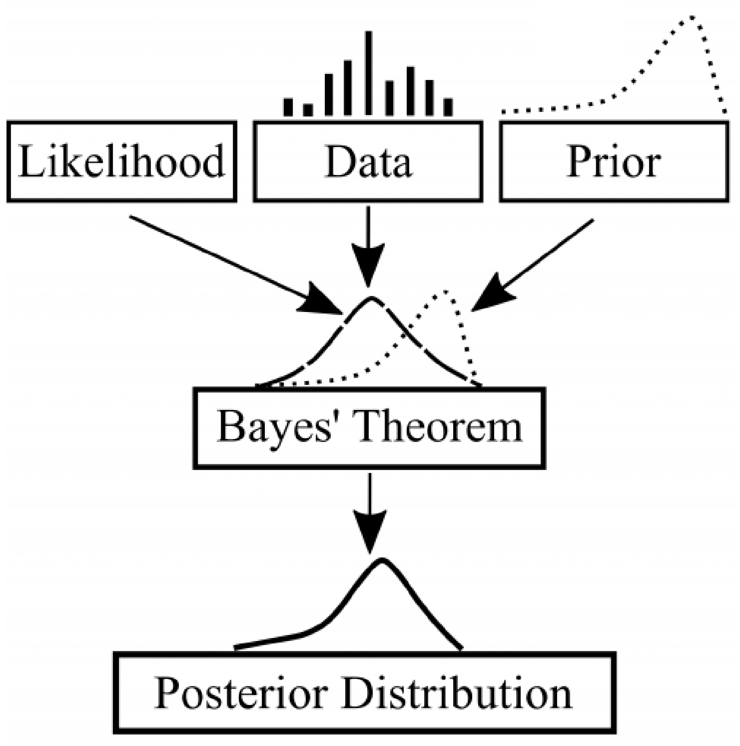
# Bayesian Analysis




PSY504, Cody Dong

# Statistical Inference

We're interested in making inferences about unknown population parameters **θ** given a (probability) model of the data (y).

Start with our model of the data:

$$y_i \sim p(y_i | \boldsymbol{\theta})$$

How is the data distributed?

- Normal: $y_i \sim \text{Normal}(\mu, \sigma^2)$

Question of Interest: What is the population mean $\mu$? Or standard deviation $\sigma$?

**<u>Classical Inference:</u>**

Unknown parameters **θ** are considered to be **fixed values**

- Point estimation: what is the single "best estimate" for **θ**?

## What if we choose the value of **θ** which makes observing our data most **likely**?

**Likelihood:** How likely was it to observe my data based on a specific parameter set (ex: mean of 30, standard deviation of 10)?

$$\Pr(y_i|\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu)^2\right\}$$

← Probability of observing one data point (normal distributed)

$$\text{Likelihood} = \Pr(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \Pr(y_i|\boldsymbol{\theta})$$

← *Joint* probability of observing your data

**Maximum Likelihood Estimate (MLE):**

$$\hat{\theta}_{MLE} = \text{argmax}_{\theta}(\Pr(\boldsymbol{y}|\boldsymbol{\theta}))$$

- For a normal model: $\widehat{\mu}_{\mathrm{MLE}} = \bar{y}$
  - The MLE for the mean is the sample mean

Great! We have our best estimate $\hat{\theta}_{MLE}$

**But** our data was only obtained from random sample from the population.

How does our point estimate $\hat{\theta}_{MLE}$ vary across random samples (uncertainty)?

- **Sampling distributions** of our statistic
  - Requires asymptotic assumptions like central limit theorem (e.g., as our sample size increases…)

- **Confidence Intervals**: 95% confidence interval $(c_1, c_2)$ for $\mu$ does **NOT** mean that $\Pr(\mu \in (c_1, c_2)) = 0.95$

"As we repeat this procedure over the long run, we expect the true (fixed) parameter value to be captured by the confidence interval 95% of the time."

