MATH 559: BAYESIAN THEORY AND METHODS
THE DE FINETTI REPRESENTATION THEOREM

Exchangeability

(i) A *finite* sequence of random variables $Y_1, Y_2, \ldots, Y_n$ taking values on $\mathcal{Y} \subseteq \mathbb{R}^d$ is *finitely exchangeable* if, for arbitrary sets $A_1, A_2, \ldots, A_n \subseteq \mathbb{R}^d$,

$$\Pr\left[\bigcap_{i=1}^{n}(Y_i \in A_i)\right] = \Pr\left[\bigcap_{i=1}^{n}(Y_i \in A_{\sigma(i)})\right] \tag{1}$$

for any permutation $(\sigma(1), \sigma(2), \ldots, \sigma(n))$ of indices $(1, 2, \ldots, n)$.

(ii) An infinite sequence, $Y_1, Y_2, \ldots$, is *infinitely exchangeable* if (1) holds for all finite subsets of size n of the sequence, for all $n \geq 1$.

Theorem ***(The de Finetti 0-1 Representation Theorem)***
If $Y_1, Y_2, \ldots$ is an infinitely exchangeable sequence of 0-1 variables, then there exists a probability function $\pi_0(.)$ such that for all $n \geq 1$, the joint mass function of $(Y_1, Y_2, \ldots, Y_n)$ can be represented

$$p_{Y_1,Y_2,\ldots,Y_n}(y_1, y_2, \ldots, y_n) = \int_0^1 \left\{\prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i}\right\} \pi_0(d\theta)$$

where $\pi_0(.)$ is defined for $0 \leq \theta \leq 1$ by

$$\int_0^\theta \pi_0(dt) = \lim_{n\to\infty} \Pr[R_n \leq \theta]$$

with

$$S_n = \sum_{i=1}^{n} Y_i \qquad R_n = \frac{S_n}{n}.$$

Furthermore, we may define

$$\theta_0 = \lim_{n\to\infty} R_n,$$

that is, $R_n \xrightarrow{a.s.} \theta_0$, so that θ_0 is the limiting relative frequency that an arbitrary Y takes the value one.

PROOF By exchangeability, for $0 \leq s_n \leq n$, we must have

$$\Pr[S_n = s_n] = \binom{n}{s_n} p_{Y_1,Y_2,\ldots,Y_n}(y_1, y_2, \ldots, y_n) \tag{2}$$

with $s_n = \sum\limits_{i=1}^{n} y_i$, as there are $\dbinom{n}{s_n}$ sequences of ys that have sum s_n.

Now consider $N \geq n$, with N finite. By the Theorem of Total Probability we may write

$$\Pr[S_n = s_n] = \sum_{s=s_n}^{N-n+s_n} \Pr[S_n = s_n | S_N = s] \Pr[S_N = s] \tag{3}$$

that is, in light of the information fact that $S_N = s$, we can identify the conditional probability for different values of S_n.

By the exchangeability assumption, $\Pr\left[S_n = s_n | S_N = s\right]$, is a **hypergeometric** probability:

- we have a (finite) exchangeable sequence of N binary variables that contain precisely s ones and $N-s$ zeros; in this sequence, the first n variables contain precisely s_n ones and $n-s_n$ zeros;
- alternatively, of the s ones in the sequence of length N, s_n are placed somewhere in the first n positions, the remaining $s-s_n$ are placed in the remaining $N-n$ positions.

Therefore

$$\Pr\left[S_n = s_n | S_N = s\right] = \frac{\dbinom{s}{s_n}\dbinom{N-s}{n-s_n}}{\dbinom{N}{n}} \qquad 0 \le s_n \le n.$$

Writing out the binomial coefficients, we have

$$\frac{\dbinom{s}{s_n}\dbinom{N-s}{n-s_n}}{\dbinom{N}{n}} = \frac{s!}{s_n!(s-s_n)!}\frac{(N-s)!}{(n-s_n)!(N-s-n+s_n)!}\frac{n!(N-n)!}{N!}$$

$$= \frac{n!}{s_n!(n-s_n)!}\frac{s!}{(s-s_n)!}\frac{(N-s)!}{(N-s-n+s_n)!}\frac{(N-n)!}{N!}.$$

For $k \ge l \ge 1$, define

$$(k)_l = \frac{k!}{(k-l)!} = k\left(k-1\right)\left(k-2\right)\ldots\left(k-l+1\right)$$

– this is the *descending factorial* function. In this notation

$$\frac{\dbinom{s}{s_n}\dbinom{N-s}{n-s_n}}{\dbinom{N}{n}} = \binom{n}{s_n}\frac{(s)_{s_n}((N-s))_{n-s_n}}{(N)_n}$$

Thus

$$\Pr\left[S_n = s_n\right] = \binom{n}{s_n}\sum_{s=s_n}^{N-n+s_n}\frac{(s)_{s_n}((N-s))_{n-s_n}}{(N)_n}\Pr\left[S_N = s\right], \tag{4}$$

This result holds for any finite N. For the result, we require that the representation holds for the infinitely exchangeable sequence, and for all $n \ge 1$, so we must consider behaviour as $N \longrightarrow \infty$.

