Inferring differences in population probabilities from differences in samples

Abstract

A fundamental question in scientific and statistical reasoning is: what is the likelihood that some property is more frequent in one population than another, given different rates of occurrence in observed sample data? Perhaps surprisingly, there is no clear and general answer to this question. Instead, this question has been subject to controversy and debate over at least 70 years, with multiple methods for calculating this probability being proposed, and no proposal generally accepted. In this paper we give an exact expression for this probability. We develop this expression by focusing on the common situation where all possible values for the two population probabilities p_1 and p_2 were equally likely prior to sampling (the case of *uniform priors*), and then generalise to a broader range of situations. We also give a simple and efficient computational technique for calculating values of this expression exactly under uniform priors and under arbitrary integer-parameterised beta-distributed priors.

keywords: 2×2 **contingency table; Binomial proportions; Comparative trials; Fisher's exact test.**

Inferring differences in population probabilities from differences in samples

In this paper we give a mathematical analysis of a very common situation in statistical decision making. Stated as generally as possible, the situation is as follows. Assume two sets or populations U_1 and U_2, each made up of some finite or infinite number of elements, and where each element may or may not also be a member of some other set A. The proportion of elements in U_1 that are also members of A we call the population probability of A in U_1 (and write as p_1); the proportion of elements in U_2 that are also members of A we call the population probability of A in U_2 (and write as p_2). From the first population U_1 we draw a random sample of size N_1 and from the second a random sample of size N_2. We find that there are K_1 members of A in the first sample and K_2 in the second sample. The question we now ask is: given these samples, what are the chances that A is more frequent or more likely in U_1 than in U_2? More formally, our question is: what is the normatively correct estimate for the probability $P(p_1 > p_2)$, given sample data N_1, K_1, N_2, K_2 ?

This question (what is the probability that some property is more frequent in one population than another, given observed sample data?) arises in many different areas of scientific investigation. Perhaps surprisingly, there is no generally accepted method for calculating this probability $P(p_1 > p_2)$, especially in the centrally important case of small samples, which arise in many drug and vaccine testing regimes (e.g. Kang and Chen, 2000; Farrington and Manning, 1990). Instead, the calculation of this probability has been subject to controversy and debate over at least 70 years, with multiple competing methods being proposed and no clear consensus being reached (Camilli, 1990; Campbell, 2007; Newcombe, 1998; Mehrotra et al., 2003). Our aim in this paper is to give the correct mathematical expression for this probability $P(p_1 > p_2)$ in one very common situation: the situation where, prior to seeing any samples from U_1 and U_2, we

have no information about p_1 and p_2 beyond the fact that they each may take on any value between 0 and 1 and so, in the absence of any further information, we must consider all possible values of p_1 and p_2 to be equally likely (we must take p_1 and p_2 to have uniform prior probabilities).

The organisation of the paper is as follows. In the first section we introduce some useful mathematical terminology, and in the second section we set the stage by deriving various results on probability estimation from samples under certain prior assumptions. In the third section we derive our exact expression for $P(p_1 > p_2)$, given sample data N_1, K_1, N_2, K_2 under the assumption of uniform priors. We also show how this expression generalises to arbitrary Beta-distributed priors. In the fourth section we present an algorithm calculating values of this expression exactly (for uniform or integer-parameterised Beta priors). We also describe computer simulations comparing calculated values for this expression against probabilities estimated via Monte Carlo methods: results confirm this exact expression as correct. In the fifth section we use these Monte Carlo methods to compare the accuracy of our expression for $P(p_1 > p_2)$ against the accuracy of two well known and general approaches to computing this probability; we also assess the performance of our method for a series of illustrative experimental results (using data from Chan and Zhang, 1999). In the sixth section we provide a useful and accurate approximation to our exact mathematical expression. In the final section we describe limitations of our results, and address some general issues related to statistical inference.

Definitions of Terms

The gamma function $\Gamma(a)$ is defined for any complex number a whose real part is positive as

$$\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}\mathrm{d}t.$$

From this, one can show that $\Gamma(a+1) = a\Gamma(a)$, which indicates that this function is a generalised factorial; in fact, for a positive integer n, $\Gamma(n) = (n-1)!$. It also allows the gamma function to be defined for all complex numbers except $a = -1, -2, \ldots$.

The gamma function may also be used to define the rising factorial (also called the Pochammer symbol) $(a)_n$, defined for any a and $n = 0, 1, 2, \ldots$ via

$$\begin{aligned}(a)_n &= \frac{\Gamma(a+n)}{\Gamma(a)} \\ &= \begin{cases} 1 & \text{for } n = 0, \\ a(a+1)\ldots(a+n-1) & \text{for } n = 1, 2, \ldots \end{cases},\end{aligned}$$

In the case where a is a positive integer, $(a)_n = (a-1+n)!/(a-1)!$, and if a is a negative integer, then $(a)_n = 0$ for all n larger than $-a$.

These rising factorials play a key role in the definition of the generalised hypergeometric functions, a class of functions which include most (if not all) of the basic elementary functions which appear in the sciences and engineering, like exponentials, sines and cosines, logarithms, Bessel functions and the like. The generalised hypergeometric function ${}_pF_q$ is a power series in a complex variable x which depends on $p+q$ parameters $a_1, \ldots, a_p$ and $b_1, \ldots, b_q$ and is defined by

$${}_pF_q\left(a_1, \ldots, a_p; b_1, \ldots, b_q; x\right) = \sum_{k=0}^{\infty} \frac{(a_1)_k \ldots (a_p)_k}{(b_1)_k \ldots (b_q)_k} \frac{x^k}{k!}.$$

In general, this is an infinite power series, but if any of the a-parameters is a non-positive integer, it becomes a finite power series.

It is useful to distinguish these generalised hypergeometric functions from the hypergeometric *distribution*. Given a set of N items of which K are members of some category A, and given that a random subset of S items have been drawn from this set, the hypergeometric distribution

$$H(k|S,N,K) = \frac{\binom{S}{k}\binom{N-S}{K-k}}{\binom{N}{K}} \tag{1}$$

gives the probability that k items in that subset are members of A. This distribution is central to Fisher's Exact Test for 2×2 contingency tables (which we will discuss below) and also comes into play in our exact expression for $P(p_1 > p_2)$.

A function related to gamma is the beta function $B(a,b)$, defined for complex numbers a and b (whose real parts are positive) by

$$B(a,b) \;=\; \int_0^1 t^{a-1}(1-t)^{b-1}\mathrm{d}t.$$

This definition leads to the alternate form

$$B(a,b) \;=\; \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

and so if m and n are positive integers, $B(m,n) = (m-1)!(n-1)!/(m+n-1)!$. We can also define the incomplete beta function $B_z(a,b)$ for any real number z to be

$$B_z(a,b) \;=\; \int_0^z t^{a-1}(1-t)^{b-1}\mathrm{d}t,$$

so (obviously) $B_1(a,b) = B(a,b)$.

Finally, this Beta function $B(a,b)$ is connected to the Beta *distribution*, written as Beta(α,β), which is central to probabilistic inference from samples. If Z is any continuous random variable such that the probability of it having a value z between p and $p+\Delta p$ is

$$P(p \le z \le p+\Delta p) \;=\; \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}\,\Delta p$$

for some positive real numbers α and β, we say that $Z \sim \text{Beta}(\alpha, \beta)$. Note that if $\alpha = 1$ and $\beta = 1$ then

$$P(p \leq z \leq p + \Delta p) \quad = \quad \frac{p^0(1-p)^0}{B(1,1)}\,\Delta p = \Delta p$$

and every possible value of z is equally likely (the chance of z falling in a given $p \ldots p + \Delta p$ range is simply equal to the size of that range): $\text{Beta}(1, 1)$ thus represents the uniform distribution.

Probability estimation from sample data

We now turn to the derivation of probability estimates from sample data, given assumed priors. Consider a situation where, prior to seeing any sample data from a given population U, we have rationally correct reasons for assuming that $P(A)$ (the probability of A in U) follows a beta distribution $P(A) \sim \text{Beta}(\alpha, \beta)$. (If, for example, all we know prior to sampling is that any value of $P(A)$ between 0 and 1 could potentially hold, then we would be rationally correct to consider $P(A)$ as following the uniform prior distribution $\text{Beta}(1, 1)$).

Suppose that A occurs K times in a sample of size N. This result could be generated by any population probability $P(A)$ between 0 and 1, but we expect that, for a given K and N, some values of $P(A)$ will be more likely generating probabilities than others. Let $X_p = \{\text{the probability for an event } A \text{ lies between } p \text{ and } p + \Delta p\}$ and $Y = \{\text{in a sample of size } N, \text{ event } A \text{ occurs } K \text{ times}\}$. The probability of an event A occurring K times out of N given that $P(A) = p$ is given by the binomial distribution; thus, if $P(A) \in [p, p + \Delta p]$, then the probability differs from the binomial distribution by at most a term of order Δp, so

$$P(Y|X_p) \quad = \quad \left(\begin{array}{c} N \\ K \end{array} \right) p^K (1-p)^{N-K} + O\left(\Delta p\right).$$

Given our prior distribution, the probability that X_p is true (that is, the probability that $P(A)$ lies between p and Δp) is

$$P(X_p) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}\,\Delta p$$

From the law of total probability $P(Y)$ is equal to $P(Y|X_p)$ across all values of p, weighted by $P(X_p)$, giving

$$\begin{aligned}
P(Y) &= \sum_{0\le p\le 1} P\left(Y|X_p\right) P\left(X_p\right) \\
&= \sum_{0\le p\le 1} \left[\left(\begin{array}{c} N \\ K \end{array} \right) p^K (1-p)^{N-K} + O\left(\Delta p\right) \right] \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}\,\Delta p \\
&\underset{\Delta p\to 0}{\longrightarrow} \left(\begin{array}{c} N \\ K \end{array} \right) \frac{\int_0^1 p^{K+\alpha-1}(1-p)^{N-K+\beta-1}\mathrm{d}p}{B(\alpha,\beta)} \\
&= \left(\begin{array}{c} N \\ K \end{array} \right) \frac{B(K+\alpha, N-K+\beta)}{B(\alpha,\beta)}.
\end{aligned}$$

