International Statistical Review (2009), 77, 2, 241–250 doi:10.1111/j.1751-5823.2008.00059.x

Exchangeability, Correlation, and Bayes' Effect

Ben O'Neill

School of Physical, Environmental and Mathematical Sciences, University of New South Wales, Australian Defence Force Academy, Canberra ACT 2600, Australia
E-mails: ben.oneill@hotmail.com, ben.oneill@adfa.edu.au

Summary

We examine the difference between Bayesian and frequentist statistics in making statements about the relationship between observable values. We show how standard models under both paradigms can be based on an assumption of exchangeability and we derive useful covariance and correlation results for values from an exchangeable sequence. We find that such values are never negatively correlated, and are generally positively correlated under the models used in Bayesian statistics. We discuss the significance of this result as well as a phenomenon which often follows from the differing methodologies and practical applications of these paradigms – a phenomenon we call Bayes' effect.

Key words: Bayesian statistics; frequentist statistics; exchangeability; independence; correlation; pseudo-correlation; Bayes' effect.

1 Independence in the Bayesian and Frequentist Paradigms

Bayesian statistics differs from frequentist statistics in its treatment of unknown values. Bayesian statistics regards probability as an epistemic concept without any necessary metaphysical analogue (see de Finetti, 1974, pp. xi–xii). This epistemic interpretation, coupled with assumptions of completeness of preferences and various measurement assumptions, implies the existence of some non-degenerate probabilistic belief for any unknown value (see Fishburn, 1986; Bernardo & Smith, 1994, pp. 13–104). Under this approach unknown parameters are given a prior probability distribution, reflecting the prior beliefs and uncertainty of the observer. This contrasts with the frequentist paradigm where parameters are regarded as unknown constants. Indeed, under the epistemic interpretation, the notion of an unknown constant is a contradiction in terms.

Standard models under both the Bayesian and frequentist paradigm treat observable values as infinitely exchangeably extendable and therefore independent and identically distributed (IID) conditional on some unknown parameter (see below for discussion). However, the different treatment of the unknown parameter leads to different results regarding the relationship between observable values. In particular, frequentists regard the observable values as independent, whereas Bayesians do not. In this paper, we examine the relationship between exchangeability and correlation under each paradigm and we discuss a phenomenon we call *Bayes' effect*. We begin with an examination of exchangeability as the basis for standard models under either paradigm.

2 Exchangeability and the Representation Theorem

We let $\mathbf{x} \equiv (x_1, x_2, x_3, \ldots)$ be a countably infinite sequence of real values and we then let $\mathbf{x}_k \equiv (x_1, x_2, \ldots, x_k)$ for all $k \in \mathbb{N}$.

Definition 1 *(Exchangeability).* If the probability measure for $\mathbf{x}_k$ is invariant under permutations of the elements of $\mathbf{x}_k$, then we say that $\mathbf{x}_k$ is *exchangeable*. If $\mathbf{x}_k$ is exchangeable for all $k \in \mathbb{N}$, then we say that $\mathbf{x}$ is *exchangeable* (also called *infinitely exchangeable*, though the latter is a redundancy since we are already referring to an infinite sequence). If $\mathbf{x}_k$ can be embedded in an exchangeable sequence $\mathbf{x}$, then we say that $\mathbf{x}_k$ is *infinitely exchangeably extendable*.

As stated above, both the Bayesian and frequentist paradigm treat observable values as infinitely exchangeably extendable. This assertion may be unfamiliar to many. After all, the standard modelling form of IID values conditional on some unknown "parameter" is often taken as the starting point for Bayesian or frequentist analysis, and exchangeability is seldom invoked (in fact, many are not even aware of what the property of exchangeability is).

This standard methodology is contrary to the *operational* approach to statistics, under which we seek to model our beliefs about observables by making structural assumptions *about those observables*. Following the operational method, Bernardo & Smith (1994) explain that:

> In much statistical writing, the starting point for formal analysis is the *assumption* of a mathematical model form, typically involving "unknown parameters", the main object of the study being to infer something about the value of these parameters. From our perspective, this is all somewhat premature and mysterious! We are seeking to represent degrees of belief about observables . . . (p. 167)

Under the operational approach, it can be shown that assumptions of *invariance* about observable values can lead us to the standard model forms of IID observations from some standard probability distribution. These invariance conditions involve an assumption that our probabilistic beliefs are invariant with respect to some aspect of the observable values. In particular, the invariance condition of exchangeability corresponds to the belief that the order of our observations is irrelevant.

