Lecture 2

A Turn to Statistics

So we've defined the parametric model as so,

$$F = \{\ p(x) : \vec{\theta} \in \Theta\ \}$$

In the real world we see x = <0, 0, 1, 0, 1, 0> = $\overline{x}$, otherwise known as data. Then we have to choose a parametric model. We might say the data fits Bernoulli. But before you can analyze the data, you have to define the parameter(s), θ. There are typically three goals of "statistical inference. Point estimation is the first goal, that is guessing the single best value for theta. The second goal is confidence sets, which is where we look to find a range of likely θs. The third goal is called theory testing, where we evaluate a theory about the value of theta.

Lets assume that the data x = <0, 0, 1, 0, 1, 0> is Bernoulli. Once you make an assumption of the parametric model, we can compute the JMF or JDF.

$$p(x;\ \theta) = \prod_{i=1}^{6}\ p(x_i;\ \theta)$$

In our context,

$$p(<0,0,1,0,1,0>;\ \theta) = (\theta^0(1-\theta)^{1-0})(\theta^0(1-\theta)^{1-0})(\theta^0(1-\theta)^{1-0})\ ...$$

Which simplifies to,

$$\theta^2(1-\theta)^4$$

The 1s in the data simplify to θ and the 0s simplify to (1-θ). This expression above represents the JMF of the Bernoulli random variable. Theta allows us to find the probability of success given the data. Because there are 2 ones and 4 zeroes, it would make sense to say that theta is less than one half.

$$if \theta = 0.5 \rightarrow p(x;\ \theta) = (0.5)^2(1-0.5)^4 = 0.0156$$

$$if\theta = 0.25 \rightarrow p(x;\ \theta) = (0.25)^2(1-0.25)^4 = 0.0198$$

Substituting in values for theta, we see that the probability that theta is 0.25 is greater than it being 0.5. But remember, theta is a constant parameter, not a variable as we are treating it to be right now. We use the *likelihood function*.

$$\mathscr{L}(\theta,\ x) = p(x;\ \theta)$$

On the right hand side of the equation above defines the probability of the data with theta known, that is the JDF/JMF. On the left hand side of the equation we have the likelihood function, which defines the probability of theta given x is known. Another way to look at the likelihood function is "the likelihood of seeing the parameter at a certain value". If you take the integral of the likelihood function across the parameter space, what value would you expect to receive? Well there is no rule.

$$\int_{\Theta} \mathscr{L}(\theta;\ x)d\theta = norule$$

$$\Sigma_(Supp[X])p(x;\ \theta) = 1$$

This process is used to find the most likely value of the parameter θ given the data. Now we will define argmax, that is the function that returns the most likely value for theta. MLE stands for Maximum Likelihood Estimator.

$$\hat{\theta}_{MLE} = argmax_{\theta\in\Theta}\{\mathscr{L}(\theta,\ x)\} = argmax_{\theta\in\Theta}\{g(\mathscr{L}(\theta,\ x))\}$$

The function g is a strictly increasing function. A very useful strictly increasing function is the log function. Let $g = ln$.

$$= argmax_{\theta\in\Theta}\{ln(\mathscr{L}(\theta,\ x))\}$$

$$\ell(\theta;\ x) = ln(\mathscr{L}(\theta,\ x))\}$$

To keep things tidy,

$$= argmax_{\theta\in\Theta}\{\ell(\theta,\ x)\}$$

We do this because of the following

$$\ell(\theta,\ x) = ln(p(x;\theta)) = ln(\prod_{i=1}^{n} p(x_i,\ \theta)) = \Sigma_{i=1}n\ ln(p(x;\ \theta))$$

Simplifying to sigma notation is much easier in the process. Taking the derivative of the sum of functions is a lot easier than taking the derivative of the product of functions. Let us now do this for our example, x = <0, 0, 1, 0, 1, 0>.