Recall that Bayes' Law says that for any two events X and Y,

$$P(X|Y) = \frac{P(X)}{P(Y)}\,P(Y|X).$$

and so we find that

$$P(X_p|Y) = \frac{\frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}\,\Delta p}{\left(\begin{array}{c} N \\ K \end{array} \right) \frac{B(K+\alpha,N-K+\beta)}{B(\alpha,\beta)}} \left[\left(\begin{array}{c} N \\ K \end{array} \right) p^K (1-p)^{N-K} + O\left(\Delta p\right) \right].$$

Dropping any terms of order Δp^2 and simplifying gives

$$P(X_p|Y) = \frac{p^{K+\alpha-1}(1-p)^{N-K+\beta-1}}{B(K+\alpha, N-K+\beta)}\,\Delta p \tag{2}$$

(and so $P(X_p|Y) \sim \mathrm{Beta}(K+\alpha, N-K+\beta)$). This expression gives us the probability that $P(A)$ is between p and $p+\Delta p$, given our observed sample data and our assumed

prior distribution for $P(A)$. We want, however, the expected value of $P(A)$ given this sample data (and prior distribution). From the definition of X_p we have

$$P(A|X_p) \;=\; p + O(\Delta p)$$

(a tautological statement: given that $P(A)$ is between p and $p + \Delta p$, the probability of A is between $p + \Delta p$). From the Law of Total Probability we have

$$P(A|Y) \;=\; \sum_{0 \le p \le 1} P(A|X_p) P(X_p|Y)$$

and substituting from Equation 2 we get

$$\begin{aligned}
P(A|Y) \;&=\; \sum_{0<p<1} p \, \frac{p^{K+\alpha-1}(1-p)^{N-K+\beta-1}}{B(K+\alpha, N-K+\beta)} \, \Delta p \\
&\underset{\Delta p \to 0}{\longrightarrow} \; \int_0^1 p \, \frac{p^{K+\alpha-1}(1-p)^{N-K+\beta-1}}{B(K+\alpha, N-K+\beta)} \, \mathrm{d}p \\
&= \; \frac{1}{B(K+\alpha, N-K+\beta)} \int_0^1 p^{K+\alpha}(1-p)^{N-K-\beta+1} \, \mathrm{d}p \\
&= \; \frac{B(K+\alpha+1, N-K+\beta)}{B(K+\alpha, N-K+\beta)}
\end{aligned}$$

From the definition of the Beta function, however, this gives

$$\begin{aligned}
P(A|Y) \;&=\; \frac{\Gamma(K+\alpha+1)\Gamma(N-K+\beta)}{\Gamma(N+\alpha+\beta+1)} \, \frac{\Gamma(N+\alpha+\beta)}{\Gamma(K+\alpha)\Gamma(N-K+\beta)} \\
&= \; \frac{\Gamma(K+\alpha+1)\,\Gamma(N+\alpha+\beta)}{\Gamma(N+\alpha+\beta+1)\,\Gamma(K+\alpha)} \\
&= \; \frac{(K+\alpha)\,\Gamma(K+\alpha)\,\Gamma(N+\alpha+\beta)}{(N+\alpha+\beta)\,\Gamma(N+\alpha+\beta)\,\Gamma(K+\alpha)} \\
&= \; \frac{K+\alpha}{N+\alpha+\beta}
\end{aligned}$$

This result means that the expected value of $P(A)$ given a sample of N events which contains K occurrences of A and under the prior distribution $P(A) \sim \mathrm{Beta}(\alpha, \beta)$, is given by

$$P(A) = \hat{p} = \frac{K+\alpha}{N+\alpha+\beta}$$

Using similar reasoning the average variance of generating probabilities around this expected value is

$$\hat{\sigma}^2 = \frac{(K+\alpha)(N-K+\beta)}{(N+\alpha+\beta+1)(N+\alpha+\beta)^2} = \frac{\hat{p}(1-\hat{p})}{(N+\alpha+\beta+1)} \tag{3}$$

Under the assumption of a uniform prior Beta$(1,1)$, the expected value of $P(A)$ given a sample of N items containing N instances of A is thus

$$P(A) = \hat{p} = \frac{K+1}{N+2} \tag{4}$$

This result is the 'Rule of Succession', first obtained by Laplace and proved in various different ways, with the strongest and most general proof being given by De Finetti (1937). As Zabell (1989), in a very interesting presentation of the history and various proofs of the Rule of Succession, notes, "[I]n order to attack [De Finitti's proof] one must attack the formidable edifice of epistemic probability itself. Modern philosophy continues to ignore it at its own peril" (for an application of the Rule of Succession to results on 'illusory correlation' in psychology and judgment, see Costello and Watts, 2018).

Comparison of Two Samples

Given this background we now consider our question of interest. Suppose we have two populations U_1 and U_2. From the first we draw a sample S_1 of size N_1 and from the second a sample S_2 of size N_2. In both samples, we look for occurrences of some event A and find that it occurs K_1 times in the first sample and K_2 times in the second. The question we now ask is: what is the chance that p_1, the population probability of A in U_1, is greater than p_2, the population probability of A in U_2?

If we assume uniform prior distributions for the population probabilities of A in U_1 and A in U_2, then from Equation 2 the chance that the U_1 population probability which

gives K_1 occurrences of A out of N_1 falls in the range $[p_1, p_1 + \mathrm{d}p_1]$ is

$$\frac{p_1^{K_1}\left(1-p_1\right)^{N_1-K_1}}{B\left(K_1+1, N_1-K_1+1\right)}\mathrm{d}p_2. \tag{5}$$

(this chance follows the Beta$(K_1, N_1 - K_1)$ distribution). Similarly, the chance that the U_2 population probability which gives K_2 occurrences of A out of N_2 falls in $[p_2, p_2 + \mathrm{d}p_1]$ is

$$\frac{p_2^{K_2}\left(1-p_2\right)^{N_2-K_2}}{B\left(K_2+1, N_2-K_2+1\right)}\mathrm{d}p_2. \tag{6}$$

(this chance follows the Beta$(K_2, N_2 - K_2)$ distribution). Assuming that our two samples are independent of one another, the chances that the U_1 probability falls in $[p_1, p_1 + \mathrm{d}p_1]$ *and* the U_2 probability falls in $[p_2, p_2 + \mathrm{d}p_2]$ is given by the product of these two values, and so the overall chance that p_1 is less than p_2 is given by

$$P\left(p_1 < p_2\right) \;=\; \int_{p_1<p_2} \frac{p_1^{K_1}\left(1-p_1\right)^{N_1-K_1} p_2^{K_2}\left(1-p_2\right)^{N_2-K_2}}{B\left(K_1+1, N_1-K_1+1\right) B\left(K_2+1, N_2-K_2+1\right)}\,\mathrm{d}p_1\mathrm{d}p_2$$

(with this expression representing the sum of the product of these values over all cases where $p_1 < p_2$). More explicitly, this gives

$$\begin{aligned} P\left(p_1 < p_2\right) &= \frac{\int_0^1 p_2^{K_2}\left(1-p_2\right)^{N_2-K_2}\left[\int_0^{p_2} p_1^{K_1}\left(1-p_1\right)^{N_1-K_1}\,\mathrm{d}p_1\right]\mathrm{d}p_2}{B\left(K_1+1, N_1-K_1+1\right) B\left(K_2+1, N_2-K_2+1\right)} \\ &= \frac{\int_0^1 p_2^{K_2}\left(1-p_2\right)^{N_2-K_2} B_{p_2}\left(K_1-1, N_1-K_1-1\right)\mathrm{d}p_2}{B\left(K_1+1, N_1-K_1+1\right) B\left(K_2+1, N_2-K_2+1\right)} \end{aligned} \tag{7}$$

If b is a nonnegative integer, the incomplete beta function may be written as the series

$$B_z(a,b) \;=\; B(a,b)z^a \sum_{\ell=0}^{b-1} \frac{(a)_\ell}{\ell!}(1-z)^\ell$$

and using this in (7) gives

$$\begin{aligned} P(p_1 < p_2) &= \frac{1}{B(K_2+1, N_2-K_2+1)} \sum_{\ell=0}^{N_1-K_1} \frac{(K_1+1)_\ell}{\ell!} \int_0^1 p_2^{K_1+K_2+1} (1-p_2)^{N_2-K_2+\ell} \, \mathrm{d}p_2 \\ &= \frac{1}{B(K_2+1, N_2-K_2+1)} \sum_{\ell=0}^{N_1-K_1} \frac{(K_1+1)_\ell}{\ell!} B(K_1+K_2+2, N_2-K_2+\ell+1) \\ &= \frac{(N_2+1)!}{K_2!(N_2-K_2)!} \sum_{\ell=0}^{N_1-K_1} \frac{(K_1+1)_\ell}{\ell!} \frac{(K_1+K_2+1)!(N_2-K_2+\ell)!}{(N_2+K_1+\ell+2)!}, \end{aligned}$$

or, since $(a)_\ell = (a+\ell-1)!/(a-1)!$ for $a = 1, 2, \ldots$,

$$P(p_1 < p_2) = \frac{(N_2+1)!(K_1+K_2+1)!}{K_1!K_2!(N_2-K_2)!} \sum_{\ell=0}^{N_1-K_1} \frac{(N_2-K_2+\ell)!(K_1+\ell)!}{\ell!(N_2+K_1+\ell+2)!}.$$

We denote the sum in this expression by σ, namely,

$$\sigma(N_1, K_1, N_2, K_2) = \sum_{\ell=0}^{N_1-K_1} \frac{(N_2-K_2+\ell)!(K_1+\ell)!}{\ell!(N_2+K_1+\ell+2)!}.$$

so that the probability we want is

$$P(p_1 < p_2) = \frac{(N_2+1)!(K_1+K_2+1)!}{K_1!K_2!(N_2-K_2)!} \sigma(N_1, K_1, N_2, K_2) \tag{8}$$

We now focus our attention on σ: for arbitrary N_1, K_1, N_2 and K_2, dividing out common factors in the fraction gives

$$\begin{aligned} \sigma(N_1, K_1, N_2, K_2) &= \sum_{\ell=0}^{N_1-K_1} \frac{(N_2-K_2+\ell)!(K_1+\ell)!}{\ell!(N_2+K_1+\ell+2)!} \\ &= \sum_{\ell=0}^{N_1-K_1} \frac{(\ell+1)(\ell+2)\ldots(\ell+N_2-K_2)}{(\ell+K_1+1)(\ell+K_1+2)\ldots(\ell+N_2+K_1+2)}. \end{aligned}$$

