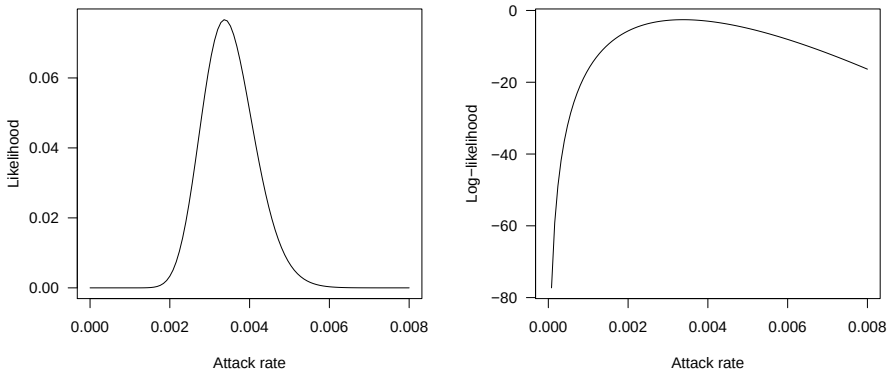
Likelihood, Bayes, and all that, in an epidemiological context

Ben Bolker

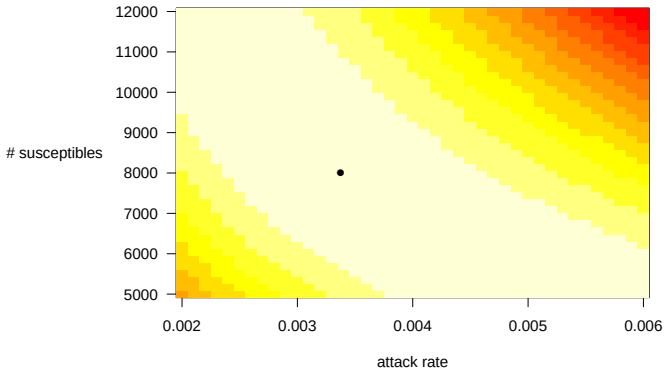
May 27, 2009

1 Likelihood

- sums of squares are fine, but may want a goodness-of-fit metric that is tied to a particular model (e.g. chain binomial) of how an epidemic works
- *likelihood*: **probability of observed data occurring given a particular model (= set of parameters)**: write this as Prob(data|model) (also sometimes stated as "probability given a hypothesis")
- Niamey example: 27 cases in week 2, commune 1. *Suppose* that there are 8000 susceptibles in the population. If the attack rate (*per capita* infection probability) is 0.002, what is the likelihood (probability of 27 cases)? *Answer*: 0.0033 (how did I get this answer (mathematically? in R?))
- the *maximum likelihood estimate* (MLE) is the value of the parameter that makes the data most likely to have occurred. In the particular case of binomial data, we could do a little bit of calculus to show that the MLE in this case is (number of cases)/(number of susceptibles), which is common sense
- the *likelihood curve* shows the likelihood for a range of possible parameter values. We often draw the *log-likelihood curve* (or the *negative log-likelihood curve*) instead, for convenience/historical reasons.



- Likelihood *surfaces*: more than one parameter (in this case number of susceptibles and force of infection)



- intuition: the shape of the likelihood surface (particularly its steepness) tells us about the uncertainty in our parameters. There are a lot of details here that we won't go into (much), because they are handled in a slightly different way by Bayesians
- we could translate this problem into a question about the contact rate β rather than the attack rate by saying $p = 1 - \exp(-\beta I(t-1)/N)$ — in this case our estimate of $\hat{p} = 0.003375$ would translate to $\hat{\beta} = -\log(1 - \hat{p})N/I(t-1) \approx \hat{p}N/I(t-1) = \hat{p} \times (300,000)/22 = 46.02$.
- to solve non-trivial problems (with more than a single data point) we usually assume that the data points are all *independent*, in which case the overall likelihood is the product of the individual likelihoods, or the overall log-likelihood is the sum of the log-likelihoods

- it also turns out that for the special case of a normal distribution, the MLE is equivalent to least-squares fitting. Suppose we have data y_i and are fitting a function to compute the expected values of μ_i (e.g. $\mu_i = a + bx_i$). The normal probability distribution is $C \cdot \exp(-(x_i - \mu_i)^2/(2\sigma^2))$, where C is some ugly stuff $(1/(\sqrt{2\pi}\sigma))$. The logarithm is $\log(C) - (x_i - \mu_i)^2/(2\sigma^2)$. If we sum up the negative log-likelihoods (so we will want to minimize rather than maximize) we get $-N\log(C) + 1/(2\sigma^2)\sum(x_i - \mu_i)^2$. If all we care about is minimizing, we can ignore the first term and the multiplier in the second term — we just have to minimize $\sum(x_i - \mu_i)^2$, which is just the sum of squared errors. (We often ignore *normalization constants* such as $1/\sqrt{2\pi}$, which don't depend on the parameters ...)

