Bayes Workshop 1

Bayesian Probability and the Likelihood Function

Corson N. Areshenkoff

University of Victoria

March 9, 2016

The Problem

Find a coin on the sidewalk. How do we know if the coin is fair?
In other words, we want to know the probability p that a coin returns heads.

The Problem

How do we get p?

- First, specify a probability model.
  - If a coin flip is random, what is its distribution?
  - If we flip the coin repeatedly, how many heads should we see?
- Then, collect data
  - Flip the coin many times and observe the proportion of heads.
  - The number of heads obtained in N independent flips of a coin is a *binomial* random variable.

Probability Mass Function

- The *probability mass function* (pmf) of a (discrete) random variable gives the probability of observing an outcome x.
- For N coin flips (with probability p of heads), the pmf gives the probability of observing k heads.

Probability Mass Function

How do we get the pmf?
Suppose we flip a fair coin 4 times ($N = 4, p = 0.5$). What is the probability that we get ONE head?

- The head happens with probability $\frac{1}{2}$
- Each tails happens with probability $\frac{1}{2}$, giving $\left(\frac{1}{2}\right)^3$
- ... but there are 4 places the head can appear.

So the probability is

$$P(1; N = 4, p = 0.5) = 4\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^3$$

Probability Mass Function

In general, the binomial pmf is given by

$$P(k; N, p) = \underbrace{\binom{n}{k}}_{\#\text{ of ways}} \overbrace{p^k}^{\text{k heads}} \underbrace{(1-p)^{n-k}}_{\text{n-k tails}}$$

Likelihood

- When we flip a fair coin N times, we talk about the *probability* of observing k heads (N and p are fixed parameters; the *outcome* is unknown).
  - Probability mass function $P(k; N, p)$
- After performing the experiment and *observing* k heads, we can treat p as unknown and ask: what is the probability that a specific value of p would give us k heads? (what is the *likelihood* that p takes a specific value?)
  - Likelihood function $\mathcal{L}(p; N, k)$

Likelihood

A practical example:

- Flip a coin 4 times and observe 3 heads. Which value of p is most likely?
- For $p = 0$, the probability of observing 3 heads is zero, so so the likelihood that $p = 0$ is zero.
- For $p = 1/2$, the probability of observing 3 heads is 0.25, so so the likelihood that $p = 1/2$ is $1/4$.
- For $p = 3/4$, the probability of observing 3 heads is 0.43, so so the likelihood that $p = 3/4$ is 0.43.
  - This is the value of p with the *highest* likelihood. In other words, $p = 3/4$ maximizes the probability of obtaining our data.

The Likelihood Function

- Every statistical model (regression, distribution fitting, etc) produces a *likelihood function*, which gives the probability of obtaining a set of data for different values of the model parameters.
- If D are the data, and Θ are the parameters of the model (say, regression coefficients), then

$$\mathcal{L}(\Theta|D) = P(D|\Theta)$$

- Common practice is to choose the parameters values which maximize the likelihood function (i.e. which maximize the probability of obtaining the observed data).
  - This is *maximum likelihood estimation*, and the resulting parameter estimates are called the *maximum likelihood estimates* (MLE's)

The Likelihood Function

Some examples:

- The MLE for p in our coin flip example is the observed proportion of heads $\hat{p} = k/N$.
- The MLE's for the mean and variance of a normal distribution are

$$\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^{N} x_i}{N}$$
$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N}(\bar{X} - x_i)^2}{N}$$

- The MLE's for a linear regression model are the usual least squares estimates.

The Likelihood Function

- The likelihood function is, arguably, the most important concept in statistical inference.
- **All of the information contained in the data about a model is contained in the likelihood function**.
- In practice, the natural logarithm of the likelihood function (the log-likelihood) is usually easier to work with.

Bayesian Probability

What does probability mean?

- "A coin flip has a %50 chance of landing on heads"

  Frequentist If we flip the coin a large number of times, approximately half of the coin flips will be heads.

  Bayesian If we're asked to call a coin flip, we have no reason to prefer heads or tails.

Frequentists define probability in terms of long-run averages.
Bayesians define probability in terms of degrees of belief.

Bayesian Estimation

Suppose that we pick up a newly minted coin and flip it $N = 4$ times, obtaining $k = 3$ heads. The MLE for the probability p is $k/N = 0.75$. Which is more plausible?

- The finely tuned machines used to mint the coin made a mistake, and somehow minted a highly unbalanced coin.
- The coin is fair, but just happened to return 3 heads.

Bayesian Estimation

1. Decide on a model (a set of distribution that characterize the data)
2. Specify prior distributions for the model parameters
3. Use data to update the prior distributions (w/ likelihood)
4. Obtain posterior

Bayesian Estimation

Or, more rigorously...

1. Define a statistical model with parameters Θ.
2. Model gives a likelihood function $\mathcal{L}(\Theta|D) = p(D|\Theta)$, which summarizes the evidence contained in the data.
3. Specify a prior distribution $p(\Theta)$, which summarizes our prior belief about the parameters.
4. Combine the likelihood and the prior to get a posterior distribution $p(\Theta|D)$.

$$p(\Theta|D) \propto p(D|\Theta)p(\Theta)$$

which summarizes our new belief about the parameter, after taking into account the data.

