MS&E 226: Fundamentals of Data Science

Lecture 16: Bayesian inference and decision making

Ramesh Johari

From causal inference to decisions

Causal inference

In causal inference, we focused on methods to estimate the *average treatment effect* (ATE):

$$\text{ATE} = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$$

We studied a range of methods to construct an estimator $\widehat{\text{ATE}}$, both from experiments, and from observational data.

We also studied frequentist approaches to quantify our uncertainty: under appropriate assumptions, $\widehat{\text{ATE}}$ has a sampling distribution that is asymptotically normal, with a standard error $\widehat{\text{SE}}$ that we can estimate from data (which can be used to construct a confidence interval for ATE).

Question: *How do we use* $\widehat{\text{ATE}}$ *to support decisions?*

Example: Online learning platform

Suppose a new online learning platform considers redesigning course descriptions to be less technical and more accessible (the "treatment").

The platform runs an experiment to measure the causal effect of treatment on the *course enrollment rate* (fraction of visitors who enroll).

The experiment randomizes a subset of *courses* to treatment (new descriptions) or control (old descriptions), and the resulting data is used to calculate $\widehat{\text{ATE}}$ (e.g., using difference-in-means, or an AIPW approach) with associated $\widehat{\text{SE}}$.

The resulting 95% confidence interval is $[\widehat{\text{ATE}} - 1.96\widehat{\text{SE}}, \widehat{\text{ATE}} + 1.96\widehat{\text{SE}}]$.

The decision problem

Question: *Should we roll out the new descriptions to all courses?*

Need to consider:

- "Uncertainty" in the treatment effect
- Implementation costs (e.g., designer time, development resources)
- Potential risks (e.g., might confuse existing users)

Does the 95% confidence interval address these considerations?

Limitations of frequentist inference for decisions

A 95% confidence interval $[\widehat{\text{ATE}} - 1.96\widehat{\text{SE}}, \widehat{\text{ATE}} + 1.96\widehat{\text{SE}}]$ tells us:

> *In 95% of "parallel universes" (frequentist repetition), the true* ATE *is in this interval.*

A key challenge is that making the decision requires understanding uncertainty around *this single decision* — not over many repetitions.

Limitations of frequentist inference for decisions

Note that in frequentist inference, ATE is *not* random; so a confidence interval *cannot* answer the following types of questions:

- What is the "probability" that ATE > 0? (That the effect is positive?)
- What is the "probability" that ATE $> c$? (That the benefit exceeds cost c?)
- What is the "expected" benefit $\mathbb{E}_{\text{ATE}}[\text{ATE} - c]$?
- How should we measure risk due to uncertainty in the ATE?

The Bayesian approach

The preceding discussion suggests we need two ingredients that are missing in frequentist inference:

1. A *probability distribution* over the treatment effect (not just a confidence interval)
2. Some way to combine this distribution with costs/benefits/risks to make optimal decisions

The *Bayesian* framework for statistical inference and decision-making provides these ingredients.

The Bayesian approach: Inference and decisions

Bayesian statistical inference:

- Treat parameters (e.g., the treatment effect ATE) as *random*
- Before experiment, model initial uncertainty over ATE via a *prior* distribution
- Combine prior with data from experiment to express uncertainty over ATE as a *posterior* distribution

The Bayesian approach: Inference and decisions

Bayesian statistical inference:

- Treat parameters (e.g., the treatment effect ATE) as *random*
- Before experiment, model initial uncertainty over ATE via a *prior* distribution
- Combine prior with data from experiment to express uncertainty over ATE as a *posterior* distribution

Bayesian decision theory:

- Formalize costs and benefits via *utilities* (or *loss functions*)
- Compute *expected utility* (or *expected loss*) under the posterior
- *Optimal decisions*: Choose decision that maximizes expected utility, or minimizes expected loss (*Bayes risk*)

In the subsequent sections we develop these components, and combine them to show how a Bayesian approach supports a coherent approach to decision-making.

Bayesian inference

Frequentist vs. Bayesian statistical inference

Frequentist approach:

- Parameters are *fixed* (deterministic)
- Data sample is *random* (drawn from population)
- Estimator uses sample to estimate parameter(s)
- Uncertainty quantified through *sampling distribution* of estimator:
  Repeat data sampling and estimation procedure ("parallel universes")

Bayesian approach:

- Parameters are *random*
- Start with a *prior distribution* on parameters (before seeing data)
- Use *Bayes' theorem* to combine prior with data → *posterior distribution*
- Uncertainty quantified through the posterior distribution

Bayes' theorem

Let's start with a simple setup with one parameter, continuous outcomes, and no covariates:

- Parameter θ
- Data sample $\mathbf{Y} = (Y_1, \ldots, Y_n)$ (independent samples from population)
- *Prior* distribution on θ: $h(\theta)$ (pdf or pmf)
- *Likelihood*: $f(\mathbf{Y}|\theta)$ (distribution of data given θ)

Bayes' theorem:

$$h(\theta|\mathbf{Y}) = \frac{f(\mathbf{Y}|\theta)h(\theta)}{f(\mathbf{Y})},$$

where $f(\mathbf{Y}) = \int f(\mathbf{Y}|\tilde{\theta})f(\tilde{\theta})d\tilde{\theta}$. (If θ is discrete, then this is a sum against the pmf, instead of an integral against the pdf.)

