Inference for a Population Proportion

Al Nosedal.
University of Toronto.

November 11, 2015

Statistical inference is drawing conclusions about an entire population based on data in a sample drawn from that population. From both frequentist and Bayesian perspectives, there are three main goals of inference: estimation, hypothesis testing, and prediction. Estimation and hypothesis testing deal with drawing conclusions about unknown and unobservable population parameters. Prediction is estimating the values of potentially observable but currently unobserved quantities.

Bayesian Inference: Summarizing the Posterior Distribution

All Bayesian inference is based on the posterior distribution, which contains all the current information about the unknown parameter. Although a plot of the posterior density gives a full graphical description, numeric summaries of the posterior are needed as well.

The Posterior Mean

The mean of the posterior distribution is often used as the Bayesian point estimate of a parameter. For a beta prior and binomial likelihood, the posterior mean is

$$E(\pi|y) = \frac{\alpha + y}{\alpha + \beta + n}$$

Example

In our example, with the beta(10, 40) prior

$$E(\pi|y) = \frac{\alpha + y}{\alpha + \beta + n} = \frac{17}{100} = 0.17$$

A comment about Posterior Mean

If we denote the posterior mean by μ_{post}, then

$$\mu_{post} = \frac{\alpha + y}{\alpha + \beta + n} = w \frac{\alpha}{\alpha + \beta} + (1 - w) \frac{y}{n}$$

where $w = \frac{\alpha+\beta}{\alpha+\beta+n}$.

Other Bayesian Point Estimates

The posterior median and posterior mode are sometimes used instead of the posterior mean as Bayesian point estimates.

The Posterior Variance

The posterior variance is one summary of the spread of the posterior distribution. The larger the posterior variance, the more uncertainty we still have about the parameter, even after learning from the current data.

Example

In our school-quitting example, with the uniform prior, the prior variance $=\frac{1}{12} = 0.083$, and posterior variance $= 0.00246$. If we instead used the Beta(10, 40) prior, the prior variance $= 0.003144$ and the posterior variance $= 0.00140$. As we would expect, the posterior variance is smaller with the informative Beta(10,40) prior than with the noninformative uniform prior.

Bayesian Posterior Intervals

Intervals called "credible sets" also are used as numeric posterior summaries. There are two commonly used kinds.

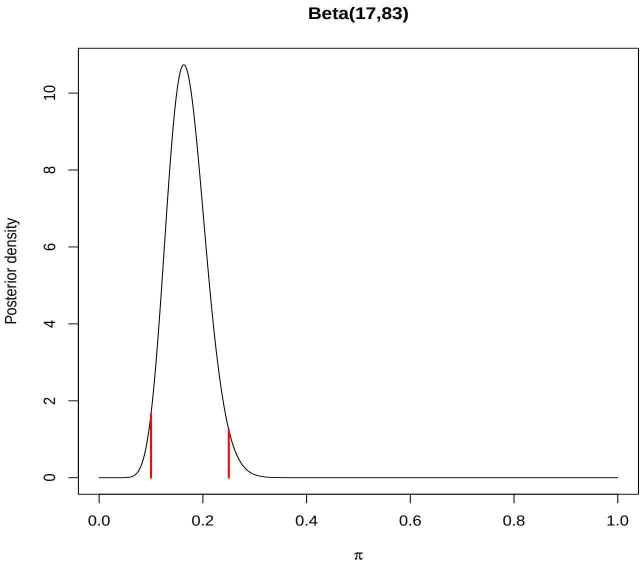
Equal-Tail Posterior Credible Sets

For example, the endpoints of a 95% equal-tail credible set are the 0.025 and the 0.975 quantiles of the posterior distribution. We can use built-in R functions to calculate them. For our quitting- school problem with the beta(10,40) prior, the posterior density was Beta(17, 83), and the qbeta function in R can be used as follows:

```
qbeta( c(0.025, 0.975), 17, 83 );
```

This interval is shown graphically in the next slide.

Equal-Tail Posterior Credible Sets



Highest Posterior Density Regions

The other kind of Bayesian posterior interval is the highest posterior density region, or HPD region. The posterior density at any point inside such an HPD region is greater than the density at any point outside it. The HPD region also is the shortest possible interval trapping the desired probability. HPD regions are preferable to equal-tail credible sets when the posterior is highly skewed or multimodal. However, they are **generally difficult to compute**. The intuition behind the computation of an HPD region is as follows. Suppose that we want a 95% posterior probability region. We begin by placing a horizontal line just touching the posterior density curve at its mode. We then slide the line downwards toward the x-axis until it cuts the density curve at points such that the area under the density curve between these points is exactly 0.95.

