Quantitative Methods for Cognitive Scientists

Bayesian Inference and Parameter Estimation

Emre Neftci

Department of Cognitive Sciences, UC Irvine,

May 21, 2019

Emre Neftci

eneftci@uci.edu

https://canvas.eee.uci.edu/courses/16991

The Limitations of Maximum Likelihood

- **Lack of uncertainty estimation.** Consider two scenarios:
  - Subject 1 responds to 9 questions out of 10 correctly
  - Subject 2 responds to 18 questions out of 20 correctly
  - Using Maximum Likelihood and a binomial distribution, our estimation of r, the rate of successes will be .9
  - However, our data from subject 2 should be more certain. Maximum Likelihood Estimation doesn't capture this uncertainty.

The Limitations of Maximum Likelihood

- **Lack of uncertainty estimation.** Consider two scenarios:
  - Subject 1 responds to 9 questions out of 10 correctly
  - Subject 2 responds to 18 questions out of 20 correctly
  - Using Maximum Likelihood and a binomial distribution, our estimation of r, the rate of successes will be .9
  - However, our data from subject 2 should be more certain. Maximum Likelihood Estimation doesn't capture this uncertainty.
- **Overfitting:** If the number of parameters is much larger than the number of data points, the model cannot generalize to new data points. We will see more of this when we study neural networks

The Limitations of Maximum Likelihood

- **Lack of uncertainty estimation.** Consider two scenarios:
  - Subject 1 responds to 9 questions out of 10 correctly
  - Subject 2 responds to 18 questions out of 20 correctly
  - Using Maximum Likelihood and a binomial distribution, our estimation of r, the rate of successes will be .9
  - However, our data from subject 2 should be more certain. Maximum Likelihood Estimation doesn't capture this uncertainty.
- **Overfitting:** If the number of parameters is much larger than the number of data points, the model cannot generalize to new data points. We will see more of this when we study neural networks
- **No way to introduce prior.** We know that dice tend to be unbiased, but the likelihood has no way of introducing this prior knowledge

Bayesian parameter estimation can solve all these problems

From Maximum Likelihood to Bayesian Parameter Estimation

- The likelihood is not a probability function. In fact in general $\int \mathrm{d}\theta P(X|\theta) \neq 1$.
- This means that we cannot associate uncertainty to our estimates.
- Bayesian parameter estimation takes a related but slightly different approach to find the uncertainty of the parameters

- Recall that the **conditional probability** of rv X_1 given X_2, denoted $P(X_1|X_2)$, is the probability of observing X_1 given that we have observed X_2.

Bayes Rule

- Recall that the **conditional probability** of rv X_1 given X_2, denoted $P(X_1|X_2)$, is the probability of observing X_1 given that we have observed X_2.
- The conditional probability is given by:

$$P(X_1|X_2)P(X_2) = P(X_1, X_2)$$

Bayes Rule

- Recall that the **conditional probability** of rv X_1 given X_2, denoted $P(X_1|X_2)$, is the probability of observing X_1 given that we have observed X_2.
- The conditional probability is given by:

$$P(X_1|X_2)P(X_2) = P(X_1, X_2)$$

- The reverse is also true and noting that $P(X_1, X_2) = P(X_2, X_1)$:

$$P(X_2|X_1)P(X_1) = P(X_1, X_2)$$

Bayes Rule

- Recall that the **conditional probability** of rv X_1 given X_2, denoted $P(X_1|X_2)$, is the probability of observing X_1 given that we have observed X_2.
- The conditional probability is given by:

$$P(X_1|X_2)P(X_2) = P(X_1, X_2)$$

- The reverse is also true and noting that $P(X_1, X_2) = P(X_2, X_1)$:

$$P(X_2|X_1)P(X_1) = P(X_1, X_2)$$

- Because both are equal to $P(X_1, X_2)$, we obtain

$$P(X_1|X_2) = \frac{P(X_2|X_1)P(X_1)}{P(X_2)}$$

Bayes Rule

- Recall that the **conditional probability** of rv X_1 given X_2, denoted $P(X_1|X_2)$, is the probability of observing X_1 given that we have observed X_2.
- The conditional probability is given by:

$$P(X_1|X_2)P(X_2) = P(X_1, X_2)$$

- The reverse is also true and noting that $P(X_1, X_2) = P(X_2, X_1)$:

$$P(X_2|X_1)P(X_1) = P(X_1, X_2)$$

- Because both are equal to $P(X_1, X_2)$, we obtain

$$P(X_1|X_2) = \frac{P(X_2|X_1)P(X_1)}{P(X_2)}$$

- This is the **Bayes Rule.** It allows to "invert" conditional probabilities. Note that $P(X_1|X_2) \neq P(X_2|X_1)$!

