Exchangeability

Sayan Mukherjee

Imagine that one observes 10 coin flips of which 9 heads. It is not unreasonable to change ones belief that the next flip will result in heads than before observing the 10 flips. From the perspective of probabilities as beliefs, the subjectivist perspective, this makes sense. From the objective perspective of defining a probability of success p and then taking iid draws it is not immediately obvious how to deal with this perspective. The idea of exchangeability will allow for an objectivist interpretation for changing the parameter p as more and more coins are flipped. The key assumption we will make is that we don't care when the heads or tails occur but just how many of them we have observed.

Definition 0.0.1 *An infinite sequence $\{X_i\}_{i=1}^{\infty}$ of random variables is exchangeable if $\forall n = 1, 2, ...$.*

$$X_1, ..., X_n \stackrel{d}{=} X_{\pi(1)}, ..., X_{\pi(n)}, \quad \forall \pi \in S(n),$$

where $S(n)$ is the symmetric group, the group of permutations.

An example would be the following

$$\mathbf{P}(1, 1, 0, 0, 0) = \mathbf{P}(1, 0, 0, 0, 1).$$

In these notes we will focus on exchangeable sequences of binary random variables but results and ideas are valid for discrete and continuous random variables. Note that $iid \Rightarrow exchangeable$ but $exchangeable \not\Rightarrow iid$.

Pólya's urn

The following is a classic process to generate exchangeable binary sequences and we will return to it several times. This is the Pólya urn model. One is given an urn with B_0 black and W_0 white balls. the following procedure is run

1) Draw a ball at random from the urn and note its color.

(2) If the ball is black then $X_i = 1$, $X_i = 0$ otherwise.

(3) Increment $i = i + 1$.

(4) Place a balls of the same color in the urn.

(5) Goto (1).

If $a = 1$ then this is an iid sequence, sampling with replacement, when $a = 0$ this is sampling without replacement and results in a hypergeometric distribution.

To see that $X_1, ..., X_n$ is exchangeable note the following

$$\begin{aligned}
\mathbf{P}(1,1,0,1) = &= \frac{B_0}{B_0 + W_0} \times \frac{B_0 + a}{B_0 + W_0 + a} \times \frac{W_0}{B_0 + W_0 + 2a} \times \frac{B_0 + 2a}{B_0 + W_0 + 3a} \\
\mathbf{P}(1,0,1,1) = &= \frac{B_0}{B_0 + W_0} \times \frac{W_0}{B_0 + W_0 + a} \times \frac{B_0 + a}{B_0 + W_0 + 2a} \times \frac{B_0 + 2a}{B_0 + W_0 + 3a}.
\end{aligned}$$

The sequence $\{X_i, i \geq 1\}$ is not iid and may not be a Markov process.

de Finetti's theorem

We start with the colloquial statement of de Finetti's theorem:

An infinite exchangeable sequence is distributed as a mixture of iiid sequences.

We now state the formal definition of the theorem for binary sequences

Theorem 0.0.1 (de Finetti 1931) *A binary sequence $\{X_n\}_{i=1}^{\infty}$ is exchangeable iff there exists a distribution function $F(p)$ on $[0,1]$ such that for any $n \geq 1$*

$$\mathbf{P}(X_1 = x_1, ..., X_n = x_n) = \int_0^1 p^{S_n}(1-p)^{n-S_n} dF(p)$$

where $S_n = \sum_i x_i$.

The distribution F is a function of the limiting frequency

$$Y = \bar{X}_\infty = \lim_{n \to \infty} \frac{\sum_i X_i}{n}, \quad \mathbf{P}(Y \leq p) = F(p),$$

and conditioning on $Y = p$ results in iid Bernoulli draws

$$\mathbf{P}(X_1 = x_1, ..., X_n = x_n \mid Y = p) = p^{S_n}(1-p)^{n-S_n},$$

and for the Pólya urn model

$$\lim_{n \to \infty} \bar{X}_n = Y \sim \text{Beta}\left(\frac{B_0}{B_0 + W_0}, \frac{W_0}{B_0 + W_0}\right).$$

The result can be interpreted from a statistical, probabilistic, and function analytic perspective. We will mention the probabilistic and function analytic perspective later in the notes.

The statistical perspective:

- A frequentist would assume that data is drawn iid from a Binomial with unknown parameter p
- A Bayesian would place a prior distribution π on the the parameter p of the Binomial distribution

de Finetti's theorem is saying that the Bayesian model is equivalent to the assumption that the observations $\{X_i, i \geq 1\}$ are exchangeable.

Analysis of the Pólya urn model

We will use de Finetti's theorem to compute the limiting distribution for the Pólya urn model

$$\lim_{n \to \infty} \bar{X}_n = Y \sim \text{Beta}\left(\frac{B_0}{B_0 + W_0}, \frac{W_0}{B_0 + W_0}\right).$$

We first define the Beta and Gamma functions

$$\beta(x,y) = \int_0^1 p^{x-1}(1-p)^{y-1} dp, \quad \Gamma(x+1) = x\Gamma(x), \quad \beta(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

The probability of observing k black balls given n draws is

$$\begin{aligned} \mathbf{P}(k \text{ black balls given } n \text{ draws}) &= \binom{n}{k} \frac{B_0(B_0+1)\cdots(B_0+k-1)W_0(W_0+1)\cdots(W_0+n-k-1)}{(W_0+B_0)(W_0+B_0+1)\cdots(W_0+B_0+n-1)} \\ &= \binom{n}{k} \frac{\beta(B_0+k, B_0+n-k)}{\beta(B_0, W_0)}. \end{aligned} \tag{1}$$

