Intro to Bayesian Epistemology

Stefano Mezzini

2025-03-28

Imagine you have a cat...

Imagine you have a cat... will they break stuff when you leave?

Exploring alternatives with truth tables

- Outline all possible outcomes
- Alternatives are explored with deductive reasoning
- Conclusions are reached using logic
- Uncertainty is either ignored or a complete block
- See William Spaniel's Logic 101 on YouTube

For example: "Broken" always implies "Guilty"

Broken	Guilty	not Broken	not Guilty	Broken or Guilty	Broken and Guilty	...
T	**T**	**F**	**F**	**T**	**T**	**...**
T	**F**	**F**	**T**	**T**	**F**	**...**
F	**T**	**T**	**F**	**T**	**F**	**...**
F	F	T	T	F	F	...

Exploring alternatives with probability tables

- Outline all possible outcomes
- Alternatives are explored with inductive reasoning
- Conclusions are reached using probability
- Uncertainty is acknowledged and leveraged in the framework

For example: "Broken" does not always imply "Guilty", and vice-versa

	Guilty	not Guilty	Total
Broken	0.33	0.01	0.34
not Broken	0.33	0.32	0.66
Total	0.67	0.33	1.00

What is probability?

- A number in the interval $[0, 1]$
- A measure that quantifies the occurrence of random events
- A measure that allows us to quantify the risk of events
- A measure of our confidence in a parameter, hypothesis, or value

	Guilty	not Guilty	Total
Broken	**0.33**	**0.01**	**0.34**
not Broken	**0.33**	**0.32**	**0.66**
Total	**0.67**	**0.33**	**1.00**

Probability fundamentals

Let B and G be two events. Then:

- $0 \leq P(B) \leq 1$
- the probability of "not B" is $P(\neg B) = 1 - P(B)$
- B **or** G: $P(B \vee G) = P(B) + P(G) - P(B \wedge G)$
- B **and** G: $P(B \wedge G) = P(B) \times P(G|B) = P(G) \times P(B|G)$

	Guilty	not Guilty	Total
Broken	0.33	0.01	0.34
not Broken	0.33	0.32	0.66
Total	0.67	0.33	1.00

Frequentist probability: the long-run frequency of events

Frequentist probability: the unobserved long-term frequency of some event (e.g., "how often does it rain, on average?")

- Datasets are random: we cannot get attached to a particular dataset
- "What if we observed a different dataset, instead?"
- "What if this dataset is an outlier dataset?"
- "We cannot know how often the cat will break something"
- "Are these random observations compatible with hypothesis H_0 or H_A?

For example, if $P(B \wedge G) = 1/3$:

- In the long run, the cat will break something 1/3 of the time
- You don't know if the cat will break something next time you leave

Bayesian probability: the relative certainty of events

Bayesian probability: the certainty in the relative occurrence of some event (e.g., "how confident am I that it is raining?")

- Datasets are fixed after observation because they inform our knowledge
- "How do we make inferences from all the data we accumulate?"
- "What if this dataset is an outlier dataset?"
- "How confident can you be that the cat will break something, next time?"
- "What range of values seems credible, given the data?"

For example, if $P(B \wedge G) = 1/3$:

- You are ~33% sure the cat will break something
- Lock the cat in a safe room if its cost is < 1/3 the expected damage


Priors: An opinion before data collection

Priors: We all have opinions, even before data collection

Priors:

- summarize our knowledge before data collection
- are independent of future data
- are distributions, not single values
- communicate our uncertainty before data colletion
- can be used to make predictions before data collection
- can vary between people

Are there priors in Frequentist statistics?

Beyesian statistics (generally) start with analyst-informed priors

Frequentist statistics:

- Hypothesis testing: H_0, H_a, and p-values
- Ignore priors because uninterested in estimating θ
- Interested in likelihood of data given θ, which is $P(\mathrm{data}|\theta) = \mathcal{L}(\theta|\mathrm{data})$
- Not interested in finding the most likely value of θ
- $\mathcal{L}(\theta|\mathrm{data})$ is a function, not a probability distribution

Likelihood: The information in the data

Likelihood: What we learn from data

The likelihood:

- summarizes the information we collected through data, D
- is independent of the prior and the analyst
- gives $P(D|\theta = \theta_i)$ for all possible values of θ_i, even unrealistic ones
- does not make statements on $P(D)$ independently of θ
- is the distribution of $P(D|\theta)$, not $P(\theta|D)$

Posteriors: Our updated beliefs

Posteriors: Our opinion after data collection

Posteriors:

- summarize our total knowledge after data collection: $P(\theta|D)$
- are independent of future data
- are distributions, not single values
- communicate our updated uncertainty after data collection
- can be used to make predictions about θ after collecting new evidence
- depend more on one's priors if the data had a small sample size

Calculating posteriros using Bayes' theorem

For a parameter θ and an observed dataset D, we have:

$$P(\theta|D) = \frac{P(\theta)\ P(D|\theta)}{P(D)}$$

where θ could be $P(Guilty)$ and D could be the number of broken objects.

- $P(\theta|D)$ is the posterior
- $P(\theta)$ is the prior, which summarizes knowledge before data collection
- $P(D|\theta)$ is the likelihood, which summarizes the data you observed
- $P(D)$ is the probability of observing the dataset, independent of θ


Less math, more cats!

Assessing degrees of belief with Bayesian Epistemology

Should you leave your cat unsupervised? If you leave, the cat will most likely do something it shouldn't, since $P(G) = 0.67 > 0.5$

	Guilty	not Guilty	Total
Broken	**0.33**	**0.01**	**0.34**
not Broken	**0.33**	**0.32**	**0.66**
Total	**0.67**	**0.33**	**1.00**

Based on $P(G)$ alone: if the cat never breaks things when supervised, you should not leave the cat unsupervised!

Assessing risks and costs with Bayesian Epistemology

- Include the cost of mischief by multiplying it by $P(G)$:

	Guilty	not Guilty	Total
Broken	**CAD 10 x 0.33 = CAD 3.33**	**CAD 1 x 0.01 = CAD 0.01**	**CAD 3.34**
not Broken	**CAD 0 x 0.33 = CAD 0**	**CAD 0 x 0.32 = CAD 0**	**CAD 0**
Total	**CAD 3.33**	**CAD 0.01**	**CAD 3.34**

- Do not leave the cat alone if cost of leaving > supervising it
- How do we convert the cat's loneliness to a dollar amount?

Collecting evidence and updating beliefs

- You decide to leave the cat unsupervised
- You come home to a broken vase

	Guilty	not Guilty	Total
Broken	**CAD 10 x 0.33 = CAD 3.33**	**CAD 1 x 0.01 = CAD 0.01**	**CAD 3.34**
not Broken	CAD 200 x 0.33 = CAD 66.6	CAD 0 x 0.32 = CAD 0	CAD 66.67
Total	CAD 70.00	CAD 0.01	CAD 70.01

- What is the probabilty that the cat is guilty, given the broken vase?

Should you scold the cat?

Conditionalize on something being broken:

	Guilty	not Guilty	Total
Broken	**0.33 / 0.34 = 0.97**	**0.01 / 0.34 = 0.03**	**(0.33 + 0.01) / 0.34 = 1**

- Cat is most likely guilty (0.97 > 0.5)
- You can be quite sure the cat broke the vase