1.1 Example

We can read the data straight from the web, if we know where it is:

```
> daturl <- "http://www.umich.edu/~kingaa/EEID/Ecology/commune.measles.csv"
> niamey <- read.csv(url(daturl), header = FALSE)
```

Or we can read it from the working directory:

```
> niamey <- read.csv("commune.measles.csv")
> cases <- niamey[, 1]
```

Suppose that we know the starting number of cases in commune 1, reconstruct the number of susceptibles at the *previous* period in each case:

```
> S0 <- 8000
> n <- length(cases)
> cumcases <- cumsum(cases)
> Susc <- S0 - c(0, cumcases[1:(n - 1)])
```

For a given value of β, the force of infection is βI and the attack rate is $1 - \exp(-\beta I \Delta t) = 1 - \exp(-\beta I)$ (because we have biweekly data, $\Delta t = 1$).

For a given value of β, say $\beta = 50$, we can compute the probability of 27 cases given 8000 susceptibles and 22 cases in the previous generation:

```
> popsize = 3e+05
> beta <- 50
> dbinom(cases[2], size = Susc[1], prob = 1 - exp(-beta * cases[1]/popsize))
```

```
[1] 6.579164e-06
```

Or to calculate the log-likelihood:

```
> dbinom(cases[2], size = Susc[1], prob = 1 - exp(-beta * cases[1]/popsize),
    log = TRUE)
```

```
[1] -11.93160
```

There are a few ways we could figure out the overall likelihood, or log-likelihood, for a given value of β. Read them over and try **one** of them.

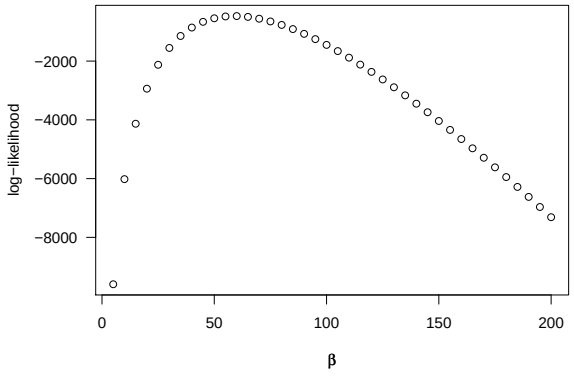
1. Define a starting value for `likelihood` of 1. Then use a `for` loop stepping from 2 to n. At each step, compute the likelihood `curr_lik` using the current cases, previous cases, and previous susceptibles as shown above, and update the likelihood: `likelihood <- likelihood * curr_lik`

2. Do the same for log-likelihoods, starting from 0 rather than 1 and adding the log-likelihood instead of multiplying likelihoods.

3. Set up a vector of cases from time 2 to n by dropping the first case: `cases[-1]` (or `cases[2:n]`). Then set up the vector of cases from time 1 to $n-1$ (`cases[1:(n-1)]` or `cases[-n]`) and the vector of susceptibles. Now you can compute $1 - \exp(-\beta I(t-1))$ (for a given value of β) in a single, vectorized statement, and you can feed this expected-attack-rate vector and the matching numbers of susceptibles and cases into `dbinom` and get a vector of attack rates: `prod` takes the product of a vector

4. as previous item, but with log-likelihoods (add the `log=TRUE` argument to the `dbinom` call) and use `sum` instead of `prod`

If you have time: write another `for` loop, over different values of β (try values ranging from 5 to 200 in steps of 5), to compute the likelihood curve for β. The result should look like this:



2 Bayes

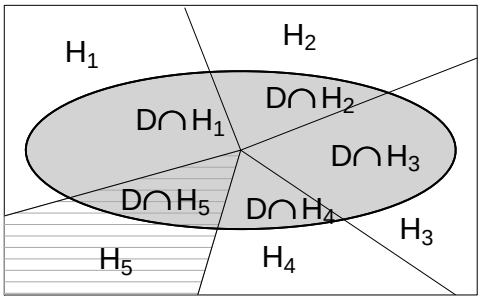
2.1 Bayes' rule

- *Bayes' rule*: just a rule about how to figure out Prob($H|D$) from Prob($D|H$) (fairly easy to derive):

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} \tag{1}$$

or

$$P(H_i|D) = \frac{P(D|H_i)P(H_i)}{\sum_j P(H_j)P(D|H_j)} \tag{2}$$



- (false positive/medical testing/forensic example, if time permits)

