An Introduction to the Bayes approach in survey sampling

Here we will consider some simple examples which demonstrate the basic calculations underlying the Bayesian approach to survey sampling.

We begin by recalling the definition of conditional probability for two events A and B, i.e.

$$P(B \mid A) = P(A \cap B)/P(A) \quad \text{or} \quad P(A \cap B) = P(A)P(B \mid A)$$

In a finite sample space let $A_1, \ldots, A_K$ be a set of mutually exclusive events whose union is the whole space. Let B be another event. Then we have

$$\begin{aligned} P(B) &= \sum_{i=1}^{K} P(A_i \cap B) \\ &= \sum_{i=1}^{K} P(A_i)P(B \mid A_j) \end{aligned}$$

and

$$\begin{aligned} P(A_j \mid B) &= P(A_j \cap B)/P(B) \\ &= \frac{P(A_j)P(B \mid A_j)}{\sum_{i=1}^{K} P(A_i)P(B \mid A_i)} \end{aligned}$$

A two urn example

As an example, suppose we have two urns where the first contains 3 white balls and 7 blue balls and the second contains 6 white balls and 4 blue balls. Consider the random experiment which selects urn 1 with probability 1/3 and urn two with probability 2/3 and then a ball is selected at random from the selected urn. Let w_1 be the event that the selected ball was white, and I and II are the events that urn 1 or urn 2 was selected. Then from the above we see that

$$\begin{aligned} P(w_1) &= P(I)P(w_1 \mid I) + P(II)P(w_1 \mid II) \\ &= (\frac{1}{3})(\frac{3}{10}) + (\frac{2}{3})(\frac{6}{10}) \end{aligned}$$

and

$$\begin{aligned} P(I \mid w_1) &= \frac{p(I \cap w_1)}{p(w_1)} \\ &= \frac{(\frac{1}{3})(\frac{3}{10})}{P(w_1)} \end{aligned}$$

Suppose now instead of drawing just one ball from the selected urn we draw two balls using srswr. Let w_2 be the event that the second ball drawn was white. Then

$$\begin{aligned} P(w_2 \mid w_1) &= \frac{P(w_1 \cap w_2)}{P(w_1)} \\ &= \frac{\frac{1}{3}(\frac{3}{10})^2 + \frac{2}{3}(\frac{6}{10})^2}{P(w_1)} \\ &= \frac{(\frac{1}{3})(\frac{3}{10})}{P(w_1)}\frac{3}{10} + \frac{(\frac{2}{3})(\frac{6}{10})}{P(w_1)}\frac{6}{10} \\ &= P(I \mid w_1)P(w_2 \mid I) + P(II \mid w_1)P(w_2 \mid II) \end{aligned}$$

Now suppose instead of sampling just 2 balls from the selected urn we sampled 5 balls using srswr. Then the conditional probability of seeing a white ball on the 5th draw given that we have seen two whites then a blue and another white is

$$\begin{aligned}
P(w_5 \mid w_1, w_2, b_3, w_4) &= \frac{P(w_1, w_2, b_3, w_4, w_5)}{P(w_1, w_2, b_3, w_4)} \\
&= \frac{\frac{1}{3}(\frac{3}{10})^4\frac{7}{10} + \frac{2}{3}(\frac{6}{10})^4\frac{4}{10}}{\frac{1}{3}(\frac{3}{10})^3\frac{7}{10} + \frac{2}{3}(\frac{6}{10})^3\frac{4}{10}} \\
&= \frac{\frac{1}{3}(\frac{3}{10})^3\frac{7}{10}}{P(w_1, w_2, b_3, w_4)}\frac{3}{10} + \frac{\frac{2}{3}(\frac{6}{10})^3\frac{4}{10}}{P(w_1, w_2, b_3, w_4)}\frac{6}{10} \\
&= P(I \mid w_1, w_2, b_3, w_4)\frac{3}{10} + P(II \mid w_1, w_2, b_3, w_4)\frac{6}{10} \\
&= P(I \mid w_1, w_2, b_3, w_4)P(w_5 \mid I) + P(II \mid w_1, w_2, b_3, w_4)P(w_5 \mid II)
\end{aligned}$$

