**Exchangeable sequences and probabilities for probabilities**

1996; modified 98–5–21 to add material on mutual information; modified 98–7–21 to add Heath-Sudderth proof of de Finetti representation; modified 99–11–24 to make the presentation clearer and more complete and 00–10–18 to include comments on the integration measure

Suppose one assigns a probability, $P(p_1,\ldots,p_N) = P(\mathbf{p})$, to the single-trial probabilities for $N$ alternatives. Then, in $L$ trials, the occurrence probability—i.e., the total probability that alternative $i$ occurs $n_i$ times, $i = 1,\ldots,N$—is given by

$$\begin{aligned} p(\mathbf{n}) = p(n_1,\ldots,n_N) &= \int d\mathbf{p}\, p(n_1,\ldots,n_N|\mathbf{p})P(\mathbf{p}) \\ &= \int d\mathbf{p}\, \frac{L!}{n_1!\ldots n_N!}\, p_1^{n_1}\ldots p_N^{n_N}\, P(\mathbf{p}) \\ &= \frac{L!}{n_1!\ldots n_N!}\, \langle p_1^{n_1}\ldots p_N^{n_N}\rangle \ . \end{aligned}$$

Here

$$L = \sum_{i=1}^{N} n_i \ ,$$

$$d\mathbf{p} = dp_1\ldots dp_N \ ,$$

and the integral runs over positive values of the single-trial probabilities. The probability on probabilities, $P(\mathbf{p})$, is restricted to the simplex; i.e., as a function on positive values of the probabilities, it is proportional to a delta function

$$\delta\left(\sum_{i=1}^{N} p_i - 1\right) .$$

Notice that, in contrast to other notes, we do not include the “inverse direction cosine” $\sqrt{N}$ in the integration measure on the simplex, and we put the $\delta$ function that restricts to the simplex in the distribution rather than in the integration measure.

The moment of single-trial probabilities,

$$\langle p_1^{n_1}\ldots p_N^{n_N}\rangle = \int d\mathbf{p}\, p_1^{n_1}\ldots p_N^{n_N}\, P(\mathbf{p}) \ ,$$

is the probability for any sequence in which occurrence numbers are given by the vector $\mathbf{n} = (n_1,\ldots,n_N)$. The last form of $p(\mathbf{n})$ thus writes the occurrence probability in the form of a moment of the single-trial probabilities. Notice that the occurrence probabilities for $L$ trials are determined by the $L$th-order moments of $P(\mathbf{p})$. In particular, the marginal probabilities for a single trial,

$$\langle p_i\rangle = \int d\mathbf{p}\, p_i P(\mathbf{p}) \ ,$$

are the first moments of $P(\mathbf{p})$.

An *exchangeable probability assignment* (or an *exchangeable sequence*) is one such that the probability for a sequence does not change under reördering; in other words, all sequences with the same occurrence vector $\mathbf{n}$ have the same probability. Any probability on probabilities leads to an exchangeable probability assignment on the multi-trial hypothesis space. This means that there is a map from probabilities on probabilities to exchangeable probability assignments. The de Finetti representation theorem asserts that any exchangeable probability assignment corresponds to a unique probability on probabilities. Another way of putting this is that the map from probabilities on probabilities to exchangeable probability assignments is one-to-one and onto.

We can get at the uniqueness (i.e., the map is one-to-one) easily. One way to proceed is to define a characteristic function

$$\begin{aligned}\Phi(\mathbf{k}) \equiv \left\langle e^{i\mathbf{k}\cdot\mathbf{p}}\right\rangle &= \int d\mathbf{p}\, e^{i\mathbf{k}\cdot\mathbf{p}} P(\mathbf{p}) \\ &= \sum_{n_1,\ldots,n_N} \frac{i^L}{n_1!\ldots n_N!} k_1^{n_1}\ldots k_N^{n_N}\langle p_1^{n_1}\ldots p_N^{n_N}\rangle \\ &= \sum_{n_1,\ldots,n_N} \frac{i^L}{L!} k_1^{n_1}\ldots k_N^{n_N} p(\mathbf{n})\;.\end{aligned}$$

That $P(\mathbf{p})$ is restricted to the simplex means that for vectors of the form $\mathbf{k} = k(1,\ldots,1)$, the characteristic function becomes $\Phi(\mathbf{k}) = e^{ik}$. Now it is clear why two different probabilities on probabilities cannot lead to the same exchangeable probability assignment: if they did, they would have the same characteristic function and thus, under the inverse Fourier transform, they would be the same.

Another way of putting this is that the polynomials $p_1^{n_1}\ldots p_N^{n_N}$ are linearly independent and complete (but not orthogonal). Thus two different probabilities on probabilities cannot lead to the same exchangeable sequence, for if they did, they would have have the same overlap with this complete set of polynomials and thus would be the same.

