Bayesian Statistics

2: Exchangeability and Conjugate Modeling

David Draper

Department of Applied Mathematics and Statistics
University of California, Santa Cruz

draper@ams.ucsc.edu

http://www.cse.ucsc.edu/~draper

© 2003 David Draper (all rights reserved)
Contents

2.1 Probability as quantification of uncertainty about observables. Binary outcomes. Case Study: Hospital-specific prediction of patient-level mortality rates

2.2 Review of frequentist modeling and maximum-likelihood inference

2.3 Exchangeability as a Bayesian concept parallel to frequentist independence

2.4 Prior, posterior, and predictive distributions

2.5 Inference and prediction. Coherence and calibration

2.6 Conjugate analysis. Comparison with frequentist modeling

2.7 Integer-valued outcomes; Poisson modeling. Case Study: Hospital length of stay for birth of premature babies

2.8 Continuous outcomes; Gaussian modeling. Multivariate unknowns; marginal posterior distributions. Case Study: Measurement of physical constants

2.9 The exponential family; conjugate priors

2.10 References
Introduction to Bayesian Modeling

2.1 Probability as quantification of uncertainty about observables. Binary outcomes

Case Study: Hospital-specific prediction of mortality rates. Let’s say you’re interested in measuring the quality of care (e.g., Kahn et al., 1990) offered by one particular hospital.

I am thinking of the Dominican Hospital (DH) in Santa Cruz, CA; you may have a different hospital in mind.

As part of this you decide to examine the medical records of all patients treated at the DH in one particular time window, say January 2000–December 2003, for one particular medical condition for which there is a strong process-outcome link, say acute myocardial infarction (AMI; heart attack).

(Process is what health care providers do on behalf of patients; outcomes are what happens as a result of that care.)

In the time window you’re interested in there will be about $n = 400$ AMI patients at the DH.
The Meaning of Probability

To keep things simple let’s ignore process for the moment and focus here on one particular outcome: death status (mortality) as of 30 days from hospital admission, coded 1 for dead and 0 for alive.

(In addition to process this will also depend on the sickness at admission of the AMI patients, but let’s ignore that initially too.)

From the vantage point of December 1999, say, what may be said about the roughly 400 1s and 0s you will observe in 2000–03?

The meaning of probability. You are definitely uncertain about the 0–1 death outcomes $Y_1, \ldots, Y_n$ before you observe any of them.

Probability is supposed to be the part of mathematics concerned with quantifying uncertainty; can probability be used here?

In part 1 I argued that the answer was yes, and that three types of probability—classical, frequentist, and Bayesian—are available (in principle) to quantify uncertainty like that encountered here.
2.2 Review of Frequentist Modeling

I will focus on the approaches with the most widespread usage—frequentist and Bayesian—in what follows.

How can the frequentist definition of probability be applied to the hospital mortality problem?

By definition the frequentist approach is based on the idea of hypothetical or actual repetitions of the process being studied, under conditions that are as close to independent identically distributed (IID) sampling as possible.

When faced with a data set like the 400 1s and 0s \((Y_1, \ldots, Y_n)\) here, the usual way to do this is to think of it as a random sample, or like a random sample, from some population that is of direct interest to you.

Then the randomness in your probability statements refers to the process of what you might get if you were to repeat the sampling over and over—the \(Y_i\) become random variables whose probability distribution is determined by this hypothetical repeated sampling.
Frequentist Modeling (continued)

Population

30-day mortality

$N = \binom{1}{i,j} \cdots$

Sample

The observed data

30-day mortality

$\bar{Y} = \hat{\pi} = 0.18$ (say)

$\bar{Y} = \hat{\pi} = 0.21$ (say)

$\mu = \hat{\pi}$

$\sigma = \sqrt{\hat{\pi}(1-\hat{\pi})}$

Here SRS = simple random sampling at random without replacement; when $n < < N$ ($n$ is a lot smaller than $N$), SRS $\equiv$ IID $= \text{at random with replacement}$.
Frequentist Modeling (continued)

On the previous page SD stands for standard deviation, the most common measure of the extent to which the observations $y_i$ in a data set vary, or are spread out, around the center of the data.

The center is often measured by the mean $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$, and the SD of a sample of size $n$ is then given by

$$SD = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2}.$$  \hfill (1)

The population size is denoted by $N$; this is often much larger than the sample size $n$.

With 0/1 (dichotomous) data, like the mortality outcomes in this case study, the population mean $\mu$ simply records the proportion $p$ of 1s in the population (check this), and similarly the sample mean $\bar{y}$ keeps track automatically of the observed death rate $\hat{p}$ in the sample.

As $N \to \infty$ the population SD $\sigma$ with 0/1 data takes on a simple form (check this):

$$\sigma = \sqrt{p(1-p)}.$$  \hfill (2)

It's common in frequentist modeling to make a notational distinction between the random variables $Y_i$ (the placeholders for the process of making IID draws from the population over and over) and the values $y_i$ that the $Y_i$ might take on (although I will abuse this notation with $\hat{p}$ below).
Frequentist Modeling (continued)

In the diagram on page 6 the relationship between the population and the sample data sets can be usefully considered in each of two directions:

- If the population is known you can think about how the sample is likely to come out under IID sampling—this is a probability question.