All of these stem from our initial consideration of $\boldsymbol{\theta}$ to be **fixed**

## What if we consider our parameters **θ** can **vary**?

**<u>Bayesian Inference</u>**

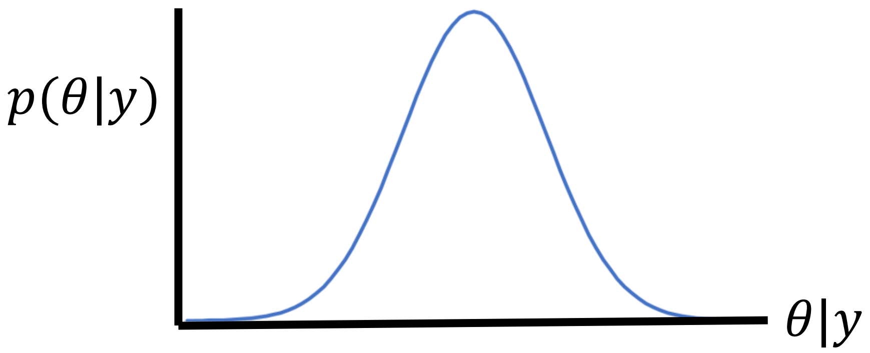
Unknown parameters **θ** are considered as **<u>random variables</u>** (vary and have a distribution)

We're interested in the **posterior distribution of θ**: $p(\boldsymbol{\theta}|\boldsymbol{y})$

- Ex: "Given our observed data, what does the distribution of our population mean look like?"

The posterior distribution is very useful:

- Tells us where the most likely parameter values are located
- Tells us how uncertain we are about what values the parameters could take



# How do we calculate the posterior distribution?

We use **Bayes Rule**:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Posterior Distribution

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta}) * p(\boldsymbol{\theta})}{p(\boldsymbol{y})}$$

Data **Likelihood**

$$\text{Likelihood} = \Pr(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \Pr(y_i|\boldsymbol{\theta})$$

**Prior** Distribution

Before seeing any data, what were your assumptions on how $\theta$ is distributed?

Marginal Likelihood of Data (**Evidence**)

Quantifies agreement between data and prior (*more on this when I talk about Bayes Factors)

$p(y) = \int p(y|\theta) * p(\theta)\, d\theta$ = Some Constant Number (since $\theta$ is the only thing that varies)

## Bayes Rule:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta}) * p(\boldsymbol{\theta})}{p(\boldsymbol{y})}$$

Often rewritten to:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta}) * p(\boldsymbol{\theta})$$

**Posterior distribution** is proportional to **data likelihood** multiplied by the **prior**

**Example:** Coin-Tossing Experiment

You're suspicious about this coin your friend pulled out:

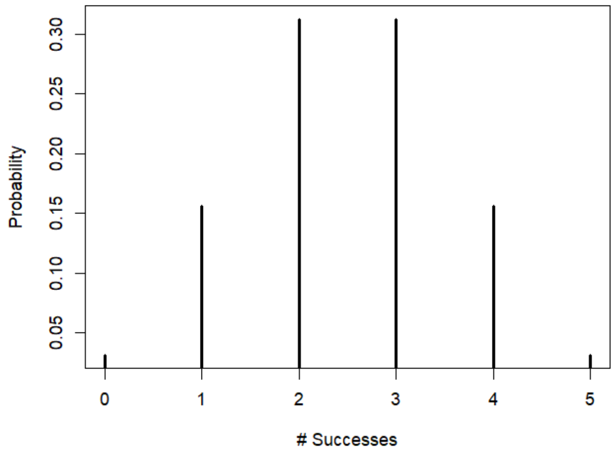


Is this a weighted coin? What is the probability of Heads (Yes)?

