Introduction to Bayesian Inference

Felipe José Bravo Márquez

January 25, 2021

Some Critics to the Frequentist Approach

- The statistical methods that we have discussed so far are known as frequentist (or classical) methods.
- The frequentist approach requires that all probabilities be defined by connection to the frequencies of events in very large samples.
- This leads to frequentist uncertainty being premised on imaginary resampling of data.
- If we were to repeat the measurement many many times, we would end up collecting a list of values that will have some pattern to it.
- It means also that parameters and models cannot have probability distributions, only measurements can.
- The distribution of these measurements is called a sampling distribution.
- This resampling is never done, and in general it doesn't even make sense.

Bayesian Inference

There is another approach to inference called Bayesian inference [Wasserman, 2013], which is based on the following postulates:

- Probability describes **degree of belief**, not limiting frequency.
  - We can make probability statements about lots of things, not just data which are subject to random variation.
  - For example, I might say that ”the probability that Albert Einstein drank a cup of tea on August 1, 1948” is .35.
  - This does not refer to any limiting frequency.
  - It reflects my strength of belief that the proposition is true.
- We can make probability statements about parameters, even though they are fixed constants.
- We make inferences about a parameter θ by producing a probability distribution for θ. Inferences, such as point estimates and interval estimates, may then be extracted from this distribution.

Bayesian Inference

- In modest terms, Bayesian data analysis is no more than counting the numbers of ways the data could happen, according to our assumptions [McElreath, 2020].
- In Bayesian analysis all alternative sequences of events that could have generated our data are evaluated.
- As we learn about what did happen, some of these alternative sequences are pruned.
- In the end, what remains is only what is logically consistent with our knowledge [McElreath, 2020].
- Warning: understanding the essence of Bayesian inference can be hard.
- The following toy example tries to explain it in a gentle way.

Counting Possibilities

- Suppose there's a bag, and it contains **four** marbles.
- These marbles come in two colors: **blue** and **white**.
- We know there are four marbles in the bag, but we don't know how many are of each color.
- We do know that there are five possibilities:
  (1) [○○○○], (2) [●○○○], (3) [●●○○], (4) [●●●○], (5) [●●●●]
- These are the only possibilities consistent with what we know about the contents of the bag. Call these five possibilities the **conjectures**.
- Our goal is to figure out which of these conjectures is most **plausible**, given some **evidence** about the contents of the bag.
- Evidence: A sequence of three marbles is pulled from the bag, one at a time, replacing the marble each time and shaking the bag, in that order.
- The sequence that emerges is:  , which is our **data**.

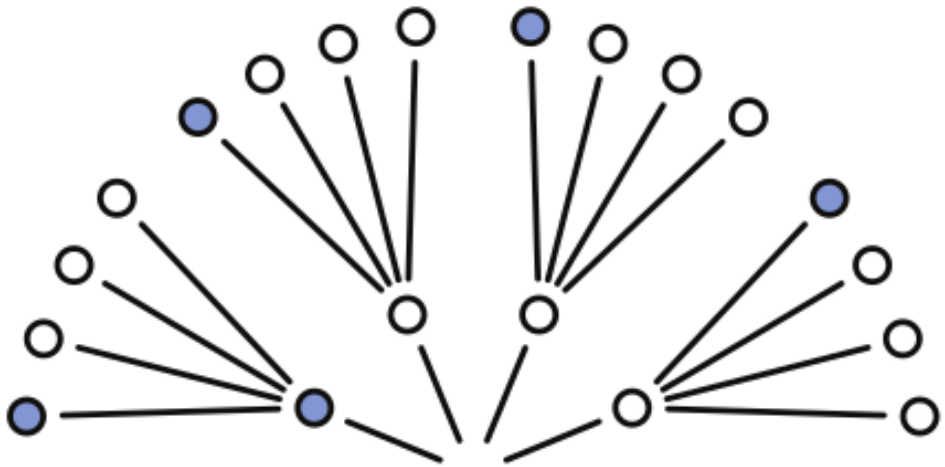
Counting Possibilities

- Now, let's see how to use the data to infer what's in the bag.
- Let's begin by considering just the single conjecture, [●○○○], that the bag contains one blue and three white marbles.
- On the first draw from the bag, one of four things could happen, corresponding to one of four marbles in the bag.

- Notice that even though the three white marbles look the same from a data perspective we just record the color of the marbles, after all they are really different events.
- This is important, because it means that there are three more ways to see ○ than to see ●.