Here in this case $p$ would be known and you’re trying to figure out the random behavior of the sample mean $\bar{Y} = \hat{p}$.

- If instead only the sample is known your job is to infer the likely composition of the population that could have led to this IID sample—this is a question of statistical inference.

In this problem the sample mean $\bar{Y} = \hat{p}$ would be known and your job would be to estimate the population mean $p$.

Suppose that $N \gg n$, i.e., that even if SRS was used you are effectively dealing with IID sampling.

Intuitively both SRS and IID should be “good”—representative—sampling methods, so that $\hat{p}$ should be a “good” estimate of $p$, but what exactly does the word “good” mean in this sentence?

Evidently a good estimator $\hat{p}$ would be likely to be close to the truth $p$, especially with a lot of data (i.e., if $n$ is large).

In the frequentist approach to inference quantifying this idea involves imagining how $\hat{p}$ would have come out if the process by which the observed $\hat{p} = 0.18$ came to you were repeated under IID conditions.

This gives rise to the imaginary data set, the third part of the diagram on page 6: we consider all possible $\hat{p}$ values based on an IID sample of size $n$ from a population with $100p\%$ 1s and $100(1-p)\%$ 0s.
Frequentist Modeling (continued)

Let \( M \) be the **number of hypothetical repetitions** in the imaginary data set.

The long-run mean (as \( M \to \infty \)) of these imaginary \( \hat{p} \) values is called the **expected value** of the random variable \( \hat{p} \), written \( E(\hat{p}) \) or \( E_{\text{IID}}(\hat{p}) \) to emphasize the mechanism of drawing the sample from the population.

The long-run SD of these imaginary \( \hat{p} \) values is called the **standard error** of the random variable \( \hat{p} \), written \( SE(\hat{p}) \) or \( SE_{\text{IID}}(\hat{p}) \).

It's natural in studying how the hypothetical \( \hat{p} \) values vary around the center of the imaginary data set to make a **histogram** of these values: this is a plot with the possible values of \( \hat{p} \) along the horizontal scale and the frequency with which \( \hat{p} \) takes on those values on the vertical scale.

It's helpful to draw this plot on the **density scale**, which just means that the vertical scale is chosen so that the total area under the histogram is 1.

The long-run histogram of the imaginary \( \hat{p} \) values on the density scale is called the **(probability) density** of the random variable \( \hat{p} \).

The values of \( E(\hat{p}) \) and \( SE(\hat{p}) \), and the basic shape of the density of \( \hat{p} \), can be determined **mathematically** (under IID sampling) and verified by **simulation**.

It turns out that

\[
E_{\text{IID}}(\hat{p}) = p \quad \text{and} \quad SE_{\text{IID}}(\hat{p}) = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}, \tag{3}
\]

and the density of \( \hat{p} \) for large \( n \) is well approximated by the **normal curve** or **Gaussian distribution** (this result is the famous **Central Limit Theorem (CLT)**).
Frequentist Modeling (continued)

Suppose the sample of size $n = 400$ had 72 1s and 328 0s, so that $\hat{p} = \frac{72}{400} = 0.18$.

Thus you would estimate that the population mortality rate $p$ is around 18%, but how much uncertainty should be attached to this estimate?

The above standard error formula is not directly usable because it involves the unknown $p$, but we can estimate the standard error by plugging in $\hat{p}$:

$$\hat{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(0.18)(0.82)}{400}} \approx 0.019. \quad (4)$$

In other words, I think $p$ is around 18%, give or take about 1.9%.

A probabilistic uncertainty band can be obtained with the frequentist approach by appeal to the CLT, which says that (for large $n$) in repeated sampling $\hat{p}$ would fluctuate around $p$ like draws from a normal curve with mean $p$ and SD (SE) 0.019, i.e.,

$$0.95 = P_F \left[ p - 1.96 \hat{SE}(\hat{p}) \leq \hat{p} \leq p + 1.96 \hat{SE}(\hat{p}) \right]$$

$$= P_F \left[ \hat{p} - 1.96 \hat{SE}(\hat{p}) \leq p \leq \hat{p} + 1.96 \hat{SE}(\hat{p}) \right]. \quad (5)$$

Thus (Neyman 1923) a 95% (frequentist) confidence interval for $p$ runs from $\hat{p} - 1.96 \hat{SE}(\hat{p})$ to $\hat{p} + 1.96 \hat{SE}(\hat{p})$, which in this case is from $0.180 - (1.96)(0.019) = 0.142$ to $0.180 + (1.96)(0.019) = 0.218$, i.e., I am "95% confident that $p$ is between about 14% and 22%".

But what does this mean?