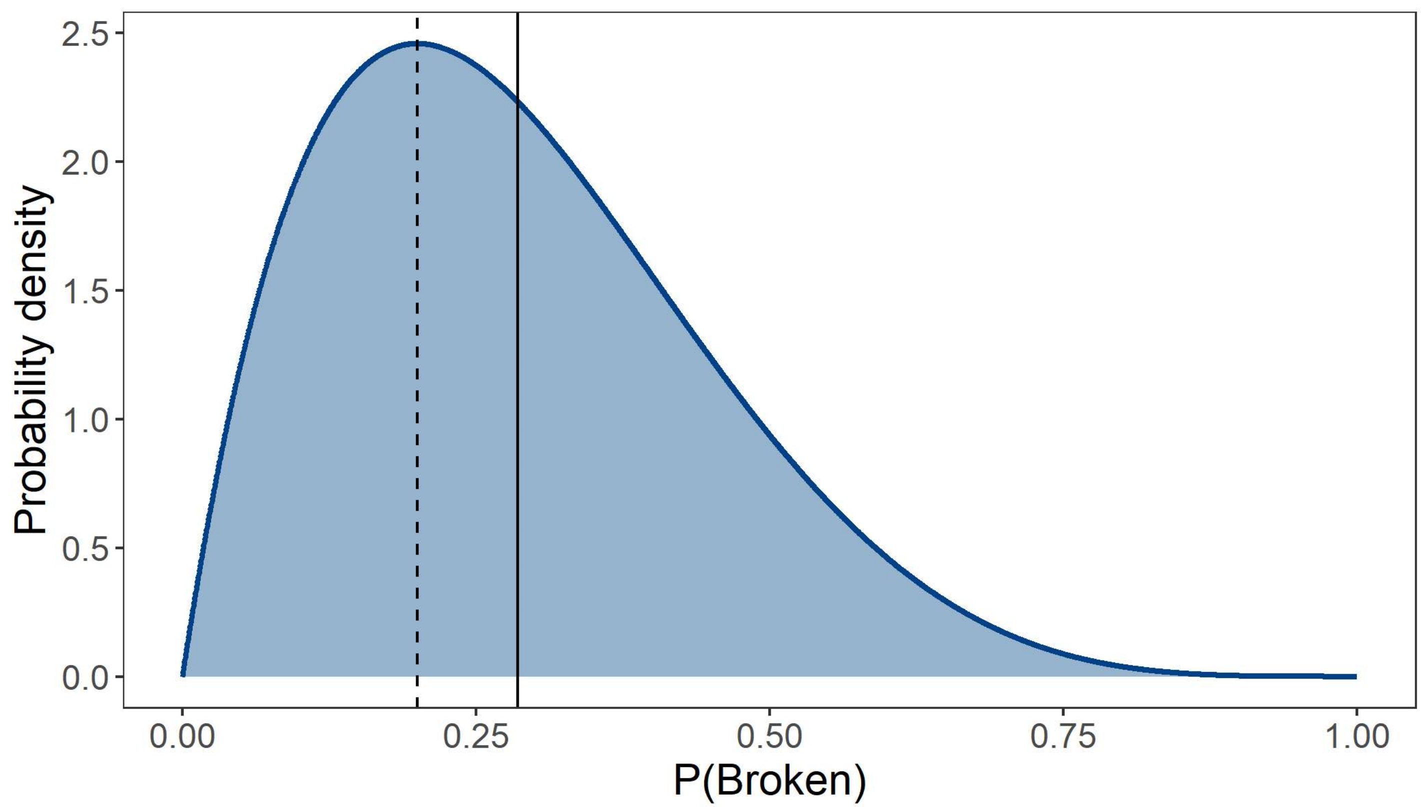
Should you change your future behavior?

- You have gained new evidence
- Update your belief based on new evidence
- Update your *behavior* based on your new beliefs:
  - remove breakable objects
  - lock the cat in a safe room when you leave