2.2 Bayesian inference

- interpret $P(H|D)$ as *the probability of the "hypothesis" given data*, where "hypothesis" can mean (as in the discussion of likelihood above) a model, or for $P(H_i|D)$, H_i is a particular value of the parameters; this is called the **posterior probability**, the distribution is the **posterior distribution**.
- then $P(D)$ is the likelihood ...
- but what the hell is $P(H_i)$? The **prior**.
- the denominator, $\sum P(H_j)P(D|H_j)$ or (for continuously distributed parameters $\int P(p)P(D|p)\,dp$) is $P(D)$, the probability of having gotten the data *somehow*
- if all of the $P(H_i)$ are the same then they all cancel out of the Bayes' rule formula (and we get $P(H_i|D) = P(D|H_i)/\sum_j P(D|H_j)$, sometimes called the *scaled likelihood*) — then the shape of the posterior distributions is the same as the shape of the likelihood curve (and the maximum (*mode*) of the posterior distribution is the same as the MLE)

Bayesians usually summarize the results of an analysis via the *posterior means* of the parameters (sometimes the *posterior modes*) and the *quantiles* or *credible intervals* of the *marginal distributions* (give details if time). Computing the marginal posterior distributions is a big pain (integrals!) but can be avoided by using the techniques Aaron will discuss.

2.2.1 Pros and cons

Three reasons to be Bayesian:

1. *philosophical*: using a Bayesian framework gives you the ability to make inferences about what you *really* (arguably) want to know, the probability of a given parameter or model, rather than jumping through the semantic hoops required by frequentist inference

2. *using prior information*: in sparse-data cases (conservation biology, wildlife disease) we can introduce extra information that we already know, in a very natural way, through the prior distribution (McCarthy (2007) shows lots of nice examples)

3. *pragmatic*: a lot of problems that are very hard to solve in classical ways become easier in a Bayesian framework. For example: "mixed" models (involving random effects); models with process and measurement error, or unobserved states; etc.

Three reasons *not* to be Bayesian:

1. *philosophical*: the Bayesian framework requires you to specify your prior distribution, which most of the time is a subjective statement of your personal belief about what the parameter might be. Some researchers hate this subjectivity (Dennis, 1996). You can try to make the priors *weak*, or *uninformative* (*flat* is another synonym), but this is hard for various reasons.

2. *pragmatic ("too hard")*: using Bayesian approaches requires more technical overhead than classical approaches (partly, but only partly, because most canned statistical software is written to apply classical approaches). Very hard problems are easier, but moderately hard problems are harder . . . the main problem is computing integrals (Aaron's lecture will talk about how to do *avoid* doing the integrals)

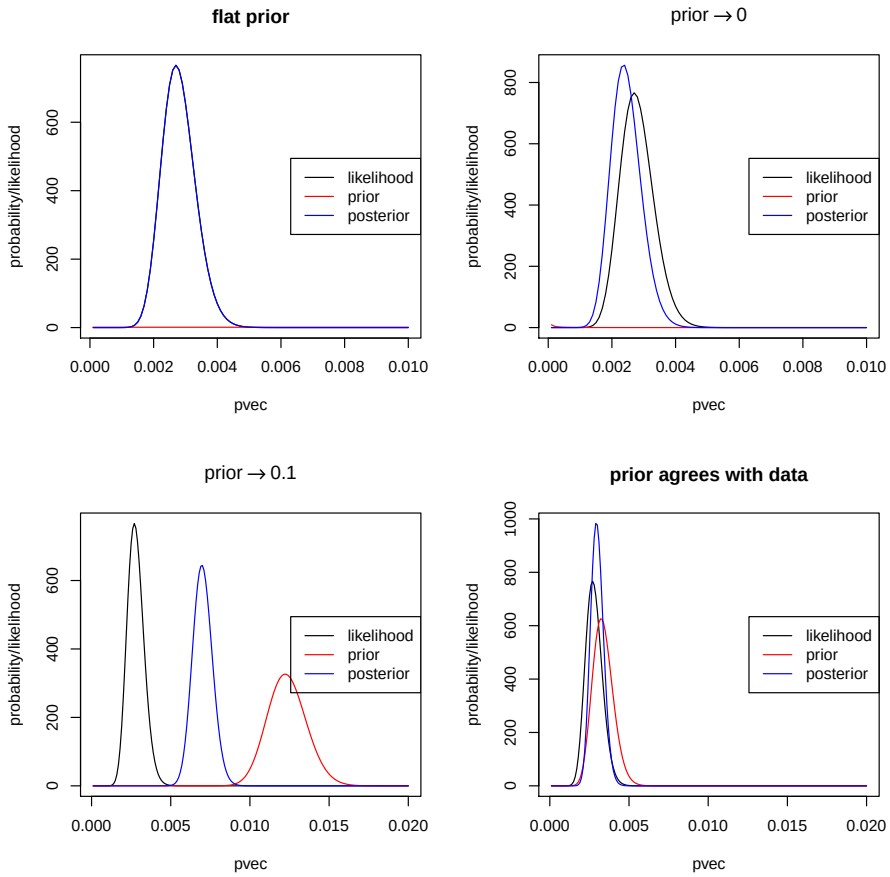
3. *pragmatic ("too easy")*: in practice, one can almost always get an answer to a problem (even very hard ones) using modern Bayesian methods, but sometimes the answers make no sense — where classical methods would just fail (!!)