Counting Possibilities

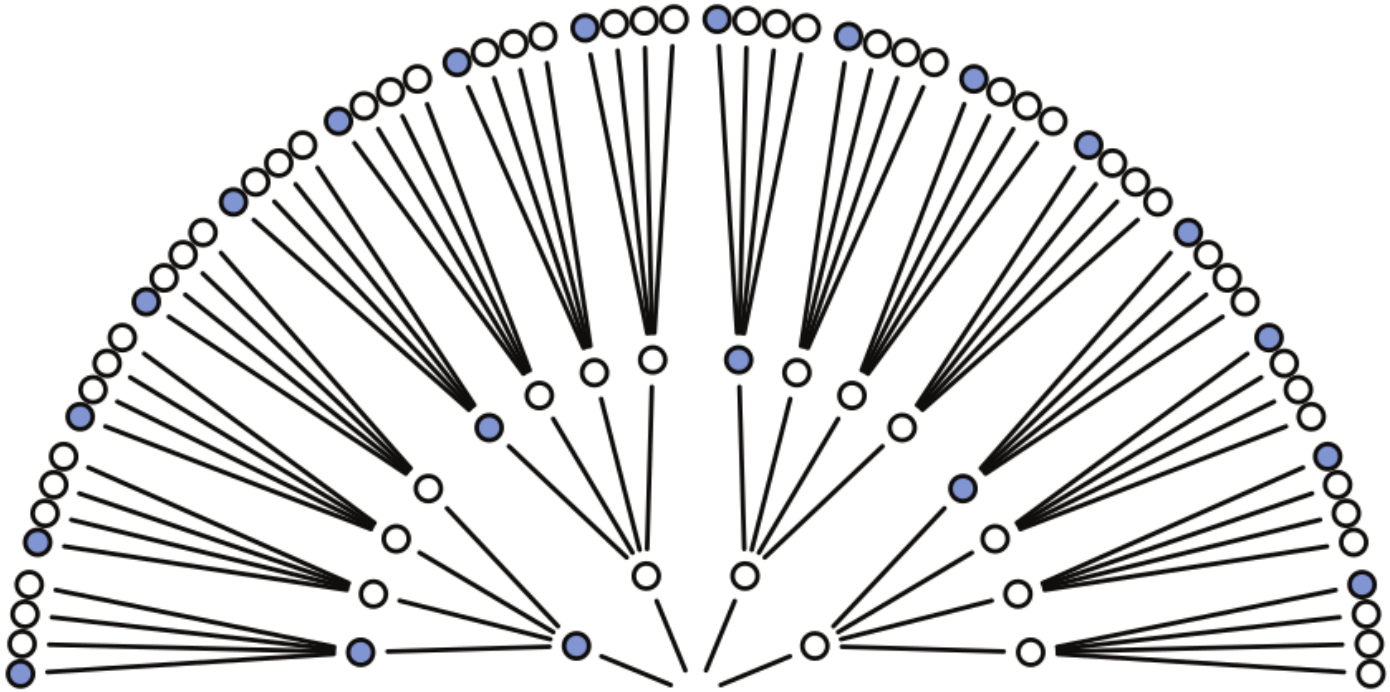
- Now consider the garden as we get another draw from the bag. It expands the garden out one layer:



- Now there are 16 possible paths through the garden, one for each pair of draws.
- On the second draw from the bag, each of the paths above again forks into four possible paths. Why?

Counting Possibilities

- Because we believe that our shaking of the bag gives each marble a fair chance at being drawn, regardless of which marble was drawn previously.
- The third layer is built in the same way, and the full garden is shown below:



- There are $4^3 = 64$ possible paths in total.

Counting Possibilities

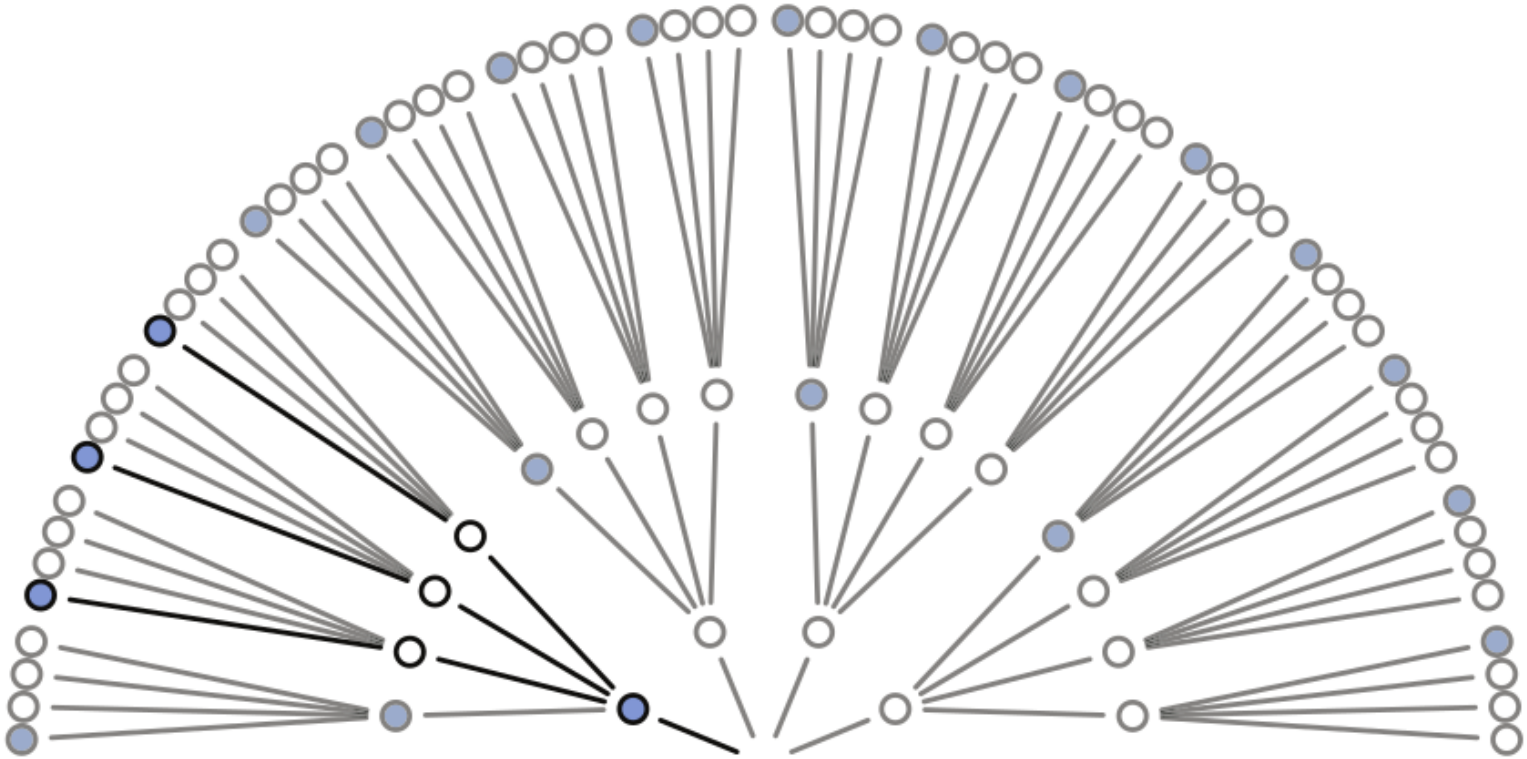
- As we consider each draw from the bag to get  , some of these paths are logically eliminated.
- The first draw tuned out to be  , recall, so the three white paths at the bottom are eliminated right away.
- If you imagine the real data tracing out a path, it must have passed through the one blue path near the origin.
- The second draw from the bag produces  , so three of the paths forking out of the first blue marble remain.

