Data Analytics
EEE 4774 & 6777

Module 2

Model Selection

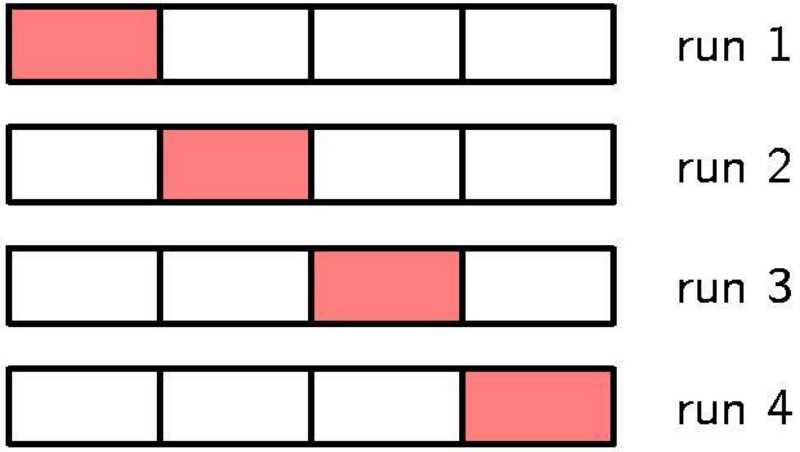
Spring 2022

Frequentist

Model Selection

K-fold Cross-Validation

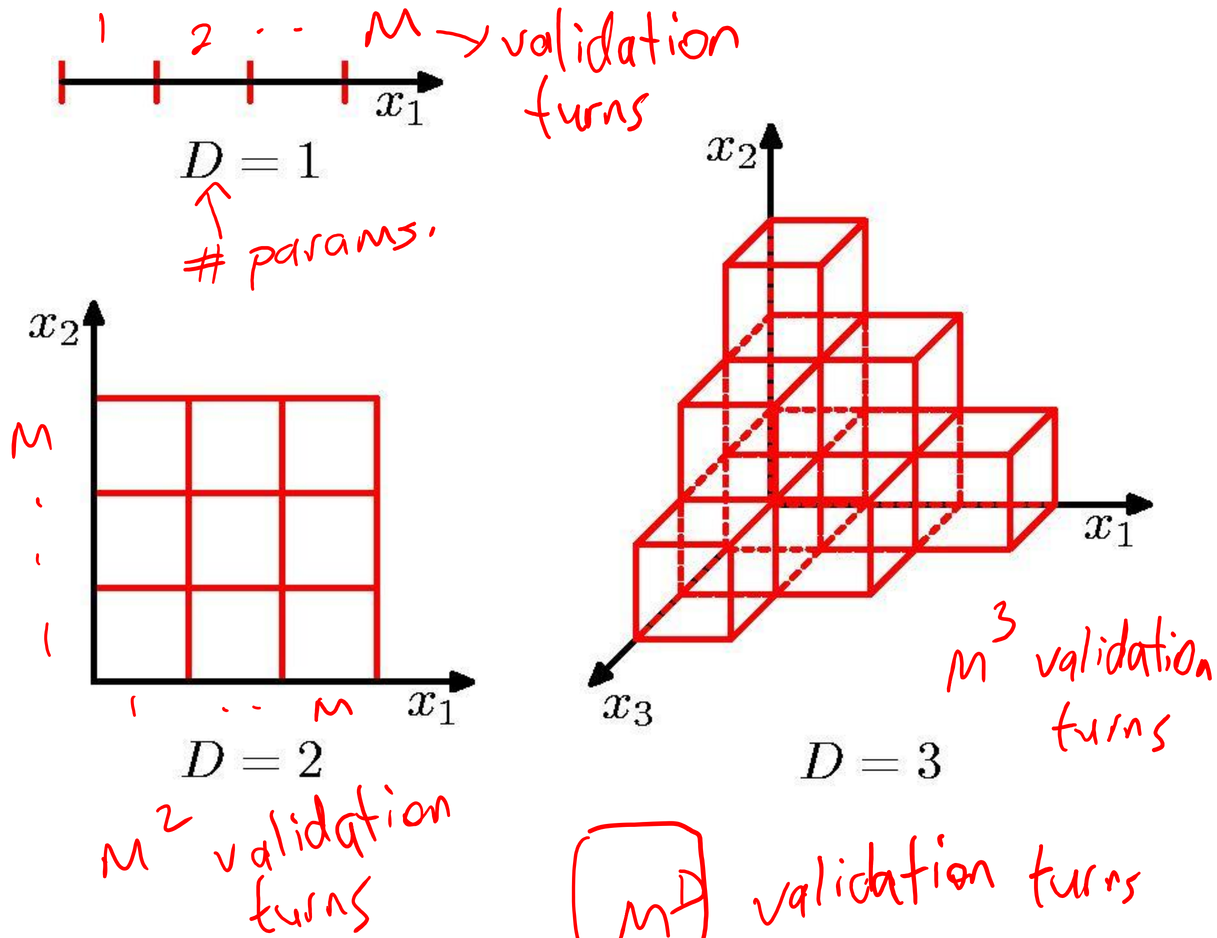
- ***Cross-Validation***



- High computational complexity
  - even a single training may be expensive
  - combinations of multiple complexity parameters to tune
- Need a better approach for complex problems!

- ***Curse of Dimensionality***

KM model training



$\boxed{M^D}$ validation turns

Model Selection

- ***Alternative approaches***
  - Akaike Information Criterion (AIC)

    likelihood

    penalty term

    $$2k - 2\log p(D|w_{ML})$$

    k: # estimated parameters in the model

    data

    ML estimate

  - Bayesian Information Criterion (BIC)

    $$\underline{k\log N} - 2\log p(D|w_{ML})$$

    N: # data points in D

  - Occam's razor
    - if several models are compatible with the observations, pick the ***simplest*** one

  - <u>Bayesian</u>

    model likelihood

    model prior

    $$p(M_i|D) \propto \underbrace{p(D|M_i)}_{\text{evidence}}\ \underbrace{p(M_i)}_{\text{prior}}$$

    model posterior

    $$\boxed{p(D|M_i) = \int p(D|w, M_i)\, p(w|M_i)\, dw}$$

    data

    data likelihood for given param. values w of model M_i

    parameter prior under model M_i

Comparing Models the Bayesian Way

Given an indexed set of models $M_1, \ldots, M_m$, and associated prior beliefs in the appropriateness of each model $p(M_i)$, our interest is the model posterior probability