However, each term in this sum is a fraction with no remaining shared factors in its numerator and denominator only if $N_2 - K_2 > K_1 + 1$, i.e. $K_1 + K_2 \geq N_2$. In contrast, if $K_1 + K_2 < N_2$, then there are still $N_2 - K_1 - K_2$ shared factors that can be cancelled out, giving

$$\sigma(N_1, K_1, N_2, K_2) = \sum_{\ell=0}^{N_1-K_1} \frac{(\ell+1)(\ell+2)\ldots(\ell+K_1)}{(\ell+N_2-K_2+1)(\ell+K_1+2)\ldots(\ell+N_2+K_1+2)}.$$

But note this can be obtained from the $K_1 + K_2 \geq N_2$ expression with the substitutions

$$\begin{aligned} N_1 &\rightarrow N_1 + N_2 - K_1 - K_2, \\ K_1 &\rightarrow N_2 - K_2, \\ N_2 &\rightarrow K_1 + K_2, \\ K_2 &\rightarrow K_2, \end{aligned} \tag{9}$$

so we only need to compute $\sigma(N_1, K_1, N_2, K_2)$ for the $K_1 + K_2 \geq N_2$ case, and then the $K_1 + K_2 < N_2$ case will be given by $\sigma(N_1 + N_2 - K_1 - K_2, N_2 - K_2, K_1 + K_2, K_2)$. Furthermore, the prefactor of σ in (8) is invariant under these substitutions, which means that if we compute $P(p_1 < p_2)$ for $K_1 + K_2 \geq N_2$, then we can get $P(p_1 < p_2)$ for $K_1 + K_2 < N_2$ by using (9).

We now want to find an expression for the series σ in the $K_1 + K_2 \geq N_2$ case using a partial fraction expansion for each term: recall that if we have a rational function of the form $f(x)/g(x)$, where (i) $g(x) = (x - x_1) \dots (x - x_r)$ has distinct roots $x_1, \dots, x_r$, (ii) the degree of $f(x)$ is less than r and (iii) none of the roots of g are also roots of f, then

$$\frac{f(x)}{g(x)} = \sum_{j=1}^{r} \frac{1}{x - x_j} \frac{f(x_j)}{\prod_{i \neq j}(x_j - x_i)}$$

is the partial fraction expansion of $f(x)/g(x)$. Thus, if we choose

$$\begin{aligned} f(\ell) &= (\ell + 1)(\ell + 2) \dots (\ell + N_2 - K_2), \\ g(\ell) &= (\ell + K_1 + 1)(\ell + K_1 + 2) \dots (\ell + N_2 + K_1 + 2) \end{aligned}$$

then each term in σ is a rational function of the form described, where the roots of g are

$\ell_j = -K_1 - j$ for $j = 1, \ldots, N_2 + 2$. It is straightforward to show that

$$\begin{aligned} f(\ell_j) &= \frac{(-1)^{N_2-K_2}(K_1+j-1)!}{(K_1+K_2-N_2+j-1)!}, \\ \prod_{i\neq j}(\ell_j - \ell_i) &= (-1)^{j-1}(j-1)!\,(N_2-j+2)!. \end{aligned}$$

Thus,

$$\begin{aligned} &\frac{(\ell+1)(\ell+2)\ldots(\ell+N_2-K_2)}{(\ell+K_1+1)(\ell+K_1+2)\ldots(\ell+N_2+K_1+2)} \\ &= \sum_{j=1}^{N_2+2} \frac{(-1)^{N_2-K_2+j-1}(K_1+j-1)!}{(j-1)!\,(N_2-j+2)!\,(K_1+K_2-N_2+j-1)!}\,\frac{1}{\ell-K_1-j} \\ &= \frac{(-1)^{N_2-K_2}K_1!}{(N_2+1)!\,(K_1+K_2-N_2)!}\sum_{n=0}^{N_2+1}\frac{(-N_2-1)_n\,(K_1+1)_n}{n!\,(K_1+K_2-N_2+1)_n}\,\frac{1}{\ell+K_1+n+1}, \end{aligned}$$

where we have changed the summation index from j to $n = j-1$ and used $(a+n)! = a!(a+1)_n$ for integer a. Since

$$\frac{1}{\ell+K_1+n+1} = \int_0^1 x^{\ell+K_1+n}\mathrm{d}x,$$

we have

$$\begin{aligned} &\frac{(\ell+1)(\ell+2)\ldots(\ell+N_2-K_2)}{(\ell+K_1+1)(\ell+K_1+2)\ldots(\ell+N_2+K_1+2)} \\ &= \frac{(-1)^{N_2-K_2}K_1!}{(N_2+1)!\,(K_1+K_2-N_2)!}\int_0^1 \sum_{n=0}^{N_2+1}\frac{(-N_2-1)_n\,(K_1+1)_n}{n!\,(K_1+K_2-N_2+1)_n}x^{\ell+K_1+n}\mathrm{d}x \\ &= \frac{(-1)^{N_2-K_2}K_1!}{(N_2+1)!\,(K_1+K_2-N_2)!}\int_0^1 x^{\ell+K_1}\,{}_2F_1\left(-N_2-1, K_1+1; K_1+K_2-N_2+1; x\right)\mathrm{d}x. \end{aligned}$$

We now have an expression for σ:

$$\begin{aligned} \sigma(N_1,K_1,N_2,K_2) &= \sum_{\ell=0}^{N_1-K_1}\frac{(\ell+1)(\ell+2)\ldots(\ell+N_2-K_2)}{(\ell+K_1+1)(\ell+K_1+2)\ldots(\ell+N_2+K_1+2)} \\ &= \frac{(-1)^{N_2-K_2}K_1!}{(N_2+1)!\,(K_1+K_2-N_2)!} \\ &\quad\times\int_0^1 \frac{\left(1-x^{N_1-K_1+1}\right)x^{K_1}}{1-x}\,{}_2F_1\left(-N_2-1, K_1+1; K_1+K_2-N_2+1; x\right)\mathrm{d}x. \end{aligned}$$

Using the well-known identity

$$ {}_2F_1(a,b;c;x) \;=\; (1-x)^{c-a-b}\,{}_2F_1(c-a,c-b;c;x) $$

gives

$$ \begin{aligned} \sigma\left(N_1,K_1,N_2,K_2\right) &= \frac{(-1)^{N_2-K_2}K_1!}{(N_2+1)!\,(K_1+K_2-N_2)!} \\ &\times \int_0^1 \left(x^{K_1}-x^{N_1+1}\right)(1-x)^{K_2}\,{}_2F_1\left(K_2-N_2,K_1+K_2+2;K_1+K_2-N_2+1;x\right)\mathrm{d}x \end{aligned} $$

and thus

$$ \begin{aligned} P\left(p_1<p_2\right) &= \frac{(-1)^{N_2-K_2}\,(K_1+K_2+1)!}{K_2!\,(N_2-K_2)!\,(K_1+K_2-N_2)!} \\ &\times \int_0^1 \left(x^{K_1}-x^{N_1+1}\right)(1-x)^{K_2}\,{}_2F_1\left(K_2-N_2,K_1+K_2+2;K_1+K_2-N_2+1;x\right)\mathrm{d}x \end{aligned} \tag{10} $$

for $K_1+K_2\geq N_2$.

The first term in the integral has a rather remarkable value:

$$ \begin{aligned} &\int_0^1 x^{K_1}(1-x)^{K_2}_2F_1\left(K_2-N_2,K_1+K_2+2;K_1+K_2-N_2+1;x\right)\mathrm{d}x \\ &= \sum_{n=0}^{N_2-K_2}\frac{\left(K_2-N_2\right)_n\left(K_1+K_2+1\right)_n}{n!\left(K_1+K_2-N_2+1\right)_n}\int_0^1 x^{K_1+n}(1-x)^{K_2}\mathrm{d}x \\ &= \sum_{n=0}^{N_2-K_2}\frac{\left(K_2-N_2\right)_n\left(K_1+K_2+1\right)_n}{n!\left(K_1+K_2-N_2+1\right)_n}\frac{(K_1+n)!K_2!}{(K_1+K_2+1+n)!} \\ &= \frac{K_1!K_2!}{(K_1+K_2+1)!}\sum_{n=0}^{N_2-K_2}\frac{\left(K_2-N_2\right)_n\left(K_1+1\right)_n}{n!\left(K_1+K_2-N_2+1\right)_n} \\ &= \frac{K_1!K_2!}{(K_1+K_2+1)!}{}_2F_1\left(K_2-N_2,K_1+1;K_1-K_2-N_2+1;1\right), \end{aligned} $$

and because ${}_2F_1(-m,b;c;1)=\left(c-b\right)_m/\left(c\right)_m$ for $m=0,1,\ldots,$

$$ \begin{aligned} &\int_0^1 x^{K_1}(1-x)^{K_2}_2F_1\left(K_2-N_2,K_1+K_2+2;K_1+K_2-N_2+1;x\right)\mathrm{d}x \\ &= \frac{K_1!K_2!}{(K_1+K_2+1)!}\frac{\left(K_2-N_2\right)_{N_2-K_2}}{\left(K_1+K_2-N_2+1\right)_{N_2-K_2}} \\ &= \frac{(-1)^{N_2-K_2}K_2!\,(N_2-K_2)!\,(K_1+K_2-N_2)!}{(K_1+K_2+1)!} \end{aligned} $$

which is exactly the inverse of the prefactor of the integral in (10). Thus,

$$P\left(p_1 < p_2\right) = 1 - \frac{(-1)^{N_2-K_2}\left(K_1+K_2+1\right)!}{K_2!\left(N_2-K_2\right)!\left(K_1+K_2-N_2\right)!} \times \int_0^1 x^{N_1+1}(1-x)^{K_2}{}_2F_1\left(K_2-N_2, K_1+K_2+2; K_1+K_2-N_2+1; x\right)\mathrm{d}x$$

or

$$P\left(p_1 > p_2\right) = \frac{(-1)^{N_2-K_2}\left(K_1+K_2+1\right)!}{K_2!\left(N_2-K_2\right)!\left(K_1+K_2-N_2\right)!} \times \int_0^1 x^{N_1+1}(1-x)^{K_2}{}_2F_1\left(K_2-N_2, K_1+K_2+2; K_1+K_2-N_2+1; x\right)\mathrm{d}x$$

