A math-light introduction to Bayesian statistics

Stefano Mezzini

2025-02-25

Contents

1 Priors: an opinion before data collection **1**

2 Likelihood: the information in the data **4**

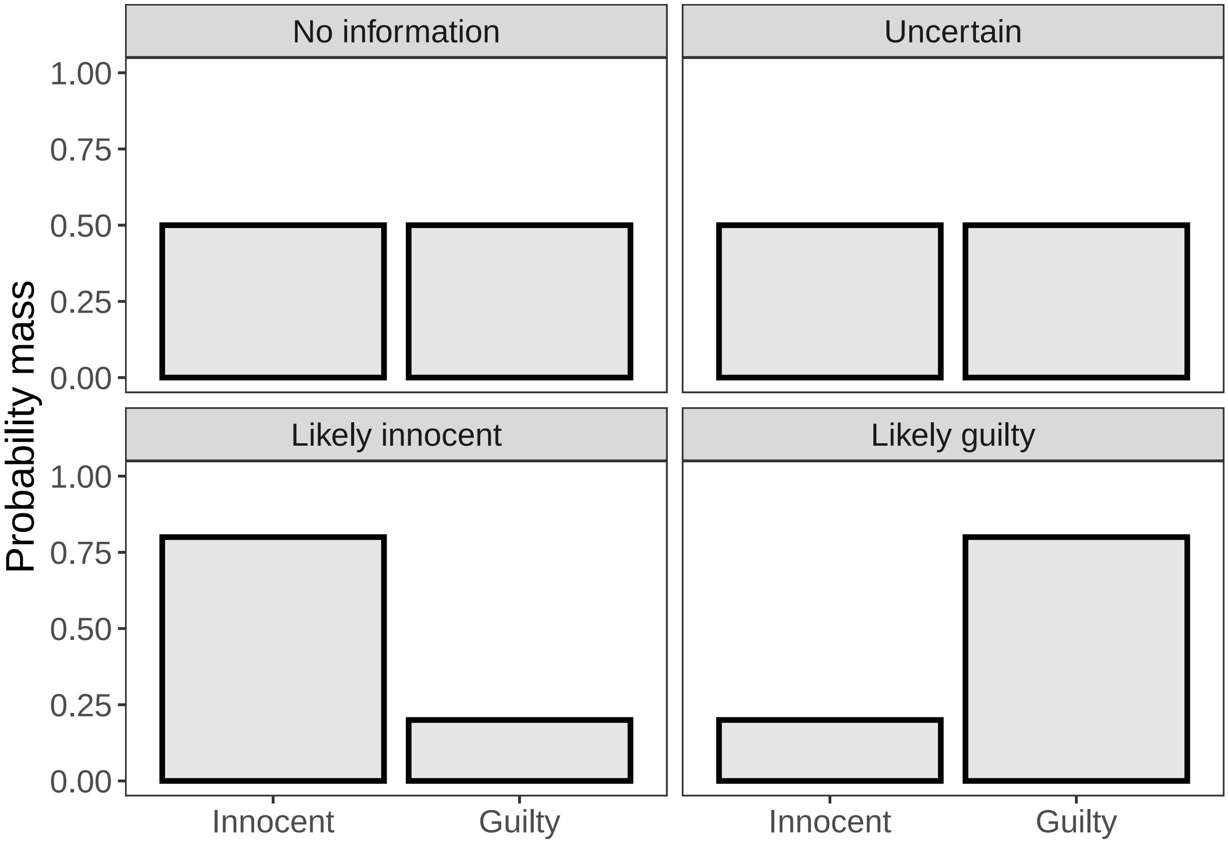
3 Posterior: your updated belief **4**

3.1 Bayesian updating: when posteriors become the new priors 7

4 Expressing uncertainty **7**

1 Priors: an opinion before data collection

Imagine you have a cat (or a dog, if that helps – just read all mentions of "cat" as "dog"). One day you come home to find a framed picture on the floor; the frame broken and the glass cracked. Did your cat break it? Before you start collecting any information, you may already have an opinion about whether or not it did. If you have no information at all, you may say the chances of it being guilty are the same as it being innocent. You may also say the same if you have some information but are highly uncertain. If you believe your cat is likely innocent, you would place a larger probability on it being innocent, and you would place less if you believe it to be guilty.



However, how do you distinguish between having no information at all about the cat and having some small amount of information? And how do you express your uncertainty in your initial guess? We can do this by expressing our prior belief as a distribution rather than a single value. You could say that this distribution is your entire belief state (the most likely value and the uncertainty around it) prior to any data collection, which we will call your **prior** for short. If use θ to indicate the **probability of your cat breaking the frame** (note: not just whether it broke it or not), P(guilty), your prior may look like one of the distributions below.