$$\max_{M_i} \quad p(M_i|\mathcal{D}) = \frac{p(\mathcal{D}|M_i)p(M_i)}{p(\mathcal{D})}$$

$$\max_{M_i} p(M_i|D) \equiv \max_{M_i} p(D|M_i)\, p(M_i)$$

where the likelihood of the data $\mathcal{D}$ is

$$p(\mathcal{D}) = \sum_{i=1}^{m} p(\mathcal{D}|M_i)p(M_i)$$

Model M_i is parameterized by θ_i, and the model likelihood, i.e., model ***evidence***, is given by

set of parameters of model M_i

$$\boxed{p(\mathcal{D}|M_i) = \int p(\mathcal{D}|\theta_i, M_i)p(\theta_i|M_i)d\theta_i}$$

In discrete parameter spaces, the integral is replaced with summation. Note that the number of parameters $\dim(\theta_i)$ need not be the same for each model.

Bayes Factor

Comparing two competing model hypotheses M_i and M_j is straightforward and only requires the Bayes Factor:

$$\frac{p(M_i|\mathcal{D})}{p(M_j|\mathcal{D})} = \underbrace{\frac{p(\mathcal{D}|M_i)}{p(\mathcal{D}|M_j)}}_{\text{Bayes' Factor}} \frac{p(M_i)}{p(M_j)}$$

> 1 choose M_i

< 1 " M_j

which does not require integration/summation over all possible models.

Caveat

$p(M_i|\mathcal{D})$ only refers to the probability relative to the set of models specified $M_1, \ldots, M_m$. This is not the *absolute* probability that model M fits 'well'.

Example: Fair or Biased coin?

$p(x = \text{heads} \mid \theta = 0.5) = 0.5$

$\theta \in (0, 1)$: prob. of heads

Two models:

$$M_{fair} : \text{The coin is fair}, \quad M_{biased} : \text{The coin is biased}$$

probability param.