for $K_1+K_2 \geq N_2$. The integral may be written in terms of a ${}_3F_2$-type hypergeometric function, as follows:

$$\begin{aligned}
&\int_0^1 x^{N_1+1}(1-x)^{K_2}{}_2F_1\left(K_2-N_2, K_1+K_2+2; K_1+K_2-N_2+1; x\right)\mathrm{d}x \\
&= \sum_{n=0}^{N_2-K_2} \frac{\left(K_2-N_2\right)_n\left(K_1+K_2+2\right)_n}{n!\left(K_1+K_2-N_2+1\right)_n}\int_0^1 x^{N_1+1+n}(1-x)^{K_2}\mathrm{d}x \\
&= \sum_{n=0}^{N_2-K_2} \frac{\left(K_2-N_2\right)_n\left(K_1+K_2+2\right)_n}{n!\left(K_1+K_2-N_2+1\right)_n}\frac{(N_1+1+n)!K_2!}{(N_1+K_2+2+n)!} \\
&= \frac{(N_1+1)!K_2!}{(N_1+K_2+2)!}\sum_{n=0}^{N_2-K_2} \frac{\left(K_2-N_2\right)_n\left(K_1+K_2+2\right)_n\left(N_1+2\right)_n}{n!\left(K_1+K_2-N_2+1\right)_n\left(N_1+K_2+3\right)_n} \\
&= \frac{(N_1+1)!K_2!}{(N_1+K_2+2)!}{}_3F_2\left(K_2-N_2, K_1+K_2+2, N_1+2; K_1+K_2-N_2+1, N_1+K_2+3; 1\right).
\end{aligned}$$

Using the identity

$$\begin{aligned}
&{}_3F_2(-n, b, c; d, e; 1) \\
&= \frac{(-1)^n\left(e-b\right)_n\left(e-c\right)_n}{\left(d\right)_n\left(e\right)_n}\,{}_3F_2(-n, 1-e-n, b+c-d-e+1; b-e-n+1, c-e-n+1; 1)
\end{aligned}$$

and doing a bit of algebra gives

$$P\left(p_1 > p_2\right) = \frac{\binom{N_1+1}{K_1}\binom{N_2+1}{K_2+1}}{\binom{N_1+N_2+2}{K_1+K_2+1}}\,{}_3F_2\left(-K_1, -(N_2-K_2), 1; N_1-K_1+2, K_2+2; 1\right) \quad (11)$$

The expression in Equation 11 technically holds only for $K_1 + K_2 \geq N_2$; however, it has exactly the same form after using the mapping given in Equation 9, and so it holds for the $K_1 + K_2 < N_2$ case as well, and thus is the correct expression for *all* values of K_1, N_1, K_2 and N_2. As an aside, note that the combinatorial prefactor in this expression is a value of the hypergeometric distribution described in Equation 1:

$$H(K_1|N_1+1, N_1+N_2+2, K_1+K_2+1) \quad = \quad \frac{\binom{N_1+1}{K_1}\binom{N_2+1}{K_2+1}}{\binom{N_1+N_2+2}{K_1+K_2+1}}$$

There may very well be an identity which allows the ${}_3F_2$-function in Equation 11 to be simplified in terms of factorials, but if so, we have been unable to find it. However, since $-K_1$ and $-(N_2 - K_2)$ are nonnegative integers, this ${}_3F_2$-function is a finite series and so may be computed via

$${}_3F_2\left(-K_1, -(N_2-K_2), 1; N_1-K_1+2, K_2+2; 1\right) \quad = \quad \sum_{i=0}^{\min(K_1, N_2-K_2)} s_i$$

where $s_0 = 1$ and

$$s_{i+1} = \frac{(i-K_1)(i-N_2+K_2)}{(i+N_1-K_1+2)(i+K_2+2)}\, s_i \qquad i = 0, 1, 2, \ldots$$

Thus we have a finite computational procedure for determining the exact probability that $P(A)$ is higher in the population U_1 than in the population U_2, given our observed sample data N_1, K_1, N_2, K_2 (and under the assumption of a uniform prior for both populations).

While we've derived Equation 11 in terms of the uniform prior, the result can be generalised to all priors described by a beta distribution $\text{Beta}(\alpha, \beta)$. This is because, as we saw in the previous section, the distribution for p given counts N and K and prior $\text{Beta}(\alpha, \beta)$ is exactly equivalent the distribution for p given counts $N + \alpha$ and $K + \beta$ and prior $\text{Beta}(1, 1)$. This means that $P(p_1 > p_2)$ given sample counts N_1, K_1 and N_2, K_2

and under the assumption of prior distributions Beta(α_1, β_2), Beta(α_2, β_2) is necessarily equal to $P(p_1 > p_2)$ given sample counts $N_1 + \alpha_1 + \beta_1 - 2, K_1 + \alpha_1 - 1$ and $N_2 + \alpha_2 + \beta_2 - 2, K_2 + \alpha_2 - 1$ under the assumption of uniform priors, and so by substitution we have

$$\begin{aligned} P\left(p_1 > p_2\right) &= \frac{\binom{N_1+\alpha_1+\beta_1-1}{K_1+\alpha_1-1}\binom{N_2+\alpha_2+\beta_2-2}{K_2+\alpha_2}}{\binom{N_1+N_2+\alpha_1+\alpha_2+\beta_1+\beta_2-2}{K_1+K_2+\alpha_1+\alpha_2-1}} \\ &\quad \times {}_3F_2\left(1-K_1-\alpha_1, 1-(N_2-K_2)-\beta_2, 1; N_1-K_1+\beta_1+1, K_2+\alpha_2+1; 1\right). \end{aligned} \tag{12}$$

as an exact expression for $P(p_1 > p_2)$ given sample counts N_1, K_1 and N_2, K_2 and priors Beta(α_1, β_2), Beta(α_2, β_2) . This also means that values of Equation 12 with positive integer parameters $\alpha_1, \beta_1, \alpha_2, \beta_2$ can be calculated using the finite computational methods described for Equation 11 above. (In the case of noninteger prior parameters, Equation 12 still holds, but the ${}_3F_2$ may now be an infinite series, and so an exact value cannot be obtained and approximation approaches must be used.)

Now for a slight digression to comment on Equation 11: It is obvious that if we switch (N_1, K_1) and (N_2, K_2) in this Equation, this simply switches p_1 and p_2, i.e.

$$P\left(p_2 > p_1\right) = \frac{\binom{N_2+1}{K_2}\binom{N_1+1}{K_1+1}}{\binom{N_1+N_2+2}{K_1+K_2+1}} {}_3F_2\left(-K_2, -(N_1-K_1), 1; N_2-K_2+2, K_1+2; 1\right).$$

Also, if we switch K_i with $N_i - K_i$, this switches p_i with $1 - p_i$ (since K_i occurrences of A with probability p_I is the same as $N_i - K_i$ occurrences of $\neg A$ with probability $1 - p_i$), and thus gives $P(1 - p_1 > 1 - p_2) = P(p_2 > p_1)$, and it is straightforward to confirm that this is identical to the above, as it must be. But since $P(p_1 > p_2) + P(p_2 > p_1)$ must equal 1, this leads to the identity

$$\begin{aligned} &\binom{N_1+1}{K_1}\binom{N_2+1}{K_2+1} {}_3F_2\left(-K_1, -(N_2-K_2), 1; N_1-K_1+2, K_2+2; 1\right) \\ +&\binom{N_2+1}{K_2}\binom{N_1+1}{K_1+1} {}_3F_2\left(-K_2, -(N_1-K_1), 1; N_2-K_2+2, K_1+2; 1\right) = \binom{N_1+N_2+2}{K_1+K_2+1}. \end{aligned}$$

More generically, for any m, n, r and s for which the combinations and ${}_3F_2$ functions are defined,

$$\begin{aligned}&\binom{m+r+1}{r+1}\binom{n+s+1}{n}\,{}_3F_2\left(-m,-n,1;r+2,s+2;1\right)\\ +&\binom{m+r+1}{r}\binom{n+s+1}{n+1}\,{}_3F_2\left(-r,-s,1;m+2,n+2;1\right) \;=\; \binom{m+n+r+s+2}{n+r+1},\end{aligned}$$

an identity the authors have not been able to find elsewhere (and thus may be a new result).

Computational implementation

In this section we describe a computational algorithm that efficiently and exactly computes the probability expression derived above, and describe simulation results testing and demonstrating the correctness of this algorithm. The algorithm (Algorithm 1) consists of three functions: functions P, F and S. The function P returns the exact value of $P(p_1 > p_2)$, given sample counts N_1, K_1 and N_2, K_2 and under the assumption of uniform priors. The function calculates this required probability by calling the function F, which calculates the required ${}_3F_2$ function by itself repeatedly calling the function S, which calculates and sums the terms in the finite series expression for this ${}_3F_2$ as in Equations given above. The function P then combines the result obtained with the additional combinatorial prefactors to give the overall probability expression. Note that all calculations in these functions are done purely in terms of integer ratios; such ratios are not reduced at any point to the standard floating-point real number representation used to approximate such ratios. This algorithm thus calculates the required probability exactly.

