an introduction to Bayesian analysis for epidemiologists

Charles DiMaggio

Departments of Anesthesiology and Epidemiology
College of Physicians and Surgeons
Columbia University
New York, NY 10032

cjd11@columbia.edu

P9489
applications of epidemiologic methods II
Spring 2014

Outline

1. The Bayesian Way
   - Why Bayes?
     - Bayes vs. Classical
     - The Benefits of Bayes
   - Bayes Theorem

2. Conjugate Single-Parameter Problems
   - Binomial Examples: Race and Promotion, Perchlorate and Thyroid Tumors
   - Poisson Example: Airline Crashes
   - Single-Parameter Normal Model
   - (More) Conjugate Examples: Drug Response, London Bombings, Heart Transplant Mortality

acknowledgments

- David Spiegelhalter
- Nicky Best
- Andrew Gelman
- Bendix Carstensen
- Lyle Gurrin
- Jim Albert
- Shane Jensen
- Statistical Horizons

A Bayesian is one who, vaguely expecting to see a horse and catching a glimpse of a donkey, strongly concludes he has seen a mule. (Senn, 1997)

The Bayesian approach is "the explicit use of external evidence in the design, monitoring, analysis, interpretation and reporting of a (scientific investigation)" (Spiegelhalter, 2004)

"you shall know them by their posteriors"

a natural and coherent approach

- theoretically correct, and now practical and doable
- advantages
    - It is flexible and can adapt to complex situations
    - It is efficient, using all available information
    - It is intuitively informative, providing *relevant* probability summaries in a way that is consistent with how we think and learn.
    - It captures additional uncertainty in predictions by allowing parameter estimates to vary.

Bayes in a nutshell

- parameters (θ) are allowed to vary randomly
  - direct probability statements about θ
- combine what you know with what you observe to update your knowledge
  - $prior + data = update$
  - $Pr[\theta|y] \propto Pr[y|\theta]Pr[\theta]$
- additional variation in predictions
  - posterior predictive distributions

there is no free lunch

specifying a prior distribution and combining it with the data likelihood will complicate our lives

statistics

1. Estimating unknown parameters (What is the mean value for some medical test in a population?)
2. Accounting for variability in estimated parameters (How much does that value vary around the mean?)
3. Testing hypotheses (Is the value for the medical test different in treated vs. untreated populations)
4. Making predictions (What would we expect the mean value to be in a new sample of patients?)

classical statistics

- parameters (means, standard deviations, regression coefficients) fixed but unknown
- the only thing that varies is the sample
    - what is a 95% CI?
    - if take 100 samples, 95 of them contain the true value
- ...so take a lot of samples
    - but not really...
- rely on asymptotics and CLT

the data likelihood

- classical emphasis on data sample → MLE
    - e.g. batting average for baseball team
    - overall probability multiply all the batting averages: $p(y|\theta = \Pi(y_i|\theta)$
    - called the likelihood function (joint probability of all observations)
- MLE - parameters that make data you observed as likely as possible
    - take derivative set it equal to zero
- intuitive results for standard distributions
- e.g. normal, $\Sigma Y_i/n$ for μ and $(y_i - \hat{y})^2/n$ for variance
- more difficult for non-standard distributions

Bayesian statistics

- parameters vary randomly (normal, binomial, Poisson)
- in *addition* to characterizing the likelihood of the data, added task characterizing parameter probability distributions, called prior distributions
- combine data likelihood ($p(y|\theta)$) with prior expectation($p(\theta)$) to update inference on parameters called posterior ($p(\theta|y)$)
- Bayes rule: $p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$, or

$$\mathbf{p}(\theta|\mathbf{y}) \propto \mathbf{p}(\mathbf{y}|\theta)\mathbf{p}(\theta) \tag{1}$$

  - ($p(y)$ drops out as normalizing term)

Bayesian weighting

- posterior is a weighted combination of the likelihood and the prior
- lots of data, likelihood "swamps" the prior
- small data, prior influential
- e.g. normal prior normal likelihood
- prior influences prior through τ
- likelihood influences through n

$$\mu \,|\, \mathbf{y} \sim \mathrm{Normal}\left(\frac{\frac{n}{\sigma^2}\,\overline{y} + \frac{1}{\tau^2}\,\alpha}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} \right)$$

Figure : Posterior Distribution for Normal Mean ($\alpha =$ prior mean, μ_0)

it's all about the priors

- alot of skepticism about priors (choice, assumptions)
- ideally, based on real prior information
- else, "non" or "minimally" informative priors
  - won't exert too much influence
  - more on that soon

so, really, why Bayes?

is it worth the hassle?