Now you can see how the “no information” prior is different from the “uncertain” prior. The “no information prior” says that all values of θ are equally as likely (including always being guilty, $\theta = 1$, and always being innocent, $\theta = 0$), while the “uncertain” prior recognizes the uncertainty by setting the most likely value to $\theta = 0.5$ while stating that it is impossible for the cat to *always* be guilty or innocent – maybe you hung the picture poorly, but sometimes cats *do* break things. Unlike Beyesian statistics, the Frequentist approach to statistics (the “traditional” hypothesis testing approach with H_0, H_a, and p-values) always starts with no prior information, i.e. the “flat prior”.

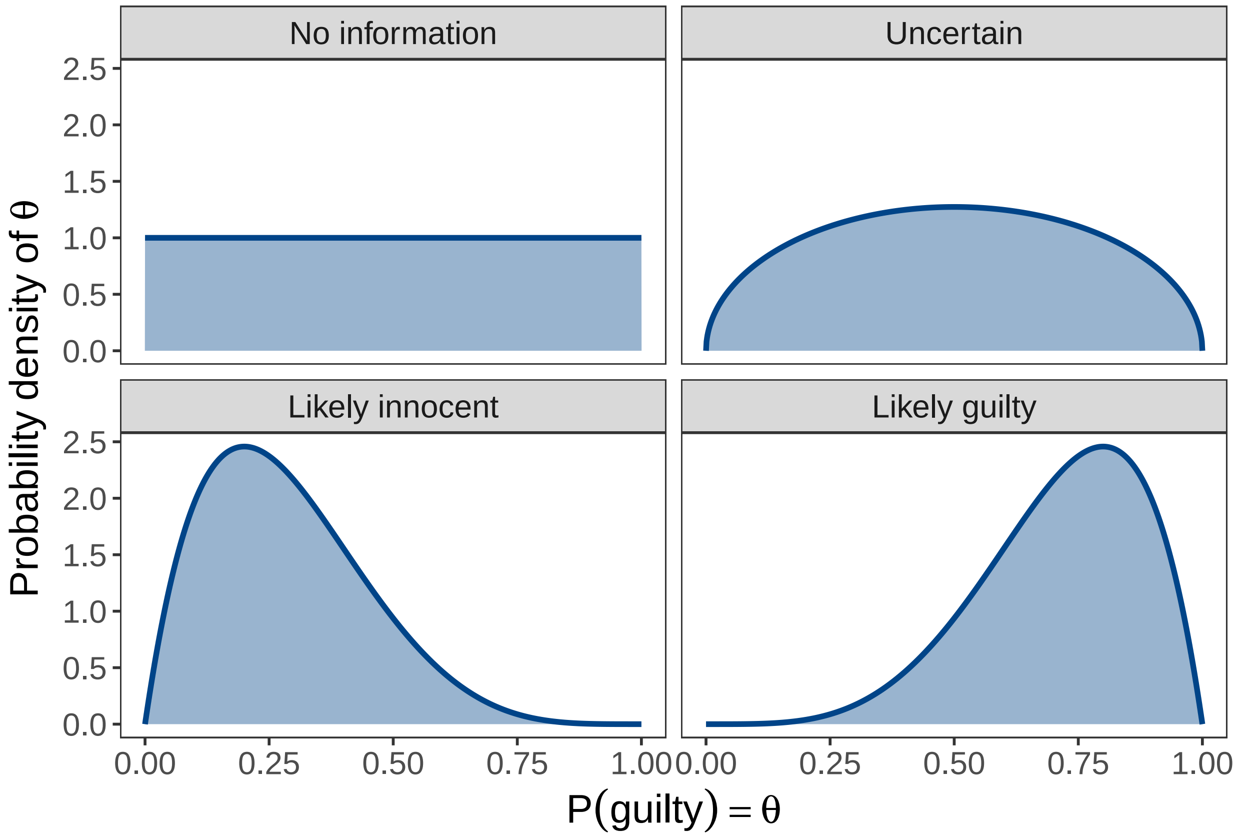


Figure 1: Examples of priors for the probability of your cat having broken the frame, for different belief states before collecting any data.

