# Bayesian Reasoning and Statistics

Workshop

Philipp Hummel

Graduate Training Center of Neurosciences, Tübingen

Please, interrupt me at any time to ask questions!

# Things I Would Like To Discuss

1. Bayesian Interpretation of probability.
2. The basic principles of Bayesian inference/statistics.
3. Applying the basic principles in an easy example.
4. Approximation strategies for applying the principles in harder settings.
5. Hierarchical Models - One of the most useful ideas.
6. Why are there two ways of doing statistics?
7. What are the key differences between Frequentist and Bayesian statistics?

# Probability Foundations

# Bayesian Idea of Probability

Let's try to formalize the process of deriving new conclusions (inference) from the knowledge we have about some propositions $A$, $B$, ...
We have some commonsensical understanding of how to derive these conclusions in simple cases but formal rules would be useful for hard cases and implementing this inference process in a computer.

One framework for formal inference is deductive logic. However, we can only apply it when we are absolutely certain about the propositions.

We would like to have a framework for formal inference when we are uncertain, i.e. when we have degrees of belief about the propositions:

# Cox Desiderata

Some desiderata for the formal system of manipulating degrees of believe in a proposition:

- Degrees of belief/plausibility are represented by real numbers
- Qualitative correspondence with commonsense
- Consistency:
  - If a conclusion can be reasoned out in more than one way, then every possible way must lead to the same result.
  - The reasoner always takes into account all relevant evidence.
  - Equivalent states of knowledge are represented by equivalent plausibility assignments.

Consequence: Degrees of belief must obey probability theory![1]

[1] Cox, Probability, Frequency and Reasonable Expectation, 1946;
Jaynes, Probability Theory: The Logic of Science, 2003, Chapter 1 & 2

# Degrees of Belief as Probability

Knowing that rational reasoners must obey probability theory (or violate one of the desiderata) is quite important:

- For understanding how agents can pursue goals in uncertain environments (Expected Utility Maximization e.g. in economics).
- For building thinking machines (Bayesian Machine Learning).
- For finding out when human reasoning is flawed (Twersky and Kahneman)
- and making our reasoning more rational.
- For deciding which hypothesis seems most likely in light of the data (Bayesian Statistics).

# Rules of Probability

Everything follows from two basic rules:

Sum Rule:

$$\underbrace{P(A)}_{\text{Marginal Probability of A}} = \sum_B \underbrace{P(A,B)}_{\text{Joint Probability of A and B}}$$

or:

$$P(A) = \int_B p(A,B)$$

Product Rule (Definition of Conditional Probability):

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

$$P(A,B) = P(A|B)P(B) = P(B|A)P(A)$$

# Bayes Rule

Reminder:
Product Rule:
$P(A, B) = P(A|B)P(B) = P(B|A)P(A))$

Sum Rule:
$P(A) = \sum_B P(A, B)$

Bayes Rule:

$$P(A|B) \overset{PR}{=} \frac{P(A,B)}{P(B)} \overset{PR}{=} \frac{P(B|A)P(A)}{P(B)} \overset{SR}{=} \frac{P(B|A)P(A)}{\sum_A P(A,B)} \overset{PR}{=} \frac{P(B|A)P(A)}{\sum_A P(B|A)P(A)}$$

Bayes Rule is a fundamental result of probability theory!

# Bayes Rule applied to Hypotheses

Reminder:
$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

Bayes Rule is so important because it tells us how to update our belief about the hypotheses in question given some new data:

$$\overset{\text{Posterior}}{P(Hypothesis|Data)} = \frac{\overset{\text{Likelihood}}{P(Data|Hypothesis)}\overset{\text{Prior}}{P(Hypothesis)}}{\underset{\text{Evidence}}{P(Data)}}$$

# Summary - Bayesian Interpretation of Probability

Probability theory is an extension of logic that formalizes rational reasoning about uncertain propositions.

Bayes rule tells us how to update our beliefs when new data is arriving.

# Bayesian Inference Foundations

# Three Inferential Goals

When doing bayesian statistics usually we have three goals:

- Estimation of parameter values
- Prediction of/for new data values
- Model Comparison (not dealt with today)

# Three Inferential Goals - Parameter Estimation

Reminder:

$$P(A|B) \overset{PR}{=} \frac{P(A,B)}{P(B)} \overset{PR}{=} \frac{P(B|A)P(A)}{P(B)} \overset{SR}{=} \frac{P(B|A)P(A)}{\sum_A P(A,B)} \overset{PR}{=} \frac{P(B|A)P(A)}{\sum_A P(B|A)P(A)}$$

For estimating the parameter values $\theta$ of a model we can directly apply Bayes Rule:

Posterior Likelihood Prior

$$p(\theta|Data) = \frac{P(Data|\theta)p(\theta)}{P(Data)} = \frac{P(Data|\theta)p(\theta)}{\int_\theta P(Data|\theta)p(\theta)}$$

Evidence

Notice how the notation introduced a lowercase $p$. Parameters are usually continuous valued and therefore each individual parameter value is infinitely unlikely. The lowercase $p$ indicates a probability density function. Real probabilities are in this case only defined for intervals.

# Three Inferential Goals - Prediction

Reminder:
Product Rule:
$P(A, B) = P(A|B)P(B) = P(B|A)P(A))$

Sum Rule:
$P(A) = \int_B P(A, B)$

For calculating the probability of new data values we can use the posterior predictive distribution:

$$\begin{aligned} P(x_{new}|Data) &= \int_\theta p(x_{new}, \theta|Data)d\theta && \text{Sum (/Integration) Rule} \\ &= \int_\theta P(x_{new}|\theta, Data)p(\theta|Data)d\theta && \text{Product Rule} \\ &= \int_\theta P(x_{new}|\theta)p(\theta|Data)d\theta && \text{assuming i.i.d.} \end{aligned}$$

# Bayesian Inference Foundations - Easy Example

# Parameter Estimation in Practice

What does this look like for an actual example?
Let's try to find the posterior distribution of the probability of a coin to show heads $\theta$ (i.e. its bias).

