Supervised Maximum Likelihood Estimation (MLE)

First scenario: (toss a ′coin′ $Z$)

Second scenario: (toss $Z$; (then A or B))

Common-sense and relative frequency

Suppose a 2-sided ′coin′ $Z$, one side labelled ′a′, other side labelled ′b′

$P(Z = a)$: probability of giving ′a′ when tossed – currently not known

$P(Z = b)$: probability of giving ′b′ when tossed – currently not known

Suppose you have data $d$ recording 100 tosses of $Z$

if there were (50 a, 50 b) in $d$, ′common-sense′ says $P(Z = a) = 50/100$

if there were (30 a, 70 b) in $d$, ′common-sense′ says $P(Z = a) = 30/100$

ie. you ′define′ or ′estimate′ the probability by the relative frequency
Data likelihood

assuming the tosses of $Z$ are all independent, can work out the probability of the observed data $d$ if $Z$’s probabilities had particular values.

let $\theta_a$ and $\theta_b$ stand for $P(Z = a)$ and $P(Z = b)$

let #($a$) be the number of ’a’ outcomes in the sequence $d$

let #($b$) be the number of ’b’ outcomes in the sequence $d$

the probability of $d$, assuming the probability settings $\theta_a$ and $\theta_b$ is

$$p(d) = \theta_a^{#(a)} \times \theta_b^{#(b)}$$

(1)

different settings of $\theta_a$ and $\theta_b$ will give different values for $p(d)$

following slides investigate this empirically

$p(d)$ for 50 $a$, 50 $b$

as $\theta_a$ is varied, data prob $p(d)$ varies
max occurs at $\theta_a = 0.5$
which is $\frac{50}{50 + 50}$

$p(d)$ for 30 $a$, 70 $b$

as $\theta_a$ is varied, data prob $p(d)$ varies
max occurs at $\theta_a = 0.3$
which is $\frac{30}{30 + 70}$

$p(d)$ for 70 $a$, 30 $b$

as $\theta_a$ is varied, data prob $p(d)$ varies
max occurs at $\theta_a = 0.7$
which is $\frac{70}{70 + 30}$
in each case, it looks like the max of the data probability occurred at the value given by the relative frequency

this suggests that in these cases, Max. Likelihood Estimator

if you wanted to find \( \theta_a \) (and \( \theta_b \)) that maximise the data probability, that is you want

\[
\arg\max_{\theta_a, \theta_b} p(\mathbf{d}; \theta_a, \theta_b)
\]

then the relative frequencies would give the answer, that is

\[
\theta_a = \frac{\#(a)}{\#(a) + \#(b)} \quad \theta_b = \frac{\#(b)}{\#(a) + \#(b)}
\]

technically expressed as: the relative frequency is a maximum likelihood estimator of the parameters

Define \( L(\theta_a) \) as \( \log(P(\mathbf{d}; \theta_a)) \). Then you get

\[
L(\theta_a) = \#(a) \log \theta_a + \#(b) \log(1 - \theta_a)
\]

need to take derivative wrt to \( \theta_a \) and set to 0, which is

\[
\frac{dL(\theta_a)}{d\theta_a} = \frac{\#(a)}{\theta_a} - \frac{\#(b)}{1 - \theta_a} = 0 \quad \Rightarrow \quad \theta_a = \frac{\#(a)}{\#(a) + \#(b)} = \frac{\#(a)}{100}
\]

so in this scenario of 100 tosses of \( Z \), we have proven that the relative frequency is always going to the maximum likelihood estimator

now want to consider slightly more complex scenario

on reflection, if you have to set parameters given data, it makes a lot of sense to set the parameters to whatever values make the data as likely as possible formula for \( p(\mathbf{d}; \theta_a, \theta_b) \) is (1), repeated below

\[
p(\mathbf{d}; \theta_a, \theta_b) = \theta_a^{\#(a)} \times \theta_b^{\#(b)}
\]

and because \( \theta_b = 1 - \theta_a \) can really write this in terms of just parameter \( \theta_a \)

\[
p(\mathbf{d}; \theta_a) = \theta_a^{\#(a)} \times (1 - \theta_a)^{\#(b)}
\]

Looking at some pics suggested a formula for the value of \( \theta_a \) that maximises this. Can we actually derive this formula?