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta}) * p(\boldsymbol{\theta})$$

**Binomial Data Likelihood:** $p(\boldsymbol{y}|\boldsymbol{\theta})$

$p(\boldsymbol{y}|\boldsymbol{\theta} = 0.5)$




**Let $\theta$ = probability of heads**

**Let $n = 5$ (we will flip the coin 5 times and observe the results)**

**Example:** Coin-Tossing Experiment

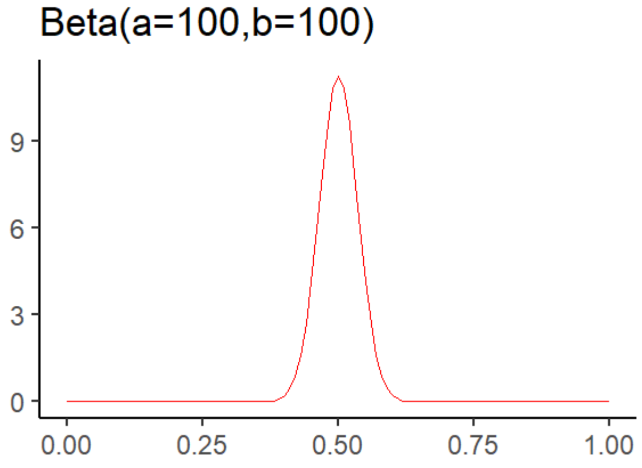

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta}) * p(\boldsymbol{\theta})$$

**Prior Distribution:** $p(\boldsymbol{\theta})$ Choice is up to you!

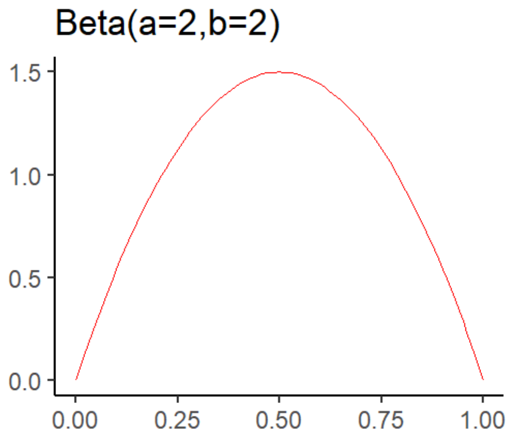
Hint: Beta distribution nicely combines (i.e., is a conjugate prior) with the binomial likelihood

Here is where you can incorporate in your **prior beliefs** about the coin

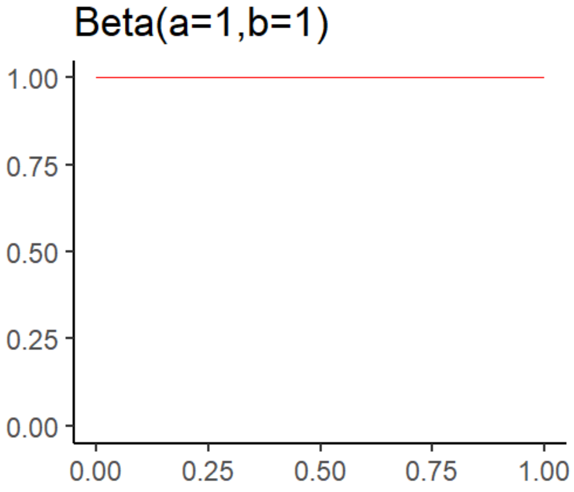
**1)** My friend would never cheat me, the coin is unbiased