The random variable S_N takes values on the set $\{0, 1, \ldots, N\}$ with probabilities

$$\Pr[S_N = 0], \Pr[S_N = 1], \ldots, \Pr[S_N = N]$$

respectively. Denote by $F_{S_N}(s)$ the cumulative distribution for S_N; in the construction, $F_{S_N}(s)$ is not pre-specified, but it is a univariate distribution function that induces the form of $p_{Y_1,Y_2,\ldots,Y_n}\left(y_1, y_2, \ldots, y_n\right)$. Now, let

$$R_N = \frac{S_N}{N} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

and denote by $F_{R_N}(r)$ the cumulative distribution for R_N; R_N takes values on the finite set

$$\mathcal{R}_N = \{0, 1/N, 2/N, \ldots, (N-1)/N, 1\}$$

with $\Pr[R_N = r] \equiv \Pr[S_N = Nr]$, for $r \in \mathcal{R}_N$, and zero otherwise. Thus we may re-write (4)

$$\Pr\left[S_n = s_n\right] = \binom{n}{s_n} \sum_{s=s_n}^{N-n+s_n} \frac{(s)_{s_n}((N-s))_{n-s_n}}{(N)_n} \Pr\left[R_N = s/N\right], \tag{5}$$

or, by changing variables in the sum from s to $r = s/N$,

$$\Pr\left[S_n = s_n\right] = \binom{n}{s_n} \sum_{r \,\in\, \mathcal{R}_N} \frac{(Nr)_{s_n}((N-Nr))_{n-s_n}}{(N)_n} \Pr\left[R_N = r\right]. \tag{6}$$

We consider the behaviour of this summation as $N \longrightarrow \infty$. First, note that $F_{R_N}(r)$ is a step-function with steps at values in $\mathcal{R}_N$. By utilizing the Riemann-Stieltjes integral notation we may re-write the sum as

$$\Pr\left[S_n = s_n\right] = \binom{n}{s_n} \int_0^1 \frac{(Nr)_{s_n}\left((1-r)\,N\right)_{n-s_n}}{(N)_n} F_{R_N}\left(dr\right). \tag{7}$$

Now when $k \gg l$,

$$(k)_l = k(k-1)(k-2)\dots(k-l+1) \simeq k^l$$

as there are l terms each of which are approximately equal to k. Thus, in the limit as $N \longrightarrow \infty$ for n and s_n fixed, for each $r \in \mathcal{R} \equiv \{a/b : a, b \in \mathbb{Z}, a \geq 0, b \geq 1, a \leq b\}$, we have that

$$\frac{(Nr)_{s_n}\left((1-r)\,N\right)_{n-s_n}}{(N)_n} \longrightarrow r^{s_n}\left(1-r\right)^{n-s_n} = \prod_{i=1}^{n} r^{y_i}\left(1-r\right)^{1-y_i}$$

with terms in N cancelling.

Now we consider the behaviour of $F_{R_N}\left(dr\right)$ as $N \longrightarrow \infty$; to do this, we appeal to *Helly's Theorem*, and related concepts of weak convergence of measure (see for example, Ash and Doléans-Dade, *Probability and Measure Theory (2nd Edition)*, section 7.2)

- the sequence of distributions $\{F_{R_N}(.)\}_{N=n+1}^{\infty}$ has a convergent subsequence $\{F_{R_{N_j}}(.)\}_{j=1}^{\infty}$;
- that is, for some distribution function $\pi_0(.)$, say, and for each c, $0 \leq c \leq 1$

  $$F_{R_{N_j}}\left(c\right) \longrightarrow \pi_0\left(c\right)$$

  as $j \longrightarrow \infty$.

