L E C T U R E 8

Example 18. White pine is one of the best known species of pines in the northeastern United States and Canada. White pine is susceptible to blister rust, which develops cankers on the bark. These cankers swell, resulting in death of twigs and small trees. A forester wishes to estimate the average number of diseased pine trees per acre in a forest.

Solution: The number of diseased trees per acre can be modeled by a Poisson distribution with mean θ. Since θ changes from area to area, the forester believes that $\theta \sim Exp(\lambda)$. Thus,

$$p(\theta) = \left(\frac{1}{\lambda}\right) e^{-\theta/\lambda}, \quad \text{if } \theta > 0, \quad \text{and } 0 \text{ elsewhere.}$$

The forester takes a random sample of size n from n different one-acre plots. The likelihood function is

$$p(y|\theta) = \prod_{i=1}^{n} \frac{\theta^{y_i}}{y_i!} e^{-\theta} = \frac{\theta^{\sum_{i=1}^{n} y_i}}{\prod y_i!} e^{-n\theta}.$$

Consequently, the posterior distribution is

$$p(\theta|y) = \frac{\theta^{\sum_{i=1}^{n} y_i} e^{-\theta(n+1/\lambda)}}{\int_0^\infty \theta^{\sum_{i=1}^{n} y_i} e^{-\theta(n+1/\lambda)} d\theta}.$$

We see that this is a Gamma-distribution with parameters $\alpha = \sum_{i=1}^{n} y_i + 1$ and $\beta = n + 1/\lambda$. Thus,

$$p(\theta|y) = \frac{(n+1/\lambda)^{\sum_{i=1}^{n} y_i + 1}}{\Gamma(\sum_{i=1}^{n} y_i + 1)} \theta^{\sum_{i=1}^{n} y_i} e^{-\theta(n+1/\lambda)}.$$

The Gamma Distributions.

The *gamma distribution* is important because it includes a wide class of specific distributions, some of which underlie fundamental statistical procedures. In addition to serving as a utility distribution, the gamma provides probabilities for yet another random variable associated with Poisson processes (the exponential distribution itself is a member of the gamma distributions).

The Gamma Function.

In order to describe the gamma distribution in detail, we must first consider a useful function, Gamma Function:

$$\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1}\, e^{-x}\, dx, \qquad \alpha > 0.$$

The symbol Γ (Greek uppercase *gamma*) is reserved for this function. The integration by parts of $\Gamma(\alpha)$ yields that

$$\Gamma(\alpha+1) = \alpha\,\Gamma(\alpha).$$

Note, that for any nonnegative integer k we have

$$\Gamma(k+1) = k!,$$

In particular, $\Gamma(1) = 1$.

An important class involves values with halves. We have

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi},$$

and for any positive integer k

$$\Gamma\left(k+\frac{1}{2}\right) = \frac{(2k-1)!!}{2^k}\,\sqrt{\pi},$$

where $(2k-1)!! = 1\cdot 3\cdot 5\cdot \ldots \cdot (2k-1)$.

The Density Function of Gamma Random Variable

The following expression gives the density function for a gamma distribution.

$$f(x) = \begin{cases} 0 & \text{if} \quad x \le 0, \\ \dfrac{\lambda^\alpha}{\Gamma(\alpha)}\, x^{\alpha-1}\, e^{-\lambda\, x} & \text{if} \quad x > 0. \end{cases}$$

The two parameters λ and α may be any positive values ($\lambda > 0$ and $\alpha > 0$).

A special case of this function occurs when $\alpha = 1$. We have

$$f(x) = \begin{cases} 0 & \text{if} \quad x \le 0, \\ \lambda\, e^{-\lambda\, x} & \text{if} \quad x > 0 \end{cases}$$

which is the density function for the exponential distribution.

The expectation and Variance of gamma distribution have the forms:

$$E\eta = \frac{\alpha}{\lambda} \qquad \text{and} \qquad \text{Var}\,(\eta) = \frac{\alpha}{\lambda^2}.$$

When α is natural, say $\alpha = n$, the gamma distribution with parameters (n, λ) often arises in practice:

$$f(x) = \begin{cases} 0 & \text{if} \quad x \le 0, \\ \dfrac{\lambda^n}{(n-1)!}\, x^{n-1}\, e^{-\lambda\, x} & \text{if} \quad x > 0. \end{cases}$$

This distribution is often referred to in the literature as the n-Erlang distribution. Note that when $n = 1$, this distribution reduces to the exponential.