As noted earlier, probabilities obtained given sample counts N_1, K_1 and N_2, K_2 and under the assumption of Beta prior distributions with positive integer parameters

Algorithm 1 Algorithm calculating $P(p_1 > p_2)$ in terms of unreduced integer ratios. Here '/' is a data structure holding numerator and denominator, an evaluated division.

function S(i, N, b, c, d, e) // Finite sum S.
  if $i = 0$ **then return** $1/1$ // Return the ratio 1/1;
  else
    $s_n \ / \ s_d \leftarrow S(i-1, N, b, c, d, e)$ // S returns integer ratio s_n / s_d.
    $r_n \leftarrow s_n \times (i - 1 - N) \times (i - 1 + b) \times (i - 1 + c)$
    $r_d \leftarrow s_d \times i \times (i + d) \times (i + e)$
    return $r_n \ / \ r_d$ // return result as integer ratio
  end if
end function
function ${}_3F_2(N, b, c, d, e)$ // ${}_3F_2$ function.
  $F_n \ / \ F_d \leftarrow S(0, -N, b, c, d, e)$
  for $i \leftarrow 1$ to $1 - N$ **do**
    $r_n \ / \ r_d \leftarrow S(i, -N, b, c, d, e)$
    $F_n \leftarrow F_n \times r_d + F_d \times r_n$
    $F_d \leftarrow F_d \times r_d$
  end for
  return $F_n \ / \ F_d$
end function
function P(N_1, K_1, N_2, K_2) // $P(p_1 > p_2)$ under uniform priors, and...
  $H_n/H_d \leftarrow \binom{N_1+1}{K_1} \times \binom{N_2+1}{K_2+1} \Big/ \binom{N_1+N_2+2}{K_1+K_2+1}$
  $F_n/ \ F_d \leftarrow {}_3F_2(-K_1, -(N_2 - K_2), 1; N_1 - K_1 + 2, K_2 + 2; 1)$
return $(F_n \times H_n) \ / \ (F_d \times H_d)$
end function
function P$(N_1, K_1, N_2, K_2, \alpha_1, \beta_1, \alpha_2, \beta_2)$ // ...under integer-parameterised Beta priors.
  return $P(N_1 + \alpha_1 + \beta_1 - 2, K_1 + \alpha_1 - 1, N_2 + \alpha_2 + \beta_2 - 2, K_2 + \alpha_2 - 1)$
end function

$\alpha_1, \beta_1, \alpha_2, \beta_2$ can be calculated directly using this algorithm via the call $P(N_1 + \alpha_1 + \beta_1, K_1 + \alpha_1, N_2 + \alpha_2 + \beta_2, K_2 + \alpha_2)$. Algorithm 1 provides a definition of P which takes both sample counts and prior parameters as inputs and computes the required value of $P(p_1 > p_2)$ via this call.

Code implementing this algorithm is available in R and makes use of the GMP package's representation of indefinite-sized integer ratios (see Appendix).

Simulation tests

In this section we test our computational implementation calculating $P(p_1 > p_2)$ via the Monte Carlo simulation shown in Algorithm 2. For a given set of sample data N_1, K_1, N_2, K_2 we run this simulation by calling the function COMPARISON. This function loops repeatedly, on each cycle randomly picking two values for population probabilities of some event A, so that p_1 represents A's probability in population U_1 and p_2 its probability in population U_2 (each p is drawn uniformly from the range $0 \ldots 1$ inclusive). For probability p_1 and sample size N_1 we call the function SAMPLE(p_1, N_1) that draws a sample of N_1 items from the population by randomly picking N_1 values q, drawn uniformly from the range $0 \ldots 1$ inclusive: cases where $q < p_1$ are counted as an instance of event A. For probability p_2 and sample size N_2 we similarly call the function SAMPLE(p_2, N_2). If SAMPLE(p_1, N_1) returns K_1 as the number of instances of A in the drawn sample and SAMPLE(p_2, N_2) returns K_2, we have randomly sampled two probabilities p_1 and p_2 that can generate the given set of sample data N_1, K_1, N_2, K_2. When this happens we increase the variable i that counts the number of such sampled probabilities obtained, and if $p_1 > p_2$ we also increase the variable G, that counts the number of such sampled probabilities for which that relationship holds. The COMPARISON function continues looping until R such 'generating' probability samples

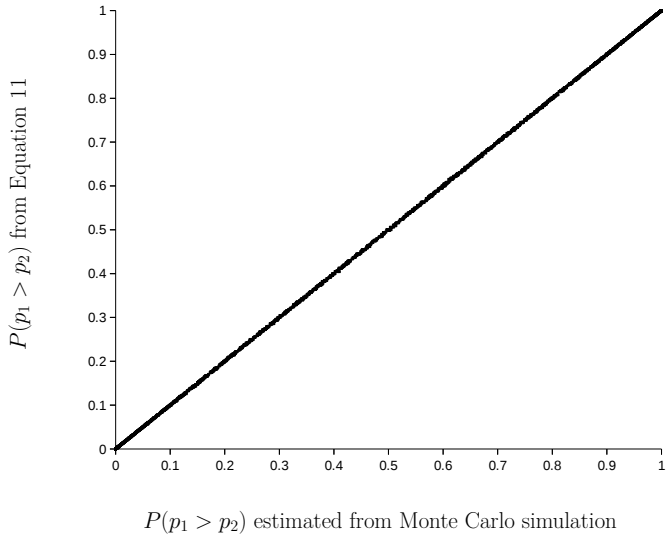

Figure 1. Scatterplot of values of probability $P(p_1 > p_2)$ estimated by the Monte Carlo simulation COMPARISON for all quadruples in the simulation, against values for the probability $P(p_1 > p_2)$ obtained by applying the expression in Equation 11 to that quadruple. The relationship is one of almost exact identity.

have been obtained. Finally, to give an estimate of the probability of p_1 being greater than p_2 given sample data N_1, K_1, N_2, K_2, the function returns the ratio G/R. This ratio can then be compared with the exact value for the probability p_1 being greater than p_2 given sample data N_1, K_1, N_2, K_2, as calculated by function $P(N_1, K_1, N_2, K_2)$ in Algorithm 1.

To test the accuracy of our expression calculating $P(p_1 > p_2)$ as in Algorithm 1 we generated every possible quadruple of values N_1, K_1, N_2, K_2 such that $0 \leq N_1 \leq 10$, $N_2 \leq 10$ and $0 \leq K_1 \leq N_1, 0 \leq K_2 \leq N_2$. For each quadruple we ran the COMPARISON function with $R = 100000$, and obtained an estimated value of $P(p_1 > p_2)$. For each quadruple we also used the code in Algorithm 1 to obtain a computed value for

Algorithm 2 Estimated probability of $p_1 > p_2$, given samples (calculated over R runs)

function SAMPLE(p,N) // samples N items from a population
  $k \leftarrow 0$ // where $P(A) = p$.
  for $i \leftarrow 1$ to N **do**
    $q \leftarrow$ uniform random number in $[0 \ldots 1]$ // generate random number.
    if $q < p$ **then** // event A has occurred.
      $k \leftarrow k + 1$ // increment count.
    end if
  end for
  return k // return number of A'sin sample.
end function
function COMPARISON(N_1, K_1, N_2, K_2, R) // required sample counts.
  $G \leftarrow 0$ // $R =$ number of runs.
  $i \leftarrow 0$ // $G =$ number of cases where $p_1 > p_2$.
  while $i < R$ **do**
    $p_1 \leftarrow$ uniform random number in $[0 \ldots 1]$ // For each cycle, generate two random
    $p_2 \leftarrow$ uniform random number in $[0 \ldots 1]$ // probabilities p_1 and p_2.
    if SAMPLE(p_1, N_1)= K_1 and SAMPLE(p_2, N_2) = K_2 **then**
      $i \leftarrow i + 1$ // Samples drawn match K_1 and K_2.
      if $p_1 > p_2$ **then**
        $G \leftarrow G + 1$ // Count cases where $p_1 > p_2$.
      end if
    end if
  end while
  RETURN(G/R) // proportion of cases where $p_1 > p_2$.
end function

$P(p_1 > p_2)$. Across all of these quadruples, the largest absolute difference between estimated and calculated $P(p_1 > p_2)$ was less than 0.002; the Root Mean Squared Difference was less than 0.001. P values varied widely, covering the full range from 0 to 1. A scatterplot of the relationship between estimated and calculated probabilities is shown in Figure 1. These results strongly support the correctness and accuracy of our expression for $P(p_1 > p_2)$.

Comparisons

In this section we compare the performance of our expression for $P(p_1 > p_2)$ (as shown in Figure 1), against the performance of the two most commonly used techniques for calculating $P(p_1 > p_2)$, the Normal approximation approach and Fisher's exact test (Campbell, 2007).

The standard approach to estimating the chance that $p_1 > p_2$, in circumstances where N_i, K_i, and $N_i - K_i$ are all 'large enough', is to assume that the probability difference $D = p_1 - p_2$ follows a Normal distribution. In this circumstance we can estimate the mean and standard deviation of p_1 and p_2 from the sample data as

$$\hat{p_1} = \frac{K_1}{N_1} \tag{13}$$

$$\hat{p_2} = \frac{K_2}{N_2} \tag{14}$$

$$\hat{\sigma_1} = \sqrt{\frac{p_1(1-p_1)}{N_1}} \tag{15}$$

$$\hat{\sigma_2} = \sqrt{\frac{p_2(1-p_2)}{N_2}} \tag{16}$$

and, given these, we can estimate the standard deviation for the difference D as

$$\hat{\sigma}_D = \sqrt{\hat{\sigma_2}^2 + \hat{\sigma_2}^2} = \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{N_1} + \frac{\hat{p_2}(1-\hat{p_2})}{N_2}} \tag{17}$$

and we have

$$Z_{\hat{D}} = \frac{\hat{p_1} - \hat{p_2}}{\hat{\sigma}_D} \tag{18}$$

as the Z-transformed difference. Now, in a one-sided significance test the probability of $p_1 > p_2$ not holding, given observed data N_1, K_1, N_2, K_2 and the assumption of normally distributed differences, is approximately equal to the area under the standard normal curve from $Z_{\hat{D}}$ to ∞. The probability of $p_1 > p_2$ holding, then, is equal to 1 minus that area, and so we have

$$P_N(p_1 > p_2) \approx \Phi\left(Z_{\hat{D}}\right) \tag{19}$$

where Φ is the cumulative function of the standard normal distribution.