**2)** I think the coin is unbiased, but I'm open to other possibilities




**3)** I will refrain from making assumptions; anything is possible

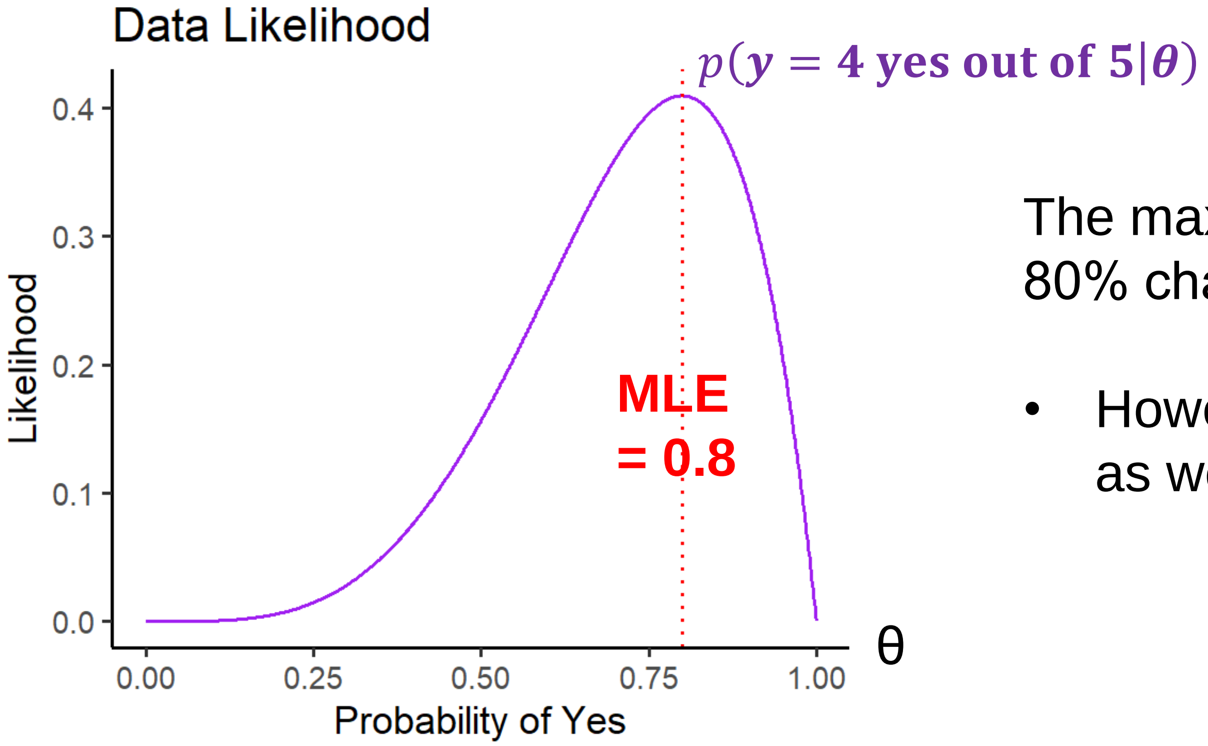


**Example:** Coin-Tossing Experiment


$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta}) * p(\boldsymbol{\theta})$$

Experiment: toss the coin 5 times:


Results: 4 yes; 1 no




The maximum likelihood estimate is as we expect, 80% chance for “Yes/Heads”.

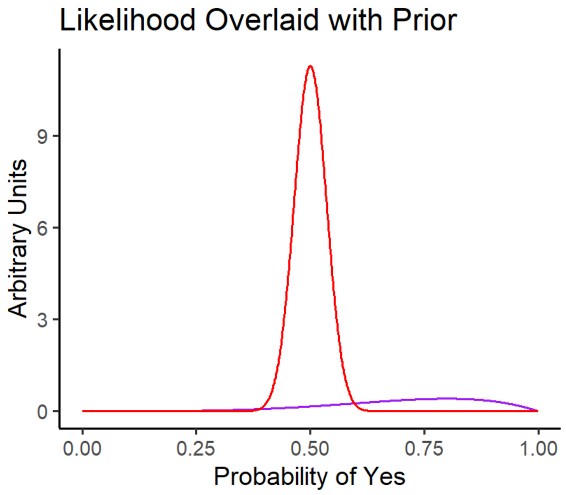
- However, other probabilities for heads seem likely as well

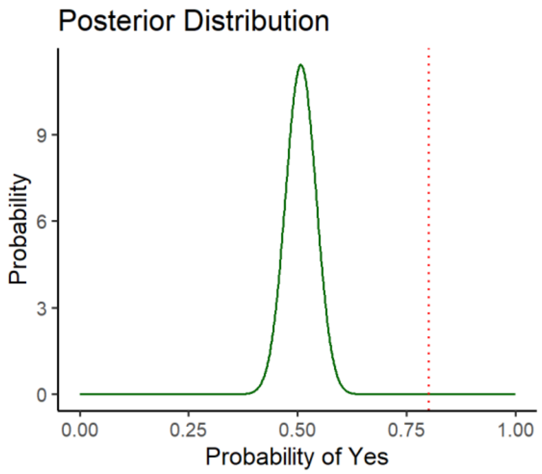
$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta}) * p(\boldsymbol{\theta})$$

