Philosophy of Bayesian Reasoning: Study Notes
Session 1: Probability as Plausible Reasoning

Based on Jaynes (2003)

Instructor: Sudeep Regmi

Session Goal

To understand the philosophical motivation behind the Bayesian interpretation of probability, particularly as articulated by E.T. Jaynes. We will explore probability not primarily as a description of frequencies in the physical world, but as a tool for logical and consistent reasoning in the face of uncertainty—an extension of deductive logic.

Key Questions for this Session

- What is probability fundamentally measuring or describing?
- Can we have a 'logic' for situations where deductive certainty is impossible?
- How does the Bayesian view of probability differ from the more common 'frequency' view taught in introductory statistics?
- What basic principles should any system of rational reasoning under uncertainty obey?
- Why is the 'background information' or 'conditioning context' so crucial in probability theory?

1 The Problem: Uncertainty and the Limits of Logic

Deductive Logic

Classical deductive logic (Aristotelian logic, Boolean algebra) provides rules for determining the truth or falsity of propositions based on the assumed truth of other propositions (premises).

- **Example:** If Premise 1 ("All humans are mortal") and Premise 2 ("Socrates is human") are true, then the Conclusion ("Socrates is mortal") is necessarily true.
- **Limitation:** Deductive logic only works when we have complete certainty in our premises and clear rules of implication. It cannot handle situations involving incomplete information, ambiguity, or degrees of certainty.

The Ubiquity of Uncertainty

In most real-world situations, especially in scientific inquiry, data analysis, and everyday decision-making, we operate with incomplete information.

- We rarely know if a hypothesis is absolutely true or false.
- Data is often noisy, limited, or indirect.
- We need to weigh evidence, update beliefs, and make judgments based on plausibility, not certainty.

Conclusion: We need a framework for reasoning that goes beyond deductive logic to handle degrees of plausibility or belief.

2 Jaynes's Central Thesis: Probability as Extended Logic

Key Idea 2.1 (Probability as Plausible Reasoning)**.** E.T. Jaynes forcefully argues that probability theory is, or ought to be understood as, an extension of deductive logic to situations of uncertainty. It's a system for conducting **plausible reasoning** in a consistent and rational manner.

- Probability quantifies a **degree of belief** (or plausibility, confidence) that a proposition is true, given a specific state of knowledge or information.
- It is fundamentally **epistemic** (relating to knowledge), rather than purely **aleatory** (relating to randomness in the physical world).

This perspective shifts the focus:

- **From:** Probability as an inherent property of a system (like the bias of a physical coin).
- **To:** Probability as a reflection of our *state of information* about the system. If our information changes, the probability changes, even if the physical system remains the same.

3 Contrasting Interpretations of Probability

Understanding the Bayesian perspective requires contrasting it with the dominant alternative, the Frequentist interpretation.

3.1 The Bayesian (Epistemic/Logical) Interpretation

- **Core Idea:** Probability measures a degree of belief, plausibility, or confidence in a proposition being true, conditional on the information currently available (I). We write this as $P(A|I)$.
- **Origin:** Rooted in the work of thinkers like Laplace, Bayes, Jeffreys, Cox, and Jaynes.
- **Nature of Probability:** Epistemic – resides in the mind of the observer (or an idealized rational agent) representing their state of knowledge.
- **Scope:** Can assign probabilities to *any* logical proposition, including hypotheses (e.g., "the probability that this specific patient has disease X," "the probability that Einstein's theory is correct given the evidence").

- **Subjectivity/Objectivity:** Often critiqued as "subjective," but proponents like Jaynes argue for an "objective" Bayesianism where probabilities are determined uniquely by the specified information and desiderata of rational thought. The belief is personal, but the rules for manipulating beliefs should be objective.

3.2 The Frequentist (Aleatory) Interpretation

- **Core Idea:** Probability is the long-run relative frequency of an event occurring in a series of repeatable, identical trials.
- **Origin:** Developed by Venn, Fisher, Neyman, Pearson, von Mises. Became dominant in the 20th century.
- **Nature of Probability:** Aleatory – considered an objective property of the physical system or data-generating process in the real world (propensity).
- **Scope:** Strictly applicable only to events that can be considered outcomes of potentially infinitely repeatable experiments (e.g., coin flips, drawing cards, sampling from a population). It is difficult (or arguably impossible) within this framework to assign a probability to a fixed hypothesis being true (e.g., "The mass of the electron *is* m") because a hypothesis is either true or false, it doesn't "happen" with a frequency.
- **Objectivity:** Claimed as a major strength – probabilities are objective features of the world, independent of observer belief.

Key Differences Summarized

Feature	Bayesian (Jaynes)	Frequentist
Definition	Degree of belief/plausibility conditional on information	Long-run relative frequency in repeatable trials
Nature	Epistemic (state of knowledge)	Aleatory (property of the world)
Applies to	Any logical proposition (incl. hypotheses)	Outcomes of repeatable experiments
Notation	$P(A\|I)$ is fundamental	$P(A)$ often implies a fixed context
Source of Rules	Desiderata of rational reasoning (Cox/Jaynes)	Axioms of measure theory (Kolmogorov)

4 Desiderata for Plausible Reasoning (Jaynes/Cox)

If probability is to be an extension of logic, its rules shouldn't be arbitrary axioms but should arise from simple, common-sense requirements for rational reasoning under uncertainty. Jaynes (following R.T. Cox) proposed desiderata (desirable properties) that any system quantifying plausibility should satisfy. We focus on the concepts here (Chapter 2 of Jaynes contains the mathematical derivation, which we are skipping):

(1) Representation of Plausibility: Degrees of plausibility are represented by real numbers. A higher number should indicate greater plausibility.

(2) Qualitative Correspondence with Logic:

- If a conclusion can be reasoned out in multiple ways, the resulting plausibility should be the same (consistency).
- The system must reduce to deductive logic in the limit of certainty (e.g., if A is known to be true given I, its plausibility should represent certainty; if A is known to be false given I, its plausibility should represent impossibility).
- Plausibility should change sensibly with new information (e.g., learning supporting evidence should increase plausibility).