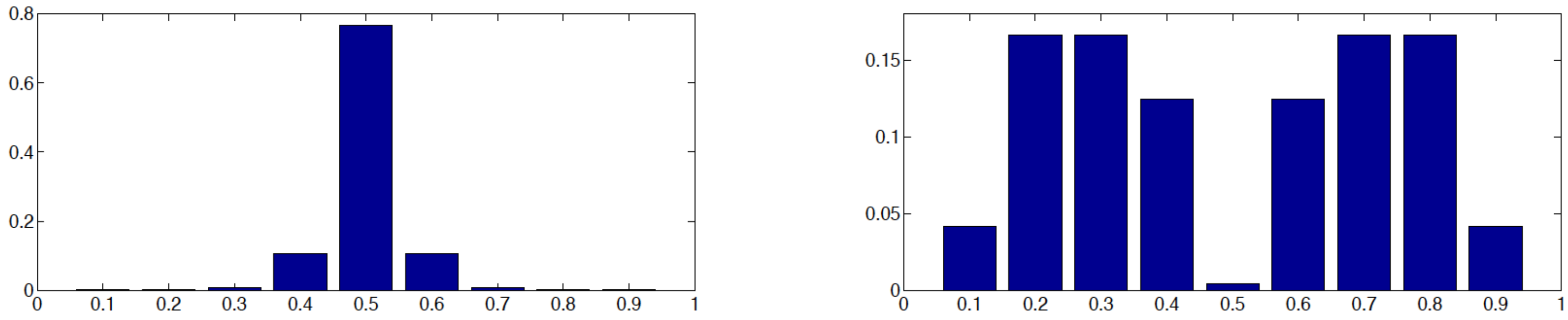
For simplicity we assume $\mathrm{dom}(\theta) = \{0.1, 0.2, \ldots, 0.9\}$.

$p(\theta|M)$

Fair Model

Biased Model

$p(\theta|m)$



θ (a) θ (b)

Figure: **(a)**: Discrete prior model of a 'fair' coin $p(\theta|M_{fair})$. **(b)**: Prior for a biased 'unfair' coin $p(\theta|M_{biased})$. In both cases we are making explicit choices here about what we consider to be a 'fair' and 'unfair'.

Example: Fair or Biased coin?

The model likelihood

For each model M, the likelihood is given by

\# heads   \# tails

model evidence

$$p(\mathcal{D}|M) = \sum_{\theta} \underbrace{p(\mathcal{D}|\theta, M)}_{\text{Data likelihood given par. } \theta \text{ value}} \overbrace{p(\theta|M)}^{\text{prior from model}} = \sum_{\theta} \underbrace{\theta^{N_H} (1-\theta)^{N_T}}_{\text{Binomial likelihood}} p(\theta|M)$$

prob. of tails

This gives

$$\underbrace{0.1^{N_H} (1-0.1)^{N_T} p(\theta = 0.1|M)}_{\theta = 0.1} + \ldots + \underbrace{0.9^{N_H} (1-0.9)^{N_T} p(\theta = 0.9|M)}_{\theta = 0.9}$$

Bayes factor

Assuming that $p(M_{fair}) = p(M_{biased})$ the posterior ratio is given by the Bayes' factor, i.e., the ratio of the two model likelihoods (evidences).

$$\frac{p(M_{fair}|\mathcal{D})}{p(M_{\text{biased}}|\text{D})} = \frac{p(\mathcal{D}|M_{fair})}{p(\mathcal{D}|M_{\text{biased}})}$$

Example: Fair or Biased coin?

Dataset 1 : 7 trials

5 Heads and 2 Tails

Here $p(\mathcal{D}|M_{fair}) = 0.00786$ and $p(\mathcal{D}|M_{biased}) = 0.0072$. The Bayes' factor is

$$\frac{p(M_{fair}|\mathcal{D})}{p(M_{biased}|\mathcal{D})} = 1.09$$

9% preferance (Bayes Factor (Fair/Biased) = 1.09) for the Fair Model. Given # of Heads (5) and # of Tails (2)
=> Biased model should be peferred but that's not the case. Why?

- # of experiments is a factor. 7 trials is not enough to make a strong conclusion.
- Frequentist approach would have overfitted.

indicating that there is little to choose between the two models.

Dataset2 : 70 trials

50 Heads and 20 Tails

Here $p(\mathcal{D}|M_{fair}) = 1.5 \times 10^{-20}$ and $p(\mathcal{D}|M_{biased}) = 1.4 \times 10^{-19}$. The Bayes' factor is

$$\frac{p(M_{fair}|\mathcal{D})}{p(M_{biased}|\mathcal{D})} = 0.109$$

Biased model can be chosen with high confidence

indicating that have around 10 times the belief in the biased model as opposed to the fair model.

'Automatic' Complexity penalization

Problem

You are told that the total score given from an unknown number of die is 9. What is the distribution of the number of die?

Model posterior

From Bayes' rule, we need to compute the posterior distribution over models

total number (data)

$$\underset{\text{posterior}}{p(n|t)} = \frac{p(t|n)p(n)}{p(t)}$$

dies

Assume $p(n)$ = const.

