SCIENCE ACADEMY

How to Choose Priors

- **Weakly-informative priors** (aka "flat", "vague", "diffuse priors")
  - Provide minimal information
    - Coefficient of linear regression centered around 0: $\beta \sim \mathit{Normal}(0, 10)$
- **Regularizing priors**
  - Known information about the parameter
    - Parameter is positive: $\sigma \sim \mathit{HalfCauchy}(0, 5)$
    - Parameter close to zero, above/below a number, or in a range
    - $\beta \sim \mathit{Laplace}(0, 1)$ (lasso prior) encourages sparsity
    - $\beta \sim \mathit{Normal}(0, 1)$ discourages extreme values
- **Informative priors**
  - Strong priors from previous knowledge (expert opinion, studies)
    - From experimental data: $\beta_1 \sim \mathit{Normal}(2.5, 0.5^2)$
    - From previous data, about 5% of cases positive: $p \sim \mathit{Beta}(2, 38)$
- **Prior elicitation**
  - Compute least informative distribution given constraints
    - Estimate distribution using maximum entropy to satisfy constraints
    - E.g., beta distribution with 90% of mass between 0.1 and 0.7

- ***Communicating a Bayesian Analysis***
- Probabilistic Programming
- Posterior-Based Decisions

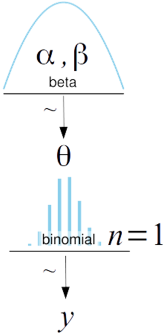
Communicating the Model of a Bayesian Analysis

1. **Communicate assumptions / hypothesis**
   - Describe priors and probabilistic models
   - E.g., coin-flip distributions:

$$\begin{cases} \theta \sim \text{Beta}(\alpha, \beta) \\ y \sim \textit{Binomial}(n = 1, p = \theta) \end{cases}$$

2. **Communicate Bayesian analysis result**
   - Describe posterior distribution
   - Summarize location and dispersion
   - Mean (or mode, median)
   - Std dev
     - Misleading for skewed distributions
   - Highest-posterior density (HPD)
     - Shortest interval containing a portion of probability density (e.g., 95% or 50%)
     - Amount is arbitrary (e.g., `ArviZ` defaults to 94%)

Kruschke diagram



Confidence Intervals vs Credible Intervals

- People confuse:
  - Frequentist **confidence intervals** with
  - Bayesian **credible intervals**
- In the frequentist framework, there is a true (unknown) parameter value
  - A **confidence interval** may or may not contain the true parameter value
  - Interpretation of a 95% confidence interval
    - No: *"There is a 95% probability that the true value is in this interval"*
    - Yes: *"If repeated many times, 95% of intervals would contain the true value"*
- In the Bayesian framework, parameters are random variables
  - Interpretation of a 95% **Bayesian credible interval**
    - *"There is a 95% probability that the true parameter lies within this interval, given the observed data"*
    - Bayesian **credible interval** is intuitive

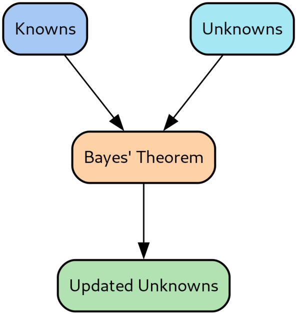
Confidence Intervals vs Credible Intervals (ELI5)

- **Confidence Interval (Frequentist)**
  - Imagine fishing in a lake without seeing the fish
  - You throw your net
  - 95% confidence interval: *"If I threw this net 100 times, about 95 nets would catch the fish."*
  - Important: Once the net is thrown, it either caught the fish or not. The 95% makes sense across many attempts
- **Credible Interval (Bayesian)**
  - Imagine a magical map showing where fish *probably* are, based on past observations
  - 95% credible interval: *"Given my map, there's a 95% chance the fish is inside this part of the lake."*
  - The fish's location is uncertain, and probability describes your belief
- **Key Difference**
  - Confidence interval (Frequentist): Probability from repeating experiments
    - It's about the procedure, not the specific interval
  - Credible interval (Bayesian): Probability describes your belief about the value, given the data
    - It's about *this interval*

- Communicating a Bayesian Analysis
- ***Probabilistic Programming***
- Posterior-Based Decisions

Bayesian Statistics

- Given:
  - The **"knows"**
    - Model structure (modeled as a graph of probability distributions)
    - Data, observations (modeled as constants)
  - The **"unknowns"**
    - Model parameters (modeled as probability distributions)
- Use Bayes' theorem to condition unknowns to knowns hoping to reduce the uncertainty about the unknowns



