Genesis of Bayesian Analysis:
Exchangeability and de Finetti's Theorem
Stat 775, 3/4/99

THE SUBJECTIVE PROBABILITY ASSESSMENT FOR A SEQUENCE OF BINARY TRIALS MAY NATURALLY ENFORCE EXCHANGEABILITY OF THE RANDOM VARIABLES. THIS TURNS OUT TO BE EQUIVALENT TO ASSERTING THAT THEY ARE CONDITIONALLY INDEPENDENT AND IDENTICALLY DISTRIBUTED BERNOULLI TRIALS IN WHICH THE SUCCESS PROBABILITY IS A RANDOM VARIABLE. IT FOLLOWS THAT BAYESIAN ANALYSIS MAY PROCEED BY PLACING PRIOR PROBABILITY DISTRIBUTIONS OVER THE PARAMETERS OF A STANDARD MODEL.

Subjective probability versus the orthodox view: Consider a sequence of trials each having a binary outcome, such as the case of tack tossing discussed in class on March 2. Denote the outcomes by $T_1, T_2, \ldots T_n$, with T_i, say, the indicator that tack i lands on its top side after being dropped (a success, say). The orthodox analysis of this system treats these T_i as *independent and identically distributed* Bernoulli trials, having some success probability, θ, say, which represents the long-run relative frequency with which a tack lands on its top side. Recall the class demonstration. I take tack 1 and prior to dropping it I ask you to evaluate your $P(T_1 = 1)$. Of course in the orthodox view this is θ, but in the subjective Bayesian view, your probability measures your uncertainty in the outcome of this trial. This probability is based on your understanding of the system, and in principle may be evaluated by you using available information. Since you may not have repeatedly dropped such tacks, you do not know θ, and thus θ cannot be your $P(T_1 = 1)$. Furthermore, different people (i.e. people with perhaps different understanding of the system) may have different values for $P(T_1 = 1)$. This is not possible in the orthodox view, since the probability is considered to be an attribute of the experiment itself rather than an expression of the uncertainty of the observer.

Learning: By observing the outcome of the first trial we have realized $T_1 = t_1$. In the class example $t_1 = 0$. Now I ask you to evaluate your probability $P(T_2 = 1|T_1 = 0)$. This first trial indeed provides you with some insight into the system, and it may be that your probability for trial 2 has changed from what it was; i.e. you have learned, and your expression of uncertainty has changed. Of course, in the orthodox view, T_1 and T_2 are independent trials, and so $P(T_2 = 1|T_1 = 0) = P(T_2 = 1) = \theta$. But subjectively we cannot assign such probability since we do not know θ.

In class we repeated several trials and I asked you to continue to update your probability for the next trial. You may have found it somewhat difficult to proceed with this updating.

There are many legitimate methods to do so, but one due to George Polya is particularly simple. Polya's learning scheme imagines the state of mind of the observer to be equivalent to an urn containing a 1's and b 0's. Uncertainty about the outcome T_1 corresponds to the probability $P(T_1 = 1) = a/(a+b)$. In this model, learning occurs by adding contents to the urn. If we observe $t_1 = 0$, then a 0 is added to the urn, otherwise a 1 is added. The updated urn represents the state of mind of the observer after having learned what happened on the first trial. As before, uncertainty about a second trial is based on the urn contents, and so $P(T_2 = 1|T_1 = t_1) = (a + t_1)/(a + b + 1)$. The degree of belief concerning this second trial has changed. For instance, if $a = b = 1$ to start with, then $P(T_1 = 1) = 1/2$ and $P(T_2 = 1|T_1 = 0) = 1/3$. Various values of a and b lead to the same $P(T_1 = 1)$. They are distinguished by the sum $a + b$ which regulates the amount by which your opinion will be changed. For example, if $a + b$ is very large, then your opinion will hardly change after realizing a trial. Learning continues through the sequence of trials as above, so that

$$P(T_i = 1|t_1, \ldots, t_{i-1}) = \left(a + \sum_{j=1}^{i-1} t_j \right) /(a + b + i - 1)$$

Notably, this probability is very close to the observed relative frequency when i is large compared to $a + b$. It is an exercise to combine these conditional probabilities and express the joint probability as:

$$p(t_1, \ldots, t_n) = \frac{\Gamma(a+b)\Gamma(s+a)\Gamma(n-s+b)}{\Gamma(a)\Gamma(b)\Gamma(a+b+n)}. \tag{1}$$

Here $s = \sum_{i=1}^{n} t_i$ and $\Gamma()$ is the Gamma function. With integer values of a and b, it suffices to know that, for example, $\Gamma(n+1) = n!$.

Exchangeable Random Variables: Two binary random variables T_1, T_2 are exchangeable if for every pair of realizations (t_1, t_2), the joint probability is invariant to permutation; that is, if $P(T_1 = t_1, T_2 = t_2) = P(T_1 = t_2, T_2 = t_1)$. Exchangeability is a property of the joint distribution of the random variables. Essentially, it is a statement that the labels on the trials are irrelevant. In the tack example, I labeled the tacks then tossed them in that order. If I think the labeling in no way affects the trial outcomes, then it is natural to assume exchangeable random variables. That is, I would prescribe the same collection of probabilities in a different experiment in which the tack labeled 2 was tossed first. It is important to note that exchangeable random variables need not be independent. Indeed, you may confirm this in the Polya sequence above. However, it is also easy to check that exchangeable random variables, though possibly correlated, have equal distributions. That is, you should be able to show that $P(T_1 = t) = P(T_2 = t)$ for all t. To do so consider the matrix containing the joint probabilities and note the symmetry in this matrix.

