2

Probability assignment

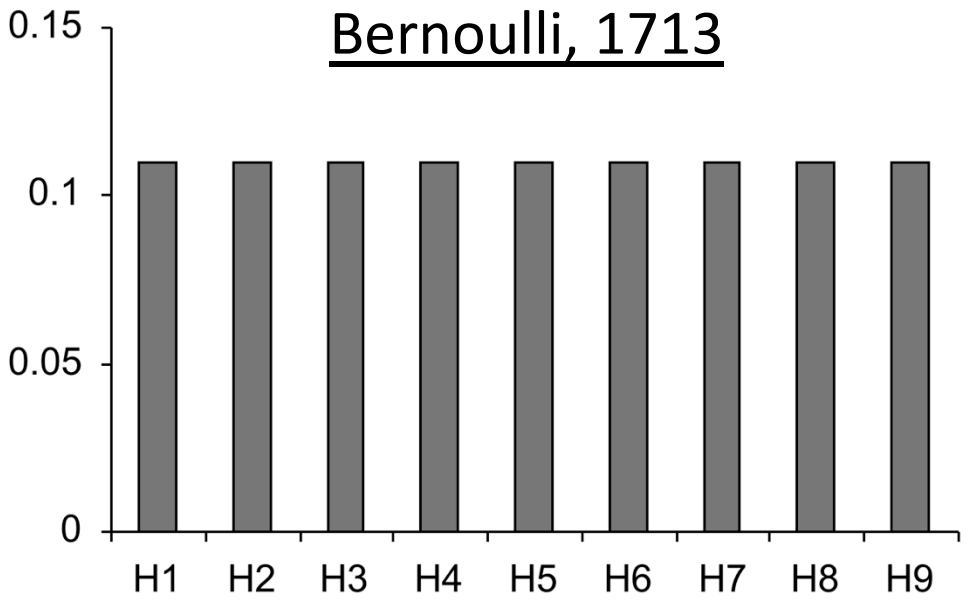
We have seen the basics of probability calculus …

… but what about probability assignment?

Principle of Insufficient Reason

- If we have a set of N exhaustive and mutually exclusive hypotheses H_i and there is no reason to prefer any of them then we must assign

$$P(H_i \mid I) = 1/N$$



- This is the principle we use to assign $P(\text{head} \mid I) = 0.5$ in unbiased coin flipping or $P(3 \mid I) = 1/6$ in dice throwing

Toy Model: Bernoulli's Urn

- An urn contains N balls, with R red and $N-R$ white balls
- What is the probability to get a red ball at the first draw?
- Everybody knows the answer: $P(R_1 \mid I) = R/N$ and this is usually taken for granted
- But where does this assignment actually come from?
- To see this, we will derive this result from scratch using nothing else but the principle of insufficient reason and the rules of probability calculus

But buckle-up: at the end of the road we will make some observations which you may find quite remarkable …

- Label each ball and define the exhaustive and exclusive set of hypotheses

$$H_i := \text{‘this ball has label } i\text{’}$$

- Use the principle of insufficient reason to assign

$$P(H_i|N, I) = \frac{1}{N}$$

- Also define the exhaustive and exclusive set

$$\begin{aligned} H_R &:= \text{‘this ball is red’} \\ H_W &:= \text{‘this ball is white’} \end{aligned}$$

- Now expand in the set H_i

$$P(H_R|I) \quad \langle H_R|I\rangle = \sum \langle H_R|H_i\rangle\langle H_i|I\rangle$$

$$= \frac{1}{N}\sum_{i=1}^{N} P(H_R|H_i, I) = \frac{R}{N}$$

- In the last step we made the trivial assignment

$$P(H_R|H_i, I) = \begin{cases} 1 & \text{if ball '}i\text{' is red} \\ 0 & \text{otherwise.} \end{cases}$$

This simple example clearly shows expansion as a powerful tool in probability assignment

What is the probability that the second ball is red?

- Drawing with replacement (contents do not change)

$$P(R_2|I) = \frac{R}{N}$$

- Drawing without replacement (contents do change)

$$P(R_2|R_1, I) = \frac{R-1}{N-1}$$

$$P(R_2|W_1, I) = \frac{R}{N-1}$$

- Here we know the outcome of the first draw, but what happens when we do not know this outcome?

- If we draw without replacement and do not know the outcome of the first draw then the probability that the second draw is red is found by expansion in the set $\{R_1, W_1\}$

$$\langle R_2|I\rangle = \langle R_2|R_1\rangle\langle R_1|I\rangle + \langle R_2|W_1\rangle\langle W_1|I\rangle$$

$$= \frac{R-1}{N-1}\frac{R}{N} + \frac{R}{N-1}\frac{W}{N} = \frac{R}{N}$$

⇒ This does not depend on the outcome of the first draw and thus not on the contents of the urn which has changed after the first draw!

- We have seen that the outcome of the first draw can influence the probability of the outcome of the second draw, in case we draw without replacement and know the color of the first ball
- The first event influences the second which is kinda natural, right?
- But now we will show that the outcome of the second draw can influence the probability of the first draw!

- The probability that the first ball is red is *R/N*
- Blindly draw the first ball, without recording its color
- Draw he second ball and it is red
- The probability that the first ball is red is now not *R/N* as Bayes' theorem shows

$$P(R_1|R_2) = \frac{P(R_2|R_1)\,P(R_1)}{P(R_2|R_1)\,P(R_1) + P(R_2|W_1)\,P(W_1)} = \frac{R-1}{N-1}$$

- The color of second ball can influence the probability of the color of the first ball!
- This is of course not a causal but a logical relationship

Here is a simple example to make it clear

- An urn contains one red and one white ball
- The probability of a red ball at the first draw is ½
- Lay the ball aside without knowing its color
- Now draw the second ball
- If the second ball is red, then the probability that the first ball is red is 0 and not ½