- **Problem**
  - Most probabilistic models are analytically intractable
- **Solution**
  - Probabilistic programming
    - Specify a probabilistic model using code
    - Solve models using numerical techniques

Probabilistic Programming Languages

- **Steps**:
  1. Specify models using code
  2. Numerical models solve inference problems without user understanding
     - Universal inference engines
     - `PyMC3`: flexible Python library for probabilistic programming
     - `Theano`: library to define, optimize, evaluate mathematical expressions using tensors
     - `ArviZ`: library to interpret probabilistic model results
- **Pros**:
  - Compute results without analytical closed form
  - Treat model solving as a black box
  - Focus on model design, evaluation, interpretation
- Probabilistic programming languages impact like Fortran on scientific computing
  - Build algorithms, ignore computational details

Coin Example: Numerical Solution (1/3)

- Assume you know the true value of θ (not true in general)
- Observe samples of the variable y
- Model the prior θ and the likelihood $y|\theta$

$$\begin{cases} \theta \sim \text{Beta}(\alpha = 1, \beta = 1) \\ Y \sim \text{Binomial}(n = 1, p = \theta) \end{cases}$$

- Run inference
- Generate samples of the posterior
- Summarize posterior
  - E.g., Highest-Posterior Density (HPD)

SCIENCE ACADEMY

Coin Example: Numerical Solution (2/3)

[18]:
```python
np.random.seed(123)
n = 4
# Unknown value.
theta_real = 0.35

# Generate some observational data.
data = stats.bernoulli.rvs(p=theta_real, size=n)
data
```

[18]:
```
array([1, 0, 0, 0])
```

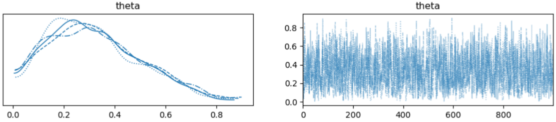
[19]:
```python
with pm.Model() as our_first_model:
    # Prior.
    theta = pm.Beta('theta', alpha=1., beta=1.)
    # Likelihood.
    y = pm.Bernoulli('y', p=theta, observed=data)
    # (Numerical) Inference to estimate the posterior distribution through samples.
    idata = pm.sample(1000, random_seed=123)
```

```
Auto-assigning NUTS sampler...
Initializing NUTS using jitter+adapt_diag...
Multiprocess sampling (4 chains in 4 jobs)
NUTS: [theta]
```

```
Sampling 4 chains, 0 divergences ━━━━━━━━━━━━━━━━━━━━ 100% 0:00:00 / 0:00:00
```

```
Sampling 4 chains for 1_000 tune and 1_000 draw iterations (4_000 + 4_000 draws total) took 1 seconds.
```

[20]:
```python
az.plot_trace(idata);
```



[21]:
```python
az.summary(idata)
```

[21]:

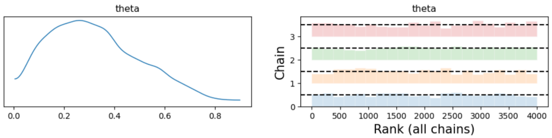
	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
theta	0.324	0.179	0.031	0.653	0.005	0.003	1500.0	1737.0	1.0

- Generate data from ground truth model
- Build PyMC model matching mathematical model
- PyMC uses NUTS sampler, computes 4 chains
- No trace diverges
- Kernel density estimation (KDE) for posterior (should be Beta)
- Traces appear “noisy” and non-diverging (good)
- Numerical summary of posterior: mean, std dev, HDI
- $\mathbb{E}[\hat{\theta}] \approx 0.324$
- $\Pr(\hat{\theta} \in [0.031, 0.653]) = 0.94$

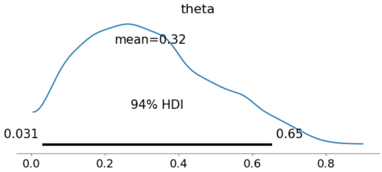
SCIENCE ACADEMY

Coin Example: Numerical Solution (3/3)

[22]:
```python
az.plot_trace(idata, kind="rank_bars", combined=True);
```



[23]:
```python
az.plot_posterior(idata);
```



- Compute single KDE for all chains
- Rank plot to check results
- Histograms should look uniform, exploring different (and all) posterior regions
- Plot single KDE with all statistics

SCIENCE ACADEMY

- Communicating a Bayesian Analysis
- Probabilistic Programming
- ***Posterior-Based Decisions***

