Introduction to Bayesian Inference

Felipe José Bravo Márquez

November 20, 2020

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**.

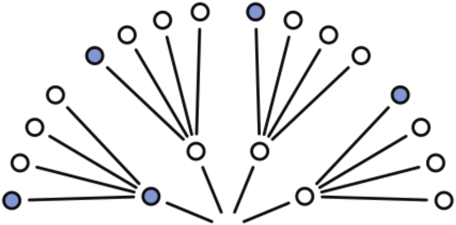
Counting Possibilities

- 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

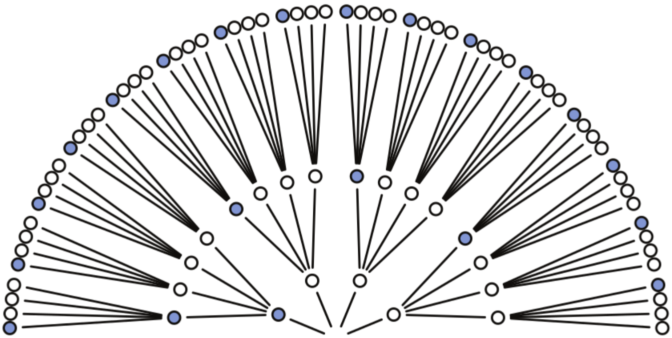
- Now consider the garden as we get another draw from the bag. It expands the garden out one layer:



- 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.

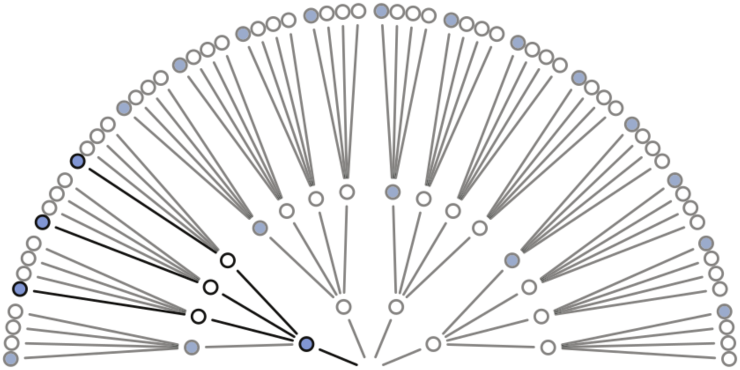
- As we consider each draw from the bag to get ●○●, some of these paths are logically eliminated.
- The first draw tuned out to be ●, recall, so the three white paths at the bottom are eliminated right away.
- If you imagine the real data tracing out a path, it must have passed through the one blue path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

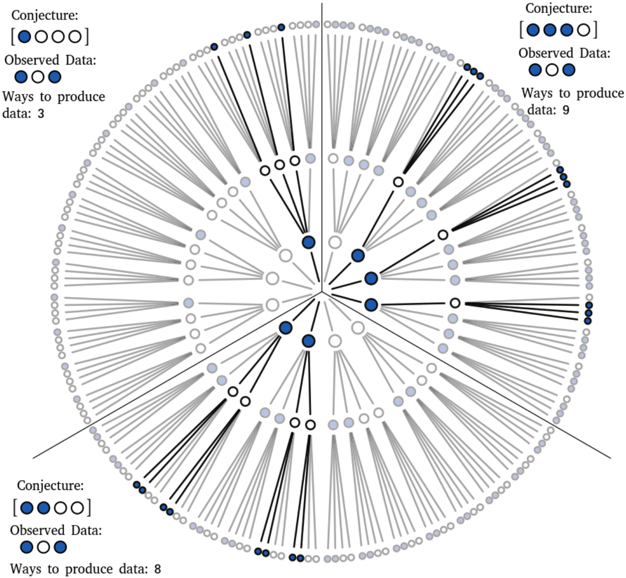


- We can't be sure which of those three paths the actual data took.
- But as long as we're considering only the possibility that the bag contains one blue and three white marbles, we can be sure that the data took one of those three paths.
- Those are the only paths consistent with both our knowledge 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:

$$\text{plausibility of } [●○○○] \text{ after seeing } ●○●$$
$$\propto$$
$$\text{ways } [●○○○] \text{ can produce } ●○●$$
$$\times$$
$$\text{prior plausibility } [●○○○]$$

- 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	$P(\theta)$	Ways to Produce Data	$P(d\|\theta)$	$P(\theta\|d) = P(d\|\theta) * P(\theta)/P(d)$
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 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 are) now in our prior assumptions of $P(\theta)$.
- 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 $P(d|\theta)$ remains unchanged.

θ	Factory count	$P(\theta)$	$P(d\|\theta)$	$P(\theta\|d) = P(d\|\theta) * P(\theta)/P(d)$
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 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 $\mathbb{P}(d|\theta)$.

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.