Scientific Research Publishing

Open Journal of Statistics, 2020, 10, 363-374
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Exact Distribution of Difference of Two Sample Proportions and Its Inferences

Keshab R. Dahal[1], Mohamed Amezziane[2]

[1]Department of Statistics, Truman State University, Kirksville, USA
[2]Department of Statistics, Central Michigan University, Mt. Pleasant, USA
Email: kdahal@truman.edu

How to cite this paper: Dahal, K.R. and Amezziane, M. (2020) Exact Distribution of Difference of Two Sample Proportions and Its Inferences. *Open Journal of Statistics*, **10**, 363-374.
https://doi.org/10.4236/ojs.2020.103024

Received: April 14, 2020
Accepted: May 6, 2020
Published: May 9, 2020

Abstract

Comparing two population proportions using confidence interval could be misleading in many cases, such as the sample size being small and the test being based on normal approximation. In this case, the only one option that we have is to collect a large sample. Unfortunately, the large sample might not be possible. One example is a person suffering from a rare disease. The main purpose of this journal is to derive a closed formula for the exact distribution of the difference between two independent sample proportions, and use it to perform related inferences such as a confidence interval, regardless of the sample sizes and compare with the existing Wald, Agresti-Caffo and Score. In this journal, we have derived a closed formula for the exact distribution of the difference between two independent sample proportions. This distribution doesn't need any requirements, and can be used to perform inferences such as: a hypothesis test for two population proportions, regardless of the nature of the distribution and the sample sizes. We claim that exact distribution has the least confidence width among Wald, Agresti-Caffo and Score, so it is suitable for inferences of the difference between the population proportion regardless of sample size.

Keywords

Statistical Inferences, Exact Distribution, Difference of Sample Proportions

1. Introduction

Comparing two population proportions, especially when the sample size is small is very challenging in statistics, and has applications in many fields. Several procedures have been suggested [One of the most popular and common methods that has been used for a long time is the Wald interval]. Due to simplicity and convenience, the first method that comes in the mind of most statisticians is the

Wald method. However, there are some disadvantages of the Wald interval. Firstly, it is based on normal approximation and for this approximation to work well, we need a large sample. Unfortunately, large samples may be costly in practice. Secondly, the coverage probability is liberal. The coverage probability with nominal 95% confidence interval is almost less than 0.5 when the sample size is small. Even for a large sample size, the coverage probability is always less than the nominal confidence level ($1-\alpha$).

Agresti and Brian Caffo (2000) [1] introduced Adjusted Wald Confidence Interval by slightly modifying Wald interval by adding one success and one failure for each group. They have also shown that the coverage probability of the Adjusted Wald Interval is reasonably greater than the regular Wald interval. However, Agresti-Caffo interval is also based on normal approximation.

Robert G. Newcombe (1998) [2] has explained eleven different methods to compare the difference between two population proportions. Some of them are conservative, like Score, while others are liberal, like Wald.

The main purpose of this journal is to derive a closed formula for the exact distribution of the difference between two independent sample proportions, and use it to perform related inferences such as a hypothesis test. The rest of the journal is organized as follows. In Section 2, we derive the closed formula for exact distribution of the difference between two independent sample proportions and break it into different cases. We obtain the support of the distribution in Section 3. In Section 4, we perform the hypothesis test. In Section 5, we compute the power of the hypothesis test. In Section 6, we compute the confidence interval and compare it to others. In Section 7, we summarize the main findings and conclude the journal.

2. Exact Distribution of Difference of Two Sample Proportions

Let $X_1, X_2, \cdots, X_m$ and $Y_1, Y_2, \cdots, Y_n$ are *iid* Bernoulli random samples from two different populations with parameters p_1 and p_2 respectively and let $\hat{p}_1 = \frac{1}{m}\sum_{i=1}^{m} X_i$ and $\hat{p}_2 = \frac{1}{n}\sum_{i=1}^{n} Y_i$ be the point estimates of the parameters p_1 and p_2 respectively. We denote the difference between $\hat{p}_1$ and $\hat{p}_2$ by D.

To obtain the exact distribution of D, we first derive the probability generating function (*pgf*) of $W = mn(D+1)$ in the following lemma.

Lemma

Let $W = mn(D+1)$, then the *pgf* of W is given by

$$p_w(z) = \sum_{s=0}^{m}\sum_{u=0}^{s}\sum_{t=0}^{n}\sum_{v=0}^{t}(-1)^{s+t+u+v}\binom{m}{s}\binom{s}{u}\binom{n}{t}\binom{t}{v}p_1^s(1-p_2)^t z^{un+vm} \quad (1)$$

Now, let $f\left(\frac{k}{m}-\frac{l}{n}\right)$ denote the probability mass function (*pmf*) of D at the point $\frac{k}{m}-\frac{l}{n}$, for $k = 0, \cdots, m$ and $l = 0, \cdots, n$.