This Normal approximation approach only applies in cases where sample sizes are 'large enough' and where p_1 and p_2 are 'far enough' away from the bounds 0 and 1 (whatever we take 'large enough' and 'far enough away' to mean). As a demonstration of the problems with using this approach, we calculated $P_N(p_1 > p_2)$ for each quadruple N_1, K_1, N_2, K_2 where N_1, N_2 and $Z_{\hat{D}}$ were all greater than 0 (that is, quadruples where calculation of $P_N(p_1 > p_2)$ as in Equations 13 to 19 did not involve division by zero). Figure 2 gives a scatterplot relating estimated values of $P(p_1 > p_2)$ from quadruples produced via Monte Carlo simulations to values of $P_N(p_1 > p_2)$ calculated as above. While a range of calculated values approximately agree with the probabilities observed in the simulation (the diagonal 45° 'stroke' of the Z in this scatterplot) a range of values show substantial deviation between observed and calculated probabilities (the two curving 'arms' of this Z). The greatest deviation between the Normal approximation probability estimates and probability values produced by Monte Carlo simulation arose when sample proportions were close to the boundaries of 0 and 1. The two 'arms' of the Z in Figure 2, for example, are made up of all and only those quadruples where one of the sample proportions K_1/N_1 or K_2/N_2 was equal to either 0 or 1, but the other was

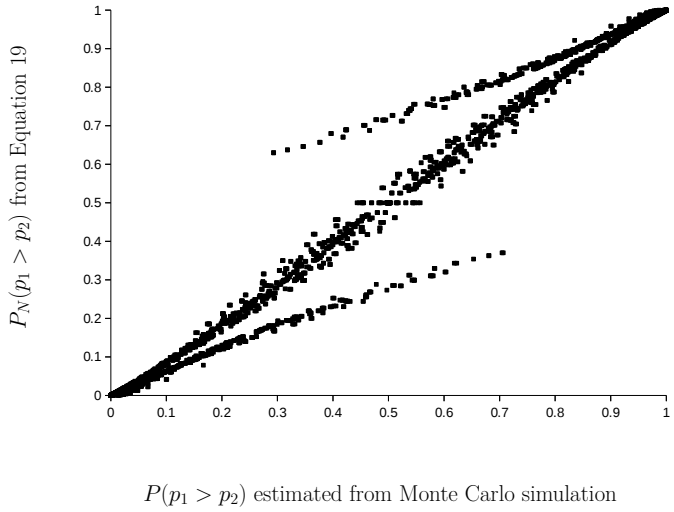

Figure 2. Scatterplot of values of probability $P(p_1 > p_2)$ estimated by the Monte Carlo simulation COMPARISON for quadruples in the simulation, against the standard Normal approximation for this probability $P_N(p_1 > p_2)$as in Equation 19. This scatterplot only shows values for quadruples where N_1 and N_2 were both greater than 0 and where either $0 < K_1/N_1 < 1$ or $0 < K_2/N_2 < 0$.

not.

Given that this normal approximation technique for estimating $P(p_1 > p_2)$ does not work well with small sample sizes, a standard approach in the literature when such estimates are required has been to use Fisher's Exact Test or variants of that test (e.g. Campbell, 2007; Mehrotra et al., 2003; Camilli, 1990). Note that this general use of Fisher's exact test is controversial, because the test only applies in the specific case where all marginal counts are fixed by experimental design, a situation which rarely holds in practice (see e.g. Rice, 1988; Suissa and Shuster, 1985, for discussion). Because the test is commonly used beyond this specific case, we consider it here.

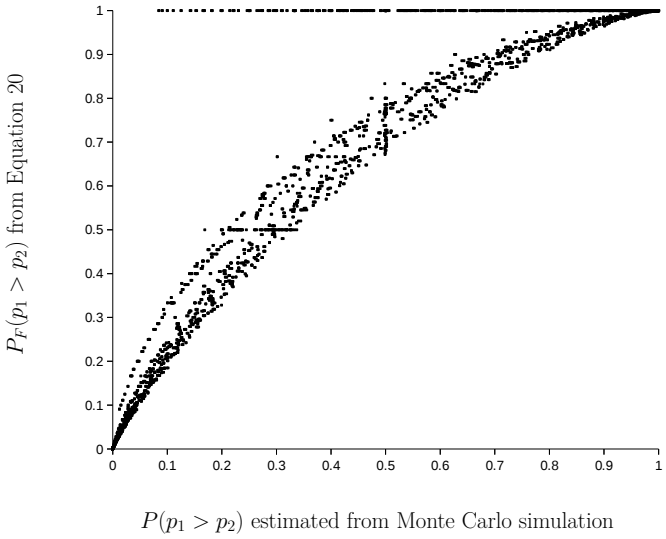

Figure 3. Scatterplot of values of probability $P(p_1 > p_2)$ estimated by the Monte Carlo simulation COMPARISON for all quadruples in the simulation, against values for the probability $P_F(p_1 > p_2)$ obtained by applying Fisher's exact test to that quadruple, as in Equation 20.

Fisher's exact test assumes fixed marginals N_1, N_2 and $K_1 + K_2 = K$. Under this assumption the value of K_2 (the number of instances of event A in our sample of N_2 items from population U_2) is fully determined by the value of K_1 (the number of instances of A in our sample of N_1 items from population U_1). We thus take as our test statistic the value S, the number of instances of A in our sample of N_1 items from U_1. Fisher's exact test computes the probability of obtaining a value of S that is equal to or more extreme than K_1 under the null hypothesis $p_1 = p_2$. The probability of obtaining a result S that is exactly equal to K_1 under the null hypothesis is given by the hypergeometric distribution $H(K_1|N_1, N, K)$ while the probability of obtaining a value of S that is one less than K_1 is given by $H(K_1 - 1|N_1, N, K)$ and so on. To estimate the

probability that $p_1 > p_2$ holds using the Fisher's Exact Test approach, we sum this hypergeometric distribution across all possible values of S less than or equal to K_1 and consistent with the fixed marginal K (that is, across all positive i's for which $0 \leq K_1 - i$ and $K - K_1 + i \leq N_2$), giving

$$P_F(p_1 > p_2) = \sum_{i=0}^{Min(K_1, N_2 - K + K_1)} H(K_1 - i | N_1, N, K) \tag{20}$$

The greater this sum is, the less likely it is that $p_1 < p_2$ holds, and so the more likely it is that $p_1 > p_2$ holds: this sum thus estimates the probability that $p_1 > p_2$.

Figure 3 shows a scatterplot comparing the probabilities $P(p_1 > p_2)$ estimated by the Monte Carlo simulation COMPARISON for all quadruples against values of $P_F(p_1 > p_2)$ for the same quadruples. There is significant and systematic disagreement, visible both in the curve of the main diagonal in this scatterplot and more strikingly in the horizontal line at 1. This line represents all quadruples where either $K_1 = N_1$ or $K_2 = K - K_1 = 0$. From Equation 20 we see that $P_F(p_1 > p_2) = 1$ necessarily holds in these cases. This is because with $K_1 = N_1$, or with $K_1 = K$, the sum in Equation 20 runs across the full set of all possible values of S consistent with the given marginals, and so must have a total probability of 1. From our simulation, however, we see that $P(p_1 > p_2)$ does not, in fact, equal 1 in those cases, but indeed can take on values as low as 0.1.

We finally compare probabilities obtained from our expression $P(p_1 > p_2)$ to the results obtained a series of confidence-interval measures for the significance of $p_1 - p_2$, using data from Chan and Zhang (1999). Chan and Zhang's research on estimates of the difference in population probabilities given sample data was motivated by a real-life problem: In a vaccine clinical trial to investigate whether a new manufacturing process provides improvement over the current process, subjects were randomised to receive

either the new or the current process materials. The preliminary data showed that the proportions of subjects responding to the vaccine were 0.94(17/18) and 0.61(11/18) for the new and current processes, respectively. The rate difference was 0.33, suggesting that the new process had noticeable improvement. Standard small-sample statistical techniques, however, yielded a 95% confidence interval of $(-0.02, 0.63)$, indicating no significant improvement. Chan and Zhang (1999) investigated this result by applying a number of other statistical techniques to this data, to other data concerning vaccine trials, and to artificial data. Table 1 shows the results, alongside the exact probabilities $P(p_1 > p_2)$ calculated as in Equation 11. The count data in Table 1 comes from a number of sources, with the first row representing Chan and Zhang's vaccine count data and the third row representing data from an influenza vaccine study by Fries et al. (1993), and the remaining rows being illustrative. We see from this Table that, while the various confidence-interval measures do not indicate a significant difference between p_1 and p_2 despite a relatively large difference in the observed population proportions (e.g. 17 out of 18 versus 11 out of 18 for the first row), our exact probability $P(p_1 > p_2)$ shows a very significant difference for these rows (a 99% chance that $p_1 > p_2$ in the first row, for example, indicating a 1% chance that there is no difference between the population proportions).

Approximating our exact expression for $P(p_1 > p_2)$

The standard approach to estimating $P(p_1 > p_2|N_1, K_1, N_2, K_2)$ is to assume that probability difference $D = p_2 - p_1$ follows the normal distribution, and to estimate D and its standard deviation via sample proportions (as in Equations 13 to 19). There are at least two reasons why this approximation deviates from the correct probabilities (as shown in Figure 2). The first and most obvious reason for this deviation is that the

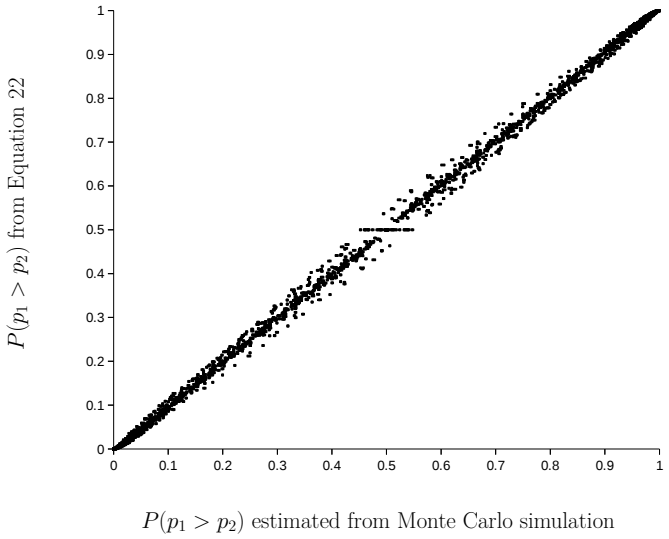

Figure 4. Scatterplot of values of probability $P(p_1 > p_2)$ estimated by the Monte Carlo simulation COMPARISON for all quadruples in the simulation, against values for the probability $P(p_1 > p_2)$ obtained by applying the standard normal approximation for this probability with a correction following the Rule of Succession, as given in Equation 22 to that quadruple.

normal distribution is unbounded (going from $-\infty$ to $+\infty$) while the difference D it is intended to approximate is bounded between -1 and $+1$. This means that the size of the normal distribution tail will differ from the size of the tail of the distribution of the difference D, with this difference increasing as D approaches its bounds of -1 and $+1$.