A practical example: updating your beliefs day by day

Your belief prior to data collection

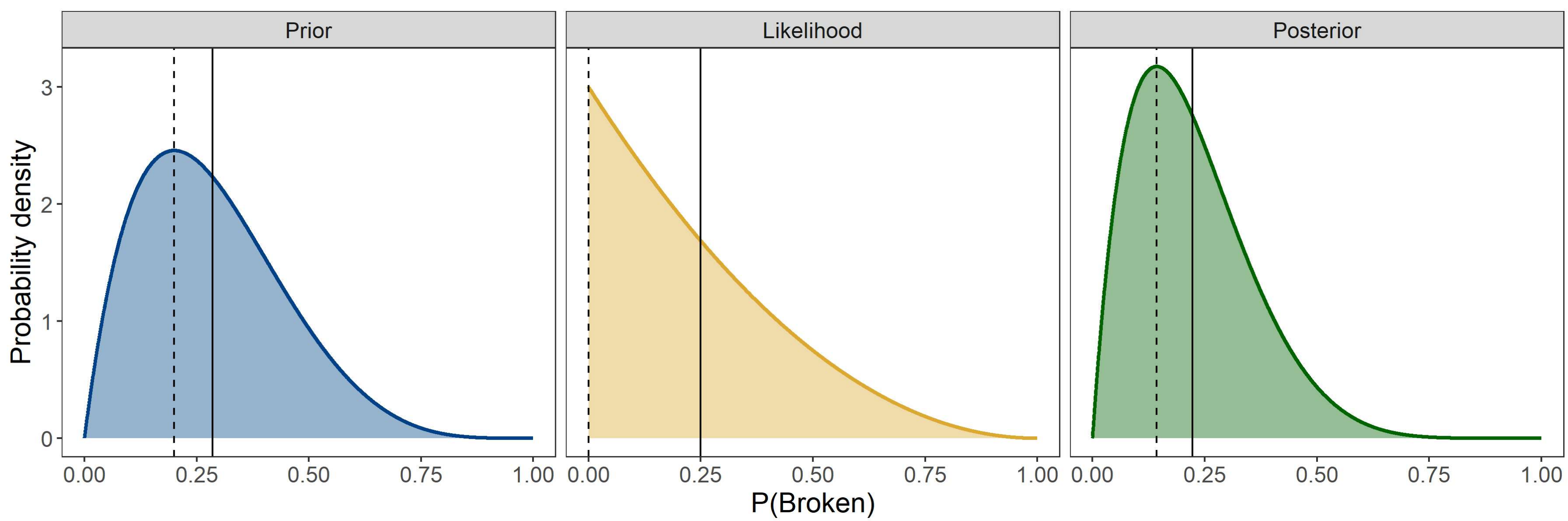
- You get a new cat
- You don't expect your cat to break things, but it could still be possible
- Prior is $\theta \sim B(2, 5)$:
  - Mean of θ is $\mathbb{E}(\theta) = 0.286$
  - Mode of θ is $M(\theta) = 0.2$



Running a test

- You leave the cat free to roam before going to campus for the day
- You come back to find nothing broken, but maybe you missed something?
- Update your prior knowledge by including the new data:

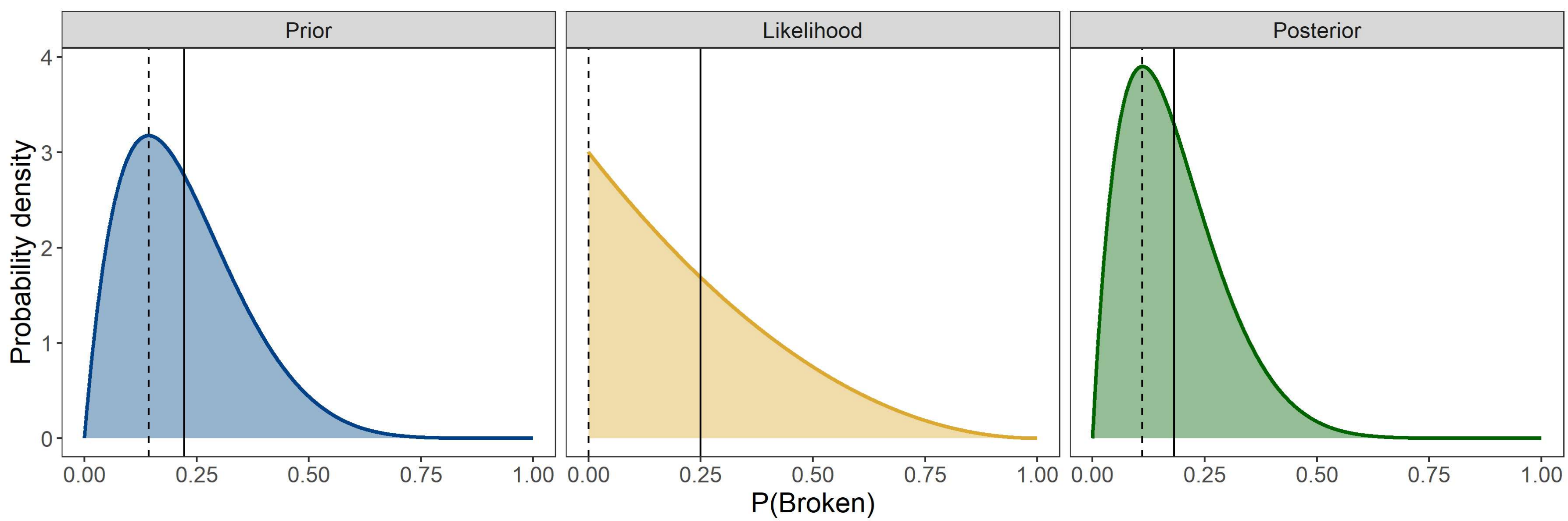
$$P(\theta|D) = \frac{P(\theta)\ P(D|\theta)}{P(D)}$$



