

STAT 238 - Bayesian Statistics

Lecture Ten

Spring 2026, UC Berkeley

Aditya Guntuboyina

11 Feb 2026

1 Additional Comments on the Kidney Cancer Data Analysis

For the kidney cancer dataset $(X_i, n_i), i = 1, \dots, N$ (N denotes the total number of counties, n_i is the average population of county i and X_i is the number of deaths due to kidney cancer in the fixed time period 1980 – 89). Consider the following three models for this dataset:

1. **Model One:** $\theta_i \stackrel{\text{i.i.d}}{\sim} \text{Beta}(0, 0)$ and $X_i | \theta_i \stackrel{\text{ind}}{\sim} \text{Bin}(n_i, \theta_i)$. This model uses an uninformative prior on θ_i , and the posterior mean estimate of each θ_i coincides with the frequentist MLE X_i/n_i . This model performs poorly for the present dataset. In particular, rankings of counties by estimated risk (e.g., the top 100 values of $\hat{\theta}_i$) are dominated by counties with small populations. Moreover, the model yields inaccurate predictions for future county-level death counts.
2. **Model Two:** $\theta_i \stackrel{\text{i.i.d}}{\sim} \text{Beta}(15, 150000)$ and $X_i | \theta_i \stackrel{\text{ind}}{\sim} \text{Bin}(n_i, \theta_i)$. This model employs a highly informative prior on θ_i , where the hyperparameters (15, 150000) may be viewed as arising from a Beta distribution fitted to historical data. The prior encodes strong prior knowledge about the prevalence of kidney cancer mortality. This leads to substantial shrinkage of the posterior estimates towards the prior mean. Unsurprisingly, given the strength of the prior information, this model yields good empirical performance on the dataset.
3. **Model Three:** $\log a, \log b \stackrel{\text{i.i.d}}{\sim} \text{uniform}(-\infty, \infty)$, $\theta_i \stackrel{\text{i.i.d}}{\sim} \text{Beta}(a, b)$, $X_i | \theta_i \stackrel{\text{ind}}{\sim} \text{Bin}(n_i, \theta_i)$. This model places an uninformative prior on the hyperparameters a and b , and hence does not inject strong prior knowledge about the prevalence of kidney cancer mortality. Compared to Models One and Two, this is a substantially more flexible and principled model, as it allows the amount of shrinkage to be learned from the data rather than fixed a priori.

When fitted to the data, the posterior distribution of (a, b) concentrates sharply around values close to (15, 150000) (see the code file for this lecture), indicating that the model successfully infers from the data that kidney cancer deaths are rate events. As a consequence, the posterior estimates of the individual θ_i are adaptively shrunk toward a common mean, with the degree of shrinkage depending on the corresponding population sizes n_i . This hierarchical borrowing of strength leads to stable estimation, sensible county-level rankings, and accurate predictions of future death counts, and overall the model performs very well.

Note that this model does not require any a priori knowledge specific to kidney cancer and is therefore broadly applicable. For example, the same hierarchical framework can be used to predict batting averages in the baseball example, which will be explored in Homework Two.

2 Normal Likelihoods

We will next discuss Bayesian inference with normal likelihoods, and also study some connections to the James-Stein estimator. Normal likelihoods are applicable even in Binomial situations, such as in the baseball data analysis example from Efron [1, Chapter 1]. Here X_i denotes the number of hits in $n_i = n = 45$ at bats for player i . The natural model is: $X_i \sim \text{Bin}(n_i, p_i)$ with parameter p_i but a normal likelihood (as opposed to Binomial likelihood) can also be used because by invoking the Central Limit Theorem, we can write

$$\frac{X_i}{n_i} \underset{\text{approx}}{\sim} N\left(p_i, \frac{p_i(1-p_i)}{n_i}\right).$$

This is because:

$$\sqrt{n} \left(\frac{X_i}{n_i} - p_i \right) \xrightarrow{\text{Law}} N(0, p_i(1-p_i)) \text{ as } n \rightarrow \infty. \quad (1)$$

Note that the unknown parameter p_i appears in the variance above. If we want to work with normal likelihoods with known variance, a clean way is to use a variance stabilizing transformation. This is justified by the use of Delta method.

Theorem 2.1 (Delta Method). *If $\sqrt{n}(T_n - p) \xrightarrow{\text{Law}} N(0, \tau^2)$ as $n \rightarrow \infty$, then*

$$\sqrt{n} (g(T_n) - g(p)) \xrightarrow{\text{Law}} N(0, \tau^2 (g'(p))^2)$$

as $n \rightarrow \infty$ provided $g'(p)$ exists and is non-zero.

Informally, the Delta method states that if T_n has a limiting Normal distribution, then $g(T_n)$ also has a limiting normal distribution and also gives an explicit formula for the asymptotic variance of $g(T_n)$. This is surprising because g can be linear or non-linear. In general, non-linear functions of normal random variables do not have a normal distribution. But the Delta method works because under the assumption that $\sqrt{n}(T_n - p) \xrightarrow{\text{Law}} N(0, \tau^2)$, it follows that T_n converges in probability to p so that T_n will be close to p at least for large n . In a neighborhood of p , the non-linear function g can be approximated by a linear function which means that g effectively behaves like a linear function. Indeed, the Delta method is a consequence of the approximation:

$$g(T_n) - g(p) \approx g'(p) (T_n - p).$$

By the Delta method and (1), we have

$$\sqrt{n_i} (g(X_i/n_i) - g(p_i)) \xrightarrow{\text{Law}} N(0, (g'(p_i))^2 p_i(1-p_i)).$$

The variance above will not depend on p_i if we choose the function g so that

$$g'(p_i) = \frac{1}{\sqrt{p_i(1-p_i)}}$$

Solving this for g , we get $g(p_i) = 2 \arcsin(\sqrt{p_i})$. This is a variance stabilizing transformation for the Binomial. Thus by the Delta method, we have

$$2\sqrt{n} \left(\arcsin(\sqrt{X_i/n_i}) - \arcsin(\sqrt{p_i}) \right) \xrightarrow{\text{Law}} N(0, 1).$$

While working with binomial counts, it thus makes sense to transform the observation X_i and parameter p_i into

$$Y_i := 2\sqrt{n_i} \arcsin(\sqrt{X_i/n_i}) \quad \text{and} \quad \theta_i := 2\sqrt{n_i} \arcsin(\sqrt{p_i})$$

respectively and then work with the likelihood

$$Y_i | \theta_i \sim N(\theta_i, 1).$$

We will study Bayesian estimation under this model in the next lecture.

References

- [1] Efron, B. (2010) *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*. Cambridge University Press, 2010.