The *posterior* $f(\theta|\mathbf{Y})$ is the distribution of θ *after* we have seen the data sample.

Bayes' theorem: In words

Bayes' theorem can be interpreted as follows:

$$\text{posterior on } \theta \propto (\text{likelihood of data given } \theta) \times (\text{prior on } \theta).$$

Here "$\propto$" means "proportional to" – with constant of proportionality $1/f(\mathbf{Y})$.

Note that the constant $1/f(\mathbf{Y})$ does not depend on θ; so to find the posterior, it suffices to compute likelihood $\times$ prior, then *normalize* (so the result is a probability distribution).

Example 1: Binary outcomes

Suppose we flip a coin 5 times, with unknown success probability q (probability of heads).

Suppose we observe: $\boldsymbol{H}, \boldsymbol{H}, \boldsymbol{T}, \boldsymbol{H}, \boldsymbol{T}$. What can we say about q?

Bayesian approach:

1. Start with a *prior* for q: $h(q)$
2. Compute the *likelihood*: $\mathbb{P}(\mathbf{Y}|q) = q^3(1-q)^2$
3. Apply Bayes' rule to get the *posterior*:

$$h(q|\mathbf{Y}) = \frac{\mathbb{P}(\mathbf{Y}|q)h(q)}{\int_0^1 \mathbb{P}(\mathbf{Y}|q')h(q')dq'}.$$

Example 1: Binary outcomes

As an example, suppose that $h(q)$ was the uniform distribution on $[0, 1]$.

Then we can show that the posterior after n flips with k H's and $n - k$ T's is:

$$h(q|\mathbf{Y}) = \frac{1}{B(k+1, n-k+1)} q^k (1-q)^{n-k},$$

the $\mathrm{B}(k+1, n-k+1)$ distribution.

(Here $B(\cdot)$ is the *beta function*.)

Example 1: Binary outcomes

More generally, suppose the *prior* is $q \sim \mathrm{B}(a, b)$.

Then we can show that after n flips with k heads and $n - k$ tails, the *posterior* is:

$$q|\mathbf{Y} \sim \mathrm{B}(a + k, b + n - k).$$

We say the Beta distribution is *conjugate* to binomially distributed data (binary outcomes):

- The prior and posterior are in the same "family" (both Beta)
- *Simple prior to posterior update rule:* Add observed counts to prior parameters

Note: The uniform prior is the special case $\mathrm{B}(1, 1)$.

Example 2: Normal data with normal prior

Suppose $Y_1, \ldots, Y_n$ are independent $\mathcal{N}(\mu, \sigma^2)$, where σ^2 is known.

Suppose the *prior* is: $\mu \sim \mathcal{N}(a, b^2)$.

Then we can show that the *posterior* is also normal: $\mu|\mathbf{Y} \sim \mathcal{N}(\hat{a}, \hat{b}^2)$, where

$$\hat{a} = c_n \overline{Y} + (1 - c_n)a$$

$$\hat{b}^2 = \frac{1}{n/\sigma^2 + 1/b^2}$$

$$c_n = \frac{n/\sigma^2}{n/\sigma^2 + 1/b^2}$$

Example 2: Normal data with normal prior

The *posterior* is $\mu|\mathbf{Y} \sim \mathcal{N}(\hat{a}, \hat{b}^2)$, where

$$\hat{a} = c_n \overline{Y} + (1 - c_n)a$$
$$\hat{b}^2 = \frac{1}{n/\sigma^2 + 1/b^2}$$
$$c_n = \frac{n/\sigma^2}{n/\sigma^2 + 1/b^2}$$

In other words, the posterior mean is a *weighted average* of sample mean $\overline{Y}$ and prior mean a.

- As $n \to \infty$, $c_n \to 1$: data dominates, $\hat{a} \to \overline{Y}$
- As $b \to \infty$ (weak prior), $c_n \to 1$: data dominates

Example 3: Posterior for ATE

Suppose after an experiment, using observed data $\mathbf{Y} = (Y_1, \ldots, Y_n)$ and assignments $\mathbf{W} = (W_1, \ldots, W_n)$, we compute $\widehat{\text{ATE}}$ and its associated standard error $\widehat{\text{SE}}$.

