Exchangeable beliefs
Example

Consider the two situations:

1. We want to estimate the proportion $p$ of people in some large, finite population who possess a characteristic $X$. We take a sample of size, $n$, with replacement, count the number of such people, and use the ratio $r/n$ to estimate $p$.

2. We want to estimate the probability, $p$, that a spun coin lands heads. We spin the coin $n$ times, count the number of heads, $r$, and use the ratio $r/n$ to estimate $p$.

- Standard methodology treats both (A) and (B) in a similar fashion: we view the observed value $r$ as having a binomial distribution $\text{Bin}(n, p)$. 
We then either estimate (or update our priors about) $p$ - essentially the same analysis for both (A) and (B)

However there is a fundamental difference

- For (A), the quantity $p$ is well-defined and in-principle observable with a definite physical meaning
- For (B), $p$ is not observable, is not clearly defined and has no immediate physical meaning - so what exactly are we learning about?

Clearly this a problem arising from relative frequency definition of probability

Even subjectively, we have to construct the quantity “the probability a coin lands head” to learn from the experience of spinning the coin
Exchangeable Beliefs and Variance Learning

Exchangeable beliefs

Even if we are only interested in tangible observable outcomes (e.g., future coin tosses), then we often suppose that there exists an underlying probability $p$ and update our beliefs having observed various spins of the coin and use the updated beliefs to make predictive statements.

So we go from $p$ being a physical quantity of intrinsic interest to a mental construct introduced to simplify an otherwise complex analysis.

This comes at a cost, as the probabilistic specification we require over the observables is usually complex.

The most careful interpretation of relative frequency probability as a mental construct comes from the notion of exchangeability.
We can make a strong case that exchangeability is the fundamental judgement which gives meaning to the kinds of assumptions and modelling which characterises the usual types of statistical analysis.

However for full exchangeability, it can be difficult to construct the probabilistic prior specifications required, even for simple problems.

If we only wish to make a minimal partial prior specification, then we introduce second-order exchangeability.
Second-order exchangeability

Definition (Second-order exchangeability)

The collection of vectors $X_1, X_2, \ldots$ is second-order exchangeable if the first- and second-order belief specification for the sequence of vectors is unaffected by any permutation of the order of the vectors, so that

(i) the mean vector and variance matrix is the same for each individual
$$E(X_i) = \mu, \quad \text{Var}(X_i) = \Sigma, \quad \forall i;$$

(ii) the covariance matrix between any two different individuals is the same,
$$\text{Cov} (X_i, X_j) = \Gamma, \quad \forall i \neq j.$$

This is the simplest useful specification that we can make for such a sequence of random vectors.
Definition (Representation Theorem for an infinite sequence of second-order exchangeable random vectors)

If \( X_1, X_2, \ldots \) is a sequence as above, then we may introduce a further random vector \( \mathcal{M}(X) \), termed the population mean vector, and also the infinite sequence \( \mathcal{R}_1(X), \mathcal{R}_2(X), \ldots \) termed the individual residual vectors, which satisfy the following properties:

(i) For each individual \( j \)

\[
X_j = \mathcal{M}(X) + \mathcal{R}_j(X)
\]

(ii) The mean and variance for \( \mathcal{M}(X) \) are

\[
E(\mathcal{M}(X)) = \mu, \quad \text{Var}(\mathcal{M}(X)) = \Gamma
\]
Definition (Representation Theorem for an infinite sequence of second-order exchangeable random vectors)

(iii) The collection $\mathcal{R}_1(X), \mathcal{R}_2(X), \ldots$ is second-order exchangeable, with, for each individual $j$

$$E(\mathcal{R}_j(X)) = 0, \quad \text{Var}(\mathcal{R}_j(X)) = \Sigma - \Gamma,$$

and the vectors $\mathcal{R}_1(X), \mathcal{R}_2(X), \ldots$ are mutually uncorrelated.

(iv) Each $\mathcal{R}_j(X)$ is uncorrelated with $\mathcal{M}(X)$. 

Every $X_j = \mathcal{M}(X) + \mathcal{R}_j(X) - \mathcal{M}(X)$ is shared among all $X_j$, whereas $\mathcal{R}_j(X)$ is unique to $X_j$. We can use the other $X$s to learn about $\mathcal{M}(X)$, but not $\mathcal{R}_j(X)$.

- Informally, the sequence $X_1, X_2, \ldots$ are “conditionally independent given $\mathcal{M}(X)$”
- $E_{\mathcal{M}(X)}(X_j) = \mathcal{M}(X)$
- $A_{\mathcal{M}(X)}(X_j) = \mathcal{R}_j(X)$
- $\mathcal{M}(X)$ is, in fact, $\lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} X_j$
Bayes linear sufficiency for sample means

- If $E_{D∪F}(B) = E_D(B)$, then $D$ is Bayes linear sufficient for $D∪F$ when adjusting $B$ (and also $Var_{D∪F}(B) = Var_D(B)$).