(3) Consistency and Universality: The rules for combining plausibilities should be universally applicable and consistent, regardless of the specific propositions involved. All available evidence relevant to a question must be used.

Key Idea 4.1 (Cox's Theorem - Conceptual Outcome)**.** Jaynes argues (based on Cox's mathematical proof) that these simple desiderata **uniquely** lead to the standard rules of probability theory (the sum rule and the product rule), up to a monotonic transformation (like using decibels instead of raw probabilities). This provides a powerful justification: the rules of probability are not arbitrary conventions but are *necessary consequences* of demanding rational, consistent reasoning under uncertainty. We will explore these rules in Session 2.

5 The Crucial Role of Conditioning Information: P(A—B)

A cornerstone of the Jaynesian/Bayesian perspective is that probability is *always* conditional.

- $P(A|B)$ reads: "The probability (plausibility) of proposition A being true, given that proposition B (representing our background knowledge, data, assumptions) is true."
- There is no such thing as an absolute or unconditional probability. A probability value only has meaning relative to a specified state of information (B).
- What might seem like an unconditional probability, $P(A)$, is implicitly $P(A|I)$ where I represents the assumed (often vast and unstated) background information.
- **Example:** The probability of rain tomorrow depends crucially on whether our information B includes today's weather patterns, satellite images, historical data, etc. $P(\text{Rain}|\text{Cloudy sky})$ is different from $P(\text{Rain}|\text{Clear sky})$.
- Recognizing the conditioning context is essential for avoiding paradoxes and misinterpretations. Many disagreements arise from different parties implicitly using different background information B.

6 Examples of Plausible Reasoning

Think about how we intuitively reason in situations without certainty:

- **Medical Diagnosis:** A doctor combines patient symptoms (evidence) with knowledge of disease prevalence and symptom patterns (background information) to assess the plausibility of different diagnoses.

- **Detective Work:** A detective evaluates the plausibility of a suspect's guilt based on accumulating clues (evidence) and general knowledge about motives and means (background information).
- **Scientific Inference:** A scientist evaluates the plausibility of a hypothesis based on experimental results (evidence) within the context of existing theories and knowledge (background information).
- **Everyday Life:** Deciding whether to bring an umbrella based on the look of the sky and the weather forecast.

The Bayesian framework aims to provide a formal, consistent calculus for performing these kinds of reasoning tasks.

7 Discussion Points

- Where does uncertainty primarily arise in your own field of study or research? Is it typically treated as randomness in the world, or lack of knowledge?
- What is your initial reaction to the idea of probability as a 'degree of belief'? Does it seem too subjective for science? Or does it capture something essential about how we reason?
- Can you think of examples where the 'Frequentist' definition of probability seems difficult or impossible to apply? (e.g., probability of a unique historical event, probability of a fundamental constant having a certain value).
- How important are the 'desiderata' (consistency, correspondence with logic)? Would you trust a system of reasoning that violated them?

Reading Guidance

- **Primary:** E.T. Jaynes, "Probability Theory: The Logic of Science," **Chapter 1: Plausible Reasoning**. Focus on the arguments for needing an extension of logic and the description of plausible reasoning.
- **Skim/Conceptual:** Jaynes, **Chapter 2: The Quantitative Rules**. Focus on understanding the *statement* of the desiderata at the beginning, and the *claim* that these lead uniquely to the probability rules. Do not worry about following the detailed mathematical derivations unless you are comfortable with functional equations. The key takeaway is the philosophical justification for the rules.
- **Optional Supplementary:** Search for introductory articles or blog posts comparing Bayesian and Frequentist interpretations of probability for additional perspectives.

Looking Ahead to Session 2

Having established the philosophical goal (probability as extended logic) and the desiderata for rational reasoning, Session 2 will delve into the specific rules that emerge from these desiderata: the Sum Rule and the Product Rule. We will see how these naturally lead to **Bayes' Theorem**, the cornerstone of Bayesian inference for updating beliefs in light of evidence.

Philosophy of Bayesian Reasoning: Study Notes
Session 2: The Rules of the Game & Bayes' Theorem

Based on Jaynes (2003)

Instructor: Sudeep Regmi

Session Goal

To understand the fundamental rules governing the manipulation of probabilities (degrees of belief) derived from the desiderata of rational reasoning discussed in Session 1. We will introduce the Product Rule and Sum Rule, show how they lead directly to Bayes' Theorem, and interpret this theorem as the core mechanism for rational belief updating in light of new evidence.

Key Questions for this Session

- What specific mathematical rules emerge from the demand for consistent plausible reasoning?
- How do we calculate the plausibility of multiple propositions being true? Or at least one of them being true?
- How does Bayes' Theorem arise naturally from these basic rules?
- What do the different terms in Bayes' Theorem represent conceptually?
- Why is Bayes' Theorem considered the fundamental rule for learning from data/evidence?

1 Recap: From Desiderata to Quantitative Rules

In Session 1, we established the Jaynesian perspective: probability theory as an extension of logic, quantifying degrees of plausibility ($P(A|I)$). We discussed the desiderata (qualitative requirements) for any such system of reasoning:

- Representation by real numbers.
- Qualitative correspondence with logic (consistency, limits to certainty).
- Consistency and universality in application.

We introduced the conceptual result of Cox's Theorem: these desiderata uniquely determine the fundamental rules for manipulating probabilities. Today, we examine these rules.

Key Idea 1.1 (Foundation of Rules)**.** The rules of probability (Sum Rule, Product Rule) are not arbitrary axioms chosen for mathematical convenience. In the Jaynes/Cox framework, they are **derived consequences** of requiring consistent plausible reasoning. This provides a strong philosophical grounding.