Likelihood: For large n, by the Central Limit Theorem, the sampling distribution of $\widehat{\text{ATE}} \approx \mathcal{N}(\text{ATE}, \widehat{\text{SE}}^2)$.

Prior: Suppose we assume that $\text{ATE} \sim \mathcal{N}(\mu_0, \sigma_0^2)$. (Later we'll discuss how one might obtain such a prior.)

Posterior: By a similar approach as the previous slide, we can show that $\text{ATE}|\text{data} \approx \mathcal{N}(\mu_n, \sigma_n^2)$, where:

$$\mu_n = c_n \widehat{\text{ATE}} + (1 - c_n)\mu_0; \quad \sigma_n^2 = \frac{1}{1/\widehat{\text{SE}}^2 + 1/\sigma_0^2}; \quad c_n = \frac{1/\widehat{\text{SE}}^2}{1/\widehat{\text{SE}}^2 + 1/\sigma_0^2}$$

Using the posterior for inference

Recall the two main goals of inference:

1. How do we estimate the parameter?
2. How do we quantify uncertainty?

Bayesian inference answers both questions through the posterior distribution.

Estimation

The posterior can be used to compute various estimators:

- *Posterior mean:* $\mathbb{E}_\theta[\theta|\text{data}]$
- *Posterior median:* $\text{median}(\theta|\text{data})$
- *Posterior mode:* The value of θ that maximizes $f(\theta|\text{data})$ (also called the *maximum a posteriori* or *MAP* estimate)

(As we have shown earlier in the class, the posterior mean is the estimate that minimizes expected squared error under the posterior; and the posterior median is the estimate that minimizes expected absolute error under the posterior.)

Quantifying uncertainty: Credible intervals

A $1-\alpha$ *credible interval* is an interval $[L, U]$ such that:

$$\mathbb{P}_\theta(L \leq \theta \leq U | \mathbf{Y}) \geq 1 - \alpha.$$

Comparison to (frequentist) confidence intervals:

- *Confidence interval:* Endpoints $[L, U]$ are *random*, parameter θ is *fixed*
- *Credible interval:* Endpoints $[L, U]$ are *fixed*, parameter θ is *random*
- A credible interval is a direct probabilistic statement about θ.

Example: For a normal posterior $\mathcal{N}(\hat{a}, \hat{b}^2)$, a 95% credible interval is:

$$[\hat{a} - 1.96\hat{b}, \hat{a} + 1.96\hat{b}].$$

Quantifying uncertainty: Taking advantage of the posterior

Because the posterior is a full probability distribution, we can do far more than just compute credible intervals.

We can answer far richer probabilistic questions about parameters, for example:

- "What is $\mathbb{P}_\theta(\theta > 0|\text{data})$?"
- "What is the 90th percentile of the posterior on θ given the data?"
- If there are multiple parameters, e.g., θ_1, θ_2, we can ask questions that involve their *joint* posterior: e.g., "What is $\mathbb{P}_{\vec{\theta}}(\theta_1 > \theta_2|\text{data})$?"

In Bayesian inference, you should not limit yourself to just simple estimates and credible intervals; interrogation of the full posterior distribution is quite valuable and often yields significant insight.

Example: Bayesian causal inference

Recall earlier example of experiment with normal prior on ATE.

Posterior is ATE|data $\sim \mathcal{N}(\mu_n, \sigma_n^2)$.

Estimation from posterior: Mean = median = mode (MAP) = μ_n.

95% credible interval: $[\mu_n - 1.96\sigma_n, \mu_n + 1.96\sigma_n]$.

Probability effect is positive:

$$\mathbb{P}_{\text{ATE}}(\text{ATE} > 0|\text{data}) = 1 - \Phi\left(-\frac{\mu_n}{\sigma_n}\right),$$

where Φ is the standard normal CDF.

This approach to causal inference is called *Bayesian causal inference*.

Comparing Bayesian and frequentist inference

MAP estimation compared to MLE

Recall: posterior $\propto$ likelihood $\times$ prior.

If the prior is *flat* (uniform), then:

$$\arg\max_{\theta} f(\theta|\mathbf{Y}) = \arg\max_{\theta} f(\mathbf{Y}|\theta).$$

(Note: $\arg\max$ means "the θ that achieves the maximum value.")
Therefore, in this case: The MAP estimate (posterior mode) *equals* the MLE.

In general, a non-uniform prior will "shift" the MAP away from the MLE, and towards the prior mode.

Example: Binary outcomes revisited

Recall: $\mathrm{B}(a, b)$ prior with k H's, $n-k$ T's $\implies$ $\mathrm{B}(a+k, b+n-k)$ posterior.

