Lecture 10: Bayesian Inference and Bayes' Rule

CS109A: Introduction to Data Science

Harvard University

■ **Course:**	CS109A: Introduction to Data Science
■ **Lecture:**	Lecture 10
■ **Instructor:**	Pavlos Protopapas, Kevin Rader, Chris Gumb
■ **Objective:**	Reviewing statistical inference, understanding the distinction between confidence and prediction intervals, learning Bayes' rule, and introducing Bayesian inference as an alternative paradigm to frequentist statistics

Contents

1 Introduction and Motivation

Lecture Overview

This lecture reviews statistical inference concepts and introduces a fundamentally different way of thinking about probability and parameters: **Bayesian inference**. The Bayesian approach treats parameters as random variables with probability distributions, rather than fixed unknown constants.

Key Topics:

- **Review of Inference**: Standard errors, confidence intervals, hypothesis testing
- **Bootstrap vs. Permutation**: When to use which resampling method
- **Confidence Interval vs. Prediction Interval**: A crucial distinction
- **Likelihood Review**: The foundation for Bayesian methods
- **Bayes' Rule**: Flipping conditional probabilities
- **Bayesian Inference**: Updating beliefs based on evidence
- **Frequentist vs. Bayesian**: Two worldviews of statistics

1.1 Continuing the Housing Example

We continue with the Cambridge/Somerville housing data from last lecture:

- **Simple regression**: Price $\sim$ Square footage
- **Estimated model**: $\hat{y} = 247.4 + 0.5898x$
- **Interpretation**: Each additional square foot is associated with roughly \$600 higher selling price

Correlation vs. Causation Reminder

"Be careful. Doesn't mean it's causal because there could be other confounders in that data set that we haven't controlled for."

The coefficient 0.5898 represents an *association*, not necessarily a causal effect. Other factors correlated with square footage (neighborhood quality, lot size, etc.) might be driving part of this relationship.

2 Review: Statistical Inference

2.1 Population vs. Sample

Definition: Two Models

Population Model (what we want):

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- β_0, β_1 are the *true* parameters for all homes in this geographic region

- $\sigma^2 = \text{Var}(\epsilon_i)$ is another unknown parameter

Estimated Model (what we have):

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- $\hat{\beta}_0, \hat{\beta}_1$ are estimates from our sample of ~500 homes
- These are one "realization" from the sampling distribution

2.2 Standard Errors: Understanding Uncertainty

The **standard error** (SE) quantifies how much our estimate would vary across different samples:

Definition: Standard Error of the Slope

$$\hat{SE}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

where $\hat{\sigma}^2$ is the estimated residual variance.

2.2.1 What Controls the Standard Error?

Looking at the formula, we can build intuition:

How to Reduce Standard Error

1. **Increase sample size (n):**
   - More observations → larger denominator
   - The sum $\sum(x_i - \bar{x})^2$ increases with n
   - **Most reliable way to reduce SE**
2. **Increase spread in X:**
   - Wider range of X values → larger $\sum(x_i - \bar{x})^2$
   - Hard to control in observational studies
   - In experiments, you can deliberately sample extreme X values
3. **Reduce residual variance ($\hat{\sigma}^2$):**
   - Better model → tighter fit around the line
   - Add relevant predictors, use transformations
   - Limited by irreducible error

2.2.2 What About Standardizing Predictors?

Professor Rader poses a "midterm type question": What happens to the SE formula if you standardize the predictor?

Example: The Standardization Trap

At first glance, standardizing X (mean 0, variance 1) might seem to reduce SE since the denominator becomes smaller.

But wait!

When you standardize X:

- The denominator $\sum(x_i - \bar{x})^2$ decreases
- BUT $\hat{\beta}_1$ itself changes (different units!)
- Interpretation changes from "per unit X" to "per standard deviation of X"

Net effect: The t-statistic remains the same!

$$t = \frac{\hat{\beta}_1}{\hat{SE}(\hat{\beta}_1)}$$

Standardizing doesn't magically make your predictor more "significant." The narrower CI is offset by a smaller $\hat{\beta}_1$.

2.3 Confidence Intervals: Formula-Based

Definition: Confidence Interval Formula

$$\text{CI for } \beta_1: \quad \hat{\beta}_1 \pm t^* \cdot \hat{SE}(\hat{\beta}_1)$$

where t^* is the critical value from the t-distribution (roughly 2 for 95% CI with large samples).