A second, perhaps less appreciated, reason for this deviation arises from assumptions about the estimation of p_1 and p_2 via the sample proportions K_1/N_1 and K_2/N_2 . Recall that the standard Normal approximation approach uses the sample proportions K_1/N_1 and K_2/N_2 to approximate the population probabilities p_1 and p_2. From the Rule of Succession we see that this is incorrect: instead, assuming uniform

priors we can better estimate the mean and standard deviation of p_1 and p_2 from the sample data as

$$\hat{p_1} = \frac{K_1 + 1}{N_1 + 2}$$

$$\hat{p_2} = \frac{K_2 + 1}{N_2 + 2}$$

and

$$\hat{\sigma_1} = \sqrt{\frac{\hat{p_1}(1 - \hat{p_1})}{N_1 + 3}}$$

$$\hat{\sigma_2} = \sqrt{\frac{\hat{p_2}(1 - \hat{p_2})}{N_2 + 3}}$$

Given these we can estimate the mean and standard deviation for the difference D as

$$\hat{D}_* = \hat{p_1} - \hat{p_2}$$

$$\hat{\sigma}_{D_*} = \sqrt{\hat{\sigma_2}^2 + \hat{\sigma_2}^2} = \sqrt{\frac{\hat{p_1}(1 - \hat{p_1})}{N_1 + 3} + \frac{\hat{p_2}(1 - \hat{p_2})}{N_2 + 3}}$$

and we have

$$Z_* = \frac{\hat{D}_*}{\hat{\sigma}_{D_*}} \tag{21}$$

as the Z-transformed difference and, reasoning just as before, we have

$$P(p_1 > p_2 | N_1, K_1, N_2, K_2) \approx \Phi\left(Z_*\right) \tag{22}$$

This approach gives a much closer and more accurate approximation of $P(p_1 > p_2)$ for small samples (see Figure 4).

The modification to the standard normal approximation described here is motivated by applying the the Rule of Succession to probability estimates from samples, and essentially amounts to adding two 'pseudo-observations' (one success, one failure) to each sample. Interestingly, this modification is exactly the same as one suggested by (Agresti and Caffo, 2000) on purely pragmatic grounds and without theoretical

motivation: Agresti and Caffo found, essentially by trial and error, that adding pseudo-observations of one success and one failure to the sample counts used in the Normal approximation approach resulted in estimates of $P(p_1 > p_2)$ that were surprisingly accurate, even for small samples. Because this modification avoids the problems associated with small sample size which limit the application of the standard normal approach to samples that are 'large enough', Agresti and Caffo suggested that this modification be taught as a standard, generally applicable technique on undergraduate statistics courses. Our application of the Rule of Succession, as described here, explains why this approach works well: it is because the addition of two 'pseudo-observations' (one success, one failure) results in a probability estimate that agrees with the normatively correct Rule of Succession in the case of uniform priors.

Conclusions

Comparing the probability of event occurrences in different populations, given samples from those populations, is a central task in science, medicine and many other areas. Statistical methods for carrying out such comparisons have long been controversial and unclear, especially in the case of small sample sizes (a particularly important case for some uses). In this paper we have given a mathematical derivation expressing the exact value of the required probability $P(p_1 > p_2)$, the chance that the population probability p_1 is greater than the population probability p_2, given a sample N_1, K_1 from population 1 and a sample N_2, K_2 from population 2. We've also provided a computational implementation calculating values of this probability precisely, with no approximations or rounding.

Our expression for $P(p_1 > p_2)$ is derived with no prior information about p_1 and p_2 apart from the assumption that these probabilities are equally likely to fall anywhere

between 0 and 1; that is, under the assumption of a uniform prior for both p_1 and p_2. It is important to stress that our expression is not correctly applicable when this assumption does not hold. If there are good reasons for believing that the uniform prior is not applicable in a given situation, this expression should not be used. To address this, we also give a similar exact mathematical expression for $P(p_1 > p_2)$ under arbitrary Beta-distributed priors Beta(α, β), and show that our computational mechanism can compute exact values of $P(p_1 > p_2)$ for Beta-distributed priors when the parameters α and β are positive integers. If suitable alternative prior distributions can be identified and justified in a given situation, our recommendation is that they be used conservatively in calculating our expression for $P(p_1 > p_2)$.

Rejection of the uniform prior

In what situations would we reject the assumption that, prior to observation, all possible population probabilities are equally likely (that is, reject the uniform prior)? One obvious reason for rejecting the uniform prior is when we know, from the structure of the population being sampled, that not all probabilities are equally likely. Consider a very extreme example: a coin which we know *a priori* is fake (being either double-headed or double-tailed) but which we have not seen. Let p represent the coin's 'probability parameter': the chance of getting a head when the coin is tossed. Before seeing a toss of this coin, what should we take as our prior distribution for p? Clearly, not the uniform distribution: Either $p = 0$ or $p = 1$ may hold (and before tossing the coin we don't know which), but any values of p between 0 and 1 cannot hold. The appropriate prior for p in this case is one with two discrete probability masses of 0.5 at 0 and at 1, and nothing in between. This is the Haldane prior distribution, equivalent to Beta$(\alpha = 0, \beta = 0)$. Recall that, from our earlier results, the expected value for $P(A)$

given sample data K, N and a beta-distributed prior with parameters α, β, is given by $P(A) = (K + \alpha)/(N + \alpha + \beta)$. Under the Haldane prior, this gives

$$P(A) = \frac{K + \alpha}{N + \alpha + \beta} = \frac{K}{N}$$

This is clearly correct for this situation: if we toss the coin once and get heads, we now know that both sides of the coin are heads and so $P(heads) = 1/1 = 1$, while if we toss the coin once and get tails, we now know that both sides of the coin are tails and so $P(heads) = 0/1 = 0$. The uniform prior would necessarily give different results, and so is clearly not appropriate here.

Another more pragmatic reason for rejecting the uniform prior is when we already have previous sample information giving us prior information about the probability of interest. To give an example, consider Chan and Zhang's vaccine example. The preliminary data showed that the proportions of subjects responding to vaccines were 0.94(17/18) and 0.61(11/18) for the new and current processes, respectively. Let p_1 represent the population probability of response for the new vaccine, and p_2 that for the current vaccine. Under the assumption of uniform priors for the effectiveness of both vaccines (that is, under the assumption that both p_1 and p_2 are equally likely to fall anywhere between 0 and 1), our expression gives

$$P(p_1 > p_2|18, 17, 18, 11) = 0.991$$

suggesting a high degree of confidence that p_1 is greater than p_2. However, the assumption of uniform priors seems unlikely to apply in this situation. First, the vaccine manufacturer presumably has a reasonably large amount of prior evidence on the effectiveness of vaccine 2, from previous clinical trials. Such evidence most likely takes the direct form of response counts from those trials. If our aim is to come to an optimal estimate of $P(p_1 > p_2)$, it is rationally correct to include this previous information in our

probability calculations, which we could do simply by adding the number of previous trial participants to N_2 and the number of participants responding in those trials to K_2, reflecting this beta-distributed prior and treating all of our information about the effectiveness of vaccine 2 as a single sample. If the response probability for vaccine 1 in the current trial was close to that seen in previous trials, this will simply reduce the variation of our estimate for p_2, and so will act to increase the value of $P(p_1 > p_2)$. For example, if there were 100 participants in a previous trial of vaccine 2, and 60 of those participants responded to that vaccine, we get

$$P(p_1 > p_2|18, 17, 18 + 100, 11 + 60) = 0.998$$

(as opposed to the 0.991 value obtained when that previous trial data was not included). This makes sense: given our previous trial data, we are now more confident that the effectiveness rate of vaccine 1 is close to 0.61, and so vaccine 2's higher effectiveness rate in our sample is even less likely to fall with the range possible for vaccine 1. If the response probability in previous trials was, however, higher than that seen in the current trial (90 out of 100, say, as opposed to the 11 out of 18 seen currently), we see our measure of the probability that p_1 is greater than p_2 will fall, in this case to

$$P(p_1 > p_2|18, 17, 18 + 100, 11 + 90) = 0.78$$

Again, this is appropriate: it seems that vaccine 2 is performing unusually badly in the current trial, and we should be wary of rejecting vaccine 2 on that basis.

Finally, we can consider a more speculative line of reasoning that would lead to rejection of the uniform prior for p_1, the probability of effectiveness of the new vaccine. In particular, we may consider that, prior to sampling, our expectations for the effectiveness of the new, previously untested vaccine 1 were unlikely to be distributed in a fully uniform way across the 0 to 1 probability range. Given that this vaccine was

developed by knowledgeable experts building on previous success in vaccine manufacturing, we may be justified in considering it extremely unlikely that the effectiveness of this new vaccine will be close to 0. Similarly, given that no perfect vaccine exists, we may be justified in considering it extremely unlikely that the effectiveness of this new vaccine will be close to 1. A natural choice for prior distribution here would be a symmetrical distribution peaked at 0.5; that is, a beta distribution with $\alpha_1 = \beta_1 = x > 1$. Such a prior distribution for vaccine 1 would decrease the probability of effectiveness rates for that vaccine that are near the extremes of 0 and 1, which seems reasonable. Such a prior would also result in a more conservative test: this symmetrical prior will 'pull' our estimates for p_1 towards the centre of the probability scale, thus lowering $P(p_1 > p_2)$. With $\alpha_1 = \beta_1 = 3$ as an illustrative example, we get a prior distribution which gives p_1 an 80% chance of falling in the central half of the probability scale (between 0.25 and 0.75) and a 20% chance of falling in the extreme half of the scale (between 0 and 0.25 or between 0.75 and 1). With this prior, we get $P(p_1 > p_2|18 + 6, 17 + 3, 18, 11) = 0.89$: a probability giving a significantly lower degree of confidence in the belief that $p_1 > p_2$.

Why use priors at all?

Some readers may feel a certain degree of scepticism about the use of priors in our form of inference about differences in population probabilities, given sample data. Surely this idea of using priors introduces too much flexibility in the inferential process? Other forms of probabilistic inference from samples seem to manage perfectly well without considering prior distributions: why do we need to consider these priors at all?