A second experiment: Bayesian updating

- You leave the cat free to roam again and find nothing broken
- **Note**: your likelihood is independent of your prior!
- Your old posterior becomes your new prior
- You can update your posterior again:

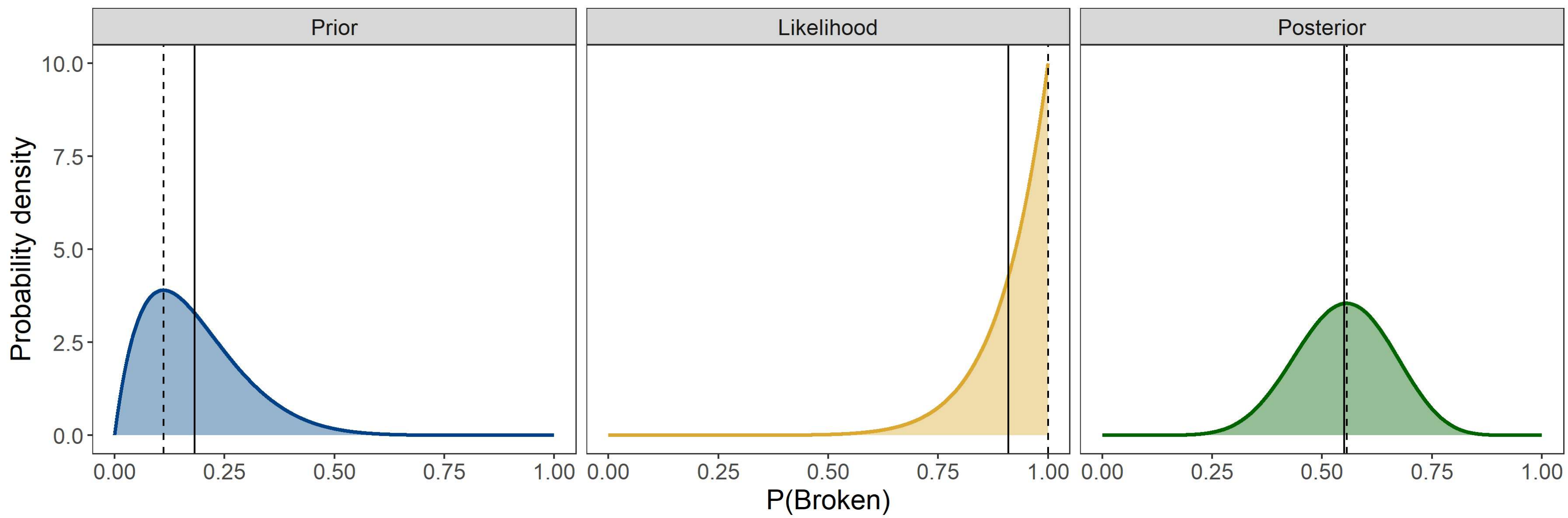
$$P(\theta|D) = \frac{P(\theta)\ P(D|\theta)}{P(D)}$$



An unfortunate third experiment

- You leave the cat free to roam again, but this time it breaks something
- You can update your posterior again:

$$P(\theta|D) = \frac{P(\theta)\ P(D|\theta)}{P(D)}$$



Bayesian updating: updating all at once is like updating each time

$$P(\theta|D_1, D_2, D_3) = \frac{P(\theta|D_1,\ D_2)\ P(D_3|\theta)}{P(D_3)}$$

$$P(\theta|D_1, D_2, D_3) = \frac{P(\theta|D_1)\ P(D_2|\theta)\ P(D_3|\theta)}{P(D_2, D_3)}$$