2 Likelihood: the information in the data

Once you have defined your prior, it is time to **collect some data** so you can update your belief based on the evidence. To do this, you may inspect the frame and the area around it. Is anything else out of order? Are other objects knocked down? Could your cat have reached the frame? The answer to each of these questions can be summarized into a new distribution: the **likelihood**. The likelihood contains all the information you gathered from inspecting the scene (or running an experiment). Formally, it is the probability of obtaining the information (or data, D) you obtained, for some value of θ, and we can write it as $P(D|\theta)$. This is what the Frequentist approach to statistics is based on: p-values are the probability of observing the dataset you observed or a more unlikely one, if θ has the value specified by the null hypothesis, which we can write as $P(D|H_0 : \theta = \theta_0)$. If you calculate $P(D|\theta = \theta_i)$ for all possible values of θ rather than just the value from H_0, you get the probability distribution of D conditional on θ, which is the likelihood $P(D|\theta)$.

When you inspect the objects around the frame, you notice that other things are knocked over, and the nail the frame was hanging from seems to have have given out due to excessive weight. In this case, the likelihood may look like something like this:

It seems reasonable to believe that the cat did indeed cause the frame to fall, since would be fairly unlikely to observe this evidence if the cat is most probably innocent ($P(\theta < 0.5)$. Consequently, most of the likelihood density is for $\theta > 0.5$, with $\theta = 0.5$ being a 50-50 change of the cat being guilty. But could this just be an accident that happened due to chance? Maybe the frame wasn't hung properly, and when it fell it scared the cat, who then knocked the other items down. Fortunately, we can combine the likelihood with our prior to estimate the probability that the cat did indeed knock the frame off the wall.

3 Posterior: your updated belief

What is the probability that your cat knocked the frame off the wall, given your guess before observing the evidence? Since both the likelihood and the prior are probability distributions, we can combine them by taking the product of the two. More specifically, we can apply **Bayes' theorem** to calculate the **posterior**, which is the probability of the cat being guilty, given the evidence (which is different different from the likelihood!). Mathematically, we can write it as

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

where $P(\theta)$ is our prior, $P(D|\theta)$ is our likelihood, and $P(D)$ is the probability of observing the evidence