The posterior mode (MAP) is $\text{MAP} = \frac{a+k-1}{a+b+n-2}$.

This can be written as:

$$\text{MAP} = c_n \cdot \text{MLE} + (1 - c_n) \cdot (\text{prior mode}),$$

where $\text{MLE} = k/n$, prior mode $= (a-1)/(a+b-2)$, and $c_n = n/(a+b+n-2)$.

Observe that for large n, $c_n \approx 1$, so MAP $\approx$ MLE.

Example: Normal data revisited

Recall: Normal prior with normal data $\implies$ Normal posterior with mean

$$\hat{a} = c_n \overline{Y} + (1 - c_n)a.$$

Here $\overline{Y}$ is the MLE, a is the prior mean.

For large n: $c_n \approx 1$, so $\hat{a} \approx \overline{Y}$ (the MLE).

(Also for large n: $\hat{b}^2 \approx \sigma^2/n = \text{SE}^2$, the variance of the MLE.)

Example: Bayesian linear regression and regularization

In the appendix, we develop a Bayesian approach to linear regression: we put a prior on the *coefficients* $\boldsymbol{\beta}$, then combine with the data to estimate a posterior distribution.

We show that if (A1)-(A4) hold, then for appropriate choices of prior that put high weight on coefficients being zero, we obtain ridge regression and lasso as the resulting MAP coefficient estimates!

This behavior is similar to the preceding examples: here Bayesian estimation "shrinks" the MAP estimated coefficients towards zero (the prior mode), and away from the OLS solution (the MLE under (A1)-(A4)).

Large sample behavior

The previous observations are more general. In fact, for "reasonable" priors, when the sample size n grows large, the posterior becomes *asymptotically normal*, with:

- Posterior mean/mode $\approx$ MLE
- Posterior variance $\approx \widehat{\text{SE}}^2$ (variance of the MLE)

In other words, when data is sufficient, the prior no longer plays a significant role, so Bayesian and frequentist parameter estimates become similar.

Warning: Despite this superficial similarity, note that regardless of the sample size, there is no frequentist analog of the Bayesian posterior! (Parameters are not random for frequentists.)

When do Bayesian methods work well?

Bayesian methods work well when *prior information matters*.

Example: Suppose we run a small A/B test of a new website feature.

- In hundreds of previous A/B tests, the typical increase in conversion rate was 1-2%.
- In this small test, we observe a 15% increase in conversion rate.
- The frequentist would estimate $\widehat{\text{ATE}} = 15\%$.
- But the Bayesian posterior would "shrink" the estimate toward 1-2%, the historical average.

Here being Bayesian "protects" against overconfident extrapolation from a single, small experiment.

When do Bayesian methods work poorly?

Bayesian methods work poorly when the *prior is not well matched to reality*.

Example: Suppose we test a new promotion to attract customers.

- If previous promotions failed spectacularly, we will necessarily have a *pessimistic* prior (low prior probability of a positive ATE).
- But if the new promotion is actually successful, our strong pessimistic prior makes us less likely to detect the success.

This is a fundamental tradeoff: the stronger the prior, the greater the protection against mistaken overconfidence; but the greater the risk of missing true, novel effects.

Comparing approaches

Frequentist strengths:

- Uses only the data for inferences: e.g., confidence intervals and p-values are "objective" summaries that depend only on the data
- Guarantees on performance of statistical procedures under repetition

Bayesian strengths:

- Leverages available prior information effectively
- Combines prior and data into single distribution (posterior)

Combining methods

In practice, it is often valuable to:

- Ask that Bayesian methods have good frequentist properties
- Ask that frequentist estimates "make sense" given prior understanding

Having both approaches in your toolkit is useful for this reason.

Data-driven priors: Empirical Bayes

Where does the prior come from?

A key question: How do we choose the prior $h(\theta)$?

Two traditional schools of thought:

- *Subjective Bayesian:* Prior encodes subjective beliefs
  - *Example:* If flipping a fair coin, prior should be strongly concentrated around $q = 0.5$
- *Objective Bayesian:* Prior should be "uninformative"
  - *Example:* "Flat" or uniform prior
  - See appendix for more on objective priors

In practice, both approaches have challenges:

- *Subjective priors:* Hard to defend, especially in science/policy
- *Objective priors:* Often not truly "uninformative"; can be overly conservative

Empirical Bayes

Empirical Bayes: Estimate the prior from related data.