The gamma distribution with $\lambda = 1/2$ and $\alpha = n/2$ (n is natural) is called χ^2_n (read "chi-squared") distribution with n degrees of freedom:

$$f(x) = \begin{cases} 0 & \text{if} \quad x \le 0, \\ \dfrac{1}{2^{n/2}\,\Gamma(n/2)}\, x^{n/2-1}\, e^{-x/2} & \text{if} \quad x > 0. \end{cases}$$

We have

$$E\chi^2_n = n \qquad \text{and} \qquad \text{Var}\chi^2_n = 2n.$$

The Beta Distribution.

A random variable is said to have a beta distribution if its density is given by

$$f(x) = \begin{cases} \dfrac{1}{B(a,b)}\, x^{a-1}\,(1-x)^{b-1} & \text{if} \quad 0 < x < 1, \\ 0 & \text{otherwise,} \end{cases}$$

where

$$B(a,b) = \int_0^1 x^{a-1}\,(1-x)^{b-1}\, dx.$$

Note that when $a = b = 1$ the beta density is the uniform density. When a and b are greater than 1 the density is bell-shaped, but when they are less than 1 it is U-shaped. When $a = b$, the beta density is symmetric about $\frac{1}{2}$. When $b > a$, the density is skewed to the left (in the sense that smaller values become more likely), and it is skewed to the right when $a > b$. The following relationship exists between beta and gamma functions:

$$B(a,b) = \frac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)}.$$

The expectation and Variance of beta distribution have the forms:

$$E\eta = \frac{a}{a+b} \qquad \text{and} \qquad \text{Var}\,(\eta) = \frac{a\,b}{(a+b)^2\,(a+b+1)}.$$

THE POISSON MODEL

Some measurements, such as a persons number of children or number of friends, have values that are whole numbers. In these cases our sample space is $Y = \{0, 1, 2, ...\}$. Perhaps the simplest probability model on Y is the Poisson model.

Poisson distribution.

Recall that a random variable Y has a Poisson distribution with mean θ if

$$P(Y = y|\theta) = Pois(y, \theta) = \theta^y e^{-\theta}/y! \quad \text{for } y \in \{0, 1, 2, ...\}.$$

For such a random variable,

$$E[Y|\theta] = \theta; \qquad Var[Y|\theta] = \theta.$$

People some times say that the Poisson family of distributions has a mean variance relationship because if one Poisson distribution has a larger mean than another, it will have a larger variance as well.

14.2.1 POSTERIOR INFERENCE.

If we model $Y_1, \ldots, Y_n$ as i.i.d. Poisson with mean θ, then the joint pdf of our sample data is as follows:

$$P(Y_1 = y_1, \ldots, Y_n = y_n|\theta) = \prod_{i=1}^{n} p(y_i|\theta) = \prod_{i=1}^{n} \frac{1}{y_i!}\theta^{y_i}e^{-\theta} = c(y_1, \ldots, y_n)\theta^{\sum_{i=1}^n y_i}e^{-n\theta}.$$

Comparing two values of θ a posteriori, we have

$$\frac{p(\theta_a|y_1, \ldots, y_n)}{p(\theta_b|y_1, \ldots, y_n)} = \frac{c(y_1, \ldots, y_n)e^{-n\theta_a}\theta_a^{\sum y_i}p(\theta_a)}{c(y_1, \ldots, y_n)e^{-n\theta_b}\theta_b^{\sum y_i}p(\theta_b)} = \frac{e^{-n\theta_a}\theta_a^{\sum y_i}p(\theta_a)}{e^{-n\theta_b}\theta_b^{\sum y_i}p(\theta_b)}$$

As in the case of the i.i.d. binary model, $\sum_{i=1}^n Y_i$ contains all the information about θ that is available in the data, and again we say that $\sum_{i=1}^n Y_i$ is a sufficient statistic. Furthermore, $\{\sum_{i=1}^n Y_i|\theta\} \sim \text{Poisson}(n\theta)$.

Conjugate prior.