The assumption of exchangeability leads us to our standard model as follows. For all $k \in \mathbb{N}$, we let F_k be the *empirical distribution of* $\mathbf{x}_k$ defined by

$$F_k(t) \equiv \frac{1}{k} \sum_{i=1}^{k} I(x_i \leq t) \text{ for all } \quad t \in \mathbb{R}.$$

We then let $F_{\mathbf{x}}$ be the *empirical distribution of* $\mathbf{x}$ defined as set out in Definition 2 in the Appendix. Under this definition we have

$$F_{\mathbf{x}}(t) = \lim_{k \to \infty} \frac{1}{k} \sum_{i=1}^{k} I(x_i \leq t) \text{ for all } \quad t \in \mathbb{R},$$

so long as this limit exists (i.e. the function $F_{\mathbf{x}}$ is an extension of this limit function).

THEOREM 1 (REPRESENTATION THEOREM). *If* $\mathbf{x}$ *is exchangeable, then it follows that the elements of* $\mathbf{x}|F_{\mathbf{x}}$ *are independent with sampling distribution* $F_{\mathbf{x}}$ *(i.e. the sampling distribution is the empirical distribution of* $\mathbf{x}$*) so that for all* $k \in \mathbb{N}$ *we have*

$$F(\mathbf{x}_k) = \int \prod_{i=1}^{k} F_{\mathbf{x}}(x_i) dP(F_{\mathbf{x}}).$$

If $F_{\mathbf{x}}$ is indexed by another parameter θ (and with densities defined), then we have

$$p(\mathbf{x}_k) = \int \prod_{i=1}^{k} p(x_i|\theta) dP(\theta).$$

Proof. This is an expression of the celebrated Representation Theorem of de Finetti (1980). See Appendix for an outline proof.

The Representation Theorem shows that if **x** is exchangeable, then the elements of **x** are IID conditional on the empirical distribution of **x** (or equivalently, conditional on any other parameter indexing this distribution).

This resulting Bayesian model form contrasts with the frequentist model form which takes the parameter as an unknown constant. Frequentists generally argue directly for such a model form on the grounds of random sampling, although implicit in such arguments is an assertion of exchangeability. Indeed, although frequentists rarely purport to justify this model on the grounds of exchangeability, it is easy to see that this model form does indeed follow from the Representation Theorem above, coupled with the frequentist practice of treating parameters as unknown constants.

COROLLARY 1. *If* **x** *is exchangeable and if $F_{\mathbf{x}}$ is an unknown constant (so that $F_{\mathbf{x}}$ has a point mass distribution on its own unknown value), then for all $k \in \mathbb{N}$ we have*

$$F(\mathbf{x}_k) = \prod_{i=1}^{k} F_{\mathbf{x}}(x_i).$$

If $F_{\mathbf{x}}$ is indexed by a parameter θ (so that θ has a point mass distribution on its own true but unknown value θ_{T}), then we have

$$p(\mathbf{x}_k) = \prod_{i=1}^{k} p(x_i|\theta = \theta_{\mathrm{T}}).$$

Corollary 1 shows us that the standard IID frequentist model can be justified by an assumption of exchangeability of the sequence of observable values, coupled with treating the unknown parameter (indexing the empirical distribution) as having a point mass distribution on its own unknown value (that is, treating it as an unknown constant). However, we can see that, under this approach, the unconditional distribution of the observable values is still a function of the unknown parameter.

The Representation Theorem demonstrates that it is our judgement of exchangeability of the sequence of observable values that underlies our use of the standard statistical model involving IID observable values based on an unknown parameter. In Bayesian modelling, we regard the unknown parameter as a random variable with the observable values being IID conditional on this parameter. In frequentist modelling, we regard the unknown parameter as a constant with the observable values being unconditionally IID.

Frequentists may retort that they do not in fact assert unconditional independence of the observable values, since they only ever refer to densities that are a function of the parameters. They may reject the notion that the parameter can be taken to have a density at all since the parameter is not repeatable and is therefore not amenable to frequentist probability statements.