Likelihood

$$\underset{\text{Model evidence}}{p(t|n)} = \sum_{s_1,\dots,s_n} p(t, s_1, \dots, s_n|n) = \sum_{s_1,\dots,s_n} p(t|s_1, \dots, s_n) \prod_i p(s_i)$$

$$= \sum_{s_1,\dots,s_n} \mathbb{I}\left[t = \sum_{i=1}^{n} s_i\right] \prod_i p(s_i)$$

indicator func.

1 if inside parantheses true
0 o.w.

where $p(s_i) = 1/6$ for all scores s_i.

‘Automatic’ Complexity penalization

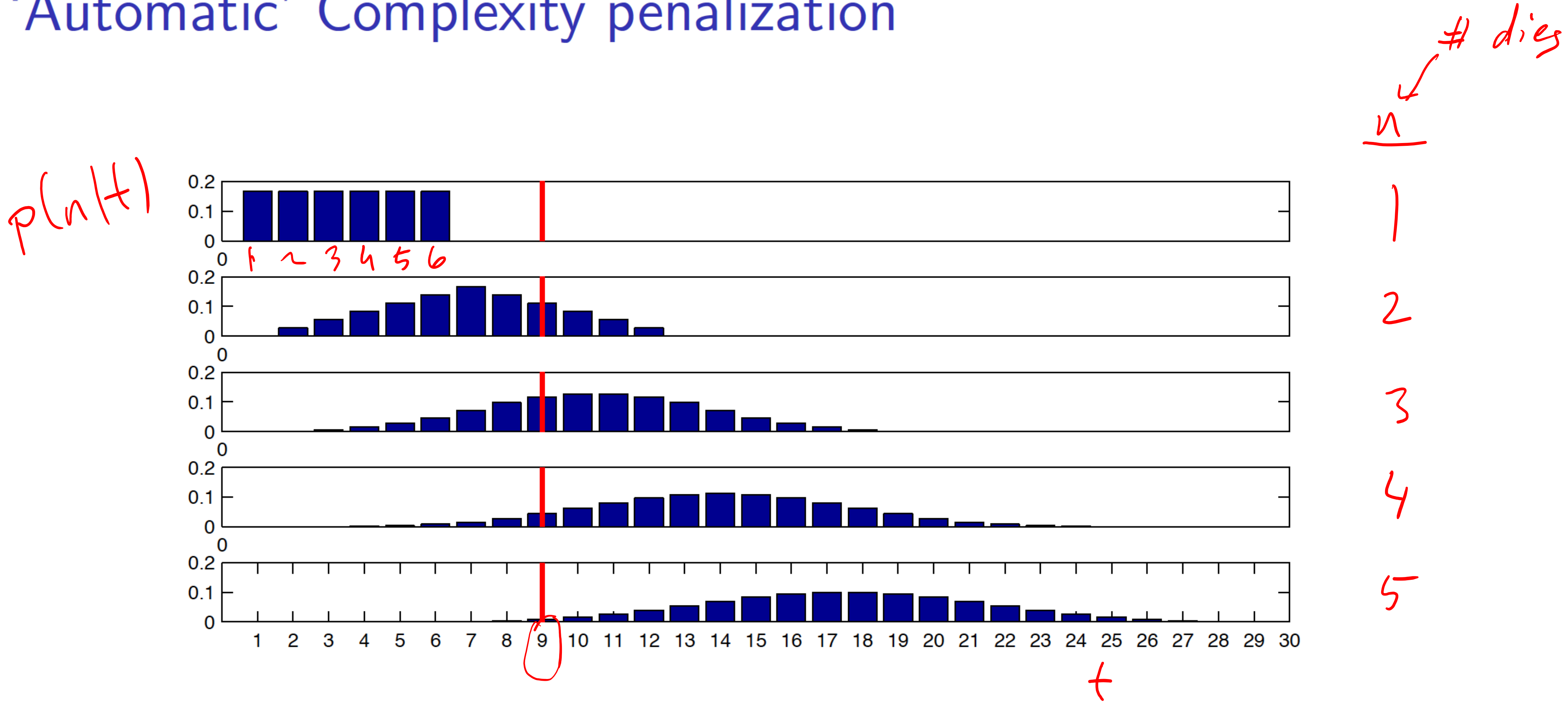
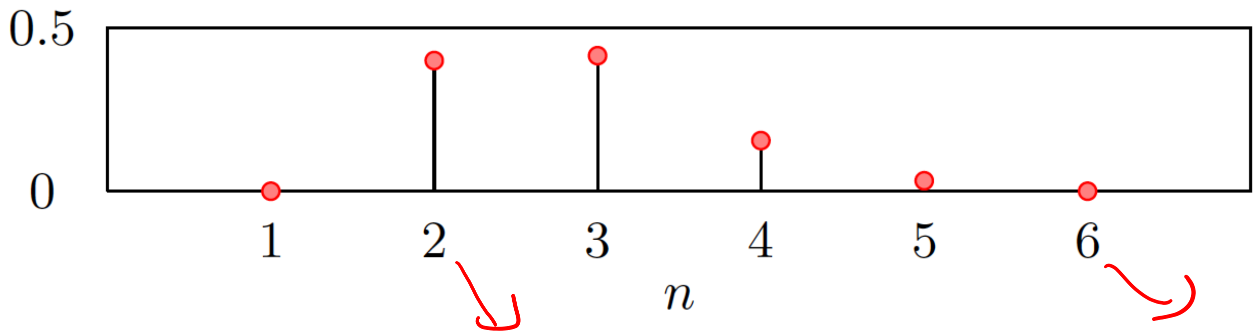