In the same way we have

$$\begin{aligned}
P(w_5, b_6 \mid w_1, w_2, b_3, w_4) &= P(I \mid w_1, w_2, b_3, w_4)(\frac{3}{10})(\frac{7}{10}) + P(II \mid w_1, w_2, b_3, w_4)(\frac{6}{10})(\frac{4}{10}) \\
&= P(I \mid w_1, w_2, b_3, w_4)P(w_5, b_5 \mid I) + P(II \mid w_1, w_2, b_3, w_4)P(w_5, b_6 \mid II)
\end{aligned}$$

A K urn example

Suppose we have an urn which contains N balls which are labeled $1, 2, \ldots, N$. Each ball is either white or blue. Let $y_j = 1$ if the jth ball is white and equal 0 if the jth ball is blue. Assume the colors were assigned to the balls using the following probability model. Let $0 < p < 1$ be fixed. Then the y_j's in the urn were iid Bernoulli(p). So for any choice of the y_j's we have

$$P(y_1, \ldots, y_N) = p^{\sum_{j=1}^N y_j}(1-p)^{N-\sum_{j=1}^N y_j}$$

Suppose the value of p is not known but it was chosen randomly from a set of K possible values, say $\{p_1, \ldots, p_K\}$ with probabilities $\{\pi_1, \ldots, \pi_K\}$. Here both the p_i's and π_i's are assumed to be known. Then under this two step random process for generating the colors of the balls in the urn the marginal probability of a vector of the y_j's is given by

$$P(y_1, \ldots, y_N) = \sum_{i=1}^K \pi_i\, p_i^{\sum_{j=1}^N y_j}(1-p_i)^{N-\sum_{j=1}^N y_j}$$

Suppose after this two step process of generating the y_j values in the urn we are allowed to take a srswor of size n from the urn and observe the colors of the selected balls. For notational simplicity we assume that $(y_1, \ldots, y_n)$ were the balls in the sample. Then it is easy to see that

$$P(y_1, \ldots, y_n) = \sum_{i=1}^K \pi_i\, p_i^{\sum_{j=1}^n y_j}(1-p_i)^{n-\sum_{j=1}^n y_j}$$

Then arguing as in the previous section we have

$$
\begin{aligned}
P(y_{n+1},\ldots,y_N \mid y_1,\ldots,y_n) &= \frac{P(y_1,\ldots,y_N)}{P(y_1,\ldots,y_n)} \\
&= \frac{\sum_{i=1}^K \pi_i\, p_i^{\sum_{j=1}^N y_j}(1-p_i)^{N-\sum_{j=1}^N y_j}}{P(y_1,\ldots,y_n)} \\
&= \sum_{i=1}^K \Big(p_i^{\sum_{j=n+1}^N y_j}(1-p_i)^{N-n-\sum_{j=n+1}^N y_j} \times \\
&\qquad \frac{\pi_i\, p_i^{\sum_{j=1}^n y_j}(1-p_i)^{n-\sum_{j=1}^n y_j}}{P(y_1,\ldots,y_n)} \Big) \\
&= \sum_{i=1}^K p_i^{\sum_{j=n+1}^N y_j}(1-p_i)^{N-n-\sum_{j=n+1}^N y_j} P(p_i \mid y_1,\ldots,y_n)
\end{aligned}
$$

Note the prior expectation of y_i is $\sum_{i=1}^K p_i\pi_i$ while the posterior expectation given $y_1,\ldots,y_n$ for a y_j with $j > n$ is $\sum_{i=1}^K p_i P(p_i \mid y_1,\ldots,y_n)$.

Infinitely many urns

To handle this case we need to use some calculus.