However, even in this case, frequentists *do* explicitly assert that the observable values are independent *regardless* of the true parameter value. It follows from an elementary application of *modus ponens* that the observable values are independent, unconditional on any assertion of the true value of the parameter.

3 Exchangeability, Correlation, and Bayesian Prediction

If $\mathbf{x}$ is exchangeable, then the elements of $\mathbf{x}$ are independent if and only if the empirical distribution (or equivalently, any other parameter indexing the empirical distribution) is almost surely constant. Of course, under the epistemic interpretation of probability, this is only the case when the empirical distribution is known. Thus, under the Bayesian paradigm, when the empirical distribution is unknown (which is always the case in problems of interest), the elements of $\mathbf{x}$ will be dependent. This immediately explains why we use observed values in our predictions of future values.

We have said that Bayesians regard conditionally independent values as predictive of one another insofar as they give information on the underlying parameters. Since any observation means that the empirical distribution is more likely to put greater mass on this observation than is the case *a priori* this suggests that unconditionally, the observable values should be positively related. Indeed, we can demonstrate that, under wide conditions, such observations will be positively correlated.

THEOREM 2. *If $\mathbf{x}$ is exchangeable, then for all $i \neq j$ we have*

$$\mathrm{Cov}(x_i, x_j) = \mathrm{Var}(\mu(\theta)) \geq 0 \text{ and } \mathrm{Corr}(x_i, x_j) = \frac{\mathrm{Var}(\mu(\theta))}{\mathrm{Var}(\mu(\theta)) + \mathrm{E}(\sigma^2(\theta))} \geq 0$$

where $\mu(\theta) \equiv \mathrm{E}(x|\theta)$ *and* $\sigma^2(\theta) \equiv \mathrm{Var}(x|\theta)$.

Proof. Since $\mathbf{x}$ is exchangeable, it follows from Theorem 1 that the elements of $\mathbf{x}$ are IID conditional upon $F_{\mathbf{x}}$, or equivalently, conditional upon θ. It follows that

$$\begin{aligned}\mathrm{Cov}(x_i, x_j) &= \mathrm{E}(x_i x_j) - \mathrm{E}(x_i)\mathrm{E}(x_j)\\ &= \mathrm{E}(\mathrm{E}(x_i x_j|\theta)) - \mathrm{E}(\mathrm{E}(x_i|\theta))\mathrm{E}(\mathrm{E}(x_j|\theta))\\ &= \mathrm{E}(\mu(\theta)^2) - \mathrm{E}(\mu(\theta))^2\\ &= \mathrm{Var}(\mu(\theta)) \geq 0,\end{aligned}$$

so that

$$\mathrm{Corr}(x_i, x_j) = \frac{\mathrm{Cov}(x_i, x_j)}{\mathrm{Var}(x)} = \frac{\mathrm{Var}(\mu(\theta))}{\mathrm{Var}(\mu(\theta)) + \mathrm{E}(\sigma^2(\theta))} \geq 0.$$

COROLLARY 2. *If $\mathbf{x}$ is exchangeable and if* $\mathrm{Var}(\mu(\theta)) > 0$, *then the elements of $\mathbf{x}$ are positively correlated.*

Theorem 2 establishes that the elements of an exchangeable sequence can never be negatively correlated and are uncorrelated only when $\mu(\theta)$ is almost surely constant (so that they are positively correlated if $\mu(\theta)$ is not almost surely constant). The former result was observed by Kingman (1978). Of course, under the frequentist paradigm, $\mu(\theta)$ is an unknown constant so that the observable values are indeed uncorrelated, and, in fact, independent (see Corollary 1).

However, under the Bayesian paradigm it is often the case that $\mathrm{Var}(\mu(\theta)) > 0$, so that the observable values are positively correlated. Of course, this is not always the case, even for standard models.

Example 1. Suppose that $\mathbf{x}$ is exchangeable with spherical symmetry and with precision θ defined by $1/\theta \equiv \lim_{k\to\infty} \sum_{i=1}^{k} x_i^2/k$. Then it follows from Lemma 9 of Freedman (1963) that $p(x|\theta) = \mathrm{N}(x|0, \theta)$ so that $\mu(\theta) = 0$ and $\sigma^2(\theta) = 1/\theta$. In this case, even under the Bayesian approach of treating θ as a random variable, the elements of $\mathbf{x}$ are uncorrelated (though they are not independent unless θ is known almost surely).

