Frustra fit per plura quod potest fieri per pauciora.
It is vain to do with more what can be done with fewer.

-- William of Ockham
The story of Mr A and Mr B

- Mr A has a theory.
- Mr B has also a theory, but with an adjustable parameter $\lambda$.

Whose theory should we prefer on the basis of data $D$?
The story of Mr A and Mr B

• Mr A has a theory.
• Mr B has also a theory, but with an adjustable parameter $\lambda$.

Whose theory should we prefer on the basis of data $D$?

Posterior ratio = \[ \frac{\text{prob}(A|D, I)}{\text{prob}(B|D, I)} \]

\[
\begin{aligned}
\text{posterior ratio} &= \\
&= \begin{cases} 
>> 1 & \text{prefer Mr A’s theory} \\
<< 1 & \text{prefer Mr B’s theory}
\end{cases}
\end{aligned}
\]
The story of Mr A and Mr B

\[
\frac{\text{prob}(A|D, I)}{\text{prob}(B|D, I)} = \frac{\text{prob}(D|A, I)}{\text{prob}(D|B, I)} \times \frac{\text{prob}(A|I)}{\text{prob}(B|I)}
\]

- What do we think about the theories before we saw any data? (ratios of the priors!)
- How well do the two theories fit the data?
  - Simple for Mr A: no adjustable parameter
  - What about Mr B? He needs $\lambda$ to do predictions.
Mr B’s prediction

$$\text{prob}(D|B, I) = \int \text{prob}(D, \lambda|B, I) d\lambda$$

$$= \int \text{prob}(D|\lambda, B, I)\text{prob}(\lambda|B, I) d\lambda$$
Lindley’s paradox revisited

- Mr A thinks that $\theta = 0.5$
- Mr B thinks that $\theta \neq 0.5$
- We obtained the following data $D$:
  - 49,581 heads
  - 48,870 tails

Whose theory should we prefer on the basis of data $D$?
Lindley’s paradox revisited

\[
\text{prob}(D|A, I) = \binom{n}{k} (0.5)^k (1 - 0.5)^{n-k} \approx 1.95 \times 10^{-4}
\]

\[
\text{prob}(D|B, I) = \int_0^1 \binom{n}{k} u^k (1 - u)^{n-k} du
\]

\[
= \binom{n}{k} B(k + 1, n - k + 1) \approx 1.02 \times 10^{-5}
\]

To be fair, we take the ratio of the priors to be unity.

\[
\frac{\text{prob}(A|D, I)}{\text{prob}(B|D, I)} = \frac{\text{prob}(D|A, I)}{\text{prob}(D|B, I)} \times \frac{\text{prob}(A|I)}{\text{prob}(B|I)}
\]

\[
= 19.12
\]
> binom.test(49581, 49581 + 48870)

    Exact binomial test

data: 49581 and 49581 + 48870
number of successes = 49581, number of trials = 98451, p-value = 0.02365
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
  0.5004826 0.5067390
sample estimates:
  probability of success
    0.5036109
Lindley’s paradox revisited

- Based on the data, Mr A’s theory is rejected with a p-value of 0.02
- Nonetheless Mr A’s theory given the data is much more likely than Mr B’s theory
- Mr A is very specific $\theta = 0.5$.
- Mr B seems to be very unspecific, he allows every value for $\theta$ except 0.5
Mr B’s theory

• $\theta$ lies within $0$ and $1$

$$\text{prob}(\theta|B, I) = \frac{1}{\theta_{max} - \theta_{min}} = 1$$

• There is value $\theta_0$, which yields the closest agreement with the measurements
Back to Parameter Estimation

\[
P = \text{prob}(\theta|D, B, I) = \frac{\text{prob}(D|\theta, B, I) \times \text{prob}(\theta|B, I)}{\text{prob}(D|B, I)}
\]

The best estimate for \( \theta = \theta_0 \) is given by the conditions

\[
\left. \frac{dP}{d\theta} \right|_{\theta_0} = 0 \text{ and } \left. \frac{d^2P}{d\theta^2} \right|_{\theta_0} < 0
\]
Back to Parameter Estimation

\[
L = \ln P \\
= L(\theta_0) + \frac{1}{2} \left. \frac{d^2 L}{d\theta^2} \right|_{\theta_0} (\theta - \theta_0)^2 + \ldots
\]

\[
\text{prob}(\theta|D, B, I) \approx A \exp \left[ \frac{1}{2} \left. \frac{d^2 L}{d\theta^2} \right|_{\theta_0} (\theta - \theta_0)^2 \right]
\]

\[
\approx \frac{1}{\sigma \sqrt{2\pi}} \exp \left[ -\frac{1}{2} \frac{(\theta - \theta_0)^2}{\sigma^2} \right]
\]
Mr B’s best estimate

\[ \theta_0 = \frac{\# \text{ of heads}}{\# \text{ of heads} + \# \text{ of tails}} \approx 0.5036 \]

\[ \sigma = \sqrt{\frac{\theta_0(1 - \theta_0)}{\# \text{ of heads} + \# \text{ of tails}}} \approx 0.0016 \]

