Introduction to Bayesian Methods

Rob Scharpf

Biostatistics Division, Department of Oncology
Johns Hopkins School of Medicine

January 20, 2015

Contact information

Instructors: Rob Scharpf (term 1) and Gary Rosner (term 2)
office hours: by appointment
email: rscharpf@jhu.edu
TA: Lei Huang lehuang@jhsph.edu
Grading: weekly homework (75%) and class project (25%)

Textbook:

- Peter D. Hoff: A First Course in Bayesian Statistical Methods (PH)

The course will follow the layout of the PH book, using many of the examples and definitions. In addition, we will adopt the notation for priors, posteriors, etc used in PH.

Course Overview

Bayesian Methods I (Ch. 1-6 of PDH):

- basic concepts: axioms of probability · marginal, joint, and conditional probabilities · priors and posteriors
- simple univariate distributions: binomial, Poisson, normal · hierarchical models
- computation: Monte Carlo approximation, Markov Chain Monte Carlo / Gibb's sampler

Bayesian Methods II:

- generalized linear models
- Bayesian nonparametrics
- special topics

Computing environment

- All examples in class will use R, and only R will be supported for questions pertaining to homework assignments.

Review

Conditional probability. Let A and B denote two events. The probability of event A given event B occured is

$$p(B|A) = \frac{p(A \cap B)}{p(A)}$$

Review

- My family has 2 children. At least 1 is a girl. What is the probability that both are girls? (Assume that the probability of a girl is $1/2$ and that the gender of the second child does not depend on the gender of the first child).

Review

- I have a *fair* deck of 52 cards, 4 of which are aces. You draw 2 cards, one at a time. What is the probability that both are aces?

Review

- You draw 3 cards, one at a time. What is the probability that all three are aces?

Review

Rules of craps (From D.A. Berry (DAB), Statistics, A Bayesian perspective).

- A player rolls two 6-sided dice. The sum on the two dice is all that matters.
- If you roll a sum of 7 or 11, you win immediately.
- You lose immediately if you roll a 2, 3, or 12.
- If you roll any other sum, this sum becomes your *point*. E.g., if you roll a 1 and a 4, your point is a 5.
  - roll repeatedly until you get your point or a sum of 7. You win if you get your point and lose if you get a 7.

What is the probability of winning? Is craps a *fair* game?

Review

Bayes' rule:

$$p(B|A) = \frac{p(A|B)p(B)}{p(A|B)p(B) + p(A|B^c)p(B^c)}$$

Review

Question posed to physicians regarding communicating risk of breast cancer for a woman with a positive mammogram test: `http://opinionator.blogs.nytimes.com/2010/04/25/chances-are/?_r=0`

The probability that one of these women has breast cancer is 0.8 percent. If a woman has breast cancer, the probability is 90 percent that she will have a positive mammogram. If a woman does not have breast cancer, the probability is 7 percent that she will still have a positive mammogram. Imagine a woman who has a positive mammogram. What is the probability that she actually has breast cancer?

(95% of the doctor's asked this question estimated the woman's probability to have breast cancer to be $\approx$ 75%.)

Review

$$
\begin{aligned}
p(C|+) &= \frac{p(+|C)p(C)}{p(+|C)p(C) + p(+|\bar{C})p(\bar{C})} \\
&= \frac{\text{sensitivity} \times \text{prevalence}}{\text{sensitivity} \times \text{prevalence} + \text{(1-specificity)} \times (1 - \mathit{prevalence})} \\
&= \frac{0.9 \times 0.008}{0.9 \times 0.008 + 0.07 \times (1 - 0.008)} \\
&= 0.09
\end{aligned}
$$