**Example:** Coin-Tossing Experiment

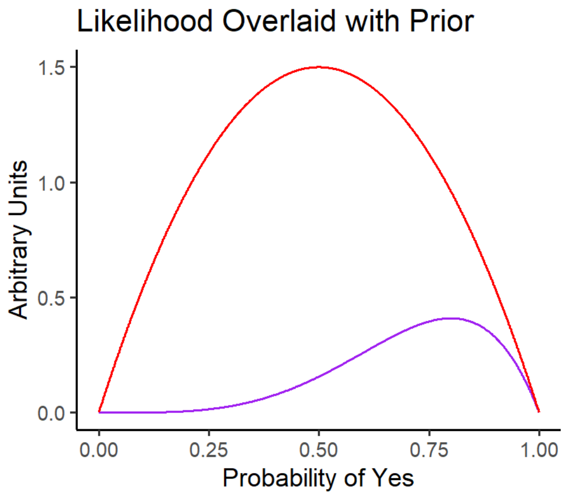

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta}) * p(\boldsymbol{\theta})$$

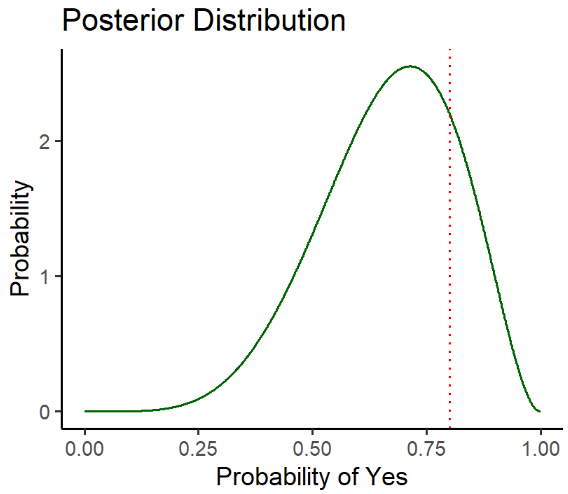
**1)** Prior: very convinced $\theta$=0.5



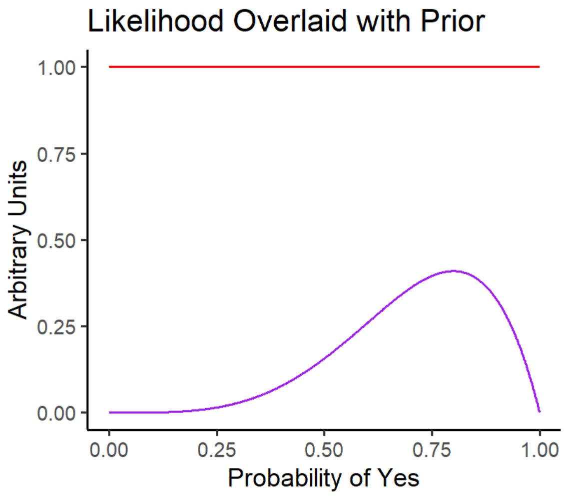



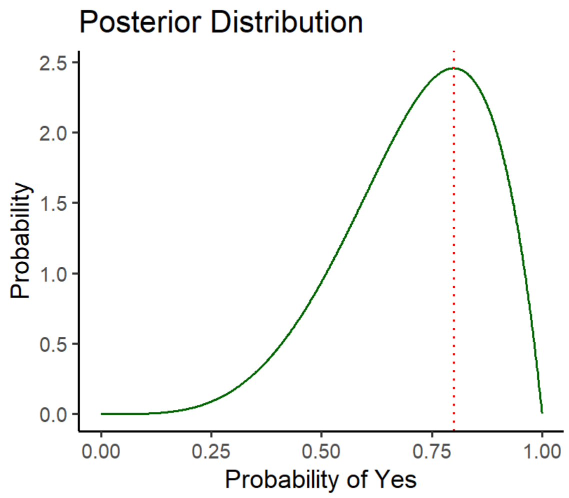

**2)** Prior: loosely convinced $\theta$=0.5



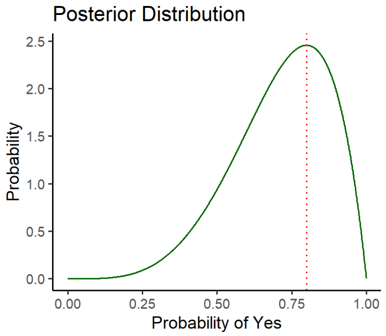

**3)** Prior: uniform; all values equally likely



**Example:** Coin-Tossing Experiment $p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta}) * p(\boldsymbol{\theta})$