4 Frequentist Prediction and Pseudo-Correlation

Under the standard frequentist model, we assume that the elements of $\mathbf{x}$ are independent with sampling density dependent upon the parameter θ. We have shown that this is equivalent to assuming that $\mathbf{x}$ is exchangeable coupled with treatment of θ as having a point mass density on its own true value θ_{T}. In either case, we obtain the model

$$p(\mathbf{x}_k) = \prod_{i=1}^{k} p(x_i|\theta = \theta_{\mathrm{T}}).$$

This model gives us the predictive density

$$p(x_{k+1}|\mathbf{x}_k) = p(x_{k+1}) = p(x_{k+1}|\theta = \theta_{\mathrm{T}}).$$

Thus, the frequentist is able to conclude that the elements of $\mathbf{x}$ are independent regardless of the true value of θ and in particular that x_{k+1} and $\mathbf{x}_k$ are independent. As such, one might expect that observed values would not be used in predicting unknown values.

However, since frequentists take the underlying parameter θ as an unknown constant, they are unable to obtain the above densities exactly. Instead, the frequentist must be content with an estimate of these densities using an estimate of θ that is based on the observed value $\mathbf{x}_k$. Thus, the frequentist estimates the predictive density as

$$\hat{p}(x_{k+1}|\mathbf{x}_k) \equiv p(x_{k+1}|\theta = \hat{\theta}(\mathbf{x}_k)),$$

where $\hat{\theta}(\mathbf{x}_k)$ is some estimate of θ based on observing $\mathbf{x}_k$. Thus, we see that, in practice, frequentists do indeed use observed values of $\mathbf{x}_k$ in predicting unknown values so that they *effectively* treat the elements of $\mathbf{x}$ as dependent notwithstanding their model assumptions to the contrary.

This result may seem incongruous. Since independence means that conditioning on observed values does not affect the probability measure for unobserved values, how can the frequentist legitimately use the former in predicting the latter? The answer appears to be that the frequentist *does* accept unconditional independence but is nonetheless forced to use only an estimate of the predictive density. This means that the estimated behaviour (including the obvious dependence between observable values) is indeed inaccurate, but this is only natural given the limited data. Indeed, if the estimator $\hat{\theta}$ is strongly consistent, then $\lim_{k\to\infty} \hat{p}(x|\mathbf{x}_k) = p(x|\theta = \theta_{\mathrm{T}})$ almost surely so that, as we obtain perfect information, our predictive density becomes less and less dependent upon the idiosyncrasies of the observed data.

We have seen that both Bayesians and frequentists use observed data in their predictions of unobserved data. Bayesians do so by virtue of an explicit assertion of dependence (usually positive correlation) between these values. Frequentists assert that the vales are independent but that they use these values in practice to estimate the unknown densities involved. Essentially, Bayesians treat the observable values as dependent *de jure*, whereas frequentists treat these values as dependent *de facto*.

This may seem like a trivial difference. In either case, observed values are used to make predictions regarding unobserved values. However, in addition to avoiding any apparent incongruity between model assumptions and results, the explicit treatment of observable values as dependent under the Bayesian paradigm allows Bayesians to *quantify* the degree to which observed values affect predictions of unobserved values.

Example 2. Suppose that $\mathbf{x}$ is exchangeable with elements having range $\{0, 1\}$. It follows from Theorem 1 and the strong law of large numbers that

$$p(\mathbf{x}_k|\theta) = \theta^{n(\mathbf{x}_k)}(1-\theta)^{k-n(\mathbf{x}_k)},$$

where $n(\mathbf{x}_k) \equiv \sum_{i=1}^{k} x_i$ and $\theta \equiv \theta(\mathbf{x}) \equiv \lim_{k\to\infty} n(\mathbf{x}_k)/k$. It can easily be shown that $\mu(\theta) = \theta$ and $\sigma^2(\theta) = \theta(1-\theta)$, so that we have

$$\mathrm{Corr}(x_i, x_j) = \frac{\mathrm{Var}(\theta)}{\mathrm{Var}(\theta) + \mathrm{E}(\theta(1-\theta))} = \frac{\mathrm{E}(\theta^2) - \mathrm{E}(\theta)^2}{\mathrm{E}(\theta) - \mathrm{E}(\theta)^2}.$$

So, if our prior beliefs are such that $\theta \sim \mathrm{Be}(\alpha, \beta)$, then it can easily be shown that $\mathrm{Corr}(x_i, x_j) = 1/(1 + \alpha + \beta)$. This gives us an explicit expression for the correlation between our observable values, and demonstrates that the correlation can potentially be quite high (if $\alpha + \beta$ is low).

