# Bayesian Estimation

# Bayes Rule

Recall **Bayes Rule**, a fundamental property for conditional probabilities:

$$Pr(A|B) = \frac{Pr(B|A) \times Pr(A)}{Pr(B)}$$

This formula allows us to update our beliefs given new evidence.

# Understanding the Joint Density

The joint density function, often represented as $f_{X_1,...,X_n}(x_1, ..., x_n|\theta)$, can be understood as the likelihood of observing our data given a set of parameters, i.e.

$$f_{X_1,...,X_n}(x_1, ..., x_n|\theta) = \mathcal{L}(\theta|x_1, ..., x_n)$$

This is analogous to the concept of conditional probability, $Pr(A|B)$, where we assess the likelihood of event $A$ occurring given that event $B$ has occurred.
In essence:

$$f_{X_1,...,X_n}(x_1, ..., x_n|\theta)$$

is like the probability of observing our data $X_1, ..., X_n$ given parameters $\theta$.

# Bayesian Estimation

Equating the joint density function $f_{X_1,...,X_n}(x_1, ..., x_n|\theta)$ to conditional probability $Pr(A|B)$ we can use Bayes rule to flip the conditioning. By treating the parameter set $\theta$ as a random variable, we can utilize Bayes rule to derive the density of $\theta$ given the data:

$$\pi(\theta|X_1, ..., X_n) = \frac{f_{X_1,...,X_n}(x_1, ..., x_n|\theta) \times \pi(\theta)}{f_{X_1,...,X_n}(x_1, ..., x_n)}$$

# Bayesian Estimates vs. Traditional Estimates

While traditional methods like Maximum Likelihood Estimation yield a single estimate, Bayesian approaches offer a probability distribution for unknown parameters.

**Example**: Suppose you are estimating $\mu$ and your data is $\{8, 9, 10\}$. An MLE might will give you $\hat{\theta}_{MLE} = 9$ while a Bayesian estimate gives you $\pi(\mu|\{8, 9, 10\})$, a whole distribution for the answer.

# Dissecting the Bayesian Formula

Understanding the parts of our Bayesian formula:

$$\pi(\theta|X_1, ..., X_n) = \frac{f_{X_1,...,X_n}(x_1, ..., x_n|\theta) \times \pi(\theta)}{f_{X_1,...,X_n}(x_1, ..., x_n)}$$

- $\pi(\theta|X_1, ..., X_n)$: Posterior - updated beliefs about parameters given data.
- $f_{X_1,...,X_n}(x_1, ..., x_n|\theta)$: Likelihood - how well our model explains the observed data.
- $\pi(\theta)$: Prior - initial beliefs about the parameters.
- $f_{X_1,...,X_n}(x_1, ..., x_n)$: Evidence - how probable our observed data is under all possible parameter values.

# Computation of Bayesian Estimates

In some cases we can use math to find the posterior distribution of the unknown parameters. Not only is that hard, but in some cases it's impossible. There is no closed form solution. Software uses the Markov chain Monte Carlo (MCMC) method to return the posterior distribution. MCMC provides an **empirical distribution** of the posterior distribution.

# Empirical Distributions: An Analogy

Consider a known distribution:

$$P(x) = \begin{cases} \frac{1}{4} & \text{if } x = -1 \\ \frac{1}{2} & \text{if } x = 0 \\ \frac{1}{4} & \text{if } x = 1 \end{cases}$$

Drawing a random value from this distribution is akin to randomly picking from the set $\{-1, 0, 0, 1\}$. This set is an empirical representation of our distribution.
Empirical distributions, especially when sufficiently large, can serve as close approximations to their continuous counterparts.

# Power of Empirical Distributions

With a sample empirical distribution like $\{1, 2, 3, 4, 5, 5, 5, 5, 5, 6\}$, we can:

- Determine the mean (expected value).
- Calculate probabilities.
- Examine the distribution of transformations of the data.

This means that MCMC's output - a large empirical distribution - can be used to infer characteristics about the underlying probability distribution.

