MLE for the Bernoulli/binomial model

Suppose $X_i \sim Ber(\theta)$ where $X_i = 1$ represents "heads", $X_i = 0$ represents "tails", and

$\theta \in [0,1]$ is the rate parameter (probability of heads). If the data are iid (identically independent distributed), the likelihood has the form

$$p(D|\theta) = \theta^{N_1}(1-\theta)^{N_0} \qquad (*)$$

Where we have $N_1 = \sum_{i=1}^{N} \mathrm{I}(x_i = 1)$ heads and $N_0 = \sum_{i=1}^{N} \mathrm{I}(x_i = 0)$ tails. ($\mathrm{I}(x)$ is indicator function). These two counts are called the sufficient statistics of the data, since this is all we need to know about D to infer θ.

More formally, we say $s(D)$ is a sufficient statistic for data D if $p(D|\theta) = p(\theta|s(data))$ .if we use a uniform prior, this is equivalent to saying $p(D|\theta \propto p(s(D|\theta)$.Consequently, if we have two datasets with the same sufficient statistics, we will infer the same value for θ.

Now suppose the data consists of the count of the number of heads N_1 observed in a fixed number $N = N_0 + N_1$ for trials. In this case, we have $N_1 \sim Bin(N,\theta)$,where Bin represents the binomial distribution, which has the following pmf (probability mass function):

$$Bin(k|n,\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

Since $\binom{n}{k}$ is a constant independent of θ,the likelihood for the binomial sampling model is the same as the likelihood for the Bernoulli model. So any inferences we make about θ will be the same whether we observe the counts, $D = (N_1, N)$, or a sequence of trials, $D = \{x_1, \dots, x_N\}$.

Optimizing of equation $(*)$:

Taking the logarithm of the sides of equation (*) and then taking the partial derivative with respect to θ,

$$\log p(D|\theta) = N_1 \log\theta + N_0 \log(1-\theta)$$

$$\frac{\partial(\log p(D|\theta))}{\partial\theta} = \frac{N_1}{\theta} - \frac{N_0}{1-\theta}$$

And by setting equal to zero, we have:

$$\frac{N_1}{\theta} = \frac{N_0}{1-\theta} \xrightarrow{or} N_1(1-\theta) = N_0\theta \xrightarrow{or} N_1 = (N_0 + N_1)\theta = N\theta \xrightarrow{or} \theta = \frac{N_1}{N}$$

If we denote maximum likelihood estimate by MLE, we conclude that,

$$\theta_{MLE} = \frac{N_1}{N}$$

(which is just the empirical fraction of heads).This θ can maximize the likelihood function in equation (*) .