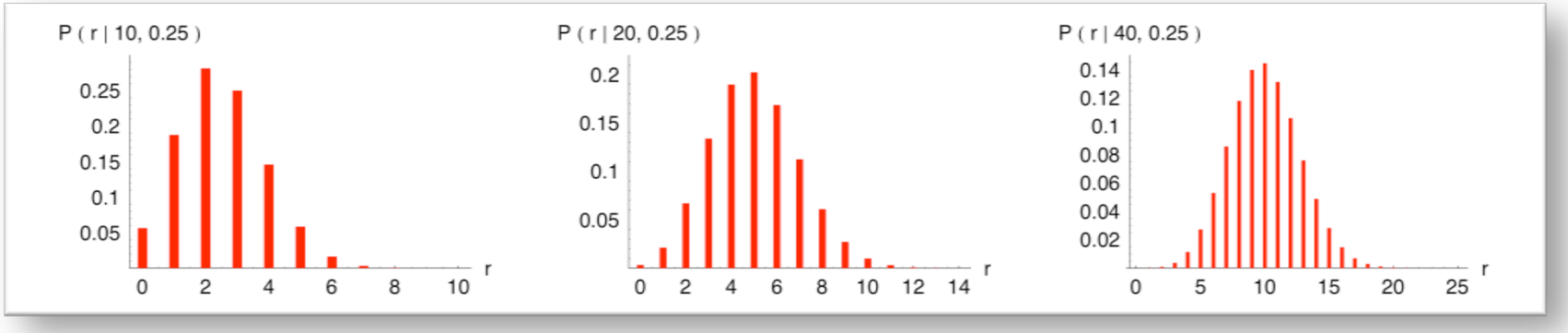
Binomial distribution

- Let *h* be the probability that a coin flip gives heads
- The probability to observe *n* heads in *N* flips is the binomial distribution (see writeup for derivation)

$$P(n|N,h) = \frac{N!}{n!(N-n)!} h^n (1-h)^{N-n}$$

- Applies to cases where the outcome is binary like yes/no, head/tail, good/bad … and where the probability *h* is the same for all trials

$$< n >= Nh \qquad < \Delta n^2 >= Nh(1-h)$$



$$\frac{< n >}{N} = h \pm \sqrt{\frac{h(h-1)}{N}} \rightarrow h \text{ for } N \rightarrow \infty$$

The fact that *n/N* converges to *h* is called the law of large numbers and provides the link between a probability (*h*) and a frequency (Bernoulli 1713)

The binomial posterior

- Bayes' theorem gives for the posterior of *h*

$$p(h,|n,N)\,\mathrm{d}h = C\,P(n|h,N)\,p(h|I)\,\mathrm{d}h$$

- A uniform prior $p(h|I) = 1$ yields upon calculation of the normalization constant *C*

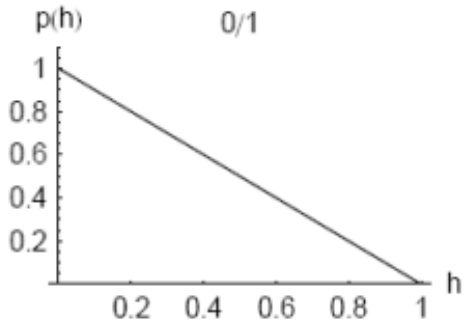
$$p(h,|n,N)\,\mathrm{d}h = \frac{(N+1)!}{n!(N-n)!}\,h^n(1-h)^{(N-n)}\,\mathrm{d}h$$

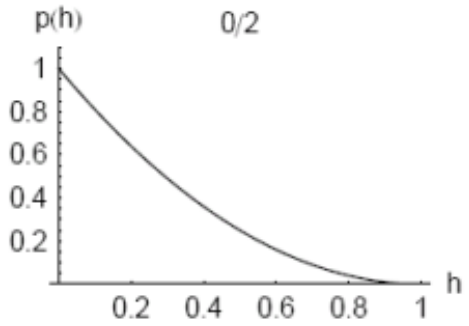
- Looks like the Binomial but isn't since it is a function of *h* not *n*
- The mode and width (inverse of the Hessian) are

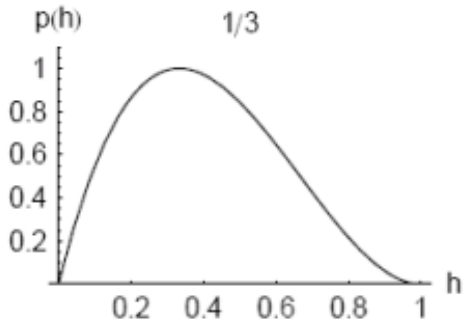
$$\hat{h} = \frac{n}{N} \pm \sqrt{\frac{\hat{h}(1-\hat{h})}{N}}$$

Posterior for the first three flips of a coin

NB: the distributions are scaled to unit maximum for ease of comparison







No head in one throw

The data exclude the possibility that $h = 1$

No head in two throws

The data favor lower values of h

One head in three throws

The data favor lower values of h and exclude that $h = 0$

Error on a 100% efficiency measurement

- One may ask the question how to calculate the error on an efficiency measurement where a counter fires *N* times in *N* events. Binomial statistics tell you that the error is zero

$$\varepsilon = \frac{N}{N} = 1 \qquad \Delta\varepsilon = \sqrt{\frac{\varepsilon(1-\varepsilon)}{N}} = 0$$