Frequentists are left in a most unsatisfactory position. They simultaneously assert the independence of observable values while using them to make predictions of one another. If pressed about this *pseudo-correlation*, they must explain that this is merely a practical measure to approximate the unknown predictive density. However, when asked about the degree to which this practical measure causes observed values to affect the predictions, they are unable to give a satisfactory answer.

5 Constants Versus Random Parameters: Bayes' Effect

Aside from the fact that Bayesians are able to make explicit assertions of positive correlation between observations, there is another difference that often arises between the Bayesian and frequentist paradigms. In dealing with parameters that we believe *a priori* are probably close to a particular value, Bayesians are able to accommodate this belief in the form of a prior measure concentrated narrowly around this value. However, since frequentists view parameters as constants, they are unable to accommodate such information in any probabilistic sense. Instead they are left with a choice: discard the information and treat the parameter as an unknown constant as normal, or use the information to treat the parameter as a *known* constant or a constant that is known to fall within some narrow range. In practice, where this prior information is strong, frequentists are often inclined to take the parameter value as known and ignore the uncertainty. This is the case in many gambling examples where frequentists generally assume that the gambling device in question is fair (that is, that the long-run proportions of various outcomes are equal) and is not subject to any possible imperfections that may cause bias.

We have seen from Theorems 1 and 2 that, for any exchangeable sequence, treating the associated empirical distribution as a random variable rather than a constant induces some dependence (usually positive correlation) between the observable values. We call this phenomenon *Bayes' effect* since it follows from the Bayesian practice of treating the parameters as random variables. Moreover, unlike the case where the frequentist uses an *unknown* constant, the positive correlation induced by treating the parameters as random variables is not merely an explicit substitute for pseudo-correlation – it is a substitute for genuine independence.

Bayes' effect explains why Bayesians are more inclined to view observable values as informative of one another. In gambling examples, frequentists regard outcomes as independent and generally take the long-run proportions of various outcomes as equal – without the possibility of bias – so that, even *in practice*, no information on future outcomes is gained from past outcomes. Bayesians on the other hand, are able to accommodate slight uncertainty in these values by the use of an appropriately concentrated prior measure.

6 A Simple Example Involving Bias in Coin Tossing

There is both theoretical and empirical evidence that suggests that flipping a coin and catching it (so that it does not bounce) results in a random process that is close to "fair", in the sense that the long-run proportions of heads and tails will be close to equal. Using a theoretical analysis based on an idealized physical model, Keller (1986) finds that this occurs so long as either the initial vertical velocity or the initial angular momentum for the coin flip is sufficiently high to ensure a large number of revolutions of the coin before it lands (see Theorem 1 of Keller, 1986 for a more precise statement of this result). Keller includes gravity in his model (or else the coin would not land) but ignores air resistance or any other form of friction which would diminish the vertical velocity or angular momentum of the coin during the flip.

In empirical experiments where the number of revolutions of the coin is of course finite, Diaconis *et al.* (2007) find a small bias for coin flipping based on the side that begins the flip facing up. Gelman & Nolan (2002) also analyse empirical trials of coin flipping and find that even weighting of the coins does not result in any significant bias. However, they also find that large biases can occur due to weighting when a coin is spun instead of flipped or when a coin is flipped onto a hard surface so that it is able to bounce.

Both Bayesians and frequentists are able to test claims of bias through empirical trials using hypothesis testing (Bayesian and frequentist hypothesis testing, respectively). Indeed, when the Belgian Euro coin was minted, there were claims that the weighting of the coin was such that it was biased, and this was initially tested using a classical hypothesis test with a small amount of data (see MacKensie, 2002).