2.3.1 Why t-distribution Instead of Normal?

t-Distribution vs. Normal Distribution

Normal (Z) distribution: Used when σ is *known* (rare in practice).

t-distribution: Used when σ must be *estimated* from data.

- Estimating σ adds extra uncertainty
- t-distribution has "fatter tails" to account for this
- As sample size $\to \infty$, t-distribution $\to$ normal distribution
- With $n \geq 50$, the difference is minimal

2.4 Hypothesis Testing Review

Hypothesis Testing for β_1

Hypotheses:

- $H_0 : \beta_1 = 0$ (no association between X and Y)
- $H_A : \beta_1 \neq 0$ (there is an association)

Test Statistic:

$$t = \frac{\hat{\beta}_1 - 0}{\hat{SE}(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\hat{SE}(\hat{\beta}_1)}$$

This measures: "How many standard errors is our estimate from zero?"
p-value: Probability of observing a $|t|$ this extreme or more extreme, assuming H_0 is true.
Decision: If p-value < 0.05, reject H_0.

2.5 Two-Sided Tests and Absolute Values

Example: Why Absolute Values in p-value Calculation?

The p-value formula often looks like:

$$\text{p-value} = P(|T| \geq |t_{\text{observed}}|)$$

Why absolute values?
Because our alternative hypothesis is $\beta_1 \neq 0$ (two-sided). We consider evidence against H_0 whether the association is positive or negative.
If we observed $t = -4$:

- Area to the left of -4 (extreme negative)
- Plus area to the right of $+4$ (equally extreme positive)

3 Bootstrap vs. Permutation Tests

3.1 When to Use Each

Very Important: Bootstrap vs. Permutation: Critical Distinction

Bootstrap: For **confidence intervals**

- Resample *with replacement* from your data
- Preserves the relationships in your data
- Estimates the sampling distribution of your statistic

Permutation Test: For **hypothesis testing (p-values)**

- Shuffle the response variable Y, keep X fixed
- **Enforces the null hypothesis** (no relationship between X and Y)
- Builds a reference distribution *under* H_0

Why not bootstrap for hypothesis testing?
Bootstrap doesn't enforce H_0. Under certain conditions (violated assumptions, multicollinearity), it can lead to **inflated Type I error**—rejecting H_0 too often when it's actually true.

Example: Permutation Testing Procedure

To test $H_0 : \beta_1 = 0$:

1. **Shuffle**: Randomly permute the Y values (break the X-Y relationship)
2. **Refit**: Calculate $\hat{\beta}_1^*$ on the shuffled data
3. **Repeat**: Do this many times (e.g., 1000 permutations)
4. **Result**: A distribution of $\hat{\beta}_1^*$ values, centered at 0 (since H_0 is enforced)

The p-value is the proportion of permuted $|\hat{\beta}_1^*|$ values that exceed your observed $|\hat{\beta}_1|$.

3.2 Comparing Bootstrap and Formula-Based CIs

From the housing data:

Method	**95% CI for β_1**	**Width**
statsmodels (formula)	(0.544, 0.636)	0.092
Bootstrap	(0.487, 0.705)	0.218

Table 1: *Comparison of confidence intervals*

Why the Difference?

The bootstrap CI is **wider** because:

- The formula-based CI assumes constant variance (homoscedasticity)
- Our data shows clear heteroscedasticity (variance fanning out with larger homes)
- When assumptions are violated, formulas give **incorrectly narrow** CIs
- Bootstrap makes fewer assumptions and captures the true variability

Lesson: When in doubt, bootstrap is safer!

3.3 Linear Regression Assumptions (LINE)

The LINE Assumptions

1. **L**inearity: The relationship is linear
2. **I**ndependence: Observations are independent (most important!)
3. **N**ormality: Residuals are normally distributed (least important if $n > 50$)
4. **E**qual variance: Constant variance (homoscedasticity)

Bootstrap: Relaxes N and E, still requires L and I

Formula-based: Requires all four

4 Confidence Interval vs. Prediction Interval

This is one of the most commonly confused distinctions in regression!