Figure: The likelihood of the total dice score, $p(t|n)$ for $n = 1$ (top) to $n = 5$ (bottom) die. Plotted along the horizontal axis is the total score t. The vertical line marks the comparison for $p(t = 9|n)$ for the different number of die. The more complex models, which can reach more states, have lower likelihood, due to normalization over t.

‘Automatic’ Complexity penalization

The posterior $p(n|t = 9)$



$2 \leq t \leq 12$

prob. distributed over 11 possibilities

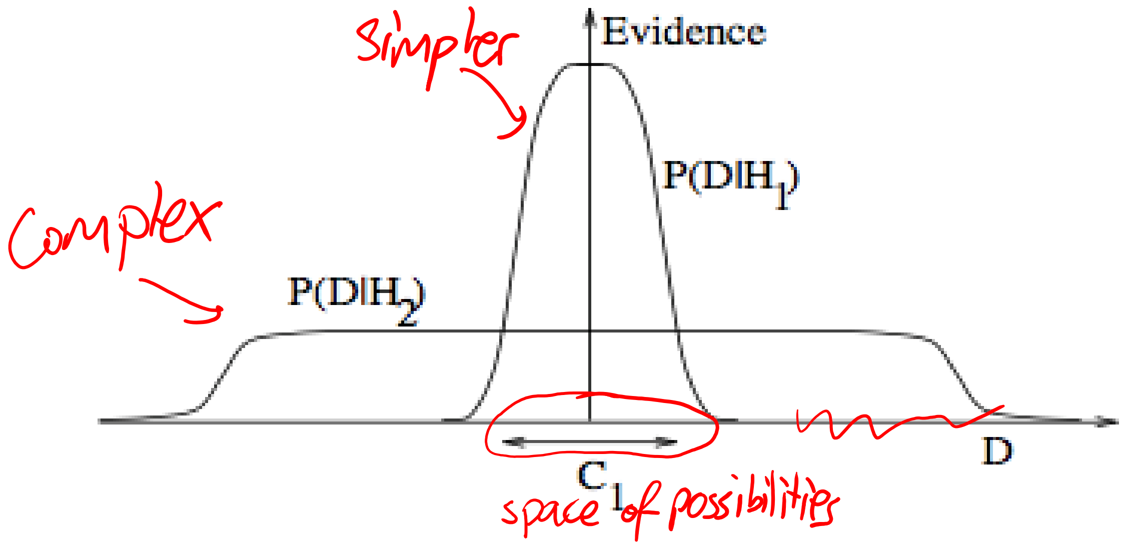
$6 \leq t \leq 36$

prob. distr. ove 31 possib.

Occam’s razor

- A posteriori, there are only 3 plausible models, namely $n = 2, 3, 4$ since the rest are either too complex, or impossible.
- As the models become more ‘complex’ (n increases), more states become accessible and the probability mass typically reduces.
- This demonstrates the Occam’s razor effect which penalizes models which are over complex.

Bayesian approach includes Occam's razor



- A simple model H_1 makes only a limited range of predictions
- A more powerful model H_2 (e.g., with more free parameters than) predicts a greater variety of datasets
- H_2 does not predict the data sets in region C_1 as strongly as H_1
- With equal priors, if the dataset falls in region C_1, the less powerful model H_1 will be the more probable