Example 3. Statisticians Tomasz Gliszczynski and Waclaw Zawadowski conducted repeated trials spinning the Belgian Euro coin to test for bias – in 250 trials they obtained 140 heads and only 110 tails (see MacKensie, 2002). Suppose that $\mathbf{x}$ is the sequence of outcomes of spins of a single Belgian Euro coin, where

$$x_i = \begin{cases} 1 \text{ if the } i\text{-th spin is a head} \\ 0 \text{ if the } i\text{-th spin is a tail} \end{cases}.$$

Let $n(\mathbf{x}_k) \equiv \sum_{i=1}^{k} x_i$ be the number of heads in the first k coin spins and let $\theta \equiv \theta(\mathbf{x}) \equiv \lim_{k\to\infty} n(\mathbf{x}_k)/k$ be the long-run proportion of heads in the sequence. If $\mathbf{x}$ is exchangeable, then we have $n(\mathbf{x}_k) \sim \text{Bi}(k, \theta)$. Suppose that, on the basis of our experience with other coins and our knowledge of the mechanics of coin spinning, we believe *a priori* that the coin is likely to be close to unbiased, but we are not certain that it is completely unbiased. With 140 heads out of 250 spins, it is unlikely that we would be convinced of the existence of bias on the basis of a classical hypothesis test, since this data yields a p-value of 0.0664, which is not particularly small. However, it is also difficult to see why this result should lead us to reject the *possibility* of bias in contradiction to our prior belief. Moreover, we would be unlikely to accept the maximum likelihood estimate $\hat{\theta}_{\text{MLE}} = 140/250 = 0.56$ as a sensible estimate of the long-run proportion of heads, because our prior beliefs suggest that such a large bias is implausible.

By treating the unknown parameter as a constant, the frequentist is left in a bit of a quandary in attempting to predict the results of future coin spins, since he cannot properly incorporate prior beliefs about that parameter into the analysis. In practice, unless a hypothesis test reveals statistically significant evidence of bias, many frequentists would simply ignore their uncertainty and assume that the coin is unbiased, whereas under the Bayesian paradigm we could instead incorporate our initial uncertainty into our analysis by using a prior density that is narrowly concentrated around our prior expectation.

Example 4. Continuing Example 3, suppose that the frequentist ignores any uncertainty about θ and assumes that $\theta = \theta_{\text{T}} = 1/2$, whereas the Bayesian models his prior uncertainty using the symmetrical beta prior density:

$$p(\theta) = p(\theta|\delta) \propto [\theta(1-\theta)]^{(1-12\delta^2)/8\delta^2}$$

for some known $0 < \delta < 1/2$, so that $\text{E}(\theta) = 1/2$, $\text{Var}(\theta) = \delta^2$ and $\text{Corr}(x_i, x_j) = 4\delta^2$ for all $i \neq j$. The frequentist regards and treats the coin tosses as independent, whereas the Bayesian regards and treats the coin tosses as positively correlated. The frequentist determines that the probability that the next toss is heads is:

$$p_F \equiv p(x_{k+1} = 1|n(\mathbf{x}_k), \theta = 1/2) = 1/2,$$

whereas the Bayesian determines that this probability is:

$$p_B \equiv p(x_{k+1} = 1|n(\mathbf{x}_k), \delta) = \frac{1/2 + 4\delta^2(n(\mathbf{x}_k) - 1/2)}{1 + 4\delta^2(k-1)}.$$

Suppose that the Bayesian is sufficiently sceptical of bias in coin spinning that he takes $\delta = 0.01$. Then, $p_F = 0.5$ and $p_B = 0.5054565$, so that the frequentist will regard heads and tails as equally likely whereas the Bayesian will regard the next toss as being slightly more likely to be heads than tails, because this is the outcome that has occurred the most frequently.

The principle of predicting the outcome that occurs the most frequently in exchangeable sequences is not contingent on the specific model form used in these examples. Indeed, O'Neill & Puza (2005) show that this result occurs for any symmetrical prior belief, regardless of its particular form. They therefore advocate that, for processes like coin tossing – which are designed to produce independent and uniformly distributed outcomes but which may be subject to imperfections leading to bias – ignorance as to the direction of the bias should lead us to predict the outcome that has occurred the most.

These examples demonstrate Bayes' effect in action. By allowing for uncertainty in the underlying parameters in situations where the frequentist is willing to ignore this uncertainty, the Bayesian judges observable values from an exchangeable sequence to be positively correlated with one another, such that he uses past observations to predict future outcomes, in full awareness of their correlation.

