



# Bayesian analysis for Data Science

*Externado para la vida*

Mathematics Department



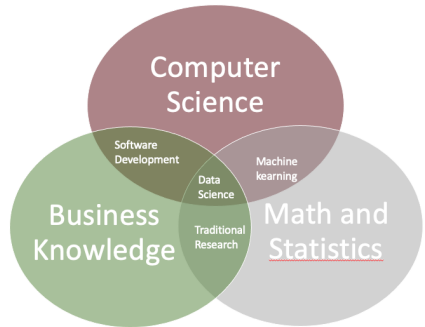
- Module Logistics.
- What is Data Science?
- Emergence of Data Science.
- Navigating a project.
- Data Science Methodology.
- Deductive and Inductive Reasoning.
- Trade-off in Data Science.
- Bayesian Inference.

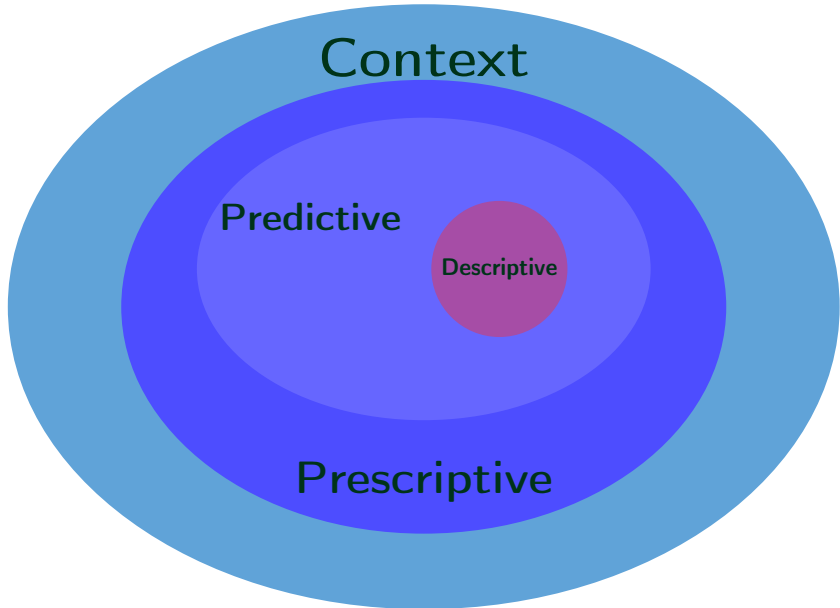


- Bayesian rule and main properties
- Univariate models: discrete and continuous
- Multivariate models and variable dependence
- Hierarchical models and Bayesian parameter estimation (Supervised models)
- Clustering analysis and Bayesian non parametric estimation (Non Supervised models)

# Data Science

- 'Data science is the scientific practice of extracting knowledge from data, usually referring to large and complex bodies of data' Pietsch (2022, p. 2).
- 'Data science could be defined as an interdisciplinary knowledge to understand the real world, using massive amounts of data to create technological artifacts which provide information to make decisions in different fields ' Andrés Martínez.





## Key Components of a Data Science Project

Most sensitive project metrics.



**Metrics**

Define the initial questions.



**Questions**

Sources of data collection.



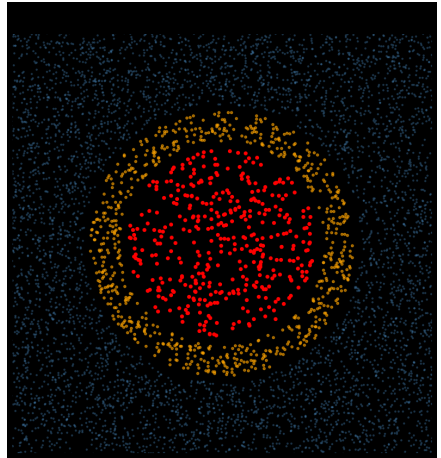
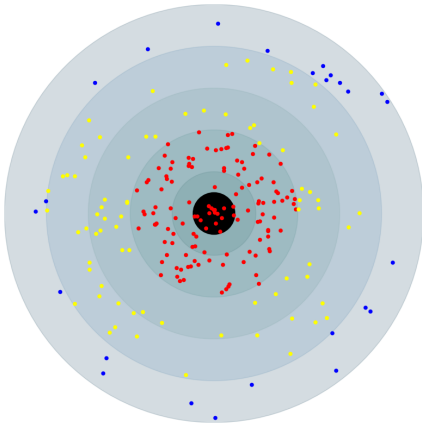
**Data**

What I cannot control.



**Factors**

# Decision Making Process



# Traditional Statistics vs. Data Science

## Traditional Statistics (Deductive Method)

- Deductive approach.
- Defines hypotheses a priori.
- Selects statistical techniques.
- Collects specific data.
- Tests hypotheses.
- Conclusions based on hypotheses.
- Uses representative samples.

## Data Science (Inductive Method)

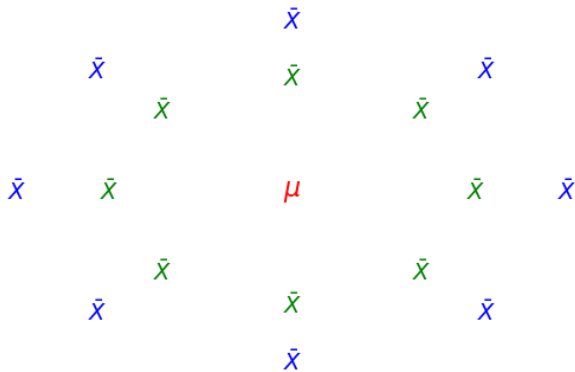
- Inductive approach.
- Explores data without bias.
- Collects large volumes of data.
- Discovers patterns and relationships.
- Conclusions emerge from the data.
- Does not rely on representative samples.



# Parameter Estimation

Estimaciones de  $\hat{\mu}$  con Diferente Precisión

- Unbiasedness
- Efficiency
- Consistency



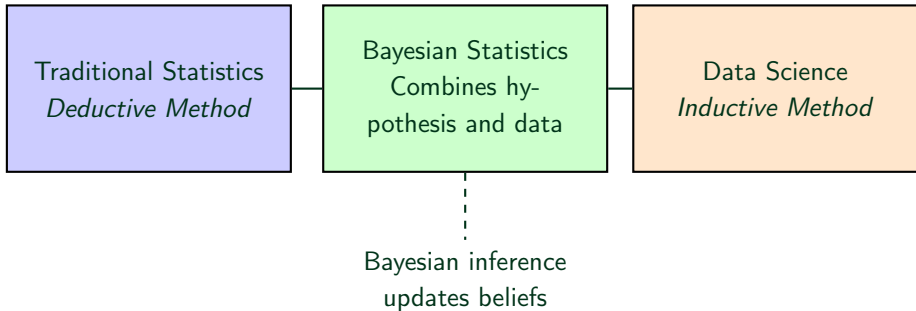
# How Important Are These Properties?