- Here is the Bayesian answer: assuming a uniform prior, the (normalised) posterior of ε for *n* hits in *N* events is

$$p(\varepsilon|n,N) = \frac{(N+1)!}{n!(N-n)!}\,\varepsilon^n(1-\varepsilon)^{(N-n)} \quad \Rightarrow$$

$$p(\varepsilon|N,N) = (N+1)\,\varepsilon^N$$

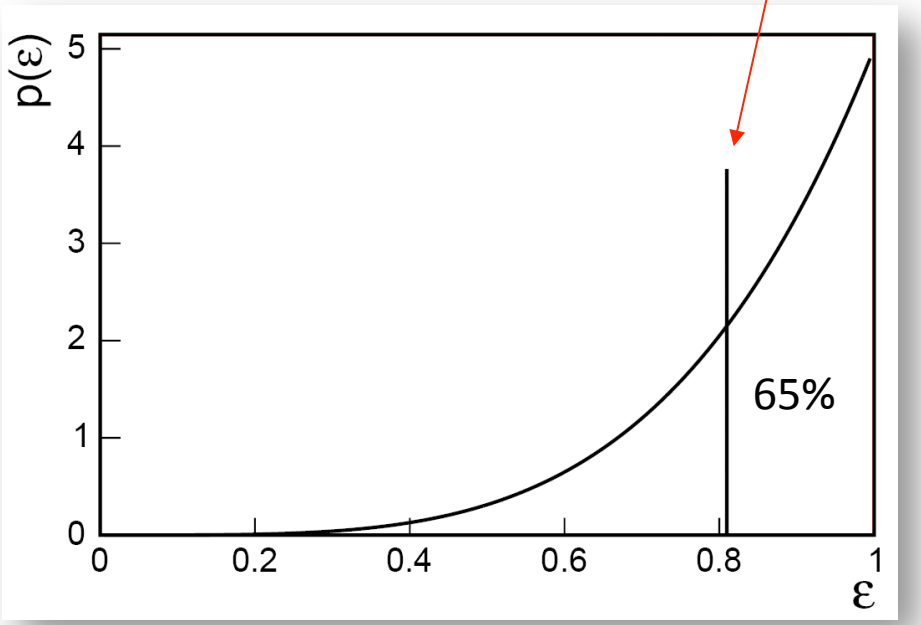
- Integrating the posterior gives the α confidence limit

$$1-\alpha = \int_0^a p(\varepsilon|N,N)\,\mathrm{d}\varepsilon = a^{(N+1)} \quad \rightarrow \quad a = (1-\alpha)^{1/(N+1)}$$

- If $N = 4$ and we choose $\alpha = 0.65$ then we find $a = 0.81$

- The 65% CL result is thus, for $N = 4$:

$$\varepsilon = 1\ ^{+0}_{-0.19}$$



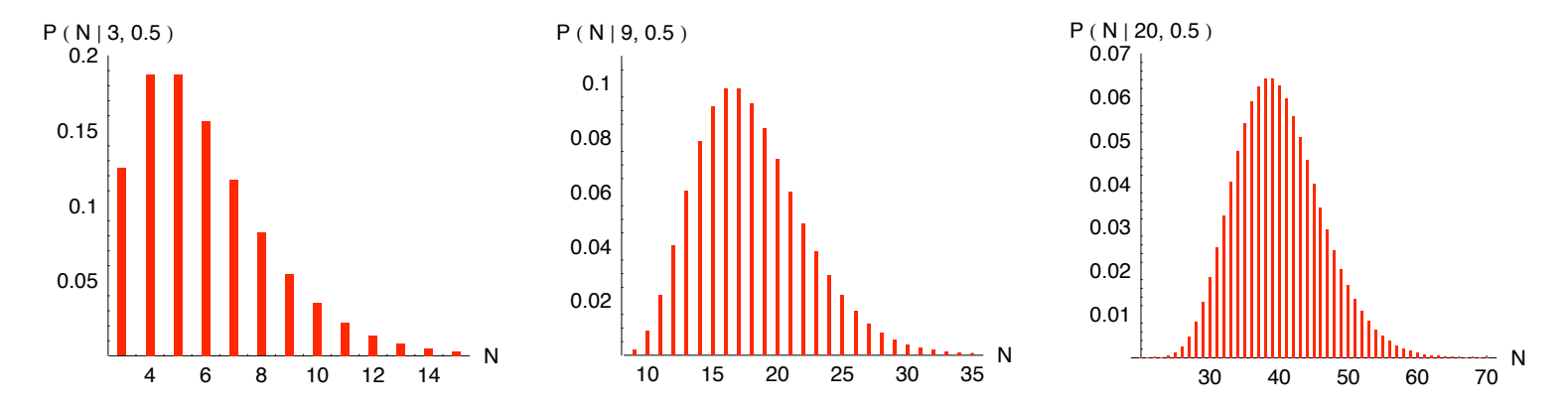
Negative binomial

- Draw *N* times ⇨ the number of heads (*n*) is the random variable with binomial distribution
- Wait for *n* heads ⇨ the number of draws (*N*) is the random variable with negative binomial distribution

$$
\begin{aligned}
P(N|n,h) &= \mathrm{Binom}(n-1|N-1,h) \times \mathrm{Binom}(1|1,h) \\
&= \frac{(N-1)!}{(n-1)!(N-n)!}\, h^N\, (1-h)^{N-n}
\end{aligned}
$$

- The likelihood for *n* heads in *N* trials thus depends on the stopping strategy

$$< N >= nq \qquad < \Delta N^2 >= nq(q-1) \qquad q \equiv 1/h$$



$$\frac{< N >}{n} = q \pm \sqrt{\frac{q(q-1)}{n}} \quad \rightarrow \quad q = \frac{1}{h} \text{ for } n \rightarrow \infty$$

Thus *N/n* converges to *1/h* but the reciprocal *n/N* does not converge to *h* (see next slide)

Stopping problem: Frequentist

- Observe *n* heads in *N* flips of a coin with bias *h*
- Define statistic *R* = *n*/*N* as an estimate for *h*
- If we stop at *N* flips then <*R*> converges to *h* because *n* follows a binomial distribution
- If we stop at *n* heads then <*R*> does not converge to *h* because *N* follows a negative binomial distribution

$$\frac{n}{\langle N \rangle} = h \quad \text{but} \quad \langle R \rangle = \langle \frac{n}{N} \rangle = n \langle \frac{1}{N} \rangle \neq \frac{n}{\langle N \rangle}$$

- The correct statistic would be *Q* = *N*/*n* that converges to 1/*h*

Given *n* heads in *N* flips but no stopping rule,
the Frequentist cannot analyse these data!