In real life we are rarely interested in the bias of a coin. But there are many interesting scenarios that are mathematically equivalent:

- Success probability of a drug
- Probability of correctness on an exam question
- Probability that a heart surgery patient will survive more than a year after surgery
- Probability that a widget on an assembly line is faulty

# Inferring the bias of a coin

For inferring the bias we also need to specify:

- The likelihood function: This is the probability distribution from which we think our samples are generated.
- A prior for the bias of the coin. Here we are in principle free to specify this in any form we think is reasonable.

# Inferring the bias of a coin – The Likelihood

Reminder:
$p(\theta|Data) = \frac{P(Data|\theta)p(\theta)}{P(Data)} = \frac{P(Data|\theta)p(\theta)}{\int_\theta P(Data|\theta)p(\theta)}$

The likelihood function in this case is the Bernoulli distribution, because we only have two outcomes: Head (coded as 1), Tail (coded as 0)
We have a dataset (e.g. $X = \{1, 0, 1\}$) and a probability for heads $\theta$:

$$P(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}$$

By assuming that all throws are independent and number of heads $h$ and number of tails $t$:

$$P(X|\theta) = \prod_{i=1}^{h+t} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^h(1-\theta)^t$$

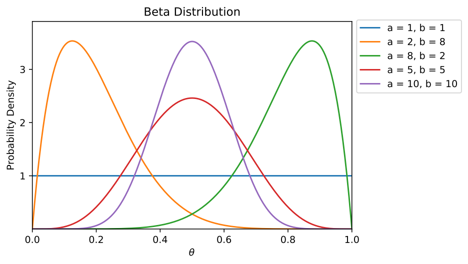
# Inferring the bias of a coin – The Prior

Reminder:
$p(\theta|Data) = \frac{P(Data|\theta)p(\theta)}{P(Data)} = \frac{P(Data|\theta)p(\theta)}{\int_\theta P(Data|\theta)p(\theta)}$

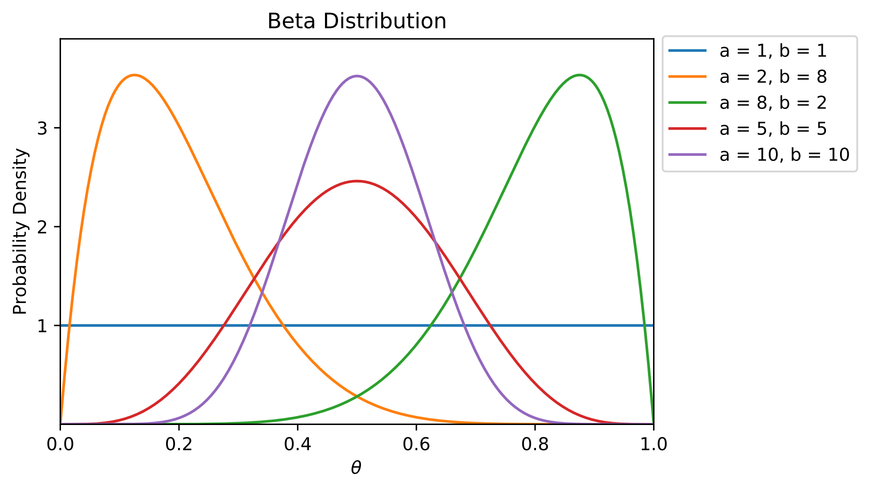
We will use the beta distribution to express our prior beliefs. It assigns a probability density to all values of $\theta \in [0, 1]$ given the shape parameters $\alpha$ and $\beta$. We can interpret these parameters as having seen $\alpha$ heads and $\beta$ tails (Pseudo-Counts).

$$Beta(\theta|\alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta} = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}$$

where $B(\alpha, \beta)$ is the normalizing constant known as the beta function.



# Visualizing the Beta Distribution



# Inferring the bias of a coin – Computing the Posterior

Reminder:

$p(\theta|Data) = \frac{P(Data|\theta)p(\theta)}{P(Data)} = \frac{P(Data|\theta)p(\theta)}{\int_\theta P(Data|\theta)p(\theta)}$

$p(\theta) = Beta(\theta|\alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)} = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta}$

$P(X|\theta) = \theta^h(1-\theta)^t$

$$\begin{aligned} p(\theta|X) = \frac{P(X|\theta)p(\theta)}{\int_\theta P(X|\theta)p(\theta)} &= \frac{\theta^h(1-\theta)^t \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}}{\int_0^1 \theta^h(1-\theta)^t \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} d\theta} \\ &= \frac{\frac{1}{B(\alpha,\beta)}\theta^{h+\alpha-1}(1-\theta)^{t+\beta-1}}{\frac{1}{B(\alpha,\beta)}\int_0^1 \theta^{h+\alpha-1}(1-\theta)^{t+\beta-1}d\theta} \\ &= \frac{\theta^{\alpha_n-1}(1-\theta)^{\beta_n-1}}{\int_0^1 \theta^{\alpha_n-1}(1-\theta)^{\beta_n-1}d\theta} \quad \text{with } \alpha_n := h+\alpha \text{ and } \beta_n := t+\beta \\ &= Beta(\theta|\alpha_n, \beta_n) \end{aligned}$$

Notice how the posterior has the same functional form as the prior. When this is the case for a pair of prior and likelihood, then the prior is said to be **conjugate** to the likelihood.

# Inferring the bias of a coin – Choosing Prior Parameters

Reminder:

$p(\theta|Data) = \frac{p(Data|\theta)p(\theta)}{p(Data)} = \frac{p(Data|\theta)p(\theta)}{\int_\theta p(Data|\theta)p(\theta)}$

$$Beta(\theta|\alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{C}$$

Say we took a coin randomly from my wallet. Because this seems to be a normal coin, we are pretty sure that it is approximately fair. We can translate this into a beta prior with parameters $\alpha = 10; \beta = 10$

$$Beta(\theta|\alpha, \beta) = \frac{\theta^{10-1}(1-\theta)^{10-1}}{C}$$

If we didn't have any prior information about the phenomenon we could use $Beta(\theta|1, 1)$.