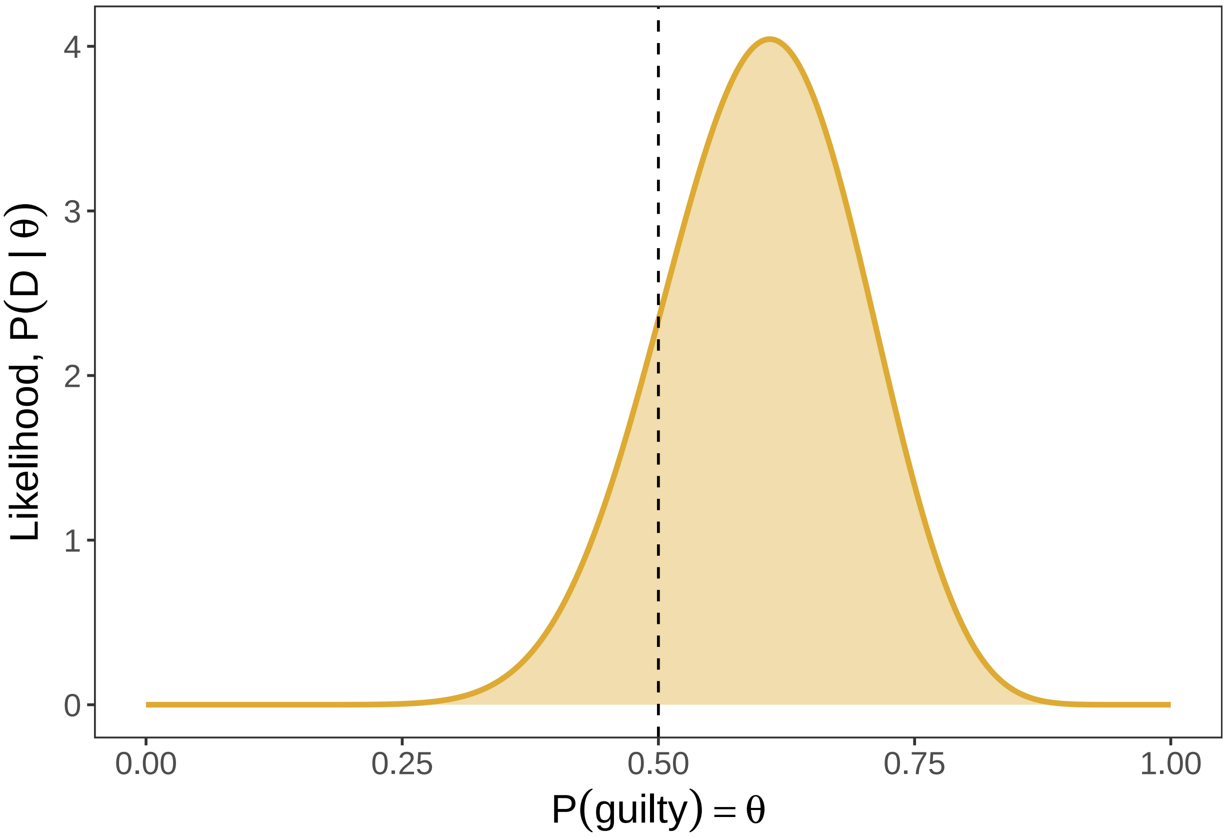


Figure 2: Likelihood of the data (i.e., the evidence observed) for different probabilities of the cat having broken the frame, θ. The scene seems unlikelt to be caused by coincidece, but it is still possible.

we observed, irrespective of θ (i.e., averaged across all possible values of θ). We haven't talked about $P(D)$ yet, but this is not an issue. The probability of observing the scene we observed is a number we can figure out and cancel out by making sure the total area for $P(\theta|D)$ is equal to 1. Mathematically, this means that $\int_D P(\theta|D)\, dD = 1$, but you don't have to worry about doing any calculations. Through some simulations and "mathematical magic" (e.g., the Metropolis–Hastings algorithm), we can approximate $P(\theta|D)$ without complex (and possibly unsolvable) integrals. If we calculate the posterior using each of our priors from before, we get the following figure:

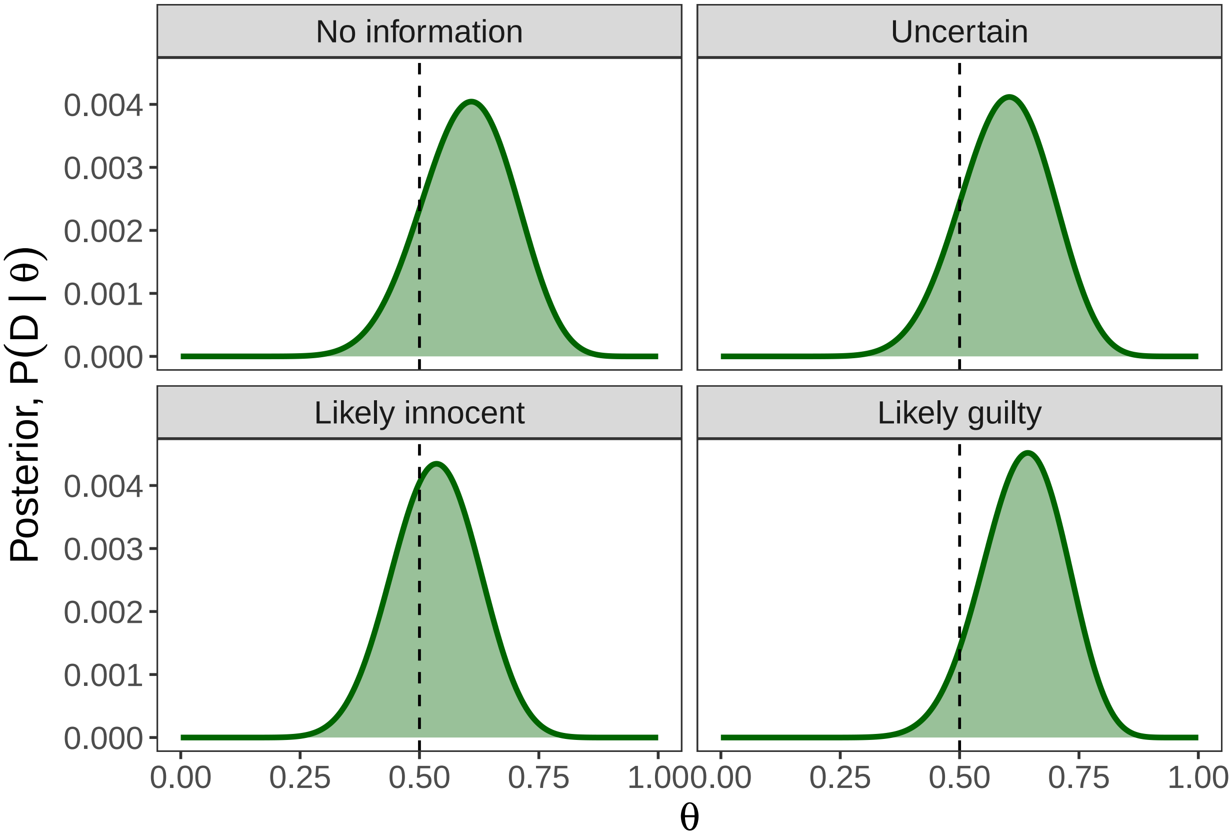


Figure 3: Examples of posteriors for the probability of your cat having broken the frame, for each of the priors shown in Figue 1.