Counting Possibilities

- As the data trace out a path, we know it must have passed through one of those three white paths (after the first blue path).
- But we don't know which one, because we recorded only the color of each marble.
- Finally, the third draw is  .
- Each of the remaining three paths in the middle layer sustain one blue path, leaving a total of three ways for the sequence  to appear, assuming the bag contains  .

Counting Possibilities

- The figure below shows the forking paths again, now with logically eliminated paths grayed out.



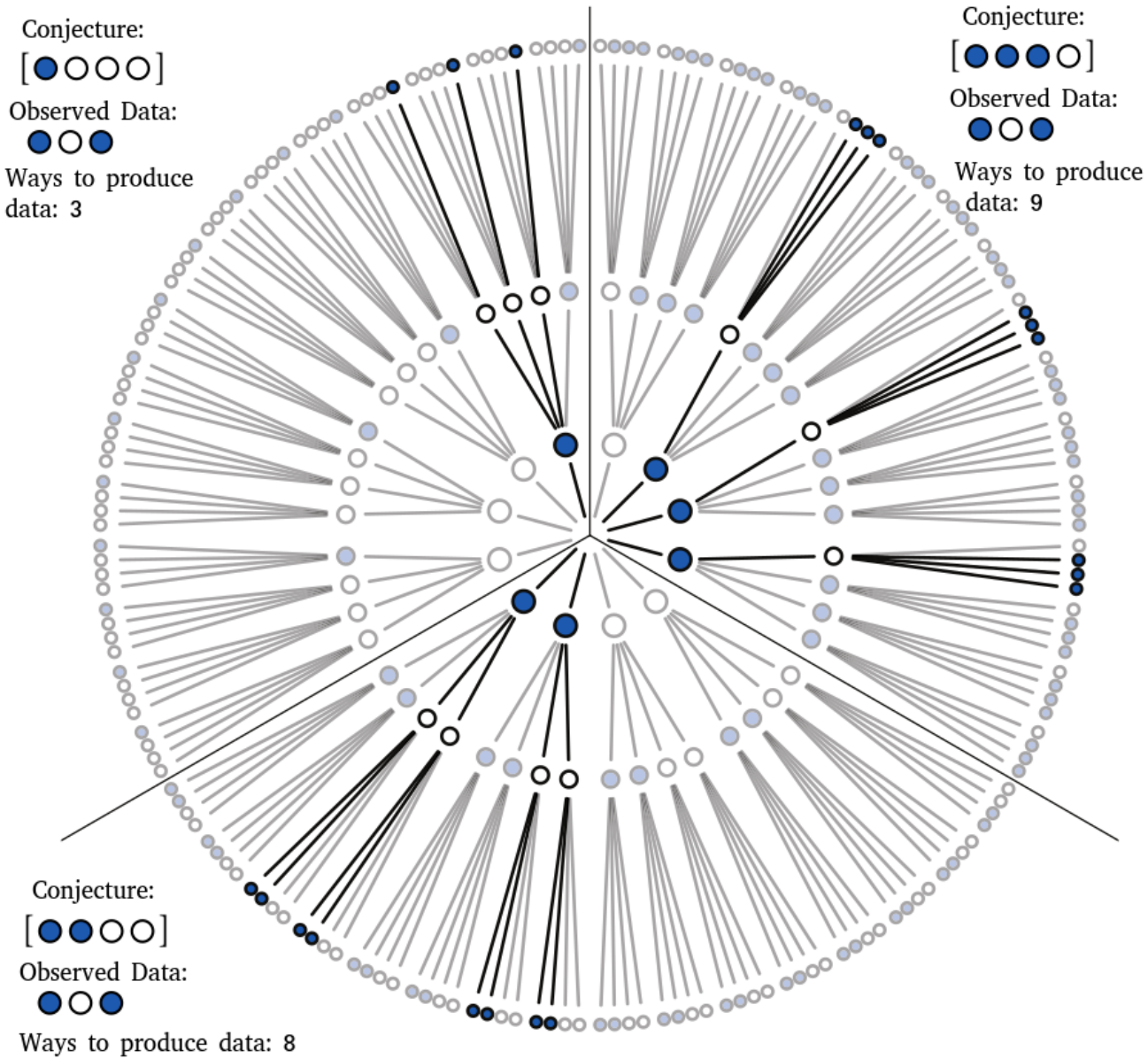
Counting Possibilities

- We can't be sure which of those three paths the actual data took.
- But as long as we're considering only the possibility that the bag contains one blue and three white marbles, we can be sure that the data took one of those three paths.
- Those are the only paths consistent with both our knowledge of the bag's contents (four marbles, white or blue) and the data ().
- This demonstrates that there are three (out of 64) ways for a bag containing to produce the data.
- We have no way to decide among these three ways.