# Concrete Numbers – The Posterior

Reminder:

$p(\theta) = Beta(\theta|10, 10) = \frac{\theta^{10-1}(1-\theta)^{10-1}}{C}$

$p(\theta|X) = Beta(\theta|\alpha_n, \beta_n) = \frac{\theta^{\alpha_n-1}(1-\theta)^{\beta_n-1}}{B(\alpha_n,\beta_n)}$

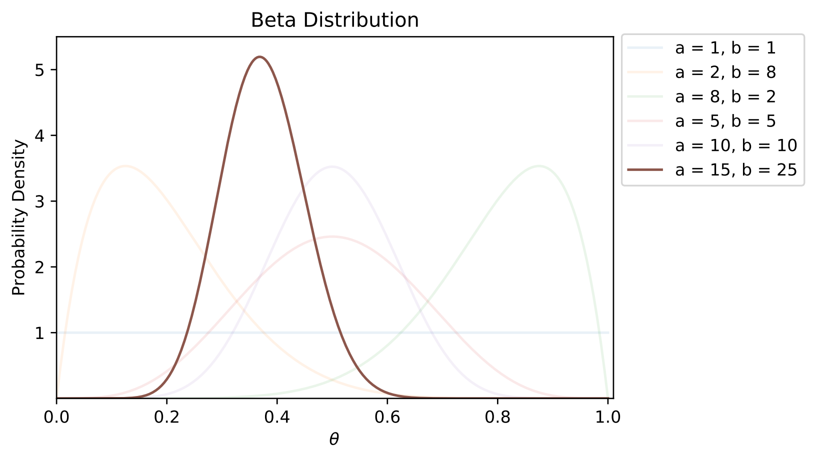
$\alpha_n := h + \alpha$ and $\beta_n := t + \beta$

Say we have observed a dataset X with $h = 5$ heads and $t = 15$ tails. Then the posterior is:

$$p(\theta|X) = \frac{\theta^{\alpha_n-1}(1-\theta)^{\beta_n-1}}{B(\alpha_n,\beta_n)} = \frac{\theta^{15-1}(1-\theta)^{25-1}}{B(15,25)} = Beta(\theta|15, 25)$$

# Practical Example – The Posterior

The posterior probability density $Beta(\theta|15, 25)$ represents our belief about the parameter $\theta$ after we have seen the data.



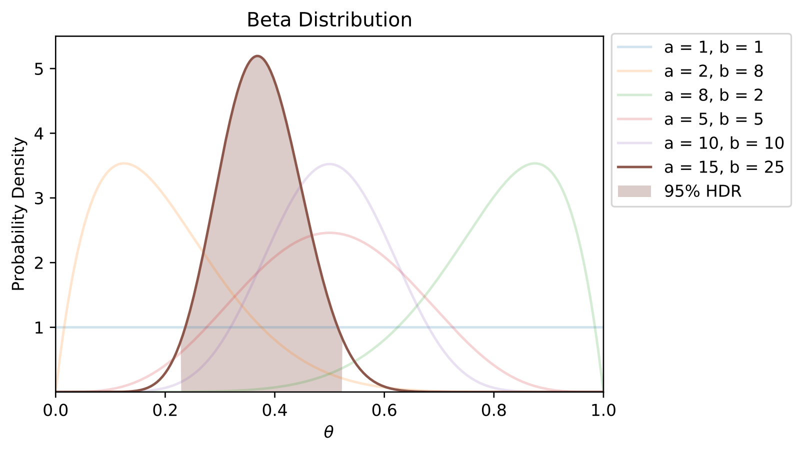
# Highest Density Region

A very useful summary statistics is the Highest Density Region (HDR);
(sometimes also called Highest Probability Density (HPD) or Highest Density Interval (HDI) in the literature)

This region can summarize a range of credible parameter values. Each parameter value in the region is more credible than any values outside the region

For example we might want to look at the 95% HDR for our coin bias:

# Highest Density Region



# Bayesian Statistics - Subjective Priors?

So far we have chosen priors pretty much randomly (the functional form and the parameters). This seeming arbitrariness has led many people to argue that we cannot use bayesian versions of statistics in science since the results are subjective.

In Principle:

- Yes, subjective. Refers to your incomplete state of knowledge.
- Still objective. Everyone with the same knowledge needs to come to the same conclusion (or violate desiderata).

In Practice:

- Perfectly translating background knowledge into numerical prior is infeasible.
- Different people have different background knowledge.

Classical Statistics ignores the problem but that doesn't make it go away.

# Solutions to Subjective Priors

How to choose priors for statistics:

- Uninformative Priors (a prior that assigns equal probability to all hypotheses)
- General World Knowledge
- Previous Experiments
- Audience-agreeable Prior
- Evaluate your results with several priors

But even more importantly:

- Report likelihood ratios between hypotheses[3]
- PUBLISH YOUR DATA (and your analysis code)!

