Original Article

DOI: 10.14235/bas.galenos.2020.4202
Bezmialem Science 2021;9(2):212-8

Analyzing the Odds Ratio Via Distribution Function

Odds Oranının Dağılım Fonksiyonuna Bağlı Olarak İncelenmesi

Mehmet GÜRCAN[1], Mehmet Onur KAYA[2], Yunus GÜRAL[1]

[1]Fırat Univeristy Faculty of Science, Division of Statistics, Elazığ, Turkey

[2]Fırat University Faculty of Medicine, Department of Biostatistics and Medical Informatics, Elazığ, Turkey

ABSTRACT

Objective: The use of cross tables was frequently seen in early literature research in the biostatistics. Furthermore, its importance in many clinical examinations is still evident today. The aim of this study is to investigate how the 2x2 type tables are perceived in probability literature and how some studies are applied in practice. Thus, different methods can be developed for the purposes of applications.

Methods: The method used to determine the distribution of a 2x2 type table is to consider one cell of a table as a random variable and calculate the probability that this variable can take the observed value. Hypergeometric distribution was taken into consideration in the study. This issue is explained in the methodology section of the study.

Results: Some of the important statistics obtained from 2x2 type tables are the numerical statistical values that direct the researcher in experimental studies such as odds ratio. Considering the distribution of the table, the probabilities of these values are a very important finding for the experimental study. In particular, a high probability value is a measure of how well the statistical value commonly used in biostatistics applications, such as the odds ratio, represents the experimental study performed.

Conclusion: According to the findings of the study, one of the observed results is the determination of the maximum probability ratio representing the experimental study, and the other is the weighted odds ratios that are used to combine odds ratios in the meta-analysis.

Keywords: Contingency table, probability distribution, hypergeometric distribution, odds ratio, relative risk, meta-analysis

ÖZ

Amaç: Biyoistatistik alanındaki erken literatür araştırmalarında dahi çapraz tabloların kullanımına sıkça rastlanmakla beraber günümüzde de önemini birçok klinik incelemedeki kolay kullanımı ile göstermektedir. Bu çalışmadaki amacımız 2x2 tipli tabloların olasılık literatüründe nasıl algılandığını ve bu konu ile ilgili literatürde yapılan birtakım çalışmaların pratikte nasıl uygulandığının araştırılmasıdır. Bu sayede uygulamalardaki amaçlar için farklı birtakım yöntemler rahatlıkla geliştirilebilir.

Yöntemler: 2x2 tipindeki bir tablonun dağılımın belirlenmesi için temelde kullanılan yöntem tablonun bir gözesinin tesadüfi değişken olarak kabul edilmesi ve bu değişkenin gözlenen değeri alma olasılığının hesaplanmasıdır. Genelliği bozmaksızın çalışmada hipergeometrik dağılım dikkate alınmıştır. Bunun nedeni çalışmanın yöntemler kısmında açıklanmaktadır.

Bulgular: 2x2 tipli bir tablodan elde edilen önemli istatistiklerden bazıları, relatif risk ve odds oranı gibi, deneysel çalışmalarda araştırmacıya yön gösteren, sayısal istatistik değerleridir. Tablonun dağılımı dikkate alındığında bu değerlerin olasılıkları yapılan deneysel çalışma için oldukça önemli bir bulgudur. Özellikle olasılığın yüksek olması, odds oranı gibi, biyoistatistik alanındaki uygulamalarda sıkça kullanılan bir istatistik değerinin yapılan deneysel çalışmayı ne kadar iyi temsil ettiğinin bir ölçüsüdür.

Sonuç: Yapılan çalışmada elde edilen bulgulara bakıldığında gözlemlenen sonuçlardan birisi deneysel çalışmayı temsil eden maksimum olasılıklı odds oranının belirlenmesi, diğeri ise meta-analizinde odds oranları birleştirilirken ağırlık olarak odds oranlarının olasılıklarının kullanılabilmesidir.

Anahtar Sözcükler: Kontenjans tablosu, olasılık dağılımı, hipergeometrik dağılım, odds oranı, relatif risk, meta-analizi

Address for Correspondence: Mehmet Onur KAYA, Fırat University Faculty of Medicine, Department of Biostatistics and Medical Informatics, Elazığ, Turkey
E-mail: monurkaya@gmail.com **ORCID ID:** orcid.org/0000-0001-8052-0484

Received: 17.02.2020
Accepted: 21.03.2020

Cite this article as: Gürcan M, Kaya MO, Güral Y. Analyzing the Odds Ratio via Distribution Function. Bezmialem Science 2021;9(2):212-8.