As you can see, your starting belief impacts whether or not you believe your cat was guilty, even after seeing the evidence, but in each case you believe them to be most probably guilty. Table 1 below summarizes how certain you are the cat is guilty, based on your prior. In each case, you are more than 50% sure that the cat did break the frame, but you are less inclined to think so if you initially believed it to be innocent. In contrast, you are more than 90% certain it is guilty if you started from with the "likely guilty" prior. Note that if you started with no information (i.e., a flat prior), your posterior looks identical to your likelihood.

Each of the other three priors "pulled" the likelihood slightly towards its own peak.

Table 1: Degree of certainty in the posterior that the cat broke the frame, conditional on each prior in Figure 1.

Prior	Certainty of cat being guilty
No information	0.85
Uncertain	0.84
Likely innocent	0.65
Likely guilty	0.93

3.1 Bayesian updating: when posteriors become the new priors

What if we collected our information about the scene piece by piece? Since our posterior contains all our knowledge about the scene, we can update our current belief state after each new piece of evidence. In this way we can start with our prior (which may also be based on data collected previously), and add one observation at a time. To do this, we decide our original prior before collecting any data from this scene, collect data on the first observation, and calculate the first posterior. When we observe the second observation, we can include our first posterior as the second prior, we continue the process until all data are collected: the second posterior becomes the third prior, the third posterior becomes the fourth prior, and so on. This process is called **Bayesian updating**, and it represents the idea of updating our beliefs gradually as we collect new information: each time what was once new information becomes old information that we use to evaluate along with the new information. Richard McElreath has a good example of this in his second lecture in the Statistical Rethinking lecture series where he is trying to estimate the proportion of the Earth that is covered by water, and a related example on the wait times at Starbucks coffee places in his twelfth lecture, although this second example is a bit more complicated because it includes learning about wait times at different Starbucks at once.

4 Expressing uncertainty

I hope I have convinced you that Bayesian statistics can be used as an intuitive framework that is similar to our daily thought process. The prior indicates our information (or belief) prior to any data collection, the likelihood contains all information about the data we observed, and the posterior is our belief after updating our prior knowledge using the new evidence. Since each of these three are probability distributions, we can use them to express our uncertainty using intuitive measures of probability. For example, a 50% interval over the prior would provide us with a range of values we believe to be realistic with 50% certainty, even without collecting any data. While this may make some people uncomfortable, most people use this thought process daily: You do not have to have tried something to have an opinion about the potential outcome. However, if your information about the topic is limited, your prior should be appropriately vague. Excessively tight priors can produce unrealistic predictions on their own, but the same can be said about excessively vague priors. For example, let's look at the possible ranges of prior predictions from each of the priors in Figure 1, which are shown in Figure 4. As you may expect, the "flat and"no information" and "uncertain" priors result in a very wide variety of frequencies of "guilty verdicts", including the cat being essentially always innocent ($\theta = 0$) or always guilty ($\theta = 1$), both of which are unrealistic. A more informed (i.e., tighter) prior would be better at constraining such extreme values of θ.