Counting Possibilities

- The inferential power comes from comparing this count to the numbers of ways each of the other conjectures of the bag's contents could produce the same data.
- For example, consider the conjecture [○○○○].
- There are zero ways for this conjecture to produce the observed data, because even one ● is logically incompatible with it.
- The conjecture [●●●●] is likewise logically incompatible with the data.
- So we can eliminate these two conjectures, because neither provides even a single path that is consistent with the data.
- The next slide's figure displays all the paths for the reamaining three conjectures: [●○○○], [●●○○], and [●●●○].

Counting Possibilities



Counting Possibilities

- The number of ways to produce the data, for each conjecture, can be computed by first counting the number of paths in each “ring” of the garden and then by multiplying these counts together.

Conjecture	Ways to produce ●○●
[○○○○]	$0 \times 4 \times 0 = 0$
[●○○○]	$1 \times 3 \times 1 = 3$
[●●○○]	$2 \times 2 \times 2 = 8$
[●●●○]	$3 \times 1 \times 3 = 9$
[●●●●]	$4 \times 0 \times 4 = 0$

- By comparing these counts, we have part a way to rate the relative **plausibility** of each conjectured bag composition.

Combining other information

- We may have additional information about the relative plausibility of each conjecture.
- This information could arise from knowledge of how the contents of the bag were generated.
- It could also arise from previous data.
- Whatever the source, it would help to have a way to combine different sources of information to update the plausibilities.
- Luckily there is a natural solution: Just multiply the counts.

Combining other information

- Suppose that each conjecture is equally plausible at the start.
- So we just compare the counts of ways in which each conjecture is compatible with the observed data: ●○●.
- This comparison suggests that [●●●○] is slightly more plausible than [●●○○], and both are about three times more plausible than [●○○○].
- Since these are our initial counts, and we are going to update them next, let's label them **prior**.
- Now suppose we draw another marble from the bag to get another observation: ●.
- How can we update our plausibilities about each conjecture based on this new evidence?
- There are two choices discussed next.

Combining other information

- Option 1: draw a forking path with four layers and do the counting again.
- Option 2: Update previous counts (0, 3, 8, 9, 0) with the new information by multiplying the new count by the old count.
- Both approach are matematically identical as long as the new observation is logically independent of the previous observations.

Conjecture	Ways to produce ●	Prior counts	New count
[○○○○]	0	0	$0 \times 0 = 0$
[●○○○]	1	3	$3 \times 1 = 3$
[●●○○]	2	8	$8 \times 2 = 16$
[●●●○]	3	9	$9 \times 3 = 27$
[●●●●]	4	0	$0 \times 4 = 0$

Combining other information

- In the previous example, the prior data and new data are of the same type: marbles drawn from the bag.
- But in general, the prior data and new data can be of different types.
- Suppose for example that someone from the marble factory tells you that blue marbles are rare.
- So for every bag containing  , they made two bags containing  and three bags containing  .
- They also ensured that every bag contained at least one blue and one white marble.

Combining other information

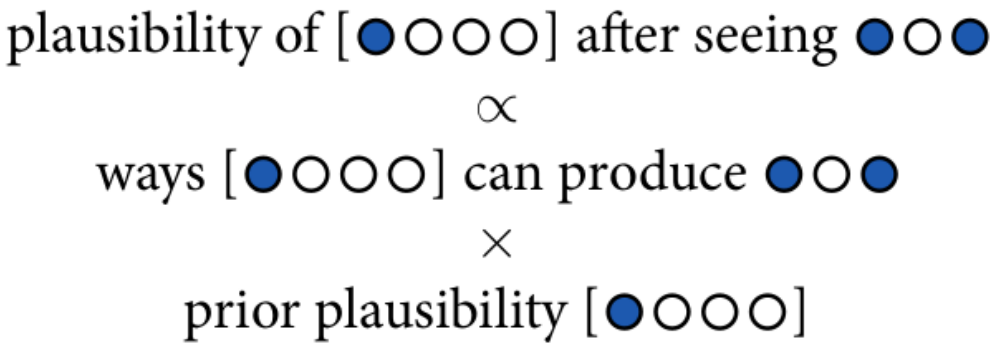
- We can update our counts again:

Conjecture	Prior count	Factory count	New count
[○○○○]	0	0	$0 \times 0 = 0$
[●○○○]	3	3	$3 \times 3 = 9$
[●●○○]	16	2	$16 \times 2 = 32$
[●●●○]	27	1	$27 \times 1 = 27$
[●●●●]	0	0	$0 \times 0 = 0$

- Now the conjecture [●●○○] is most plausible, but barely better than [●●●○].
- Is there a threshold difference in these counts at which we can safely decide that one of the conjectures is the correct one?
- We will explore this question next.

From counts to probability

- So far, we have defined the updated plausibility of each possible composition of the bag, after seeing the data, as:



- The problem of representing plausibilities as counts is that these numbers grow very quickly as the amount of data grows.
- It is better to standardize them to turn them into probabilities.

From counts to probability

- Now we will formalize the Bayesian framework using probabilities.
- Let index our conjecture with a parameter θ defined as the fractions of marbles from the bag that are blue:

  $\theta = 0 \rightarrow$ [○○○○], $\theta = 0.25 \rightarrow$ [●○○○], $\theta = 0.5 \rightarrow$ [●●○○], $\theta = 0.75 \rightarrow$ [●●●○], $\theta = 1 \rightarrow$ [●●●●].
- Let's call our data ●○● d.
- We construct probabilities by standardizing the plausibility so that the sum of the plausibilities for all possible conjectures will be one.

$$\text{plausibility of } \theta \text{ after } d = \frac{\text{ways } \theta \text{ can produce } d \times \text{prior plausibility } \theta}{\text{sum of products}} \tag{1}$$

- This is essentially the Bayes theorem:

$$\mathbb{P}(\theta|d) = \frac{\mathbb{P}(d|\theta) \times \mathbb{P}(\theta)}{\mathbb{P}(d)} \tag{2}$$

From counts to probability

- The denominator $\mathbb{P}(d)$ (that standardizes values to sum one) can be expressed by the law of total probabilities as:

$$\mathbb{P}(d) = \sum_{\theta} \mathbb{P}(d|\theta) \times \mathbb{P}(\theta) \tag{3}$$

- Let's consider the prior assumptions that all conjectures are equally plausible at the start, then $\mathbb{P}(\theta)$ is uniformly distributed.