Hence, limiting form of equation (7) as $N \longrightarrow \infty$ is obtained as

$$\Pr\left[S_n = s_n\right] = \binom{n}{s_n} \int_0^1 \frac{(Nr)_{s_n}\left((1-r)\,N\right)_{n-s_n}}{(N)_n} F_{R_N}\left(dr\right) \longrightarrow \binom{n}{s_n} \int_0^1 r^{s_n}(1-r)^{n-s_n}\,\pi_0\left(dr\right).$$

and the result follows by rearrangement. We may reserve the special notation θ to represent the integrating variable and thus have

$$p_{Y_1,Y_2,\dots,Y_n}(y_1, y_2, \dots, y_n) = \int_0^1 \theta^{s_n}(1-\theta)^{n-s_n}\,\pi_0\left(d\theta\right)$$

Finally, as $n \longrightarrow \infty$, it is evident from the strong law of large numbers that $R_n \stackrel{a.s.}{\longrightarrow} \theta_0$, say, for some $\theta_0 \in [0, 1]$.

Corollary : Posterior Predictive Distributions

For $m, n \geq 1$, under infinite exchangeability, the predictive distribution of $Y_{n+1}, \ldots, Y_{n+m}$ given $Y_1, Y_2, \ldots, Y_n$ takes the form

$$p_{Y_{n+1},Y_{n+2},\ldots,Y_{n+m}|Y_1,Y_2,\ldots,Y_n}\left(y_{n+1}, y_{n+2}, \ldots, y_{n+m}|y_1, y_2, \ldots, y_m\right)$$

$$= \frac{p_{Y_1,Y_2,\ldots,Y_{n+m}}\left(y_1, y_2, \ldots, y_{n+m}\right)}{p_{Y_1,Y_2,\ldots,Y_n}\left(y_1, y_2, \ldots, y_n\right)} = \int_0^1 \left\{\prod_{i=n+1}^{n+m} \theta^{y_i}\left(1-\theta\right)^{1-y_i}\right\} \pi_n\left(d\theta\right) \quad (8)$$

where

$$\pi_n(d\theta) = \frac{\left\{\prod\limits_{i=1}^{n} \theta^{y_i}\left(1-\theta\right)^{1-y_i}\right\}\pi_0\left(d\theta\right)}{\displaystyle\int_0^1 \left\{\prod\limits_{i=1}^{n} t^{y_i}\left(1-t\right)^{1-y_i}\right\}\pi_0\left(dt\right)}$$

is the *posterior* distribution, which is effectively and updated version of the *prior* distribution $\pi_0(d\theta)$.

Finally, for all $n \geq 0, m \geq 1$, let

$$S_{n+1,n+m} = \sum_{i=n+1}^{n+m} Y_i \qquad\qquad R_{n+1,n+m} = \frac{S_{n+1,n+m}}{m}$$

so that $S_{1,n} \equiv S_n$ from above. Then from (8) we can deduce that

$$\Pr[S_{n+1,n+m} = s|S_{1,n} = s_{1,n}] = \int_0^1 \binom{m}{s} t^s\left(1-t\right)^{m-s}\pi_n\left(dt\right) \qquad s \in \{0, 1, 2, \ldots, m\}$$

using t now as the integrating variable. It follows immediately that

$$\Pr[R_{n+1,n+m} = r|S_{1,n} = s_{1,n}] = \Pr[S_{n+1,n+m} = mr|S_{1,n} = s_{1,n}] \qquad r \in \mathcal{R}_m \equiv \{0, 1/m, 2/m, \ldots, 1\}$$

We consider the behaviour as $m \longrightarrow \infty$, with n fixed; the set of values that r can take becomes the interval $[0, 1]$, and we therefore consider the distribution function for $R_{n+1,n+m}$: for $0 \leq \theta \leq 1$

$$\Pr[R_{n+1,n+m} \leq \theta \mid S_{1,n} = s_{1,n}] = \Pr[S_{n+1,n+m} \leq m\theta|S_{1,n} = s_{1,n}] = \sum_{j=0}^{\lfloor m\theta \rfloor} \int_0^1 \binom{m}{j} t^j\left(1-t\right)^{m-j}\pi_n\left(dt\right)$$

$$= \int_0^1 \left\{\sum_{j=0}^{\lfloor m\theta \rfloor} \binom{m}{j} t^j\left(1-t\right)^{m-j}\right\}\pi_n\left(dt\right).$$

The term in the central curly bracket can be interpreted as follows: if $X \sim Binomial(m, t)$,

$$\sum_{j=0}^{\lfloor m\theta \rfloor} \binom{m}{j} t^j\left(1-t\right)^{m-j} \equiv \Pr[X \leq \lfloor m\theta \rfloor].$$