$$P(X_1|X_2) \neq P(X_2|X_1)$$

Many bad decision are made by confusing $P(X_1|X_2)$ and $P(X_1|X_2)$

- "The probability of a person being pregnant given that they are female differs from the probability of a person being female given that they are pregnant"

$$P(X_1|X_2) \neq P(X_2|X_1)$$

Many bad decision are made by confusing $P(X_1|X_2)$ and $P(X_1|X_2)$

- “The probability of a person being pregnant given that they are female differs from the probability of a person being female given that they are pregnant”
- “The probability of death given a shark attack is different than the probability of a shark attack given death”

We can *marginalize* out random variables by summing over them:

$$P(X_1) = \sum_x P(X_1, X_2 = x)$$

In the continuous case, we must integrate over them:

$$P(X_1) = \int P(X_1, X_2 = x)\mathrm{d}x$$

- In assignment 3, question 2, you did exactly this when you over over all branches of the decision tree to calculate $P(Orangeball)$.

Joint and Marginal Probabilities, Example

Table 6.1 Joint and marginal probabilities

	Gender (a)		
Education (b)	Male	Female	Marginal
High School	40	30	70
College	40	40	80
Graduate degree	30	22	52
Marginal	110	92	202

- This table summarizes a hypothetical survey of 202 people who are classified according to their gender and level of education
- These numbers can be converted into probabilities: for example, the probability of someone in the sample being male AND having a high school diploma is 40/202 = 0.198 (this is the joint probability)
- This marginal probability is obtained by ignoring level of education, which in the present case means summing across the outcomes for education: P(a = male) = (40 + 40 + 30)/202 = 0.54.

Bayes Theorem in its most useful Form

$$P(X_1|X_2) = \frac{P(X_2|X_1)P(X_1)}{P(X_2)}$$

- The denominator $P(X_2)$ is a marginal probability. We can express the denominator by first unpacking it into its constituents, and by then re-expressing the individual joint probabilities by their equivalent conditional probabilities

$$P(X_2) = \sum_x P(X_1 = x, X_2) = \sum_x P(X_2|X_1 = x)P(X_1 = x)$$

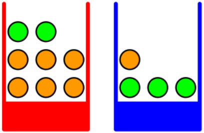
- Substituting for the denominator in the Bayes Rule we get:

$$P(X_1|X_2) = \frac{P(X_2|X_1)P(X_1)}{\sum_x P(X_2|X_1 = x)P(X_1 = x)}$$

- "The posterior distribution is given by the fraction formed by the probability of the particular outcome that was observed in an experiment given our prior knowledge of the parameters, compared to all possible outcomes that could have been observed"

Example

Balls and Boxes



Assume we pick a ball from the red box with probability 40% and from the blue box with probability 60%.

- Given that we've picked a green ball, what is the probability that we picked from the blue box

Summary

- The Likelihood function is not a probability distribution: it does not give us any sense of uncertainty
- Marginal Probabilities: $P(X_1) = \sum_x P(X_1, X_2 = x)$
- Bayes Rule: $P(X_1|X_2) = \frac{P(X_2|X_1)P(X_1)}{P(X_2)}$
- Bayes Rule in its most useful form:

$$P(X_1|X_2) = \frac{P(X_2|X_1)P(X_1)}{\sum_x P(X_2|X_1 = x)P(X_1 = x)}$$

- $P(X_1|X_2) \neq P(X_2|X_1)$

Summary

- The Likelihood function is not a probability distribution: it does not give us any sense of uncertainty
- Marginal Probabilities: $P(X_1) = \sum_x P(X_1, X_2 = x)$
- Bayes Rule: $P(X_1|X_2) = \frac{P(X_2|X_1)P(X_1)}{P(X_2)}$
- Bayes Rule in its most useful form:

$$P(X_1|X_2) = \frac{P(X_2|X_1)P(X_1)}{\sum_x P(X_2|X_1 = x)P(X_1 = x)}$$

- $P(X_1|X_2) \neq P(X_2|X_1)$
- Shark attacks are of little concern

The Bayes rule can be used for parameter estimation by replacing X_1 with the model and X_2 with the data. Let's call them Θ and $\mathbf{D}$. For discrete Θ:

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

For continuous Θ:

$$P(\Theta|D) = \frac{P(D|\Theta)P(\theta)}{\int P(D|\Theta=\theta)P(\Theta=\theta)\mathrm{d}\theta}$$

Terminology

$$P(\Theta|D) = \frac{P(D|\Theta)P(\theta)}{\int P(D|\Theta = \theta)P(\Theta = \theta)\mathrm{d}\theta}$$

- $P(\Theta|D)$ is called the **Posterior distribution.** The posterior distribution describes the distribution of θ after the fact (hence posterior) that we have observed D.
- $P(\Theta)$ is called the **prior.** It reflects our prior knowledge (bias) about the model.
- $P(D|\Theta)$ is the **Likelihood function** (as seen earlier), describing how likely the data D is given the parameter is equal to Θ.

Memory Recall: Uniform Prior

k is the number of correct answers out of $n = 10$, $r(t)$ is the correct answer rate after t days

- Likelihood: $P(k|r,n) = \binom{n}{k} r^k (1-r)^{n-k}$

Memory Recall: Uniform Prior

k is the number of correct answers out of $n = 10$, $r(t)$ is the correct answer rate after t days

- Likelihood: $P(k|r,n) = \binom{n}{k} r^k (1-r)^{n-k}$
- Prior: For now, let's consider a uniform prior, meaning that $P(r) = 1$. Note that $P(r)$ is a probability distribution over a continuous random variable r.

Memory Recall: Uniform Prior

k is the number of correct answers out of $n = 10$, $r(t)$ is the correct answer rate after t days

- Likelihood: $P(k|r, n) = \binom{n}{k} r^k (1-r)^{n-k}$
- Prior: For now, let's consider a uniform prior, meaning that $P(r) = 1$. Note that $P(r)$ is a probability distribution over a continuous random variable r.
- Data: k=9

k is the number of correct answers out of $n = 10$, $r(t)$ is the correct answer rate after t days

- Likelihood: $P(k|r, n) = \binom{n}{k} r^k (1 - r)^{n-k}$
- Prior: For now, let's consider a uniform prior, meaning that $P(r) = 1$. Note that $P(r)$ is a probability distribution over a continuous random variable r.
- Data: k=9

The posterior becomes

$$P(r|n, k) = \frac{\binom{n}{k} r^k (1 - r)^{n-k}}{\int \binom{n}{k} p^k (1 - p)^{n-k} dp}$$

Memory Recall: Uniform Prior

The posterior becomes

$$P(r|n,k) = \frac{\binom{n}{k} r^k (1-r)^{n-k}}{\int \binom{n}{k} p^k (1-p)^{n-k} dp}$$

Memory Recall: Uniform Prior

The posterior becomes

$$P(r|n,k) = \frac{\binom{n}{k} r^k (1-r)^{n-k}}{\int \binom{n}{k} p^k (1-p)^{n-k} dp}$$

$$P(r|n,k) = \frac{\binom{n}{k} r^k (1-r)^{n-k}}{\binom{n}{k} \int p^k (1-p)^{n-k} dp}$$

The posterior becomes

$$P(r|n,k) = \frac{\binom{n}{k} r^k (1-r)^{n-k}}{\int \binom{n}{k} p^k (1-p)^{n-k} dp}$$

$$P(r|n,k) = \frac{\binom{n}{k} r^k (1-r)^{n-k}}{\binom{n}{k} \int p^k (1-p)^{n-k} dp}$$

$$P(r|n,k) = \frac{r^k (1-r)^{n-k}}{\int p^k (1-p)^{n-k} dp}$$

The denominator can be evaluated (not straightforward). It is $\frac{k!(n-k)!}{(n+1)!}$, where $k! = 1 * 2 * 3 * 4 * ... * k$, So the final posterior becomes:

Memory Recall: Uniform Prior

The posterior becomes

$$P(r|n,k) = \frac{\binom{n}{k} r^k (1-r)^{n-k}}{\int \binom{n}{k} p^k (1-p)^{n-k} dp}$$

$$P(r|n,k) = \frac{\binom{n}{k} r^k (1-r)^{n-k}}{\binom{n}{k} \int p^k (1-p)^{n-k} dp}$$

$$P(r|n,k) = \frac{r^k (1-r)^{n-k}}{\int p^k (1-p)^{n-k} dp}$$

The denominator can be evaluated (not straightforward). It is $\frac{k!(n-k)!}{(n+1)!}$, where $k! = 1 * 2 * 3 * 4 * ... * k$, So the final posterior becomes:

$$P(r|n,k) = \frac{(n+1)!}{k!(n-k)!} r^k (1-r)^{(n-k)}$$

Memory Recall: Uniform Prior

The posterior becomes

$$P(r|n,k) = \frac{\binom{n}{k} r^k (1-r)^(n-k)}{\int \binom{n}{k} p^k (1-p)^(n-k) dp}$$

$$P(r|n,k) = \frac{\binom{n}{k} r^k (1-r)^(n-k)}{\binom{n}{k} \int p^k (1-p)^(n-k) dp}$$

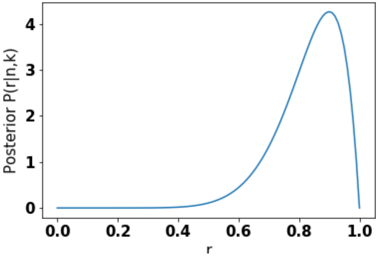
$$P(r|n,k) = \frac{r^k (1-r)^(n-k)}{\int p^k (1-p)^(n-k) dp}$$

The denominator can be evaluated (not straightforward). It is $\frac{k!(n-k)!}{(n+1)!}$, where $k! = 1 * 2 * 3 * 4 * ... * k$, So the final formula becomes:

$$P(r|n,k) = \frac{(n+1)!}{k!(n-k)!} r^k (1-r)^{(n-k)}$$

Memory Recall: Uniform Prior

$$P(r|n,k) = \frac{(n+1)!}{k!(n-k)!} r^k (1-r)^{(n-k)}$$

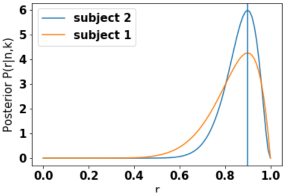


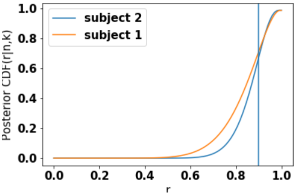
- Now we have a full distribution over the parameter r

Back to the first example:

$$P(r|n,k) = \frac{(n+1)!}{k!(n-k)!} r^k (1-r)^{(n-k)}$$

- Subject 1 responds to 9 questions out of 10 correctly
- Subject 2 responds to 18 questions out of 20 correctly



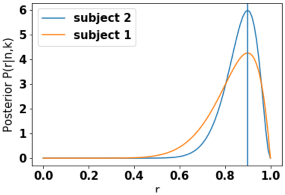


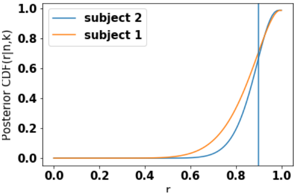
Memory Recall: Uniform Prior

Back to the first example:

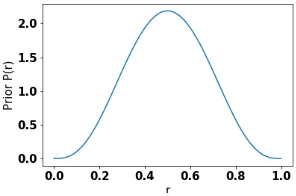
$$P(r|n,k) = \frac{(n+1)!}{k!(n-k)!} r^k (1-r)^{(n-k)}$$

- Subject 1 responds to 9 questions out of 10 correctly
- Subject 2 responds to 18 questions out of 20 correctly
- Now we can estimate the probability of r. For example what is the probability P(r>9|subject 1)? What is P(r>9|subject 2)?