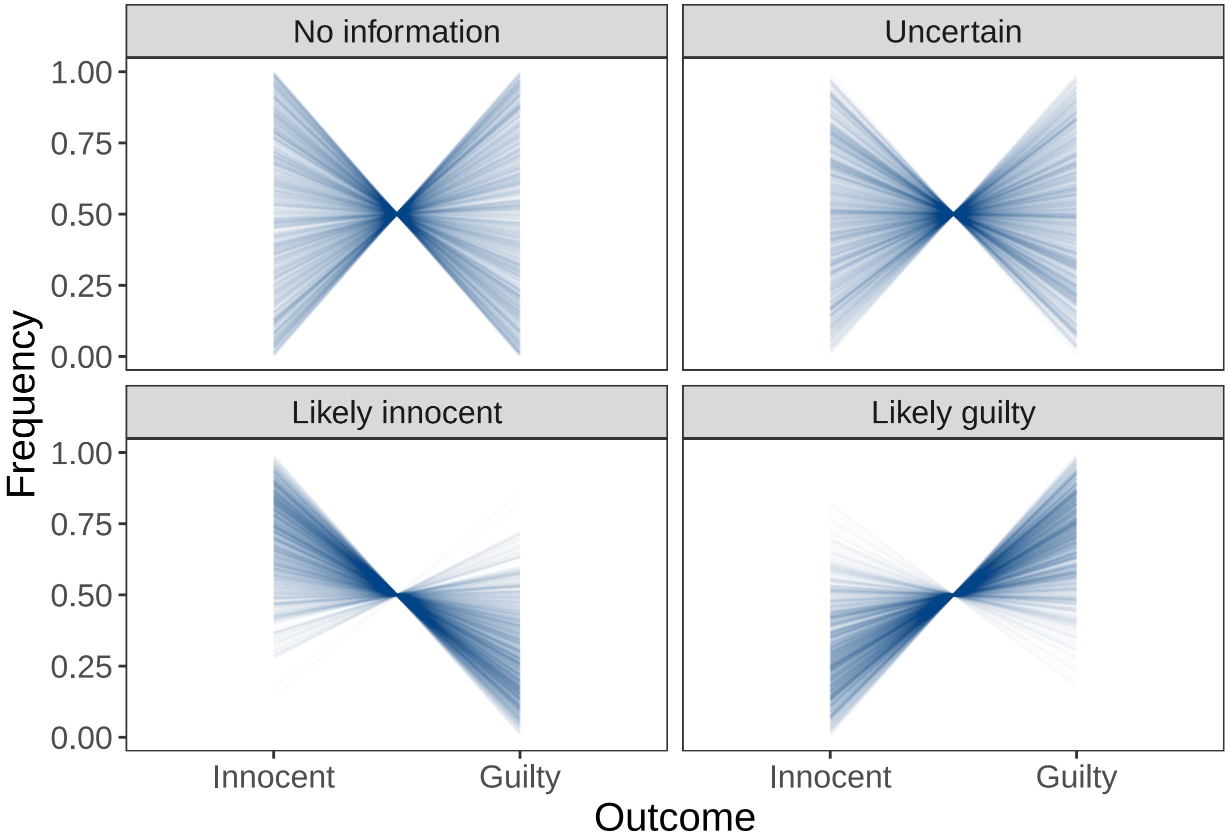


Figure 4: A thousand prior predictive simulations for each of the priors presented in Figure 1.

Using the simulations above, we can construct **credible intervals** for each prior that tell us the range of credible (i.e., believable) values with some degree of credibility. For example, the ribbons in Figure 5 below indicate the ranges of values with 50% credibility, meaning that the results are believable with 50%

certainty, given the prior that we started with.