Theorem

Let the greatest common divisor: $gcd(m,n) = r$, and m' and n' be such

that $m = rm'$ and $n = rn'$. The *pmf* of D is given by

$$f\left(\frac{k}{m}-\frac{l}{n}\right)=(-1)^{k+n-l}\sum_{s=0}^{m}(-1)^{s}\binom{m}{s}p_1^{s}\sum_{t=0}^{n}(-1)^{t}\binom{n}{t}(1-p_2)^{t}$$
$$\cdot\sum_{i\in S_{m,n}(s,t)}(-1)^{i(m'-n')}\binom{s}{k+im'}\binom{t}{(n-l)-in'},$$

for $k = 0,\cdots,m$ and $l = 0,\cdots,n$, where

$$S_{m,n}(s,t)=\left[\max\left(-\frac{k}{m'},\frac{(n-l)-t}{n'}\right),\min\left(\frac{s-k}{m'},\frac{(n-l)}{n'}\right)\right]\cap\mathbb{Z}.$$

From the Theorem above, we derive the next results by corresponding them to different relations between m and n.

Corollary 1

If $gcd(m,n)=1$, then the exact distribution of D is given by:

$$Pr\left(D=\frac{k}{m}-\frac{l}{n}\right)=\binom{m}{k}\binom{n}{l}p_1^{k}p_2^{l}(1-p_1)^{m-k}(1-p_2)^{n-l}$$

for $\frac{k}{m}-\frac{l}{n}\neq 0$, while $Pr(D=0)=(1-p_1)^{m}(1-p_2)^{n}+p_1^{m}p_2^{n}$.

Corollary 2

If $m = n$ and $k = l$ then the exact distribution of D is given by

$$Pr(D=0)=(-1)^{n}\sum_{s=0}^{n}(-1)^{s}\binom{n}{s}p_1^{s}\sum_{t=0}^{n}(-1)^{t}\binom{n}{t}(1-p_2)^{t}\sum_{u=n-t}^{s}\binom{s}{u}\binom{t}{n-u}$$

Corollary 3

The exact distribution of D is given by

$$Pr\left(D=\frac{k}{m}-\frac{l}{n}\right)$$
$$=\sum_{s=0}^{m}\sum_{t=0}^{n}\sum_{(u,v)\in S_{s,t}}(-1)^{s+t+u+(k-u)\frac{n}{m}+n-l}\binom{m}{s}\binom{s}{u}\binom{n}{t}\binom{t}{(k-u)\frac{n}{m}+n-l}p_1^{s}(1-p_2)^{t}$$

for $k = 0,\cdots,m$ and $l = 0,\cdots,n$ where,

$$S_{s,t}=\left\{\left(u,(k-u)\frac{n}{m}+n-l\right)\in\mathbb{N}^2:\max\left(0,k+(n-l-t)\frac{m}{n}\right)\right.$$
$$\left.\leq u\leq\min\left(s,k+(n-l)\frac{m}{n}\right)\right\}$$

Corollary 4

The exact distribution of D is symmetrical about zero if $m = n$ and $p_1 = p_2$.

3. Support of the Distribution

Support of the exact distribution is denoted by $D(m,n)$. For small values of m and n, it can be derived manually. However, for larger values of m and n, it is tedious and time consuming, so the software such as R is used.

For $m = n = 2, D=\frac{k}{2}-\frac{l}{2}$. Where $k = 0,1,2$ and $l = 0,1,2$.

Case	k	l	$D=d$	Case	k	l	$D=d$
1	0	0	0.0	6	1	2	−0.5
2	0	1	−0.5	7	2	0	1.0
3	0	2	−1.0	8	2	1	0.5
4	1	0	0.5	9	2	2	0.0
5	1	1	0.0				

Thus the support for $m=n=2$ is $-1,-0.5,0,0.5,1$.

The graphs of the Probability mass function for exact distribution for the difference of two population proportion for m = n and $p_1 = p_2$ are plotted in **Figure 1**. These graphs (**Figure 1**) are the evidence to support corollary 4.