- direct probability statements = fewer logical gymnastics
- less reliance on asymptotics = more inference from small data
- principled and logical approach to scientific learning
    - when we actually do know something, we don't have to ignore or throw it out
- more conservative predictions
- posterior predictive distribution: $p(y * | Y) = p(y * |\theta)p(\theta|y)$
    - includes variability in data *and* variability in the parameter
    - classical regression prediction uses model (regression line)
    - Bayesian regression prediction includes variability in the regression line itself

Bayes theorem for discrete outcomes

Bayes Theorem says that if we know $Pr[A|B]$ we can get at $Pr[B|A]$.

$$\begin{aligned} Pr[A \cap B] &= Pr[B \cap A] \\ Pr[A \cap B] &= Pr[A|B]Pr[B] \\ Pr[B \cap A] &= Pr[B|A]Pr[A] \\ Pr[A|B]Pr[B] &= Pr[B|A]Pr[A] \\ Pr[A|B] &= \frac{Pr[B|A]Pr[A]}{Pr[B]} \\ Pr[B|A] &= \frac{Pr[A|B]Pr[B]}{Pr[A]} \end{aligned}$$

Bayes theorem for parameter distributions

$$Pr[\theta|y] = \frac{Pr[y|\theta]Pr[\theta]}{\int dBPr[y|\theta]Pr[\theta]}$$

integration in denominator can be a bear, so

$$Pr[\theta|y] \propto Pr[y|\theta]Pr[\theta]$$

remove normalizing constant in denominator (makes it sum to 1)
form the same (only size changes)

a first example: student sleep habits

Jim Albert

- what proportion of students get 8 or more hours sleep?
- intuition says somewhere between 0 and 50%, but close to about 30% (the prior)
- class survey says $\frac{11}{27} = .47$ (the likelihood)

how can we combine using Bayes rule to update our prior?

the likelihood

- binomially distributed, $\theta^k * (1-\theta)^{n-k}$, where,
- θ is the probability of sleeping more than 8 hours
- k is the number of students who said they slept more than 8 hours
- n is the number of students surveyed.

the prior

trickier

- "discrete" approach.
- list plausible values
- weight them by how probable we think they are
- convert the weights to a probability distribution that sums to one by dividing through by the sum of the weights.

creating the discrete prior

- plausible values for proportion of heavy sleepers (theta)
  - 0.05, 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.86, 0.95
- weights for those plausible values
  - 0, 1, 2, 3, 4, 2, 1, 0, 0, 0
- convert to probabilities
  - 0.00, 0.08, 0.15, 0.23, 0.31, 0.15, 0.08, 0.00, 0.00, 0.00

calculating the posterior

- multiply each value of the posterior by the likelihood of that value (involves logs etc...)
- use *pdisc()* from "LearnBayes"

```
library(LearnBayes)
data<-c(11, 16)  # number successes and failures
theta<-seq(0.05, 0.95, by = 0.1)
weights<-c(0, 1, 2, 3, 4, 2, 1, 0, 0,0)
prior<-weights/sum(weights)
plot(theta, prior, type="h", ylab="Prior Probability")
post<-pdisc(theta, prior, data)
round(cbind(theta, prior, post), 2)
par(mfrow=c(3,1))
plot(c(0,1),c(11/(11+16),16/(11+16)),  type="h", ylim=c(0,.7),
        lwd=5, main="data" )
plot(theta, prior, type="h", ylab="Prior Probability",
        ylim=c(0,.7),  lwd=5, main="prior")
plot(theta, post, type="h", ylab="Posterior Probability",
        ylim=c(0,.7),  lwd=5, main="posterior")
```