Contrast: Kolmogorov Axioms The standard mathematical approach (Kolmogorov, 1933) defines probability based on measure theory, starting with axioms about sets:

1. Non-negativity: $P(E) \geq 0$ for any event E.
2. Normalization: $P(\Omega) = 1$ where Ω is the sample space (all possibilities).
3. Additivity (for mutually exclusive events $E_1, E_2, ...$): $P(E_1 \cup E_2 \cup ...) = \sum P(E_i)$.

Conditional probability is then typically *defined* as $P(A|B) = P(A \cap B)/P(B)$. While mathematically powerful and consistent with the derived rules, Jaynes argued this approach lacks the direct philosophical motivation connecting probability to reasoning and information.

2 The Fundamental Rules of Probability Logic

Let A, B, C represent propositions. C represents the background information or context, which is always present.

- AB (or $A \wedge B$): Represents the conjunction "A and B are both true".
- $A + B$ (or $A \vee B$): Represents the disjunction "At least one of A or B is true".
- $\overline{A}$ (or $\neg A$): Represents the negation "A is false".

2.1 The Product Rule (Rule for Conjunction)

Theorem 2.1 (The Product Rule)**.** The plausibility of both A and B being true, given C, can be decomposed as:

$$\begin{aligned} P(AB|C) &= P(A|BC)P(B|C) \\ &= P(B|AC)P(A|C) \end{aligned}$$

- **Interpretation:** To find the plausibility that both A and B are true (given C), we first consider the plausibility of B being true (given C). Then, *assuming* B is true (hence the context becomes BC), we consider the plausibility of A also being true. The overall plausibility is the product of these two stages. The second line shows symmetry – we could start with A and then consider B given A.
- **Connection to Conditional Probability:** If $P(B|C) \neq 0$, we can rearrange the first line to define the conditional probability of A given B (and C):

  $$P(A|BC) = \frac{P(AB|C)}{P(B|C)}$$

  This matches the conventional definition but arises here as part of the fundamental rule derived from desiderata, rather than being a definition itself.
- **Independence:** Two propositions A and B are said to be *independent* given C if learning B does not change the plausibility of A. Mathematically: $P(A|BC) = P(A|C)$. In this specific case, the product rule simplifies to $P(AB|C) = P(A|C)P(B|C)$. *Caution:* Independence is relative to the background information C!

2.2 The Sum Rule (Rule for Disjunction)

Theorem 2.2 (The Sum Rule)**.** The plausibility that at least one of A or B is true, given C, is:

$$P(A + B|C) = P(A|C) + P(B|C) - P(AB|C)$$

- **Interpretation:** We add the plausibilities of A and B individually, but then subtract the plausibility of their conjunction (both being true) to avoid double-counting the situation where A and B overlap. Think of Venn diagrams.
- **Mutually Exclusive Events:** If A and B cannot both be true given C (i.e., they are mutually exclusive), then $P(AB|C) = 0$. The sum rule simplifies to:

  $$P(A + B|C) = P(A|C) + P(B|C) \quad \text{(if AB = False given C)}$$

  This is the basis of Kolmogorov's third axiom, but here it's a consequence of the more general sum rule.
- **Negation:** A proposition A and its negation $\overline{A}$ are mutually exclusive ($P(A\overline{A}|C) = 0$) and exhaustive ($A + \overline{A}$ is always true, representing certainty). The sum rule applied to $A + \overline{A}$ gives:

  $$P(A + \overline{A}|C) = P(A|C) + P(\overline{A}|C)$$

  Since $A + \overline{A}$ represents certainty, its plausibility must be 1 (by correspondence with logic). Therefore:

  $$1 = P(A|C) + P(\overline{A}|C) \quad \Longrightarrow \quad P(\overline{A}|C) = 1 - P(A|C)$$

  The plausibility of A being false is one minus the plausibility of A being true.

3 Deriving Bayes' Theorem

Bayes' Theorem is not a new rule, but a direct consequence of applying the product rule in two ways.

Key Idea 3.1 (Bayes' Theorem Derivation)**.** Start with the two symmetric forms of the product rule for the conjunction AB given C:

1. $P(AB|C) = P(A|BC)P(B|C)$
2. $P(AB|C) = P(B|AC)P(A|C)$

Since both right-hand sides equal $P(AB|C)$, they must equal each other:

$$P(A|BC)P(B|C) = P(B|AC)P(A|C)$$

Assuming $P(B|C) > 0$, we can divide both sides by $P(B|C)$ to isolate $P(A|BC)$:

$$P(A|BC) = \frac{P(A|C)P(B|AC)}{P(B|C)}$$

This elegant result is Bayes' Theorem.

4 Interpreting Bayes' Theorem for Inference

To make the theorem's application to inference clearer, let's relabel the propositions:

- H: A specific hypothesis we are interested in.
- E: New evidence or data relevant to the hypothesis.
- I: The background information or context (previously denoted by C).

Substituting these into the derived equation $P(A|BC) = P(A|C)P(B|AC)/P(B|C)$, we get the most common form of Bayes' Theorem used in statistical inference:

Theorem 4.1 (Bayes' Theorem)**.**

$$P(H|E,I) = \frac{P(H|I)P(E|H,I)}{P(E|I)}$$

Or, often written focusing on proportionality (since $P(E|I)$ is a normalizing constant independent of H):

$$\boxed{P(H|E,I) \propto P(H|I)P(E|H,I)}$$

Let's break down the terms:

- $P(H|E,I)$: The **Posterior Probability**. This is the updated degree of belief (plausibility) in hypothesis H *after* considering the evidence E, within the context of background information I. This is typically what we want to compute.
- $P(H|I)$: The **Prior Probability**. This is the initial degree of belief in hypothesis H *before* considering the specific evidence E, based only on the background information I. This represents our state of knowledge at the start. (Topic of Session 3!)
- $P(E|H,I)$: The **Likelihood**. This is the probability (plausibility) of observing the evidence E *if* the hypothesis H were true, within the context I. It quantifies how well the hypothesis H predicts or explains the data E. Note: This is a probability of the *data* given the *hypothesis*, not the other way around!
- $P(E|I)$: The **Evidence Probability** (also called Marginal Likelihood or Normalizing Constant). This is the overall probability of observing the evidence E, considering all possible hypotheses, weighted by their prior probabilities. It is calculated using the law of total probability:

  $$P(E|I) = \sum_j P(E|H_j,I)P(H_j|I)$$

  where the sum is over all mutually exclusive and exhaustive hypotheses H_j. Its main role is to ensure that the posterior probabilities $P(H|E,I)$ sum to 1 over all possible H. In many applications where we only compare relative plausibilities of different H's, this term can be ignored (hence the proportionality relation).