Showing that every exchangeable assignment corresponds to a probability on probabilities (the map is onto) requires more work. Suppose, for example, that one uses the occurrence probabilities $p(\mathbf{n})$ to define a characteristic function and then inverts the Fourier transform to get a function $P(\mathbf{p})$. The normalization of the occurrence probabilities implies that $\Phi(\mathbf{k}) = e^{ik}$ for $\mathbf{k} = k(1,\ldots,1)$, which in turn implies that $P(\mathbf{p})$ is restricted to the surface $\sum_i p_i = 1$. The difficulty is that one can't tell from this procedure that $P(\mathbf{p})$ is restricted to positive values of the probabilities—i.e., restricted to the simplex—or, even worse, that it is positive. This difficulty has to be remedied by using some other method. The simplest proof seems to be one due to David Heath and William Sudderth [*The American Statistician* **30**(4), 188–189 (November 1976)], which I sketch here for the case of binary alternatives, the case considered in their paper.

Let $X_1, X_2, \ldots, X_M$ denote the results of $L$ trials of a binary quantity taking on values 0 and 1, and let $p(n,K)$, $K \le L$, be the probability for $n$ 1s in $K$ trials. Exchangeability guarantees that

$$p(n,K) = \binom{K}{n} p(X_1 = 1,\ldots,X_n = 1, X_{n+1} = 0,\ldots,X_K = 0)\;.$$

We can condition the probability on the right on the occurrence of $m$ 1s in all $L$ trials:

$$p(n,K) = \binom{K}{n} \sum_{m=0}^{L} p(X_1 = 1, \ldots, X_n = 1, X_{n+1} = 0, \ldots, X_K = 0 \mid m, L) p(m, L) \ .$$

Given $m$ 1s in $L$ trials, the $\binom{L}{m}$ sequences are equally likely. Thus the situation is identical to drawing without replacement from an urn that has $m$ 1s on $L$ balls, and we have that

$$\begin{aligned}
&p(X_1 = 1, \ldots, X_n = 1, X_{n+1} = 0, \ldots, X_K = 0 \mid m, L) \\
&\qquad = \frac{m}{L}\frac{m-1}{L-1} \cdots \frac{m-(n-1)}{L-(n-1)} \frac{L-m}{L-n} \frac{L-m-1}{L-n-1} \cdots \frac{L-m-(K-n-1)}{L-(K-1)} \\
&\qquad = \frac{(m)_n (L-m)_{K-n}}{(L)_K} \ ,
\end{aligned}$$

where

$$(r)_q \equiv \prod_{j=0}^{q-1} (r-j) = r(r-1)\cdots(r-q+1) = \frac{r!}{(r-q)!} \ .$$

Therefore, we have the main result that

$$p(n,K) = \binom{K}{n} \sum_{m=0}^{L} \frac{(m)_n (L-m)_{K-n}}{(L)_K} p(m,L) \ .$$

The de Finetti representation theorem fails for sequences that are exchangeable for a finite number of trials $L$: for finite exchangeable sequences that can be derived from a probability on probabilities, the probability on probabilities is not unique, and there are finite exchangeable sequences—in particular, anticorrelated sequences such as drawing from an urn without replacement—that cannot be derived from a probability on probabilities. Yet the Heath-Sudderth proof establishes that all finite exchangeable sequences can be derived from mixtures of urn probabilities.

What remains is to take the limit $L \to \infty$. We can write $p(n,K)$ as an integral

$$p(n,K) = \binom{K}{n} \int_0^1 dz \, \frac{(zL)_n \big((1-z)L\big)_{K-n}}{(L)_K} P_L(z) \ ,$$

where

$$P_L(z) = \sum_{m=0}^{L} p(zL, L) \delta(z - m/L)$$

is a distribution concentrated at the $L$-trial frequencies $m/L$. In the limit $L \to \infty$, $P_L(z)$ converges to a continuous distribution on the simplex, and the other term in the integrand goes to $z^n (1-z)^{K-n}$, giving

$$p(n,K) = \binom{K}{n} \int_0^1 dz \, z^n (1-z)^{K-n} P_\infty(z) \ .$$

What we have shown is that if $P(n,K)$ is derived from an infinite exchangeable sequence, then it has a de Finetti representation in terms of a probability distribution on the simplex. The result can readily be extended to nonbinary variables. The conclusion is that a probability on probabilities is just a convenient shorthand for specifying occurrence probabilities on a multi-trial hypothesis space.