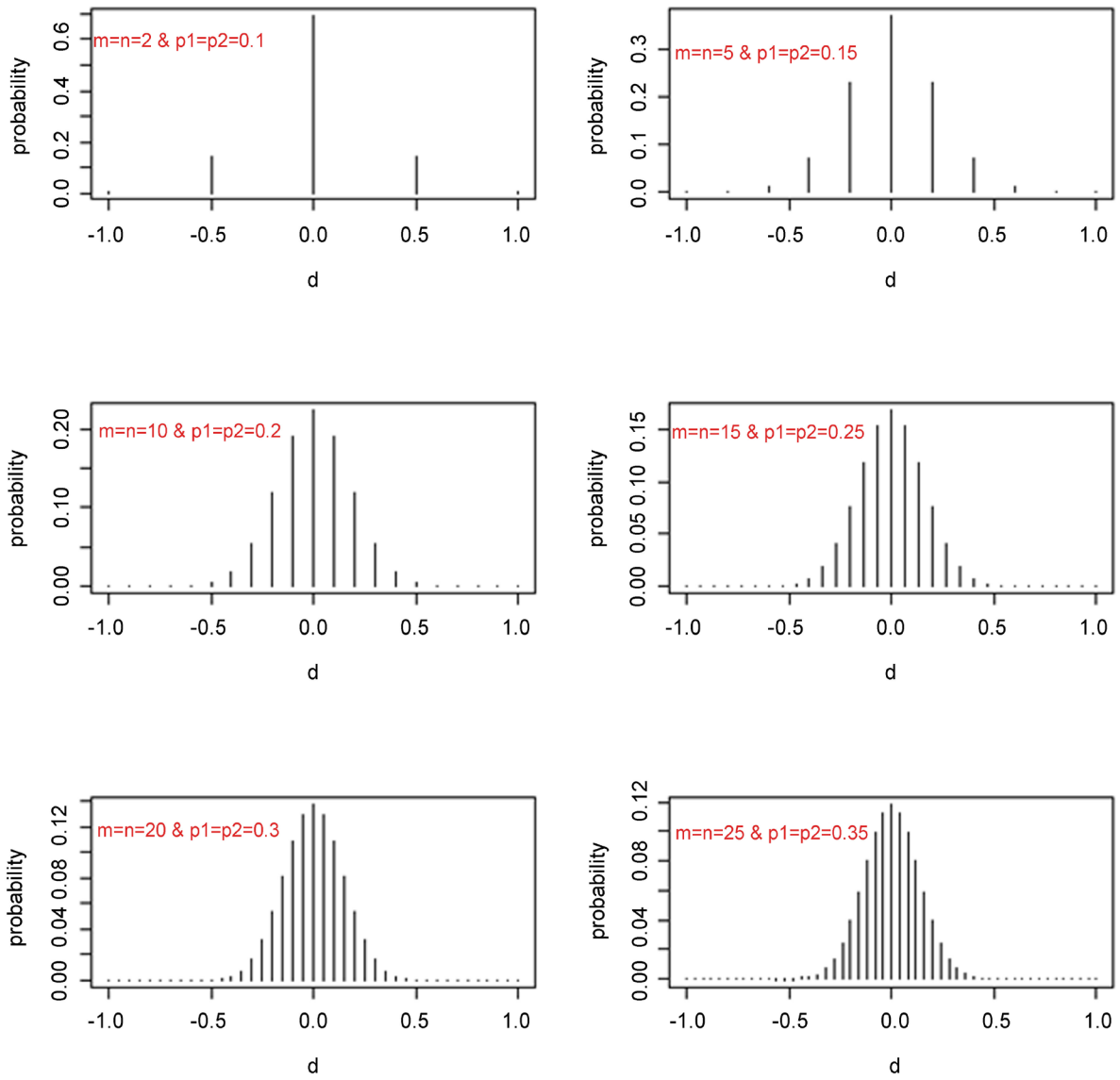

Figure 1. Probability mass function for exact distribution for the difference of two population proportion for $m=n$ and $p_1 = p_2$.

4. Hypothesis Testing

To test $H_0 : p_1 = p_2 = p$ against $H_1 : p_1 - p_2 = \delta \neq 0$, we use D as a test statistic. Let $p\left(D = \frac{k}{m} - \frac{l}{n} \mid H_0\right) = f_0\left(\frac{k}{m} - \frac{l}{n}\right)$. Then the null distribution of D is given by

$$f_0\left(\frac{k}{m} - \frac{l}{n}\right) = (-1)^{k+n-l} \sum_{s=0}^{m} (-1)^s \binom{m}{s} p^s \sum_{t=0}^{n} (-1)^t \binom{n}{t} (1-p)^t \cdot \sum_{i \in S_{m,n}(s,t)} (-1)^{i(m'-n')} \binom{s}{k+im'} \binom{t}{(n-l)-in'},$$

for $k = 0, \cdots, m$ and $l = 0, \cdots, n$, where

$$S_{m,n}(s,t) = \left[\max\left(-\frac{k}{m'}, \frac{(n-l)-t}{n'}\right), \min\left(\frac{s-k}{m'}, \frac{(n-l)}{n'}\right)\right] \cap \mathbb{Z}.$$

