Notes on Bayesian Analysis of Difference in Prevalence

Jim W. Kay[1] and Robin A. A. Ince[2]

[1]Department of Statistics, University of Glasgow, UK
[2]Institute of Neuroscience and Psychology, University of Glasgow, UK

July 7, 2020

Introduction

The aim of this work is to extend the theory developed in *BayesianPrevalence.pdf* to the situation where two prevalences are being compared by using the difference in prevalence. We consider two different experimental designs:

Case 1: A test procedure is applied with two distinct groups of units and the difference in prevalence between the two groups is estimated by using Bayesian posterior inference;

Case 2: Two different test procedures are applied with a single group of units and the difference is prevalence between the two tests is estimated by using Bayesian posterior inference.

Case 1

Within each of the two distinct populations, the units are of two types: a unit either does or does not possess a definable effect. In population 1, a proportion, γ_1 possess the definable effect, while the proportion of this population which do not possess this effect is $1-\gamma_1$. Similarly, in population 2 a proportion, γ_2, possess a definable effect, while the proportion of this population which do not possess this effect is $1-\gamma_2$.

A random sample of n_i units is selected from the ith population $(i = 1, 2)$ and each unit undergoes a test procedure, in which the presence of the defined effect is investigated using a significance test. It is assumed that for each unit the significance level for the ith test is a_i (*1 - specificity*), or *false positive rate*, and the power, or *sensitivity*, of the ith test is b_i $(0 < a_i < b_i \leq 1)$. Thus, for the ith test, the probability that a randomly selected unit from the population who does not possess the defined effect will produce a significant result is a_i, whereas the probability that a randomly selected unit from the population who does possess the defined effect will produce a significant result is b_i.

A binary variable – *shows a significant effect* or *does not show a significant effect* – is recorded for each unit in each of the samples and we suppose that the total number of units who show a significant effect, out of the n_i tested, is k_i, $(i = 1, 2)$. Let θ_i be the probability that a randomly selected unit from the ith population would show a significant effect. Then

$$\theta_i = (1-\gamma_i)a_i + \gamma_i b_i = a_i + (b_i - a_i)\gamma_i, \quad (i = 1, 2) \tag{1}$$

We will develop the modelling in terms of the parameters θ_1, θ_2, and later use (1) to find appropriate results in terms of the prevalence difference, $\gamma_1 - \gamma_2$.

In designed studies, the aim would be to set the levels of sensitivity and specificity to be equal for both tests, i.e. $a_1 = a_2 \equiv a$ and $b_1 = b_2 \equiv b$. Otherwise the comparison of the prevalences would be biased, a priori. The general situation where specificity and sensitivity, respectively, are not assumed to be equal for both tests will be described in the sequel.

Modelling

For each of the tests, we assume that the test results on the performance of the units are independent and that the parameter θ_i is the same for all units who undertake the test. Let the random variable X_i denote the number of units out of the n_i tested which show a significant effect at significance level a. Then X_i follows a binomial distribution and

$$\Pr(X_i = k_i|\theta_i) = \binom{n_i}{k_i}\theta_i^{k_i}(1-\theta_i)^{n_i-k_i}, \quad k_i = 0, 1, \ldots, n_i, \quad (0 < \theta_i < 1, i = 1, 2). \tag{2}$$

Also X_1 and X_2 are independent given θ_1, θ_2.

We now define prior distributions to characterise the prior uncertainty about the θ_i. First, we note that under the uncontroversial assumption that $b_i > a_i$ for the ith test, we find from (1) that $\theta_i > a_i$. Also, since $\gamma_i < 1$, we find that $\theta_i < b_i$. The claim regarding the assumption that $b_i > a_i$ is perfectly reasonable since it would make no sense to employ a test procedure for which the power is less than the significance level. It follows that $a_1 < \theta_1 < b_1$ and $a_2 < \theta_2 < b_2$

The conjugate prior for θ_i is the beta distribution so, bearing in mind the constraint on θ_i, we assume that the prior distribution for θ_i is the following truncated beta distribution with probability density function

$$p(\theta_i|r_i, s_i) = \frac{1}{B(r_i, s_i)}\frac{\theta_i^{r_i-1}(1-\theta_i)^{s_i-1}}{[F(b; r_i, s_i) - F(a; r_i, s_i)]}, \quad a_i < \theta_i < b_i, \quad (r_i > 0, s_i > 0, i = 1, 2), \tag{3}$$

where $F(x; r_i, s_i)$ is the cumulative distribution function (cdf) of θ_i.

