

Bayesian statistics

Kristian Wichmann

July 17, 2016

This document is based primarily on chapter 19 from Wonnacott and Wonnacott's "Introductory Statistics" (fifth edition) and the Coursera course "Bayesian Statistics" offered by Duke University.

1 Bayesian inference

1.1 The Bayesian view of probability

Bayesian statistics, as opposed to frequentist statistics, views probabilities merely as current opinion regarding the true state of the world. As new data is brought to light such opinion will be revised to reflect the new evidence. The way to update probabilities is prescribed by Bayes' theorem.

So Bayesian probabilities are subjective, in contrast to the objective probabilities of frequentist statistics.

1.2 Prior and posterior probabilities

Assume we have a given probability of an event A happening. This probability $P(A)$ is known as the *prior* in Bayesian terminology. Now, assume we know that event B has occurred - this constitutes evidence. We should now adjust our probability of event A to the conditional probability $P(A|B)$, which is known as the *posterior*. The two are related by Bayes' theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

1.2.1 Example: Radio quality

A given corporation produces radios. Of the last 200 truckloads of radios, 128 have been "bad" and 72 "good"; In the bad truckloads 44% of the radios

were defective. In the good truckloads only 15%. Now, we're faced with determining whether a new truckload of radios is good or bad. Initially, since all we have is the information that 128 out of 200 truckloads have been bad, our prior probabilities would be:

$$P(B) = \frac{128}{200} = 64\%, \quad P(G) = \frac{72}{200} = 36\% \quad (2)$$

Here, B refers to the event "Bad truckload", and G to the event "Good truckload". However, we now sample one of the radios from the truckload. This radio turns out to be defective. What are the updated, posterior probabilities of the truckload being good or bad? To answer this, we need Bayes' theorem:

$$P(B|D) = \frac{P(D|B)P(B)}{P(D)} \quad (3)$$

Here D refers to the event "Defective radio". $P(D|B)$ is the probability of a radio in a bad truckload being defective. We know that this is 44%. We know that $P(B) = 64\%$. But what is $P(D)$? By the law of total probability, this is:

$$P(D) = P(D|B)P(B) + P(D|G)P(G) = 44\% \cdot 64\% + 15\% \cdot 36\% = 33.56\% \quad (4)$$

Now, we can insert into equation (3):

$$P(B|D) = \frac{44\% \cdot 64\%}{33.56\%} = 83.9\% \quad (5)$$

By symmetry, the posterior probability of a good truckload has shrunk to $P(G|D) = 100\% - 83.9\% = 16.1\%$. The knowledge that the sample radio is defect makes us update our view of the world.

1.2.2 Radio quality with odds

One could also reformulate the example above using *odds*. The odds of a bad truckload is:

$$\frac{P(B)}{P(G)} = \frac{64\%}{36\%} = 1.78 \quad (6)$$

These are the prior odds. After the reveal of the defect radio, odds are:

$$\frac{P(B|D)}{P(G|D)} = \frac{83.9\%}{16.1\%} = 5.21 \quad (7)$$

These are the posterior odds. How are the two related? Let's use Bayes' theorem to find out:

$$\frac{P(B|D)}{P(G|D)} = \frac{P(D|B)P(B)/P(D)}{P(D|G)P(G)/P(D)} = \frac{P(D|B)P(B)}{P(D|G)P(G)} \quad (8)$$

So the prior odds times the quantity $\frac{P(D|B)}{P(G|B)}$, which can be interpreted as a *likelihood ratio*. The relation between these three quantities can be written:

$$\text{posterior odds} = \text{likelihood ratio} \cdot \text{prior odds} \quad (9)$$

1.3 Calculating posterior probabilities

In general, let θ be a parameter (or a vector of parameters) used to describe an event. In the example above, θ covered two options: Good or Bad truckload. In general, θ may represent many - even infinitely many - different outcomes. Let X represent new evidence. By Bayes' theorem:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} \quad (10)$$

For a given set of evidence X we wish to update on, $P(X)$ is a constant, so this may also be expressed:

$$P(\theta|X) \propto P(X|\theta)P(\theta) \quad (11)$$

Or reworded to resemble equation (9):

$$\text{posterior probability} \propto \text{likelihood} \cdot \text{prior probability} \quad (12)$$