$$\ell(\theta,\ x) = \Sigma_{i=1}^{6}\ ln(\theta^{x_i}(1-\theta)^{1-x_i}) = \Sigma_{i=1}^{6}\ (x_i ln(\theta) + (1-x_i)(ln(1-\theta)$$

$$= (\Sigma x_i)ln(\theta) + (6 - \Sigma x_i)ln(1-\theta)$$

Note: $\bar{x} = \frac{1}{n}\Sigma x_i \rightarrow \Sigma x_i = n\bar{x}$

$$= 6\bar{x}ln(\theta) + (6 - 6\bar{x})ln(1-\theta) = 6(\bar{x}ln(\theta) + (1-\bar{x})ln(1-\theta)$$

Now we need to find the argmax of this function. Theta double hat, denoted $\hat{\hat{\theta}}_{MLE}$, represents a realization from the estimator. We have to take the derivative of the log likelihood with respect to theta and set the result equal to zero and solve.

$$\hat{\hat{\theta}}_{MLE} = \frac{d}{dx}[\ell(\theta;\ x)] = 6(\frac{\bar{x}}{\theta} - \frac{1-\bar{x}}{1-\theta}) = 0$$

Dividing by 6 on both sides ...

$$\frac{\bar{x}}{\theta} - \frac{1-\bar{x}}{1-\theta} = 0$$

Simplifying ...

$$\frac{\bar{x}}{\theta} = \frac{1-\bar{x}}{1-\theta}$$

Cross multiplying ...

$$\bar{x}(1-\theta) = (1-\bar{x})\theta$$

$$\bar{x} - \bar{x}\theta = \theta - \bar{x}\theta$$

$$\bar{x} = \theta$$

$$\hat{\hat{\theta}}_{MLE} = \bar{x} = \frac{2}{6}$$

The estimator, $\hat{\theta}_{MLE} = \bar{x}$ is a random variable whose realizations are estimate. These are a few benefits of using $\hat{\theta}_{MLE}$.

1. $\hat{\theta}_{MLE}$ is consistent. This means that this estimator can provide arbitrary precision on theta given enough n. The more n there are in the data set, the more accurate the prediction is.

2. $\hat{\theta}_{MLE} \stackrel{.}{\sim} \mathbb{N}(\theta, SE[\hat{\theta}_{MLE}]^2)$, in English, theta hat MLE is normally distributed.

3. "Efficiency" means that among all consistent estimators, it has minimum variance.

Lets examine MLE property 2. In the curly-F Bernoulli case,

$$\hat{\theta}_{MLE} \dot{\sim} \mathbb{N}(\theta, SE[\hat{\theta}_{MLE}]^2) = N(\theta, \sqrt{\frac{\theta(1-\theta)}{n}})$$

$$\hat{\theta}_{MLE} = \bar{X}, SE[\hat{\theta}_{MLE}] = SE[\bar{X}] = \sqrt{Var[\bar{X}]} = \sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{\theta(1-\theta)}{n}}$$

In the curly-F Geometric case,

$$\hat{\theta}_{MLE} = \frac{1}{\bar{X}+1}, SE[\frac{1}{\bar{X}+1}] = ?$$

We now use property 2 to attack the other goals of inference:

Confidence Sets: Sometimes there may be a range of possible valid thetas, so we use a method called the "confidence interval":

$$CI_{\theta,1-\alpha} = [\hat{\hat{\theta}}_{MLE} \pm Z\frac{\alpha}{2} SE[\hat{\theta}_{MLE}]]$$

For the iid (independently and identically distributed) Bernoulli case,

$$CI_{\theta,1-\alpha} = [\bar{x} \pm Z\frac{\alpha}{2}\sqrt{\frac{\theta(1-\theta)}{n}}$$

Letting $1-\alpha = 95\%$, then $\alpha = 0.05\%$.

$$CI_{\theta,95\%} = [\bar{x} \pm 1.96\sqrt{\frac{\theta(1-\theta)}{n}}$$