[3] See this blog post for why this might be a good idea

# Inferring the bias of a coin – Posterior Predictive Distribution

Reminder:

$P(x_{new}|Data) = \int_\theta P(x_{new}|\theta)p(\theta|Data)d\theta$

$p(\theta|X) = Beta(\theta|\alpha_n, \beta_n) = \theta^{\alpha_n - 1}(1-\theta)^{\beta_n - 1}$

$P(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}$

We can use the posterior predictive distribution to calculate the probabilities for new data values in our coin example:

$$\begin{aligned} p(x_{new}|Data) &= \int_\theta p(x_{new}|\theta)p(\theta|Data)d\theta \\ &= \int_\theta \theta^{x_{new}}(1-\theta)^{1-x_{new}} Beta(\theta|\alpha_n, \beta_n) \\ &= \begin{cases} \frac{\alpha_n}{\alpha_n+\beta_n}, & \text{if } x_{new} = 1. \qquad \big(\text{Expected value of beta distribution}\big) \\ \frac{\beta_n}{\alpha_n+\beta_n}, & \text{otherwise.} \end{cases} \end{aligned}$$

# Inferring the bias of a coin – Posterior Predictive Distribution

Reminder:
$p(\theta|X) = Beta(\theta|\alpha_n, \beta_n) = Beta(\theta|15, 25)$

$$\begin{aligned} p(x_{new}|Data) &= \int_\theta p(x_{new}|\theta)p(\theta|Data)d\theta \\ &= \int_\theta \theta^{x_{new}}(1-\theta)^{1-x_{new}} Beta(\theta|\alpha_n, \beta_n) \\ &= \begin{cases} \frac{\alpha_n}{\alpha_n+\beta_n}, & \text{if } x_{new} = 1. \qquad \text{(Expected value of beta distribution)} \\ \frac{\beta_n}{\alpha_n+\beta_n}, & \text{otherwise}. \end{cases} \end{aligned}$$

In our current example we would predict $x_{new}$ with probability:

$$P(x_{new}|Data) = \begin{cases} \frac{\alpha_n}{\alpha_n+\beta_n} = \frac{15}{15+25} = \frac{3}{8}, & \text{if } x_{new} = 1. \\ \frac{\beta_n}{\alpha_n+\beta_n} = \frac{25}{15+25} = \frac{5}{8}, & \text{otherwise}. \end{cases}$$

# Summary - The basic principles of Bayesian inference

Parameter Estimation:

$$p(\theta|Data) \quad = \quad \frac{P(Data|\theta)p(\theta)}{P(Data)} \quad = \quad \frac{P(Data|\theta)p(\theta)}{\int_\theta P(Data|\theta)p(\theta)}$$

Prediction:

$$P(x_{new}|Data) = \int_\theta P(x_{new}|\theta)p(\theta|Data)d\theta$$

# Break

Let's have a break!

# First Half

So far we have seen:

- Probability theory can be interpreted as an extension of logic that formalizes rational reasoning about uncertain propositions.
- Two easy equations for parameter estimation: $p(\theta|Data) = \frac{P(Data|\theta)p(\theta)}{\int_\theta P(Data|\theta)p(\theta)}$ and prediction: $P(x_{new}|Data) = \int_\theta P(x_{new}|\theta)p(\theta|Data)d\theta$
- Some ideas for how to choose priors.

# Second Half

Now we will deal with:

- Why we need approximate approaches for doing bayesian inference and what these approaches are
- The idea of hierarchical models
- Why there are two ways of doing statistics and what their differences are

# Approximate Approaches

# Why do we need approximations?

The integral in bayes rule is only analytically solvable in very easy cases and if we choose the prior to be conjugate to the likelihood:

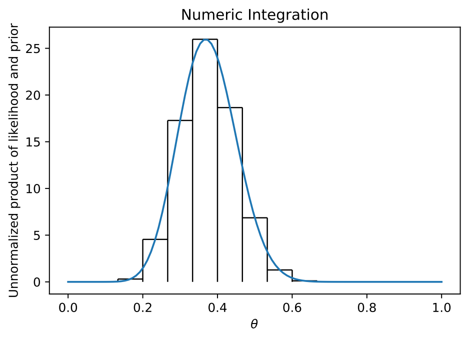
$$p(\theta|Data) = \frac{p(Data|\theta)p(\theta)}{p(Data)} = \frac{p(Data|\theta)p(\theta)}{\int_{\theta} p(Data|\theta)p(\theta)}$$

For almost any model of interest, e.g. (bayesian) linear regression, this integral is not analytically solvable. So we need good and efficient approximation methods if we want to do bayesian inference for these models.
I would like to show you two ways of doing it:

# Approximate Approaches - Numeric Integration

# Numeric Integration

One way to tackle the problem of this integral is to just use numeric integration techniques. We can approximately find the integral of a function $f(x)$ in an interval $[a,b]$, $\int_a^b f(x)dx$, by breaking the interval up into small rectangular bins. The height of a bin is determined by the function value at the midpoint of the bin. Calculating the area of the bins is easy and the total integral is just the sum of all of the bins.



# Numeric Integration - Problems

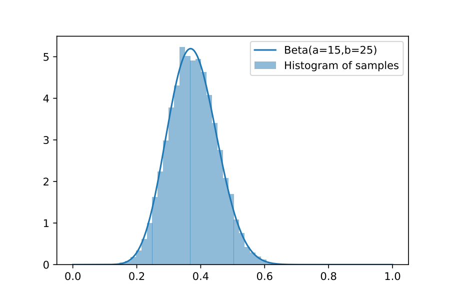
Numeric integration works fine for very low dimensional problems. However, the computational effort grows exponentially with the dimensionality of the problem. As a rule of thumb: do not attempt to use numeric integration when you have more than five parameters.

Luckily, there is a very useful family of indirect approximation methods:

# Approximate Approaches - Sampling with Markov Chain Monte Carlo Methods

# Markov Chain Monte Carlo

The idea is to use many random samples to approximate a distribution and statistics of interest of this distribution.



True Mean: 0.375
Sample Mean: 0.376

However, often is not easy to obtain samples from the posterior distribution directly.