- **Unbiasedness:** Focused on expected accuracy across samples.
- **Efficiency:** Seeks minimal variance among unbiased estimators.
- **Consistency:** Ensures estimates converge to the true value as data increases.

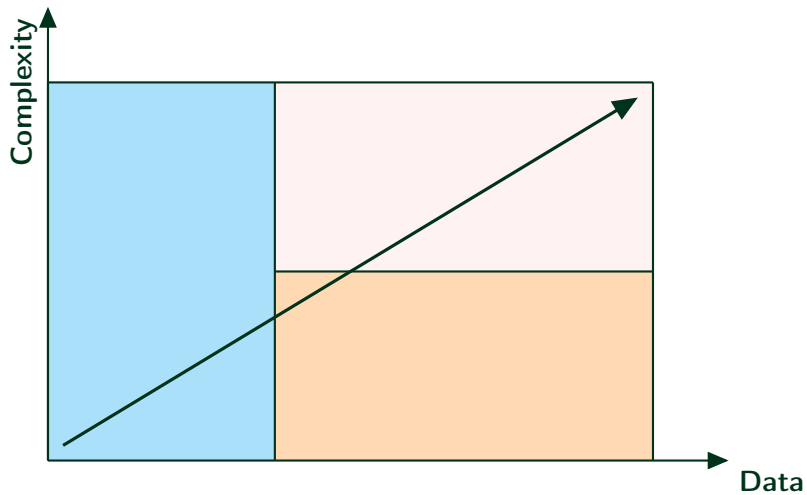
## Their importance across contexts:

Property	Frequentist	Bayesian	Data Science
Unbiasedness	Core principle	Not central	Often sacrificed
Efficiency	Optimality goal	Less emphasized	Secondary to speed
Consistency	Highly valued	Depends on priors	Desirable, not critical

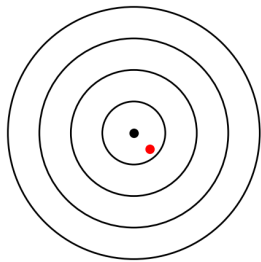
# Bayesian Statistics as a Bridge Between Deduction and Induction



# Model Complexity

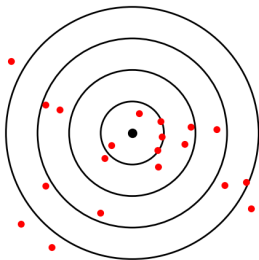


# Frequentist, Bayesian, and Data Science Perspectives



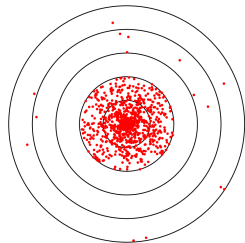
## Frequentist

Small sample, high variance



## Data Science

Massive data, exploratory,  
patterns emerge



## Bayesian

Prior + data = precise  
inference

# Frequentist vs. Bayesian Statistics

## Frequentist Statistics

- Parameters are fixed but unknown.
- Data are random; once observed, inference is done.
- No probability is assigned to parameters.
- Requires large, representative samples.
- Inference relies on sampling distributions.
- Hypotheses are tested using p-values and confidence intervals.
- Cannot formally include prior knowledge.
- Point estimates and interval estimates.

## Bayesian Statistics

- Parameters are random variables with prior distributions.
- Data are fixed once observed; beliefs are updated.
- Assigns probability distributions to parameters.
- Can work with small or incomplete samples.
- Combines prior knowledge and data into posterior.
- Inference via credible intervals and posterior probabilities.
- Prior information can be subjective or objective.
- Full probability model of uncertainty.

# Classical vs. Bayesian Estimation

## Classical Estimation (Frequentist)

- Parameters are fixed and unknown.
- Likelihood is a function of  $\theta$  given  $x$ .
- Uses Maximum Likelihood Estimation (MLE):

$$\hat{\theta} = \frac{x}{n}$$

- Evaluates estimator properties: bias, variance, consistency.
- No probability assigned to  $\theta$ .

## Bayesian Estimation

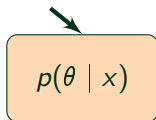
- Parameters are random variables.
- Data are fixed once observed.
- Inference is based on Bayes' Theorem:

$$p(\theta | x) = \frac{p(x | \theta) \cdot p(\theta)}{p(x)}$$

- Prior knowledge is incorporated through  $p(\theta)$ .
- Result is a full posterior distribution.

# Bayes Theorem Components

Posterior Distribution



A diagram illustrating the components of Bayes' Theorem. It features three colored boxes: an orange box on the left containing the expression  $p(\theta | x)$ , a purple box in the middle containing  $p(x | \theta)$ , and a gray box on the right containing  $p(\theta)$ . An arrow points from the text 'Posterior Distribution' to the orange box. Below the orange box is the text 'Likelihood Distribution'. Below the purple box is the text 'Prior Distribution'. A symbol  $\propto$  is placed between the orange and purple boxes. Arrows also point from the text 'Likelihood Distribution' to the purple box and from the text 'Prior Distribution' to the gray box.

$$p(\theta | x)$$

$\propto$

$$p(x | \theta)$$

$$p(\theta)$$

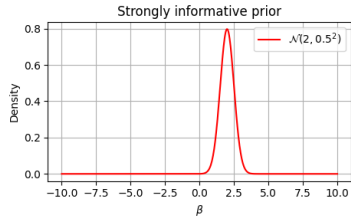
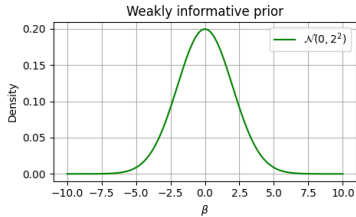
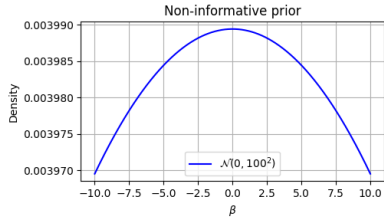
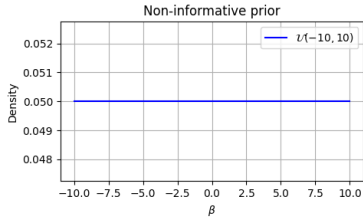
Likelihood Distribution