Key Idea 4.1 (Bayes' Theorem as Belief Updating)**.** The formula Posterior $\propto$ Prior $\times$ Likelihood shows precisely how rational belief should change:

- Start with a prior belief $P(H|I)$.

- Observe evidence E.
- Evaluate how likely that evidence was under hypothesis H (the likelihood $P(E|H, I)$).
- Multiply the prior by the likelihood to get the (unnormalized) posterior belief.
- Normalize (divide by $P(E|I)$) if absolute probabilities are needed.

Bayes' Theorem is thus the engine of learning from experience in a logically consistent way.

5 Simple Example: Medical Diagnosis (Conceptual)

Imagine a patient presents with a symptom E (e.g., a cough). We are interested in the hypothesis H that the patient has a specific disease (e.g., bronchitis). Our background information I includes general medical knowledge.

- **Prior** $P(H|I)$: What is the general prevalence of bronchitis in the relevant population? This is our belief before seeing the patient's specific symptom. Maybe it's relatively low, say 5
- **Likelihood** $P(E|H, I)$: How likely is a cough (E) if the patient *does* have bronchitis (H)? Let's say it's quite likely, e.g., 75
- **Evidence** $P(E|I)$: How likely is a cough in general in this population? This requires considering bronchitis *and* other causes (flu, allergy, etc.). We need $P(E|H, I)P(H|I)$ (cough given bronchitis) + $P(E|\overline{H}, I)P(\overline{H}|I)$ (cough given no bronchitis). Let's say the probability of a cough if the patient *doesn't* have bronchitis is 10
- **Posterior** $P(H|E, I)$: Now we apply Bayes' Theorem:

$$P(H|E, I) = \frac{P(H|I)P(E|H, I)}{P(E|I)} = \frac{0.05 \times 0.75}{0.1325} = \frac{0.0375}{0.1325} \approx 0.283$$

  Our belief in the patient having bronchitis increased from 5

This example illustrates how Bayes' Theorem combines prior knowledge with new evidence in a quantitative way.

6 Discussion Points

- Does the derivation of the probability rules from simple desiderata seem more compelling than simply stating them as axioms?
- Can you rephrase the Product Rule and Sum Rule in plain English using concepts like "assuming X is true" or "at least one"?
- Consider the components of Bayes' Theorem. Which part represents your initial assumptions? Which part represents the new information from data? Which part represents the updated belief?
- Why is the likelihood $P(E|H, I)$ not the same as the posterior $P(H|E, I)$? Why is this distinction important (and often confused)?
- Can you think of a simple reasoning process in your field that could be modeled (even qualitatively) as an application of Bayes' Theorem?

Reading Guidance

- **Primary:** Jaynes, **Chapter 2: The Quantitative Rules**. Review the statements of the Product Rule (often called Rule 1) and Sum Rule (often called Rule 2). Follow the derivation of Bayes' Theorem from the Product Rule (Section 2.3 in older drafts, search for "Bayes' Theorem").
- **Primary:** Jaynes, **Chapter 4: Elementary Hypothesis Testing**. Read the introductory sections (e.g., 4.1) where he sets up Bayes' Theorem as the tool for comparing hypotheses based on evidence.
- **Optional Supplementary:** Look for online tutorials or articles explaining Bayes' Theorem with simple examples (like the medical diagnosis or coin flip examples). Focus on understanding the conceptual role of each term.

Looking Ahead to Session 3

We have seen how Bayes' Theorem provides the mechanism for updating beliefs. However, this process requires a starting point: the Prior Probability $P(H|I)$. Session 3 will tackle the challenging but crucial question of how to assign these prior probabilities, addressing the philosophical debates around subjectivity vs. objectivity and introducing principles like Maximum Entropy.

Philosophy of Bayesian Reasoning: Study Notes
Session 3: The Elephant in the Room: Prior Probabilities

Based on Jaynes (2003)

Instructor: Sudeep Regmi

Session Goal

To confront the most debated aspect of Bayesian inference: the selection of prior probabilities ($P(H|I)$). We will explore the necessity of priors, the philosophical debates surrounding subjectivity versus objectivity, and examine various principles proposed for assigning priors, with a focus on Jaynes's preferred method, the Principle of Maximum Entropy.

Key Questions for this Session

- Why is a prior probability essential for applying Bayes' Theorem?
- What makes the choice of prior probabilities so controversial? Isn't science supposed to be objective?
- How did Jaynes view the role and nature of prior probabilities?
- What principles or methods can be used to assign priors when prior knowledge is vague or minimal?
- How can we represent existing knowledge within a prior distribution?
- What is the Principle of Maximum Entropy and why did Jaynes champion it?

1 Recap: The Role of the Prior in Bayes' Theorem

From Session 2, Bayes' Theorem is the engine for updating beliefs:

$$\underbrace{P(H|E,I)}_{\text{Posterior}} \propto \underbrace{P(H|I)}_{\text{Prior}} \times \underbrace{P(E|H,I)}_{\text{Likelihood}}$$

The theorem shows that our updated belief (Posterior) depends mathematically on both the evidence (E, via the Likelihood) and our initial state of belief *before* considering E (the Prior, $P(H|I)$).

Key Idea 1.1 (Necessity of Priors)**.** Within the Bayesian framework of probability as extended logic representing states of knowledge, a prior probability distribution is not optional. It represents the **initial state of information (I)** regarding the hypothesis (H) before incorporating the specific data (E) under analysis. Without a prior, the updating process described by Bayes' Theorem cannot begin.