Theorem

The sample mean vector $\bar{X}_n = \frac{1}{n} \sum_{j=1}^{n} X_j$ is Bayes linear sufficient for the collection $X_1, \ldots, X_n$ for adjusting both $\mathcal{M}(X)$ and any values $X_i$, $i > n$, namely

$$E_{X_1,\ldots,X_n}(\mathcal{M}(X)) = E_{\bar{X}_n}(\mathcal{M}(X)), \ Var_{X_1,\ldots,X_n}(X) = Var_{\bar{X}_n}(\mathcal{M}(X))$$

and, for any $i > n$

$$E_{X_1,\ldots,X_n}(X) = E_{\bar{X}_n}(\mathcal{M}(X)) = E_{\bar{X}_n}(\mathcal{M}(X))$$

$$Var_{X_1,\ldots,X_n}(X) = Var_{\bar{X}_n}(\mathcal{M}(X)) = Var_{\bar{X}_n}(\mathcal{M}(X)) + Var(\mathcal{R}_i(X))$$
Summary of the scary maths

- If a bunch of vectors $X_j$ are SOE then they have the same expectation and variance, and every pair has the same covariance.

- If $X_j$ are SOE, then we can write each vector in the form shared mean vector + individual residual.

- The residuals are all uncorrelated with each other, and are uncorrelated with the mean vector.

- Given the mean vector, the $X_j$ are "conditionally independent."

- We can learn about the mean vector $M(X)$ from the $X_j$.

- If we want to learn about $M(X)$ or predict a future $X_i$, then all we need is the sample mean $\bar{X}_n$ from the $n$ vectors we have already observed.

- All this just from exchangeability!
Variance Learning
Variance learning

- We know how to use **second-order exchangeability** judgements to adjust our beliefs about the population mean.
- We can apply the same technique to learn about **variances** and **covariances**.
- Suppose $X = \{X_1, X_2, \ldots\}$ is an infinite SOE sequence of scalar random quantities.
- Let $E(X_j) = \mu$, $\text{Var}(X_j) = \sigma^2$, and $\text{Cov}(X_j, X_k) = \gamma$.
- $X$ is SOE, so we can write $X_j = \mathcal{M}(X) + R_j(X)$ where

  $$E(\mathcal{M}(X)) = \mu, \quad \text{Var}(\mathcal{M}(X)) = \gamma,$$

  $$E(R_j(X)) = 0, \quad \text{Var}(R_j(X)) = \sigma^2 - \gamma = \omega_R.$$
Magic moments

- Suppose (for now) that $M(X)$ is known (ie $\gamma = 0$)
- Let $V_k = (X_k - \mu)^2 = R_k(X)^2$ and suppose that we judge the sequence $V_1, V_2, \ldots$ to also be SOE
- Then we have $V_k = M(V) + R_k(V)$ where

  $$E(M(V)) = \omega_R, \quad Var(M(V)) = \omega_{M(V)},$$
  $$E(R_j(V)) = 0, \quad Var(R_j(V)) = \omega_{R(V)}.$$

- So we must specify $Var(V_k) = \omega_{M(V)} + \omega_{M(V)}$, ie variances of variances – 4th order moments!

- $\omega_R$ is the expectation of the variance ($\sigma^2$)
- $\omega_{M(V)}$ – how quickly we learn about $M(V)$ given larger samples
- $\omega_{R(V)}$ – shape of the distribution (kurtosis)
Assessing variance with known mean

- Just as for the sample mean, $\bar{X}_n^{(2)} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$ is Bayes linear sufficient for all the $X_i^2$ for adjusting $\mathcal{M}(V)$

- Given specifications for $\omega_R$, $\omega_{M(V)}$, and $\omega_{R(V)}$ we may evaluate the adjusted expectation and variance for $\mathcal{M}(V)$ as follows

$$E_n(\mathcal{M}(V)) = \frac{\omega_{M(V)} \bar{X}_n^{(2)} + \frac{1}{n} \omega_{R(V)} \omega_{R}}{\omega_{M} + \frac{1}{n} \omega_{R(V)}}$$

$$\text{Var}_n(\mathcal{M}(V)) = \frac{\omega_{M(V)} \times \frac{1}{n} \omega_{R(V)}}{\omega_{M} + \frac{1}{n} \omega_{R(V)}}$$
Assessing variances with unknown means is also possible, but mathematically more complex.

We can also learn about variances when the $X$s are correlated (ie not SOE) by orthogonalising to remove the correlation.

This framework can also be used to learn about covariances and correlations (by considering the variance of differences).
Exchangeability is the fundamental judgement which should underpin statistical analysis and modelling.

Second-order exchangeability is a useful tool which requires less stringent specifications.

Bayes linear methods and second-order exchangeability provide a tractable approach to variance learning.

Which was nice!