Prior Distribution

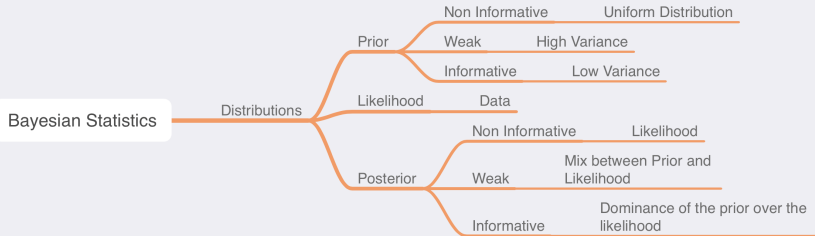
**Figure 1:** Bayes Theorem Components



# Priors Distribution



# Summary

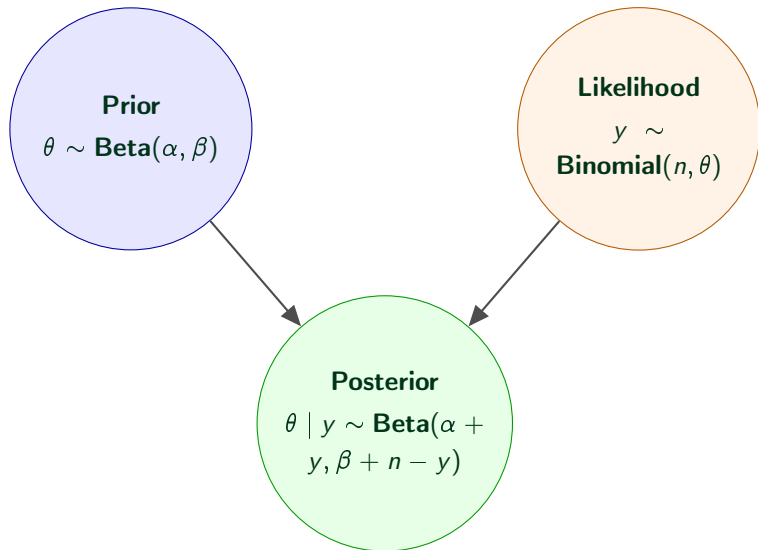


# Conjugate Priors: Beta Distribution Case

- A conjugate prior is one in which the posterior has the same distributional form as the prior.
- This allows for closed-form analytical solutions and efficient computation.
- A classic case: the Beta prior with a Binomial likelihood.

**Model:**  $Y \mid \theta \sim \text{Bin}(n, \theta)$  with prior  $\theta \sim \text{Beta}(a, b)$ .

# Bayesian Inference for Binomial Model



# Expected Value and Variance

## Prior

$$\mathbb{E}[\theta] = \frac{\alpha}{\alpha + \beta}$$

$$\text{Var}[\theta] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

## Likelihood

$$\mathbb{E}[y] = n\theta$$

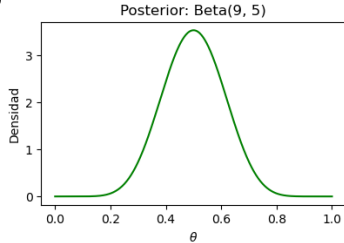
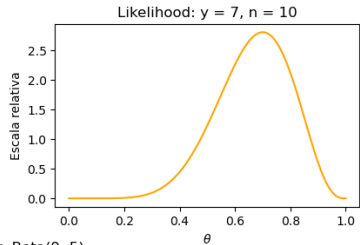
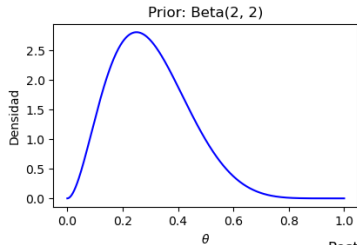
$$\text{Var}[y] = n\theta(1 - \theta)$$

## Posterior

$$\mathbb{E}[\theta \mid y] = \frac{\alpha + y}{\alpha + \beta + n}$$

$$\text{Var}[\theta \mid y] = \frac{(\alpha + y)(\beta + n - y)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)}$$

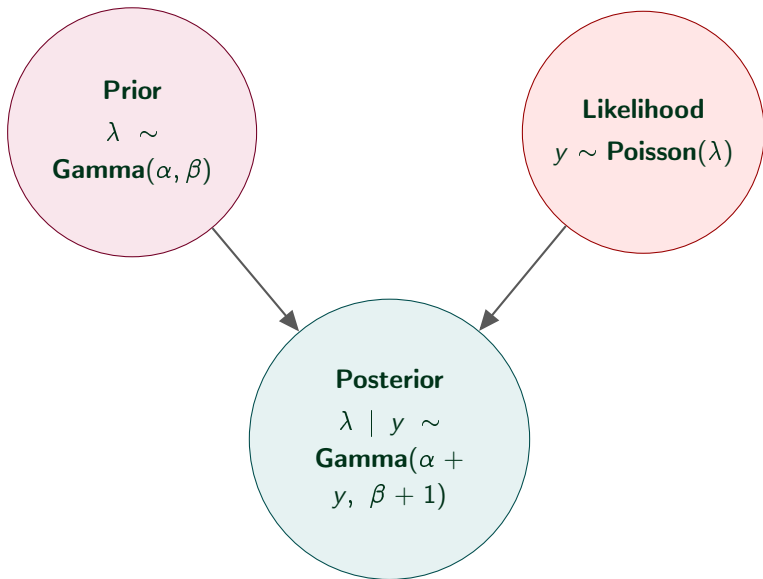
# Densities



# Poisson Model: Motivation

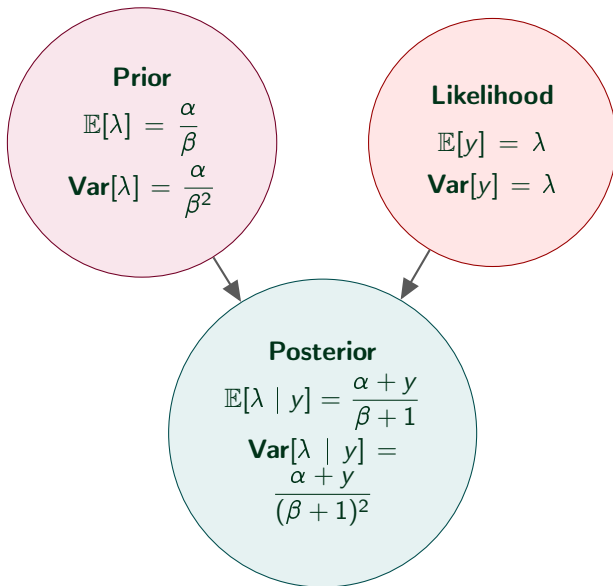
- Used for count data (e.g., number of arrivals, defects, events).
- If  $Y \sim \text{Poisson}(\theta)$ , then:
  - $P(Y = y \mid \theta) = \frac{\theta^y e^{-\theta}}{y!}$
  - $\mathbb{E}[Y] = \text{Var}(Y) = \theta$
- Poisson is appropriate when events are independent and occur at a constant average rate.