Key idea:

- Use historical information: previous experiments; related data scientific studies; meta-analyses; etc.
- Estimate prior from this historical data
- Results in priors that are data-driven

Empirical Bayes

Example: For treatment effect ATE on a new experiment:

- Collect data from m previous experiments on similar interventions
- Observed treatment effects: $\widehat{\text{ATE}}_1, \ldots, \widehat{\text{ATE}}_m$
- Estimate prior: ATE $\sim \mathcal{N}(\hat{a}, \hat{b}^2)$, where $\hat{a}$ is sample mean and $\hat{b}^2$ is sample variance of $\widehat{\text{ATE}}_1, \ldots, \widehat{\text{ATE}}_m$[1]

[1] In practice, an important but subtle issue is that the resulting $\hat{b}^2$ may be too large (i.e., the resulting prior is too "diffuse"), because $\hat{b}^2$ also uncertainty due to finite sample sizes in each experiment. This can be corrected by reducing the prior variance lower than $\hat{b}^2$ in a data-driven manner (beyond the scope of this lecture).

Example: Empirical Bayes for course descriptions

Suppose our online learning platform has run many previous A/B tests:

- User interface changes, email campaigns, recommendation tweaks, etc.
- *For each:* Suppose we estimated $\widehat{\text{ATE}}_m$ – effect on enrollment rate

The empirical Bayes approach constructs a prior from this previous experimental data.

Note that we might want to *weight* some experiments more heavily in this prior construction: e.g., those that specifically focused on changes to course descriptions.

Benefits of empirical Bayes

Empirical Bayes has several key features:

1. *Objectivity*: It is relatively objective, in the sense that the prior is grounded in historical data and observed outcomes.
2. *Shrinkage*: Posterior estimates are "pulled" towards the prior mean or mode ("shrunk" towards the historical norm), with a strength that depends on the sample size.
3. *Flexibility*: The prior distribution can be quite complex/flexible, both to better match historical data, and to select historical experiments that are most relevant to the current experiment.

The winner's curse [$*$]

The "shrinkage" effect of empirical Bayes directly addresses a key issue in experimentation, known as the "winner's curse":

Suppose we run an experiment comparing multiple variations of a web page. If we pick the "best" based on the observed estimated effects, we will have *overestimated* the true ATE of the "winner" on average.

(This is a similar effect seen in prediction: the validation set error of the winning model underestimates that model's true generalization error.)

Empirical Bayes shrink all estimates toward the historical mean, but in particular, the "winner" gets shrunk the *most* (because it is further from the historical mean).

This has the consequence of mitigating the "winner's curse" bias.

Decisions

From inference to decisions

Recall our course platform that ran an experiment with new course descriptions.

Using the previous sections, we can construct a posterior for ATE (the enrollment rate per course).

Should we roll out the new course descriptions? This isn't just inference; it's a *decision*.

Need to simultaneously consider:

- Benefits
- Costs
- Uncertainty about the true ATE

Example: Course description rollout decision

To be concrete, consider the following situation:

- $K = 1000$ courses
- ATE represents additional enrollment rate, i.e., expected additional enrollments per course
- Suppose that experiment produces $\widehat{\text{ATE}} = 0.02$, and $\widehat{\text{SE}} = 0.015$.
- Note significant uncertainty: 95% frequentist confidence interval is $\approx[-0.01, 0.05]$, which includes zero $\implies$ do not reject the null hypothesis $H_0 : \text{ATE} = 0$.

Example: Course description rollout decision

Let's translate this to a Bayesian decision problem:

- Posterior from experiment: ATE|data $\sim \mathcal{N}(0.02, 0.015^2)$
- Suppose revenue from each enrollment: $\$400$
- Suppose implementation cost platform-wide: $C = \$5000$
- *Decision:* Roll out or not?

Formalizing the decision

Decisions: roll out, or don't roll out, new course descriptions platform-wide
Unknown parameter: True effect ATE

Consequences: Financial outcomes that depend on action and true ATE

Utilities:

$$u(\text{roll out}, \text{ATE}) = K \cdot \text{ATE} \cdot 400 - C \qquad (\text{gain } K \cdot \text{ATE} \cdot 400, \text{ pay cost } C)$$

$$u(\text{don't roll out}, \text{ATE}) = 0 \qquad (\text{no benefit, no cost})$$

The decision rule

Since ATE is uncertain, evaluate each action by its *expected utility* under the posterior.

Expected utility of rolling out:

$$\begin{aligned}\mathbb{E}_{\text{ATE}}[u(\text{roll out}, \text{ATE}) \mid \text{data}] &= \mathbb{E}_{\text{ATE}}[K \cdot \text{ATE} \cdot 400 - C \mid \text{data}] \\ &= K \cdot \mathbb{E}_{\text{ATE}}[\text{ATE} \mid \text{data}] \cdot 400 - C \\ &= 1000 \times 0.02 \times 400 - 5000 = \$3000\end{aligned}$$

Expected utility of not rolling out:

$$\mathbb{E}_{\text{ATE}}[u(\text{don't rollout}, \text{ATE}) \mid \text{data}] = 0$$

Decision: Choose action maximizing expected utility $\implies$ *roll out*.