4.1 Two Different Questions

Definition: CI vs. PI

Confidence Interval (CI): Uncertainty about the *mean* response

- Question: "What is the **average** selling price of *all* homes with 2860 sqft?"
- Only includes uncertainty in the regression line ($\hat{\beta}$'s)

Prediction Interval (PI): Uncertainty about a *single* new observation

- Question: "What will *this specific new home* with 2860 sqft sell for?"
- Includes uncertainty in the line **AND** the individual noise (σ^2)

4.2 Why PI is Always Wider

Two Sources of Uncertainty for Predictions

Source 1: Uncertainty in the regression line

- We estimated $\hat{\beta}_0$ and $\hat{\beta}_1$ from a sample
- Different samples would give different estimates
- This is what the CI captures

Source 2: Individual observation variability (irreducible error)

- Even if we knew the exact population line, individual homes vary around it
- This is σ^2 (the variance of ϵ)
- No matter how much data you collect, this never goes away!

Prediction Interval:

$$\text{PI} = \hat{y} \pm t^* \cdot \sqrt{(\text{SE of fit})^2 + \hat{\sigma}^2}$$

The extra $\hat{\sigma}^2$ term is why PI is always wider than CI.

Example: Game Time Question

Use this output to predict with 95% uncertainty the selling price of a home that is 2860 sqft:

```
Intercept: 247.4,  sqft coef: 0.5898,  SE(sqft): 0.023
```

Options:

A. $0.5898 \pm 2 \times 0.023$ (CI for slope)

B. $247.4 + 0.5898 \times 2860$ (point prediction only)

C. $247.4 + 0.5898(2860) \pm 2 \times 0.023$ (CI for mean response)

D. $247.4 + 0.5898(2860) \pm 2\sqrt{0.023^2 + \hat{\sigma}^2}$ (PI for new observation)

Answer: **D**!

Since we're predicting "a home"—a single new observation—we need the **Prediction Interval**, which includes both the model uncertainty and the residual variance $\hat{\sigma}^2$.

5 Likelihood Review

Before diving into Bayesian inference, let's solidify our understanding of likelihood.

5.1 Flipping the Perspective

Definition: Likelihood Function

PDF/PMF: $f(x|\theta)$ — Given parameters, what's the probability of the data?
Likelihood: $L(\theta|x)$ — Given data, how plausible is each parameter value?
Mathematically: They're the same function! The difference is conceptual:

- PDF: θ is fixed, x varies
- Likelihood: x is fixed (observed), θ varies

5.2 Why Products Become Sums

Log-Likelihood

For independent observations $x_1, \ldots, x_n$:

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

Taking the log:

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta)$$

Why use log?

1. Products $\rightarrow$ sums (much easier calculus!)
2. Numerical stability (avoids underflow from multiplying many small numbers)
3. Same maximum: $\arg\max L(\theta) = \arg\max \ell(\theta)$

5.3 The OLS-MLE Connection (Review)

OLS

If we assume $\epsilon_i \sim N(0, \sigma^2)$, then maximizing the likelihood is equivalent to minimizing the sum of squared errors.
Why? The negative log-likelihood becomes:

$$-\ell(\beta_0, \beta_1) = \text{constant} + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Minimizing this with respect to β_0, β_1 is the same as minimizing SSE!

6 Bayes' Rule: Flipping Conditional Probabilities

6.1 The Basic Formula

Definition: Bayes' Rule

For events A and B:

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

Extended version (using Law of Total Probability for denominator):

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|A^c) \cdot P(A^c)}$$

6.2 Example: CS and STAT Concentrators

Example: Conditional Probability Practice

In a hypothetical CS109A class (among undergrads):

- 40% are STAT concentrators: $P(\text{STAT}) = 0.40$
- 60% are CS concentrators: $P(\text{CS}) = 0.60$
- 20% are both (joint/double): $P(\text{STAT} \cap \text{CS}) = 0.20$

Question 1: Among STAT concentrators, what fraction are also CS?

$$P(\text{CS}|\text{STAT}) = \frac{P(\text{STAT} \cap \text{CS})}{P(\text{STAT})} = \frac{0.20}{0.40} = 0.50$$

Question 2: Among CS concentrators, what fraction are also STAT?

$$P(\text{STAT}|\text{CS}) = \frac{P(\text{STAT} \cap \text{CS})}{P(\text{CS})} = \frac{0.20}{0.60} = 0.333$$

Key Insight: $P(\text{CS}|\text{STAT}) \neq P(\text{STAT}|\text{CS})$!

These are fundamentally different questions—don't confuse them.

6.3 Independence and Dependence

Are STAT and CS Independent?