\[ \theta = 0.5036 \pm 0.0016 \]
Mr B’s likelihood function

\[
\text{prob}(D|B, I) = \int \text{prob}(D|\theta, B, I) \text{prob}(\theta|B, I) d\theta
\]

\[
\text{prob}(D|\theta, B, I) = \text{prob}(D|\theta_0, B, I) \times \exp \left[ -\frac{(\theta - \theta_0)^2}{2\sigma^2} \right]
\]

\[
\text{prob}(D|\theta_0, B, I) = \binom{n}{k} (\theta_0)^k (1 - \theta_0)^{n-k}
\]

\[
\text{prob}(D|B, I) = \frac{1}{\theta_{max} - \theta_{min}} \text{prob}(D|\theta_0, B, I) \int_{\theta_{min}}^{\theta_{max}} \exp \left[ -\frac{(\theta - \theta_0)^2}{2\sigma^2} \right] d\theta
\]
\[ P(D|\theta, B, I) \]
Back to the posterior ratio

\[
\frac{\text{prob}(A|D, I)}{\text{prob}(B|D, I)} = \frac{\text{prob}(A|I)}{\text{prob}(B|I)} \times \frac{\text{prob}(D|A, I)}{\text{prob}(D|\theta_0, B, I)} \times \frac{\theta_{\text{max}} - \theta_{\text{min}}}{\sigma \sqrt{2\pi}}
\]

= 19.21125
\[ \theta = 0.501 \]

\[ N = 0.501 \]
\[ \frac{1}{\sigma \sqrt{2\pi}} \]

\[ \text{prob}(D|\theta_0, B, I) \]

\[ \text{prob}(D|A, I) \]
Back to the posterior ratio

\[
\frac{\text{prob}(A|D, I)}{\text{prob}(B|D, I)} = \frac{\text{prob}(A|I)}{\text{prob}(B|I)} \times \frac{\text{prob}(D|A, I)}{\text{prob}(D|\theta_0, B, I)} \times \frac{\theta_{\text{max}} - \theta_{\text{min}}}{\sigma \sqrt{2\pi}}
\]

= 19.21125

Ockham Factor
Model selection is not parameter estimation

\[
\frac{\text{prob}(A|D, I)}{\text{prob}(B|D, I)} = \frac{\text{prob}(A|I)}{\text{prob}(B|I)} \times \frac{\text{prob}(D|A, I)}{\text{prob}(D|\theta_0, B, I)} \times \frac{\theta_{\text{max}} - \theta_{\text{min}}}{\sigma \sqrt{2\pi}}
\]

= 19.21125

- Mr A’s prob(D|A, I) = 1.95 \times 10^{-4}
- Mr B’s prob(D|\theta_0, B, I) = 2.54 \times 10^{-3}
  - Mr B’s theory gives a better fit to the data
  - But Mr B’s theory has a lot more uncertainty regarding the actual value of his additional parameter
Model selection is not parameter estimation

\[ P = \text{prob}(\theta \mid D, B, I) = \frac{\text{prob}(D \mid \theta, B, I) \times \text{prob}(\theta \mid B, I)}{\text{prob}(D \mid B, I)} \]

- Parameter estimation: find the maximal posterior estimate of a parameter
- Model selection: average the likelihood over the specified parameter range
Illustration

- Mr A has a parameter $\mu$
- Mr B has a parameter $\theta$

\[
\frac{\text{prob}(A|D, I)}{\text{prob}(B|D, I)} = \frac{\text{prob}(A|I)}{\text{prob}(B|I)} \times \frac{\text{prob}(D|\mu_0, A, I)}{\text{prob}(D|\theta_0, B, I)} \times \frac{\sigma_\mu(\theta_{max} - \theta_{min})}{\sigma_\theta(\mu_{max} - \mu_{min})}
\]
Model selection is Hypothesis testing

• Posterior probability that hypothesis $H_i$ is true

\[
\text{prob}(H_1|D, I) = \frac{\text{prob}(D|H_1, I) \times \text{prob}(H_1|I)}{\text{prob}(D|I)}
\]

\[
\text{prob}(H_2|D, I) = \frac{\text{prob}(D|H_2, I) \times \text{prob}(H_2|I)}{\text{prob}(D|I)}
\]

• Posterior ratio

\[
\frac{\text{prob}(H_1|D, I)}{\text{prob}(H_2|D, I)} = \frac{\text{prob}(D|H_1, I)}{\text{prob}(D|H_2, I)} \times \frac{\text{prob}(H_1|I)}{\text{prob}(H_2|I)}
\]
Model selection is Hypothesis testing

• Suppose the hypothesis $H_2$ is that $H_1$ is false

$$\text{prob}(H_1|D, I) = \frac{\text{prob}(D|H_1, I) \times \text{prob}(H_1|I)}{\text{prob}(D|I)}$$

$$\text{prob}(H_2|D, I) = \frac{\text{prob}(D|H_2, I) \times \text{prob}(H_2|I)}{\text{prob}(D|I)}$$

— How do we calculate the likelihood for $H_2$? We need a model of $H_2$!

— This leads naturally to model selection!