Posterior-Based Decisions

- Sometimes describing the posterior is not enough
  - You need to make decisions based on our inference
- E.g., is the coin fair ($\theta = 0.5$) or biased?
  - Since $\mathbb{E}[\hat{\theta}] = 0.324$ it seems that the coin is biased
  - You can't rule out that the coin in unbiased since
    - $HPI = [0.03, 0.65]$
    - $0.5 \in HPI$
- If you want a sharper decision, you need to:
  - Collect more data to reduce the spread of the posterior
  - Define a more informative prior

Savage-Dickey Density Ratio

- The Savage-Dickey ratio tests a point null-hypotheses in Bayesian inference
- **Idea**: compare prior and posterior densities at a single point θ_0

$$BF_{01} = \frac{p(\theta_0|H_1)}{p(\theta_0|\mathcal{D}, H_1)}$$

where:

- $p(\theta_0|H_1)$ is the *prior* density θ under the alternative hypothesis H_1, evaluated at θ_0
- $p(\theta_0|\mathcal{D}, H_1)$ is the *posterior* density θ under H_1 evaluated at θ_0

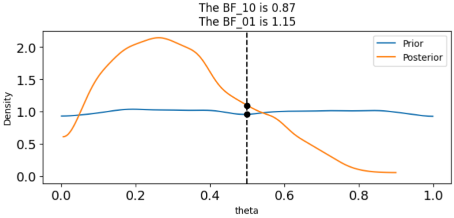
Bayes Factor (BF)	Interpretation
1 - 3	Not enough evidence
3 - 10	Substantial evidence
10 - 100	Strong evidence
$>$ 100	Decisive evidence

- **Intuition**: this ratio shows how much data changes belief about θ_0
  - If posterior density at θ_0 is much smaller than prior density, Bayes factor suggests strong evidence against H_0

SCIENCE ACADEMY

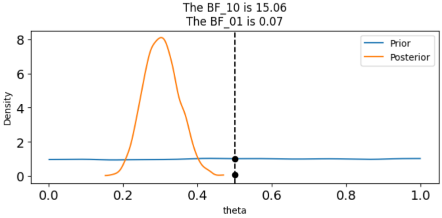
Savage-Dickey Density Ratio: Example

```
[29]: az.plot_bf(idata1, var_name="theta", prior=np.random.uniform(0, 1, 10000), ref_val=0.5);
```




```
[30]: az.plot_bf(idata2, var_name="theta", prior=np.random.uniform(0, 1, 10000), ref_val=0.5);
```

```
The reference value is outside of the posterior. This translate into infinite support for H1
```



- H_0: “coin is fair”
- The prior for H_0 is 0.87
- The posterior for H_0 is 1.15
- $BF_{10} = $

ROPE: Region of Practical Equivalence

- **ROPE** = an interval for a parameter where all values inside are considered "equivalent"
  - H_0: *"coin is fair"* iff $\theta = 0.5$ is impractical
  - ROPE: $\theta \in [0.45, 0.55]$ is equivalent to 0.5
- **Hypothesis testing with ROPE and HPI**
  - Compare ROPE (Region Of Practical Equivalence) with HPI (Highest-Posterior Interval)
    - If HPI is within ROPE, no effect: H_1 is rejected
    - If HPI is outside ROPE, there is an effect: H_0 is rejected
    - If HPI overlaps with ROPE, result is inconclusive
- Decide ROPE before analysis based on domain knowledge
  - Picking it after analysis is like picking the p-value threshold after seeing the p-value

Loss Function: Motivation

- You need to make decisions based on our inference
- For many problems, decision cost is asymmetric
  - E.g., cost of a bad decision $>$ benefit of a good decision
  - E.g., vaccines may cause overreaction, but benefits outweigh risks
- To make the best decision, measure:
  - Benefits of a correct decision
  - Cost of a mistake
  - Decide trade-off between benefits and costs using a loss function
  - Use loss we function for decisions
- Loss quantifies *"how bad is an estimation mistake?"*
  - Larger loss indicates worse estimation

Loss Function

- Aka "cost function"
  - The inverse is known as "objective", "fitness", "utility function"
- Use a function to measure the difference between:
  - The true value θ; and
  - The estimated value $\hat{\theta}$

Loss	Expression	Point estimate
Quadratic loss	$(\theta - \hat{\theta})^2$	Mean of posterior
Absolute loss	$\lvert\theta - \hat{\theta}\rvert$	Median of posterior
1-0 loss	$I(\theta \neq \hat{\theta})$	Mode of posterior

- Making decisions in Bayesian statistics using loss function
  - Goal: pick a single value $\hat{\theta}$
  - You don't know the true value θ
  - Estimate θ in terms of the posterior distribution
  - Find the value $\hat{\theta}$ that minimizes the expected loss function