Introduction

One of the problems encountered in scientific research is the inadequacy of data. This can be due to the rarity of data, as well as the lack of time and cost or the lack of specialized personnel. For this reason, especially in health researches, clinical trials and studies are undertaken on a limited number of units. Sometimes, it is necessary to work with small samples for ethical purposes. In such a case, combining studies with similar characteristics by different researchers may make the study findings more meaningful. For these reasons, developing suitable combination methods is necessary.

The most striking example of this is the combination of odds ratios (ψ). Odds ratio combining methods in the literature are Mantel-Haenszel, Peto, General Variance, and DerSimonian-Laird methods. Detailed information on these methods can be found in Katz et al. (1) and Morris and Gardner (2). In these studies, important information is given about establishing confidence intervals of odds ratio. The normal distribution was used to establish the confidence interval. However, the condition of normal distribution may not always be possible. In this case, it is important to determine the distribution of odds ratio. No study in the literature has reported the distribution of odds ratio. However, a distribution that can be used in contingency tables has been examined by Patnaik (3) and Stevens (4). Studies of these researchers will be given with examples in the following sections. These examples were very useful in calculating the distribution of odds ratio. The distribution of odds ratio will be shown in the example in the Results section of our study. In addition, the distribution of combined odds ratios will be calculated in real data application.

Some Probabilistic Notes on Contingency Tables

In biostatistics, the statistical methods which are frequently used in both retrospective and prospective studies are based on statistics such as relative risk and odds ratio obtained from the information in Table 1. Therefore, it is very important to examine the probabilistic features of this table.

As retrospective study is limited to observed data, experimental values are fixed. However, it does not mean that it cannot vary depending on the retrospective follow-up period or other reasons. Thus, the value of *a* in the table is the observation value of X_1. The same applies to the control group. The probability $\Pr\{X_1 = a\}$ can be calculated by the ratio of desired states to all possible states, as in the hypergeometric probability. The number of possible states is written as follows,

$$\frac{N!}{a!b!c!d!}$$

The number of desired states can be calculated as follows,

$$\binom{m}{a}\binom{n}{d}\binom{r}{c}\binom{s}{b} = \frac{m!n!r!s!}{a!(m-a)!d!(n-d)!c!(r-c)!b!(s-b)!}$$

We have

$$Pr\{X_1 = a\} = \frac{m!n!r!s!}{N!a!b!c!d!}$$

where $max\{0, m + r - N\} \leq X_1 \leq min\{m, r\}$.

Sample 1. Consider the data in Table 2

In this example, the variable X_1 takes values between $0 \leq X_1 \leq 10$. Let us show the possible states and probabilities of variable X_1 in Table 3.

According to Table 3, the probability that X_1 takes the value of 5 is the highest probability. The graph of the probability values in Table 3 is as follows,

Methodology

In literature, the first study about this probability belongs to P. B. Patnaik in 1948. In the study, the common cell of the case and the positive effect was accepted as a random variable, and it was shown by P. B. Patnaik that it has a hypergeometric distribution. This makes it easier to obtain the term representing the odds ratio from the conditional probability of the hypergeometric distribution. Therefore, hypergeometric distribution was taken into consideration in the study. Patnaik calculated the mean and variance of the distribution with the help of the hypergeometric distribution as $\mathrm{E}X_1 = \mathrm{mr/N}$ and $\mathrm{Var}X_1 = \mathrm{mnrs/N^2(N-1)}$ [3]. The mean $\mathrm{E}X_1$ calculated by Patnaik is used as the expected value of the cells in the chi-square relationship test. This was

Table 1. Structure of contingency table

	Positive effect	Negative effect	Total
Case	$X_1 = a$	b	m
Control	$X_2 = c$	d	n
Total	r	s	N

Table 2. Sample data

	Positive effect	Negative effect	Total
Case	$X_1 = a = 8$	$b = 2$	$m = 10$
Control	$X_2 = c = 10$	$d = 15$	$n = 25$
total	$r = 18$	$s = 17$	$N = 35$

followed by W. L. Stevens. Stevens assumed the conditional probability of the variable X_1 as a function of a under the condition that all marginal totals are known. As follows [4],

$$Pr\{X_1 = a|X_1 + X_2 = r, m\} = C\binom{m}{a}\binom{n}{r-a}\psi^a$$

where $\psi = ad/bc$ is odds ratio. The conditional probability mentioned above can be obtained as the multiplication of two binomial probabilities,