Review

$$\begin{aligned} p(B|A) &= \frac{p(A|B)p(B)}{p(A|B)p(B) + p(A|\bar{B})p(\bar{B})} \\ &= \frac{1}{1 + \frac{p(A|\bar{B})}{p(A|B)} \frac{p(\bar{B})}{p(B)}} \end{aligned}$$

$$p(\bar{B}|A) = 1 - p(B|A) = \frac{\frac{p(A|\bar{B})}{p(A|B)} \frac{p(\bar{B})}{p(B)}}{1 + \frac{p(A|\bar{B})}{p(A|B)} \frac{p(\bar{B})}{p(B)}}$$

and so

$$\frac{p(B|A)}{p(\bar{B}|A)} = \frac{p(A|B)}{p(A|\bar{B})} \times \frac{p(B)}{p(\bar{B})}.$$

Review

A bowl contains 5 syringes, each with a different vaccine. Assume the vaccines are either 100% protective or 0% protective for the flu. There are 6 possible models for the number of protective vaccines in the bowl. Assume the models are equally likely *a priori*.

- Experiment: You select a syringe from the bowl at random. A week later you are sick with the flu. What is your updated probabilities (posterior probabilities) for each of the models?

Review

- (HW1) Given the information from the first experiment (first syringe had a vaccine that was not protective for the flu), what is the probability that the next syringe selected will have a vaccine that works?

Bayesian methods

- Bayes' rule: a *rational* method for updating beliefs about characteristics of a population (parameters denoted by θ) after observing data (y)
- Bayesian inference: the process of *inductive* learning via Bayes' rule
- Bayesian methods: data analysis tools derived from the principles of Bayesian inference.

Bayes' learning

Bayes' learning begins with a numerical formulation of joint beliefs about y and θ, expressed in terms of probability distributions over $\mathcal{Y}$ and Θ.

1. *prior distribution* $p(\theta)$: describes our belief that θ represents the true population characteristics
2. sampling model $p(y|\theta)$: describes our belief that y would be the outcome of the study if we knew θ. In this term, we will only consider parametric models.
3. *posterior distribution* $p(\theta|y)$: describes our belief that θ is true having observed dataset y.

Importance of the prior

- The importance of the prior distribution is *not* that θ is generated from some distribution p nor its treatment as a random variable.
- The prior distribution is the best way to summarize available information (or the lack thereof) about θ, allowing the incorporation of its uncertainty in the decision (or estimation) process.

- Christian P Robert: “The Bayesian Choice”

Bayes' rule

The posterior distribution is obtained from the prior distribution and our sampling model via Bayes' rule:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_{\Theta} p(y|\tilde{\theta})p(\tilde{\theta})d\tilde{\theta}}$$

Example (from The Bayesian Choice)

(Bayes 1764) A billiard ball W is rolled on a line of length one, with a uniform probability of stopping anywhere. It stops at θ. A second ball is then rolled n times under the same assumptions and X denotes the number of times the ball stopped on the left of W. Given X, what inference can we make on θ?

Example (from The Bayesian Choice)

- The Bayesian approach is to derive the posterior distribution of $\theta|x$ when the prior distribution on θ is uniform on $[0, 1]$.
- Omitting the details, it can be shown that the posterior distribution is $\text{beta}(x + 1, n - x + 1)$.
- We can easily sample from this distribution, plot the density, etc. using R (see `dbeta, rbeta, qbeta`).

PH 1.2.1: estimating the probability of a rare event

To estimate the prevalance (θ) of a rare disease in a small city, we randomly sample 20 individuals.