Why expected utility? [∗]

Under reasonable axioms, it can be shown that expected utility maximization is the *coherent* way to make decisions under uncertainty.

Formal foundation: Von Neumann-Morgenstern axioms (see appendix).

- Basic rationality requirements on preferences: completeness, transitivity, continuity, independence
- These imply preferences can be represented by maximizing expected utility

We won't derive the axioms here, but this provides the normative justification for expected utility maximization.

General Bayesian decision framework

Now abstract from the example:

Elements:

- *Actions:* $a \in \mathscr{A}$
- *Unknown parameter:* θ
- *Utility:* $u(a, \theta)$ (or *loss* $L(a, \theta) = -u(a, \theta)$)
- *Posterior:* $h(\theta \mid \text{data})$

Bayes-optimal decision:

$$\begin{aligned} d^* &= \arg\max_d \mathbb{E}_\theta[u(d, \theta) \mid \text{data}] \\ &= \arg\min_d \mathbb{E}_\theta[L(d, \theta) \mid \text{data}] \end{aligned}$$

The expected loss $\mathbb{E}_\theta[L(d, \theta)|\text{data}]$ is called the *Bayes risk*.

Incorporating risk aversion

Our simple decision rule: roll out if $\mathbb{E}_{\text{ATE}}[\text{ATE}|\text{data}]$ is large relative to cost ratio.

But this ignores *risk*: Large uncertainty σ_n should make us more cautious...

Approaches to incorporate risk:

1. Use a nonlinear utility function: a *concave* $u(x)$, e.g., $u(x) = \sqrt{x}$, will make the decision maker *risk averse*
2. Add a penalty for variance: $\mathbb{E}_\theta[u(d,\theta)|\text{data}] - \lambda \operatorname{Var}_\theta(u(d,\theta)|\text{data})$
3. Require higher confidence: roll out only if $\mathbb{P}_{\text{ATE}}(\text{ATE} > C/N|\text{data}) > 0.95$

The first of these is the usual way to incorporate risk aversion in a decision-theoretic framework; the second and third do not strictly correspond to expected utility maximization, but can be useful practical heuristics.

Comparison to frequentist approach

In industrial practice, it is common to rely on frequentist hypothesis testing to make decisions:

- Test H_0 : ATE $= 0$ vs. H_1 : ATE $\neq 0$
- Reject H_0 if p-value $< \alpha$ (e.g., $\alpha = 0.05$)
- If H_0 rejected and $\widehat{\text{ATE}}$ is positive, roll out; otherwise, do not roll out

But this approach has many limitations:

- Ignores prior (historical) knowledge
- Does not incorporate costs or benefits, and ignores uncertainty in the magnitude of the effect
- Choice of α not explicitly related to risk preferences

All of these issues are coherently addressed in a Bayesian decision-making framework.

Complete Bayesian decision-making workflow

1. *Prior:* Use empirical Bayes (historical experiments) to obtain prior ATE $\sim \mathcal{N}(\mu_0, \sigma_0^2)$
2. *Experiment:* Run A/B test, obtain $\widehat{\text{ATE}}$ and $\widehat{\text{SE}}$
3. *Posterior:* Compute ATE|data $\sim \mathcal{N}(\mu_n, \sigma_n^2)$
4. *Decision:* Compute expected utilities, choose decision that maximizes expected utility

Additional (advanced) steps can include *sensitivity analysis* (checking robustness of the decision to choice of prior), and calculating the *value of information* (i.e., determining whether gathering more data would be worth the cost).

Computational considerations

Computation: Modern tools

Even simple Bayesian problems can be computationally challenging:

- Computing posterior involves integration over the full parameter space, which can be complex
- Conjugate priors are special cases; most problems don't have closed form posterior calculations

However, modern computational techniques such as Markov chain Monte Carlo and generative modeling have meant that computation is no longer a barrier to Bayesian methodology in most applied domains.

Markov Chain Monte Carlo (MCMC) [$*$]

MCMC is a computational methods for sampling from the posterior:

- Generate samples $\theta_1, \ldots, \theta_B \sim f(\theta|\text{data})$
- Use samples to approximate posterior mean, credible intervals, etc.
- Conceptual idea: construct a Markov chain whose stationary distribution is the posterior

Modern generative models [$*$]

More recent advances in AI models have also had significant impact on Bayesian inference.

