Handout 17: Bayesian Statistics I

Interpretation of Probability Both the mathematical framework for probability that you learned in 329 as well as most of the theoretical results in 330 are founded on a framework developed in a series of works by Andrey Kolmogorov (1925; 1931; 1933). While the mathematical framework is almost universally accepted, there are several different interpretations of how the probability of an event should be understood as a model or reality. Many sources suggest that there are roughly four different families of interpretation:[1]

[1] The philosophy of probability is a huge field. If you are interested in learning more, I suggest starting with the excellent article on the subject published by the *Stanford Encyclopedia of Philosophy*. I have included a few names and dates to help give some context, though note that these are just a few key works that fit into longer histories and conversations that continue into the present day.

- **naïve**: also known as the classical interpretation; the probability of an event as the proportion of total outcomes in which this event occurs (Laplace 1799)
- **frequentist**: the probability of an event is the proportion of times it will occur in an infinite number of independent repetitions of an experiment (Poisson 1837; Bernoulli, 1713)
- **subjective**: the probability is a measurment of certainty; an extention of Boolean (true/false) logic (Bayes 1763; Laplace 1812)
- **propensity**: a physical definition based on causes (Pierce 1878; Popper 1954)

We started with the naïve interpretation in 329, before largely using the frequentist interpretation throughout the remainder of the semester and up until now in 330. As I tried to explain in class last week, however, there are some challenges with the frequentist interpretation. Today, we will see how the subjective interpretation of probability leads to some novel approaches to statistical inference.

Bayesian Estimation for Binomial Using the frequentist interpretation of probability, we have treated our data (X) as a random variable defined by a distribution that has one or more fixed but unknown parameters. Our goal is to estimate features of the parameters through point estimators, confidence intervals, and/or hypothesis tests. The probabilistic properties of these tasks (for example, the bias of a point estimator or confidence level of a hypothesis test) are in terms of repeating an experiment many times; they do not say anything concrete about a particular run of an experiment.

Subjective probabilities open the door to a very different approach: we can treat the unknown parameters as random variables as well, where probabilities indicate uncertainty rather than long-term frequencies. Today, we will see how this works with a specific case, before moving to a more general framework next time.

Consider the task of estimating the parameter p from a random variable X taken from a $Bin(n, p)$ distribution with a known value of n. Let's assume that before observing any data, we think that any value of p is equally likely. We could write this by defining our unknown parameter p to be a random variable with a uniform distribution:

$$p \sim Unif(0, 1).$$

This is called the **prior distribution**, because it reflects our knowledge of p prior to observing any data. Now, when describing the random variable X, we have to give its distribution conditioned on a specific value of the random variable p. That is, we need to write this:[2]

$$X|p \sim Bin(n, p).$$

[2] The distribution of $X|p$ is called the **likelihood** following the notation from the MLE.

Now, the important thing is describing our knowledge about p *after* observing the data. That is, we want to know the distribution of $p|X$. Bayes rule tells us that we can calculate this as:

$$f_{p|X}(p|x) = \frac{f_{X|p}(x|p) \times f_p(p)}{f_X(x)}.$$

This quantity is called the **posterior distribution**. Determining the form of the posterior distribution is the key task in generating Bayesian estimators. One simplifying step is to notice that the denominator does not depend on p, so we can replace it with a constant, adding it back later (if needed) by whatever number makes the posterior a proper distribution (in other words, it integrates to 1). This gives the following standard form:[3]

$$f_{p|X}(p|x) \propto f_{X|p}(x|p) \times f_p(p).$$

[3] I will try to keep the subscripts on the density functions f for clarity in the notes. However, on the board I will almost always drop them. Feel free to do the same in your work.

When doing computations, you can always drop any additive or multiplicative factor that depends only on X and not p.

Now let's actually find the posterior distribution for this specific example. We have the following form of the density function (keep in mind that this is a function of p; we can remove any constants that depend only on x and n):

$$\begin{aligned} f_{p|X}(p|X) &\propto \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x} \cdot (1) \\ &\propto p^x \cdot (1-p)^{1-x} = p^{(x+1)-1} \cdot (1-p)^{(n-x+1)-1}. \end{aligned}$$

The last step may seem unusual, but if you look at the distribution table it becomes more clear. This is a Beta distribution, with $\alpha = (x+1)$ and $\beta = (n - x + 1)$. So, the posterior is given by:

$$p|X \sim Beta(x+1, n-x+1).$$

This new distribution represents our knowledge and uncertainty about the parameter P. We will talk more next time about how we can use the posterior distribution to generate Bayesian point estimates and confidence intervals.