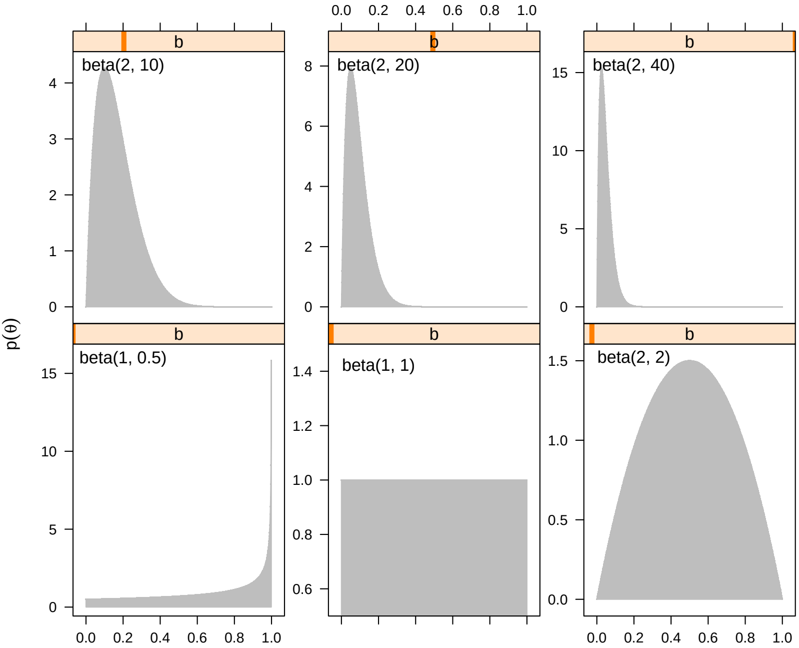
A sampling model for Y given θ:

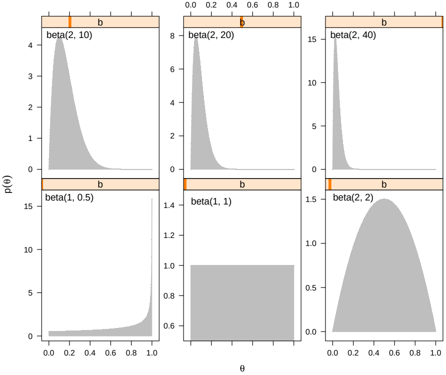
$$y|\theta \sim \text{binomial}(20, \theta).$$

Prior

- Infection rates from comparable cities range from 0.05 to 0.20 with an average prevalance of 0.10.
- To estimate the infection rate in a city of interest, we can incorporate the infection rates from other cities as prior knowledge
- Claim: probability distributions can be used to represent prior knowledge

```
A <- c(1, 1, 2, 2, 2, 2)
B <- c(0.5, 1, 2, 10, 20, 40)
x <- seq(0,1, by=0.001)
ylist <- foreach(a=A, b=B) %do% dbeta(x, a, b)
df <- data.frame(theta=rep(x, length(B)), p=unlist(ylist),
                 a=rep(A, each=length(x)),
                 b=rep(B, each=length(x)))
fig <- xyplot(p~theta | b, df, pch=".", col="gray",
              xlab=expression(theta),
              ylab=expression(p(theta)),
              scales=list(y=list(relation="free", rot=0)),
              shape1=df$a,
              shape2=df$b,
              panel=function(x,y, shape1, shape2, subscripts,...){
                      b <- shape2[subscripts[1]]
                      a <- shape1[subscripts[1]]
                      panel.xyplot(x, y, ...)
                      x <- x[is.finite(y)]
                      y <- y[is.finite(y)]
                      lpolygon(c(x, rev(x)),
                               y=c(rep(0, length(y)), rev(y)),
                               col="gray",
                               border="gray")
```


b
beta(2, 10)
b
beta(2, 20)
b
beta(2, 40)
b
beta(1, 0.5)
b
beta(1, 1)
b
beta(2, 2)
p(θ)
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
0
2
4
6
8
0
5
10
15
0.6
0.8
1.0
1.2
1.4
0.0
0.5
1.0
1.5