θ	$\mathbb{P}(\theta)$	Ways to Produce Data	$\mathbb{P}(d\|\theta)$	$\mathbb{P}(\theta\|d) = \mathbb{P}(d\|\theta) * \mathbb{P}(\theta)/\mathbb{P}(d)$
0	1/5	0	0/64	$\frac{0/64*1/5}{0.0625}$ =0
0.25	1/5	3	3/64	$\frac{3/64*1/5}{0.0625}$ = 0.15
0.5	1/5	8	8/64	$\frac{8/64*1/5}{0.0625}$ = 0.4
0.75	1/5	9	9/64	$\frac{9/64*1/5}{0.0625}$ = 0.45
1	1/5	0	0/64	$\frac{0/64*1/5}{0.0625}$ = 0

- where $\mathbb{P}(d)$ = 1/5 * 0/64 + 1/5*3/64 + 1/5*8/64+1/5*9/64+1/5*0/64 = 0.0625

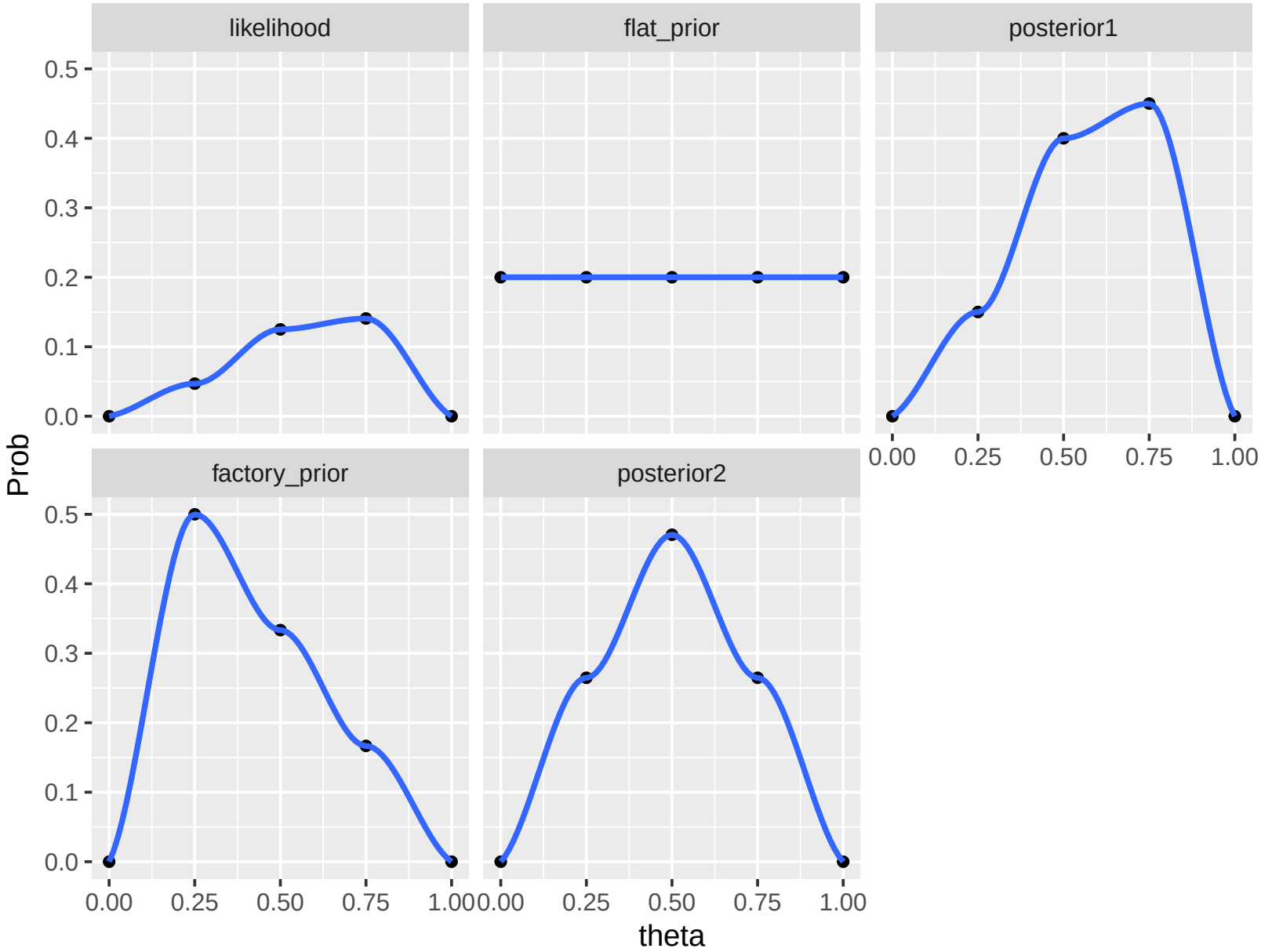
From counts to probability

- Let's use the factory counts information (blue marbles are rare) now in our prior assumptions of $\mathbb{P}(\theta)$.
- This can be done by normalizing the factory counts.
- Notice that this new prior assumption doesn't affect the ways each conjecture can generate the data and $\mathbb{P}(d|\theta)$ remains unchanged.