Many statistical practitioners switch back and forth among approaches depending on what works best in a particular case. Fewer and fewer *real* statisticians are rabid about the distinction, although some have strong preferences . . . for an amusing recent entry into the debate, see Gelman (2008).

2.2.2 Priors and conjugate priors

For a single-parameter problem, we can get the general shape of the posterior (ignoring the denominator) easily. Let's go back to the attack rate problem we started with (for 27 cases out of 10,000 susceptibles, what can we say about the attack rate?) Say we use a flat prior, where the prior probability of any attack rate is equal: then (as stated above) the prior terms cancel out, and the posterior is equal to the scaled likelihood.

Let's try varying the prior. Location of the peak tells us something about our estimate, width of the peak tells us something about our certainty.



Crome et al. (1996) give a nice example of contrasting the effects of different priors in a conservation context (effects of logging on bird communities).

2.3 Conjugate priors

Conjugate priors are special forms of the priors such that the combination of the prior and the likelihood (i.e. the outcome of Bayes' rule) ends up having the same distribution as the prior. For example, if you start with a normally distributed prior mean, and your data are normal, then the posterior distribution of the

mean is also normal — but with different (updated) parameters, in a sensible way

$$\mu_{\text{post}} = (\mu_{\text{dat}}/\sigma^2_{\text{dat}} + \mu_{\text{prior}}/\sigma^2_{\text{prior}})/(1/\sigma^2_{\text{dat}} + 1/\sigma^2_{\text{prior}}) \quad (3)$$

Data	**Prior**	**Meaning**
Binomial (p)	Beta	$a \to a + k$=# successes, $b \to b + (N - k)$=# failures
Poisson (λ)	Gamma	mean (shape $\cdot$ scale) = prior mean intensity, shape = # counts
Normal (μ)	Normal	Mean and standard dev of prior obs.
Normal (σ^2)	Inverse-gamma	

Also see `http://en.wikipedia.org/wiki/Conjugate_prior#Table_of_conjugate_distributions`

Conjugate priors are useful for illustrating/understanding the effects of priors (see above); they are most useful as *components* of more complicated Bayesian solutions.

2.3.1 Example

Play around with conjugate priors for the binomial distribution (Beta). Use the `curve` function: for example, consider binomial data with 5 successes and 10 failures (total sample $N = 15$, # successes $k = 5$, failures $N - k = 10$) and a prior distribution of 10 successes and 5 failures ($a = 10$, $b = 5$). The posterior is $a = 15$, $b = 15$ (R uses `shape1` and `shape2` to denote these parameters.)

```
> curve(dbeta(x,shape1=5,shape2=10),col=1,from=0,to=1,
      ylim=c(0,5)) ## prior
> curve(dbinom(5,prob=x,size=15),col=2,add=TRUE) ## likelihood
> curve(dbeta(x,shape1=15,shape2=15),col=4,add=TRUE) ## posterior
> legend("topleft",c("prior","likelihood","posterior"),
       col=c(1,2,4),lty=1)
```

Experiment with:

- weak priors ($a = 1$, $b = 1$)
- strong priors that agree with the data (e.g. $a = 5$, $b = 10$ (mean prior prob=2/3), $k = 5$, $N = 15$)
- strong priors that disagree with the data (e.g. $a = 10$, $b = 5$ (mean prior prob=1/3), $k = 5$, $N = 15$)
- data much stronger than prior (e.g. as above two examples but use $k = 25$, $N = 75$)

References

Crome, F. H. J., M. R. Thomas, and L. A. Moore. 1996. A novel Bayesian approach to assessing impacts of rain forest logging. Ecological Applications **6**:1104–1123.

Dennis, B. 1996. Discussion: Should ecologists become Bayesians? Ecological Applications **6**:1095–1103.

Gelman, A. 2008. Objections to Bayesian statistics. Bayesian Analysis **3**:445–450. URL http://ba.stat.cmu.edu/journal/2008/vol03/issue03/gelman.pdf.

McCarthy, M. 2007. Bayesian methods for ecology. Cambridge University Press, Cambridge, England.