2 The Controversy: Subjectivity vs. Objectivity

The need for a prior distribution is the primary source of criticism directed at Bayesian methods, particularly from the frequentist perspective which strives for purely "objective" procedures based only on data frequencies.

2.1 The Critique

- If the prior $P(H|I)$ reflects a personal degree of belief, then won't different researchers starting with different priors reach different conclusions (different posteriors $P(H|E, I)$) even when analyzing the *same data* E?
- Doesn't this inject undesirable subjectivity into the scientific process, which aims for objective knowledge?
- How can we justify a particular choice of prior, especially when genuine prior information is scarce?

2.2 The Bayesian Response (esp. Jaynes)

- **Probability is Epistemic:** Jaynes argues that since probability fundamentally represents a state of knowledge, it is inherently tied to the information (I) available. In this sense, it's "subjective" (pertaining to the subject's knowledge state) but should not be arbitrary.
- **Striving for Objectivity:** The goal of "Objective Bayesianism" (as advocated by Jaynes) is not to deny the role of the information state I, but to develop methods for translating a given state of information I into a prior probability distribution $P(H|I)$ in a *consistent and unbiased* way. The rules should be objective, even if the information state I might differ between individuals.
- **Honesty and Transparency:** Bayesian analysis forces prior assumptions (I translated into $P(H|I)$) to be made explicit. This is arguably more scientifically honest than procedures where assumptions might be hidden within the choice of model or method.
- **Ignoring Information is Illogical:** Refusing to specify a prior amounts to refusing to state the initial information I. Jaynes argues that consciously incorporating prior information (or acknowledging its absence in a principled way) is more logical than pretending it doesn't exist or influence reasoning.
- **Convergence:** In many (but not all) situations, as more and more data (E) accumulate, the likelihood function $P(E|H, I)$ becomes sharply peaked and eventually dominates the prior $P(H|I)$. Different reasonable priors will then lead to very similar posterior conclusions. The prior matters most when data is weak or scarce.

3 Principles for Assigning Prior Probabilities

How can we translate a state of information I into a prior $P(H|I)$ in a non-arbitrary way? Several principles have been proposed:

3.1 Principle of Indifference (Laplace)

Principle 3.1 (Indifference)**.** If there are N mutually exclusive and exhaustive possibilities, say $(H_1, ..., H_N)$ and the available information I does not provide any reason to prefer one over another, assign equal probabilities to each:

$$P(H_i|I) = \frac{1}{N} \quad \text{for all } i = 1, ..., N$$

- **Example:** Assigning probability 1/6 to each face of a die presumed fair, based only on the information that it has 6 faces.
- **Strengths:** Simple, intuitive for discrete, symmetric cases.
- **Weaknesses/Paradoxes:** * Ambiguity in defining "possibilities." (Bertrand's paradox: asking for the probability of a random chord in a circle being longer than the side of an inscribed equilateral triangle gives different answers depending on how "random chord" is defined). * Parameterization dependence: If we apply indifference to a parameter θ, we generally don't get indifference for a transformed parameter $\phi = f(\theta)$ (e.g., indifference on side length vs. indifference on area). This suggests the choice of variable matters, violating a sense of consistency.

3.2 Principle of Maximum Entropy (MaxEnt) (Jaynes)

Jaynes proposed this as the fundamental principle for translating constraints specified by information I into a probability distribution $p(x) = P(X = x|I)$.

Principle 3.2 (Maximum Entropy)**.** Given certain constraints on a probability distribution (e.g., known mean value, known range), choose the distribution $p(x)$ that maximizes the Shannon Entropy $H(p) = -\sum_x p(x) \log p(x)$ (or integral form for continuous variables), subject to those constraints.

- **Rationale (Information Theory):** Shannon Entropy measures the amount of uncertainty represented by a probability distribution. Maximizing entropy means maximizing uncertainty, subject to the known constraints.
- **Jaynes's Argument:** The MaxEnt distribution is maximally non-committal with respect to information not explicitly included in the constraints. It uses the information we *do* have (the constraints) but avoids assuming anything extra. It is the "most honest" representation of the state of knowledge defined by I.
- **Connection to Indifference:** If the only constraint is that probabilities sum to 1 over N possibilities, MaxEnt yields the uniform distribution $p_i = 1/N$, recovering the Principle of Indifference as a special case.
- **Other Examples (Conceptual):** * If we only know the range $[a, b]$ of a continuous variable, MaxEnt gives the uniform distribution $U(a, b)$. * If we know the mean μ and variance σ^2 of a continuous variable on $(-\infty, \infty)$, MaxEnt gives the Normal (Gaussian) distribution $N(\mu, \sigma^2)$. * If we know the mean λ of counts on $[0, \infty)$, MaxEnt gives the Exponential distribution (or Poisson for discrete counts).
- **Strengths:** Provides a unique, consistent method for many situations. Yields many standard distributions used in physics (e.g., Boltzmann distribution in statistical mechanics) and engineering. Avoids parameterization paradoxes of indifference when formulated correctly.

- **Weaknesses:** Can be mathematically complex to implement. Requires clearly defined constraints from I. Requires specifying the underlying measure or sample space correctly.

3.3 Transformation Groups / Invariance Principles (Jeffreys)

Principle 3.3 (Invariance)**.** Choose a prior distribution that is invariant (in some sense) under transformations that leave the structure of the problem unchanged (e.g., changes in units of measurement).

- **Example: Jeffreys Prior:** For a parameter θ, the Jeffreys prior is proportional to $\sqrt{\det I(\theta)}$, where $I(\theta)$ is the Fisher Information matrix. This prior has the property of being invariant under reparameterization. * For a location parameter μ $(-\infty, \infty)$, Jeffreys prior is uniform $(\propto 1)$. * For a scale parameter σ $(0, \infty)$, Jeffreys prior is $\propto 1/\sigma$.
- **Strengths:** Offers a form of "objectivity" through invariance. Provides specific recommendations for common parameter types.
- **Weaknesses:** Can lead to "improper" priors (that don't integrate to 1, e.g., uniform on $(-\infty, \infty)$), requiring careful handling. Sometimes complex to calculate. Justification is more mathematical than directly epistemic like MaxEnt.