# Gamma-Poisson Conjugacy

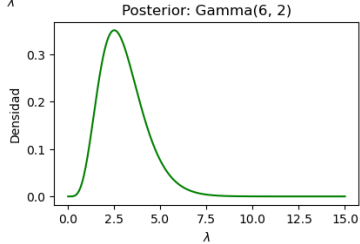
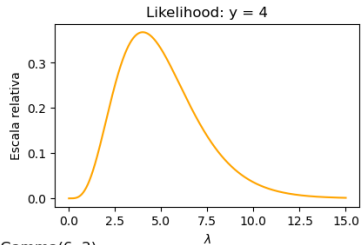
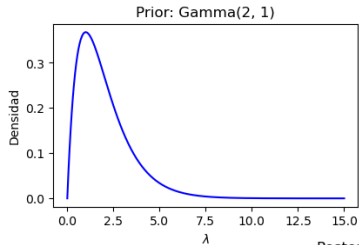




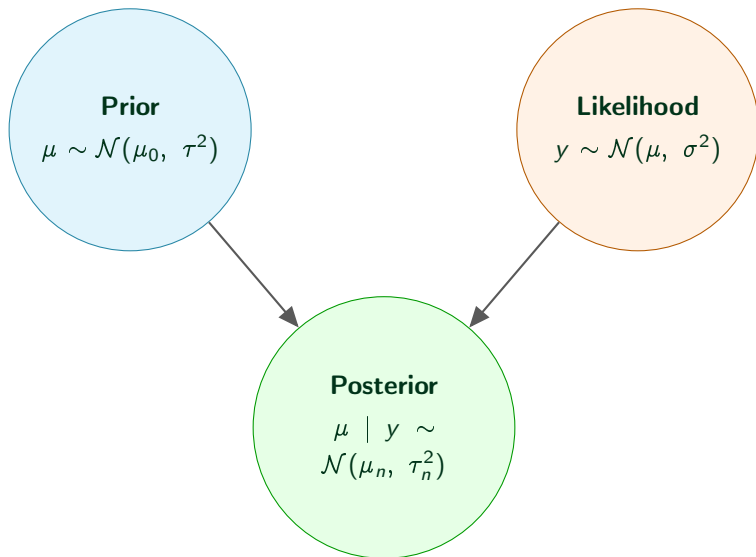
## Expected Value and Variance



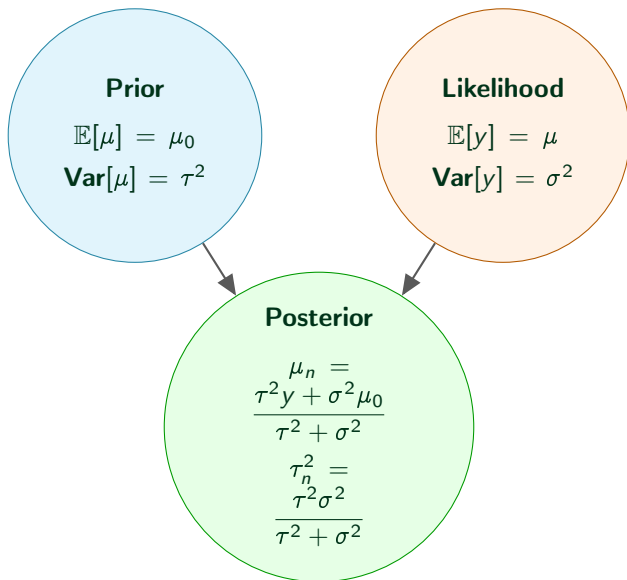
# Densities



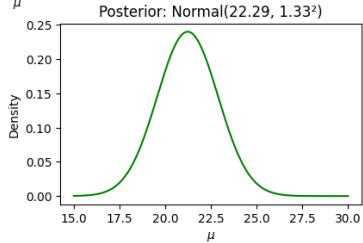
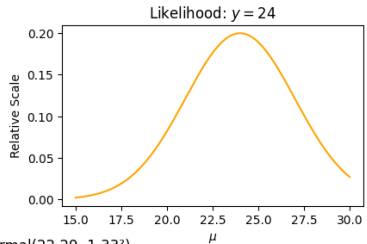
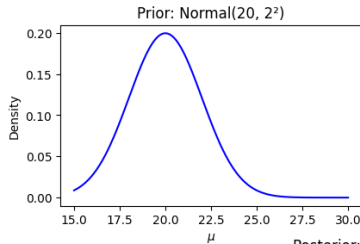
# Normal-Normal Conjugacy