Single Parameter Binomial Example: Race and Promotion at a State Agency

- 26/48 Black vs. 206/259 White applicants passed test
- What is the probability of a Black applicant passing the test compared to a White applicant?
- What is the probability of a future Black applicant passing the test?
- binomial likelihood for the theta (like in sleep example)
- use a more realistic continuously distributed prior (rather than discrete) that can combine with a binomial likelihood
    - if we are being "agnostic" about it, need a prior that does not influence the data

Beta prior for binomial likelihood

- Beta distributions have properties that make them easy to combine with binomial distributions

$$\begin{aligned} &\sim Beta(\alpha, \beta) \\ \mu =& \frac{\alpha}{\alpha + \beta} \\ \sigma^2 =& \frac{\alpha\beta}{(\alpha\beta)^2(\alpha + \beta) + 1} \end{aligned}$$

- Beta(1,1) is flat on range 0 to 1

posterior distribution for beta-binomial

- $Binomial(y, n)$ combines with $Beta[\alpha, \beta]$ to produce $Beta[y + \alpha, n - y + \beta]$
- Posterior distribution for Black applicants is Beta(27,23), for Whites Beta(207,54)
- Simulate many times from these distributions, draw inferences, compare

code for race and promotion example

```
    # data
y.black <- 26; n.black <- 48
y.white <- 206; n.white <- 259
   # likelihood for black applicants
?rbinom
likelihood.black<-rbinom(10000, 48 ,(26/48))
plot(density(likelihood.black))
   #  Beta(1,1) prior
?rbeta
prior<- rbeta(10000,1,1)
plot(density(prior))
  # posterior from updated uniform Beta(1,1) by adding 1 to number of successes and 1 to number of failures
  # 10000 simulation for blacks and whites
theta.black <- rbeta(10000,y.black+1,n.black-y.black+1)
theta.white <- rbeta(10000,y.white+1,n.white-y.white+1)
   # plot densities
old.par<-par()
par(mfrow=c(2,1))
plot(density(theta.black), xlim=c(0,1), main="probability of blacks passing")
plot(density(theta.white),  col="red", xlim=c(0,1), main="probability of whites passing")
   # plot histograms
mintheta <- min(theta.black,theta.white)
maxtheta <- max(theta.black,theta.white)
hist(theta.black,col="gray",xlim=c(mintheta,maxtheta))
hist(theta.white,col="gray",xlim=c(mintheta,maxtheta))
par(old.par)
  # compute posterior probability blacks scoring less than whites
  # proportion times in the 10000 simulations blacks scored less than whites
(prob <- sum(theta.black <= theta.white)/10000)
  # essentially 100%
```

prediction

future pool of 100 Black applicants

- sample a large number (10000) of θ's from the Beta(27,23) posterior
- for each of those values of θ_i, sample a single y* from Binomial(100, θ_i)
- (classical approach might be a parametric bootstrap, sampling y* from Binomial(100, θ) from single value of θ)

about that Beta prior...

- Beta(1,1) like adding two observations to the data, one success and one failure
- subtly ”pulls” posterior estimate to center
- this kind of ”compromise” between data and prior is a characteristic of Bayesian analyses
- allows us to draw inferences even when little data
  - e.g. say zero outcomes, Binomial(0,25)
  - classical CI of 1.96 plus minus $\sqrt{pq/n}$ collapses to infinity
  - combining Beta(1,1) with Binomial(0,25) mostly near zero

about "non-informative" priors

- all priors carry information and assumptions
  - even if the assumption is that you know nothing
- for small data sets (5 or 6 observations), prior will have an influence
- flat prior on Binomial may make sense if constrained range 0,1
  - flat prior on normal, $(-\infty, +\infty)$ means we are so unsure we believe it ranges across infinite values (?)
- 1950'a Sir Harold Jeffreys described an approach or general scheme for selecting minimally informative priors
  - Jeffreys prior, set prior equal to the square root of the expected Fisher information
  - Jeffreys prior for binomial data is Beta(0.5,0.5) for θ