# Markov Chain Monte Carlo – Metropolis-Hastings Algorithm

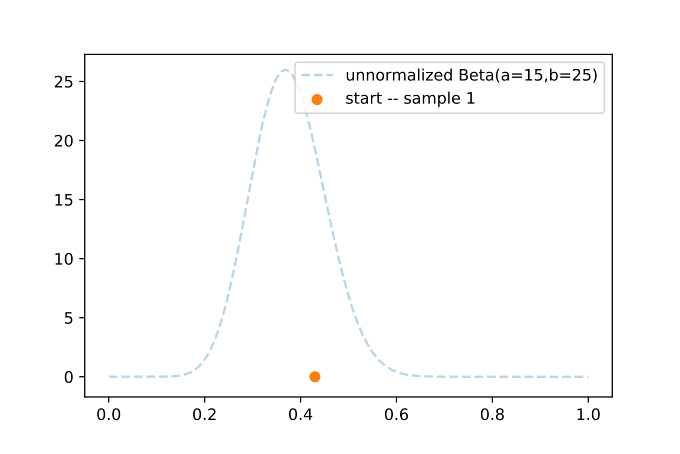
In the cases where we cannot obtain samples directly from the posterior, we can use a proposal distribution $q(z^{t+1}|z^t)$ to explore the posterior and generate samples from it.
The Metropolis-Hastings Algorithm:

1. Start somewhere in parameter space – this is our first sample $z^1$.
2. Use the proposal distribution $q(z^{t+1}|z^t)$ to propose a new sample.
3. Evaluate the unnormalized posterior density $\tilde{p}$ at $z^{t+1}$ and $z^t$.
   We can do this by using $\tilde{p}(z) = P(Data|\theta = z)p(\theta = z)$.
4. Accept the new point with acceptance probability
   $A(z^{t+1}, z^t) = \min(1, \frac{\tilde{p}(z^{t+1})q(z^t|z^{t+1})}{\tilde{p}(z^t)q(z^{t+1}|z^t)})$
   Otherwise the current point is used as a sample again.
5. Repeat from 2.

The samples form a markov chain, because the value of the sample at time $t$ is dependent on the value of the sample at time $t-1$.

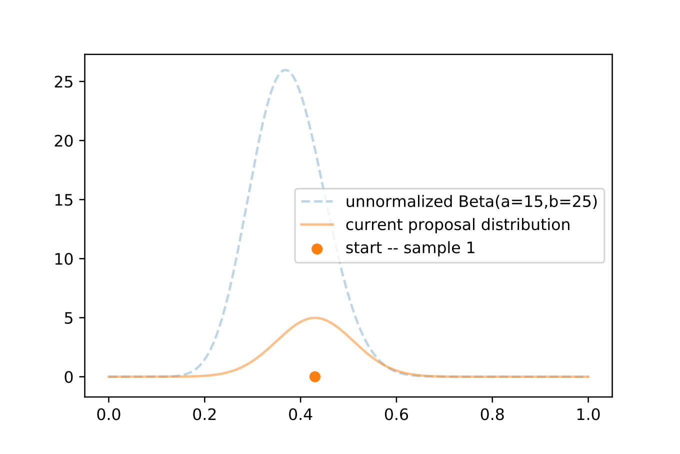
# The Metropolis-Hastings Algorithm

We use an unnormalized version of our $\mathit{Beta}(\alpha = 15, \beta = 25)$ to visualize the metropolis-hastings algorithm.



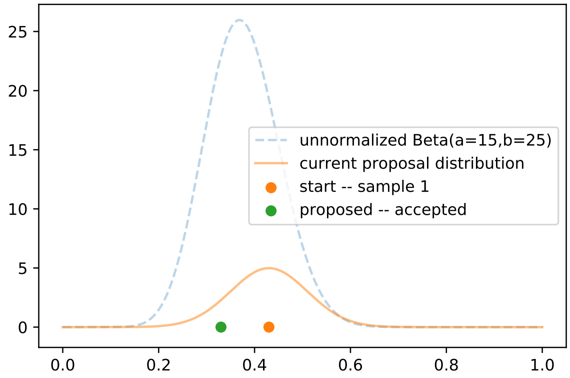
# The Metropolis-Hastings Algorithm

We use an unnormalized version of our $Beta(\alpha = 15, \beta = 25)$ to visualize the metropolis-hastings algorithm.



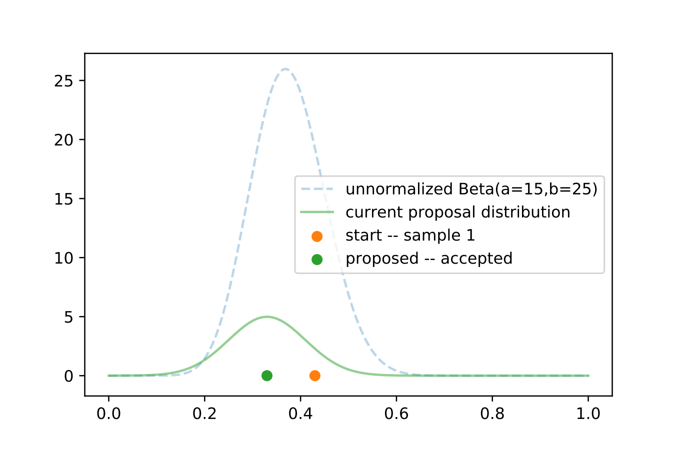
# The Metropolis-Hastings Algorithm

We use an unnormalized version of our *Beta*$(\alpha = 15, \beta = 25)$ to visualize the metropolis-hastings algorithm.



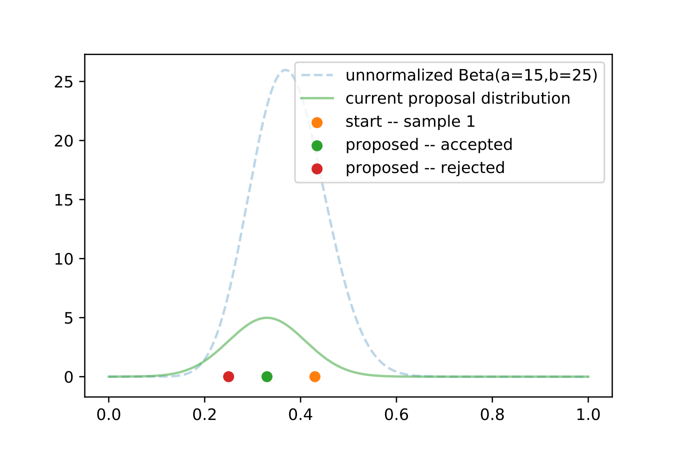
# The Metropolis-Hastings Algorithm

We use an unnormalized version of our *Beta*$(\alpha = 15, \beta = 25)$ to visualize the metropolis-hastings algorithm.