3.4 Conjugate Priors

Definition 3.1 (Conjugate Prior)**.** A prior distribution is called conjugate to a likelihood function if the resulting posterior distribution belongs to the same family of distributions as the prior.

- **Example:** For a Binomial likelihood (modelling successes in n trials), the Beta distribution is a conjugate prior. If the prior is $\text{Beta}(\alpha, \beta)$ and we observe k successes in n trials, the posterior is also a Beta distribution: $\text{Beta}(\alpha + k, \beta + n - k)$.
- **Strengths:** Mathematical convenience – simplifies calculations significantly, posterior is easy to compute and interpret.
- **Weaknesses:** May not accurately reflect actual prior knowledge. Choice is driven by convenience, not necessarily fundamental principles of representing information state I. Using a conjugate prior is an explicit modelling choice that should be justified.

3.5 Informative Priors

- **Idea:** When genuine prior information exists (from previous studies, physical laws, expert opinion), it *should* be incorporated into the prior $P(H|I)$.
- **Example:** Estimating the effect of a new drug. Previous studies on similar drugs might provide information about the likely range or average effect size, which can be encoded in the prior (e.g., a Normal distribution centered near zero with a certain plausible standard deviation).
- **Challenge:** Eliciting and formalizing expert knowledge or past results into a precise probability distribution can be difficult. Requires careful consideration and sensitivity analysis (checking how results change with different plausible priors).

3.6 "Uninformative" or Reference Priors

- **Goal:** To choose a prior that lets the data "speak for itself" as much as possible, intended for situations where prior knowledge is genuinely minimal or when seeking a "default" or "objective" analysis.
- **Methods:** Often uses principles like indifference (flat priors), invariance (Jeffreys prior), or limiting cases of conjugate priors (e.g., Beta(1,1) which is uniform). MaxEnt with minimal constraints also fits here.
- **Challenges:** True uninformativeness is elusive. A prior "uninformative" in one parameterization might be informative in another. Improper priors often arise. Jaynes was skeptical, arguing MaxEnt was the proper way to represent lack of information beyond specified constraints.

4 Discussion Points

- Consider a problem in your field. What prior information (I) exists before collecting new data (E)? How might you try to represent this information (even qualitatively) as a prior $P(H|I)$?
- Does the idea of using Maximum Entropy to set priors seem compelling? What are its potential benefits and difficulties?
- How comfortable are you with the "subjectivity" inherent in choosing a prior? Is transparency about the prior sufficient justification?
- When might using a conjugate prior be justified for convenience, and when might it be misleading?
- How much should a prior influence the final conclusion? When should we be most concerned about the choice of prior?

Reading Guidance

- **Primary:** Jaynes, **Chapter 11: The Principle of Maximum Entropy**. Focus on the motivation and the examples showing how it leads to common distributions based on different constraints (esp. Sections 11.1-11.4). Don't get bogged down in complex derivations unless interested; grasp the concept.
- **Primary:** Jaynes, **Chapter 12: Ignorance Priors and Transformation Groups**. Discusses issues with indifference and introduces invariance arguments (related to Jeffreys priors). Highlights the challenges of representing "ignorance."
- **Skim:** Jaynes, other chapters where priors are discussed in examples (e.g., Chapter 6 introduction). Note how he approaches prior specification in practice.
- **Optional Supplementary:** Search for articles/posts discussing "objective Bayesian priors," "Jeffreys prior," or "conjugate priors" for broader context. Look for discussions of Bertrand's paradox.

Looking Ahead to Session 4

Having established the core logic (Session 1), the rules and Bayes' Theorem (Session 2), and tackled the challenge of priors (Session 3), Session 4 will focus on how Bayesian inference works in practice. We'll contrast its interpretation of results (like parameter estimates and hypothesis tests) with classical frequentist approaches, drawing on Jaynes's critiques and synthesis of probability as the unified logic of science.

Philosophy of Bayesian Reasoning: Study Notes
Session 4: Bayesian Inference, Interpretation & The Logic of Science

Based on Jaynes (2003)

Instructor: Sudeep Regmi

Session Goal

To synthesize the concepts from the previous sessions by examining how Bayesian inference operates in practice, focusing on the interpretation of its results (parameter estimates, credible intervals, model comparisons). We will critically contrast these interpretations with those from classical frequentist statistics, drawing on E.T. Jaynes's arguments for probability as the unified and consistent logic of scientific reasoning.

Key Questions for this Session

- How is the Bayesian framework (Logic + Rules + Priors) actually used to draw conclusions from data?
- How do we interpret the outputs of a Bayesian analysis, such as posterior distributions or credible intervals?
- What are the fundamental differences between Bayesian and Frequentist interpretations of statistical results (e.g., confidence vs. credible intervals, p-values vs. posterior probabilities)?
- What were Jaynes's main criticisms of frequentist methods based on logical consistency?
- Does the Bayesian approach offer a more coherent framework for scientific reasoning, as Jaynes claimed?

1 Recap: The Bayesian Framework

We've built the Bayesian approach from the ground up:

- **Session 1:** Probability as extended logic, quantifying plausibility based on information (I). $P(A|I)$.
- **Session 2:** Derived the rules (Sum, Product) from desiderata, leading to Bayes' Theorem for belief updating: $P(H|E,I) \propto P(H|I)P(E|H,I)$.
- **Session 3:** Addressed the assignment of prior probabilities $P(H|I)$, representing the initial state of knowledge, using principles like MaxEnt.

Now, we use this framework to learn from evidence E.