The selection of values for the parameters r_i, s_i depends on prior information about θ_i. In the absence of any prior information about θ_i we will use the choice $r_i = 1, s_i = 1$ in practical applications, while keeping the notation general in the formulation. This corresponds to the *a priori* assumption that the prior uncertainty regarding θ_i can be represented by a uniform distribution on the interval (a_i, b_i), $(i = 1, 2)$. We also assume *a priori* that θ_1 and θ_2 are independent. Given the conditional independence of X_1 and X_2, given θ_1, θ_2, this means that θ_1 and θ_2 are independent *a posteriori* given the binomial data from the test results. This means that the posterior distribution of (θ_1, θ_2) given the binomial data factorises into a product of two truncated beta distributions.

Defining $m_{i1} \equiv k_i + r_i, m_{i2} \equiv n_i - k_i + s_i, (i = 1, 2)$, the posterior distribution for (θ_1, θ_2) given the binomial data is

$$p(\theta_1, \theta_2|k_1, k_2, r_1, r_2, s_1, s_2) = C\,\theta_1^{m_{11}-1}(1-\theta_1)^{m_{12}-1}\theta_2^{m_{21}-1}(1-\theta_2)^{m_{22}-1}, \quad a_1 < \theta_1 < b_1, a_2 < \theta_2 < b_2,$$

where

$$C = \frac{1}{\text{Beta}(m_{11}, m_{12})[F(b; m_{11}, m_{12}) - F(a; m_{11}, m_{12})]\text{Beta}(m_{21}, m_{22})[F(b; m_{21}, m_{22}) - F(a; m_{21}, m_{22})]} \tag{4}$$

and $F(x; \lambda, \mu)$ is the cdf of a Beta distribution having parameters λ, μ.

Posterior density of prevalence difference by simulation

We wish to compute a HPD interval for the difference between the probabilities of a significant result in the two tests, $\theta_1 - \theta_2$ and then convert it into a corresponding HPD interval for the prevalence difference,

$$\gamma_1 - \gamma_2 = \frac{\theta_1 - a_1}{b_1 - a_1} - \frac{\theta_2 - a_2}{b_2 - a_2}, \tag{5}$$

making use of (1). Since the posterior distributions of θ_1 and θ_2 are independent, we can draw values from truncated beta distributions independently for θ_1 and for θ_2.

How can we make a random draw from a truncated distribution? The following procedure achieves this:

- First, draw a random number from the uniform distribution on the interval

$$[F(a_1; m_{11}, m_{12}), F(b_1; m_{11}, m_{12})]$$

- Second, apply the inverse cdf method to find the corresponding random number which follows the Beta(m_{11}, m_{12}) distribution, truncated by $a_1 < \theta_1 < b_1$. This gives the first simulated value for θ_1. Repeat these steps for the required number of times.

Similarly, simulated values of θ_2 are obtained as follows:

- First, draw a random number from the uniform distribution on the interval

$$[F(a_2; m_{21}, m_{22}), F(b_2; m_{21}, m_{22})]$$

- Second, apply the inverse cdf method to find the corresponding random number which follows the Beta(m_{21}, m_{22}) distribution, truncated by $a_2 < \theta_2 < b_2$. This gives the first simulated value for θ_2. Repeat these steps for the required number of times.

The simulated values of the prevalence difference, $\gamma_1 - \gamma_2$ are then available from (5) for each simulated pair (θ_1, θ_2).

Case II

In this case, two different test procedures are applied to a sample of n units. Within the population of units there is a prevalence γ_1 in relation to Test 1 and a prevalence γ_2 in relation to Test 2. Let θ_i be the probability that a randomly selected unit from the population will show a significant result on the ith test $(i = 1, 2)$ Then

$$\theta_i = (1 - \gamma_i)a_i + \gamma_i b_i = a_i + (b_i - a_i)\gamma_i, \quad (i = 1, 2). \tag{6}$$

Modelling

Each unit provides one of four mutually exclusive results, and we denote the observed data by a vector $\mathbf{k} = \{k_{11}, k_{120}, k_{01}, k_{00}\}$, the elements of which are defined as follows:

- k_{11} is the number of subjects which have a significant result on both tests;
- k_{10} is the number of subjects which have a significant result on Test 1 and a non-significant result on Test 2;
- k_{01} is the number of subjects which have a non-significant result on Test 1 and a significant result on Test 2;
- k_{00} is the number of subjects which have a non-significant result on both tests;

and these observed frequencies sum to n, i.e. $\sum_{i,j} k_{ij} = n$.