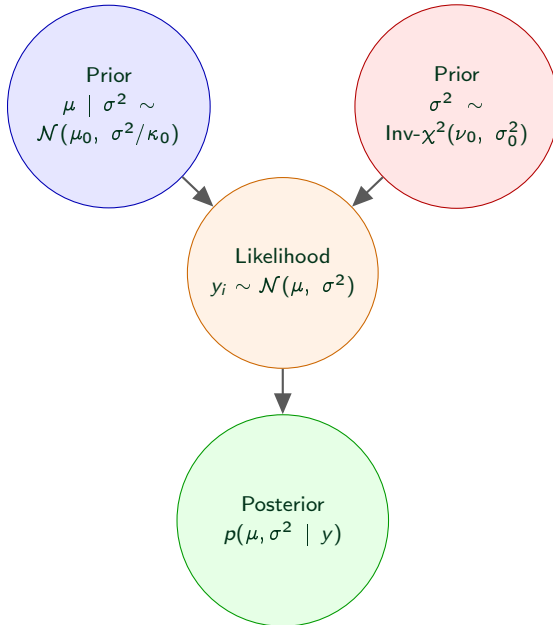
# Expected Value and Variance



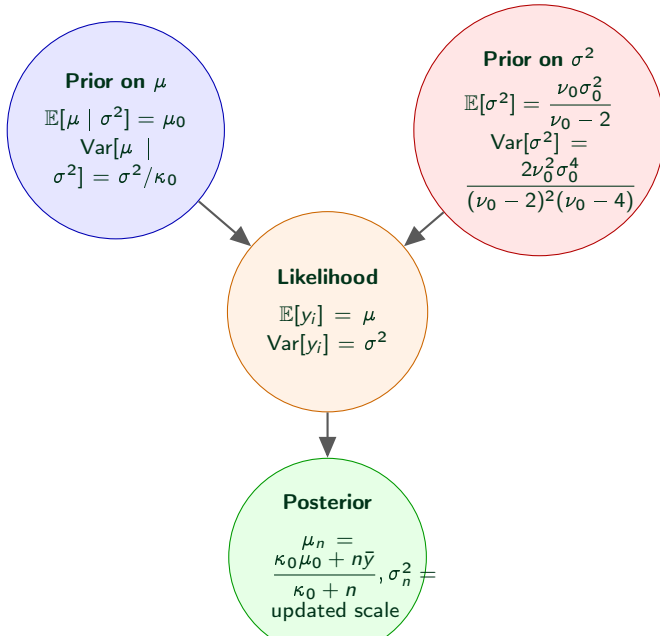
# Densities



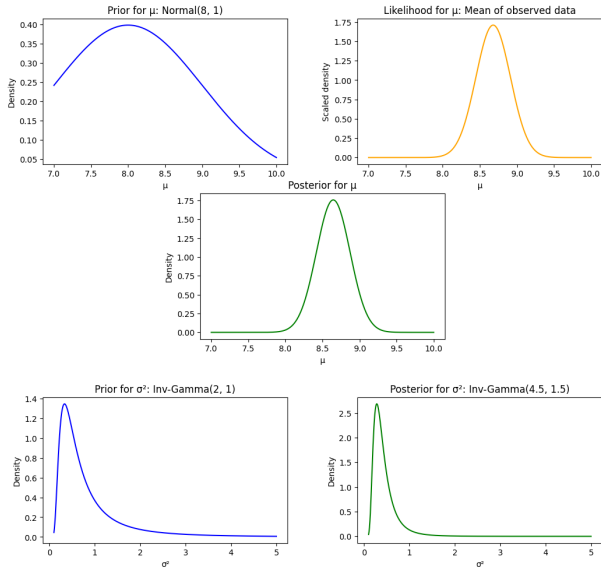
# Normal with Unknown Variance



# Expected Value and Variance



# Densities





# What is Linear Regression?

- A method to model the relationship between a dependent variable  $y$  and one or more explanatory variables  $x_1, x_2, \dots, x_p$ .
- The simple linear regression model assumes:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where  $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ .

- Goal: Estimate  $\beta_0, \beta_1$  to predict  $y$  or understand the effect of  $x$  on  $y$ .
- Estimation is usually done via Ordinary Least Squares (OLS) (Frequentist):

$$\min_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

# Bayesian Linear Regression: Priors and Likelihood

- In the Bayesian framework, we specify prior distributions for all unknown parameters.

- Model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- Priors:

$$\beta_0 \sim \mathcal{N}(0, 10^2), \quad \beta_1 \sim \mathcal{N}(0, 10^2), \quad \sigma \sim \text{Half-Normal}(0, 5)$$

- Likelihood:

$$y_i \mid \beta_0, \beta_1, \sigma \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$$

- The posterior distribution is then computed via Bayes' Rule:

$$p(\beta_0, \beta_1, \sigma \mid y, x) \propto p(y \mid \beta_0, \beta_1, \sigma, x) \cdot p(\beta_0) \cdot p(\beta_1) \cdot p(\sigma)$$

# Assumptions of the Linear Regression Model

- **Linearity:** The response  $y$  is a linear function of predictors:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- **Gaussian Errors:** The error term follows a normal distribution:

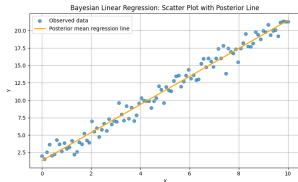
$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- **Independent Observations:** Each data point  $(x_i, y_i)$  is independent.
- **Homoscedasticity:** Constant variance across all levels of  $x$ :

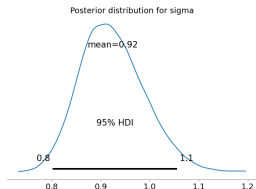
$$\text{Var}(y_i) = \sigma^2$$

- **Priors are specified:** Prior distributions for  $\beta_0, \beta_1, \sigma^2$  must be defined.

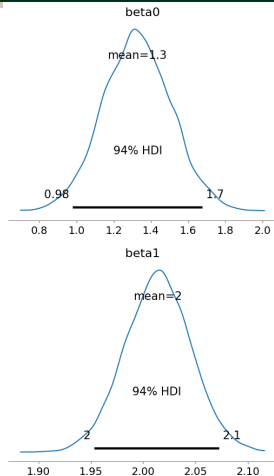
# Bayesian Linear Regression: Visualization



**Figure 2:** Regression Line and Observations



**Figure 3:** Posterior distribution of  $\sigma^2$



**Figure 4:** Posterior Distributions of  $\beta_0$  and  $\beta_1$

# What is Logistic Regression?

- Used to model a binary outcome  $y_i \in \{0, 1\}$  given predictor(s)  $x_i$ .
- The logistic model assumes:

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_i \quad \text{where} \quad p_i = \mathbb{P}(y_i = 1 \mid x_i)$$

- Hence,

$$y_i \sim \text{Bernoulli}(p_i)$$

- The inverse-logit function maps real values to probabilities:

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i)}}$$

- Goal: Estimate  $\beta_0, \beta_1$  to classify or predict probabilities.

# Bayesian Logistic Regression: Priors and Likelihood

- In the Bayesian framework, priors are placed on the parameters:

$$\beta_0 \sim \mathcal{N}(0, 10^2), \quad \beta_1 \sim \mathcal{N}(0, 10^2)$$

- The likelihood is based on the Bernoulli distribution:

$$y_i \sim \text{Bernoulli} \left( \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i)}} \right)$$

- The posterior:

$$p(\beta_0, \beta_1 \mid y, x) \propto \prod_{i=1}^n p(y_i \mid \beta_0, \beta_1, x_i) \cdot p(\beta_0) \cdot p(\beta_1)$$

- Estimation via MCMC (e.g., NUTS in PyMC).

## Connection with Beta-Binomial Model

- The Beta-Binomial model is the conjugate prior-posterior structure for binomial data:

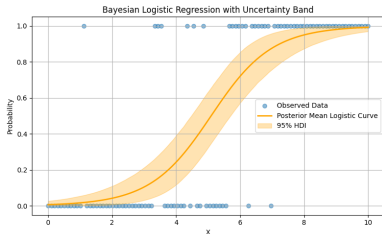
$$y \sim \text{Binomial}(n, p), \quad p \sim \text{Beta}(\alpha, \beta)$$

- In logistic regression, we generalize this to:

$$y_i \sim \text{Bernoulli}(p_i), \quad p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i)}}$$

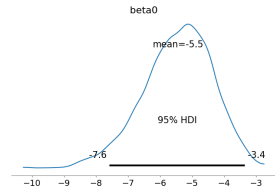
- Instead of a prior on  $p$ , we place priors on  $\beta_0, \beta_1$ , which define  $p_i$  via the logit link.
- Thus, logistic regression can be seen as a flexible, covariate-driven generalization of the Beta-Binomial.