For now we will work with a class of conjugate prior distributions that will make posterior calculations simple. Recall that a class of prior densities is conjugate for a sampling model $p(y_1, \ldots, y_n|\theta)$ if the posterior distribution is also in the class. For the Poisson sampling model, our posterior distribution for θ has the following form:

$$p(\theta|y_1, \ldots, y_n) \propto p(\theta) \times p(y_1, \ldots, y_n|\theta) \propto p(\theta) \times \theta^{\sum y_i} e^{-n\theta}.$$

This means that whatever our conjugate class of densities is, it will have to include terms like $\theta^{c_1} e^{-c_2\theta}$ for numbers c_1 and c_2. The simplest class of such densities includes only these terms, and their corresponding probability distributions are known as the family of gamma distributions.

Gamma distribution.

An uncertain positive quantity θ has a $gamma(a, b)$ distribution if

$$p(\theta) = gamma(\theta, a, b) = \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta}, \quad \text{for } \theta, a, b > 0.$$

For such a random variable,

$$E[\theta] = a/b; \quad Var[\theta] = a/b^2; \qquad mode[\theta] = \begin{cases} (a-1)/b, & \text{if } a > 1 \\ 0, & \text{if } a \leq 1 \end{cases}.$$

Posterior distribution of θ.

Suppose $Y_1, \ldots, Y_n|\theta \sim$ i.i.d. Poisson(θ) and $p(\theta) = gamma(\theta, a, b)$. Then

$$p(\theta|y_1, \ldots, y_n) = p(\theta) \times p(y_1, \ldots, y_n|\theta)/p(y_1, \ldots, y_n) = \{\theta^{a-1} e^{-b\theta}\} \times \{\theta^{\sum y_i} e^{-n\theta}\} \times c(y_1, \ldots, y_n, a, b)$$

$$= \{\theta^{a+\sum y_i - 1} e^{-(b+n)\theta}\} \times c(y_1, \ldots, y_n, a, b).$$

This is evidently a gamma distribution, and we have confirmed the conjugacy of the gamma family for the Poisson sampling model:

$$\theta \sim gamma(a, b), \quad Y_1, \ldots, Y_n|\theta \sim Poisson(\theta) \Rightarrow \{\theta|Y_1, \ldots, Y_n\} \sim gamma(a + \sum_{i=1}^{n} Y_i, b + n).$$

Estimation and prediction proceed in a manner similar to that in the binomial model. The posterior expectation of θ is a convex combination of the prior expectation and the sample average:

$$E[\theta|y_1,\ldots,y_n] = \frac{a+\sum y_i}{b+n} = \frac{b}{b+n}\,\frac{a}{b} + \frac{n}{b+n}\,\frac{\sum y_i}{n}.$$

b is interpreted as the number of prior observations;

a is interpreted as the sum of counts from b prior observations.

For large n, the information from the data dominates the prior information:

$$n >> b \Rightarrow E[\theta|y_1,\ldots,y_n] \approx \overline{y}, \quad Var[\theta|y_1,\ldots,y_n] \approx \frac{\overline{y}}{n}.$$

Predictions about additional data can be obtained with the posterior predictive distribution:

$$p(\tilde{y}|y_1,\ldots,y_n) = \int_0^\infty p(\tilde{y}|\theta,y_1,\ldots,y_n)\,p(\theta|y_1,\ldots,y_n)d\theta = \int_0^\infty p(\tilde{y}|\theta)\,p(\theta|y_1,\ldots,y_n)d\theta$$

$$= \int Pois(\tilde{y},\theta)\,gamma(\theta, a+\sum y_i, b+n)\,d\theta$$

$$= \int \left\{\frac{1}{\tilde{y}!}\theta^{\tilde{y}}e^{-\theta}\right\}\left\{\frac{(b+n)^{a+\sum y_i}}{\Gamma(a+\sum y_i)}\theta^{a+\sum y_i-1}\,e^{-(b+n)\theta}\right\}d\theta =$$

$$\frac{(b+n)^{a+\sum y_i}}{\Gamma(\tilde{y}+1)\,\Gamma(a+\sum y_i)}\int_0^\infty \theta^{a+\sum y_i+\tilde{y}-1}\,e^{-(b+n+1)\theta}d\theta.$$

Evaluation of this complicated integral looks daunting, but it turns out that it can be done without any additional calculus. Lets use what we know about the gamma density:

$$1 = \int_0^\infty \frac{b^a}{\Gamma(a)}\theta^{a-1}\,e^{-b\theta}\,d\theta \quad \text{for any values } \; a,b>0.$$

This means that

$$\int_0^\infty \theta^{a-1}\,e^{-b\theta}\,d\theta = \frac{\Gamma(a)}{b^a} \quad \text{for any values } \; a,b>0.$$