Stopping problem: Bayesian

- Observe *n* heads in *N* flips of a coin with bias *h*
- If we stop at *N* flips then the posterior of *h* becomes

$$p(h|n,N) = C\ \mathrm{Binom}(n|N,h)\ p(h|I) = C\ h^n(1-h)^{N-n}\ p(h|I)$$

- If we stop at *n* heads then the posterior of *h* becomes

$$p'(h|n,N) = C'\ \mathrm{Negbin}(N|n,h)\ p(h|I) = C'\ h^n(1-h)^{N-n}\ p(h|I)$$

- Normalisation gives *C = C'* and thus *p = p'* which means that Bayesian inference is not sensitive to the stopping rule!

This shows that Bayesian inference automatically discards irrelevant information in accordance with Cox' desideratum 3[a] which states that only relevant information should play a role

Other standard distributions are discussed in the write-up and are not presented here

- Multinomial distribution
- Poisson distribution
- Gauss distribution from the central limit theorem

Up to now, assignments are based on the modeling of random processes which unambiguously determines the probability distribution

4.5 Multinomial distribution

A generalisation of the binomial distribution is the **multinomial distribution** which applies to N independent trials where the outcome of each trial is among a set of k alternatives with probability p_i. Examples are drawing from an urn containing balls with k different colours, the throwing of a dice ($k = 6$) or distributing N independent events over the bins of a histogram.

The multinomial distribution can be written as

$$P(\boldsymbol{n}|\boldsymbol{p}, N) = \frac{N!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k} \tag{4.17}$$

where $\boldsymbol{n} = (n_1, \ldots, n_k)$ and $\boldsymbol{p} = (p_1, \ldots, p_k)$ are vectors subject to the constraints

$$\sum_{i=1}^{k} n_i = N \qquad \text{and} \qquad \sum_{i=1}^{k} p_i = 1. \tag{4.18}$$

The multinomial probabilities are just the terms of the expansion

$$(p_1 + \cdots + p_k)^N$$

from which the normalisation of $P(\boldsymbol{n}|\boldsymbol{p}, N)$ immediately follows. The average, variance and covariance are given by

$$\begin{aligned} <n_i> &= Np_i \\ <\Delta n_i^2> &= Np_i(1 - p_i) \\ <\Delta n_i \Delta n_j> &= -Np_ip_j \quad \text{for} \quad i \neq j. \end{aligned} \tag{4.19}$$

Marginalisation is achieved by adding in (4.17) two or more variables n_i and their corresponding probabilities p_i.

Exercise 4.5: Use the addition rule above to show that the marginal distribution of each n_i in (4.17) is given by the binomial distribution $P(n_i|p_i, N)$ as defined in (4.5).

The conditional distribution on, say, the count n_k is given by

$$P(\boldsymbol{m}|n_k, \boldsymbol{q}, M) = \frac{M!}{n_1! \cdots n_{k-1}!} q_1^{n_1} \cdots q_{k-1}^{n_{k-1}}$$

where

$$\boldsymbol{m} = (n_1, \ldots, n_{k-1}), \quad \boldsymbol{q} = \frac{1}{s}(p_i, \ldots, p_{k-1}), \quad s = \sum_{i=1}^{k-1} p_i \quad \text{and} \quad M = N - n_k.$$

Exercise 4.6: Derive the expression for the conditional probability by dividing the joint probability (4.17) by the marginal (binomial) probability $P(n_k|p_k, N)$.

30

- Until now we have not paid much attention to the role of the prior in Bayesian inference

$$p(y|x,I) = \frac{p(x|y,I)\ p(y|I)}{\int p(x|y,I)\ p(y|I)\mathrm{d}y}$$

- To get an idea, lets flip some coins

Is this coin biased?

- Denote bias by $0 < h < 1$
  - $h = 0.0$: two tails
  - $h = 0.5$: unbiased coin
  - $h = 1.0$: two heads
- The likelihood for n heads in N throws is

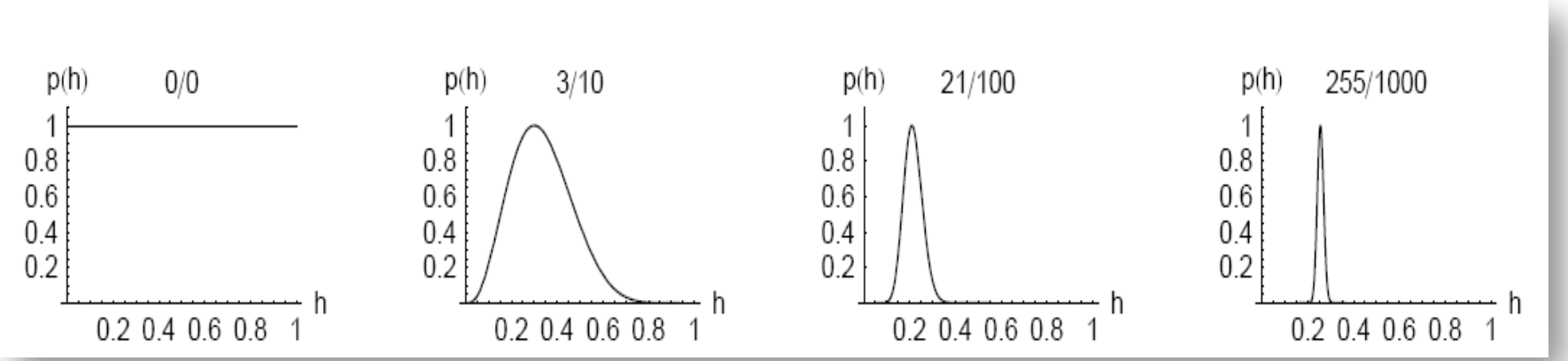
$$P(n|h,N) = \frac{N!}{n!(N-n)!} h^n (1-h)^{N-n}$$