1.3.1 Discrete example: Coin throws

Consider a coin which may or may not be biased. We initially think it's 50% likely to be fair, but still consider the possibility of the probability of getting heads p as 20%, 40%, 60%, and 80% possible, and assign each possibility an equal share of the remaining 50%. In other words, our prior is:

p_0	0.2	0.4	0.5	0.6	0.8
$P(p = p_0)$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{2}$	$\frac{1}{8}$	$\frac{1}{8}$

Now, we toss the coin three times and get all heads. What is the posterior distribution in this case? As usual, we can use Bayes' theorem to answer this:

$$P(p = p_0 | 3 \text{ heads}) = \frac{P(3 \text{ heads} | p = p_0)P(p = p_0)}{P(3 \text{ heads})} \quad (13)$$

Here, p_0 may take on any of our five considered values. The likelihood $P(3 \text{ heads} | p = p_0)$ is a binomial distribution:

$$P(3 \text{ heads} | p = p_0) = \binom{3}{3} p_0^3 (1 - p_0)^0 = p_0^3 \quad (14)$$

The probability $P(3 \text{ heads})$ can be calculated using the law of total probability:

$$\begin{aligned} P(3 \text{ heads}) &= \sum_{p_0} P(3 \text{ heads} | p = p_0) P(p = p_0) = \sum_{p_0} (1 - p_0)^3 P(p = p_0) = \\ &= (1 - 0.2)^3 \frac{1}{8} + (1 - 0.4)^3 \frac{1}{8} + (1 - 0.5)^3 \frac{1}{2} + (1 - 0.6)^3 \frac{1}{8} + (1 - 0.8)^3 \frac{1}{8} = \\ &= 0.1625 \end{aligned}$$

Now, we can calculate a posterior probability for each p_0 . For instance:

$$P(p = 0.2 | 3 \text{ heads}) = \frac{(1 - 0.2)^3 \cdot \frac{1}{8}}{0.1625} \approx 39.4\% \quad (15)$$

Repeating for the remaining 4 possibilities gives us the following table of the posterior distribution:

p_0	0.2	0.4	0.5	0.6	0.8
$P(p = p_0 3 \text{ heads})$	39.4%	16.6%	38.5%	4.9%	0.6%

Of course, we could also have disregarded the constants independent of p_0 - the denominator and the binomial coefficient (in general it will be different from 1) - and eventually normalize the resulting distribution.

1.3.2 Discrete example: The "two-armed bandit"

We have two slot machines - also known as "one-armed bandits" - to play. Let's call them M_1 and M_2 . We know that one of them is "good", in that the win rate is $\frac{1}{2}$, and that one of them is "bad", having only a win rate of $\frac{1}{3}$. But initially, we do not know which is which. So our prior is:

$$P(M_1 \text{ good}) = P(M_2 \text{ bad}) = \frac{1}{2}, \quad P(M_1 \text{ bad}) = P(M_2 \text{ good}) = \frac{1}{2} \quad (16)$$

But we do know the conditional probabilities for winning

$$P(\text{Win on } M_1 | M_1 \text{ good}) = P(\text{Win on } M_2 | M_2 \text{ good}) = \frac{1}{2} \quad (17)$$

$$P(\text{Win on } M_1 | M_1 \text{ bad}) = P(\text{Win on } M_2 | M_2 \text{ bad}) = \frac{1}{3} \quad (18)$$

And losing:

$$P(\text{Lose on } M_1 | M_1 \text{ good}) = P(\text{Lose on } M_2 | M_2 \text{ good}) = \frac{1}{2} \quad (19)$$

$$P(\text{Lose on } M_1 | M_1 \text{ bad}) = P(\text{Lose on } M_2 | M_2 \text{ bad}) = \frac{2}{3} \quad (20)$$

Now, let's say that we play M_1 and we win. What is the posterior probability of M_1 being good? We use Bayes' theorem:

$$P(M_1 \text{ good} | \text{Win on } M_1) = \quad (21)$$

$$\frac{P(\text{Win on } M_1 | M_1 \text{ good})P(M_1 \text{ good})}{P(\text{Win on } M_1 | M_1 \text{ good})P(M_1 \text{ good}) + P(\text{Win on } M_1 | M_1 \text{ bad})P(M_1 \text{ bad})} \quad (22)$$

This is equal to:

$$\frac{\frac{1}{2} \cdot \frac{1}{2}}{\frac{1}{2} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{2}} = \frac{1/4}{5/12} = \frac{3}{5} = 60\% \quad (23)$$