- Generative AI models can be used to sample from complex posteriors, without explicit distributional representation
- Examples: diffusion models, neural samplers

Appendix: Bayesian linear regression [$*$]

Bayesian linear regression [$*$]

For simplicity, assume that (A1)-(A4) hold: $Y_i = \mathbf{X}_i\boldsymbol{\beta} + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

In *Bayesian linear regression*, we also assume a *prior distribution* on $\boldsymbol{\beta}$.

As we'll see, the choice of prior is connected to *regularization*:

- Normal prior $\Rightarrow$ ridge regression
- Laplace prior $\Rightarrow$ lasso regression

(Technical note: We'll assume σ^2 is known. In practice, we use an estimate $\hat{\sigma}^2$ from the data.)

Ridge regression and lasso [$*$]

Recall that ridge regression and lasso are regularized regression techniques:
Ridge regression chooses $\hat{\boldsymbol{\beta}}$ to minimize:

$$\sum_{i=1}^{n}(Y_i - \mathbf{X}_i\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Lasso chooses $\hat{\boldsymbol{\beta}}$ to minimize:

$$\sum_{i=1}^{n}(Y_i - \mathbf{X}_i\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Ridge regression as Bayesian regression [$*$]

Suppose the prior on $\boldsymbol{\beta}$ is:

$$\boldsymbol{\beta} \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{\lambda\sigma^2}\mathbf{I}\right).$$

This prior encodes the belief that coefficients should be close to zero, with strength controlled by λ.

Key result: The MAP estimator (posterior mode) is exactly the ridge regression solution.
Interpretation: The MAP estimate balances the likelihood (data fit) with the prior (regularization toward zero). Higher λ = stronger prior = more "shrinkage" of coefficients.

Lasso as Bayesian regression [$*$]

Suppose the prior on each β_j is independent *Laplace*:

$$h(\beta_j) = \left(\frac{\lambda}{2\sigma}\right) \exp\left(-\frac{\lambda|\beta_j|}{\sigma}\right).$$

The Laplace distribution is symmetric around zero, but *more peaked at zero* than the normal distribution.

Key result: The MAP estimator is exactly the lasso solution.

Appendix: Objective priors [$*$]

Objective Bayesian inference [$*$]

Goal: Choose priors that are "uninformative" or "objective".

Motivation:

- Subjectivity should not enter scientific conclusions
- Want conclusions driven by data, not prior beliefs

Approaches:

- Flat (uniform) priors
- Improper priors
- Jeffreys' priors (transformation-invariant)

Flat priors [$*$]

A *flat prior* is uniform over the parameter space: $f(\theta) = c$ (constant).

Example: For coin bias $q \in [0, 1]$, flat prior is $f(q) = 1$ (uniform on $[0, 1]$).

When the prior is flat:

$$\text{posterior} \propto \text{likelihood} \times \text{constant} \propto \text{likelihood}.$$

So the MAP estimate equals the MLE!

Problem: What if the parameter space is unbounded (e.g., $\theta \in \mathbb{R}$)?

- A flat prior $f(\theta) = c$ cannot be a probability distribution
- But we might still be able to define a posterior…

Improper priors [$*$]

An *improper prior* is a "prior" that is not a valid probability distribution.

Example: $Y_1, \ldots, Y_n \sim \mathcal{N}(\mu, 1)$, with flat prior $f(\mu) = 1$ (constant) on $\mathbb{R}$.

The posterior is:

$$\begin{aligned} f(\mu|\mathbf{Y}) &\propto f(\mathbf{Y}|\mu) f(\mu) \\ &\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(Y_i - \mu)^2\right) \times 1 \\ &\propto \exp\left(-\frac{n}{2}(\mu - \overline{Y})^2\right). \end{aligned}$$

This is proportional to $\mathcal{N}(\overline{Y}, 1/n)$, which *is* a valid distribution!

So even though the prior is improper, the posterior is well-defined.

Improper priors: Interpretation [$*$]

Using a flat improper prior:

- Encodes "no information" about θ before seeing data
- Posterior is entirely driven by likelihood
- MAP = MLE

Conservative approach:

- Most conservative thing is to assume no knowledge except from data
- Flat prior is meant to encode this

Caution: Not all improper priors yield well-defined posteriors!

- Need to check that posterior can be normalized

Jeffreys' priors [$*$]

Problem with flat priors: Not invariant to transformations.

Example: If $f(\mu) = 1$ (flat), what is the prior on μ^2?

$$f(\mu^2) = \frac{f(\mu)}{2\mu} = \frac{1}{2\mu}.$$

This is *not* flat!

Jeffreys' prior: Choose $f(\theta) \propto \sqrt{I(\theta)}$, where $I(\theta)$ is the Fisher information.