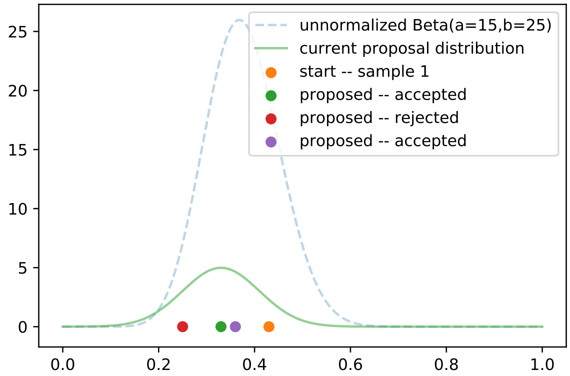
# The Metropolis-Hastings Algorithm

We use an unnormalized version of our $Beta(\alpha = 15, \beta = 25)$ to visualize the metropolis-hastings algorithm.



# The Metropolis-Hastings Algorithm

We use an unnormalized version of our *Beta*$(\alpha = 15, \beta = 25)$ to visualize the metropolis-hastings algorithm.



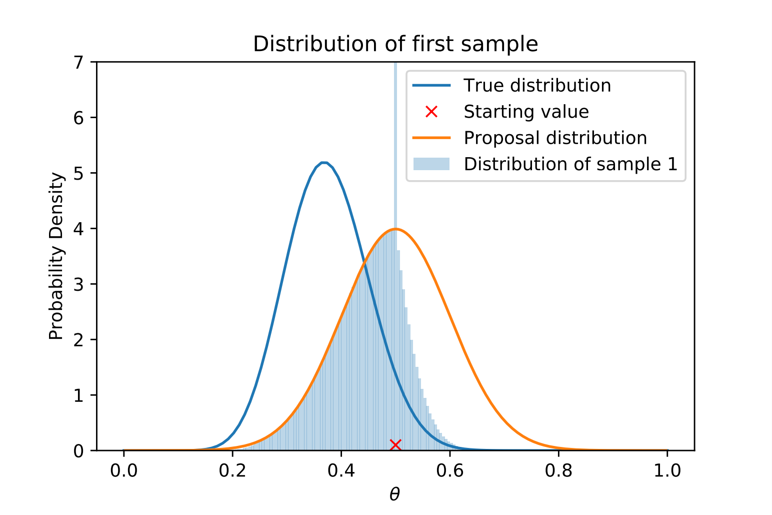
# The Metropolis-Hastings Algorithm - Convergence

Does the Metropolis-Hastings algorithm produce samples that are distributed like the posterior?

What is the distribution of sample *n* (if we were to construct many markov chains of length *n*)?

# The Metropolis-Hastings Algorithm - Convergence

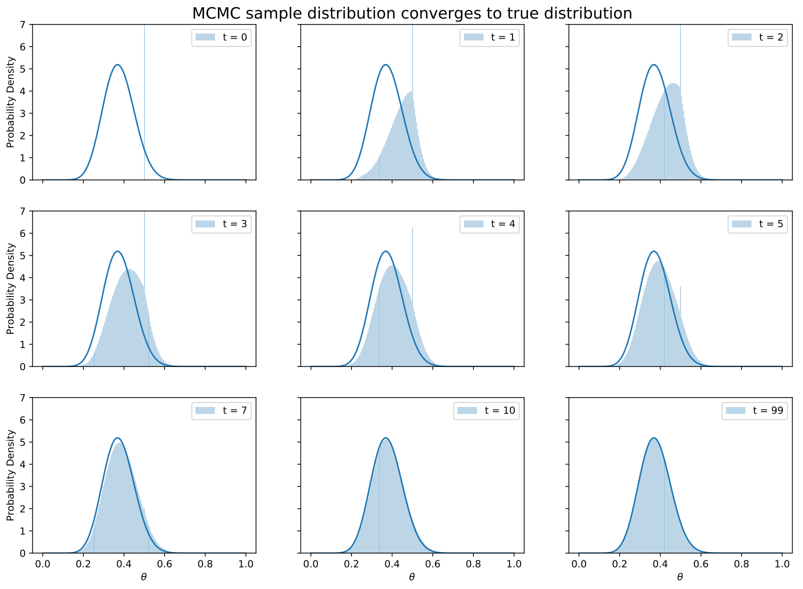
The distribution of sample one is clearly not close to the true posterior distribution:



# The Metropolis-Hastings Algorithm - Convergence

The Metropolis-Hastings algorithm is proven to construct a markov chain whose distribution of sample $n$ converges to the true distribution for $n \to \infty$. We can visualize that for our example:

# The Metropolis-Hastings Algorithm - Convergence

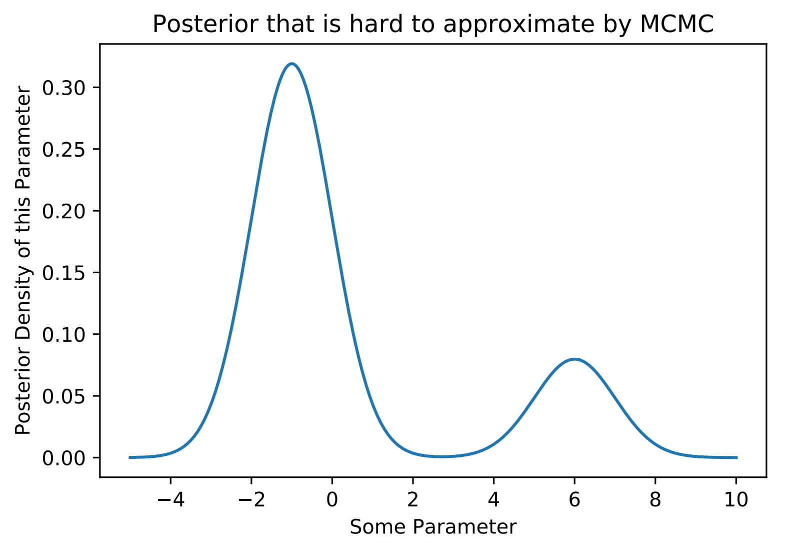


# Markov Chain Monte Carlo

Using Markov Chain Monte Carlo methods we are guaranteed to produce samples from the true posterior distribution when we let our chains run for long enough. The initial period during which the samples generated do not yet correspond to the true distribution is called **burn-in** period.