The critical region can be obtained by finding $c_{\alpha/2}$ and $c_{1-\alpha/2}$ such that:

$$\max\left\{D : pr\left(D \leq c_{\frac{\alpha}{2}} \mid H_0\right) \leq \frac{\alpha}{2}\right\} \text{ and } \min\left\{D : pr\left(D \geq c_{1-\frac{\alpha}{2}} \mid H_0\right) \leq \frac{\alpha}{2}\right\}.$$

This means that:

$$\sum_{(k,l) \in E_{\frac{\alpha}{2}}} f_0\left(\frac{k}{m} - \frac{l}{n}\right) \leq \frac{\alpha}{2} \text{ and } \sum_{(k,l) \in E_{1-\frac{\alpha}{2}}} f_0\left(\frac{k}{m} - \frac{l}{n}\right) \leq \frac{\alpha}{2}.$$

where

$$E_{\frac{\alpha}{2}} = \left\{(k,l) \in \mathbb{N}^2 : 0 \leq k \leq m, 0 \leq l \leq n, \frac{k}{m} - \frac{l}{n} \leq c_{\frac{\alpha}{2}}\right\}$$

and

$$E_{1-\frac{\alpha}{2}} = \left\{(k,l) \in \mathbb{N}^2 : 0 \leq k \leq m, 0 \leq l \leq n, \frac{k}{m} - \frac{l}{n} \geq c_{1-\frac{\alpha}{2}}\right\}$$

Example: Gender Discrimination

The table below shows the gender distribution of the promoted files.

Gender	Promoted	Not promoted	Total
Male	21	3	24
Female	14	10	24
Total	35	13	48

Data Source:

https://www2.stat.duke.edu/courses/Spring12/sta101.1/lec/lec14S.pdf.

In this question, we will investigate whether or not gender discrimination is associated with the promotion of the employees. In other words, we would like to conduct the following hypothesis test.

H_0: There is no gender discrimination in promotion **vs** H_1: There is gender discrimination in promotion.

We run the R program for exact distribution for $m=24$, $n=24$, $\hat{p}_1=\frac{21}{24}$, and $\hat{p}_2=\frac{14}{24}$, obtain the test statistic, and p-value to 0.291667 and 0.03286628 respectively. Since p-value is less than α, we reject the null hypothesis and conclude that there is gender discrimination in promotion. However the p-value is slightly less than α, so there is moderate gender discrimination for the promotion of the employees.

5. Power Calculation

If $c_{\frac{\alpha}{2}}$ and $c_{1-\frac{\alpha}{2}}$ are the left and right critical values and if the Null hypothesis is rejected for the test statistic, $d=\hat{p}_1-\hat{p}_2$ then the power of the corresponding hypothesis test is given by:

$$1-\beta=2\min\left\{pr\left(D\le d\mid H_\alpha\right),pr\left(D\ge d\mid H_\alpha\right)\right\}=2\sum_{(k,l)\in E_\alpha}f\left(\frac{k}{m}-\frac{l}{n}\right)$$

where

$$E_\alpha=\left\{(k,l)\in\mathbb{N}^2:0\le k\le m,0\le l\le n,\frac{k}{m}-\frac{l}{n}\le d \text{ or } \frac{k}{m}-\frac{l}{n}\ge d\right\}$$

Continuation of the example: Gender Discrimination

In this example, we have rejected null hypothesis with the significance level $\alpha=0.05$. Now we want to find power of the hypothesis test for $p_1=\hat{p}_1=\frac{21}{24}, p_2=\hat{p}_2=\frac{14}{24}$, and $\alpha=0.05$. We run the R program for the power calculation of exact distribution and obtain that the power of the hypothesis test equals to 0.5657226.

6. Confidence Interval

Point estimator of p_1-p_2 is $D=\hat{p}_1-\hat{p}_2$, which can be obtained by the given samples. Let $L_{\alpha/2}$ and $U_{\alpha/2}$ are lower and upper bound for $1-\alpha$ confidence coefficient for p_1-p_2. We obtain $L_{\alpha/2}$ and $U_{\alpha/2}$ as follows:

$$L_{\alpha/2}=\max\left\{D:pr\left(D\le L_{\frac{\alpha}{2}}\right)\le\frac{\alpha}{2}\right\}$$

$$U_{\alpha/2}=\min\left\{D:pr\left(D\ge U_{\frac{\alpha}{2}}\right)\le\frac{\alpha}{2}\right\}.$$

Thus, $(1-\alpha)100\%$ confidence interval for p_1-p_2 is $\left(L_{\alpha/2},U_{\alpha/2}\right)$.