Now substitute in $a+\sum y_i+\tilde{y}$ instead of a and $b+n+1$ instead of b to get

$$\int_0^\infty \theta^{a+\sum y_i+\tilde{y}-1}\,e^{-(b+n+1)\theta}\,d\theta = \frac{\Gamma(a+\sum y_i+\tilde{y})}{(b+n+1)^{a+\sum y_i+\tilde{y}}}.$$

After simplifying some of the algebra, this gives

$$p(\tilde{y}|y_1,\ldots,y_n) = \frac{\Gamma(a+\sum y_i+\tilde{y})}{\Gamma(\tilde{y}+1)\Gamma(a+\sum y_i)}\left(\frac{b+n}{b+n+1}\right)^{a+\sum y_i}\left(\frac{1}{b+n+1}\right)^{\tilde{y}}$$

for $\tilde{y} \in \{0,1,2,\ldots\}$. This is a negative binomial distribution with parameters $(a+\sum y_i, b+n)$, for which

$$E[\tilde{Y}|y_1,\ldots,y_n] = \frac{a+\sum y_i}{b+n} = E[\theta|y_1,\ldots,y_n];$$

$$Var[\tilde{Y}|y_1,\ldots,y_n] = \frac{a+\sum y_i}{b+n}\,\frac{b+n+1}{b+n} = Var[\theta|y_1,\ldots,y_n]\times(b+n+1) = E[\theta|y_1,\ldots,y_n]\times\frac{b+n+1}{b+n}.$$

Let's try to obtain a deeper understanding of this formula for the predictive variance. Recall, the predictive variance is to some extent a measure of our posterior uncertainty about a new sample $\tilde{Y}$ from the population. Uncertainty about $\tilde{Y}$ stems from uncertainty about the population and the variability in sampling from the population. For large n, uncertainty about θ is small $((b+n+1)/(b+n) \approx 1)$ and uncertainty about $\tilde{Y}$ stems primarily from sampling variability, which for the Poisson model is equal to θ. For small n, uncertainty in $\tilde{Y}$ also includes the uncertainty in θ, and so the total uncertainty is larger than just the sampling variability $((b+n+1)/(b+n) > 1)$.

Example 19. Combining Independent Unbiased Estimators. Let d_1 and d_2 denote independent unbiased estimators of θ, having known variances σ_1^2 and σ_2^2. That is, for $i = 1, 2$,

$$E[d_i] = \theta, \quad Var(d_i) = \sigma_i^2.$$

Any estimator of the form

$$d = \lambda d_1 + (1-\lambda)d_2$$

will also be unbiased. To determine the value of λ that results in d having the smallest possible mean square error, note that

$$r(d,\theta) = Var(d) = \lambda^2 Var(d_1) + (1-\lambda)^2 Var(d_2) = \lambda^2\sigma_1^2 + (1-\lambda)^2\sigma_2^2.$$

Differentiation yields that

$$\frac{d}{d\lambda}r(d,\theta) = 2\lambda\sigma_1^2 - 2(1-\lambda)\sigma_2^2.$$

To determine the value of λ that minimizes $r(d, \theta)$ – call it $\hat{\lambda}$ – set this equal to 0 and solve for λ to obtain

$$2\hat{\lambda}\sigma_1^2 = 2(1 - \hat{\lambda})\sigma_2^2$$

or

$$\hat{\lambda} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} = \frac{1/\sigma_1^2}{1/\sigma_1^2 + 1/\sigma_2^2}.$$

In words, the optimal weight to give an estimator is inversely proportional to its variance (when all the estimators are unbiased and independent).

For an application of the foregoing, suppose that a conservation organization wants to determine the acidity content of a certain lake. To determine this quantity, they draw some water from the lake and then send samples of this water to n different laboratories. These laboratories will then, independently, test for acidity content by using their respective titration equipment, which is of differing precision. Specifically, suppose that d_i, the result of a titration test at laboratory i, is a random variable having mean θ, the true acidity of the sample water, and variance σ_i^2, $i = 1, ..., n$. If the quantities σ_i^2, $i = 1, .., n$ are known to the conservation organization, then they should estimate the acidity of the sampled water from the lake by

$$d = \frac{\sum_{i=1}^{n} d_i/\sigma_i^2}{\sum_{i=1}^{n} 1/\sigma_i^2}.$$