References

Bernardo, J.M. & Smith, A.F.M. (1994). *Bayesian Theory*. Chichester: John Wiley.
Chow, Y.S. & Teicher, H. (1988). *Probability Theory*, 2nd ed. Berlin: Springer.
de Finetti, B. (1974). *Theory of Probability*. Chichester: Wiley.

de Finetti, B. (1980). Foresight; its logical laws, its subjective sources. In *Studies in Subjective Probability*, Eds. H.E. Kyberg and H.E. Smokler, pp. 93–158, New York: Dover.
Diaconis, P., Holmes, S. & Montgomery, R. (2007). Dynamical bias in the coin toss. *Soc. Ind. Appl. Math. Rev.*, **49**, 211–235.
Fishburn, P.C. (1986). The axioms of subjective probability. *Statist. Sci.*, **1**(3), 335–345.
Fortini, S., Ladelli, L. & Regazzini, E. (2000). Exchangeability, predictive distributions and parametric models. *Sankhyā Indian J. Stat.*, **62A**, 86–109.
Freedman, D.A. (1963). Invariants under mixing which generalize de Finetti's theorem: Continuous time parameter. *Ann. Math. Stat.*, **34**(4), 1194–1216.
Galambos, J. (1982). Exchangeability. In *Encyclopedia of Statistical Sciences*, Eds. S. Kotz & N.L. Johnson, pp. 573–577, New York: Wiley.
Gelman, A. & Nolan, D. (2002). You can load a die, but you can't bias a coin. *Amer. Statist.*, **56**(4), 308–311.
Keller, J.B. (1986). The probability of heads. *Amer. Math. Monthly*, **93**(3), 191–197.
Kingman, J.F.C. (1978). Uses of exchangeability. *Ann. Probab.*, **6**(2), 183–197.
MacKensie, D. (2002). Euro coin accused of unfair flipping. *New Sci.* Available at http://www.newscientist.com/article/dn1748-euro-coin-accused-of-unfair-flipping.html.
Morisson, T.J. (2000). *Functional Analysis: an Introduction to Banach Space Theory*. New York: Wiley.
O'Neill, B. & Puza, B.D. (2005). In defence of the reverse gambler's belief. *Math. Sci.*, **30**(1), 13–16.

Résumé

Nous examinons la différence entre les statistiques Bayesiennes et fréquentistes dans des propositions sur la relation entre valeurs observées. Nous démontrons comment les modèles normaux dans les deux cas peuvent être basés sur la supposition d'échangeabilité, et nous obtenons quelques résultats utiles sur la covariance et la corrélation pour des valeurs dans une suite échangeable. Ces valeurs ne sont jamais corrélées négativement, et sont en général corrélées positivement dans les modèles Bayesiens. Nous discutons la signification de ce résultat, ainsi que celui du phénomène qui s'ensuit lorsqu'on emploie ces deux méthodologies, un phénomène que nous appelons l'effet de Bayes.

Appendix

Definition 2. It is tempting to define $F_{\mathbf{x}} = \lim_{k\to\infty} F_k$, so that $F_{\mathbf{x}}$ is the Cesàro limit of the indicator functions $I(x_1 \le t), I(x_2 \le t), \ldots$, which is given by

$$C_{\mathbf{x}}(t) \equiv \lim_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} I(x_k \le t) \text{ for all } \quad t \in \mathbb{R}.$$

Unfortunately, this limit does not exist for all possible values of $\mathbf{x}$. Instead, we can define $F_{\mathbf{x}}$: $\mathbb{R} \to [0,1]$ as any Banach limit that extends the Cesàro limit $C_{\mathbf{x}}$ which can be shown to exist for all possible values of $\mathbf{x}$. Detailed discussion of the Banach limit and the Hahn–Banach theorem is beyond the scope of this paper, but can be found in most texts on functional analysis including Morisson (2000).