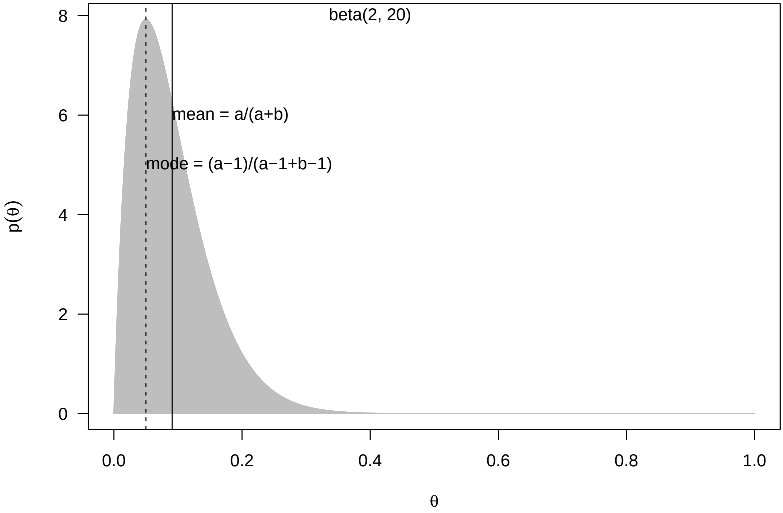


- Which probability distribution(s) best reflect the prevalence in comparable cities?
- Are probability distributions a reasonable means to represent our prior beliefs?

prior: dbeta(a=2, b=20)

```
pdf("../Figures1/beta_2_20.pdf", width=8, height=6)
par(las=1)
x <- seq(0,1,by=0.001)
y <- dbeta(x, 2, 20)
plot(x, y, type="l", col="gray", ylab=expression(p(theta)), xlab=expression
polygon(c(x, rev(x)), y=c(rep(0, length(x)), rev(y)), col="gray",
        border="gray")
text(0.4, 8, "beta(2, 20)")
abline(v=2/(2+20))
text(2/(2+20), y=6, "mean = a/(a+b)", adj=0)
abline(v=1/20,lty=2)
text(1/(1+19), y=5, "mode = (a-1)/(a-1+b-1)", adj=0)
dev.off()
```

prior: dbeta(a=2, b=20)



From Bayes' rule, we have

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_\Theta p(y|\tilde{\theta})p(\tilde{\theta})d\tilde{\theta}}.$$

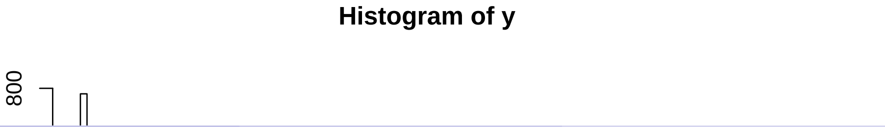
For

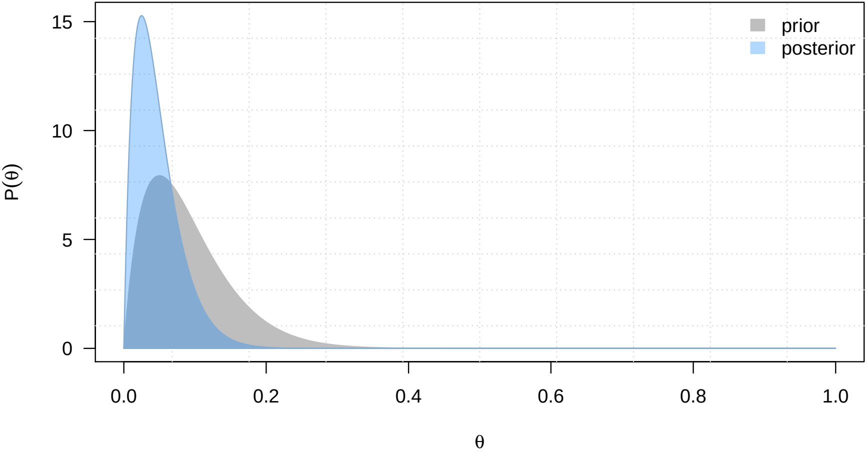
$$\begin{aligned} p(y|\theta) &= \text{binomial}(n, \theta) \text{ and} \\ p(\theta) &= \text{beta}(a, b), \text{then} \\ p(\theta|y) &= \text{beta}(a + y, b + n - y) \text{ (proof later).} \end{aligned}$$