# Markov Chain Monte Carlo - Problems

Problems with finitely long chains can occur e.g. when the posterior is strongly bimodal:



# Markov Chain Monte Carlo

There are some more problems involved with having finitely long chains and there are some partial solutions to these problems but we won't have time to cover them here.[4]

I will just say that some problems have been overcome by clever sampling algorithms:
You should check out the **No U-turn Sampler (NUTS)**

and that there are now very convenient software frameworks around to do Markov Chain Monte Carlo for you:
I will show you very shortly one framework called **PyMC**.

[4] But if you want to find out more, read
Kruschke, Doing Bayesian Data Analysis, 2014, Chapter 7 & 14.1, for a very intuitive introduction and
Barber, Bayesian Reasoning and Machine Learning, 2012, Chapter 27.3 - 27.5, for a more mathematical introduction.

# Markov Chain Monte Carlo - Code

I will show you some PyMC example code![5]

[5] PyMC has a tutorial section: https://docs.pymc.io/nb_tutorials/index.html

# Summary - Approximation Strategies

When you can't find the posterior distribution analytically you can:

- Use numeric integration to solve the integral in bayes rule when you have less than five parameters.
- Approximate the posterior distribution with samples indirectly generated by Markov Chain Monte Carlo methods e.g. with PyMC or Stan[6]
- Use variational inference to approximate the posterior distribution[7]

[6] https://mc-stan.org/

[7] You can learn about it in this coursera course or read Murphy, Machine Learning: A Probabilistic Perspective, 2012, Chapter 21 + 22

# Hierarchical Models

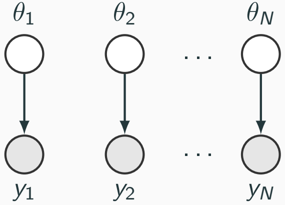
# Multiple Coins from the same Mint

Suppose we are given data from $N$ coins (with biases $\theta_i$) that come from the same mint. We might want to know whether some individual coins are biased and/or whether the mint produces biased coins in general. We have several possibilities how to model this scenario:

# Unpooled Model

Denoting the complete data we have for one coin (i.e. $h$ heads out of $n$ throws) by $y_i$.

We could model each coin bias individually:

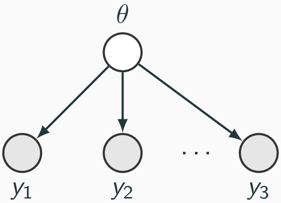


For coins with few data it will be hard to estimate the bias correctly. We are disregarding information from the other coins. Also it is not straightforward to make claims about the mint.

# Pooling the Model

Denoting the complete data we have for one coin (i.e. $h$ heads out of $n$ throws) by $y_i$.

We could pool all coins together:

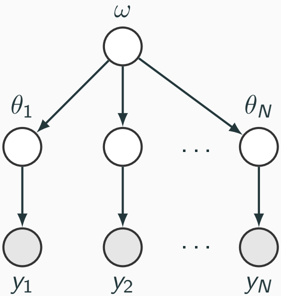


No claims about individual coins are possible. We might now make claims about the mint being biased. But what if some coins are tails biased while some others are heads biased?

# Best of both Worlds - Hierarchical Model

Denoting the complete data we have for one coin (i.e. $h$ heads out of $n$ throws) by $y_i$.

It feels most natural to assume that each bias is drawn from a distribution of biases (parameterized by $\omega$) that the mint produces:



Thereby, the data from every coin indirectly influences our estimate of every other coin. We can make better claims about coins for which we have little data and we can make direct claims about the mint.

# Hierarchical Models - Final

This kind of hierarchical structure appears very often in experiments.

- Estimating whether having a basement affects radon levels in households in different counties.[8]
- Estimating how likely a rat develops a tumor given a drug; with data from 71 experiments.[9]

Hierarchical models allow us to make the most of the available data.[10]

[8] https://twiecki.io/blog/2014/03/17/bayesian-glms-3/

[9] https://docs.pymc.io/notebooks/GLM-hierarchical-binominal-model.html

[10] Learn more about them in general in:
Kruschke, Doing Bayesian Data Analysis, 2014, Chapter 9

# Differences between the Frequentist and the Bayesian approach

# Cox Desiderata

Remember:

Some desiderata for the formal system of manipulating degrees of believe in a proposition:

- Degrees of belief/plausibility are represented by real numbers
- Qualitative correspondence with commonsense
- Consistency:
    - If a conclusion can be reasoned out in more than one way, then every possible way must lead to the same result.
    - The reasoner always takes into account all relevant evidence.
    - Equivalent states of knowledge are represented by equivalent plausibility assignments.

Consequence: Degrees of belief must obey probability theory![11]

[11] Cox, Probability, Frequency and Reasonable Expectation, 1946;
Jaynes, Probability Theory: The Logic of Science, 2003, Chapter 1 & 2

# Interpretations of Probability

Some people claim that probability as degree of belief makes no sense. Then we could not assign probabilities to hypotheses and propositions.

Actually, the interpretation of what probabilities represent is quite disputed.
Wikipedia[12] lists two broad categories subdivided into six main interpretations.