There is a vector $\boldsymbol{\theta}$ of population parameters defined as follows:

- θ_{11} is the population proportion of subjects which have a significant result on both tests;
- θ_{10} is the population proportion of subjects which have a significant result on Test 1 and a non-significant result on Test 2;
- θ_{01} is the population proportion of subjects which have a non-significant result on Test 1 and a significant result on Test 2;
- θ_{00} is the population proportion of subjects which have a non-significant result on both tests;

with $\theta_{ij} > 0$ and $\sum_{i,j} \theta_{ij} = 1$, so that $\boldsymbol{\theta}$. Let the random vector $\mathbf{X}$ describe the observed counts, k_{ij}. Then $\mathbf{X}$ follows a multinomial model with pmf

$$\Pr(\mathbf{X} = \mathbf{k}|\boldsymbol{\theta}) \propto \theta_{11}^{k_{11}} \theta_{10}^{k_{10}} \theta_{01}^{k_{01}} (1 - \theta_{11} - \theta_{10} - \theta_{01})^{k_{00}}.$$

We take the prior on $\boldsymbol{\theta}$ to be a Dirichlet distribution which is defined on the 3-simplex:

$$p(\boldsymbol{\theta}) \propto \theta_{11}^{r_{11}} \theta_{10}^{r_{10}} \theta_{01}^{r_{01}} (1 - \theta_{11} - \theta_{10} - \theta_{01})^{r_{00}}.$$

Then the posterior distribution of $\boldsymbol{\theta}$, given the observed data $\mathbf{k}$ is

$$p(\boldsymbol{\theta}|\mathbf{k}) \propto \theta_{11}^{m_{11}} \theta_{10}^{m_{10}} \theta_{01}^{m_{01}} (1 - \theta_{11} - \theta_{10} - \theta_{01})^{m_{00}}, \tag{7}$$

where $m_{ij} = k_{ij} + r_{ij}$, $(i = 0, 1)$.

In the absence of specific prior information, we assume in the simulations that $r_{ij} = 1$ for $i = 0, 1$, so that the prior distribution is uniform on the 3-simplex; clearly other values could be used, depending on the available prior information. Indeed, other forms of prior distribution could be used in general.

The marginal probabilities θ_1, θ_2 may be expressed in terms of the components of $\boldsymbol{\theta}$ as

$$\theta_1 = \theta_{11} + \theta_{10}, \quad \theta_2 = \theta_{11} + \theta_{01}, \tag{8}$$

so that from (6), (8)

$$\gamma_1 - \gamma_2 = \frac{\theta_{11} + \theta_{10} - a_1}{b_1 - a_1} - \frac{\theta_{11} + \theta_{01} - a_2}{b_2 - a_2}. \tag{9}$$

The marginal probabilities θ_1, θ_2 are subject to the constraints

$$a_1 < \theta_1 < b_1, \quad a_2 < \theta_2 < b_2 \tag{10}$$

which are in terms of the elements of $\boldsymbol{\theta}$:

$$a_1 < \theta_{11} + \theta_{10} < b_1, \quad a_2 < \theta_{11} + \theta_{01} < b_2. \tag{11}$$

So the posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{k}$ is the truncated Dirichlet distribution defined by the pdf in (7) subject to the constraints in (11). As in Case I, we determine the pdf of the prevalence difference by making use of Monte Carlo simulation. We first describe the 'stick-breaking' method (Wikipedia, article on Dirichlet processes) for simulation from a standard Dirichlet distribution, and then we extend this method to deal with simulation from a *truncated* Dirichlet distribution.

Simulating random Dirichlet data

When the parameters in $\boldsymbol{\theta}$ are constrained only by the usual 'simplex' constraints, a simple approach if given by the 'stick-breaking' method. This is based on the following standard distributional results.

The marginal distribution of θ_{11} given the data is

$$\theta_{11} \sim \text{Beta}(m_{11}, m_{10} + m_{01} + m_{00}).$$

Consideration of the conditional distribution of θ_{10} given θ_{11} and the data leads to

$$\frac{\theta_{10}}{1 - \theta_{11}} \sim \text{Beta}(m_{10}, m_{01} + m_{00})$$

Consideration of the conditional distribution of θ_{01} given θ_{11}, θ_{10} and the data leads to

$$\frac{\theta_{10}}{1 - \theta_{11} - \theta_{10}} \sim \text{Beta}(m_{01}, m_{00})$$

Finally, θ_{00} is computed by using

$$\theta_{00} = 1 - \theta_{11} - \theta_{10} - \theta_{01}.$$

The resulting simulation scheme follows.