The Heath-Sudderth proof is based on the fact that if the multi-trial probabilities are derived from a probability on probabilities $P(p)$, i.e.,

$$p(n,K) = \int_0^1 dp \binom{K}{n} p^n (1-p)^{K-n} P(p) \;,$$

then in the limit of large $N$,

$$\frac{p(n,K)}{1/K} = P(p = n/K) \;;$$

i.e., the probability $p(m,L)$ that in the Heath-Sudderth proof becomes the probability on probabilities is just what it ought to be.

It is interesting to investigate how much information one gains from $L$ trials about the single-trial probabilities $\mathbf{p} = (p_1, \ldots, p_N)$. This information is quantified by the mutual information

$$H(D_L;\mathbf{p}) = H(D_L) - H(D_L|\mathbf{p}) \;,$$

where

$$H(D_L) = -\sum_{\text{sequences}} \langle p_1^{n_1} \ldots p_N^{n_N} \rangle \log\langle p_1^{n_1} \ldots p_N^{n_N} \rangle = -\sum_{n_1,\ldots,n_N} p(\mathbf{n}) \log\langle p_1^{n_1} \ldots p_N^{n_N} \rangle$$

is the Shannon information of the data gathered in $L$ trials and

$$H(D_L|\mathbf{p}) = \int d\mathbf{p}\, P(\mathbf{p}) \left( -L \sum_{i=1}^N p_i \log p_i \right) = -L \sum_{i=1}^N \langle p_i \log p_i \rangle$$

is the conditional information in the $L$-trial data, given the single trial probabilities $\mathbf{p}$. Notice that

$$-L \sum_{i=1}^N \langle p_i \rangle \log\langle p_i \rangle \geq H(D_L) \geq H(D_L|\mathbf{p}) \;,$$

where the first term is the Shannon information for $L$ trials drawn from an i.i.d. governed by the single-trial marginal probabilities $\langle p_i \rangle$. The first inequality is a consequence of the subadditivity of Shannon information.

When the number of trials is small, it is hard to make general statements about the mutual information. If $P(\mathbf{p})$ is concentrated at several widely separated single-trial probabilities $\mathbf{p}$, then it takes only a few trials to begin getting information about which of the widely separated probabilities is generating the data. In contrast, suppose $P(\mathbf{p})$ is concentrated at a particular $\mathbf{p}$ within a small range $\Delta$ for each alternative. In this case it takes many trials to begin getting much information about which single-trial probabilities within the range are generating the data. We can estimate the number of trials required in the following way, where we consider only two alternatives ($N = 2$) for simplicity. After $L$ trials, the data is able to determine $p_1$ to within an uncertainty given roughly by $\sqrt{p_1 p_2/L}$. Thus one would expect to begin getting information about the value of $p_1$ when $\sqrt{p_1 p_2/L} \simeq \Delta$, i.e., when $L \simeq p_1 p_2/\Delta^2$. As $L$ becomes even bigger, i.e., $L \gg p_1 p_2/\Delta^2$, the data is able to distinguish roughly $\Delta/\sqrt{p_1 p_2/L} = \sqrt{L\Delta^2/p_1 p_2}$ values of $p_1$, and the mutual information should be roughly the logarithm of this number of values, i.e.,

$$H(D_L; \mathbf{p}) \sim \log\sqrt{\frac{L\Delta^2}{p_1 p_2}} \,.$$

We can put these considerations on a firm footing by considering the Gaussian approximation to the binomial distribution $p(\mathbf{n}|\mathbf{p})$. The Gaussian approximation requires that for each alternative $i$, the number of trials is large enough that $\sqrt{p_i/L} \ll p_i$, i.e., $p_i L \gg 1$, for all probabilities $\mathbf{p}$ that have substantial support in $P(\mathbf{p})$. If we further assume that the number of trials is large enough that the data can distinguish all the features of $P(\mathbf{p})$—i.e., for each alternative, $P(\mathbf{p})$ does not vary significantly on the scale $\sqrt{p_i/L}$—then it is a tedious, but straightforward computation to show that

$$\langle p_1^{n_1} \cdots p_N^{n_N} \rangle = \left(\frac{n_1}{L}\right)^{n_1} \cdots \left(\frac{n_N}{L}\right)^{n_N} P\Big(\mathbf{p} = \frac{\mathbf{n}}{L}\Big) \left(\frac{2\pi}{L}\right)^{(N-1)/2} \sqrt{\frac{n_1}{L} \cdots \frac{n_N}{L}}$$

and

$$p(\mathbf{n}) = \frac{L!}{n_1! \ldots n_N!} \langle p_1^{n_1} \cdots p_N^{n_N} \rangle = \frac{1}{L^{N-1}} P\Big(\mathbf{p} = \frac{\mathbf{n}}{L}\Big) \,,$$