A relatively easy approach to compare the difference between population proportions (p_1-p_2) is confidence interval. We calculate the sample proportions $\hat{p}_1$ and $\hat{p}_2$ from respective samples. Once $\hat{p}_1$ and $\hat{p}_2$ are calculated, we use them to construct confidence interval with nominal confidence coefficient $1-\alpha$. If the confidence interval does not include 0, we reject the null hypothesis. Otherwise fail to reject null hypothesis.

Table 1. 95% confidence interval for Exact, Wald, Agresti-Caffo, and Score.

(a)

k	m	l	n	a	L. Exact	U. Exact	L. Wald	U. Wald	L. AC	U. AC	L. Score
14	20	2	20	0.05	0.40	0.85	0.360	0.840	0.304	0.787	0.313
14	20	3	20	0.05	0.35	0.85	0.295	0.805	0.247	0.753	0.252
15	20	2	20	0.05	0.50	0.90	0.419	0.881	0.356	0.826	0.365
17	20	4	20	0.05	0.50	0.90	0.415	0.885	0.353	0.829	0.358
18	20	5	20	0.05	0.60	0.90	0.419	0.881	0.356	0.826	0.365

(b)

U. Score	Exact. CI. Width	Wald. CI. Width	AC. CI. Width	Score. CI. Width
0.790	0.45	0.480	0.484	0.477
0.754	0.50	0.509	0.505	0.503
0.825	0.40	0.462	0.470	0.461
0.826	0.40	0.470	0.476	0.469
0.825	0.30	0.462	0.470	0.461

For the purpose of this comparison, we have constructed some confidence intervals including respective confidence width for Exact, Wald, Agresti-Caffo and Score for $m = n = 20$ and 95% confidence coefficient (Table 1).

The last four columns of the above table are the confidence widths for Exact, Wald, Agrest-Caffo, and Score. It can be seen that the confidence width of Exact has the least amount.

7. Conclusion

Inferences of the difference of the population proportion are a very basic problem in statistics. Standard Wald interval has been used universally. Standard Wald interval is persistently chaotic, and has unacceptably poor coverage probabilities when either the sample sizes are small or one proportion is very large and the other is very small. Several intervals have been suggested but their level of performance is not satisfactory when the sample size is small. We have been shown that our distribution does not depend on sample size. We have also shown that exact distribution has the least confidence width among Wald, Agresti-Caffo and Score, so it is suitable for inferences of the difference between the population proportion regardless of sample size.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Agresti, A. and Caffo, B. (2000) Simple and Effective Confidence Interval for Pro-

portions and Differences of Proportions Result from Adding Two Successes and Two Failures. *American Statistical Association*, **54**, 280-288. https://doi.org/10.2307/2685779

[2] Newcombe, R.G. (1998). Interval Estimation for the Difference between Independent Proportions: Comparison of Eleven Methods. John Wiley & Sons, Ltd., Hoboken. https://doi.org/10.1002/(SICI)1097-0258(19980430)17:8<873::AID-SIM779>3.0.CO;2-I

Appendix

Proof of lemma

If we define $Z_j = (1 - Y_j)$, then W can be written as $W = n\sum_{i=1}^{m} X_i + m\sum_{j=1}^{n} Z_j$. The *pgf* of W can be written as $p_w(z) = \prod_{i=1}^{m} E\left(z^{nX_i}\right) \prod_{i=1}^{n} E\left(z^{mZ_j}\right)$ since the two samples are independent of each other and the observations in each sample are independent and identically distributed.

Since $X_i \overset{iid}{\sim} Ber(p_1)$ for $i = 1, \cdots, m$, then $E\left(z^{nX_i}\right) = 1 - p_1\left(1 - z^n\right)$ and

$$\begin{aligned} E\left(\prod_{i=1}^{m} z^{nX_i}\right) &= \left(1 - p_1\left(1 - z^n\right)\right)^m = \sum_{s=0}^{m} (-1)^s \binom{m}{s} p_1^s \left(1 - z^n\right)^s \\ &= \sum_{s=0}^{m}\sum_{u=0}^{s} (-1)^{s+u} \binom{m}{s}\binom{s}{u} p_1^s z^{un}. \end{aligned} \tag{2}$$

Similarily, since $Y_i \sim Ber(p_2)$ for $j = 1, \cdots, n$, then

$$E\left(\prod_{j=1}^{n} z^{mZ_j}\right) = \sum_{t=0}^{n}\sum_{v=0}^{t} (-1)^{t+v} \binom{n}{t}\binom{t}{v} (1 - p_2)^t z^{vm}. \tag{3}$$

We multiply the RHS' of 2 and 3 to obtain 1.