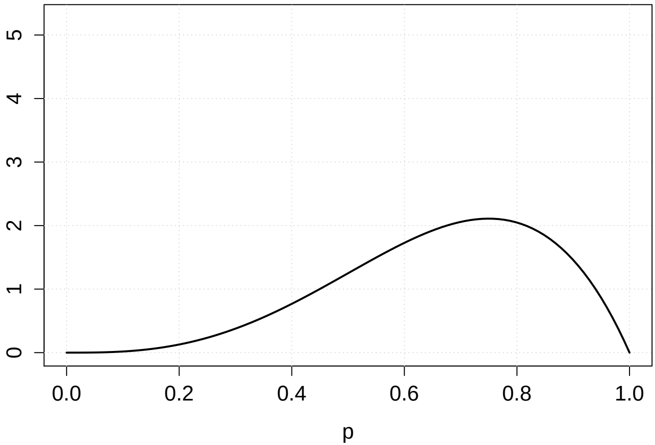

Figure: Binomial likelihood function for $N = 4$ and $k = 3$.

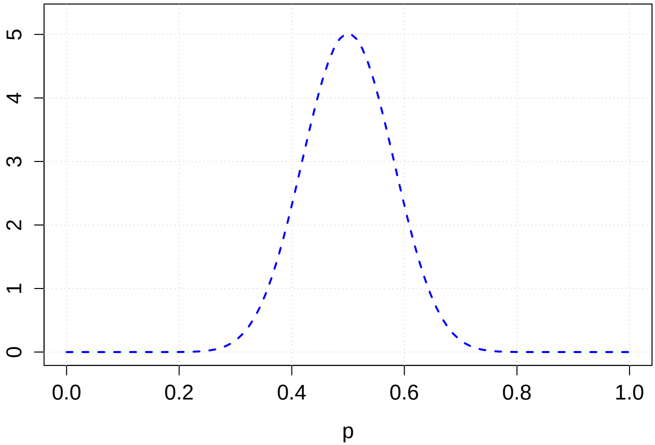

Figure: Beta(20,20) prior distribution for binomial parameter p.

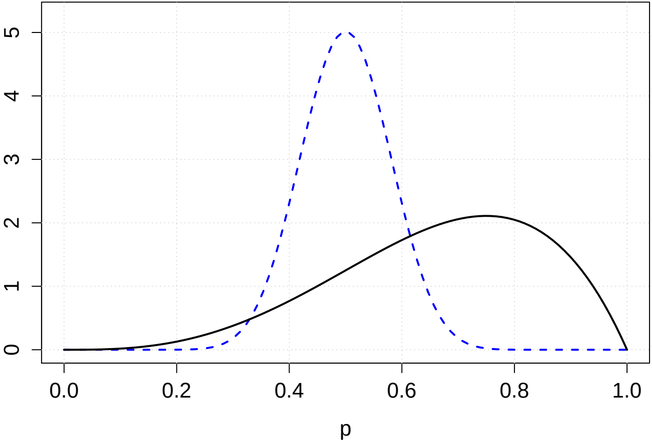

Figure: Likelihood and prior together.

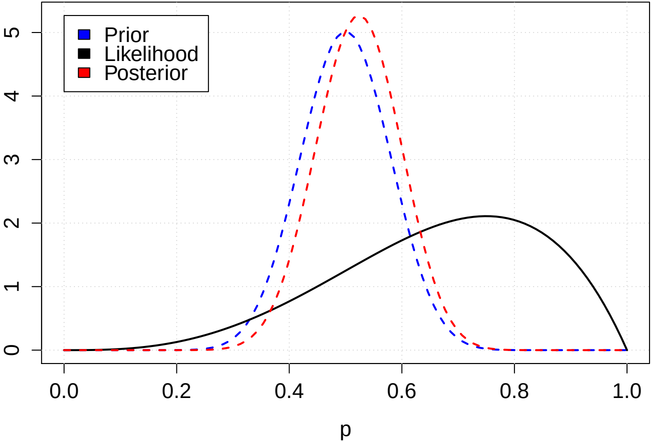

Figure: Relationship between prior, likelihood, and posterior distribution.

In general, computing posterior distributions is difficult or impossible, so we have to approximate them using e.g. MCMC.
There are several software packages to do this (e.g. BUGS, JAGS, Stan).
We use Stan, for reasons that we'll get into next time.

The **data** block specifies the data that will be passed to Stan:

```
data {
    int<lower=0> N; // Number of coin flips
    int k;          // Number of heads
}
```

The **parameters** block specifies the parameters of our model:

```
parameters {
    real<lower=0, upper = 1> p;
}
```

The **model** block specifies the likelihood/priors:

```
model {
    h ~ binomial(N, p);  // Likelihood
    p ~ beta(1,1);       // Prior
}
```

The complete .stan file is then:

```
data {
    int<lower=0> N; // Number of coin flips
    int k;          // Number of heads
}

parameters {
    real<lower=0, upper = 1> p;
}

model {
    h ~ binomial(N, p);  // Likelihood
    p ~ beta(1,1);       // Prior
}
```