Empirical Bayesian Analysis Applied to Manufacturing

Daniel Woodie

Problem Statement and Setup

I manage the quality of fishing lines that go through multiple factories. Each fishing line gets tested before ever being packaged and shipped out to a customer. I'd like to understand failure rates to better understand overall quality of my lines in the factories. Failures and passes can be considered a binomial distribution. It takes the form:

$$L\big(p\big) = \frac{n!}{x!(n-x!)}p^y(1-p)^{n-y}$$

In this case, p is the probability of a fishing line failing. The conjugate prior for this scenario is a Beta distribution.

$$\pi\big(p\big) = \frac{1}{B\big(\alpha\big)}p^{\alpha-1}(1-p)^{\beta-1}$$

$$p \sim Beta(\alpha, \beta)$$
$$y_1, ..., y_K \sim Binom(n, p)$$

Posterior

$$\propto p^{\alpha+y-1}(1-p)^{\beta+n-y-1}$$

The posterior, can be summarized as:

$$Beta(\alpha + y, \beta + n - y)$$

The expected value for this distribution is:

$$E[p] = \frac{\alpha + y}{\alpha + \beta + n}$$

Sample Calculation

Let's look at two hypothetical fishing lines. First, let's set α and β.

$$p \sim Beta(\alpha = 14, \beta = 6)$$

Therefore,

$$E[p] = \frac{\alpha}{\alpha + \beta} = \frac{14}{14 + 6} = .7$$

When you update this estimate for a fishing line that has 4 failures out of 11 this becomes:

$$E[p] = \frac{\alpha + 4}{\alpha + \beta + 4 + 7} = \frac{14 + 4}{14 + 6 + 4 + 7} = \frac{18}{31} = .58$$

What about a fishing line that has 4000 failures out of 11000:

$$E[p] = \frac{\alpha + 4000}{\alpha + \beta + 4000 + 7000} = \frac{18 + 4000}{14 + 6 + 4000 + 7000} = \frac{4018}{11021} = .365$$

The Plot Thickens

One aspect that we haven't considered here is that different factories may have different failure rates. For example, the quality of a particular fishing line could be great at Factory A but at Factory B the quality is absolutely terrible. As such, it may be important to evaluate whether improving quality issues should be focused on a fishing line across the entire supply chain or at a specific factory.

Defining the Linear Mixed Model

Y_{ij} is the failure count in Factory i for Line j. N_{ij} is number of units received. Model is something like:

$$Y_{ij} \sim Binom(p = \mu + F_i + L_j, \ n = N_i j)$$

$$where \ \mu \sim Normal(...)$$

$$F \sim Normal(0, \sigma_F), \ \sigma_F \ HalfNormal(...)$$

$$L \sim Normal(0, \sigma_L), \sigma_L \ HalfNormal(...)$$

As an aside here, I think it might make more sense to treat the priors on lines, factories, and the combination therein as betas instead of normals centered on zero.

To vary this over time, we simply restructure the model to be of the form:

$$\pi_n(\theta_j) \propto \pi(\theta_j) L(\theta_j)$$

where you're calculating a new θ for each time point j.

Background

Maximum Likelihood Estimation

You have some density where is x is a function of θ:

$$f(x|\theta)$$

and some data that comes from this density:

$$x_1, x_2, ..., x_n$$

θ can be estimated by maximizing the likelihood function:

$$\hat{\theta} = arg \max_{\theta} L(\theta) = arg \max_{\theta} \Pi_{i=1}^{n} f(x_i|\theta)$$

Bayesian Statistics Bayes Theorem:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

Traditional Statistics:

$$P(X_1, ..., X_n|\theta)$$

Bayesian Statistics:

$$P(\theta|X_1, ..., X_n)$$

You use Bayes theorem then to come up with an estimate:

$$P(\theta|X_1, ..., X_n) = \frac{P(\theta)P(X_1, ..., X_n|\theta)}{P(X_1, ..., X_n)}$$

Because the denominator does not involve θ, most people simplify this to be:

$$P(\theta|X_1, ..., X_n) \propto P(\theta)P(X_1, ..., X_n|\theta)$$

Let's say, then, you have some prior knowledge about where θ arises:

$$\pi(\theta)$$

This information can be included and inferences can be made on the resulting distribution, the posterior:

$$\pi_n(\theta) \propto \pi(\theta) L(\theta)$$

Now your estimate for θ becomes:

$$\hat{\theta} = arg \max_{\theta} \pi_n(\theta)$$

Stars can be considered a multinomial distribution. The multinomial distribution is a generalization of the binomial distribution. Instead of 2 outcomes (0 or 1), there are k outcomes. It takes the form:

$$L\left(p_1, ..., p_k\right) = \frac{n!}{x_1!...x_k!} p_1^{y_1} \times ... \times p_k^{y_k}$$

In this case, p_1 is the probability of getting a 1 star rating and p_k is the probability of getting a k star rating. The conjugate prior for this scenario is a Dirichlet distribution with k different parameters.

$$\pi\left(p\right) = \frac{1}{B\left(\alpha\right)} \Pi_{i=1}^k p_i^{\alpha_i - 1}$$

$$p_1, ..., p_K \sim Dir(\alpha_1, ..., \alpha_K)$$
$$y_1, ..., y_K \sim Mult(p_1, ..., p_K)$$

Posterior

$$\propto \Pi_{j=1}^K p_j^{a_j - 1} \Pi_{y_i}^n \Pi_{j=1}^K p_j^{y_i^{(j)}} = \Pi_{j=1}^K p_j^{\alpha_j - 1 + \sum_{y_i=1}^n y_i^{(j)}}$$

This density is exactly that of a Dirichlet distribution, except we have

$$\alpha_j^{'} = \alpha_j + \sum_{y_i=1}^n y_i^{(j)}$$

The posterior, can be summarized as:

$$Dir(\alpha_1^{'}, ..., \alpha_K^{'})$$