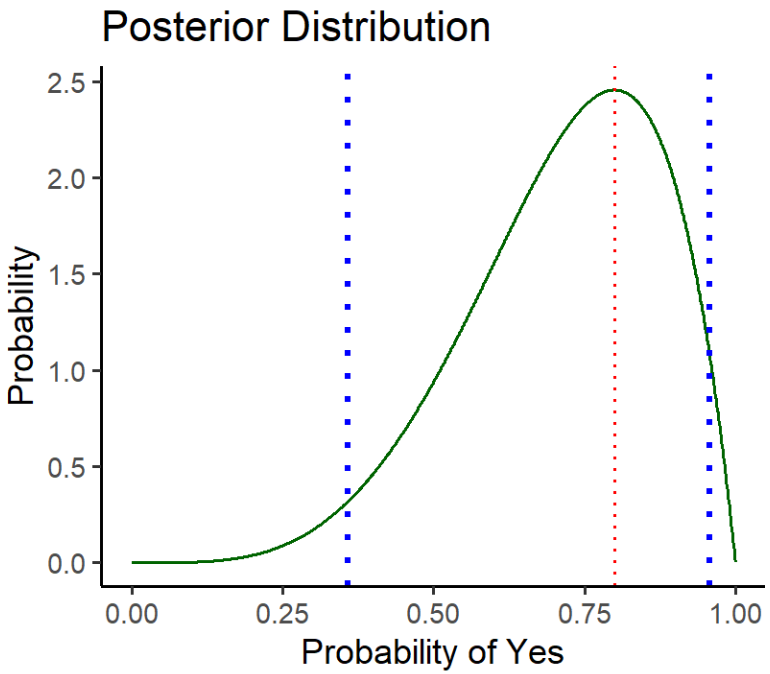
Now that we have the posterior distribution, we have all the information we want!




Ex: we can calculate **credible intervals**:

What interval contains 95% of the possible parameter values?

[0.359, 0.957]



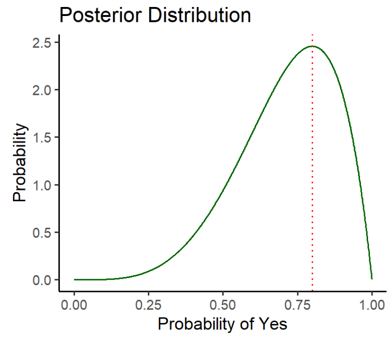

Based on the observed data, there is a 95% probability that the true probability of heads lies between 0.359 and 0.957.