Single Parameter Binomial Example: Perchlorate and Thyroid Tumors

David Dunson

- Percholorate - ground water contaminant associated with thyroid tumors
- sparse data - 2/30 exposed rats develop tumors vs. 0/30 control
- Classical approach - Fisher exact test

```
(rat.dat<-matrix(c(2,0,28,30), nrow = 2))
fisher.test(rat.dat)
```

Bayesian approach

```
  # data
y.perchlorate <- 2; n.perchlorate <- 30
y.control <- 0; n.control<- 30
  # update Beta(1,1) prior for exposed and unexposed
theta.perchlorate <- rbeta(10000,y.perchlorate+1,n.perchlorate-y.perchlorate+1)
theta.control <- rbeta(10000,y.control+1,n.control-y.control+1)
  # graphically compare exposed and unexposed
par(mfrow=c(2,1))
plot(density(theta.perchlorate), xlim=c(0,1), main="probability of tumor in exposed rats")
plot(density(theta.control),  col="red", xlim=c(0,1), main="probability of tumor in control rats")
  # probability that exposed have more tumors than unexposed
sum(theta.perchlorate >= theta.control)/10000

theta.diff<-theta.perchlorate-theta.control
  # 95% credible interval
quantile(theta.diff, probs=c(0.05,0.95))
  # plot differences
plot(density(theta.diff))
```

- Beta(1,1) prior exerts considerable influence
- 87% simulations, perchlorate exposed rats developed more thyroid tumors
- note - can now calculate probability interval for difference (most of probability to R of zero)

what about a more informative prior?

- likely some prior evidence (else why are we doing the study?)
- prior studies suggest Beta(0.11, 2.6) reasonable

```
theta.perchlorate <- rbeta(10000,y.perchlorate+.11,
      n.perchlorate-y.perchlorate+2.6)
theta.control <- rbeta(10000,y.control+.11,
      n.control-y.control+2.6)
theta.diff<-theta.perchlorate-theta.control
quantile(theta.diff, probs=c(0.05,0.95))
```

Airline Crashes

Poisson-Gamma Model

- airline crash data 1976 to 1985:
  $y = (24, 25, 31, 31, 22, 21, 26, 20, 16, 22)$
- Poisson data likelihood
  - no underlying number of "trials" as in Binomial
- assume 10 realizations of Poisson process with same underlying rate (will explore this more later)

$$p(y_i|\theta) = \frac{\theta^{y_i} e^{-\theta}}{y_i!}$$

Gamma prior for Poisson Likelihood

- Gamma is analytically convenient prior for Poisson data

$$\sim\Gamma(\alpha,\beta),$$
$$\mu = \frac{\alpha}{\beta}, \sigma^2 = \frac{\alpha}{\beta^2}$$

  - "looks" like Poisson $p(y_i|\theta) = \theta^{\alpha-1}e^{-beta*\theta}$
- $Poisson(\frac{\sum y_i}{n})$ likelihood * $Gamma(\alpha,\beta)$ prior →

$$\mathbf{Gamma}(\mathbf{y}+\alpha, \mathbf{n}+\beta)$$

- α as number of outcomes prior is "worth", β as number of "units"
- Jeffreys prior for Poisson-gamma is improper Gamma(0.5,0), use Gamma(0.5,0.0001)

code for airline crash example

```
   # data

years <- c(1976,1977,1978,1979,1980,1981,1982,1983,1984,1985)
crashes <- c(24,25,31,31,22,21,26,20,16,22)
numyears <- length(years)
sumcrashes <- sum(crashes)

  # posterior from updated noninformative (Jeffrey) prior
  theta <- rgamma(10000,shape=(sumcrashes+0.5),rate=(numyears+0.0001))
plot(density(theta))

   # posterior predictive distribution for crashes in next year

y.star <- rep(NA,10000) # vector to hold simulations

   # sample one observation from the posterior distribution
for (i in 1:10000){
y.star[i] <- rpois(1,theta[i])
}

  # plot histograms for data, posterior and posterior predictive on same scale
par(mfrow=c(3,1))
hist(crashes,col="gray",xlim=c(0,50),breaks=10)
hist(theta,col="gray",xlim=c(0,50))
hist(y.star,col="gray",xlim=c(0,50))

posterior distribution vs. the posterio predictive distribution
mean(theta)
quantile(theta, probs=c(0.05,0.95))
mean(y.star)
quantile(y.star, probs=c(0.05,0.95))

sum(theta>30)/10000
sum(y.star>30)/10000
```