Yes ⇒ take the log of this – the log-likelihood and use calculus to maximize that w.r.t. \( \theta_a \) – this turns out to be (relatively) easy

a more complex scenario

suppose \( D \) repetitions of toss disc \( Z \), to choose one of two coins A or B then toss chosen coin 10 times

Suppose 9 repetitions gave

\[
\begin{array}{ccccccccccc}
\text{d} & Z & X: \text{tosses of chosen coin} & \text{H counts} \\
1 & A & H & H & H & H & H & H & T & T & (8H) \\
2 & B & T & T & T & H & T & T & T & T & (2H) \\
3 & A & H & T & H & T & H & H & H & H & (7H) \\
4 & A & H & T & H & T & H & H & H & H & (8H) \\
5 & B & T & T & T & T & T & T & T & T & (1H) \\
6 & A & H & H & T & H & H & H & H & H & (9H) \\
7 & A & T & H & T & H & H & H & H & H & (7H) \\
8 & A & H & H & H & T & H & H & H & H & (9H) \\
9 & B & H & H & T & T & T & T & T & T & (3H)
\end{array}
\]

Let \( \theta_a \) be \( Z \)’s probability of giving A
Let \( \theta_{a|A} \) be A’s probability of giving H
Let \( \theta_{a|B} \) be B’s probability of giving H
\[
\begin{align*}
\text{for } \theta_a, \text{ need } & \frac{\text{count of } Z = A \text{ cases}}{\text{count of all } Z \text{ cases}}, \text{ ie.} \\
\text{est}(\theta_a) &= \frac{\sum_{d:Z=A} 1}{D} = \frac{6}{9} = 0.66 \\
\text{for } \theta_{h|a}, \text{ need } & \frac{\text{count of } H \text{ when } A \text{ chosen}}{\text{count of all tosses when } A \text{ chosen}}, \text{ ie.} \\
\text{est}(\theta_{h|a}) &= \frac{\sum_{d:Z=A} \#(d,h)}{\sum_{d:Z=A} 10} = \frac{48}{60} = \frac{4}{5} = 0.8 \\
\text{for } \theta_{h|b}, \text{ need } & \frac{\text{count of } H \text{ when } B \text{ chosen}}{\text{count of all tosses when } B \text{ chosen}}, \text{ ie.} \\
\text{est}(\theta_{h|b}) &= \frac{\sum_{d:Z=B} \#(d,h)}{\sum_{d:Z=B} 10} = \frac{6}{30} = \frac{1}{5} = 0.2
\end{align*}
\]

it turns out that in this scenario also, the 'common-sense', relative-frequency answers are also maximum likelihood estimators ie. values which maximise the probability of the data, and again it is (relatively) easy to show this by taking logs and using calculus.

the formula for \( p(d; \theta_a, \theta_h|a, \theta_t|a, \theta_h|b, \theta_t|b) \)
\[
p(d) = \prod_{d:Z=A} \left[ \theta_a \theta_h^{(d,h)} \theta_t^{(d,t)} \right] \prod_{d:Z=B} \left[ \theta_b \theta_h^{(d,h)} \theta_t^{(d,t)} \right]
\]

and its log comes out as
\[
\sum_{d:Z=A} \left[ \log \theta_a + \#(d,h) \log \theta_{h|a} + \#(d,t) \log \theta_{t|a} \right] + \\
\sum_{d:Z=B} \left[ \log \theta_b + \#(d,h) \log \theta_{h|b} + \#(d,t) \log \theta_{t|b} \right]
\]
call this \( L(\theta_a, \theta_{h|a}, \theta_{h|b}) \)
to make the comparison with the hidden variable version which will come up later, its worth noting that we can formulate all the restricted sums \( \sum_{d:Z=A} (\Phi(d)) \) with unrestricted sums if we put a so-called Kronecker-delta indicator function inside the sum \( \sum_d (\delta(d,A) \Phi(d)) \) where \( \delta(d,A) = 1 \) if datum \( d \) had \( Z = A \), and is 0 otherwise.
\[
\text{est}(\theta_a) = \frac{\sum_d \delta(d,A)}{D} = \frac{\sum_{d:Z=A} \#(d,h)}{\sum_{d:Z=A} 10} = \frac{48}{60} = \frac{4}{5} = 0.8
\]
\[
\text{est}(\theta_{h|a}) = \frac{\sum_d \delta(d,A) \#(d,h)}{\sum_d \delta(d,A) 10} = \frac{\sum_{d:Z=A} \#(d,h)}{\sum_{d:Z=A} 10} = \frac{48}{60} = \frac{4}{5} = 0.8
\]
hence

\[
\frac{\partial L(\theta_a)}{\partial \theta_a} = 0 \implies \theta_a = \frac{\sum_{d: Z = a} 1}{\sum_{d: Z = a} 1 + \sum_{d: Z = b} 1}
\]

\[
\frac{\partial L(\theta_{h|a})}{\partial \theta_{h|a}} = 0 \implies \theta_{h|a} = \frac{\sum_{d: Z = a} #(d, h)}{\sum_{d: Z = a} #(d, h) + \sum_{d: Z = b} #(d, t)}
\]

\[
\frac{\partial L(\theta_{h|b})}{\partial \theta_{h|b}} = 0 \implies \theta_{h|b} = \frac{\sum_{d: Z = b} #(d, h)}{\sum_{d: Z = b} #(d, h) + \sum_{d: Z = a} #(d, t)}
\]

finally the denominators of these turn into \(D, \sum_{d: Z = a} \cdot 10\) and \(\sum_{d: Z = b} \cdot 10\) respectively and so are exactly the 'common sense' formulae we started with in (2), (3), (4)