The **gamma function** is defined as

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1}\,\exp^{-u}\,du \qquad \alpha > 0$$

This function has the following property

$$\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1) \qquad \alpha > 0$$

A random variable p taking on values between 0 and 1 has a **beta distribution with parameters** $\alpha > 0$ **and** $\beta > 0$ if its probability density function is

$$f(p \mid \alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \qquad \text{for } 0 < p < 1$$

Its mean and variance are

$$E(p) = \frac{\alpha}{\alpha+\beta} \qquad \text{and} \qquad V(P) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Consider the probability model

$$
\begin{aligned}
y_1,\ldots,y_N \mid p &\sim \text{iid Bernoulli}(p) \\
p &\sim \text{Beta with parameters } \alpha \text{ and } \beta
\end{aligned}
$$

then the marginal distribution of the y_i's is

$$P(y_1,\ldots,y_N) = \int_0^1 p^{\sum_{i=1}^N y_i}(1-p)^{N-\sum_{i=1}^N y_i}\, f(p \mid \alpha,\beta)\, dp$$

Let $n < N$ and suppose we have observed $(y_1, \ldots, y_n)$. Then

$$\begin{aligned} f(p \mid y_1, \ldots, y_n) &= \frac{f(p, y_1, \ldots, y_n)}{f(y_1, \ldots, y_n)} \\ &= \frac{p^{\sum_{i=1}^n y_i}(1-p)^{n-\sum_{i=1}^n y_i} f(p)}{f(y_1, \ldots, y_n)} \\ &\propto p^{\sum_{i=1}^n y_i + \alpha - 1}\,(1-p)^{n-\sum_{i=1}^n y_i + \beta - 1} \end{aligned}$$

Hence if the prior distribution for p is Beta with parameters α and β then after the data has been observed the posterior distribution for p is Beta with parameters $\alpha + \sum_{i=1}^n y_i$ and $\beta + n - \sum_{i=1}^n y_i$. From this it follows that

$$\begin{aligned} P(y_{n+1} = 1 \mid y_1, \ldots, y_n) &= \frac{P(y_1, \ldots, y_n, y_{n+1} = 1)}{P(y_1, \ldots, y_n)} \\ &= \frac{\int_0^1 p\, p^{\sum_{i=1}^n y_i}(1-p)^{n-\sum_{i=1}^n y_i} f(p)\, dp}{\int_0^1 p^{\sum_{i=1}^n y_i}(1-p)^{n-\sum_{i=1}^n y_i} f(p)\, dp} \\ &= \int_0^1 p f(p \mid y_1, \ldots, y_n)\, dp \\ &= E(p \mid y_1, \ldots, y_n) \\ &= \frac{\alpha + \sum_{i=1}^n y_i}{\alpha + \beta + n} \\ &= \frac{\alpha + \beta}{\alpha + \beta + n}\frac{\alpha}{\alpha + \beta} + \frac{n}{\alpha + \beta + n}\bar{y}_n \\ &= E(y_{n+1} \mid y_1, \ldots, y_n) \end{aligned}$$

More generally one finds that

$$\begin{aligned} P(y_{n+1}, \ldots, y_N \mid y_1, \ldots, y_n) &= \frac{P(y_1, \ldots, y_N)}{P(y_1, \ldots, y_n)} \\ &= \int_0^1 p^{\sum_{i=n+1}^N y_i}(1-p)^{N-n-\sum_{i=n+1}^N y_i}\, \frac{p^{\sum_{i=1}^n y_i}(1-p)^{n-\sum_{i=1}^n y_i}}{P(y_1, \ldots, y_n)}\, f(p)\, dp \\ &= \int_0^1 p^{\sum_{i=n+1}^N y_i}(1-p)^{N-n-\sum_{i=n+1}^N y_i}\, f(p \mid y_1, \ldots, y_n)\, dp \end{aligned}$$

One can check that for any $j > n$ we have

$$\begin{aligned} E(y_j \mid y_1, \ldots, y_n) &= E(y_{n+1} \mid y_1, \ldots, y_n) \\ &= \frac{\alpha + \sum_{i=1}^n y_i}{\alpha + \beta + n} \end{aligned}$$

So the Bayes estimate of the population total given $y_1, \ldots, y_n$ is

$$\begin{aligned} E(\sum_{i=1}^N y_i \mid y_1, \ldots, y_n) &= \sum_{i=1}^n y_i + \sum_{j=n+1}^N E(y_j \mid y_i, \ldots, y_n) \\ &= \sum_{i=1}^n y_i + (N-n)\frac{\alpha + \sum_{i=1}^n y_i}{\alpha + \beta + n} \end{aligned}$$

The store example

Consider the store example from the beginning of chapter 6 in the text.