Proof of Theorem

Notice that, even though the support of D and W are different, their *pmf*'s have the same probabilities: $Pr(W = kn + (n-l)m) = Pr\left(D = \frac{k}{m} - \frac{l}{n}\right)$ for $k = 0, \cdots, m$ and $l = 0, \cdots, n$. The *pmf* of W can be obtained from the *pgf* as follows:

$$Pr(W = kn + (n-l)m) = \frac{1}{(kn + (n-l)m)!} \frac{\mathrm{d}^{kn+(n-l)m}}{\mathrm{d}z^{kn+(n-l)m}} p_w(z)\Big|_{z=0}.$$

Therefore,

$$\begin{aligned} &Pr(W = kn + (n-l)m) \\ &= \sum_{s=0}^{m}\sum_{u=0}^{s}\sum_{t=0}^{n}\sum_{v=0}^{t} (-1)^{s+t+u+v} \binom{m}{s}\binom{s}{u}\binom{n}{t}\binom{t}{v} p_1^s (1 - p_2)^t \delta_{kn+(n-l)m}(un + vm), \end{aligned} \tag{4}$$

where $\delta_a(x) = 1$ if $x = a$ and 0 otherwise.

To simplify the formula 4, we use the fact that $\delta_{kn+(n-l)m}(un + vm) = 1$ is equivalent to $kn + (n-l)m = un + vm$ which, in its turn, is equivalent to $(u-k)n' = (n-l-v)m'$. From this last equality, we conclude that $u - k = im'$ and $n - l - v = in'$ for some $i \in \mathbb{Z}$ because m' and n' are relative prime to each other. The values of *i* are hence obtained by solving the following system of equations:

$$\begin{cases} u - k = im' \\ (n-l) - v = in' \\ 0 \le u \le s \\ 0 \le v \le t \\ i \in \mathbb{Z} \end{cases}$$

This leads to the following simplified system: $\begin{cases} -\frac{k}{m'} \le i \le \frac{s-k}{m'} \\ \frac{(n-l)-t}{n'} \le i \le \frac{(n-l)}{n'} \\ i \in \mathbb{Z} \end{cases}$. which corresponds to the values of i that forms the set

$$S_{m,n}(s,t) = \left[\max\left(-\frac{k}{m'}, \frac{(n-l)-t}{n'} \right), \min\left(\frac{s-k}{m'}, \frac{(n-l)}{n'} \right) \right] \cap \mathbb{Z}.$$

Proof of Corollary 1

Since m and n are relatively prime to each other, the support of D becomes:

$$S_{m,n}(s,t) = \left[\max\left(-\frac{k}{m}, \frac{(n-l)-t}{n} \right), \min\left(\frac{s-k}{m}, \frac{(n-l)}{n} \right) \right] \cap \mathbb{Z}.$$

when $\frac{k}{m} - \frac{l}{n} \neq 0$, we have $(k,l) \notin \{(0,0),(m,n)\}$, hence $-1 < \max\left(-\frac{k}{m}, \frac{(n-l)-t}{n} \right) < 1$ and $-1 < \min\left(\frac{s-k}{m}, \frac{(n-l)}{n} \right) < 1$. Therefore $S_{m,n}(s,t) = \{0\}$. Now from Theorem above we get,

$$\begin{aligned} Pr\left(D = \frac{k}{m} - \frac{l}{n} \right) &= \sum_{s=k}^{m} \sum_{t=n-l}^{n} (-1)^{s+t+k+n-l} \binom{m}{s} \binom{s}{k} \binom{n}{t} \binom{t}{n-l} p_1^s (1-p_2)^t \\ &= (-1)^{k+n-l} \left[\sum_{s'=0}^{m-k} (-1)^{s'+k} \binom{m}{s'+k} \binom{s'+k}{k} p_1^{s'+k} \right] \\ &\quad \cdot \left[\sum_{t'=0}^{l} (-1)^{t'+n-l} \binom{n}{t'+n-l} \binom{t'+n-l}{n-l} (1-p_2)^{t'+n-l} \right] \\ &= p_1^k (1-p_2)^{n-l} \left[\sum_{s'=0}^{m-k} (-1)^{s'} \binom{m}{k} \binom{m-k}{s'} p_1^{s'} \right] \left[\sum_{t'=0}^{l} (-1)^{t'} \binom{n}{l} \binom{l}{t'} (1-p_2)^{t'} \right] \\ &= \binom{m}{k} \binom{n}{l} p_1^k p_2^l (1-p_1)^{m-k} (1-p_2)^{n-l} \end{aligned}$$