1. Draw u_{11} randomly from the Beta$(m_{11}, m_{10} + m_{01} + m_{00})$ distribution. Set $\theta_{11} = u_{11}$.
2. Draw u_{10} randomly from the Beta$(m_{10}, m_{01} + m_{00})$ distribution. Set $\theta_{10} = (1 - u_{11})u_{10}$.
3. Draw u_{01} randomly from the Beta(m_{01}, m_{00}) distribution. Set $\theta_{01} = (1 - u_{11} - u_{10})u_{01}$.
4. Set $\theta_{00} = 1 - \theta_{11} - \theta_{10} - \theta_{01}$.

Simulating random truncated Dirichlet data

Given the nature of the constraints on the parameters in $\boldsymbol{\theta}$, a new approach is required to define an appropriate method of simulation. We adapt the 'stick-breaking' method to our requirements and

develop a 'stick-breaking' method for Dirichlet data under truncation given by the constraints in (11), as follows.

1. Set limits for θ_{11}: lo = 0, hi = $\min(b_1, b_2)$.
2. Make a random draw, z_{11}, from the uniform distribution on the interval
$$[F(\text{lo}; m_{11}, m_{10} + m_{01} + m_{00}), F(\text{hi}; m_{11}, m_{10} + m_{01} + m_{00})].$$
3. Find the corresponding u_{11}, by the inverse cdf method, which follows the required truncated beta distribution. Set $\theta_{11} = u_{11}$.
4. Set limits for θ_{10}: lo = $\max((a_1 - \theta_{11})/(1 - \theta_{11}), 0)$, hi = $(b_1 - \theta_{11})/(1 - \theta_{11})$.
5. Make a random draw, z_{10}, from the uniform distribution on the interval
$$[F(\text{lo}; m_{10}, m_{01} + m_{00}), F(\text{hi}; m_{10}, m_{01} + m_{00})].$$
6. Find the corresponding u_{10}, by the inverse cdf method, which follows the required truncated beta distribution. Set $\theta_{10} = (1 - u_{11})u_{10}$.
7. Set limits for θ_{01}: lo = $\max((a_2 - \theta_{11})/(1 - \theta_{11} - \theta_{10}), 0)$, hi = $\min((b_2 - \theta_{11})/(1 - \theta_{11} - \theta_{10}), 1)$.
8. Make a random draw, z_{01}, from the uniform distribution on the interval
$$[F(\text{lo}; m_{01}, m_{00}), F(\text{hi}; m_{01}, m_{00})].$$
9. Find the corresponding u_{01}, by the inverse cdf method, which follows the required truncated beta distribution. Set $\theta_{01} = (1 - u_{11} - u_{10})u_{01}$.
10. Set $\theta_{00} = 1 - \theta_{11} - \theta_{10} - \theta_{01}$.
11. Then $(\theta_{11}, \theta_{10}, \theta_{01}, \theta_{00})$ is a random draw from the Dirichlet distribution in (7) subject to the constraints in (11).
12. Compute an estimate of the difference in prevalence:
$$\gamma_1 - \gamma_2 = \frac{\theta_{11} + \theta_{10} - a_1}{b_1 - a_1} - \frac{\theta_{11} + \theta_{01} - a_2}{b_2 - a_2}.$$

For given data, this simulate procedure will provide an estimate of the posterior distribution for $\gamma_1 - \gamma_2$, from which other posterior quantities can be estimated.

We finally consider how, given information about the true prevalences, a multinomial data set can be randomly generated.

Simulating random multinomial data, given prevalence information

In this case, two different test procedures are applied to a sample of n units. Within the population of units there are four different prevalences:

γ_{11} the proportion of units in the population that possesses the 'definable effect' on both tests

γ_{10} the proportion of units in the population that possesses the 'definable effect' on test 1 but not test 2

γ_{01} the proportion of units in the population that possesses the 'definable effect' on test 2 but not test 1

γ_{00} the proportion of units in the population that possesses the 'definable effect' on neither test

It is of particular interest to estimate the difference between the marginal prevalences:

$$\gamma_1 = \gamma_{11} + \gamma_{10}, \quad \gamma_2 = \gamma_{11} + \gamma_{01}$$

for Test 1 and for Test 2, respectively.

In order to simulate a random vector from the multinomial distribution, We require to express the θ_{ij}'s in terms of the γ_{ij}'s.