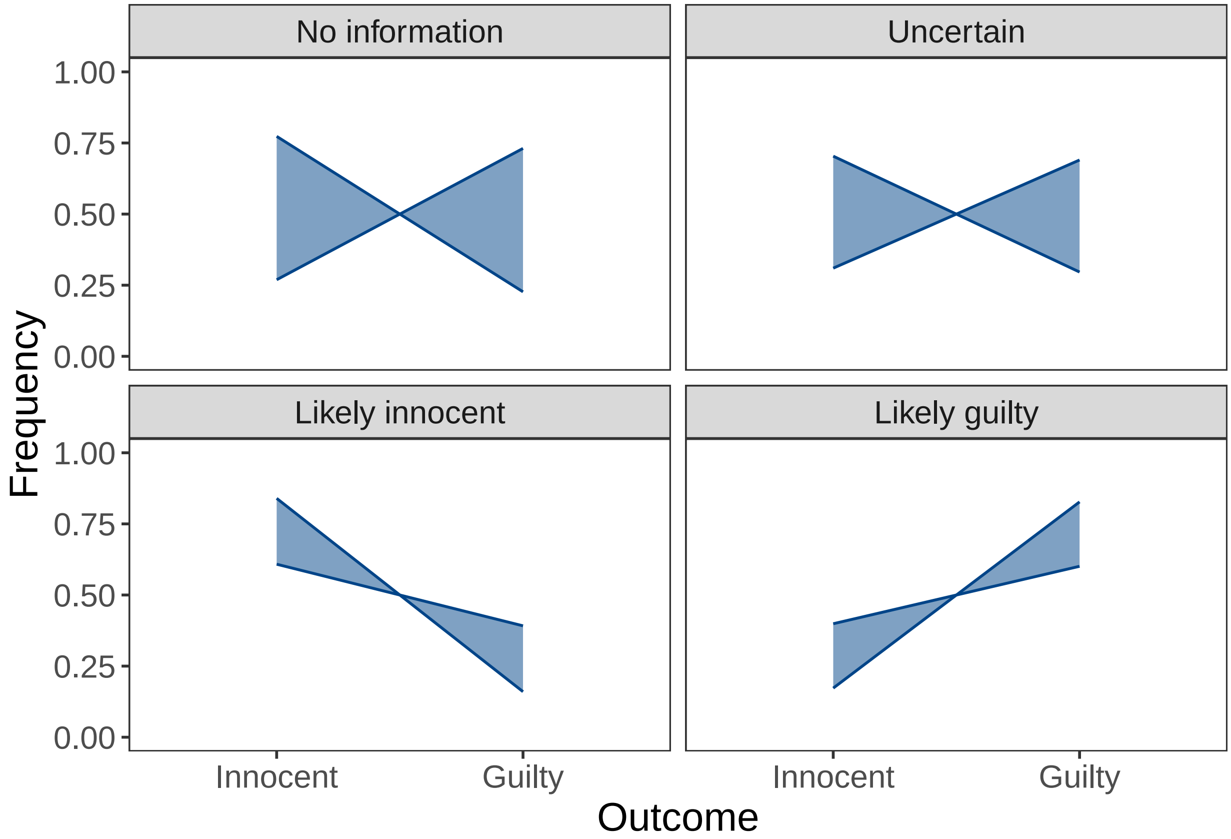


Figure 5: 50% credible intervals created using the 25% and 75% percentiles of the posteriors generated using each of the priors in Figure 1.

I constructed the intervals using the 0.25 and 0.75 quantiles of the priors, but I could use any other quantiles, too. Figure 6 shows the 50% credible intervals for quantiles 0.1 to 0.6.

There is nothing specifically wrong about either set of intervals, since both include ranges of values with 50% credibility, but the two figures tell different stories. You should choose the intervals you use carefully, and base your choice on what you are trying to show. Ideally, include entire distributions (like Figures 1, 2, and 3) rather than just point estimates and intervals. For example, if we wanted to summarize the posterior distributions in Figure 3 we could use something like Figure 7 below.