- The posterior is

$$p(h|n,N)\mathrm{d}h = CP(n|h,N)\,p(h|I)\,\mathrm{d}h$$

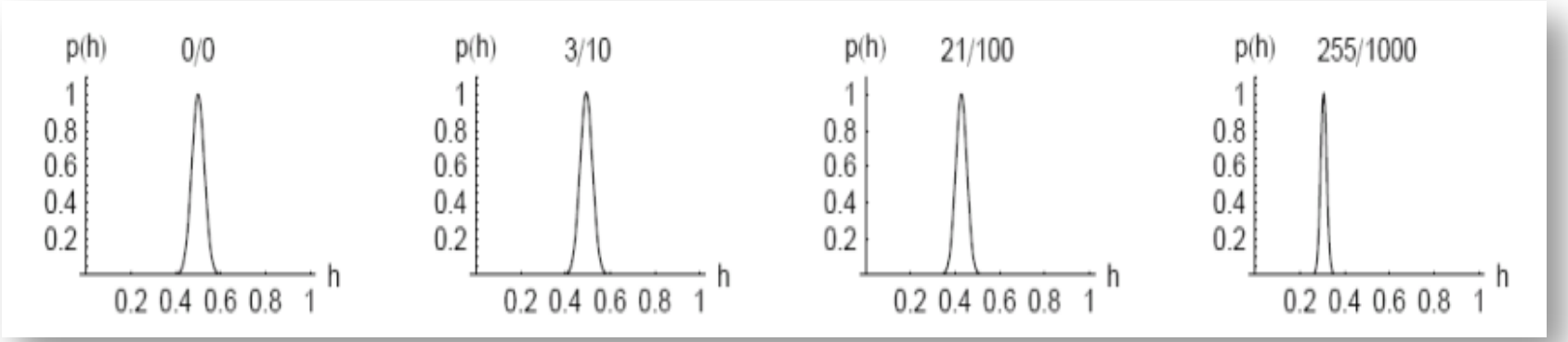
- How does the posterior behave for different priors?

1. Flat prior



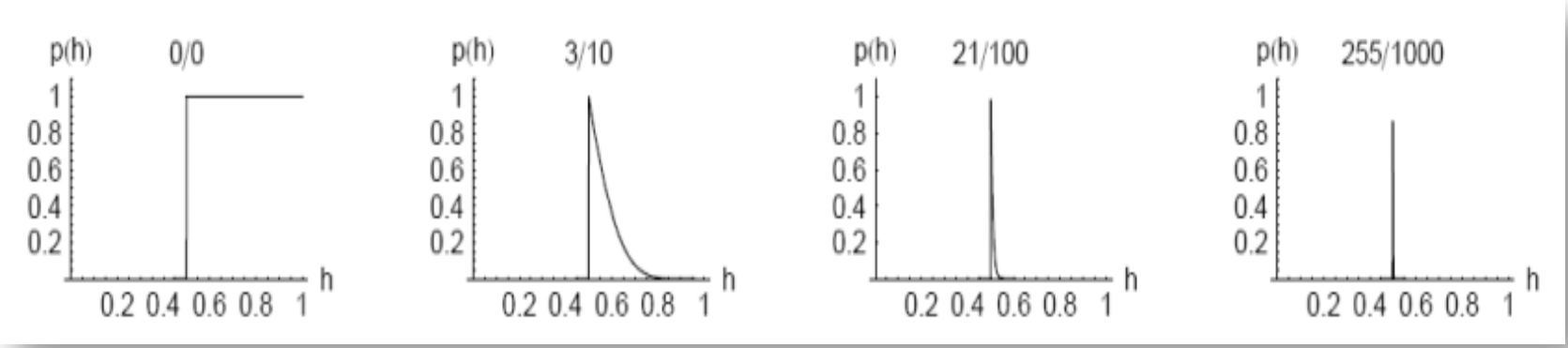
Posterior converges to $h = 0.25$ when the number of throws increases

2. Strong prior preference for a fair coin



Posterior converges to $h = 0.25$ but slower than for a flat prior ⇨ it takes a lot of evidence to change a strong prior belief

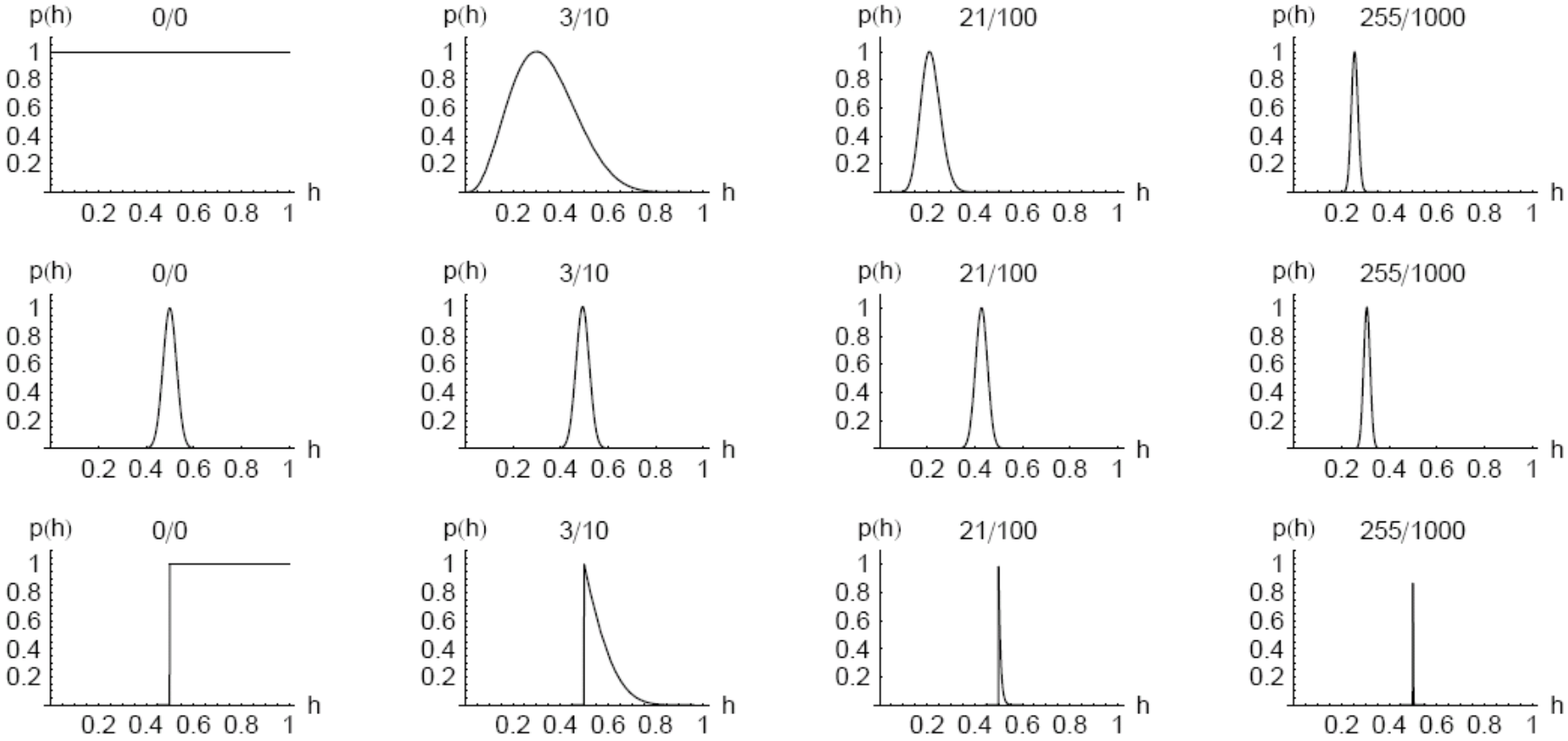
3. Exclude the possibility that $h < 0.5$



