theta)) = 6(\frac{\bar{x}}{\theta} - \frac{1-\bar{x}}{1-\theta})$$

If we set it equal to 0,

$$(1-\theta)\bar{x} - \theta(1-\bar{x}) = 0 \to \hat{\theta}_{MLE} = \bar{x}$$

Note: For our convenience, we use the natural log to differentiate $\prod$ to $\sum$. It is easier to differentiate sums rather than products.

Definition 0.1. Maximum Likelihood Estimation: $\hat{\theta}_{MLE} = \bar{X}$ where $\bar{X}$ is a random variable and has properties

Definition 0.2. Maximum Likelihood Estimate: $\hat{\theta}_{MLE} = \bar{x}$ where $\bar{x}$ has a numerical value

Example 0.2. Let $x_1, \ldots, x_n \overset{iid}{\sim} \text{Geom}(\theta) = (1-\theta)^x\theta$ where x is the number of failures before stopping success. $\text{Supp}(X) = \{0, 1, \ldots\} = \mathbb{N}$ and $\Theta = (0,1)$. Then:

$$\begin{aligned}
p(x_i, \ldots, x_n) &= \mathcal{L}(\theta; x_i, \ldots, x_n) \\
&= \prod_{i=1}^{n} (1-\theta)^{x_i}\theta
\end{aligned}$$

Therefore

$$\begin{aligned}
l(\theta; x) &= \sum \ln(1-\theta)^{x_i}\theta \\
&= \ln(1-\theta)\sum x_i + n\ln(\theta)
\end{aligned}$$

We will now differentiate this function to solve for $\hat{\theta}_{MLE}$.

$$\begin{aligned} l'(\theta; x) &= \frac{n}{\theta} - \frac{n\bar{x}}{1-\theta} = 0 \\ \frac{1}{\theta} &= \frac{\bar{x}}{1-\theta} \\ \frac{1}{\theta - 1} &= \bar{x} \\ \hat{\theta}_{MLE} &= \frac{1}{\bar{x}+1} \end{aligned}$$

Properties of MLE:

1. There exists $\varepsilon > 0$ such that

$$\lim_{n\to\infty} P(|\hat{\theta}_{MLE} - \theta| \geq \varepsilon) = 0$$

2. Asymptotic Normaling: As n increases, the the parameters behave like a normal distribution

$$\hat{\theta}_{MLE} \xrightarrow{d} N(\hat{\theta}_{MLE}, SE(\hat{\theta}_{MLE})^2)$$

3. Efficiency: $\hat{\theta}_{MLE}$ has the lowest standard error theoretically possible

Inference with MLE:

- Point Estimate: $\hat{\theta}_{MLE}$
- Confidence Set: $CI_{\theta,1-\alpha} = [\hat{\theta}_{MLE} \pm z_{\frac{\alpha}{2}} SE(\hat{\theta}_{MLE})]$
  Here, θ is the parameter of interest whereas $1-\alpha$ is the confidence level.
- Hypothesis Testing: $H_0 : \theta = \theta_0$, $H_A : \theta \neq \theta$ - fail to reject if $\hat{\theta}_{MLE}$ is in the region of $[\theta_0 \pm z_{\frac{alpha}{2}} SE(\hat{\theta}_{MLE})]$

We must observe data, then pick a parametric model $\mathcal{F}$, do inference with MLE. The problem with this is that

1. If all data values taken are 0 and we take $\mathcal{F} = \text{Bern}(\theta)$, then $\hat{\theta}_{MLE} = \bar{x} = 0$ and $SE(\bar{\theta}_{MLE}) = \sqrt{\bar{\theta}_{MLE}(1-\bar{\theta}_{MLE})} = 0$. This gives no information and thus is a big problem. No confidence set, no hypothesis testing.
2. What if we have prior knowledge about Θ? We can't use it because only data set can be used.
3. Frequentist Confidence Interval Interpretation: Let's say we found $CI_{\theta,1-\alpha} = [0.42, 0.47]$. If the experiment is repeated "many" times, then a confidence level of 95% will cover θ and $1-\alpha$ is contained in the set. But given just an interval, we can only say that a certain value will either fall in the interval or not. We can't claim that the probability that the interval contains θ is $1-\alpha$.

4. Hypothesis testing: not satisfactory since we do not know if data values are far from being retained yet rejected or near rejection (extremeness). How good is the rejection? What is $P(H_0|x)$, or H_0 given x?

5. Boundary Issues: Let's say $x = \langle 0, 0, 1, 0, 1, 0\rangle$ and $\hat{\theta}_{MLE} = \frac{1}{3}$. We want a confidence set at the 95% confidence level: $CI_{\theta,95\%} = (\frac{1}{3} \pm 2\sqrt{\frac{1}{3}\frac{2}{3}}) = (-0.6, 1.26)$. In this confidence interval, we have both a negative value and one that's greater than 1. This is no good. This happened because our data set is only composed of 6 values. Thus it cannot converge to normality. We cannot use the normal distribution to construct the interval and since we did, it came out looking wrong.

Good news: The Bayesian approach will not cause any of these issues.

Definition 0.3. Conditional Probability: $P(B|A)$, the probability of B occurring given A occurs

$$P(B|A) = \frac{P(A, B)}{P(A)}$$

Note: There is a proportionality between $P(A, B)$, the intersection of two events, and $P(B|A)$, the probability of B occurring given A occurs. Thus we can write

$$P(A, B) \propto P(B|A)$$

or

$$P(A, B) = cP(B|A)$$

Definition 0.4. Baye's Rule:

$$P(B|A) = \frac{P(A, B)}{P(A)}$$

We know from previous probability courses that $P(A, B) = P(B, A)$. We also know that $P(A, B) = P(B|A)P(A)$ and $P(B, A) = P(A|B)P(B)$. Let's set them equal to each other.

$$\begin{aligned} P(A, B) &= P(B, A) \\ P(B|A)P(A) &= P(A|B)P(B) \end{aligned}$$

This is another form of Baye's rule.

Definition 0.5. Law of Total Probability: the probability of event A occurring is sum of the probability of the intersection of event A and event B and the probability of the intersection of event A and not event B (complement of B)

$$P(A) = P(A, B) + P(A, B^C)$$

Let's combine the two equations from above.

$$\begin{aligned} P(A) &= P(A, B) + P(A, B^C) \\ &= P(A|B)P(B) + P(A|B^C)P(B^C) \\ P(B|A) &= \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^C)P(B^C)} \end{aligned}$$

This is another form of Baye's rule.

Note:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

The LHS is the posterior probability where B is the parameter of interest, A is the evidence/data, and $B|A$ is the targeted estimation. On the RHS, $P(A|B)$ is the likelihood or probability of data and $P(B)$ is a prior probability.

Finding $P(B|A)$ using A(data) and applying it to $P(B)$ is called Bayesian conditioning.