# Working with Empirical Distributions

Given the empirical distribution: $1, 2, 3, 4, 5, 5, 5, 5, 5, 6$,

- **Mean (Expected Value):**

$$\frac{1+2+3+4+5+5+5+5+5+6}{10} = 4.1$$

- **Probability of a Value (e.g., $P(X \leq 4)$):**

$$\frac{\text{Number of observations} \leq 4}{\text{Total number of observations}} = \frac{4}{10} = 0.4$$

- **Transformation (e.g., $X^2$):**
  Resulting distribution: $1, 4, 9, 16, 25, 25, 25, 25, 25, 36$

Empirical distributions provide a hands-on way to understand and analyze data distributions without complex calculations.

# Linear Models in Bayesian Framework

The standard linear regression model:

$$Y_j \sim N(\beta_0 + \beta_1 x_{1j}, \sigma)$$

becomes Bayesian when we treat its parameters as distributions, updated with the data we have. MCMC will generate empirical distributions for $\beta_0$, $\beta_1$, and $\sigma$.

# Bayesian Linear Regression Framework

Considering our linear regression model:

$$Y_j \sim N(\beta_0 + \beta_1 x_{1j}, \sigma)$$

The Bayesian posterior distribution for the parameters is:

$$\pi(\beta_0, \beta_1, \sigma | y_1, ..., y_n) = \frac{f_{y_1,...,y_n}(y_1, ..., y_n | \beta_0, \beta_1, \sigma) \times \pi(\beta_0, \beta_1, \sigma)}{f_{y_1,...,y_n}(y_1, ..., y_n)}$$

# Introduction

- We'll evaluate a Bayesian Linear Model using the 'bambi' and 'arviz' packages.
- The syntax is different than those used in the scikit-learn models

# Initialize and Fit the Bayesian Linear Model

```
import bambi as bmb
import arviz as az

model = bmb.Model('y ~ x1 + x2',data)
results=model.fit(draws=1000,chains=4)
```

- Setting up a model with $X_1$ and $X_2$ predicting $Y$.
- Run 4 MCMC chains to assess the model's convergence. Because it is a numerical method, running multiple chains helps ensure the method works as intended.

# Visualize Initial Results

```
az.plot_trace(results)
plt.show()
```

- Recall that the posterior distribution is basically a bunch of numbers that forms an empirical distribution
- Plots the posterior distribution for each parameter as a density plot
- Also plots something called a trace plot, which is a plot of the posterior distribution values by algorithm iteration number. This should look straight. If it slopes up or down or has stretches where it dips or rises the MCMC may have failed.

# Summary Statistics

```
az.summary(results)
```

- Provides a summary of the posterior distribution
- Insight into mean, sd, and other statistics of the posterior distribution for each parameter.

# Linear Models in Bayesian Framework

Challenges specific to Bayesian Linear Regression:

- **Specifying Priors:** We must define priors for each parameter: the intercept ($\beta_0$), the slope ($\beta_1$), and the error variance ($\sigma$).
- **Evidence Calculation:** The denominator, which represents the likelihood of the data over all parameter values, can be intricate. However, the MCMC technique simplifies this by sampling from the posterior directly without explicitly calculating the denominator.

# Deep Dive: The Role of Priors in Bayesian Analysis

- **Definition:** A *prior* represents our existing knowledge or beliefs about a parameter *before* seeing the current data. It is a probability distribution over possible parameter values.
- **Importance:** Priors allow us to incorporate external information, expert opinion, or results from previous studies into our current analysis.
- **Types of Priors:**
  - *Informative Prior:* Based on specific, known information.
  - *Non-informative or Weak Prior:* Represents a lack of strong initial beliefs. Often broad or flat, implying all parameter values are equally likely.
- **Influence on Posterior:** A strong prior can significantly influence the posterior, especially when the data is limited or weak.
- **Choice Matters:** Rationale for selecting a particular prior should be clearly articulated and justified.

# Art of Choosing Priors

Priors give our model a "starting point" based on existing knowledge. Always using normal priors in this class:

$$\beta_0 \sim N(\mu_{\beta_0}, \sigma_{\beta_0})$$

$$\beta_1 \sim N(\mu_{\beta_1}, \sigma_{\beta_1})$$

Where:

- $\mu_{\beta_i}$: Best initial guess, based on prior information but not the current data.
- $\sigma_{\beta_i}$: Uncertainty around our guess. Larger values indicate greater uncertainty.

# Guidelines for Priors

General tips for defining priors:

- Best guesses ($\mu_{\beta_i}$) should be based on logical reasoning, scientific understanding, or prior research.
- $\sigma_{\beta_i}$ conveys confidence in the initial guess.
- A common approach: set values so there's high confidence that the true parameter lies within $\mu_{\beta_i} \pm 2\sigma_{\beta_i}$.