[12] https://en.wikipedia.org/wiki/Probability_interpretations

# Frequentist vs. Bayesian Statistics

Accordingly, the two categories of statistics are divided along their interpretation of probability:

| Frequentist | Bayesian |
|---|---|
| Long-run relative Frequency | Degree of Belief |
| • If you throw a fair coin (corresponding to a bias of 0.5) one million times it is very likely that the relative frequency of heads is close to 0.5.<br>• Uncertainty/Probability is a property of things in the outside world. | • How strongly should a rational agent belief that the coin is approximately fair given that she has seen some flips and her background knowledge?<br>• Uncertainty/Probability is a property of a reasoner's state of knowledge. |

Bayesian methods were computationally intractable when frequentist statistics was invented.

# Frequentist vs. Bayesian reasoning about hypotheses

Reminder:

$P(Hypothesis|Data) = \frac{P(Data|Hypothesis)P(Hypothesis)}{P(Data)}$

In the bayesian framework it makes sense to say something like $P(Hypothesis) = 0.5$.
But a single hypothesis, is either true or false. Therefore, in the frequentist setting $P(Hypothesis)$ does not even make sense.
The bayesian setting allows computing the posterior probability of the hypothesis given the data, because probability is interpreted as the degree of belief in a statement.

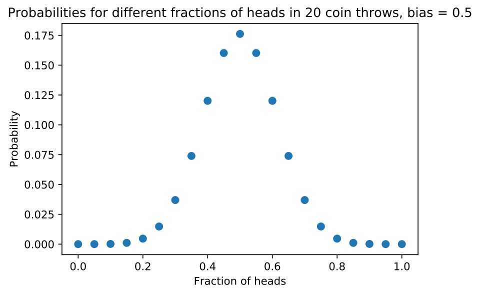
$\Longrightarrow$ In the bayesian framework we can apply Bayes rule to more cases. Especially also when reasoning about individual hypotheses.

# Frequentist reasoning about hypotheses

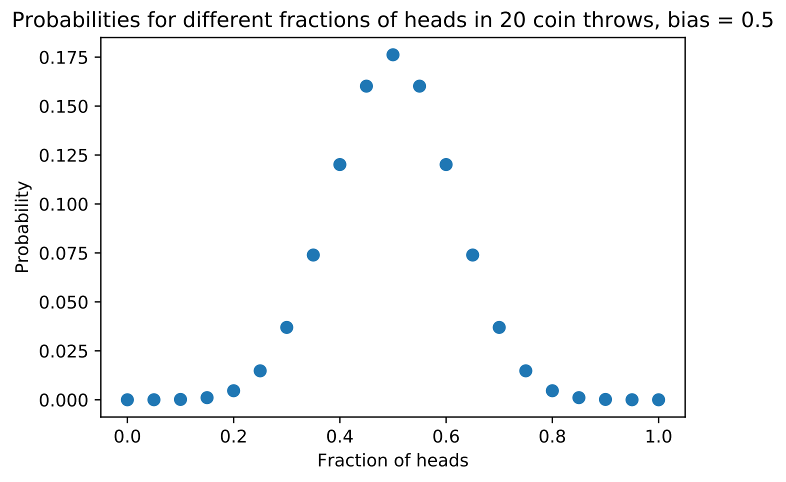
In the frequentist setting probabilities cannot be assigned to hypotheses directly, so people invented indirect methods to reason about hypotheses.

A p-value is the probability of seeing the observed data or more extreme (with respect to the alternative hypothesis) data assuming that the Null-Hypothesis were true.

p-value: $P(MoreExtremeData|H_0)$



# Probability of Data given $H_0$ - Visualization



# p-value problems

It is of course permissible in principle to use p-values but there are many practical misunderstandings around them.

p-value: $P(\mathit{MoreExtremeData}|H_0)$

Fallacies:

$$\begin{aligned}\text{p-value} &\neq P(H_0|\mathit{Data})\\ &\neq P(\neg H_1|\mathit{Data})\end{aligned}$$

The p-value does not indicate the size of the effect.[14]

[14] See the Wikipedia article about misunderstandings of the p-value: https://en.wikipedia.org/wiki/Misunderstandings_of_p-values

# Frequentist Subjectivity

The p-value depends on (the sampling distribution which depends on) the experimenter's state of mind when doing research[15].

- One-tailed and two-tailed tests
- The experimenter's stopping criterion
- Multiple Testing

With the choice of prior, bayesian versions of statistics show openly what assumptions they are making.

[15] See this blog post or Kruschke, Doing Bayesian Data Analysis, 2014, Chapter 11.

# Which Interpretation is Correct?

Which interpretation of probability is the correct one?

Basic desiderata about how to reason rationally about propositions lead to probability theory.
Additionally, if an agent's degrees of belief do not satisfy the rules of probability theory, one can construct a series of bets in which the agent will lose for sure. (**Dutch Book**)

The bayesian interpretation of probability seems valid and useful.

What about the frequentist interpretation?

# Which Interpretation is Correct?

The frequentist interpretation, that probability is a property of things, seems quite intuitive.

But imagine the following:
We have an apparatus build to flip coins in a way that $P(Heads) = 0.5$. Now we build an advanced machine learning algorithm with very good computer vision and understanding of the mechanics of the flipping apparatus. By looking at the apparatus' setting one second before the flip, the algorithm can predict the outcome with 90% certainty. What is the probability of the coin to land heads?

No matter these philosophical discussion, frequentist methods are useful in the domains for which they were developed.

# Final Summary

- Probability theory can be interpreted as an extension of logic that formalizes rational reasoning about uncertain propositions.
- Two easy equations for
  parameter estimation: $p(\theta|Data) = \frac{P(Data|\theta)p(\theta)}{\int_\theta P(Data|\theta)p(\theta)}$
  and prediction: $P(x_{new}|Data) = \int_\theta P(x_{new}|\theta)p(\theta|Data)d\theta$
- Some ideas for choosing priors.
- The integral in Bayes rule is often intractable.
- Markov Chain Monte Carlo methods can approximate distributions.
- Hierarchical Models can naturally represent the structure present in the data and draw better conclusions.
- Frequentists' interpretation of probabilities as long-run relative frequencies.
- p-value: the probability of seeing more extreme data if the Null-Hypothesis were true.