prediction in the airline crash example

- simulating from the posterior predictive distribution
- includes variation in parameter
    - in addition to the usual variation in the data
- 95% posterior predictive interval wider
- 10% probability crashes in following year will exceed 30

single parameter normal model

- characterized by two parameters, so odd to assume know one but not the other
  - but, if we did...
- analytically tractable prior for normal data likelihood is also normal
  - why not Poisson for Poisson, or Binomial for Binomial?
    - analytically intractable (don't combine nicely...)
    - gamma-gamma *will* combine, by heteroskedasticity (mean linked to variance...)
- confidence in prior (small τ) up weights the prior, accumulating evidence (large *n*) up weights the data
- as prior variance $\rightarrow$ *infty*, results $\rightarrow$ MLE estimates

$$\mu \,|\, \mathbf{y} \sim \text{Normal}\left(\frac{\frac{n}{\sigma^2}\,\bar{y} + \frac{1}{\tau^2}\,\alpha}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} \right)$$

Figure : Posterior Distribution for Normal Mean ($\alpha =$ prior mean, μ_0)

about conjugacy

- analyses so far have been conjugate
- prior and likelihood from same "family" of distributions
- analytically convenient, introduce concepts, but restrictive
- will soon need other approaches, e.g. MCMC

Likelihood	Prior	Posterior
Normal	Normal	Normal
Binomial	Beta	Beta
Poisson	Gamma	Gamma

beta-binomial model

- binomial likelihood $Pr(y|\theta) = \binom{n}{k} p^k q^{n-k}$
  - "n choose k" $\frac{n!}{k!(n-k)!}$
- minimally informative prior $\sim Beta(1,1)$
- posterior

$$
\begin{aligned}
Pr(\theta|y) = Pr(\theta|k,n) &\propto (\theta^k)(1-\theta)^{n-k} * 1 \\
&\sim Beta(1+k, 1+n-k)
\end{aligned}
$$

drug response example

- believe somewhere between 0.2 and 0.6 of patients will respond
- $\mu = 0.4$, $\sigma^2 = 0.1$
- corresponds to a $Beta(9.2, 13.8)$
- what is the probability that 15/20 patients will respond?
- this is pure simulation or Monte Carlo (no data likelihood yet)
  - will simulate from beta, and plug results into binomial, plot and tally results

code for simple simulation for drug response

```
N=1000
theta<-rbeta(N,9.2,13.8)
x<-rbinom(N,20, theta)
y<-0
accept<-ifelse(x>14.5, y+1, y+0)
plot(density(accept))
(prob<-sum(accept)/N)
```

sneak peek at BUGS:

```
   #binomial monte carlo
Model{
y~dbin(theta , 20)          #sampling dstn
theta ~ dbeta (9.2, 13.8)  #parameter from sampling dstn
p.crit <- step(y-14.5)      # indicator, 1 if y>=15, 0 else
```

add data to drug response example

- suppose, rather than trying to guess, enroll and treat 20 patients, and 15 respond
- now, instead of pure simulation, we are updating the prior with the likelihood
- Beta(9.2, 13.8) → Beta (9.2+15, 13.8+20-15) = Beta (24.2, 18.8)
    - μ =24.2/24.2+18.8 = 0.56 (closed conjugate, no need for simulation)
- how likely to see 25 successes in additional 40 patients?
- will simulate from posterior predictive distribution

code for drug response prediction

```
theta.drug<-rbeta(10000, 24.2, 18.8)
mean(theta.drug) # check mean close to analytic

x<-rbinom(N,40, theta.drug)
y<-0
accept<-ifelse(x>24.5, y+1, y+0)
prob<-sum(accept)/N
prob
```

see website for how to run this in JAGS...

where did that Beta prior come from?