Probabilities of the Tests' outcomes

We denote the test outcomes by O. Then O can be $++, +-, -+, --$, which denote significant result on both tests, significant result on Test 1 but not Test 2 etc. For each of these outcomes there are four possible ground truth situations. Let G denote the ground truth. Then G has values $++, +-, -+, --$, which denote the possibilities 'has the definable effect on both tests', 'has the definable effect only on Test 2', 'has the definable effect only on Test 2', 'has teh definable effect on neither test.

We assume that the test statistics for the two tests are conditionally independent given each value of G. Then the conditional probability that the result is $++$ given that the ground truth is $+-$ is given by $b_1 b_2$. In fact there are sixteen possible combinations of outcomes and ground truths. We illustrate how to obtain θ_{11}, which is the (unconditional) probability of obtaining a significant result on both tests, i.e. the outcome $++$. Then

$$\Pr(O = ++ | G = ++) = b_1 b_2, \quad \Pr(O = ++ | G = +-) = b_1 a_2,$$

$$\Pr(O = ++ | G = -+) = a_1 b_2, \quad \Pr(O = ++ | G = --) = a_1 a_2.$$

The ground truth probabilities are

$$\Pr(G = ++) = \gamma_{11}, \quad \Pr(G = +-) = \gamma_{10}, \quad \Pr(G = -+) = \gamma_{01}, \quad \Pr(G = --) = \gamma_{00}.$$

It follows from the law of total probability that

$$\theta_{11} = b_1 b_2 \gamma_{11} + b_1 a_2 \gamma_{10} + a_1 b_2 \gamma_{01} + a_1 a_2 \gamma_{00}.$$

By using similar arguments, we obtain

$$\theta_{10} = b_1(1 - b_2)\gamma_{11} + b_1(1 - a_2)\gamma_{10} + a_1(1 - b_2)\gamma_{01} + a_1(1 - a_2)\gamma_{00},$$

which may be expressed as

$$\theta_{10} = a_1 + (b_1 - a_1)\gamma_{1.} - \theta_{11}.$$

Also

$$\theta_{01} = (1 - b_1)b_2\gamma_{11} + (1 - b_1)a_2\gamma_{10} + (1 - a_1)b_2\gamma_{01} + (1 - a_1)a_2\gamma_{00},$$

which may be written as

$$\theta_{01} = a_2 + (b_2 - a_2)\gamma_{.1} - \theta_{11}.$$

θ_{00} is found by subtraction as $1 - \theta_{11} - \theta_{10} - \theta_{01}$.

It is worth noting that we recover equations of the marginal proportions of significant results for the two tests, as

$$\theta_1 = \theta_{11} + \theta_{10} = a_1 + (b_1 - a_1)\gamma_{1.},$$

and

$$\theta_2 = \theta_{11} + \theta_{01} = a_2 + (b_2 - a_2)\gamma_{.1}.$$

Independent prevalences

If the prevalences are assumed to be independent in the sense that

$$\gamma_{11} = \gamma_1\gamma_2, \quad \gamma_{10} = \gamma_1(1 - \gamma_2), \quad \gamma_{01} = (1 - \gamma_1)\gamma_2, \quad \gamma_{00} = (1 - \gamma_1)(1 - \gamma_{.2})$$

then, after some algebra, we find that

$$\theta_{11} = \theta_1\theta_2, \quad \theta_{10} = \theta_1(1 - \theta_2), \quad \theta_{01} = (1 - \theta_1)\theta_2, \quad \theta_{00} = (1 - \theta_1)(1 - \theta_2).$$

It is worth considering the posterior distribution for $\boldsymbol{\theta}$ in this case, which is proportional to

$$\theta_1^{m_{11}+m_{10}}(1 - \theta_1)^{m_{01}+m_{00}} \times \theta_2^{m_{11}+m_{01}}(1 - \theta_2)^{m_{10}+m_{00}}.$$

In other words it factorises into the product of two beta pdfs – a form that is the same as in Case 1. This indicates that making the assumption of independence in Case 2 doesn't make sense, since this just reduces to the case of two different groups and one test.

Defining general prevalences for data simulation

There is a simple way to express the general dependence among the the prevalences.