Remember: $\beta_0$ and $\beta_1$ in simple linear regression represent the intercept and slope. This knowledge can guide our choices for priors.

# The $\sigma$ Prior

**The Prior for $\sigma$:**

- In Bayesian linear regression, not only do we have priors for our regression coefficients, but also for the error variance $\sigma^2$.
- Our software of choice, *bambi*, comes with default priors for both $\beta$s and $\sigma$.

**A Little Secret:** These default priors, though convenient, use the data *twice* - once in defining the prior and once in the likelihood. In the statistical world, that's kinda like looking at your cards twice in a game of poker... a tad bit cheeky!

**Our Approach:** While we'll meticulously craft our priors for the $\beta$s, we'll turn a blind eye (just this once!) and use the default for $\sigma$. Let's just call it our little statistical indulgence!

*Remember:* It's all fun and games until someone double dips in the data!

# Choosing Priors: High Certainty Scenario

**Research Context:** Imagine we're investigating the relationship between years of education and annual salary. From previous robust studies, we're quite certain about the following:

- Initial salary (with no education) is roughly around \$20,000.
- Each additional year of education increases salary by approximately \$5,000.

**Priors:**

$$\beta_0 \sim N(20,000, 5,000)$$
$$\beta_1 \sim N(5,000, 1,000)$$

These priors reflect strong beliefs based on previous research, with a relatively small standard deviation indicating our confidence.

# Interpreting the High Certainty Priors using the 2-SD Rule

**Intercept,** $\beta_0$:

- Mean ($\mu_{\beta_0}$): \$20,000
- Standard Deviation ($\sigma_{\beta_0}$): \$5,000
- 95% Confidence Interval: \$10,000 to \$30,000

Given the 2-SD rule, we believe there's a high probability that the true intercept lies between \$10,000 and \$30,000.

**Slope,** $\beta_1$:

- Mean ($\mu_{\beta_1}$): \$5,000
- Standard Deviation ($\sigma_{\beta_1}$): \$1,000
- 95% Confidence Interval: \$3,000 to \$7,000

Again, based on the 2-SD rule, we believe there's a high probability that the true slope lies between \$3,000 and \$7,000.

# The Pitfalls of Poorly Chosen Priors

**What could go wrong?**

- **Choosing a Bad Mean:** If we incorrectly believe that the initial salary without education is around \$40,000 (based on wrong data or assumptions), our entire model could be biased.
- **Being Overconfident:** If we were to choose a very small standard deviation (say \$500 for $\beta_0$ and \$100 for $\beta_1$), we would be expressing undue confidence in our priors. This can overshadow the data, leading to skewed or incorrect inferences.

**Bad Priors Example:**

$$\beta_0 \sim N(40,000, 500)$$
$$\beta_1 \sim N(5,000, 100)$$

Such priors might not only bias the model but also restrict its ability to learn effectively from the data.

# Choosing Priors: Low Certainty Scenario

**Research Context:** Now, let's consider a study examining the relationship between consumption of a new health supplement and improvement in cognitive test scores ranging from 0 to 100. There's limited prior research, and findings are mixed.

**Priors:** Given the uncertainty:

- We're unsure about the base cognitive score without the supplement.
- We also don't have a clear estimate for the effect of the supplement.

**Priors:**

$$\beta_0 \sim N(50, 25)$$
$$\beta_1 \sim N(0, 50)$$

These priors reflect our lack of strong initial beliefs, with wider standard deviations indicating greater uncertainty.

# Deep Dive into Low Certainty Priors

**Intercept,** $\beta_0$:

- **Mean ($\mu_{\beta_0}$):** 50
  Given the cognitive test scores range from 0 to 100, we're making a neutral assumption that without the supplement, the average score is mid-scale at 50.
- **Standard Deviation ($\sigma_{\beta_0}$):** 25
  This large value indicates our substantial uncertainty about the base cognitive score without the supplement. The wide spread suggests scores could vary significantly.
- **2 SD Interval:** 0 to 100
  Based on the 2-SD rule, we believe there's a high likelihood that the actual base cognitive score without the supplement lies between these scores.