Posterior cannot go below $h = 0.5$ ⇨ no amount of data can change a prior certainty

Posterior for different priors

This coin is definitely biased with $h = ¼$: do you know how to make such a coin?



Remark: all posteriors are scaled to unit height for ease of comparison

What did we learn from this coin flipping experiment?

1. Conclusion depends on prior information, which always enters into inference (like axioms in mathematics)

   For instance, we investigate the coin beforehand and conclude that it is unbiased. Then 255 heads in 1000 throws tells us

   - that we have witnessed a rare event
   - or that something went wrong with the counting
   - or that the coin has been exchanged
   - or that some mechanism controls the throws
   - …

   but not that the coin is biased!

2. Unsupported information should not enter into the prior because it may need a lot of data to converge to the correct result in case this prior turns out to be wrong

3. No amount of data can ever change a prior certainty

When are priors important?

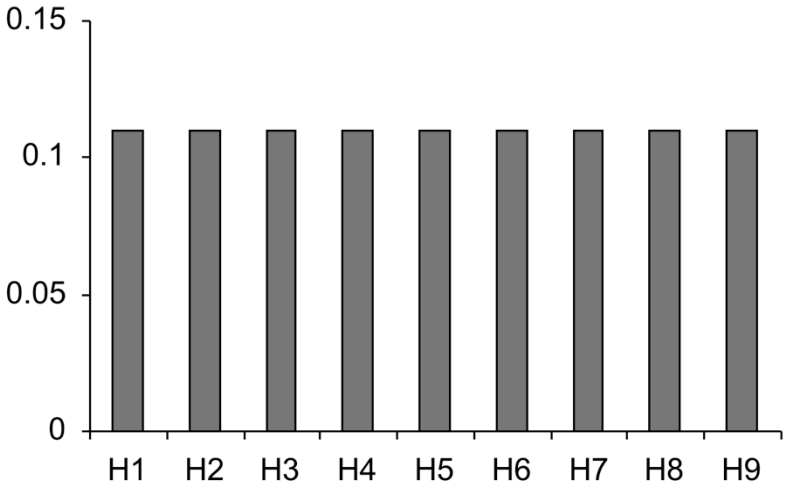
- If the prior is very wide compared to the likelihood then obviously it doesn't matter what prior we take
- If the width of the likelihood is comparable to that of any reasonable choice of prior then it does matter
- In this case the experiment does not carry much information so that answers become dominated by prior assumptions (perhaps you should, if possible, look for better data!)
- The likelihood peaks near a physical boundary or resides in an unphysical region; the information on the boundary is then contained in the prior and not in the data (neutrino mass measurements are a famous example)

Least informative priors

- Many assignments are based on expansion into a set of elementary probabilities, which are perhaps again expanded, until one hits assignment by the principle of insufficient reason
- Extension of this principle to continuous variables implies that one takes a uniform distribution as maximally un-informative
- But a coordinate transformation can make it non-uniform (informative) so that we have to look elsewhere for a least informative probability assignment
  - ⇒ Invariance (symmetry) arguments
  - ⇒ Principle of maximum entropy (MAXENT)
  - ⇒ Just make some reasonable *ansatz*
  - ⇒ …

Insufficient reason from symmetry

- Suppose we have an enumerable set of hypothesis but no other information
- Plot the probability assigned to each hypothesis in a bar chart



- It should not matter how they would be ordered in the chart
- But our bar chart can only be invariant under permutations when all the probabilities are the same
- Hence Bernoulli's principle of insufficient reason

Translation invariance

- Let *x* be a location parameter and suppose that nothing is known about this parameter
- The probability distribution of *x* should then be invariant under translations (otherwise something would be known about *x*)

$$p(x)\mathrm{d}x = p(x+a)\mathrm{d}(x+a) = p(x+a)\mathrm{d}x$$

- But this can be only satisfied when *p*(*x*) is a constant

Scale invariance

- Let *r* be a positive definite scale parameter of which nothing is known
- Scale invariance implies that for $r > 0$ and $\alpha > 0$

$$p(r)\mathrm{d}r = p(\alpha r)\mathrm{d}(\alpha r) = \alpha\, p(\alpha r)\mathrm{d}r$$