θ	Factory count	$\mathbb{P}(\theta)$	$\mathbb{P}(d\|\theta)$	$\mathbb{P}(\theta\|d) = \mathbb{P}(d\|\theta) * \mathbb{P}(\theta)/\mathbb{P}(d)$
0	0	0/6	0/64	$\frac{0/64*0/6}{0.08854167}$ =0
0.25	3	3/6	3/64	$\frac{3/64*3/6}{0.08854167}$ = 0.2647059
0.5	2	2/6	8/64	$\frac{8/64*2/6}{0.08854167}$ = 0.4705882
0.75	1	1/6	9/64	$\frac{9/64*1/6}{0.08854167}$ = 0.2647059
1	0	0/6	0/64	$\frac{0/64*0/6}{0.08854167}$ = 0

- where $\mathbb{P}(d)$= 0/6 * 0/64 + 3/6*3/64 + 2/6*8/64+1/6*9/64+0/6*0/64 = 0.08854167
- Two different prior assumptions led us to different values of $\mathbb{P}(\theta|d)$.

From counts to probability



Bayesian Components

Now we will introduce the names of the components of our Bayesian model.

Density and Mass functions

Because the Bayesian framework applies to both discrete and continuos random variables, we will use function f (instead of $\mathbb{P}$) to refer to both probability mass and density functions.

- **Parameter** θ: A way of indexing possible explanations of the data. In our example θ is a conjectured proportion of blue marbles.
- **Likelihood** $f(d|\theta)$: The relative number of ways that a value θ can produce the data. It is derived by enumerating all the possible data sequences that could have happened and then eliminating those sequences inconsistent with the data.
- **Prior probability** $f(\theta)$: The prior plausibility of any specific value of θ.
- **Posterior probability** $f(\theta|d)$: The new, updated plausibility of any specific θ.
- **Evidence** or **Average Likelihood** $f(d)$: the average probability of the data averaged over the prior. It's job is just to standardize the posterior, to ensure it sums (integrates) to one.

All these components are connected by the Bayes theorem!

Bayesian Components

- It is important to remark that in the Bayesian setting a parameter θ is random a variable, so we can make probability staments about it.
- Whereas in the frequentist approach parameters are considered unknown quantities.
- This is an important property of Bayesian inference: despite θ is an **unobseved variable** we can treat is a random variable and calculate $f(\theta)$ or $f(\theta|d)$.
- The likelihood function $f(d|\theta)$ is very similar to the likelihood function in the frequentist approach $f(d;\theta)$ but now we can condition of θ instead of just using it as function parameter.
- All the probability functions of a Bayesian model can correspond to either 1) a probability mass or 2) a density functions depending if the variable (observed or unobserved) is discrete or continous.
- A general equation that relates all Bayesian components (for both density and mass functions) is the following:

$$f(\theta|d) = \frac{f(d|\theta) \times f(\theta)}{f(d)} \tag{4}$$

Bayesian Components

- In the marble example θ is discrete so the prior and the posterior are probability mass functions.
- When θ is continous, the prior and the posterior are density functions, and the **evidence** is calculated with an integral called **marginal**

$$f(d) = \int_{\theta} f(d|\theta) f(\theta) d\theta \tag{5}$$

- In most cases this integral doesn't have a closed solution.
- However, there nice computational methods available that can efficiently approximate the posterior even when the evidence cannot be calculated.
- Next, we will go deeper into these concepts by building another Bayesian toy model.

A Globe Model

- We have a globe representing our planet.
- We want to estimate much of the surface is covered in water.
- We adopt the following strategy: we toss the globe up in the air, we catch it, record whether or not the surface under your right index finger is water or land.
- Then we toss the globe up in the air again and repeat the procedure.
- The first nine samples are: W L W W W L W L W where W indicates water and L indicates land.
- We observed 6 W and 3 L. This is our data.

A Globe Model

Designing a simple Bayesian model benefits from a design loop with three steps.

1. Data story: Motivate the model by narrating how the data might arise.
2. Update: Educate your model by feeding it the data.
3. Evaluate: All statistical models require supervision, leading to model revision.

A Globe Model

- You can motivate your data story by trying to explain how each piece of data is born.
- This usually means describing aspects of the underlying reality as well as the sampling process.
- The data story in this case is simply a restatement of the sampling process:
  1. The true proportion of water covering the globe is p.
  2. A single toss of the globe has a probability p of producing W and $1 - p$ of producing L.
  3. Each toss of the globe is independent of the others.
- The data story is then translated into a formal probability model where we assign distributions to our Bayesian components.
- Keep in mind that distribution functions are esentially shortcuts to the process of counting forking paths of the previous example.

A Globe Model

Let's define the variables of our model:

- The first variable is the unobserved parameter p, the proportion of water on the globe which is our target of inference.
- The other variables are observed in our data: the count of water W and the count of land L.
- The sum of these two variables is the number of globe tosses: N = W + L

Now, we can assign a **likelihood** function to our observed variables given the parameter that respects the two assumptions of our data story:

1. Every toss is independent of the other tosses.
2. The probability of W is the same on every toss.

A Globe Model

- The binomial distribution is the de facto discrete distribution for this kind of "coin tossing" problem:

$$f(W, L|p) = \frac{(W+L)!}{W!L!} p^W (1-p)^L$$

- Next, we need to assign initial probability values (our beliefs before observing data) for each possible value of p using a **prior** distribution.
- Recall that p (the proportion of water) can take any real value between 0 and 1.
- We will assume that all possible values of p are equally likely, which implies that p follows a **continous Uniform distribution** between 0 and 1.

$$f(p) = \text{Uniform}(0,1) = 1/(b-a) \text{ where b=1, and a=0} = 1$$

- This flat prior assumes that $p = 0$, $p = 0.5$ and $p = 1$ are all equally plausible.
- This is not the best prior information we can declare, considering that we already know that the earth cannot be completely covered by land ($p = 0$) or by water ($p = 1$).