# Bayesian Logistic Regression: Visualization

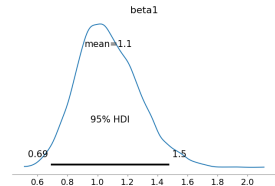


**Figure 5:** Regression Line and Observations

Posterior distributions for beta0 and beta1 (Logistic Regression)



Posterior distributions for beta0 and beta1 (Logistic Regression)



**Figure 6:** Posterior Distributions of  $\beta_0$  and  $\beta_1$



# Interpretation of Odds in Bayesian Logistic Regression

- In logistic regression, the model is:

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_i \quad \Rightarrow \quad p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i)}}$$

- The odds are defined as:

$$\text{odds}(x_i) = \frac{p_i}{1 - p_i} = e^{\beta_0 + \beta_1 x_i}$$

- Therefore, if  $\beta_1 > 0$ , an increase in  $x$  raises the probability  $p_i$ ; if  $\beta_1 < 0$ , it decreases it.
- A one-unit increase in  $x$  multiplies the odds by  $e^{\beta_1}$ . If  $\beta_1 > 0$ , the odds increase; if  $\beta_1 < 0$ , the odds decrease.
- In the Bayesian framework, the posterior distribution of  $\beta_1$  can be transformed to interpret the odds ratio:

$$\text{Posterior odds ratio} = e^{\text{posterior samples of } \beta_1}$$

# Hierarchical Models and Pooling Strategies

**Hierarchical models** Also known as *multilevel*, *mixed effects*, or *random effects* models, are designed for data with group or nested structure. These terms are often used interchangeably in the literature, though some authors distinguish them to emphasize subtle modeling differences.

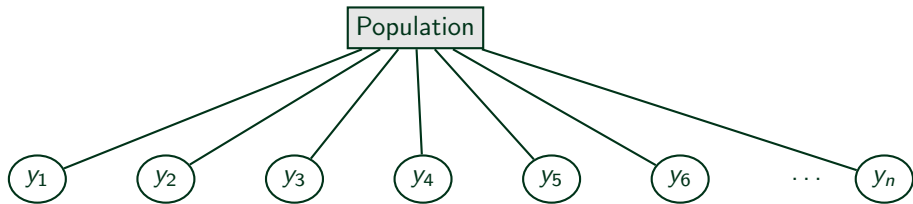
## Pooling Strategies in Inference:

- **Complete pooling:** Assumes all groups share the same parameters. Ignores group-specific variation.
- **No pooling:** Each group is modeled separately with no shared information. Estimates may be unstable for small groups.
- **Partial pooling (Hierarchical):** Group-specific parameters are drawn from a common distribution, allowing the model to share information across groups.

## Why Use Hierarchical Models?

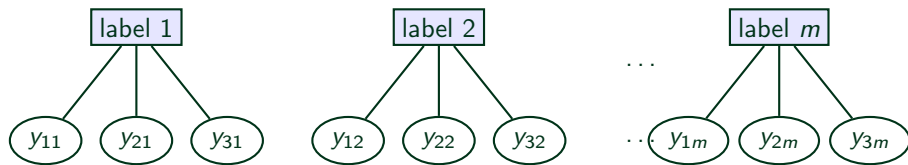
They balance individual group estimates with overall trends, improving stability and inference especially when data are sparse within some groups.

# Complete Pooling



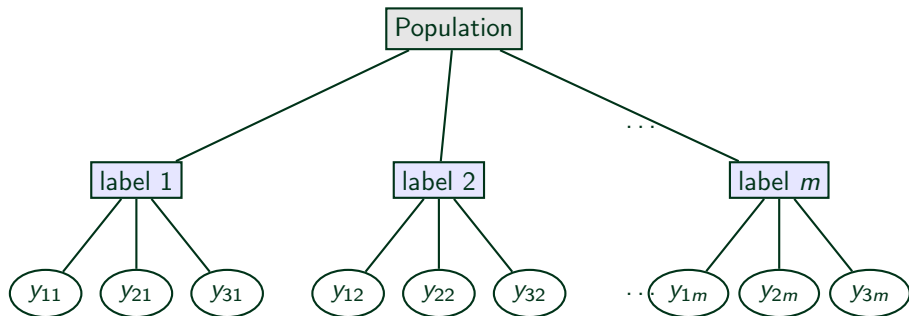
The diagram above represents **complete pooling**, where all observations  $y_1, y_2, \dots, y_n$  depend on a single population level parameter, there are no group-specific parameters involved.

# No Pooling



In a **no pooling** model, each group has its own parameter  $\theta_i$  and generates multiple observations  $y_{ij}$  independently of other groups. There is no shared structure across groups.

# Partial Pooling (Hierarchical)



In **partial pooling**, each observation  $y_i$  has its own parameter  $\theta_i$ , but these are drawn from a common population distribution. This allows for sharing information across groups, leading to more stable and balanced estimates.

# Why Hierarchical Models?

## Motivation for Hierarchical Data Collection

- It is often more practical and cost-efficient to collect multiple observations within fewer groups than single observations across many.
- Hierarchical data reveal:
  - **Within-group variability:** e.g., consistency in a hotel's yearly performance.
  - **Between-group variability:** e.g., differences in average performance between hotels.

**Illustration with labels:** Though all models use the same data:

- Complete pooling hides group differences.
- No pooling captures differences but overfits small data.
- Hierarchical (partial pooling) balances individuality and shared strength.

# Hierarchical Estimation in Tourism: A Colombian Example

**Context:** We want to estimate the average daily spending of tourists in three cities: Cartagena, Medellín, and Bogotá. We survey **5 tourists per city**.

## Modeling Strategies:

- **Complete Pooling:** Ignores city structure. Combines all data into a single estimate of tourist spending.
- **No Pooling:** Estimates a separate mean for each city using only its own data. Risk of overfitting when sample size is small.
- **Partial Pooling (Hierarchical):** Each city's estimate is adjusted ("shrunk") toward the overall mean. Balances local and global information.

**Goal:** Show how hierarchical modeling improves estimation by sharing information across structurally related groups (cities), especially when observations are limited.

# Tourism Spending Model: Complete vs No vs Partial Pooling

**Objective:** Predict total tourist spending as a function of length of stay across three cities: Cartagena, Medellín, Bogotá.

## Variables:

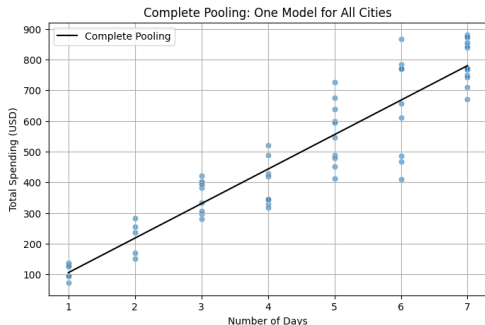
- `days` – Number of nights spent
- `total_spending` – Total money spent = `daily` × `days`
- `city` – Categorical grouping variable

## Modeling Approaches:

- **Complete Pooling:** One regression line for all cities.
- **No Pooling:** One regression line per city, independently.
- **Partial Pooling:** One model with city-specific effects drawn from a common distribution (hierarchical).