Proof of Theorem 1 *(Representation Theorem).* For brevity, we give only a heuristic proof based on Chow & Teicher (1988) and adapted from Bernardo & Smith (1994), pp. 177–179. A similar outline proof can be found in Galambos (1982). A more rigorous treatment can be found in Fortini *et al.* (2000). Since $\mathbf{x}$ is exchangeable, it follows that the indicators $I(x_1 \le t)$, $I(x_2 \le t), \ldots$ must also be exchangeable. It then follows that for all $k \in \mathbb{N}$ and $m \in \mathbb{N}$ we have

$$\mathrm{E}[(F_k(t) - F_{k+m}(t))^2] = \frac{m}{k+m}\frac{1}{k}[P(x_1 \le t) - P(x_1 \le t, x_2 \le t)] \le \frac{1}{k} \text{ for all } \quad t \in \mathbb{R}.$$

This implies that the sequence $F_1, F_2, F_3, \ldots$ is a Cauchy sequence which converges in probability to some random function. But since $F_{\mathbf{x}} = C_{\mathbf{x}} = \lim_{k\to\infty} F_k$ wherever the latter exists, it follows that $F_1, F_2, F_3, \ldots$ converges to $F_{\mathbf{x}}$ and therefore also converges in probability to $F_{\mathbf{x}}$. It then follows that for all $k \in \mathbb{N}$ and $m \in \mathbb{N}$, we have

$$\int \prod_{i=1}^{k} F_{\mathbf{x}}(x_i) dP(F_{\mathbf{x}}) = \lim_{m\to\infty} \int \prod_{i=1}^{k} F_{k+m}(x_i) dP(F_{k+m}).$$

Let $I_\alpha(\mathbf{x}_k) \equiv I(x_{\alpha_1} \leq x_1, \ldots, x_{\alpha_k} \leq x_k)$ and let $\mathcal{N}_{k,m} \equiv \{\alpha \in \mathbb{N}_{k+m}^k : (\forall i \neq j) : \alpha_i \neq \alpha_j\}$. For all $k \in \mathbb{N}$ and $m \in \mathbb{N}$, we have

$$\prod_{j=1}^{k} F_{k+m}(x_j) = \left(\frac{1}{k+m}\right)^k \prod_{j=1}^{k} \sum_{i=1}^{k+m} I(x_i \leq x_j) = \left(\frac{1}{k+m}\right)^k \sum_{\alpha \in \mathbb{N}_{k+m}^k} I_\alpha(\mathbf{x}_k) = \frac{B(m) + B^*(m)}{(k+m)^k},$$

where $B(m) \equiv \sum_{\alpha \in \mathbb{N}_{k+m}^k - \mathcal{N}_{k,m}} I_\alpha(\mathbf{x}_k)$ and $B^*(m) \equiv \sum_{\alpha \in \mathcal{N}_{k,m}} I_\alpha(\mathbf{x}_k)$. We have

$$B\,(m) \leq \text{card}\left(\mathbb{N}_{k+m}^k - \mathcal{N}_{k,m}\right) = (k+m)^k \left(1 - \prod_{j=1}^{k} \left(\frac{m+j}{m+k}\right)\right),$$

so that $\lim_{m\to\infty} B(m)/(k+m)^k = 0$. Now, since $\mathbf{x}$ is exchangeable, we also have $\int I_\alpha(\mathbf{x}_k) dP(F_{k+m}) = F(\mathbf{x}_k)$ for all $\alpha \in \mathcal{N}_{k,m}$, so that

$$\frac{\int B^*\,(m)\,dP\,(F_{k+m})}{(k+m)^k} = F(\mathbf{x}_k) \prod_{j=1}^{k} \left(\frac{m+j}{m+k}\right).$$

Using these results, we then have

$$\begin{aligned}\int \prod_{i=1}^{k} F_{\mathbf{x}}(x_i) dP\,(F_{\mathbf{x}}) &= \lim_{m\to\infty} \int \prod_{i=1}^{k} F_{k+m}(x_i) dP\,(F_{k+m}) \\ &= \lim_{m\to\infty} \int \frac{B\,(m) + B^*\,(m)}{(k+m)^k} dP\,(F_{k+m}) \\ &= F(\mathbf{x}_k) \lim_{m\to\infty} \prod_{j=1}^{k} \left(\frac{m+j}{m+k}\right) = F(\mathbf{x}_k).\end{aligned}$$

This demonstrates that the elements of $\mathbf{x}|F_{\mathbf{x}}$ are independent with sampling distribution $F_{\mathbf{x}}$ which was to be shown.

[*Received March 2008, accepted September 2008*]