which leads to a mutual information

$$H(D_L; \mathbf{p}) = -\int d\mathbf{p}\, P(\mathbf{p}) \log\big(P(\mathbf{p})\mathcal{V}(\mathbf{p})\big) \,, \tag{1}$$

where

$$\mathcal{V}(\mathbf{p}) = \left(\frac{2\pi}{L}\right)^{(N-1)/2} \sqrt{p_1 \cdots p_N}$$

is a probability-dependent volume element on the probability simplex, which can be thought of as the distinguishability volume determined by $L$ trials. The mutual information (1) has the following interpretation: bin the probabilities $\mathbf{p}$ according to the volume element $\mathcal{V}(\mathbf{p})$; the mutual information is the Shannon information for the discrete distribution obtained by replacing the continuous distribution $P(\mathbf{p})$ by the distribution of probabilities for the bins. Another way of saying this is that the mutual information (1) is the entropy of $P(\mathbf{p})$ relative to a position-dependent measure $m(\mathbf{p}) = 1/\mathcal{V}(\mathbf{p})$, which describes the position-dependent distinguishability of distributions $\mathbf{p}$.

In the aforementioned example, where $P(\mathbf{p})$ is concentrated at a particular $\mathbf{p}$, each probability having a small range $\Delta$ of possible values, the mutual information (1) becomes

$$H(D_L; \mathbf{p}) = \log\bigg(\frac{\Delta^{N-1}}{\mathcal{V}(\mathbf{p})}\bigg) = \log\bigg(\frac{(L\Delta^2/2\pi)^{(N-1)/2}}{\sqrt{p_1 \cdots p_N}}\bigg) \,,$$

which simplifies to the estimate above for $N=2$. Actually, this example is flawed because it requires one probability, say $p_N$, to vary over a range $(N-1)\Delta$. We can do a better job of taking into account the volume on the simplex by using a Gaussian

$$P(\mathbf{p}) = \frac{\sqrt{N}}{(2\pi\Delta^2)^{(N-1)/2}} \exp\left( -\sum_{i=1}^{N} \frac{(p_i - q_i)^2}{2\Delta^2} \right) ,$$

in which case the mutual information (1) becomes

$$H(D_L;\mathbf{p}) = \log\left( \frac{(2\pi e\Delta^2)^{(N-1)/2}/\sqrt{N}}{\mathcal{V}(\mathbf{q})} \right) = \log\left( \frac{(eL\Delta^2)^{(N-1)/2}}{\sqrt{N q_1 \dots q_N}} \right) .$$

In the first form here, the numerator within the logarithm can be thought of as the volume occupied by the Gaussian. The $\sqrt{N}$ is the correction to the volume that comes from projecting onto the simplex.

Notice that another neat way to write the mutual information (1) comes from introducing a Wootters distinguishability metric

$$ds^2 = 4\sum_{i=1}^{N} (d\sqrt{p_i})^2 = \sum_{i=1}^{N} \frac{dp_i^2}{p_i} .$$

The volume element for the Wootters metric is

$$\begin{aligned} d^W\mathbf{p} &= \delta\left( \sum_{i=1}^{N} (\sqrt{p_i})^2 - 1 \right) 2^N d\sqrt{p_1} \dots d\sqrt{p_N} \\ &= \delta\left( \sum_{i=1}^{N} p_i - 1 \right) \frac{dp_1 \dots dp_N}{\sqrt{p_1 \dots p_N}} \\ &= \frac{d\mathbf{p}}{\sqrt{p_1 \dots p_N}} . \end{aligned}$$

Redefining the probability $P(\mathbf{p})$ in terms of the Wootters metric,

$$P^W(\mathbf{p}) d^W\mathbf{p} = P(\mathbf{p}) d\mathbf{p} ,$$

gives

$$P^W(\mathbf{p}) = \sqrt{p_1 \dots p_N} P(\mathbf{p}) \quad \Longrightarrow \quad P(\mathbf{p})\mathcal{V}(\mathbf{p}) = \left( \frac{2\pi}{L} \right)^{(N-1)/2} P^W(\mathbf{p}) ,$$

and the mutual information (1) becomes

$$H(D_L;\mathbf{p}) = -\int d^W\mathbf{p}\, P^W(\mathbf{p}) \log\left( \left( \frac{2\pi}{L} \right)^{(N-1)/2} P^W(\mathbf{p}) \right) .$$

Since the Wootters metric is based on distinguishability from data in many trials, the mutual information becomes the information of $P^W(\mathbf{p})$ relative to an $L$-dependent, but position-independent measure $m^W(\mathbf{p}) = (L/2\pi)^{(N-1)/2}$.