In some ways we agree with these points. Arbitrary or post-hoc selection of priors is clearly wrong and misleading, and such flexibility should be avoided. However, our

results show that the use of rationally justifiable priors (such as the uniform prior in cases where, prior to sampling, we must assume that every possible value of p is equally likely) lead to accurate, correct and useful results. Avoiding this approach altogether because of potential problems with flexibility and overfitting does not seem like a good choice.

We don't agree, however, with the idea that other forms of probabilistic inference do not involve prior distributions. From our perspective all forms of probabilistic inference from sample data involve assumptions about priors, either explicitly (as in our expressions, for example) or implicitly. Consider, for example, the assumption that the sample proportion K/N is an accurate estimate of the population probability p. This assumption lies at the heart of the standard Normal approximation approach to estimating $P(p_1 > p_2)$, as described earlier. In the previous subsection, however, we saw that the assumption that the population probability p is accurately estimated by K/N corresponds to the use of the Haldane prior (Beta($\alpha = 0, \beta = 0$)), which assumes that prior to sampling p is either 0 or 1. By assuming that $p = K/N$ in the standard Normal approximation approach, therefore, we are implicitly choosing to use the Haldane prior. We can see this choice reflected in the behaviour of $p = K/N$: if $K = 0$ or $K = N$ this estimate tells us that p is definitely 0 or that p is definitely 1, no matter how small our sample size N; a direct reflection of the Haldane prior assumption. As we saw earlier, the Normal approximation approach with $p = K/N$ (equivalent to assuming a Haldane prior) performs poorly in matching the probabilities generated in our simulations, which assumed that p_1 and p_2 were distributed uniformly, while the same n What this means, at least from our perspective, is that users of standard 'prior-free' statistical approaches (such as the normal approximation approach) are, in fact, committing to implicit priors: priors that may be unsuitable for the task at hand (just as the Haldane prior is

unsuitable to general probabity-estimation task). We hope that our expression for $P(p_1 > p_2)$ go some way towards the demonstrating the usefulness of considering prior distributions explicitly in statistical inference. We also hope this expression will provide a effective tool in situations requiring decision-making from small samples and limited information.

Appendix

```
# R code calculating exact values for the probability P(p1 > p2| N1,K1,N2,K2)
# as described in the main text.  Only runs when N1,K1,N2,K2 are all integers.
#
# Requires the gmp package, installed in ubuntu via
#
# <terminal>$ sudo apt-get install libgmp3-dev
#
# R > install.packages("gmp")



# use the gmp library

library("gmp")


# calculate value of the hypergeometric distribution H, returns an integer ratio.

H <- function(k,S,N,K){
   return(Hh(k,S-k,K-k,N-S-K+k));
}


# helper function for H: takes as arguments the 4 integer values
# in the contingency table defined by the hypoergeometric distribution
# marginals, and runs over necessary sums making up the combinatorial
# expression for H.  Only accepts non-negative integer arguments.

Hh <- function(a,b,c,d){
   stopifnot(0 <= a, 0 <= b, 0 <= c, 0 <= d);
   stopifnot(a+b+c+d == as.integer(a+b+c+d));
   if(d<b) return (Hh(c,d,a,b));
   r <- as.bigq(1);
   if(a>0){
      for(x in 1:a)
         r <- r/x;
     for(x in (c+1):(c+a))
         r <- r*x;
   }
   if(c>0)
      for(x in (d+1):(d+c))
         r <- r*x;
   if (d > 0)
      for(x in (b+1):(b+d))
         r <- r*x;
   if(c+d > 0)
      for(x in (a+b+1):(a+b+c+d))
         r <- r/x;
   return(r);
}
```

```
# calculates exact values for the 3_F_2 function via calls to S.
# Returns an integer ratio, only accepts integer arguments.

F <- function(N,b,c,d,e){
   r <- S(0,-N,b,c,d,e);
   if(N<0)
      for(i in 1:(1-N)){
         s <- S(i,-N,b,c,d,e);
         r= r + s;
   }
   return(r);
}


# Recursive expression for the S sum, returns an integer ratio,
# only accepts integer arguments.

S <- function(i,N,b,c,d,e){
   stopifnot(N+b+c+d+e==as.integer(N+b+c+d+e));
   if(i == 0) {
      return(as.bigq(1,1));
   } else {
      s = S(i-1,N,b,c,d,e);
      r = s * as.bigq( (i-N-1) * (i+b-1)* (i+c-1) , i * (i+d-1) * (i+e-1) );
      return(r);
   }
}


# returns P(p1>p2 | N1,K1,N2,K2) under uniform priors, as an integer ratio.
# Only accepts non-negative arguments with K <= N. P(p1>p2 | N1,K1,N2,K2)
# under beta distributed priors with integer parameters a1,b1,a2,b2 can be
# obtained by adding a1+b1 to N1, b1 to K1, a2+b2 to N2, and b2 to K2
# and calling this function with the resulting values.

P <- function(N1,K1,N2,K2){
   stopifnot(0 <= K1, K1 <= N1,0 <= K2, K2 <= N2);
   return( H(K1,N1+1,N1+N2+2,K1+K2+1)*F(-K1,-(N2-K2),1,N1-K1+2,K2+2) );
}
```

Table 1

Exact probabilities $P(p_1 > p_2)$ for six sets of sample data calculated as in Equation 11, alongside 95% confidence intervals for $p_1 > p_2$ computed by 4 different techniques searching over the entire domain of possible values for the difference, or over a restricted domain for that difference. These results are taken from Table 1 of Chan and Zhang (1999); the confidence interval statistics used are described in the same paper. Notice that in almost all cases the 'no difference' value falls within these 95% intervals, indicating that the difference is not significant at the 0.05 level according to these statistics. Exact probabilities $P(p_1 > p_2)$ calculated as in Equation 11 show a significant difference at the 0.05 or 0.01 level for the first three data sets.

Sample data				Exact	Search over entire domain ($\gamma = 0$)				Search over restricted domain ($\gamma = 0.001$)			
				probability	Simple	Binomial Z	δ-projected	Likelihood	Simple	Binomial Z	δ-projected	Likelihood
				$P(p_1 > p_2)$	Statistc	Statistic	Z statistic	ratio statistic	Statistic	Statistic	Z statistic	ratio statistic
K_1	N_1	K_2	N_2		(S)	(Z_1)	(Z_2)	(LR)	(S)	(Z_1)	(Z_2)	(LR)
17	18	11	18	0.99	(-0.02 , 0.63)	(0.06 , 0.67)	(0.05 , 0.59)	(0.05 , 0.59)	(-0.02 , 0.63)	(0.06 , 0.66)	(0.05 , 0.59)	(0.05 , 0.59)
9	10	5	10	0.97	(-0.09 , 0.76)	(-0.02 , 0.76)	(-0.02 , 0.74)	(-0.05 , 0.74)	(-0.09 , 0.76)	(-0.02 , 0.76)	(-0.03 , 0.74)	(-0.05 , 0.74)
12	15	7	15	0.97	(-0.06 , 0.65)	(-0.02 , 0.65)	(-0.02 , 0.64)	(-0.03 , 0.64)	(-0.06 , 0.66)	(-0.02 , 0.65)	(-0.03 , 0.64)	(-0.03 , 0.64)
0	10	0	10	0.5	(-0.46 , 0.46)	(-0.46 , 0.46)	(-0.31 , 0.31)	(-0.31 , 0.31)	(-0.42 , 0.42)	(-0.42 , 0.42)	(-0.31 , 0.31)	(-0.31 , 0.31)
0	10	0	20	0.66	(-0.39 , 0.39)	(-0.39 , 0.39)	(-0.19 , 0.31)	(-0.17 , 0.31)	(-27 , 0.38)	(-0.27 , 0.38)	(-0.19 , 0.31)	(0.19 , 0.31)
1	4	0	4	0.78	(-0.51 , 0.83)	(-0.39 , 0.83)	(-0.39 , 0.81)	(-0.39 , 0.81)	(-0.52 , 0.83)	(-0.39 , 0.83)	(-0.39 , 0.81)	(-0.39 , 0.81)
2	4	1	4	0.74	(-0.51 , 0.83)	(-0.51 , 0.82)	(-0.51 , 0.83)	(-0.60 , 0.83)	(-0.52 , 0.83)	(-0.51 , 0.82)	(-0.52 , 0.83)	(-0.60 , 0.83)

References

Agresti, A. and Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *The American Statistician*, 54(4):280–288.

Camilli, G. (1990). The test of homogeneity for 2× 2 contingency tables: A review of and some personal opinions on the controversy. *Psychological Bulletin*, 108(1):135.

Campbell, I. (2007). Chi-squared and Fisher–Irwin tests of two-by-two tables with small sample recommendations. *Statistics in medicine*, 26(19):3661–3675.

Chan, I. S. and Zhang, Z. (1999). Test-based exact confidence intervals for the difference of two binomial proportions. *Biometrics*, 55(4):1202–1209.

Costello, F. and Watts, P. (2018). The rationality of illusory correlation. *Psychological Review*. In press.

De Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. volume 7, pages 1–68.

Farrington, C. P. and Manning, G. (1990). Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in medicine*, 9(12):1447–1454.

Fries, L. F., Dillon, S. B., Hildreth, J. E., Karron, R. A., Funkhouser, A. W., Friedman, C. J., Jones, C. S., Culleton, V. G., and Clements, M. L. (1993). Safety and immunogenicity of a recombinant protein influenza A vaccine in adult human volunteers and protective efficacy against wild-type H1N1 virus challenge. *Journal of Infectious Diseases*, 167(3):593–601.

Kang, S.-H. and Chen, J. J. (2000). An approximate unconditional test of non-inferiority between two proportions. *Statistics in Medicine*, 19(16):2089–2100.

Mehrotra, D. V., Chan, I. S., and Berger, R. L. (2003). A cautionary note on exact unconditional inference for a difference between two independent binomial proportions. *Biometrics*, 59(2):441–450.

Newcombe, R. G. (1998). Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in medicine*, 17(8):873–890.

Rice, W. R. (1988). A new probability model for determining exact P-values for 2 x 2 contingency tables when comparing binomial proportions. *Biometrics*, pages 1–22.

Suissa, S. and Shuster, J. J. (1985). Exact unconditional sample sizes for the 2× 2 binomial trial. *Journal of the Royal Statistical Society. Series A (General)*, pages 317–327.

Zabell, S. L. (1989). The rule of succession. *Erkenntnis*, 31(2):283–321.