2 The Bayesian Inference Workflow (Conceptual)

The core process of Bayesian inference involves these steps:

1. **Model Building:**
   - Define the hypothesis space or parameter space (H or θ).
   - Specify the **Prior Distribution** $P(H|I)$ or $P(\theta|I)$: Quantifies initial beliefs about the hypotheses or parameters before seeing the data E.
   - Specify the **Likelihood Function** $P(E|H, I)$ or $P(E|\theta, I)$: Describes the probability of observing the data E for each possible hypothesis or parameter value. This links the unknown parameters/hypotheses to the observable data.
2. **Computation / Applying Bayes' Theorem:**
   - Combine the prior and likelihood using Bayes' Theorem to calculate the **Posterior Distribution**:

$$P(H|E, I) = \frac{P(H|I)P(E|H, I)}{P(E|I)} \quad \text{or} \quad P(\theta|E, I) = \frac{P(\theta|I)P(E|\theta, I)}{P(E|I)}$$

     where $P(E|I) = \sum_H P(H|I)P(E|H, I)$ or $\int P(\theta|I)P(E|\theta, I)d\theta$.
   - This step often involves complex calculations, frequently requiring computational methods like Markov Chain Monte Carlo (MCMC) in modern practice (though the conceptual framework predates these methods).
3. **Interpretation and Inference:**
   - Analyze the **Posterior Distribution** $P(H|E, I)$ or $P(\theta|E, I)$. This distribution encapsulates *all* information about the hypotheses/parameters after considering both the prior knowledge I and the data E.
   - Summarize the posterior: calculate point estimates (mean, median, mode), credible intervals, probabilities of specific hypotheses, etc.
   - Perform model comparison or hypothesis testing if needed (see below).

3 Interpreting Bayesian Results

The key strength of Bayesian inference lies in the directness of its interpretations:

3.1 Posterior Distributions

- The posterior $P(\theta|E, I)$ is the complete result for parameter estimation. It's a probability distribution directly representing our state of belief about the possible values of the parameter θ *after* seeing the data E and incorporating prior information I.
- We can visualize it (plot the distribution) or summarize it numerically.

3.2 Credible Intervals (or Credible Regions)

Definition 3.1 (Credible Interval)**.** A 95% credible interval for a parameter θ is an interval $[a, b]$ such that the posterior probability of θ lying within that interval is 0.95:

$$P(a \leq \theta \leq b|E, I) = \int_a^b P(\theta|E, I)d\theta = 0.95$$

- **Interpretation:** "Given the data and model, there is a 95% probability that the true value of the parameter θ lies within the interval $[a, b]$."
- This is a direct statement about the parameter's location, reflecting our updated state of belief. This is often what people *intuitively want* interval estimates to mean.

3.3 Hypothesis Testing / Model Comparison

- **Posterior Probabilities:** If we have a discrete set of competing hypotheses $H_1, ..., H_N$, we can calculate the posterior probability $P(H_i|E, I)$ for each. We can directly compare these probabilities: "Given the data, hypothesis H_1 is k times more plausible than hypothesis H_2." $(k = P(H_1|E, I)/P(H_2|E, I))$.
- **Bayes Factors:** The ratio of posterior odds to prior odds is called the Bayes Factor (BF):

$$\underbrace{\frac{P(H_1|E, I)}{P(H_2|E, I)}}_{\text{Posterior Odds}} = \underbrace{\frac{P(H_1|I)}{P(H_2|I)}}_{\text{Prior Odds}} \times \underbrace{\frac{P(E|H_1, I)}{P(E|H_2, I)}}_{\text{Bayes Factor } (BF_{12})}$$

  The Bayes Factor BF_{12} measures the extent to which the data E support H_1 relative to H_2. A $BF_{12} > 1$ means the evidence favors H_1; $BF_{12} < 1$ means it favors H_2. It quantifies the *strength of evidence* provided by the data.

4 Contrasting with Frequentist Interpretations

The philosophical differences between Bayesian and Frequentist approaches lead to stark contrasts in how results are interpreted. Frequentist methods focus on the properties of procedures in hypothetical repetitions, not on belief about the specific case at hand.

Comparison 4.1 (Parameter Estimation: Intervals)**.**

- **Bayesian Credible Interval (95%):** $P(\theta \in [a, b]|E, I) = 0.95$. *Interpretation:* Direct probability statement about the parameter θ based on the observed data E and prior I. The parameter is treated as the random variable (reflecting our uncertainty); the interval $[a, b]$ is fixed once calculated.
- **Frequentist Confidence Interval (95%):** A procedure for generating intervals $[L(E), U(E)]$ such that $P(\theta \in [L(E), U(E)]; \theta) = 0.95$. * *Interpretation:* "If we were to repeat this *entire procedure* (sampling data E and calculating the interval) many times, 95% of the *intervals generated* would contain the true, fixed value of θ." * *Crucial Difference:* This is a statement about the long-run performance of the *procedure*, not about the specific interval $[L(E_{obs}), U(E_{obs})]$ calculated from the *observed* data E_{obs}. Once observed, a frequentist interval either contains θ or it doesn't (probability is 0 or 1). It does NOT mean there is a 95% probability that θ is in the specific interval calculated. The interval is treated as random (dependent on the random sample E); the parameter θ is fixed but unknown.

Comparison 4.2 (Hypothesis Testing).

- **Bayesian Approach (Posterior Odds / Bayes Factor):** * Compares the plausibility of two (or more) hypotheses H_1, H_2 directly, given the data: $P(H_1|E, I)$ vs $P(H_2|E, I)$. * Answers: "How much more plausible is H_1 than H_2, given the evidence?" * Requires specifying priors and likelihoods for *both* hypotheses being compared.

- **Frequentist Approach (p-value):**
  * Calculates $p = P(\text{Data as extreme or more extreme than observed}|H_0 \text{ is true})$. * Focuses only on the null hypothesis (H_0). Does not typically require explicit specification of an alternative hypothesis H_A (though the choice of test statistic often implicitly relates to potential alternatives). * Answers: "How surprising is the observed data (or more extreme data), assuming the null hypothesis is true?"