- Likelihood: Same as before
- Prior: let's now consider a non-uniform prior that favors values of r close to $.5$. Although any distribution can be chosen, it is common to choose the Beta Distribution. Let's choose $Beta(r|4,4)$, which has the following shape:



Memory Recall: Non-uniform Prior (continued)

The posterior becomes

$$P(r|n,k) = \frac{\binom{n}{k} r^k (1-r)^(n-k) Beta(r|4,4)}{\int \mathrm{d}p \binom{n}{k} p^k (1-p)^(n-k) Beta(r|4,4)}$$

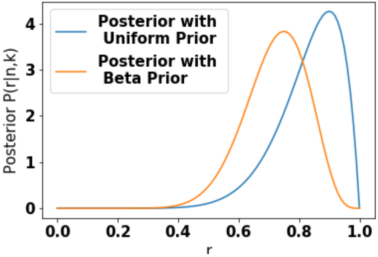
Because we chose the Beta distribution, the posterior can be evaluated (not shown here), and it equal to a new Beta distribution:

$$P(r|n,k) = Beta(r|4+k, 4+n-k)$$

In other words, if the prior is a Beta distribution with parameters α and β, then the posterior is also a Beta with parameters $\alpha + k$ and $\beta + n - k$. The property that the prior and posterior distribution belong to the same family is known as *conjugacy*.

Memory Recall: Non Uniform (Beta) Prior

$$P(r|n,k) = Beta(r|.1 + k, .5 + n - k)$$



- Now we have a full distribution over the parameter r

Graphical Representations of Bayesian Models

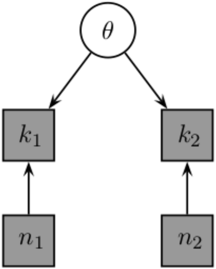
The Bayesian model with non-uniform priors can be represented graphically

$$\theta \sim \text{Beta}(1, 1)$$

$$k \sim \text{Binomial}(\theta, n)$$

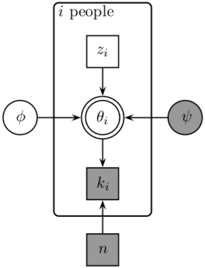
(θ in this figure is the r in the previous slides)

Model for inferring the common rate θ of two binomial processes.



$$k_1 \sim \mathrm{Binomial}(\theta, n_1)$$
$$k_2 \sim \mathrm{Binomial}(\theta, n_2)$$
$$\theta \sim \mathrm{Beta}(1, 1)$$

Hierarchical Bayesian Modeling*



$$
\begin{aligned}
z_i &\sim \text{Bernoulli}(0.5) \\
\phi &\sim \text{Uniform}(0.5, 1) \\
\psi &\leftarrow 0.5 \\
\theta_i &\leftarrow \begin{cases} \phi \text{ if } z_i = 1 \\ \psi \text{ if } z_i = 0 \end{cases} \\
k_i &\sim \text{Binomial}(\theta_i, n)
\end{aligned}
$$

Lee and Wagenmakers. Bayesian Cognitive Modeling: A Practical Course, 2013

Assume there are two different groups of people: one is the guessing group having a probability of 0.5, the other is the knowledge group having a probability greater than 0.5. Whether each person belongs to the first or the second group is an unobserved variable that can take just two values. The goal is to infer to which group each person belongs and the rate of success for the knowledge group.

Generally, posterior distributions cannot be derived analytically.
The following two methods are commonly used:

Numerical Methods for Computing Posteriors

Generally, posterior distributions cannot be derived analytically. The following two methods are commonly used:

Table 7.1 Summary of all approaches to Bayesian parameter estimation that are discussed in this chapter. The table identifies what it is that must be known or obtainable for each approach.

Knowledge required	Analytic Methods (Chapter 6)	Monte Carlo Methods (Section 7.1)	Approximate Bayesian Computation (Section 7.3)
Prior distribution	Assumed	Assumed	Assumed
Likelihood	Computed and known	Computed and known	Cannot be computed but results can be simulated
Posterior distribution	Derived analytically • $p(\theta\|y)$ can be fully evaluated and integrated	Sampled by MCMC • $p(\theta\|y)$ can be evaluated up to a proportionality constant	Sampled by comparing data to candidate simulation results • neither $p(\theta\|y)$ nor $p(y\|\theta)$ need to be computable

- When likelihoods are known: Monte Carlo Sampling `https://chi-feng.github.io/mcmc-demo`
- When likelihoods are unknown: Approximate Bayesian Computing