when $\frac{k}{m} - \frac{l}{n} = 0$, we have $(k,l) \in \{(0,0),(m,n)\}$ and hence:

$$\begin{aligned} S_{m,n}(s,t) &= \left(\left[\max\left(0, \frac{n-t}{n} \right), \min\left(\frac{s}{m}, 1 \right) \right] \cap \mathbb{Z} \right) \\ &\quad \cup \left(\left[\max\left(-1, \frac{-t}{n} \right), \min\left(\frac{s-m}{m}, 0 \right) \right] \cap \mathbb{Z} \right) \\ &= \left(\left[\frac{n-t}{n}, \frac{s}{m} \right] \cap \mathbb{Z} \right) \cup \left(\left[-\frac{t}{n}, \frac{s-m}{m} \right] \cap \mathbb{Z} \right) \end{aligned}$$

For this case, $\frac{-k}{m}$ is either 0 or -1 and $\frac{n-l}{n}$ is either 0 or 1 so, now from the theorem we get,

$$\begin{aligned} &Pr(D=0) \\ &= \sum_{s=k}^{m} \sum_{t=n-l}^{n} (-1)^{s+t+k+n-l} \binom{m}{s} \binom{s}{k} \binom{n}{t} \binom{t}{n-l} p_1^s (1-p_2)^t \end{aligned}$$

$$
\begin{aligned}
&= \sum_{s=0}^{m}\sum_{t=n}^{n}(-1)^{s+t+n}\binom{m}{s}\binom{s}{0}\binom{n}{t}\binom{t}{n}p_1^s(1-p_2)^t \\
&\quad + \sum_{s=m}^{m}\sum_{t=0}^{n}(-1)^{s+t+m+n-n}\binom{m}{s}\binom{s}{m}\binom{n}{t}\binom{t}{n-n}p_1^s(1-p_2)^t \\
&= \sum_{s=0}^{m}(-1)^{s+n+n}\binom{m}{s}\binom{n}{n}\binom{n}{n}p_1^s(1-p_2)^n \\
&\quad + \sum_{t=0}^{n}(-1)^{m+t+m}\binom{m}{m}\binom{m}{m}\binom{n}{t}p_1^m(1-p_2)^t \\
&= \sum_{s=0}^{m}(-1)^s\binom{m}{s}p_1^s(1-p_2)^n + \sum_{t=0}^{n}(-1)^t\binom{n}{t}p_1^m(1-p_2)^t \\
&= (1-p_1)^m(1-p_2)^n + p_1^m\left(1-(1-p_2)\right)^n \\
&= (1-p_1)^m(1-p_2)^n + p_1^m p_2^n
\end{aligned}
$$

Proof of corollary 2

For $m=n$ and $k=l$, the theorem reduces to,

$$Pr(D=0) = (-1)^n\sum_{s=0}^{n}(-1)^s\binom{n}{s}p_1^s\sum_{t=0}^{n}(-1)^t\binom{n}{t}(1-p_2)^t\sum_{i\in S_{n,n}(s,t)}\binom{s}{k+in}\binom{t}{n-k-in},$$

where,

$$i\in S_{n,n}(s,t) = \left[\max\left(-\frac{k}{n},\frac{n-k-t}{n}\right),\min\left(\frac{s-k}{n},\frac{n-k}{n}\right)\right]\cap\mathbb{Z}.$$

$$in\in S_{n,n}(s,t) = \left[\max(-k,n-k-t)),\min(s-k,n-k)\right)\right]\cap\mathbb{Z}.$$

$$k+in\in S_{n,n}(s,t) = \left[\max(0,n-t),\min(s,n)\right].$$

Now we replace $k+in$ by u and obtain the following result:

$$
\begin{aligned}
&Pr(D=0) \\
&= (-1)^n\sum_{s=0}^{n}(-1)^s\binom{n}{s}p_1^s\sum_{t=0}^{n}(-1)^t\binom{n}{t}(1-p_2)^t\sum_{u=\max(0,n-t)}^{\min(s,n)}\binom{s}{u}\binom{t}{n-u}. \\
&= (-1)^n\sum_{s=0}^{n}(-1)^s\binom{n}{s}p_1^s\sum_{t=0}^{n}(-1)^t\binom{n}{t}(1-p_2)^t\sum_{u=n-t}^{s}\binom{s}{u}\binom{t}{n-u}
\end{aligned}
$$