- **Common Misinterpretations of p-value (Jaynes's Critiques):** * Fallacy: Interpreting p as $P(H_0|E)$ (probability the null is true given the data). This is incorrect – it's $P(\text{Data}|H_0)$. Bayes' Theorem is needed to get $P(H_0|E)$. * Fallacy: Thinking $p = 0.05$ means there's only a 5* Fallacy: Thinking a non-significant result ($p > 0.05$) provides evidence *for* H_0. Absence of evidence against H_0 is not evidence for H_0. Bayesian methods can quantify evidence for H_0 (relative to H_A).

5 Jaynes's Critiques and Perspective

Jaynes was a sharp critic of frequentist methods, arguing they often violated basic principles of logical consistency. His critiques weren't just about interpretation, but about the procedures themselves.

- **Violation of the Likelihood Principle:** Frequentist results (like p-values, confidence intervals) often depend on aspects of the experimental design or sampling rule (e.g., stopping intentions) that didn't affect the observed data E. Bayesian inference, operating via $P(\theta|E, I) \propto P(\theta|I)P(E|\theta, I)$, depends only on the likelihood function $P(E|\theta, I)$ for the *actually observed data* E (and the prior). Jaynes argued this makes Bayesian inference more rational – conclusions should depend on what happened, not what might have happened. (See discussion of optional stopping).

- **Handling of Nuisance Parameters:** Frequentist methods often struggle to eliminate nuisance parameters (parameters needed for the model but not of direct interest) in a universally agreed-upon way. Bayesian inference handles them naturally by integrating them out of the posterior distribution using the rules of probability:
  $P(\theta_{interest}|E, I) = \int P(\theta_{interest}, \theta_{nuisance}|E, I)d\theta_{nuisance}$.

- **Ad Hoc Nature:** Jaynes viewed many frequentist procedures as a collection of "ad hoc" methods developed for specific problems, lacking the unified logical foundation provided by probability theory interpreted as extended logic. He saw p-values, confidence intervals, different estimators, etc., as disconnected recipes rather than applications of a single coherent principle.

- **Ockham's Razor Automated:** Jaynes emphasized (e.g., Chapter 20) that Bayesian model comparison automatically incorporates Ockham's Razor (prefer simpler explanations). Complex models with many parameters can fit the data well, but they spread their prior probability thinly. The evidence term $P(E|I) = \int P(\theta|I)P(E|\theta, I)d\theta$ tends to be smaller for overly

complex models unless the extra complexity is truly warranted by the data, thus naturally penalizing them.

6 Synthesis: The Logic of Science?

Jaynes argued passionately that Bayesian probability theory, understood as the logic of plausible reasoning, provides the *single* coherent and consistent framework for all scientific inference.

- **The Journey:** We started with desiderata for rational thought under uncertainty, derived the unique rules of probability (Sum, Product), obtained Bayes' Theorem as the mechanism for updating belief, addressed prior specification through principles like MaxEnt, and arrived at a framework for inference with direct, intuitive interpretations.
- **Strengths of Bayesianism (Jaynesian view):** * Philosophically coherent foundation (extended logic). * Unified framework for diverse problems (estimation, testing, model choice). * Direct, intuitive interpretation of results (probabilities of parameters/hypotheses). * Incorporates prior knowledge explicitly and transparently. * Obeys the likelihood principle; handles nuisance parameters consistently. * Automatically incorporates Ockham's Razor.
- **Challenges/Weaknesses in Practice:** * Prior specification can still be difficult and controversial. * Computational complexity can be high (though modern algorithms help greatly). * Requires careful model checking (as does any statistical approach). * Acceptance and understanding in some fields still lags behind frequentism.

Key Idea 6.1 (Probability as the Logic of Science)**.** The ultimate claim is that the rules derived from the simple desiderata of logical consistency are all that is needed for inductive reasoning in science. Bayesian inference isn't just one way to do statistics; from this perspective, it is the *only* way that is fully consistent with the principles of plausible reasoning.

7 Discussion Points

- Which interpretation of an interval estimate (Confidence vs. Credible) seems more useful or intuitive for scientific reporting? Why?
- Does the frequentist reliance on hypothetical repetitions make sense for unique events or fixed parameters?
- Are Jaynes's critiques of p-values (confusion with $P(H_0|E)$, dependence on hypothetical data) convincing?
- Is the explicit nature of priors in Bayesian analysis a strength (transparency) or a weakness (subjectivity)?
- Does the idea of a single, unified logic of science (Bayesian probability) seem plausible, or are different tools needed for different scientific questions?
- What are the biggest hurdles (philosophical or practical) to adopting a fully Bayesian approach in your field?

Reading Guidance

- **Primary:** Jaynes, **Chapter 4: Elementary Hypothesis Testing** and **Chapter 5: Queer Uses for Probability Theory**. Focus on his framing of hypothesis testing and his critiques of alternatives (esp. p-values).
- **Primary:** Jaynes, **Chapter 17: Paradoxes of Probability Theory**. Discusses issues like optional stopping and contrasts Bayesian/Frequentist handling. Focus on the conceptual arguments regarding consistency.
- **Skim:** Jaynes, **Chapter 20: Model Comparison**. Understand the concept of how Bayesian methods quantify evidence and relate to Ockham's Razor (via the Evidence term $P(E|M)$).
- **Optional Supplementary:** Search for articles/blogs clearly explaining the difference between Confidence and Credible intervals, or common p-value fallacies. Compare explanations of the Likelihood Principle.

Concluding Thought for the Course

This short course aimed to introduce the philosophical foundations of Bayesian reasoning, heavily influenced by E.T. Jaynes. By viewing probability not just as frequency, but as a measure of plausible belief derived from logical desiderata, we arrive at Bayes' Theorem as the consistent way to update knowledge. While practical implementation has challenges (priors, computation), the framework offers a potentially unified and coherent approach to the logic of science.