Interpretation of Bayesian Intervals

Recall that the posterior distribution represents our updated subjective probability distribution for the unknown parameter. Thus, for us, the interpretation of the 95% credible set is that the probability that the true π is in that interval is 0.95. For example, if the Beta(10,40) had been a true representation of our prior beliefs or

$$P(0.103 < \pi < 0.249) = 0.95.$$

Using the Posterior Distribution to Test Hypotheses

$H_1 : \pi \leq 0.10$

vs

$H_2 : \pi > 0.10$

Using the Posterior Distribution to Test Hypotheses

We simply need the posterior probabilities of these two ranges of values for π. Suppose that the Beta(10, 40) had been our true prior, so our posterior distribution is Beta(17, 83). We can use a built-in R function to obtain $P(\pi \leq 0.1|y)$.

```
pbeta(0.1, 17, 83);
```

(you should get 0.0187, roughly).

With this prior, we would conclude that $P(\pi \leq 0.1|y) \approx 0.019$.

Using the Posterior Distribution to Test Hypotheses

Note that different people, approaching the question with different prior information, will end up with different (subjective) posterior probabilities on H_0. Different people also will have different views on how small $P(\pi \leq 0.1|y)$ has to be in order for it to be appropriate to go before the regents, for instance.

Bayes Factor

A Bayesian can test competing hypotheses H_1 and H_2 by examining the Bayes factor (Jeffreys 1961)

$$BF(H_1/H_2|y) = \frac{P_{posterior}(H_1|y)/P_{posterior}(H_2|y)}{P_{prior}(H_1|y)/P_{prior}(H_2|y)}$$

The Bayes factor is the odds ratio of the posterior odds to the prior odds of the hypothesis H_1 relative H_2.

Bayes Factor

Jeffreys recommends the following guidelines for degrees of evidence for H_1 and H_2, based on ranges of Bayes factor values:

1. $BF(H_1/H_2|y) < 1/100$: decisive evidence for H_2.
2. $1/100 < BF(H_1/H_2|y) < 1/10$: strong evidence for H_2.
3. $1/10 < BF(H_1/H_2|y) < 1/\sqrt(10)$: substantial evidence for H_2.
4. $1/\sqrt(10) < BF(H_1/H_2|y) < 1$: minimal evidence for H_2.
5. $1 < BF(H_1/H_2|y) < \sqrt(10)$: minimal evidence for H_1.
6. $\sqrt(10) < BF(H_1/H_2|y) < 10$: substantial evidence for H_1.
7. $10 < BF(H_1/H_2|y) < 100$: strong evidence for H_1.
8. $BF(H_1/H_2|y) > 100$: decisive evidence for H_1.

Example

In our school-quitting example, $H_1 : \pi \leq 0.10$ and $H_2 : \pi > 0.10$. The prior odds for H_1/H_2 based on a flat noninformative prior $Beta(1,1)$, is $P_{prior}(H_1)/P_{prior}(H_2) = 0.10/0.90 \approx 0.1111$. A random sample survey of $n = 50$ students with the binomial model provide $y = 7$ (yesses) and a proportion estimate of $\hat{\pi} = \frac{7}{50} = 0.14$.

Example (cont.)

The posterior distribution based on the conjugate prior is $Beta(8, 44)$ with posterior odds $P_{posterior}(H_1|y)/P_{posterior}(H_2|y) = 0.1330/0.8670 = 0.1534$ and Bayes factor

$$BF(H_1/H_2|y) = \frac{P_{posterior}(H_1|y)/P_{posterior}(H_2|y)}{P_{prior}(H_1|y)/P_{prior}(H_2|y)}$$

$$= \frac{0.1534}{0.1111} = 1.3807$$

providing minimal evidence for $H_1 : \pi \leq 0.10$.