If $\mathrm{Y} = 0$ in our city, then

$$p(\theta|y) = \text{beta}(2, 40).$$

```
pdf("../Figures1/betaposterior.pdf", width=8, height=5)
par(las=1)
x <- seq(0,1,by=0.001)
prior <- dbeta(x, 2, 20)
posterior <- dbeta(x, 2, 40)
plot(x, posterior, type="l", col="gray", ylab=expression(P(theta)),
     xlab=expression(theta), ylim=c(0, max(posterior)))
grid(nx=10, ny=10)
polygon(c(x, rev(x)), y=c(rep(0, length(x)), rev(prior)), col="gray",
        border="gray")
polygon(c(x, rev(x)), y=c(rep(0, length(x)), rev(posterior)),
        col=rgb(0, 0.5, 1, alpha=0.3),
        border=rgb(0, 0.5, 1, alpha=0.3))
legend("topright", border=c("gray", rgb(0, 0.5, 1, alpha=0.3)),
       fill=c("gray", rgb(0, 0.5, 1, alpha=0.3)),
       legend=c("prior", "posterior"), bty="n")
dev.off()
```





15
10
5
0
P(θ)
0.0
0.2
0.4
0.6
0.8
1.0
θ
prior
posterior

The posterior expectation is a weighed average of the prior expectation, θ_0, and the sample average:

$$
\begin{aligned}
E[\theta|y] &= \frac{a+y}{a+b+n} \\
&= \frac{a}{a+b+n} + \frac{y}{a+b+n} \\
&= \frac{a+b}{a+b+n}\frac{a}{a+b} + \frac{n}{a+b+n}\frac{y}{n} \\
&= w\theta_0 + (1-w)\bar{y}
\end{aligned}
$$

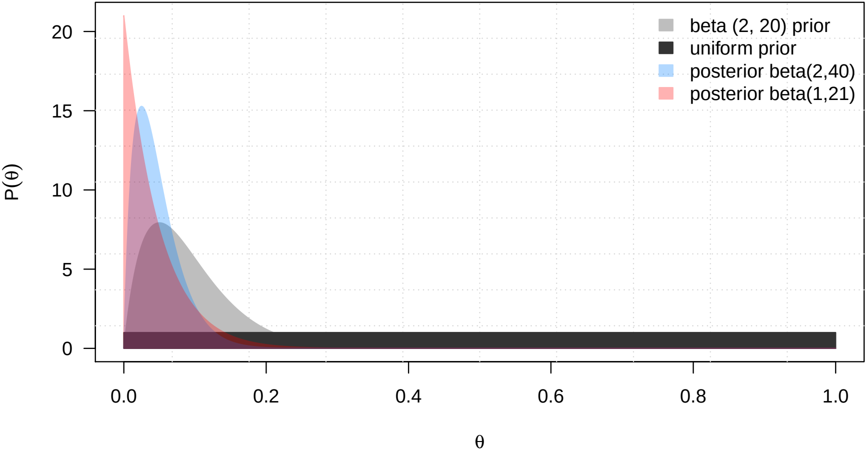
As the sample size increases, the posterior expectation is approximately the same as the sample average.

What if we had used a less informative prior?

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

```r
a <- 1; b <- 1; n <- 20; y <- 0
x <- seq(0, 1, by=0.001)
prior.uniform <- dbeta(x, a, b)
posterior.uniform <- dbeta(x, a+y, b+n-y)
```

```
pdf("../Figures1/uniformposterior.pdf", width=8, height=5)
par(las=1)
plot(x, posterior, type="n", col="gray", ylab=expression(P(theta)),
     xlab=expression(theta), ylim=c(0, max(posterior.uniform)))
grid(nx=10, ny=10)
polygon(c(x, rev(x)), y=c(rep(0, length(x)), rev(prior)), col="gray",
        border="gray")
polygon(c(x, rev(x)), y=c(rep(0, length(x)), rev(prior.uniform)),
        col="gray20",
        border="gray20")
polygon(c(x, rev(x)), y=c(rep(0, length(x)), rev(posterior)),
        col=rgb(0, 0.5, 1, alpha=0.3),
        border=rgb(0, 0.5, 1, alpha=0.3))
polygon(c(x, rev(x)), y=c(rep(0, length(x)), rev(posterior.uniform)),
        col=rgb(1, 0, 00, alpha=0.3),
        border=rgb(1, 0, 0, alpha=0.3))
cols <- c("gray", "gray20",
          rgb(0, 0.5, 1, alpha=0.3),
          rgb(1, 0, 0, alpha=0.3))
legend("topright", border=cols, fill=cols,
       legend=c("beta (2, 20) prior", "uniform prior", "posterior beta(2,40
       "posterior beta(1,21)"), bty="n")
dev.off()
```