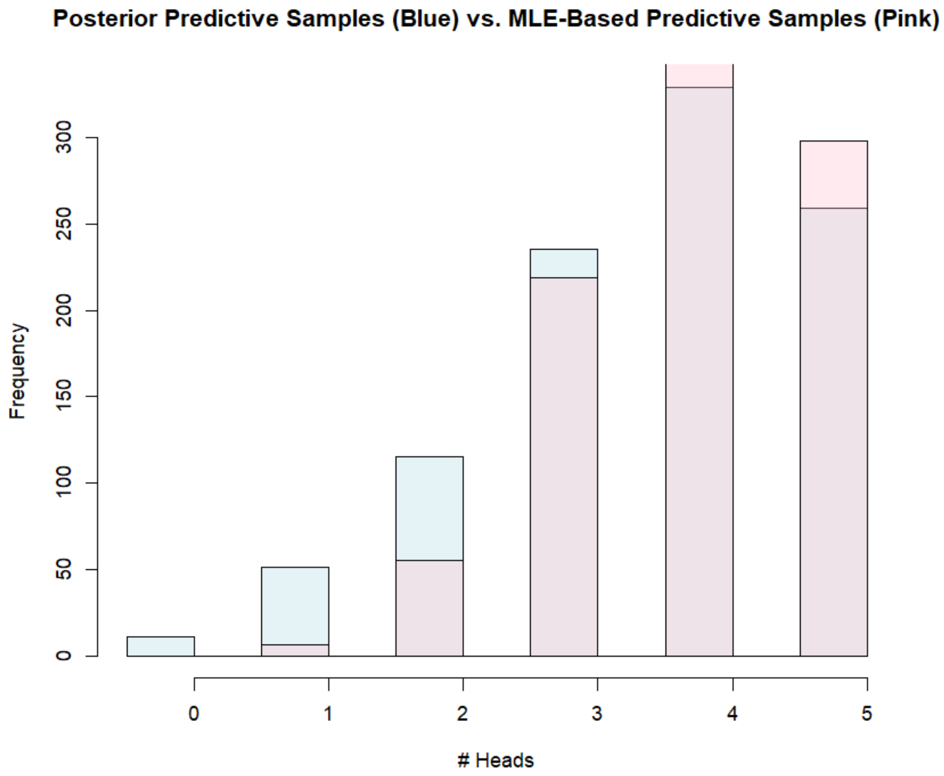
Note: credible intervals are **direct probability statements**!

**Example:** Coin-Tossing Experiment $p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta}) * p(\boldsymbol{\theta})$

Ex: we can calculate **posterior predictive samples**:

Generate sample data, accounting for the fact that there is variability in the estimated probability of heads

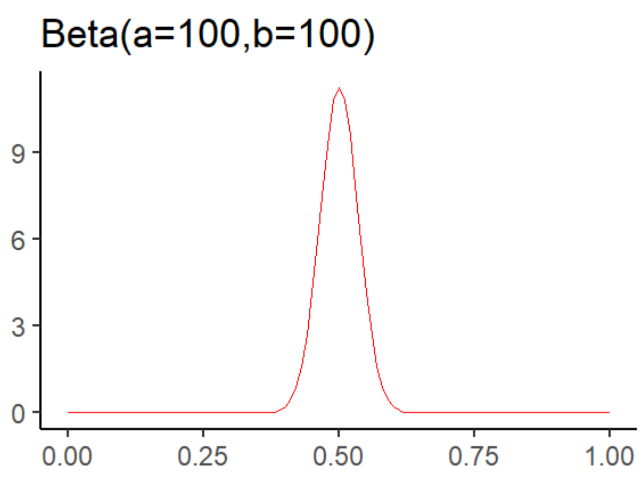


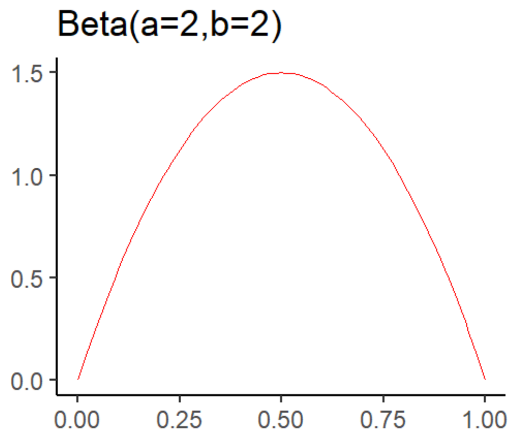

Overall, the light blue **(Bayesian posterior predictive samples) show more variability** because the classical predictive samples assume that the probability of heads is fixed (at the MLE value of 0.8) whereas the Bayesian version does not.