Re-consider the ground truth information. We define two binary variables to represent this. Let G_1 equal 1 when the defined effect associated with Test 1 is present in the population, and zero otherwise. Let G_2 equal 1 when the defined effect associated with Test 2 is present in the population, and zero otherwise. Then, the joint distribution of G_1 and G_2 is

$$\Pr(G_1 = 1, G_2 = 1) = \gamma_{11}, \quad \Pr(G_1 = 1, G_2 = 0) = \gamma_{10},$$
$$\Pr(G_1 = 0, G_2 = 0) = \gamma_{01}, \quad \Pr(G_1 = 0, G_2 = 0) = \gamma_{00},$$

and the marginal distribution are

$$\Pr(G_1 = 1) = \gamma_1\,, \quad \Pr(G_1 = 0) = 1 - \gamma_1\,,$$
$$\Pr(G_2 = 1) = \gamma_2\,, \quad \Pr(G_2 = 0) = 1 - \gamma_2\,.$$

It seems natural to consider the correlation between G_1 and G_2. Denote this by ρ_{12}. This is the correlation between the presences of the defined effects associated with Test 1 and Test 2 in the population. We now derive an expression for this correlation.

Using the probability distributions of G_1 and G_2, it follows that

$$\mathbb{E}(G_1) = \gamma_1\,, \quad \mathbb{E}(G_1^2) = \gamma_1\,,$$
$$\mathbb{E}(G_2) = \gamma_2\,, \quad \mathbb{E}(G_2^2) = \gamma_2\,.$$

Hence, we may write the variances of G_1 and G_2, as well as the covariance of G_1 and G_2 as follows.

$$\begin{aligned}
\text{var}(G_1) &= \mathbb{E}(G_1^2) - \mathbf{E}(G_1)^2 = \gamma_1(1-\gamma_1), \\
\text{var}(G_2) &= \mathbb{E}(G_2^2) - \mathbf{E}(G_2)^2 = \gamma_2(1-\gamma_2), \\
\text{cov}(G_1, G_2) &= \mathbb{E}(G_1 G_2) - \mathbf{E}(G_1)\mathbb{E}(G_2) = \gamma_{11} - \gamma_1\gamma_2\,.
\end{aligned}$$

Therefore, when $0 < \gamma_1 < 1$ and $0 < \gamma_2 < 1$ we can expression the correlation between G_1 and G_2 as

$$\rho_{12} = \text{cor}(G_1, G_2) = \frac{\text{cov}(G_1, G_2)}{\sqrt{\text{var}(G_1)\text{var}(G_2)}} = \frac{\gamma_{11} - \gamma_1\gamma_2}{\sqrt{\gamma_1(1-\gamma_1)\gamma_2(1-\gamma_2)}}.$$

We see that the correlation is equal to zero when $\gamma_{11} = \gamma_1\gamma_2$, i.e. the prevalences are 'independent'. The correlation is equal to 1 when $\gamma_{10} = \gamma_{01} = 0$, and then the marginal prevalences are equal. The correlation is equal to -1 when $\gamma_{11} = \gamma_{00} = 0$, and the marginal prevalences are typically unequal, but can be equal.

Note, however, that in the 'edge cases' in which one or both of the marginal prevalences, γ_1, γ_2, is equal to zero there can be no correlation between G_1 and G_2, since their covariance is equal to zero, and so $\rho_{12} = 0$.

In order to set up a simulation of the multinomial data we require to specify the marginal prevalences, $\gamma_1\,, \gamma_2$, as well the correlation between the presences of the defined effects, ρ_{12}. Then we set

$$\gamma_{11} = \gamma_1\gamma_2 + \rho_{12}\sqrt{\gamma_1(1-\gamma_1)\gamma_2(1-\gamma_2)}$$

and also

$$\begin{aligned}
\gamma_{10} &= \gamma_1 - \gamma_{11} \\
\gamma_{01} &= \gamma_2 - \gamma_{11} \\
\gamma_{00} &= 1 - \gamma_{11} - \gamma_{10} - \gamma_{01}
\end{aligned}$$

This formula is also valid when either or both of the marginal prevalences is equal to zero.

We compute the θ_{ij} as follows

$$
\begin{aligned}
\theta_{11} &= b_1 b_2 \gamma_{11} + b_1 a_2 \gamma_{10} + a_1 b_2 \gamma_{01} + a_1 a_2 \gamma_{00}, \\
\theta_{10} &= a_1 + (b_1 - a_1)\gamma_{1.} - \theta_{11}, \\
\theta_{01} &= a_2 + (b_2 - a_2)\gamma_{.1} - \theta_{11}, \\
\theta_{00} &= 1 - \theta_{11} - \theta_{10} - \theta_{01}.
\end{aligned}
$$

These values of θ_{ij} can then be used to generate random multinomial counts for the required number of participants.