- we said Beta(9.2, 13.8) consistent with a mean response 0.4 and sd 0.1. Why?
- general approach described in this informative stackexchange response
- can use this little function:

```
estBetaParams <- function(mu, var) {
alpha <- ((1 - mu) / var - 1 / mu) * mu ^ 2
beta <- alpha * (1 / mu - 1)
return(params = list(alpha = alpha, beta = beta))
```

- or the *betaselect()* function in "LearnBayes"

the Gamma-Poisson model

- Poisson likelihood for count data $Pr[k] = e^{-\lambda} * \lambda^k / k!$
- conjugate Gamma(a,b) prior
  - $\mu = \frac{a}{b}$ and $\sigma^2 = \frac{a}{b^2}$
- Gamma posterior $Gamma(a + n\bar{x}, b + n)$
  - compromise between the prior mean ($\frac{a}{b}$) and the MLE of the mean from the likelihood ($\bar{x}$)

London bombings during WWII

- count bomb hits in $36km^2$ area S. London partitioned into $0.25km^2$ grid
- 537 events ($\Sigma x_i * n_i = 537$ total hits), over 576 observations ($\Sigma n_i = 576$ areas)
-

Hits (x)	0	1	2	3	4	7
Areas (n)	229	211	93	35	7	1

conjugate analysis of London bombing data

- Poisson data likelihood, with Jeffreys prior (improper) Gamma (0.5, 0)
- $$\begin{aligned} p(\theta|y) = \Gamma(537+0.5, 576+0) = \Gamma(537.5, 576) \\ \mu = 537.5/576 = 0.933 \\ \sigma^2 = 537.5/576^2 = 0.0016 \end{aligned}$$
- vs. data $\mu = 537/576 = 0.932$
- In general, as the sample size increases, the posterior mean approaches the MLE mean, and the posterior s.d. approaches the MLE s.d.

heart transplant mortality example

Jim Albert

- interested 30-d heart transplant mortality
- SMR is $\lambda = \frac{y}{e}$
- unstable when sparse data
- use Bayesian approach to incorporate evidence from comparable hospitals
- $Gamma(\alpha, \beta)$ prior
  - α sum 30-d deaths 10 nearby hospitals, β sum procedures
- Hospital A, 1 death 66 surgeries; Hospital B, 4 deaths 1767 surgeries. comparison 16 deaths 15,174 procedures

results of heart transplant mortality analysis

- unadjusted MLE estimates 1.5% (95% CI 0.08%, 9.3%) hospital A, and 0.2% (95% CI 0.07%, 0.6%) hospital B
- Bayesian "smoothed" estimates little or no difference
- posterior hospital A closer to prior (more influence of prior)

code for heart transplant mortality

```
y_A<-1
n_A<-66
prop.test(y_A, n_A)

y_B<-4
n_B<-1767
prop.test(y_B, n_B)

y_T<-16
n_T<-15174
prop.test(y_T, n_T)

 # conjugate analysis

lambda_A<-rgamma(1000, shape=y_T+y_A, rate=n_T+n_A)
lambda_B<-rgamma(1000, shape=y_T+y_B, rate=n_T+n_B)

summary(lambda_A)
summary(lambda_B)
relevant
t.test(lambda_A, lambda_B)

par(mfrow = c(2, 1))
plot(density(lambda_A), main="HOSPITAL A", xlab="lambda_A", lwd=3)
curve(dgamma(x, shape = y_T, rate = n_T), add=TRUE)
legend("topright",legend=c("prior","posterior"),lwd=c(1,3))
plot(density(lambda_B), main="HOSPITAL B", xlab="lambda_B", lwd=3)
curve(dgamma(x, shape = y_T, rate = n_T), add=TRUE)
legend("topright",legend=c("prior","posterior"),lwd=c(1,3))
```

conclusions about Bayesian analysis

- philosophically coherent
- provides intuitive and directly relevant results
- uses all available information
- captures additional uncertainty in predictions
- deserves greater application in epidemiological analyses