- But this is only possible when

$$p(r) \propto 1/r$$

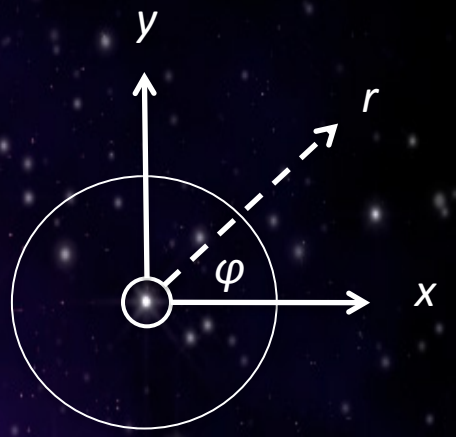
- This is called a Jeffrey's prior
- A Jeffrey's prior is uniform in ln(*r*) and assigns equal probability per decade instead of per unit interval

Improper distributions

- The uniform prior cannot be normalised on the range $[-\infty,+\infty]$ and the Jeffrey's prior not on $(0,\infty]$
- Such distributions are called improper
- Should be dealt with by defining them in a finite range $[a,b]$ and then take the limit at the end of the calculation
- If the posterior is still improper it means that the likelihood cannot sufficiently constrain the prior: your data simply do not carry enough information!

Gauss distribution from invariance

Uncertainty in star position (Herschel 1815)



- He postulated

  ⇒ Position in *x* does not yield information on position in *y*

  ⇒ Distribution does not depend on φ

- From this alone he derived the Gauss distribution (writeup p.38)

$$p(x, y|I) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right)$$

Incomplete information

Suppose we know that a dice is fair

Then we can use the principle of insufficient reason to assign $P(k|I) = 1/6$ to throw face k

We also know that $< k > = 3.5$

Now suppose we are given that $< k > = 4.5$ but nothing else

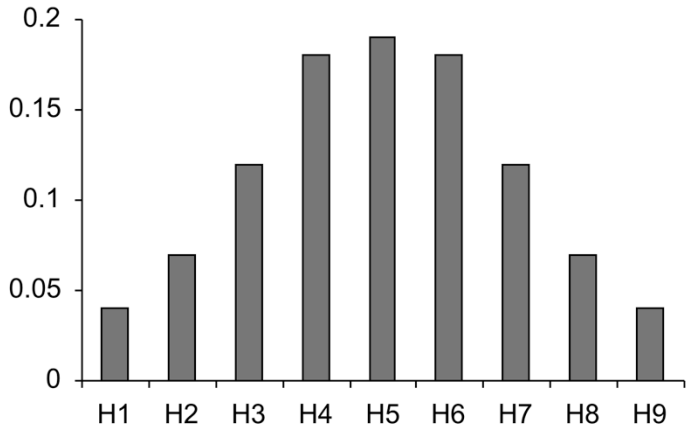
Then there are infinitely many probability assignments that satisfy this constraint

Armed with this incomplete information, what probability should we assign to throw face k ?

Principle of maximum entropy

- Jaynes (1957) has proposed to select the least informative probability distribution by maximising the entropy, subject to the constraints imposed by the available information
- For a set of discrete hypotheses

$$S = -\sum_i P_i \ln\left(\frac{P_i}{m_i}\right)$$



- Larger entropy means less information content

Information entropy

- Information entropy (Shannon 1948)

$$S(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \ln\left(\frac{p_i}{m_i}\right) \quad \text{(discrete case)}$$

$$S(p) = -\int p(x) \ln\left[\frac{p(x)}{m(x)}\right] \mathrm{d}x \quad \text{(continuous case)}$$

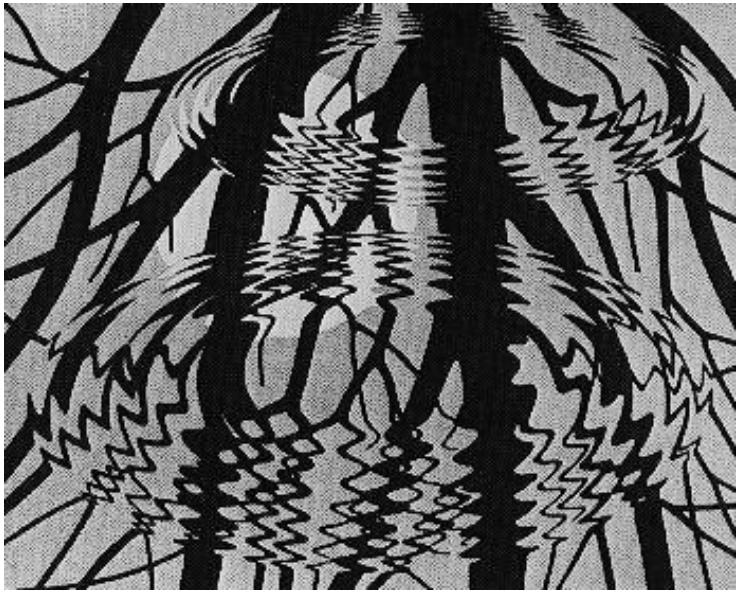
- Here *m* is the Lebesgue measure that satisfies

$$\sum_{i=1}^{n} m_i = 1 \quad \text{or} \quad \int m(x)\,\mathrm{d}x = 1$$

- Lebesgue measure assigns a `size' to any subset of the sample space and makes *S* invariant under coordinate transformations

Lebesgue measure

- Suppose we measure annual rainfall by collecting water in ponds of different size



- The amount of water collected must obviously be normalised to the surface of each pond