Note the proportion of black balls at any stage n of the process as

$$\rho_n = \frac{B_n}{W_n + B_n}, \quad \rho_\infty = \lim_{n \to \infty} \frac{B_n}{W_n + B_n}.$$

We know that

$$\mathbf{P}(k \text{ black balls given } n \text{ draws} \mid \rho_\infty = p) = \binom{n}{k} p^k (1-p)^{n-k},$$

and if $\rho_\infty \sim F(p)$ then

$$\begin{aligned} \mathbf{P}(k \text{ black balls given } n \text{ draws}) &= \int_0^1 \mathbf{P}(k \text{ black balls given } n \text{ draws} \mid \rho_\infty = p) dF(p) \\ \mathbf{P}(k \text{ black balls given } n \text{ draws}) &= \binom{n}{k} \int_0^1 p^k (1-p)^{n-k} dF(p) \end{aligned} \tag{2}$$

If we set (1) and (2) to be equal we obtain

$$\begin{aligned}\int_0^1 p^k(1-p)^{n-k}dF(p) &= \frac{\beta(B_0+k, B_0+n-k)}{\beta(B_0+W_0)} \\ &= \frac{1}{\beta(B_0+W_0)}\int_0^1 p^{B_0+k-1}(1-p)^{B_0+n-k-1}dp \\ &= \int_0^1 p^k(1-p)^k\frac{p^{B_0-1}(1-p)^{W_0-1}}{\beta(B_0,W_0)}dp,\end{aligned}$$

which implies that

$$f(p) = \frac{1}{\beta(B_0,W_0)}p^{B_0-1}(1-p)^{W_0-1}.$$

Sketch of proof of de Finetti's theorem

$\{X_i, i \geq 1\}$ is a sequence of exchangeable binary random variables and $(x_1,...,x_n) \in \{0,1\}^n$ for $n \leq N$ and $S_n = \sum x_i$ the joint probability

$$\mathbf{P}(X_1 = x_1, ..., X_n = x_n) = \sum_{M=S_n}^{N-n+S_n} \frac{\binom{N-n}{M-S_n}}{\binom{N}{M}}\mathbf{P}(S_n = M),$$

so we ask about the limit of the above quantity

$$\mathbf{P}(X_1 = x_1, ..., X_n = x_n) = \lim_{N\to\infty} \sum_{M=S_n}^{N-n+S_n} \frac{\binom{N-n}{M-S_n}}{\binom{N}{M}}\mathbf{P}(S_n = M).$$

This quantity converges in distribution to

$$\int_0^1 p^{S_n}(1-p)^{n-S_n}dF(p),$$

however we do not yet have the tools to prove this, we need to define conditional expectations and Martingales.

Finite exchangeable sequences

The result for finite exchangeable sequences can be very different than the infinite sequence setting. The following example illustrates this. We look at the following exchangeable sequence on (X_1, X_2)

$$\begin{aligned}\mathbf{P}(X_1 = 0, X_2 = 1) = \mathbf{P}(X_1 = 1, X_2 = 0) &= 1/2 \\ \mathbf{P}(X_1 = 0, X_2 = 0) = \mathbf{P}(X_1 = 1, X_2 = 1) &= 0,\end{aligned}$$

this is the case of sampling a black and white ball without replacement. For a version of de Finetti's theorem to hold there must exist a density $f(p)$ such that

$$0 = \int_0^1 p^2 f(p)dp = \int_0^1 (1-p)^2 f(p)dp,$$

and there is no such density $f(p)$.

The above result states that there is no distribution$F(p)$ such that the joint distribution given by sampling without replacement can be stated as a mixture of Binomials. However, intuitively one would think that if we are only concerned with the joint distribution of short sequences in a very long sequence then an approximate version of de Finetti's theorem should hold.

Define the joint distribution of the first k random variables in the sequence $X_1, ..., X_k, X_{k+1}, ..., X_n$ as $\mathbf{P}_k$. This subsequence is k-exchangeable, the order of the random variables does not effect the joint distribution $\mathbf{P}_k$. Let $x = \{0,1\}$ and x^k is the set of k-tuples $\{0,1\}^k$ now we define $\mathbf{P}_k$ as

$$\mathbf{P}_k(x_1, ..., x_k) = \mathbf{P}_k(x_{\pi(1)}, ..., x_{\pi(n)}),$$

we define p^k as the distribution of k independent draws of probability p so

$$p^k(x_1, ..., x_k) = \prod_{j=1}^k p(x_j).$$

We define π as a probability on $[0, 1]$, the Borel subsets of the unit interval. We now define $\mathbf{P}_{\pi k}$ as

$$\mathbf{P}_{\pi k}(x_1, ..., x_k) = \int_0^1 \left[\prod_{j=1}^{k} p(x_j) \right] \pi(dp).$$

We can now state how far $\mathbf{P}_{\pi k}$ is from $\mathbf{P}_k$. This says how close to a de Finetti theorem can one get. The formal statement is that there exists a prior π such that

$$\|P_k - \mathbf{P}_{\pi k}\| \leq 4k/n, \quad \|Q - V\| = 2 \sup_A |Q(A) - V(A)|,$$

the above is the total variation distance.

Perspectives of exchangeability

We now provide a few different perspectives on exchangeability. We state the results with greater generality than for the binary sequence. Take a space S for example $S = \mathbb{R}$ State as $\mathcal{P}(S)$ as the space of measures on S, state as $\mu \in \mathcal{P}(S)$ and state as α