# Deep Dive into Low Certainty Priors

**Slope,** $\beta_1$:

- **Mean ($\mu_{\beta_1}$):** 0
  Given the mixed findings and lack of clear evidence, we're assuming that, on average, the supplement might not have any effect. Hence, the mean is set to zero.
- **Standard Deviation ($\sigma_{\beta_1}$):** 50
  This value shows that we are very uncertain about the supplement's effect on the cognitive scores. It could be beneficial, detrimental, or neutral.
- **95% Confidence Interval:** -100 to 100
  Based on the 2-SD rule, we believe the effect of the supplement could lie anywhere in this wide range. The supplement might drastically improve, degrade, or have no effect on scores.

# Interpreting Coefficients: Key to Setting Priors

**Understanding Coefficients:** The meaning of regression coefficients is foundational for constructing informed priors. Priors should align with our beliefs and knowledge about the relationships being modeled.

**Coefficient Interpretations:**

- **Slope in Simple Linear Regression:** Represents the change in the response variable for a one-unit increase in the predictor. E.g., for every additional year of education, how much does the annual salary change?
- **Intercept:** The expected value of the response variable when all predictors are zero. E.g., the expected annual salary for someone with no years of education.
- **Coefficient in Multiple Regression:** Represents the change in the response for a one-unit increase in the predictor, holding all other predictors constant.

# Interpreting Coefficients: Key to Setting Priors

- **Binary Variable Coefficient:** Indicates the difference in the expected response between the two categories of the binary variable. E.g., the difference in test scores between those who took a supplement and those who didn't.
- **Dummy Variable for Categorical Variable:** Compares the mean of the response for each category to the mean of the reference category. If coding "season" with "Spring" as reference, the coefficient for "Summer" indicates the difference in the response between Summer and Spring.

Remember: Properly interpreting coefficients aids in setting meaningful and appropriate priors, ensuring the model aligns with our domain knowledge.

# Setting Custom Priors

```
priors = {
"Intercept": bmb.Prior("Normal", mu=[MuValue],
    sigma=[SigmaValue]),
"x1": bmb.Prior("Normal", mu=[MuValue], sigma=[
    SigmaValue]),
"x2": bmb.Prior("Normal", mu=[MuValue], sigma=[
    SigmaValue])
}
model.set_priors(priors=priors)
results = model.fit(draws=1000, chains=4)
```

- We create a dictionary of priors. Use "Intercept" for $\beta_0$ and the variable name for other variables.
- We apply the priors using `model.set_priors`
- The rest works the same as before

# Posterior Analysis

```
posterior = results.posterior.stack({"draws":["
    chain","draw"]})
variable_post = posterior["x1"].values
```

- If you are interested in the posterior distribution of $\beta_1$, this will give you an array of these values to work with.
- From this you can plot, evaluate your own summaries, or find probabilities. Again, this is an empirical distribution and you can do anything you want with those.

```
# Find the probability that beta_1 is less than
    0
print(sum(variable_post < 0)/len(variable_post))
# Find the probability that beta_1 is greater
    than 0
print(sum(variable_post > 0)/len(variable_post))
```

# Why Use a Bayesian Model?

**Advantages of Bayesian Approaches:**

- **Incorporate Prior Knowledge:** Bayesian models uniquely allow the integration of prior beliefs or results from previous studies, making them especially powerful in fields with accumulated domain knowledge.
- **Improved Predictions with Good Priors:** A well-chosen prior can refine and improve model predictions, particularly when data is sparse or noisy.
- **Maintains Probability Structure:** Bayesian methods provide full probability distributions for parameters, not just point estimates, preserving the inherent uncertainty.
- **Flexibility with Complex Models:** Bayesian methods can be adapted to intricate hierarchical structures and non-standard models, offering flexibility that some classical methods can't match.

# Why Use a Bayesian Model?

- **Parameter Uncertainty:** Bayesian models provide a natural framework for accounting for and propagating parameter uncertainty, making predictions and inferences more robust.
- **Transparent Assumptions:** Bayesian methods make assumptions explicit through priors, enabling clear scrutiny and understanding of the model's foundations.
- **Adaptive Learning:** As more data becomes available, Bayesian models can be updated iteratively, making them adaptable in dynamic environments.

In essence, while Bayesian models require careful consideration of priors, their benefits in terms of flexibility, transparency, and ability to incorporate prior information make them a powerful tool in a data scientist's toolbox.