Now, by standard central limit theorem arguments, for large m,

$$\Pr[X \leq \lfloor m\theta \rfloor] = \Pr\left[\frac{X - mt}{\sqrt{mt(1-t)}} \leq \frac{\lfloor m\theta \rfloor - mt}{\sqrt{mt(1-t)}}\right] \doteqdot \Phi\left(\frac{m\theta - mt}{\sqrt{mt(1-t)}}\right) = \Phi\left(\sqrt{m}\frac{(\theta - t)}{\sqrt{t(1-t)}}\right)$$

where $\Phi(.)$ is the standard Normal cdf. However, for fixed r and t, as $m \longrightarrow \infty$

$$\Phi\left(\sqrt{m}\frac{(\theta-t)}{\sqrt{t(1-t)}}\right) \longrightarrow \begin{cases} 0 & \theta < t \\ 1 & \theta > t \end{cases}$$

Thus

$$\lim_{m\longrightarrow\infty} \Pr\left[R_{n+1,n+m} \leq \theta \mid S_{1,n} = s_{1,n}\right] = \int_0^\theta \pi_n(dt).$$

which reveals that the posterior distribution $\pi_n(d\theta)$ is in fact the *limiting predictive distribution* for $R_{n+1,n+m}$ as $m \longrightarrow \infty$ with n fixed.

Interpretation: The de Finetti Representation for binary variables

$$p_{Y_1,Y_2,\ldots,Y_n}\left(y_1,y_2,\ldots,y_n\right) = \int_0^1 \left\{\prod_{i=1}^n \theta^{y_i}\left(1-\theta\right)^{1-y_i}\right\} \pi_0\left(d\theta\right)$$

can be interpreted in the following way;

- The joint marginal distribution of the observable quantities $Y_1, Y_2, \ldots, Y_n$ can be represented via a conditional/marginal decomposition. The conditional distribution is

$$\left\{\prod_{i=1}^n \theta^{y_i}\left(1-\theta\right)^{1-y_i}\right\}$$

  formed **as if it were a likelihood** for data $Y_1, Y_2, \ldots, Y_n$ **conditional** on a quantity θ.
- θ is a quantity that parameterizes the model whose true value θ_0 is **defined by**

$$\frac{1}{n}\sum_{i=1}^n Y_i \xrightarrow{a.s.} \theta_0$$

  that is, a strong law limit of observable quantities.
- $\pi_0(.)$ defines a probability measure for θ which term the **prior** probability measure.
- In the corollary, for the posterior predictive distribution, we have

$$p_{Y_{n+1},\ldots,Y_{n+m}|Y_1,\ldots,Y_n}\left(y_{n+1},y_{n+2},\ldots,y_{n+m}|y_1,\ldots,y_m\right) = \int_0^1 \left\{\prod_{i=n+1}^{n+m} \theta^{y_i}\left(1-\theta\right)^{1-y_i}\right\} \pi_n\left(d\theta\right)$$

  where the **posterior** distribution for θ is

$$\pi_n(d\theta) = \frac{\left\{\prod\limits_{i=1}^n \theta^{y_i}\left(1-\theta\right)^{1-y_i}\right\}\pi_0\left(d\theta\right)}{\displaystyle\int_0^1 \left\{\prod\limits_{i=1}^n t^{y_i}\left(1-t\right)^{1-y_i}\right\}\pi_0\left(dt\right)}$$

  that is, the prior distribution **updated** in light of the data $y_1, \ldots, y_n$.

Thus, from a very simple and natural assumption (exchangeability) about observable random quantities, we have a theoretical justification for using Bayesian methods, and a natural interpretation of parameters as limiting quantities. The theorem can be extended from the simple 0-1 case to very general situations

Theorem ***The de Finetti General Representation Theorem***
If $Y_1, Y_2, \ldots$ is an infinitely exchangeable sequence of random variables with joint probability measure P_Y, then there exists a distribution function $\pi_0(.)$ on $\mathcal{F}$, the set of all distribution functions on $\mathbb{R}$, such that the joint distribution of $(Y_1, Y_2, \ldots, Y_n)$ is specified by

$$\Pr\left[\bigcap_{i=1}^{n}(Y_i \leq y_i)\right] = \int_{\mathcal{F}} \left\{\prod_{i=1}^{n} F(y_i)\right\} \pi_0(dF)$$

where F parameterizes the model and is an unknown/unobservable distribution function with true value F_0 where

$$F_0\left(y\right) = \lim_{n\to\infty} F_n((-\infty, y]) = \lim_{n\to\infty} \left\{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,y]}(Y_i)\right\}$$

is a probability measure on the space of functions $\mathcal{F}$, defined as a limiting measure as $n \longrightarrow \infty$ on the ***empirical distribution function*** *F_n.*