- This is the role of the Lebesgue measure m_i in

$$S = -\sum_i P_i \ln\left(\frac{P_i}{m_i}\right)$$

Another view on the m_i

- *N* probabilities with nothing known except normalisation $\sum_{i=1}^{N} P_i = 1$
- The information entropy is $S = -\sum_{i=1}^{N} P_i \ln\left(\frac{P_i}{m_i}\right)$
- To maximise the entropy we have to solve, using Lagrange multipliers $\delta\left[\sum_{i=1}^{N} P_i \ln\left(\frac{P_i}{m_i}\right) + \lambda\left(\sum_{i=1}^{N} P_i - 1\right)\right] = 0$
- This leads to (see writeup) $P_i = m_i$
- The Lebesgue measure is the least informative distribution when nothing is known, except the normalisation constraint; it is a kind of Ur-Prior

Testable information

- In the previous slide we have maximised the entropy while satisfying the normalisation constraint
- Additional constraints like specifying means, variances or higher moments are called testable information (because you can test that your distribution satisfies such constraints)
- Maximising the entropy now becomes more complicated

$$\delta\left[\sum_{i=1}^{n} p_i \ln\left(\frac{p_i}{m_i}\right) + \lambda_0\left(\sum_{i=1}^{n} p_i - 1\right) + \sum_{k=1}^{m}\lambda_k\left(\sum_{i=1}^{n} f_{ki}p_i - \beta_k\right)\right] = 0$$

Normalisation constraint

Testable constraints

MAXENT recipe for continuous distributions

- $f_k(x)$ is a set of functions with given expectation values

$$\int f_k(x)\, p(x|I)\, \mathrm{d}x = \beta_k$$

- The partition function (normalisation integral) is

$$Z(\boldsymbol{\lambda}) = \int m(x)\, \exp\left[-\sum_k \lambda_k f_k(x)\right]$$

- The MAXENT distribution satisfying the constraints is

$$p(x|I) = \frac{1}{Z} m(x)\, \exp\left[-\sum_k \lambda_k f_k(x)\right]$$

Substitute this solution back into the constraint equations to solve for the Lagrange multipliers λ_k (often numerically)

In the write-up you can find how MAXENT is used to derive a few well known distributions

- Uniform distribution
- Exponential distribution
- Gauss distribution
- Poisson distribution

Here the Gauss distribution is of particular interest …

For a continuous distribution $p(x|I)$, the above reads as follows. Let

$$\int f_k(x)\, p(x|I)\, \mathrm{d}x = \beta_k \qquad k = 1, \ldots, m \tag{5.16}$$

be a set of m testable constraints. The distribution that maximises the entropy is then given by

$$p(x|I) = \frac{1}{Z}\, m(x)\, \exp\left[-\sum_{k=1}^{m} \lambda_k f_k(x)\right]. \tag{5.17}$$

Here the partition function Z (normalisation integral) is defined by

$$Z(\lambda_1, \ldots, \lambda_m) = \int m(x)\, \exp\left[-\sum_{k=1}^{m} \lambda_k f_k(x)\right] \mathrm{d}x. \tag{5.18}$$

The values of the Lagrange multipliers λ_i are either found by solving (5.15), or by substituting (5.17) back into (5.16).

Exercise 5.2: Prove (5.15) by differentiating the logarithm of the partition function (5.14) or (5.18).

5.4 MAXENT distributions

In this section we will derive from the maximum entropy principle a few well known distributions: the uniform, exponential, Gauss, and Poisson distributions. The fact that they can be derived from MAXENT sheds some new light on the origin of these distributions; it means that they are not necessarily related to some underlying random process, as is assumed in Frequentist theory (and also in Section 4) but that they can also be viewed as least informative distributions. With the MAXENT assignment, we indeed have moved far away from random variables, repeated observations, and the like.

If there are no constraints, $f(x) = 0$ in (5.16) so that it immediately follows from (5.17), (5.18) and (5.7) that

$$p(x|I) = m(x).$$

For a sample space without structure this gives a uniform distribution, in accordance with the continuum limit of Bernoulli's principle of insufficient reason.

Let us now consider a continuous distribution defined on $[0, \infty]$ and impose a constraint on the mean

$$< x > = \int_0^\infty x p(x|I)\, \mathrm{d}x = \mu \tag{5.19}$$

so that $f(x) = x$ in (5.16). From (5.17) and (5.18) we have, assuming a uniform Lebesgue measure,

$$p(x|I) = e^{-\lambda x} \left[\int_0^\infty e^{-\lambda x}\, \mathrm{d}x\right]^{-1} = \lambda e^{-\lambda x}.$$

Substituting this into (5.19) leads to

$$\int_0^\infty x\, \lambda e^{-\lambda x}\, \mathrm{d}x = \frac{1}{\lambda} = \mu$$

40

Noise with constraint on the variance

- Constraint $$\int_{-\infty}^{\infty} (x-\mu)^2 \, p(x|I) \, \mathrm{d}x = \sigma^2$$

- Calculating the normalisation integral and feeding back the MAXENT distribution into the constraint equation to solve for the Lagrange multiplier gives

⇒ Gauss distribution

$$p(x|\mu,\sigma,I) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

⇒ We don't need to invoke the Central Limit Theorem to justify a Gauss!

To summarise …

- Well known distributions like uniform, exponential and Gauss can be seen as least informative distributions
- In particular the Gauss distribution is the best way to describe noise of which noting else it known than its level (given by the variance σ)
- We do not need to invoke the central limit theorem to justify a Gauss!
- Note that MAXENT assignment is quite a far cry from frequentist ensembles, repeated observations, random variables and so on!

What we have learned

- Basic assignment through the principle of insufficient reason + probability calculus
- Priors are important when
  - Likelihood is wide (data do not carry much information)
  - Likelihood resides near an physical boundary, or even outside (information on boundary is in the prior and not in the data)
- Priors should not contain unsupported information
- Assignment of least informative probabilities by:
  - Principle of insufficient reason for an enumerable, exhaustive and mutually exclusive set of hypotheses
  - Symmetry (invariance) arguments
  - Principle of maximum entropy (MAXENT)

- ✓ Lecture 1

  Basics of logic and Bayesian probability calculus
- ✓ Lecture 2

  Probability assignment
- Lecture 3

  Parameter estimation
- Lecture 4

  Glimpse at a Bayesian network, and model selection