beta (2, 20) prior
uniform prior
posterior beta(2,40)
posterior beta(1,21)
P(θ)
20
15
10
5
0
0.0
0.2
0.4
0.6
0.8
1.0
θ

Prediction

Suppose we are planning to sample an additional person, $\tilde{y}$, from the population and test for disease. Given y (and assuming $\tilde{y}$ is independent of y given θ), we would like to know the predictive density. $\tilde{y}$ is also binomial: $p(\tilde{y}|y)$.

Prediction

With a uniform prior, we have

$$
\begin{aligned}
Pr(\tilde{Y} = 1|y) &= \int Pr(\tilde{Y} = 1, \theta|y) d\theta \\
&= \int Pr(\tilde{Y} = 1|\theta, y) p(\theta|y) d\theta \\
&= \int \theta p(\theta|y) d\theta \\
&= E[\theta|y] \\
&= \frac{a + \sum_i^n y_i}{a + b + n} \\
&= \frac{1 + 0}{1 + 1 + 20} \\
&= 0.045
\end{aligned}
$$

Predictive distribution

- the predictive distribution does not depend on any unknowns
- the predictive distribution depends on the observed data

Likelihood principle

The information about θ from data y is entirely contained in the likelihood function $l(\theta|y)$.

If y_1 and y_2 are two observations depending on the same parameter θ, there exists a constant c satisfying

$$l_1(\theta|y_1) = cl_2(\theta|y_2)$$

for every θ. The observations carry the same information about θ and lead to the same inference.

Example:

While working on the audience share of a TV series, $0 \leq \theta \leq 1$ representing the part of the TV audience, an investigator found 9 viewers and 3 non-viewers. If no additional information is available on the experiment, two probability models can be proposed:

1. the investigator questioned 12 persons, observing binomial$(12, \theta)$ with $y = 9$.

2. the invesigator questioned N persons until 3 nonviewers were obtained, with $N \sim$ negative binomial$(3, 1 - \theta)$ and $N = 12$.

Christian P. Roberts: *The Bayesian Choice*

Frequentist approach

$$
\begin{aligned}
L_1(\theta) &= \binom{n}{y}\theta^y(1-\theta)^{n-y} \\
&= \binom{12}{9}\theta^9(1-\theta)^3.
\end{aligned}
$$

The p-value for the rejection region:

$$
\begin{aligned}
\alpha_1 = P_{\theta=0.5}(X \geq 9) &= \sum_{j+9}^{12}\binom{12}{j}\theta^j(1-\theta)^{12-j} \\
&= 0.075.
\end{aligned}
$$

$$
\begin{aligned}
L_2(\theta) &= \binom{r+y-1}{y}\theta^y(1-\theta)^r \\
&= \binom{11}{9}\theta^9(1-\theta)^3.
\end{aligned}
$$

The p-value for the rejection region:

$$
\begin{aligned}
\alpha_2 = P_{\theta=0.5}(X \geq 9) &= \sum_{j+9}^{12}\binom{2+j}{j}\theta^j(1-\theta)^{12-j} \\
&= 0.033.
\end{aligned}
$$

Bayesian approach

$$
\begin{aligned}
p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta} \\
&= \frac{l(\theta|y)p(\theta)}{\int l(\theta|y)p(\theta)d\theta}
\end{aligned}
$$

which depends on y only through the likelihood.

- A multiplicative constant that does not depend on θ will change the scale of the posterior distribution but not its shape.

- Since the likelihoods l_1 and l_2 are proportional to $\theta^9(1-\theta)^3$ differing only by a multiplicative constant, the posterior inference will be the same regardless of the sampling model.