This also means, that the posterior probability of M_2 being good is 40%. If we had instead played M_1 and lost the corresponding posterior would be:

$$P(M_1 \text{ good} | \text{Lose on } M_1) = \quad (24)$$

$$\frac{P(\text{Lose on } M_1 | M_1 \text{ good})P(M_1 \text{ good})}{P(\text{Lose on } M_1 | M_1 \text{ good})P(M_1 \text{ good}) + P(\text{Lose on } M_1 | M_1 \text{ bad})P(M_1 \text{ bad})} \quad (25)$$

This is equal to:

$$\frac{\frac{1}{2} \cdot \frac{1}{2}}{\frac{1}{2} \cdot \frac{1}{2} + \frac{2}{3} \cdot \frac{1}{2}} = \frac{1/4}{7/12} = \frac{3}{7} \approx 42.9\% \quad (26)$$

1.3.3 Continuous example: Radio quality revisited

The radio manufacturing company from above have made further inquiries into the distribution of percentages of defective radios in a truckload. In this situation, the general parameter θ from last section corresponds to the defective percentage, which we will call π . It turns out, that this percentage seems to follow a beta distribution with parameters $\alpha = 2$ and $\beta = 4$:

$$P(\pi) \propto \pi(1 - \pi)^3 \quad (27)$$

Since beta distributions are continuous, this is a probability *density* of finding a truckload with a defective percentage of π , rather than a probability - this is our prior. We're now presented with a new truckload of radios. Without any further information, our best bet is the prior distribution. However, we now inspect the truckload, taking a random sample of 5 radios. It turns out, that 3 of those are defective. How do we account for this new information to update our distribution? According to equation (11), we need to find

the likelihood $P(X|\pi)$. Here X refers to the evidence of getting 3 out of 5 defective radios. In other words, this is a binomial likelihood¹:

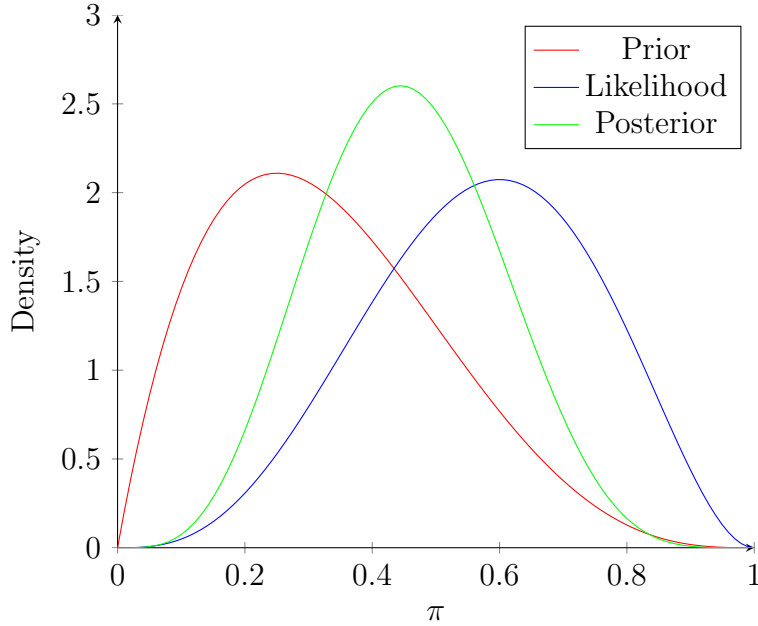
$$P(X|\pi) = \binom{5}{3} \pi^3 (1 - \pi)^2 \propto \pi^3 (1 - \pi)^2 \quad (28)$$

So the posterior distribution also turns out to be a beta distribution:

$$P(\pi|X) \propto \pi^4 (1 - \pi)^5 \quad (29)$$

This is a beta distribution with parameters $\alpha = 5$ and $\beta = 6$. When both distributions are of the same type, we speak of *conjugate distributions*. Given a likelihood function, a prior that makes the distributions conjugate is known as a *conjugate prior*.

The graph below shows the three distributions for comparison². The prior is relatively optimistic regarding quality, and the likelihood less so. The posterior is a compromise between the two.



2 Bayesian inference for the beta distribution

We've already seen the beta distribution in action in the example above. The distribution is useful, because it's appropriate to assign to something that is

¹This assumes that the truckload is large enough that the consecutive picking of radios to test does not affect π significantly.

²The likelihood isn't a distribution, since it does not sum to 1. However, for clarity, in this figure it has been scaled as if it did.

itself a probability.

3 Bayesian inference for the normal distribution