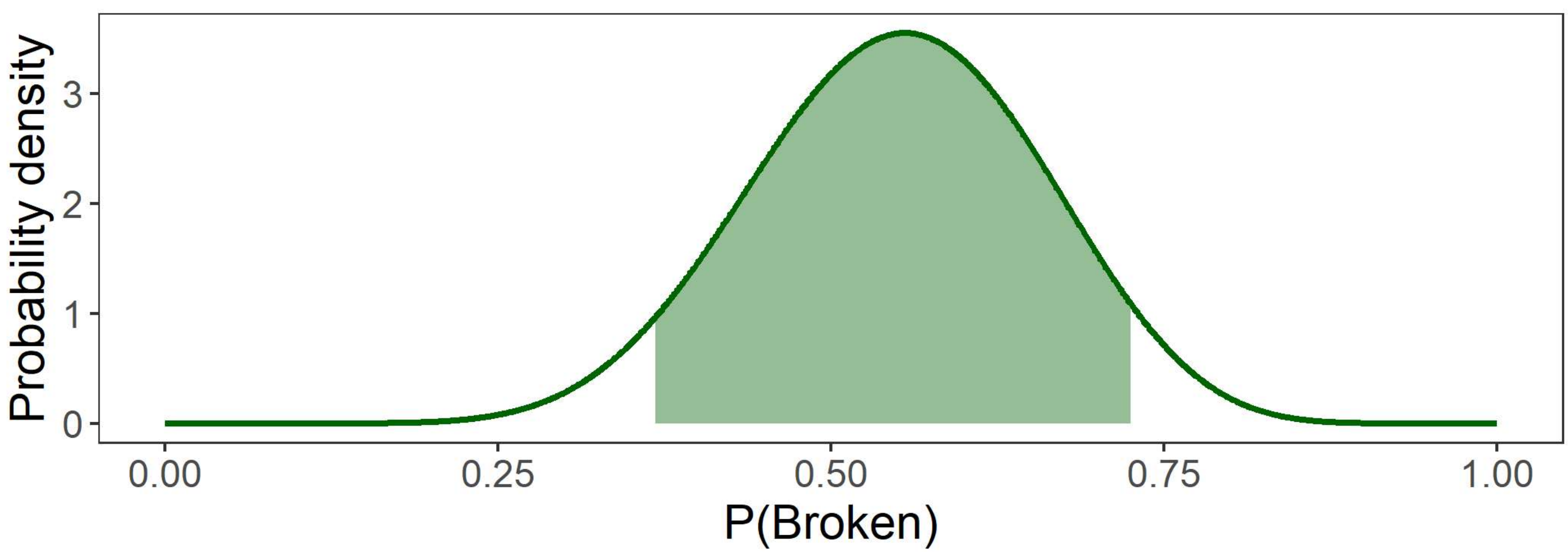
$$P(\theta|D_1, D_2, D_3) = \frac{P(\theta)\ P(D_1|\theta)\ P(D_2|\theta)\ P(D_3|\theta)}{P(D_1, D_2, D_3)}$$

$$P(\theta|D_1, D_2, D_3) = \frac{P(\theta)\ \prod_i P(D_i|\theta)}{\prod_i P(D_i)}$$

Summarizing the posterior: Credible intervals

Credible intervals:

- Are a range of believable (i.e., credible) values at some certainty level
- Account for the system you obtained data from
- Do not risk being irrational, as long as your prior was appropriate
- Depend strongly on the prior if the sample size of the data was small:
  - Useful if we want reasonable predictions despite scarce data
  - Problematic if we do not have good priors



Confidence Intervals vs Credible Intervals

Clarifying Confidence Intervals

- Confidence is in the data under H_0, not in θ
- Give good coverage across all values of θ, even impossible ones
- Are not a measure of our confidence in a hypothesis
- May be best renamed to "Compatability Intervals"
- Are highly susceptible to odd datasets
- Work well as $n \to \infty$

Contrasting Credible Intervals

- Confidence is in the parameters and hypotheses of interest, given the data
- Give good coverage across reasonable values of θ, not unlikely ones
- Can give poor coverage if data contrasts prior beliefs, especially low n
- Are a measure of our confidence in a hypothesis
- Are robust to to odd datasets
- Work well with small n
- Are suitable for search theory
- Also see this Stack Exchange post on Confidence vs Credible Intervals


Summary

Bayesian statistics mirror how many of us learn

- Prior: $P(\theta)$, belief before data collection
- Likelihood: $P(D|\theta)$, probability of observing the data, given θ
- Posterior: $P(\theta|D)$, belief after data collection
- Posterior allows us to make inferences on the credible value of θ
- Bayesian inference guarantees realistic (but possibly subjective) results
- Bayesian inference constructs a coherent, probabilistic epistemology