Store	Size a_i	Sales y_i	y_i/a_i
1	100	11,000	110
2	200	20,000	100
3	300	24,000	80
4	1000	245,000	245

For the probability model we assume

$$\begin{aligned}(y_i/a_i)'\text{s} \mid \lambda \sim &\quad \text{are independent r.v.'s with } y_i/a_i \sim \text{Normal}(\lambda, v_i)\\ \lambda \sim &\ \text{Normal}(m', v')\end{aligned}$$

where the v_i's, m' and v' are constants that must be specified by the statistician. Under this model the prior expectation of y_i is

$$E(y_i/a_i) = E(E(y_i/a_i \mid \lambda)) = E(\lambda) = m' \quad \text{or} \quad E(y_i) = a_i m'$$

From this we see that the prior expectation of the population total is $(a_1 + a_2 + a_3 + a_4) \times m'$.

Suppose the statistician can select one store to observe its y_i value and then must estimate $\sum_{i=1}^{4} y_i$. We now find the Bayes estimate of the population total under this model.

For ease of notation assume

$$y/a \sim \text{Normal}(\lambda.v) \quad \text{and} \quad \lambda \sim \text{Normal}(m', v')$$

where a is a fixed known constant. Now

$$f(\lambda \mid y) \propto f(y \mid \lambda)\, f(\lambda)$$

which, as we shall see, is a normal distribution. For this reason we only need to look at the exponent of the above when finding this posterior distribution. The exponent will be a quadratic function of λ for which we will complete the square and thus identifying the particular normal distribution. Note

$$\text{exponent of } \ f(y \mid \lambda)\, f(\lambda) = -\frac{1}{2}\Big\{\frac{(\lambda - y/a)^2}{v} + \frac{(\lambda - m')^2}{v'}\Big\}$$

and then

$$\begin{aligned}\Big\{\quad\Big\} &= \frac{\lambda^2 - 2(y/a)\lambda + (y/a)^2}{v} + \frac{\lambda^2 - 2m'\lambda + (m')^2}{v'}\\ &= (\frac{1}{v} + \frac{1}{v'})\lambda^2 - 2\big(\frac{y/a}{v} + \frac{m'}{v'}\big) + \text{terms not depending on } \lambda\\ &= \frac{v + v'}{vv'}\lambda^2 - 2\big(\frac{v'(y/a) + vm'}{vv'}\big)\lambda + \text{terms not depending on } \lambda\\ &= \frac{v + v'}{vv'}\Big(\lambda^2 - 2\frac{v'(y/a) + vm'}{v + v'}\lambda\Big) + \text{terms not depending on } \lambda\\ &= \frac{v + v'}{vv'}\Big(\lambda - \frac{v'(y/a) + vm'}{v + v'}\Big)^2 + \text{terms not depending on } \lambda\end{aligned}$$

So we see that

$$\lambda \mid y \sim \text{Normal}\Big(\frac{v'(y/a) + vm'}{v + v'}, \frac{v}{v + v'}v'\Big)$$

If y_4 is observed then the posterior mean of λ is

$$m'' = \frac{v'(y_4/a_4) + v_4 m'}{v_4 + v'}$$

Since

$$f(y_1, y_2, y_3 \mid y_4) = \int_{-\infty}^{\infty} f(\frac{y_1}{a_1}, \frac{y_2}{a_2}, \frac{y_3}{a_3} \mid \lambda) f(\lambda \mid y_4)\, d\lambda$$

we see for $i = 1, 2, 3$ the posterior expectation of y_i is $a_i m''$ and the Bayes estimate of the population total is

$$y_4 + (a_1 + a_2 + a_3)m''$$

Looking at the values of the y_i/a_i's in the table (note that this would not be known in practice but some prior information about sales could be available) one possible sensible choice for the parameters are

$$v_i = a_i \quad m' = 150 \quad v' = 2500$$

With these values the prior expectation of the population's total sales is 240,00 while the Bayes estimate of total sales becomes 375,714.

If instead we had observed y_3 then the posterior mean is

$$m'' = \frac{v'(y_3/a_3) + a_3 m'}{a_3 + v'} = 87.5$$

and the Bayes estimate becomes

$$est = 24,000 + (100 + 200 + 1,000) \times (87.5) = 150,750$$