Events are **independent** if $P(A \cap B) = P(A) \cdot P(B)$.

Check: $P(\text{STAT}) \cdot P(\text{CS}) = 0.40 \times 0.60 = 0.24$

But $P(\text{STAT} \cap \text{CS}) = 0.20 \neq 0.24$

Conclusion: They are **dependent**!

Meaning: Knowing someone's CS status gives you information about whether they're also STAT. The probability changes once you have additional information.

7 Bayes' Rule in Diagnostic Testing

This is a classic application that will connect directly to classification later.

Example: Pregnancy Test Example

A pregnancy test has:

- **Sensitivity**: $P(\text{Test} + | \text{Pregnant}) = 0.97$ (true positive rate)
- **Specificity**: $P(\text{Test} - | \text{Not Pregnant}) = 0.99$ (true negative rate)

Among people taking this test, about 30% are actually pregnant: $P(\text{Pregnant}) = 0.30$

Question: If someone tests positive, what's the probability they're actually pregnant?

We want: $P(\text{Pregnant}|\text{Test}+)$

We have: $P(\text{Test} + | \text{Pregnant})$

We need to **flip the conditional**—use Bayes' Rule!

7.1 Applying Bayes' Rule

$$
\begin{aligned}
P(\text{Preg}|T+) &= \frac{P(T+|\text{Preg}) \cdot P(\text{Preg})}{P(T+)} \\
&= \frac{P(T+|\text{Preg}) \cdot P(\text{Preg})}{P(T+|\text{Preg}) \cdot P(\text{Preg}) + P(T+|\text{Not Preg}) \cdot P(\text{Not Preg})} \\
&= \frac{0.97 \times 0.30}{0.97 \times 0.30 + 0.01 \times 0.70} \\
&= \frac{0.291}{0.291 + 0.007} = \frac{0.291}{0.298} \approx 0.977
\end{aligned}
$$

Belief Update: Prior to Posterior

- **Prior probability** of being pregnant: 30%
- **Evidence**: Positive test result
- **Posterior probability** of being pregnant: 97.7%

The positive test dramatically updated our belief from 30% to 97.7%!

This is the essence of Bayesian inference: Using evidence to update our beliefs.

8 Bayesian Inference: A New Paradigm

8.1 From Events to Parameters

Now we apply this same logic to statistical parameters:

Definition: Bayesian Inference Formula

$$f(\theta|X) = \frac{f(X|\theta) \cdot f(\theta)}{f(X)}$$

Components:

- $f(\theta|X)$ — **Posterior Distribution**: Our updated belief about θ after seeing data
- $f(X|\theta)$ — **Likelihood**: How probable is our data given θ?
- $f(\theta)$ — **Prior Distribution**: Our initial belief about θ before seeing data
- $f(X)$ — **Evidence/Marginal Likelihood**: A normalizing constant

Simplified:

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

8.2 The Bayesian Philosophy

Very Important: Frequentist vs. Bayesian Worldviews

8.3 Example: Three Coins

Example: Discrete Bayesian Inference

You have three coins in your pocket:

- Coin A: $P(\text{Heads}) = 0.1$ (biased toward tails)
- Coin B: $P(\text{Heads}) = 0.5$ (fair)
- Coin C: $P(\text{Heads}) = 0.9$ (biased toward heads)

You randomly select one coin (equal probability) and flip it 4 times.

Data observed: 3 Heads, 1 Tail

Question: Which coin did you probably pick?

Step 1: Prior

Before seeing any flips, each coin is equally likely:

$$P(p = 0.1) = P(p = 0.5) = P(p = 0.9) = \frac{1}{3}$$

Step 2: Likelihood

Calculate probability of "3H, 1T" for each coin using binomial distribution:

$$L(p = 0.1) = \binom{4}{3}(0.1)^3(0.9)^1 = 4 \times 0.001 \times 0.9 = 0.0036$$

$$L(p = 0.5) = \binom{4}{3}(0.5)^3(0.5)^1 = 4 \times 0.125 \times 0.5 = 0.2500$$

$$L(p = 0.9) = \binom{4}{3}(0.9)^3(0.1)^1 = 4 \times 0.729 \times 0.1 = 0.2916$$

Step 3: Posterior (unnormalized)

$$P(p = 0.1|\text{data}) \propto 0.0036 \times \frac{1}{3} = 0.0012$$
$$P(p = 0.5|\text{data}) \propto 0.2500 \times \frac{1}{3} = 0.0833$$
$$P(p = 0.9|\text{data}) \propto 0.2916 \times \frac{1}{3} = 0.0972$$