R code

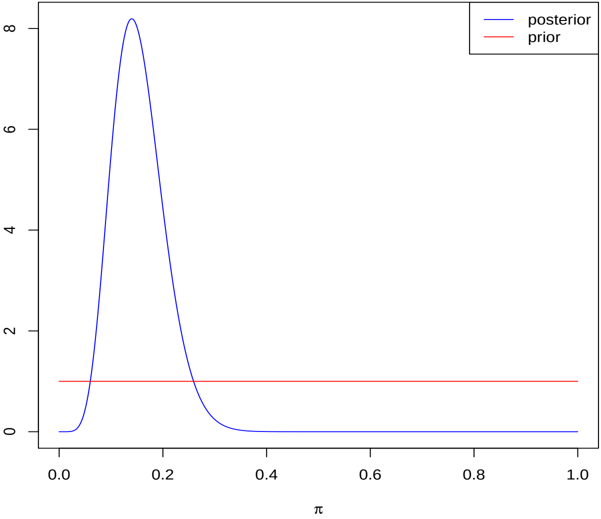
The probabilities for the posterior Beta distribution are calculated using R cumulative distribution command
for H_1.

```
pbeta(0.10,8,44);
```

for H_2

```
1-pbeta(0.10,8,44);
```

Prior and Posterior



```
## Noninformative prior and posterior;

p=seq(0,1,by=0.001);

prior=dbeta(p,1,1);

posterior=dbeta(p,8,44);

plot(p,posterior,col="blue",xlab=expression(pi),
ylab=" ", type="l");

lines(p,prior,col="red");

legend("topright",c("posterior","prior"),
col=c("blue","red"),lty=c(1,1));

# lty =1 tells R to draw a regular line;
```

Example (informative prior)

In our school-quitting example, $H_1 : \pi \leq 0.10$ and $H_2 : \pi > 0.10$. The prior odds for H_1/H_2 based on a prior $Beta(10, 40)$, is $P_{prior}(H_1)/P_{prior}(H_2) = 0.0215/0.9785 \approx 0.02197$. A random sample survey of $n = 50$ students with the binomial model provide $y = 7$ (yesses) and a proportion estimate of $\hat{\pi} = \frac{7}{50} = 0.14$.

Example (cont.)

The posterior distribution based on the conjugate prior is $Beta(17, 83)$ with posterior odds $P_{posterior}(H_1|y)/P_{posterior}(H_2|y) = 0.0188/0.9812 = 0.0191$ and Bayes factor

$$BF(H_1/H_2|y) = \frac{P_{posterior}(H_1|y)/P_{posterior}(H_2|y)}{P_{prior}(H_1|y)/P_{prior}(H_2|y)}$$

$$= \frac{0.0191}{0.02197} \approx 0.87$$

providing minimal evidence for $H_2 : \pi > 0.10$.

Posterior Predictive Distributions

In many studies, the research question of interest is predicting values of a future sample from the same population. For example, suppose we are considering interviewing another sample of 50 UI students in the hope of getting more evidence to present to the regents, and we would like to get an idea of how it is likely to turn out before we go to the trouble of doing so!

Posterior Predictive Distributions

More generally, we are considering a new sample of sample size n^* and want to estimate the probability of some particular number y^* of successes in this sample. We need the probability of getting y^* successes in a future sample given the information in our current data y, not given some particular value of π. Recall that all of our current knowledge about π is contained in the posterior distribution obtained using the original survey.

Posterior Predictive Distributions

Thus, the posterior predictive probability of getting some particular value of y^* in a future sample of size n^* is

$$p(y^*|y) = \int_0^1 p(y^*|\pi)p(\pi|y)d\pi, \quad Y = y^* = 0, 1, ..., n.$$

where y denotes the data from the original sample and $p(\pi|y)$ is the posterior distribution based on that sample.

Posterior Predictive Distributions

This is particularly easy to compute if $n^* = 1$, in which case the probability of getting 1 success is π, so

$$\begin{aligned} P(y^* = 1|y) &= \int_0^1 p(y^* = 1|\pi) p(\pi|y) d\pi \\ &= \int_0^1 \pi p(\pi|y) d\pi = E(\pi|y) \end{aligned}$$

because, by definition, the expected value of a random variable is obtained by integrating the random variable over its density. This is just the posterior mean of π.

Example

If we had used the Beta(10,40) prior, resulting in the posterior density being Beta(17,83), then $Pr(y^* = 1|y) = \frac{17}{100} = 0.17$.