A Globe Model

- Now that we have defined our model: variables, likelihood and prior, we can feed it with our data to to obtain the **posterior distribution** of p.
- The posterior distribution encodes updated plausabilities (or beliefs) for all parameter values condioned on the data.
- As we have already seen, it can be obtained using the Bayes formula:

$$f(p|W,L) = \frac{f(W,L|p) * f(p)}{f(W,L)}$$

- Where the denominator (evidence) makes sure that the posterior is a valid density function that integrates to one.
- It is not always possible to compute the posterior analytically unless we constrain our prior to special forms that are easy to do mathematics with.
- But bear in mind that in many of the interesting models in contemporary science we will need to approximate the posterior using computational techniques such as Markov Chain Montecarlo.
- This example is one the cases where the posterior can be found analytically as shown next.

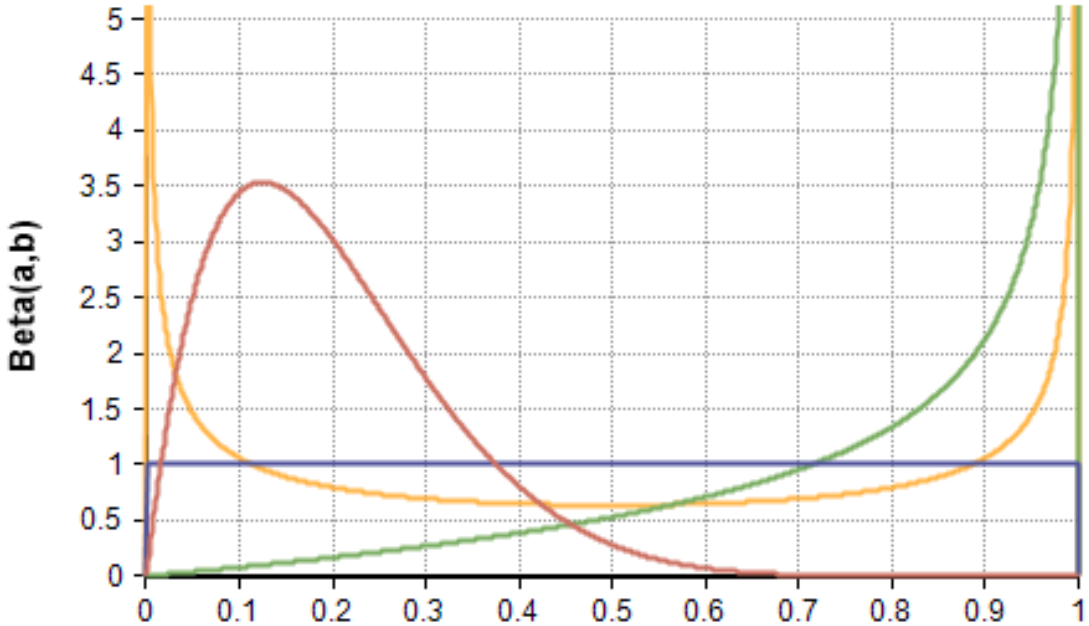
A Globe Model

- The posterior of our globe model with binomial likelihood and uniform prior has a closed form which is a Beta distribution.

$$\mathbb{P}(p|W, L) = \text{Beta}(W + 1, L + 1)$$

- This distribution is defined on the interval [0, 1] and is parameterized by two positive shape parameters, denoted by α and β.
- The Beta distribution is a continous distribution on probabilities.

$$\text{Beta}(\alpha, \beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}}$$



A Globe Model

- For any positive integer (such as W and L) the gamma function $\Gamma(n) = (n-1)!$
- Hence,

$$\text{Beta}(W+1, L+1) = \frac{p^W(1-p)^L}{\frac{\Gamma(W+1)\Gamma(L+1)}{\Gamma(W+1+L+1)}} = \frac{p^W(1-p)^L}{\frac{W!L!}{(W+L)!}} = \frac{(W+L)!}{W!L!}p^W(1-p)^L$$

- This surprisingly looks identical to the binomial distribution.
- This is because both distributions are very similar. The binomial distribution models the number of successes (W) and the beta distribution models the probaility p of sucess.
- Let's build our posterior from the likelihood and the prior:

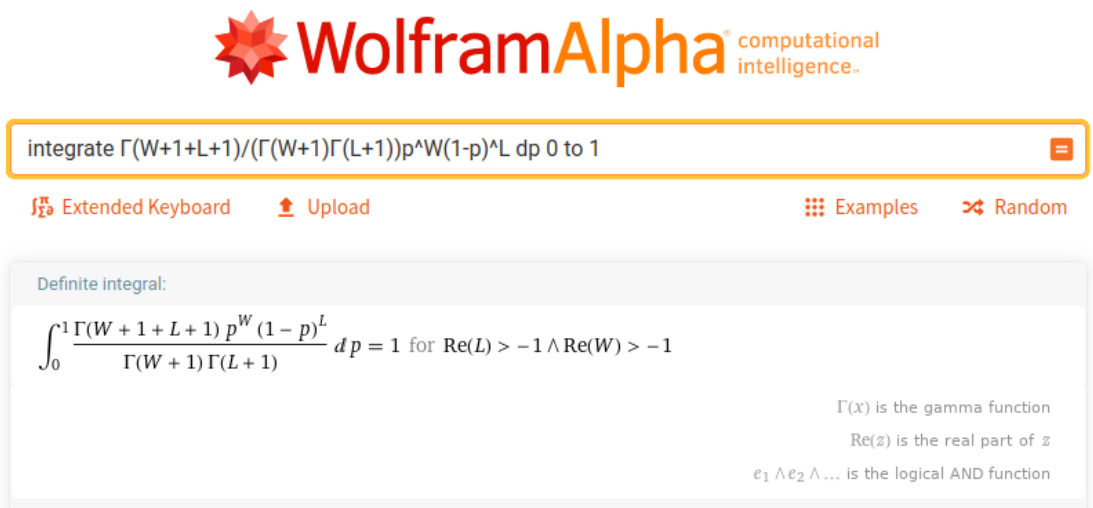
$$\mathbb{P}(p|W,L) = \frac{\mathbb{P}(W,L|p) * \mathbb{P}(p)}{\mathbb{P}(W,L)} = \frac{\mathbb{P}(W,L|p) * \mathbb{P}(p)}{\int_0^1 \mathbb{P}(W,L|p) * \mathbb{P}(p)dp}$$