# Complete Pooling: One Model for All Cities

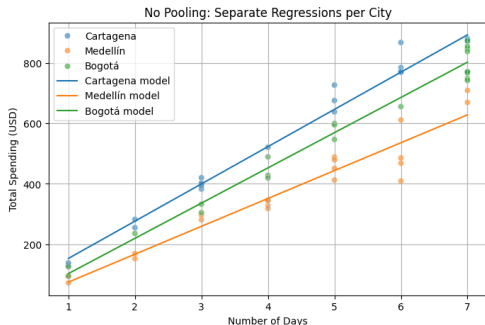


## Description:

- A single regression line is fitted across all data.
- Ignores group (city) differences.
- Assumes all tourists behave the same regardless of location.
- Simple and stable, but can underfit group-specific trends.

**Interpretation:** Useful when group differences are small or data is sparse, but may overlook important heterogeneity in local behaviors.

# No Pooling: Separate Models per City

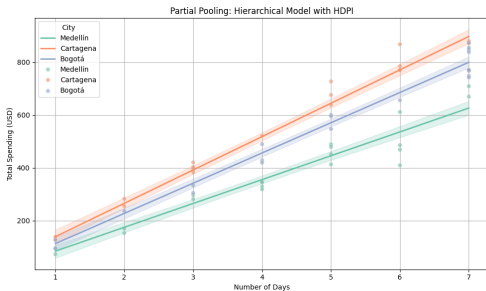


## Description:

- Fits an independent regression line for each city.
- No information is shared across groups.
- Captures group-specific behavior accurately.
- Risk of overfitting, especially with limited data per city.

**Interpretation:** This approach models each city as entirely independent, which works well when group sizes are large and behaviors are truly distinct.

# Partial Pooling: Hierarchical Model

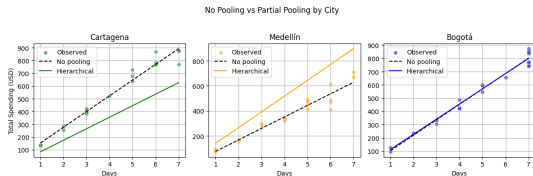


## Description:

- Each city has its own regression line, but parameters are informed by a shared distribution.
- Balances flexibility and stability by combining local and global information.
- Allows borrowing strength across groups, reducing overfitting.
- Captures both group-level (city) and population-level patterns.

**Interpretation:** Partial pooling shrinks estimates toward the global mean, especially useful when group data is limited or noisy.

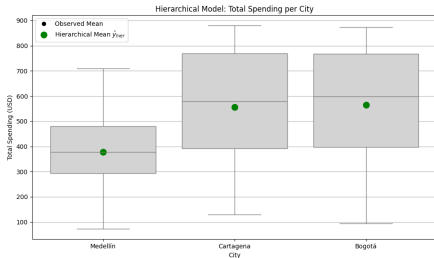
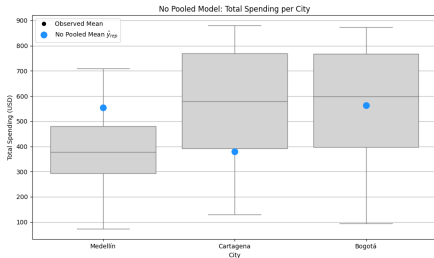
# Comparing No Pooling vs Partial Pooling



Why does the hierarchical (partial pooling) line look different in the two visualizations?

- Both plots are based on the same Bayesian hierarchical model.
- The first plot (not shown here) uses the **mean prediction from each posterior draw**, combining thousands of samples, and includes **credible intervals (HDPI)**.
- The plot above uses only the **posterior mean of the parameters** ( $\mu_i$ ,  $\beta_i$  for each city) to create the hierarchical line.
- Therefore, the difference arises from the **use of full posterior draws** vs a **single point estimate** (posterior mean), which smooths the line and reduces uncertainty.
- The full posterior approach captures *uncertainty across draws*, while the plot above only reflects *central tendency*.

# Comparing Posterior Predictive Means: No Pooling vs Hierarchical



- **No Pooling Model (left):** Fits a separate model for each city. The predicted means ( $\hat{y}_{rep}$ ) are more affected by local noise and extreme values.
- **Hierarchical Model (right):** Performs *partial pooling* across cities. The predicted means ( $\hat{y}_{hier}$ ) are shrunk toward a global average, reducing the influence of extremes.
- The hierarchical model mitigates overfitting and provides more stable estimates when data is scarce within groups.

# Complete Pooling Model

**Model Assumption:** All cities share the same intercept and slope.

$$y_i = \mu + \beta \cdot \text{days}_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

**Description:**

- Assumes a single linear relationship for all observations, ignoring city-specific differences.
- Most efficient when groups (cities) are homogeneous.
- Can lead to bias if group-level heterogeneity is important.

# No Pooling Model

**Model Assumption:** Each city has its own intercept and slope.

$$y_{ij} = \mu_j + \beta_j \cdot \text{days}_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

**Description:**

- Fits separate regressions for each group (city).
- Captures group-specific effects well.
- May overfit if data is sparse in some groups.

# Partial Pooling: Hierarchical Model

**Model Assumption:** Group-specific parameters are drawn from a common distribution.

$$y_{ij} = \mu_j + \beta_j \cdot \text{days}_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

$$\mu_j \sim \mathcal{N}(\mu_0, \sigma_\mu^2)$$

$$\beta_j \sim \mathcal{N}(\beta_0, \sigma_\beta^2)$$

## Description:

- Allows each city to have its own parameters while sharing strength across cities.
- Balances between flexibility (like no pooling) and generalization (like complete pooling).
- Reduces overfitting and improves estimation in small-sample groups.



# Model Comparison: New City with One Observation

Can we predict for a new city with only one data point?

## No Pooling (Separate Regressions)

- Each city has its own independent regression.
- Parameters  $\mu_j$  and  $\beta_j$  are not shared across cities.
- **Problem:** A new city (e.g., Cali) has no estimated parameters.
- **Cannot predict** without retraining the model on Cali.

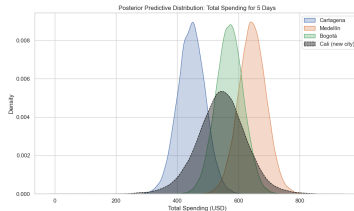
## Partial Pooling (Hierarchical Model)

- City-level parameters are drawn from common distributions:

$$\mu_j \sim \mathcal{N}(\mu_0, \sigma_\mu), \quad \beta_j \sim \mathcal{N}(\beta_0, \sigma_\beta)$$

- **Benefit:** Even with just one observation, the model borrows strength from other cities.
- **Can predict** for Cali using the hierarchical prior.

# Posterior Prediction for a New City (Partial Pooling)



## Posterior Predictive Distributions for Total Spending over 5 Days

- This graph shows the predicted total spending for a 5-day stay in four cities: Cartagena, Medellín, Bogotá, and Cali.
- The first three cities were used in the training data and have many observations.
- Cali, on the other hand, is a new city with only one observation.
- **Partial pooling** uses the hyperpriors to estimate Cali's trend, shrinking the prediction toward the overall population average.
- This approach provides a reasonable estimate for new groups with little or no data, unlike the *no pooling* model, which cannot estimate for unseen categories.