We say that a sequence of random variables $T_1, \ldots, T_n$ is exchangeable if for every realization $(t_1, \ldots, t_n)$ and for every permutation $\pi = (\pi_1, \ldots, \pi_n)$ of the integers $(1, 2, \ldots, n)$, we have equality of the following joint probabilities:

$$P(T_1 = t_1, T_2 = t_2, \ldots, T_n = t_n) = P(T_1 = t_{\pi_1}, T_2 = t_{\pi_2}, \ldots, T_n = t_{\pi_n}). \tag{2}$$

For instance, the probability of observing $(0, 1, 1, 0, 1)$ in five trials, say, must equal the probability of $(0, 0, 1, 1, 1)$ and also the probability of $(1, 1, 0, 0, 1)$, and so on. Exchangeability places many constraints on the joint distribution, but it is in some ways a natural and quite primitive assumption. It says that the probability you assign to the outcomes of a sequence of trials does not depend on the order in which the trials occur.

Observe that if the joint probability happens to be a function of $(t_1, \ldots, t_n)$ only through $s = \sum_{i=1}^n t_i$, then the random variables must be exchangeable by the commutative property of addition. Further observe that in Polya's learning scheme, $T_1, \ldots, T_n$ are exchangeable; see equation (1). This is somewhat surprising considering the asymmetry of its construction.

Mixing Bernoulli Trials: There is a simple recipe for constructing exchangeable binary sequences. Let Z be a random variable with range $[0, 1]$, and let $T_1, T_2, \ldots, T_n$ be conditionally independent and identically distributed given $Z = z$. (Draw the DAG!) Specifically, let each one be a Bernoulli trial with success probability z. Marginally over Z, the sequence $T_1, \ldots, T_n$ is exchangeable. We proved this in class in the discrete case by noting

$$\begin{aligned} p(t_1, \ldots, t_n) &= \sum_z p(t_1, \ldots, t_n, z) \\ &= \sum_z p(t_1, \ldots, t_n | z) p(z) \\ &= \sum_z \left(\prod_i p(t_i | z) \right) p(z) \\ &= \sum_z z^s (1 - z)^{n-s} \, p(z) \end{aligned}$$

where $s = \sum_i t_i$. Note where the conditional independence and identical conditional distribution enter above. Being a function of the sum s, the random variables are exchangeable by the observation above. We will say that the sequence $T_1, \ldots, T_n$ so formed is a mixture of conditionally iid Bernoulli trials.

An important example is when Z has a continuous distribution, specifically the Beta(a, b) distribution. That is, the probability density of Z is

$$p(z) = c z^{a-1} (1 - z)^{b-1}$$

and c is a normalizing constant $c = \Gamma(a + b)/[\Gamma(a)\Gamma(b)]$. It is an exercise to work out the marginal distribution for $T_1, \ldots, T_n$ in this case.

De Finetti's Theorem: The late Italian probabilist Bruno de Finetti championed the subjective view of probability and established a deep connection between subjective assessments and the orthodox view discussed above. Consider the indefinitely long sequence of binary random variables $T_1, T_2, \ldots$. Think of their joint distribution as subjectively determining your uncertainty, such as in the Polya learning scheme. De Finetti proved that these variables are exchangeable for each n if and only if they are a mixture of conditionally iid Bernoulli trials. We already saw above that a mixture of conditionally iid Bernoulli trials is exchangeable. The difficult and important result is that any exchangeable sequence must be a mixture of conditionally iid Bernoulli trials. This result characterizes exchangeable binary sequences. In other words, if you assess your uncertainty about a sequence of trials in such a way that labeling of the trials is irrelevant, then you are equivalently asserting the existence of a random variable Z such that given $Z = z$, the trials are iid Bernoulli(z). But the orthodox view is precisely that the trials are iid Bernoulli(θ), where the unknown θ represents the long-run relative frequency of successes. The connection between the subjective view and the orthodox view is complete if we simply add a probability distribution over θ. That is, if we treat the long-run relative frequency itself as a random variable (called Z, say). A subjective assessment of a sequence of binary trials which takes the primitive exchangeability assumption is equivalent to placing a *prior* distribution on the long-run success probability within the orthodox view. Bayesian analysis rests on this formulation. That is, we accept the orthodox view of iid trials, but we place probability distributions over the parameters of the *sampling* distributions.

As an exercise, you should show that by taking a Beta(a, b) prior for θ, the marginal distribution of $T_1, \ldots, T_n$ is precisely the same as that of a Polya sequence, i.e. (1). So if you consider learning about future tack tossings via the Polya learning scheme, it is as if you characterize your uncertainty about the long-run success probability with a Beta(a, b) distribution, and you view the subsequent tosses as conditionally iid Bernoulli trials.