A Globe Model

- Since $f(p) = 1$ (uniform prior) we have that

$$f(p|W, L) = \frac{f(W, L|p)}{\int_0^1 f(W, L|p)dp}$$

- The integral of the denominator is equal to 1 (essentially we are integrating a Beta distribution over its complete space of p):



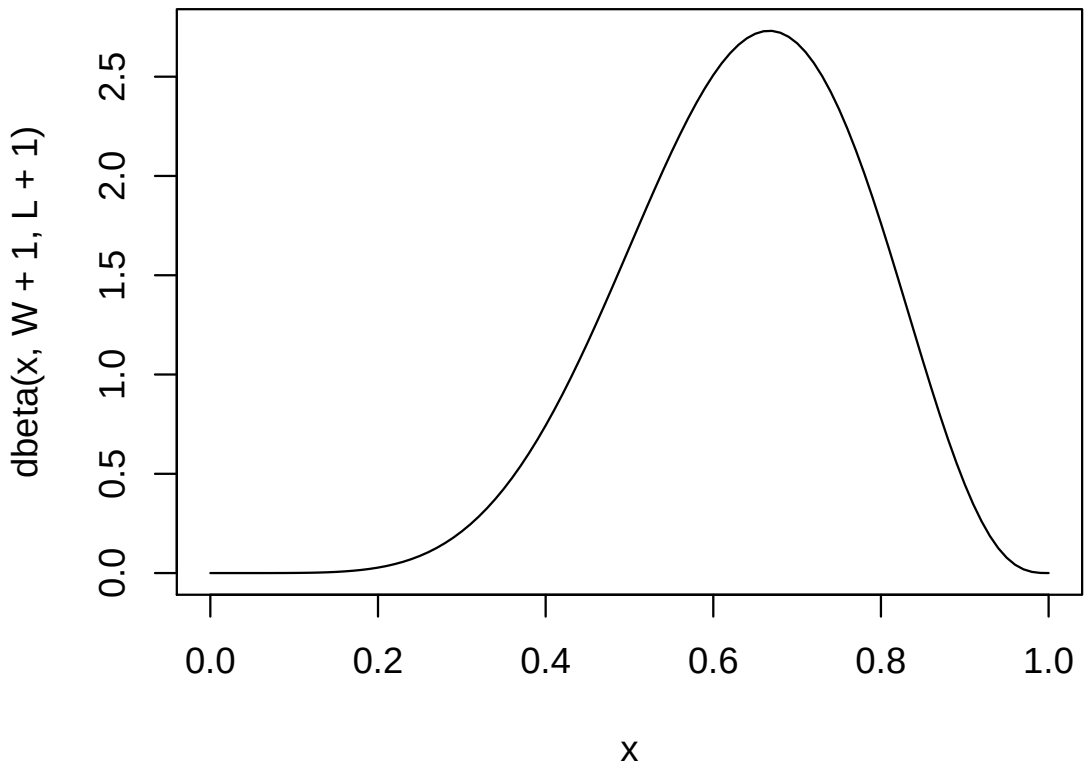
- So, we get

$$f(p|W, L) = \frac{(W + L)!}{W!L!}p^W(1 - p)^L = \text{Beta}(W + 1, L + 1)$$

A Globe Model

- So, in our globe tossing model ($W = 6, L = 3$) we can calculate the posterior distribution analytically

$$f(p|W = 6, L = 3) = \text{Beta}(7, 4)$$



A Globe Model

- We could calculate the posterior analytically because a property called **conjugate priors**.
- First of all, we must understand the Beta distribution is very flexible and can model a uniform distributions by setting α and β to 1, Beta(1,1)=Uniform
- We can change now our prior to a more general one using a Beta distribution $f(p) =$Beta(α, β).
- Now we can consider values of α, β that are more in line with our prior beliefs.
- The nice thing here is that when the prior follows a Beta distribution and the likelihood $f(W, L|p) =$ a Binomial one, he posterior takes the form of another Beta distribution with parameters $(\alpha + W, \beta + L)$.

A Globe Model

- The Beta distribution is conjugate distribution to binomial distribution, which means that the posterior distribution in the same probability distribution family as the prior.
- In simple words there are some families of conjugate distributions that can be used to calculate posterior distributions analytically.
- A very complete table of conjugate distributions is given in `https://en.wikipedia.org/wiki/Conjugate_prior`.
- In essence, conjugate priors constrain our choice of prior to special forms that are easy to do mathematics with.
- However, there are numerical techniques that allow us to accommodate any prior that is most useful for our inference problem, such as Markov Chain Monte Carlo (MCMC).

References I


McElreath, R. (2020).
Statistical rethinking: A Bayesian course with examples in R and Stan.
CRC press.


Wasserman, L. (2013).
All of statistics: a concise course in statistical inference.
Springer Science & Business Media.