**Principles for Choosing Priors:** The tricky part is you have to choose

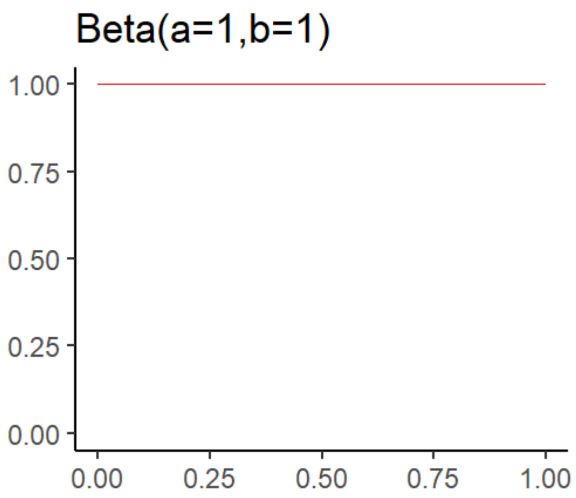
**1)** My friend would never cheat me, the coin is unbiased




**2)** I think the coin is unbiased, but I'm open to other possibilities




**3)** I will refrain from making assumptions; anything is possible




**1)** and **2)** are called **informative priors.** They build in prior information you have, which will affect the posterior

**3)** Is a **non-informative** prior (in this case, the uniform distribution). This is when you don't want to bias the posterior estimates

For this reason in experimental work, **we usually default to non-informative priors** unless you have past information you want to incorporate.

Fun Fact: In the context of regression, LASSO and Ridge regression are equivalent to using informative priors centered around 0

## **Bayes Factors:** Model Comparisons, Hypothesis Testing

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta}) * p(\boldsymbol{\theta})}{p(\boldsymbol{y})}$$

$$p(y) = \int p(y|\theta) * p(\theta)\, d\theta$$

Marginal Likelihood

- Compute $p(y)$ given different models (or priors):

$$p(\theta|y, M) = \frac{p(y|\theta, M) * p(\theta|M)}{p(y|M)}$$

$$\frac{\boldsymbol{p}(\boldsymbol{M_1}|\boldsymbol{y}) = \dfrac{p(y|M_1) * p(M_1)}{p(y)}}{\boldsymbol{p}(\boldsymbol{M_2}|\boldsymbol{y}) = \dfrac{p(y|M_2) * p(M_2)}{p(y)}}$$

$$\Rightarrow$$

$$\frac{\mathrm{p}(\mathrm{M}_1|\mathrm{y})}{\mathrm{p}(\mathrm{M}_2|\mathrm{y})} = \boxed{\frac{\mathrm{p}(\mathrm{y}|\mathrm{M}_1)}{\mathrm{p}(\mathrm{y}|\mathrm{M}_2)}} * \frac{\mathrm{p}(\mathrm{M}_1)}{\mathrm{p}(\mathrm{M}_2)}$$

**Bayes Factor**

**Bayes Factors:**

$$\frac{p(M_1|y)}{p(M_2|y)} = \frac{p(y|M_1)}{p(y|M_2)} * \frac{p(M_1)}{p(M_2)}$$

- Odds Ratio of Posterior Probabilities: how likely model 1 is compared to model 2 given the data
- Ratio of Prior Probabilities: how much did we believe in model 1 compared to model 2 prior to the data
- Ratio of Marginal Likelihoods: **Bayes Factor**
  - **After getting the data, how much should I change or update my prior belief**

**Bayes Factors > 1** would indicate evidence towards Model 1
**Bayes Factors < 1** would indicate evidence towards Model 2

The paper we read put $H_0$ as model 1 and $H_1$ as model 2, so BF > 1 would be evidence for Null

## **Summary:** Pros vs. Cons

Pros:

- Bayesian inference allows you to make **direct probability statements** for things that you are interested in
- It allows you to **incorporate prior information** you have in a formal way (via the prior distribution)
- Don't need sampling distributions and those assumptions (ex: the large sample assumptions for CLT) since the **posterior distribution incorporates the variability in the parameter** $\boldsymbol{\theta}$

Cons:

- Need to specify priors (which may be a pro or a con); Could be dangerous with small sample sizes
  - However, **for large samples, the posterior distribution will be dominated by the data likelihood**, so prior specification is less important
  - Using non-informative priors will pretty much get you similar results to classical analyses

- **COMPUTATION:** Computing posterior estimates because *very* computationally intensive especially for complicated models where you need to use Markov-Chain Monte-Carlo algorithms (e.g., Metropolis-Hastings, Gibbs Sampler, etc.) to get posterior estimates