In general, if a Bayesian analysis has been done to estimate a population proportion π, using a $Beta(\alpha, \beta)$ prior and a dataset with y successes in a sample of size n, then the posterior density $P(\pi|y)$ is $Beta(\alpha_{post}, \beta_{post})$, where $\alpha_{post} = \alpha + y = \alpha^*$ and $\beta_{post} = \beta + n - y = \beta^*$, and the predictive probability of getting y^* successes in a future sample of size n^* is

$$P(y^*|y) = \int_0^1 \binom{n^*}{y^*} \pi^{y^*} (1-\pi)^{n^*-y^*} \frac{\Gamma(\alpha^* + \beta^*)}{\Gamma(\alpha^*)\Gamma(\beta^*)} \pi^{\alpha^*-1} (1-\pi)^{\beta^*-1} d\pi$$

So the posterior predictive probability will be

$$P(y^*|y) = \binom{n^*}{y^*} \frac{\Gamma(\alpha^* + \beta^*)}{\Gamma(\alpha^*)\Gamma(\beta^*)} \frac{\Gamma(y^* + \alpha^*)\Gamma(n^* - y^* + \beta^*)}{\Gamma(\alpha^* + \beta^* + n^*)}$$

(This distribution is known as the **Beta-Binomial distribution**).

Easy Example

Find the posterior predictive probability when $\alpha^* = 1$, $\beta^* = 1$, and $n^* = 3$.

Easy Example

Solution.
If $\alpha^* = 1$, $\beta^* = 1$, and $n^* = 3$, then we have

$$\begin{aligned} P(y^* = 0|y) &= \binom{3}{0} \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)} \frac{\Gamma(1)\Gamma(3-0+1)}{\Gamma(1+1+3)} \\ &= (1)\frac{(1!)}{(0!)(0!)}\frac{(0!)(3!)}{(4!)} = \frac{1}{4} \end{aligned}$$

Easy Example (cont.)

Doing something similar we have that

$$P(y^* = 1|y) = \binom{3}{1} \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)} \frac{\Gamma(2)\Gamma(3-1+1)}{\Gamma(1+1+3)} = \frac{1}{4}$$
$$P(y^* = 2|y) = \binom{3}{2} \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)} \frac{\Gamma(3)\Gamma(3-2+1)}{\Gamma(1+1+3)} = \frac{1}{4}$$
$$P(y^* = 3|y) = \binom{3}{3} \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)} \frac{\Gamma(4)\Gamma(3-3+1)}{\Gamma(1+1+3)} = \frac{1}{4}$$

Gibbs Sampling Procedure

First, let the parameter vector be partitioned into blocks

$$\theta = (\theta_1, \theta_2, ..., \theta_J)$$

where θ_J is the Jth block of parameters. Each block contains one or more parameters. Let $\theta_{-\mathbf{J}}$ be the set of all the other parameters not in block J. The proportional form of Bayes theorem,

$$g(\theta_1, \theta_2, ..., \theta_J | y) \propto f(y | \theta_1, \theta_2, ..., \theta_J) \times g(\theta_1, \theta_2, ..., \theta_J)$$

gives the shape of the joint posterior density of all the parameters, where
$f(y|\theta_1, \theta_2, ..., \theta_J)$ and $g(\theta_1, \theta_2, ..., \theta_J)$
are the joint likelihood and the joint prior density for all parameters. This gives us the shape of the joint posterior, not its scale.

Gibbs Sampling Procedure

Gibbs sampling requires that we know the full conditional distribution of each block of parameters θ_J, given all the other parameters $\theta_{-\mathbf{J}}$ and the data y. Let the full conditional distribution of block θ_J be denoted

$$g(\theta_J|\theta_{-\mathbf{J}} = g(\theta_J|\theta_1, ..., \theta_{J-1}, \theta_{J+1}, ..., \theta_J, y)$$

These full conditional distributions may be very complicated, but we must know them to run the Gibbs sampler. In Gibbs sampling, we will cycle through the parameter blocks in turn, drawing each one from its full conditional distribution given the most recent values of the other parameter blocks, and all the observed data.

Steps of the Gibbs Sampler

- At time $n = 0$ start from an arbitrary point in the parameter space $\theta^{\mathbf{0}} = (\theta_1^0, \theta_2^0, ..., \theta_J^0)$
- For $n = 1, 2, ..., N$.
  For $j = 1, 2, ..., J$, draw $\theta_J^{(n)}$ from
  $g(\theta_j | \theta_1^{(n)}, ..., \theta_{j-1}^{(n)}, \theta_j^{(n)}, ..., \theta_J^{(n-1)}, y)$
- The long-run distribution of $\theta^{(\mathbf{N})} = (\theta_1^{(N)}, ..., \theta_J^{(N)})$ is the true posterior $g(\theta_1, ..., \theta_J | y)$.