Properties:

- Transformation-invariant
- "Uninformative" in a well-defined sense
- See Wasserman, *All of Statistics*, Section 11.6 for details

Appendix: Decision theory foundations [$*$]

Decisions under uncertainty [$*$]

In our decision framework, we choose decisions whose consequences depend on unknown parameters.

Example from lecture: Should we roll out new course descriptions?

- Consequence depends on true ATE (unknown)
- Posterior gives us probabilities: ATE | data $\sim \mathcal{N}(\mu_n, \sigma_n^2)$

In this appendix, we frame such choices in the language of *decision theory*, and use this framing to motivate expected utility maximization as optimal decision-making.

Decisions under uncertainty [$*$]

General setting: A *lottery* L specifies probabilities over possible *outcomes*.

- *Outcomes* represent possible "states of the world", i.e., actual ground truth.
- A *lottery* represents probability distributions *over* outcomes. Note that the base (deterministic) outcomes are also representable as "lotteries" – they are probability distributions that put probability 1 on a deterministic outcome.
- A decision maker has *preferences* over lotteries, represented by $\succ$. In particular, $L_1 \succ L_2$ means the lottery L_1 is strictly preferred to the lottery L_2. (We write $L_1 \succsim L_2$ if $L_1 \succ L_2$, or if the decision-maker is indifferent between L_1 and L_2.)

In our case: Outcomes are determined by the true ATE, as well as any other associated costs and/or benefits of implementation. The lottery over outcomes is determined by the *posterior*, since this gives a distribution over possible values of ATE.

"Rational" preferences [$*$]

Decision theory imposes constraints of preferences motivated by "rational" behavior.

Example: Consider preferences over three options A, B, C (could be lotteries, or just fixed outcomes).

Suppose your preferences are:

- A is preferred over B ($A \succ B$)
- B is preferred over C ($B \succ C$)
- C is preferred over A ($C \succ A$)

In each case, specifically suppose that the preferred option is greater than $\$100$ more preferred than the less preferred option.

"Rational" preferences [$*$]

These preferences are problematic due to their "circularity." In particular, consider the following sequence of trades:

- You currently own A.
- I offer to swap A for C, charging you $\$100$, and you accept.
- I offer to swap C for B, charging you $\$100$, and you accept.
- I offer to swap B for A, charging you $\$100$, and you accept.
- You're back to A, but you've paid out $\$300$!

Decision theory introduces axiomatic constraints on preferences to ensure *coherence*, i.e., to rule out problematic behavior such as *sure losses*.

Coherent preferences: The vNM axioms [$*$]

Von Neumann and Morgenstern (1944) identified conditions on preferences that prevent such incoherence in preferences.

Four axioms on preferences $\succsim$ over lotteries:

1. *Completeness:* For any L_1, L_2: either $L_1 \succsim L_2$ or $L_2 \succsim L_1$ (can always compare)
2. *Transitivity:* If $L_1 \succsim L_2$ and $L_2 \succsim L_3$, then $L_1 \succsim L_3$ (prevents circularity)
3. *Continuity:* If $L_1 \succsim L_2 \succsim L_3$, then there exists $p\,(0 \leq p \leq 1)$ such that $L_2 \sim pL_1 + (1-p)L_3$
4. *Independence:* If $L_1 \succsim L_2$ and $0 \leq p \leq 1$, and L_3 is any lottery, then $pL_1 + (1-p)L_3 \succsim pL_2 + (1-p)L_3$.

Coherent preferences: The vNM theorem [$*$]

Using the preceding axioms, we have the following theorem.
Theorem (vNM): If $\succsim$ satisfies the preceding four axioms, then there exists a utility function u over outcomes x such that:

$$L_1 \succsim L_2 \iff \mathbb{E}_{L_1}[u(x)] \geq \mathbb{E}_{L_2}[u(x)].$$

In other words Coherent preferences correspond to ranking lotteries by *expected utility* (for some corresponding utility function).

Connection to Bayesian decision making [$*$]

In our setting:

- We have a posterior $h(\theta \mid \text{data})$ over unknown parameter θ.
- Each pair of a possible decision d and parameter θ determines the outcome.
- The posterior gives us probabilities over θ, so each decision d corresponds to a lottery over outcomes.

Connection to Bayesian decision making [$*$]

With the previous setting, the vNM axioms apply:

If our preferences over decisions satisfy the four axioms, then there is a utility function u such that the optimal decision *maximizes expected utility*:

$$\text{Choose } d^* = \arg\max_d \mathbb{E}_\theta[u(d, \theta) \mid \text{data}].$$

This is exactly the Bayesian decision rule we used in the lecture.

In other words, expected utility maximization isn't just reasonable qualitatively—in the quantitative sense justified by the vNM axioms, it's the only coherent way to make decisions under uncertainty (given the posterior).