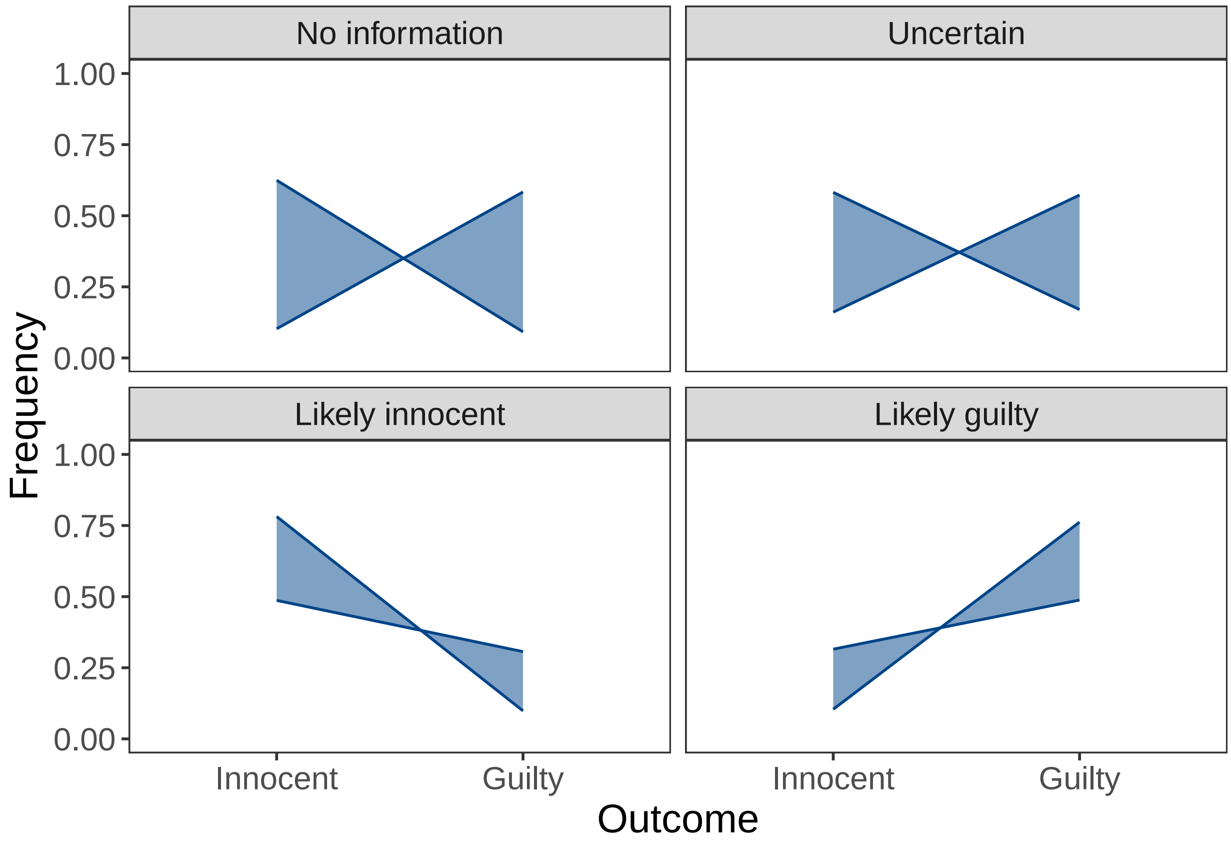


Figure 6: 50% credible of the priors intervals created using the 10% and 60% percentiles of the posteriors generated using each of the priors in Figure 1.

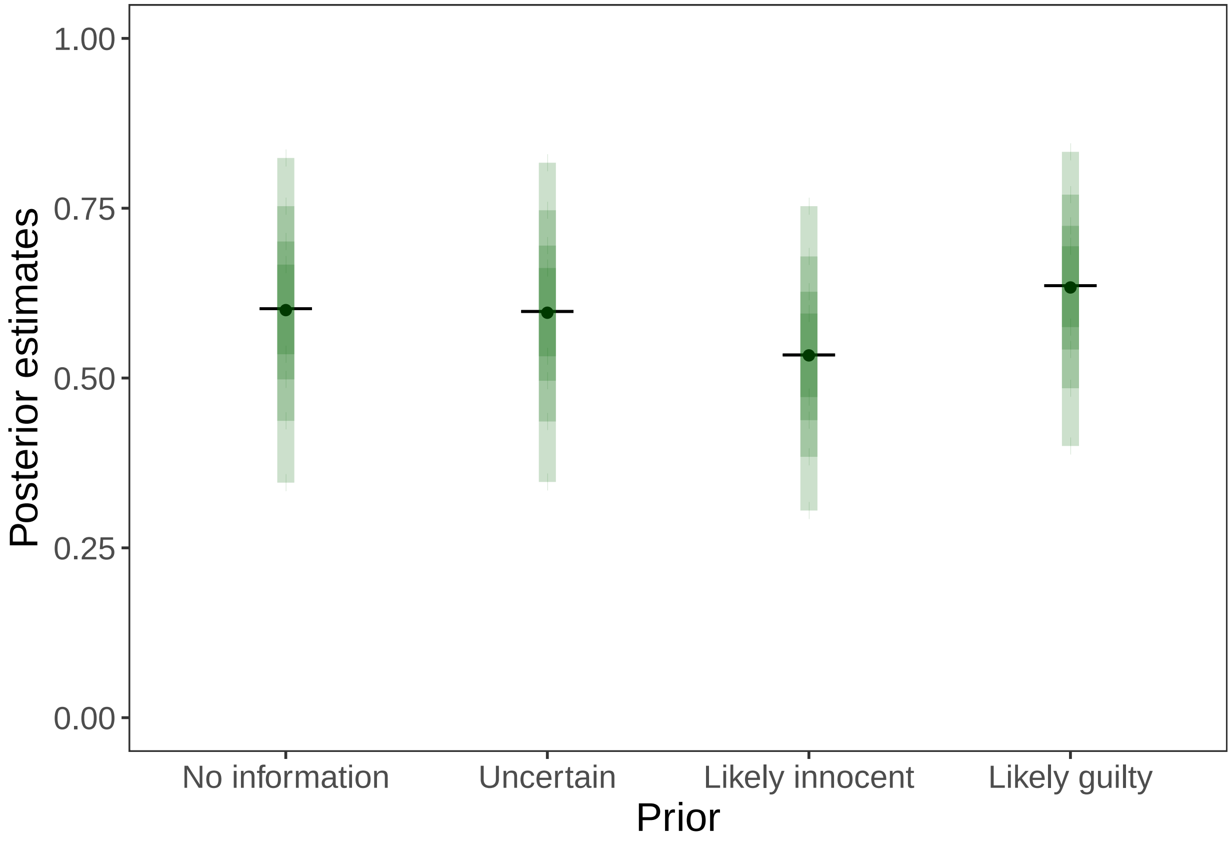


Figure 7: Summaries of the posterior distributions derived using each of the four priors in Figure 1. Points indicate the posterior mean, horizontal lines indicate the posterior median, and the green lines indicate the 50%, 70%, 90%, and 99% credible intervals.