Step 4: Normalize (sum = 0.0012 + 0.0833 + 0.0972 = 0.1817)

$$P(p = 0.1|\text{data}) = \frac{0.0012}{0.1817} \approx 0.007 \quad (0.7\%)$$
$$P(p = 0.5|\text{data}) = \frac{0.0833}{0.1817} \approx 0.458 \quad (45.8\%)$$
$$P(p = 0.9|\text{data}) = \frac{0.0972}{0.1817} \approx 0.535 \quad (53.5\%)$$

Conclusion: Our belief shifted from (33.3%, 33.3%, 33.3%) to (0.7%, 45.8%, 53.5%). We now believe we most likely picked the biased-toward-heads coin!

9 Game Time: Bootstrap Sample Size

Example: Class Question

What happens to the bootstrap distribution when B (number of bootstrap samples) increases?
Options:

A. The distribution becomes more normal

B. The variance decreases

C. The confidence intervals become narrower

D. The distribution gets smoother

Correct Answer: **D** (and arguably A)

What increasing B does:

- Gets a more *precise estimate* of the true sampling distribution
- Makes the histogram smoother (less jagged)
- Does NOT change the underlying variability of your estimate!

What increasing B does NOT do:

- Narrow confidence intervals
- Reduce the variance of $\hat{\beta}$

To actually narrow CIs, you need to increase n (original sample size), not B!

10 Quick Reference Summary

Lecture 10 Quick Reference Card

1. CI vs. PI

- **CI**: Uncertainty in mean response (where is the line?)
- **PI**: Uncertainty in single observation (includes $\hat{\sigma}^2$)
- PI is **always wider**!

2. Bootstrap vs. Permutation

- **Bootstrap**: Confidence intervals (resample with replacement)
- **Permutation**: Hypothesis tests (shuffle Y, enforces H_0)

3. Bayes' Rule

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

Allows "flipping" conditional probabilities!

4. Bayesian Inference

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

$$f(\theta|X) \propto f(X|\theta) \cdot f(\theta)$$

5. Frequentist vs. Bayesian

- **Frequentist**: θ is fixed constant, probability = long-run frequency
- **Bayesian**: θ has a distribution, probability = belief

11 Common Questions and Answers

Q: The prior seems subjective. Isn't that a problem?

A: It can be, but:

1. Use "informative priors" based on previous research or domain knowledge
2. Use "uninformative/flat priors" when you have no prior knowledge
3. Most importantly: **With enough data, the prior gets overwhelmed**. The posterior converges to the same answer regardless of (reasonable) prior choices.

Q: When should I use Bayesian vs. Frequentist methods?

A: Both have their place:

- **Frequentist**: Simpler, faster computation, standard in clinical trials

- **Bayesian**: Can incorporate prior knowledge, gives full posterior distribution, better for small samples

Modern data science often uses both and compares results.

Q: Why does increasing bootstrap samples B not narrow the CI?

A: B controls how accurately you *estimate* the sampling distribution. But the *width* of the sampling distribution depends on your original sample size n. More bootstrap samples give you a smoother histogram of the same width.

12 The Monty Hall Problem (Bonus)

Professor Rader mentions this classic probability puzzle:

Example: The Monty Hall Problem

Setup:

- 3 doors: 1 car (prize), 2 goats (losers)
- You pick a door (say, Door 1)
- Host (who knows where the car is) opens another door revealing a goat
- You're offered: "Do you want to switch to the remaining door?"

Counterintuitive Result:

- Stay: Win probability = 1/3
- Switch: Win probability = 2/3

Intuition (100 doors version):

- Pick 1 door out of 100. Probability you picked the car = 1/100.
- Host reveals 98 goats, leaving 2 doors.
- If you switch, you win unless you initially picked the car (probability 99/100)!

This demonstrates how conditional probability can be counterintuitive!

13 Looking Ahead

In the next lecture, we'll:

- Work through the three-coins example in more detail
- Explore continuous priors and posteriors
- See how Bayesian inference connects to regression
- Start applying these ideas in Python

The Bayesian framework will also be crucial when we move to classification, where we'll care about $P(\text{Class}|\text{Features})$—a natural application of Bayes' Rule!