# Appendix

- The Beta distribution is defined for  $\theta \in [0, 1]$ :

$$f(\theta) = \frac{1}{\text{Beta}(a, b)} \theta^{a-1} (1 - \theta)^{b-1} \quad (1)$$

- Where  $\text{Beta}(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$
- The Beta function ensures normalization:

$$\int_0^1 f(\theta) d\theta = 1 \quad (2)$$

# Bayes' Rule for Posterior

Given i.i.d. binary data  $Y_1, \dots, Y_n$ :

$$p(\theta | y) = \frac{p(y | \theta)p(\theta)}{p(y)} \quad (3)$$

$$p(y | \theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y} \quad (4)$$

$$p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1 - \theta)^{b-1} \quad (5)$$

# Posterior Derivation

Combining the terms:

$$p(\theta | y) \propto \theta^{a-1+y} (1-\theta)^{b-1+n-y} \quad (6)$$

$$\propto \theta^{a+y-1} (1-\theta)^{b+n-y-1} \quad (7)$$

Thus, the posterior is:

$$\theta | y \sim \text{Beta}(a+y, b+n-y) \quad (8)$$

**Conclusion:** The Beta distribution is the conjugate prior for the Binomial likelihood.

- The kernel is the part of the PDF that depends on  $\theta$ .
- Constant terms (normalization constants) are absorbed during derivation.
- Posterior expectation:

$$\mathbb{E}[\theta] = \frac{a + y}{a + b + n} \quad (9)$$

- For uniform prior,  $\text{Beta}(1, 1)$  yields:

$$\theta \mid y \sim \text{Beta}(1 + y, 1 + n - y) \quad (10)$$

- For i.i.d. observations  $Y_1, \dots, Y_n \sim \text{Poisson}(\theta)$ :

$$\begin{aligned} P(Y_1, \dots, Y_n \mid \theta) &= \prod_i \frac{\theta^{y_i} e^{-\theta}}{y_i!} \\ &= \theta^{\sum y_i} e^{-n\theta} \prod_i \frac{1}{y_i!} \end{aligned}$$

- The sufficient statistic is  $\sum y_i$
- Inference only requires the total count, not individual values



# Bayesian Update with Gamma Prior

- Choose a Gamma prior:  $\theta \sim \text{Gamma}(a, b)$

- Gamma PDF:

$$p(\theta) = \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta}$$

- Conjugate prior: posterior is also Gamma

$$p(\theta | y) \propto \theta^{a-1+\sum y_i} e^{-(b+n)\theta}$$

- Posterior:  $\theta | y \sim \text{Gamma}(a + \sum y_i, b + n)$

# Posterior Expectation and Interpretation

- Posterior expectation:

$$\begin{aligned}\mathbb{E}[\theta \mid y] &= \frac{a + \sum y_i}{b + n} \\ &= \underbrace{\frac{b}{b+n} \cdot \frac{a}{b}}_{\text{Weighted Prior Mean}} + \underbrace{\frac{n}{b+n} \cdot \frac{\sum y_i}{n}}_{\text{Weighted Sample Mean}}\end{aligned}$$

- This is a convex combination of the prior mean and the sample mean.
- As  $n$  increases, the posterior mean shifts towards the sample mean.

# Continuous Distributions and the Normal Case

- For continuous variables, uncertainty is modeled using a probability density function (PDF).
- The normal distribution is widely used for estimating average values of parameters  $\theta$ .
- Central Limit Theorem justifies its frequent application.

$$Y \sim \mathcal{N}(\theta, \sigma^2)$$

$$p(y \mid \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \cdot \frac{(y - \theta)^2}{\sigma^2}\right)$$

# Joint Posterior Distribution

We aim to estimate the joint posterior:

$$p(\theta, \sigma^2 \mid y_1, \dots, y_n) = \frac{p(y_1, \dots, y_n \mid \theta, \sigma^2) p(\theta, \sigma^2)}{p(y_1, \dots, y_n)}$$

Assume prior factorization:

$$p(\theta, \sigma^2) = p(\theta \mid \sigma^2) p(\sigma^2)$$

Use conjugate priors:

$$\theta \mid \sigma^2 \sim \mathcal{N}(\mu_0, \sigma^2/\kappa_0), \quad 1/\sigma^2 \sim \text{Gamma}(\nu_0/2, \nu_0\sigma_0^2/2)$$

# Interpreting the Priors

Gamma prior on precision:

$$1/\sigma^2 \sim \text{Gamma}(a = \nu_0/2, b = \nu_0\sigma_0^2/2)$$

This implies:

$$\mathbb{E}[\sigma^2] = \frac{\nu_0\sigma_0^2}{\nu_0 - 2}, \quad \text{Mode} = \frac{\nu_0\sigma_0^2}{\nu_0 + 2}$$

Prior on  $\theta$ :

$$\theta \mid \sigma^2 \sim \mathcal{N}(\mu_0, \sigma^2/\kappa_0)$$

where  $\kappa_0$  acts as a prior sample size.

# Posterior Distributions (Conditional Form)

Conditional posterior of  $\theta$ :

$$\theta \mid \sigma^2, y \sim \mathcal{N} \left( \mu_n = \frac{\kappa_0 \mu_0 + n \bar{y}}{\kappa_n}, \sigma^2 / \kappa_n \right)$$

Updated parameters:

$$\kappa_n = \kappa_0 + n$$

Posterior mean is a weighted average of  $\mu_0$  and  $\bar{y}$ .

# Posterior for $\sigma^2$

Marginal posterior:

$$\sigma^2 \mid y \sim \text{Inverse-Gamma}(\nu_n/2, \nu_n \sigma_n^2/2)$$

With:

$$\nu_n = \nu_0 + n, \quad \sigma_n^2 = \frac{1}{\nu_n} \left[ \nu_0 \sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_n} (\bar{y} - \mu_0)^2 \right]$$

Three sources of variability: prior, sample variance, and mean shift.

# Summary of Posterior Parameters

- $\mu_n = \frac{\kappa_0 \mu_0 + n \bar{y}}{\kappa_0 + n}$
- $\kappa_n = \kappa_0 + n$
- $\nu_n = \nu_0 + n$
- $\sigma_n^2 = \frac{1}{\nu_n} \left[ \nu_0 \sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_n} (\bar{y} - \mu_0)^2 \right]$

Posterior:

$$\theta \mid \sigma^2, y \sim \mathcal{N}(\mu_n, \sigma^2 / \kappa_n)$$
$$1/\sigma^2 \mid y \sim \text{Gamma}(\nu_n/2, \nu_n \sigma_n^2/2)$$