Proof of corollary 3

The exact distribution of *D*, using lemma, is given by;

$$Pr\left(D=\frac{k}{m}-\frac{l}{n}\right) = \sum_{s=0}^{m}\sum_{u=0}^{s}\sum_{t=0}^{n}\sum_{v=0}^{t}(-1)^{s+t+u+v}\binom{m}{s}\binom{s}{u}\binom{n}{t}\binom{t}{v}p_1^s(1-p_2)^t\delta_{kn+(n-l)m}(un+vm)$$

where $\delta_a(x)=1$ if $x=a$ and 0 otherwise. Let us define a set $H_{s,t}$ as follows:

$$
\begin{aligned}
H_{s,t} &= \left\{(u,v)\in\mathbb{N}^2 : 0\le u\le s, 0\le v\le t, un+vm = kn+(n-l)m\right\} \\
&= \left\{(u,v)\in\mathbb{N}^2 : 0\le u\le s, 0\le v\le t, v=(k-u)\frac{n}{m}+n-l\right\} \\
&= \left\{\left(u,(k-u)\frac{n}{m}+n-l\right)\in\mathbb{N}^2 : 0\le u\le s, 0\le (k-u)\frac{n}{m}+n-l\le t\right\} \\
&= \left\{\left(u,(k-u)\frac{n}{m}+n-l\right)\in\mathbb{N}^2 : 0\le u\le s, -t\le (u-k)\frac{n}{m}+l-n\le 0\right\}
\end{aligned}
$$

$$= \left\{ \left(u, (k-u)\frac{n}{m} + n - l \right) \in \mathbb{N}^2 : 0 \le u \le s, n - l - t \le (u-k)\frac{n}{m} \le n - l \right\}$$

$$= \left\{ \left(u, (k-u)\frac{n}{m} + n - l \right) \in \mathbb{N}^2 : 0 \le u \le s, k + (n-l-t)\frac{m}{n} \le u \le k + (n-l)\frac{m}{n} \right\}$$

$$= \left\{ \left(u, (k-u)\frac{n}{m} + n - l \right) \in \mathbb{N}^2 : \max\left(0, k + (n-l-t)\frac{m}{n} \right) \right.$$

$$\left. \le u \le \min\left(s, k + (n-l)\frac{m}{n} \right) \right\}$$

$$= S_{s,t}$$

$$Pr\left(D = \frac{k}{m} - \frac{l}{n} \right)$$

Thus,

$$= \sum_{s=0}^{m} \sum_{t=0}^{n} \sum_{u \in S_{s,t}} (-1)^{s+t+u+(k-u)\frac{n}{m}+n-l} \binom{m}{s} \binom{s}{u} \binom{n}{t} \binom{t}{(k-u)\frac{n}{m} + n - l} p_1^s (1-p_2)^t$$

Proof of corollary 4

Using Corollary (3), the exact distribution of D for $m = n$ and $p_1 = p_2$ is given by

$$Pr\left(D = \frac{k}{n} - \frac{l}{n} \right)$$

$$= \sum_{s=0}^{n} \sum_{t=0}^{n} \sum_{u \in S_{s,t}} (-1)^{s+t+u+(k-u)\frac{n}{n}+n-l} \binom{n}{s} \binom{s}{u} \binom{n}{t} \binom{t}{(k-u)\frac{n}{n} + n - l} p_1^s (1-p_1)^t$$

$$= \sum_{s=0}^{n} \sum_{t=0}^{n} \sum_{u \in S_{s,t}} (-1)^{s+t+u+k-u+n-l} \binom{n}{s} \binom{s}{u} \binom{n}{t} \binom{t}{k-u+n-l} p_1^s (1-p_1)^t$$

$$= \sum_{s=0}^{n} \sum_{t=0}^{n} \sum_{u \in S_{s,t}} (-1)^{s+t+n+k-l} \binom{n}{s} \binom{s}{u} \binom{n}{t} \binom{t}{k-l+u+n} p_1^s (1-p_1)^t$$

where,

$$S_{s,t} = \left\{ (u, k-l+n-u) \in \mathbb{N}^2 : \max(0, k-l+n-t) \le u \le \min(s, k-l+n) \right\}$$

Since both k and l run from 0 to n so $Pr\left(D = \frac{k}{n} - \frac{l}{n} \right) = Pr\left(-D = \frac{l}{n} - \frac{k}{n} \right)$.