Example

We shall illustrate the procedure with our easy example. We suppose that π and y^* have the joint distribution

$$p(y^*, \pi) = \binom{n^*}{y^*} \pi^{y^* + \alpha^* - 1} (1 - \pi)^{n^* - y + \beta^* - 1}$$

and that we are interested in the marginal distribution of y^*. Rather than integrating with respect to π, which would show that y has a Beta-Binomial distribution, we proceed to find the required distribution from the two conditional distributions:

$$y^* | \pi \sim Binomial(n^*, \pi),$$

$$\pi | y^* \sim Beta(y^* + \alpha^*, n^* - y^* + \beta^*)$$

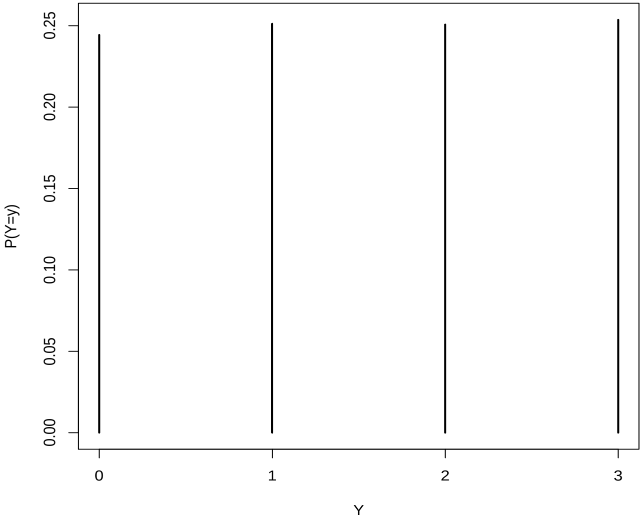
```
iter=100000;

#initializing matrix;

both=matrix(0,nrow=iter,ncol=2);

# parameters;

alpha=1;

beta=1;

m=3;
```

```
# initializing sampling;

set.seed(2015);

# first pi;

both[1,1]=runif(1);

# first y;

both[1,2]=rbinom(1,size=m,both[1,1]);
```

```
for(i in 2:iter )
{
# new y;

both[i,1]=rbeta(1,both[i-1,2]+alpha,m-both[i-1,2]+beta);

# new pi;

        both[i,2]=rbinom(1,size=m,both[i,1]);

}
```

```
# Probability distribution of Y;

dist=table(both[ ,2])/iter;

dist


# Graph of pmf;

plot(dist,xlab="Y",ylab="P(Y=y)");
```

Graph of Beta-Binomial



Negative Binomial (Pascal) Distribution

The probability mass function (pmf) for the distribution of the number of the trial on which the rth success occurs in independent Bernoulli trials of an experiment in which π denotes the probability of success on a single trial, is given by:

$$f(y|\pi, r) = \binom{y-1}{r-1}\pi^r(1-\pi)^{y-r}, \quad y = r, r+1, ...$$

Vague prior (noninformative)

A vague prior for π is given by the uniform prior density:

$$g(\pi) = 1, \quad 0 < \pi < 1.$$

The kernel of the posterior distribution for π is given by

$$h(\pi|y, r) \propto \pi^{(r+1)-1}(1-\pi)^{(y-r+1)-1}$$

Bayes' estimator of π

The Bayes' estimator of π is given by

$$\hat{\pi} = E(\pi|y,r) = \frac{r+1}{r+1+y-r+1} = \frac{r+1}{y+2}$$

Conjugate prior

A natural conjugate prior for π is given by the Beta prior density kernel:

$$g(\pi) \propto \pi^{\alpha-1}(1-\pi)^{\beta-1},\ \alpha > 0,\ \beta > 0,\ 0 < \pi < 1.$$

The posterior density kernel is found as

$$h(\pi|y,r) \propto \pi^{(\alpha+r)-1}(1-\pi)^{(\beta+y-r)-1}$$

Bayes' estimator of π

The Bayes' estimator of π is given by

$$\hat{\pi} = E(\pi|y, r) = \frac{\alpha + r}{(\alpha + r) + (\beta + y - r)} = \frac{\alpha + r}{\alpha + \beta + y}$$