$$Pr\{X_1 = a|X_1 + X_2 = r, m\} = Pr\{X_1 = a\}Pr\{X_2 = c\}$$

$$= \binom{m}{a} p_1^{\ a} q_1^{\ m-a} \binom{n}{c} p_2^{\ c} q_2^{\ n-c}$$

$$= \frac{q_1^m q_2^n p_2^r}{q_2^r} \binom{m}{a}\binom{n}{r-a}\psi^a$$

where p_1 and p_2 are the probability of success in case and control groups, respectively. In addition, the ratio p_1/p_2 is called relative risk. As a result, this equation ensures that conditional probability can be written as a function of *a*. This is an important result for 2x2 tables. If the observation value of variable X_1 is smaller than some values in the possible order, ψ will remain smaller than odds ratios of these values, otherwise vice versa. Using this feature, Jerome Cornfield formed confidence interval with $1-\alpha$ probability for odds ratio in his study done in 1956 (5). Cornfield obtained the lower limit ψ_1 for ψ from the solution of the following equation,

$$\sum_{y=X_1}^{m} C\binom{m}{a}\binom{n}{r-a}\psi^a = \frac{\alpha}{2}$$

Similarly, he obtained the upper limit ψ_2 for ψ from the solution of the following equation,

$$\sum_{y=0}^{X_1} C\binom{m}{a}\binom{n}{r-a}\psi^a = \frac{\alpha}{2}$$

Thus, the confidence interval can be written as $Pr\{\psi_1 \leq \psi \leq \psi_2\} = 1-\alpha$.

Results and Discussion

Here, conditional probability is obtained as the multiplication of two binomial distributions by independent variables X_1 and X_2. Then normal distribution test procedures can be used in hypothesis tests since the limit distributions X_1 and X_2 approach normal distribution. However, this may be the case if the marginal totals are large enough. Otherwise, it may cause incorrect interpretations. It is more accurate to obtain the exact distribution and to test with nonparametric method when an exact test statistic for ψ is desired to be created. In order for the mean and variance of ψ to be real, it is sufficient for the cells to satisfy the conditions of $a < m$ and $c > 0$. In this case, it is necessary to obtain the conditional distribution of ψ depend on these conditions. Therefore, many researchers use the normal distribution approach. The conditional distribution can be obtained by dividing binomial probabilities to the probability of $Pr\{X_1 < m\}$ for X_1 and to the probability of $Pr\{X_2 > 0\}$ for X_2. In the following example, we show the possible values and possibilities of ψ.

Sample 2. Let be the sample data as follows,

The multiplication probability table and the probability table of ψ can be formed with the data in Table 4 using the conditional probabilities of variables X_1 and X_2,

The graph of multiplication probabilities in Table 5 is as follows,

When the probability in Table 6 is taken into consideration, the variable is seen to have the highest probability at $\psi = 4$. It is seen that odds ratio in the experimental data in Table 4 would take high probability value between 2 and 5 ($Pr\{2 \leq \psi \leq 5\} = 0.443$). Such probabilistic information can also be supported in

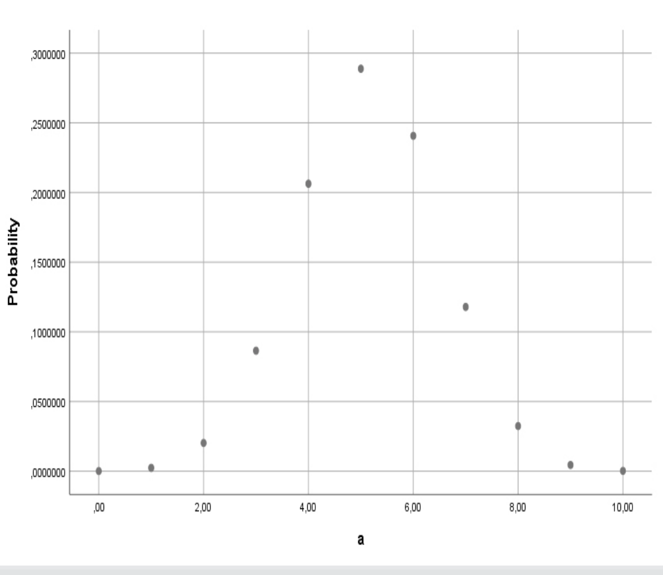

Figure 1. Graph of the probability values in Table 3

Table 3. Probability table

a	$Pr\{X_1 = a\}$	a	$Pr\{X_1 = a\}$
0	0.0001059	6	0.2406714
1	0.0023836	7	0.1178798
2	0.0202606	8	0.0324169
3	0.0864452	9	0.0045023
4	0.2062898	10	0.00023836
5	0.2888057	Total	1

statistical terms by creating rejection and acceptance zones from the distribution obtained at α significance level. Moreover, the mean $E\psi = 6.7133$ obtained from the distribution is an important statistic for ψ. The graph of probabilities in Table 6 is as follows.

Table 6 shows the distribution of ψ. The distribution of ψ can be easily obtained when multiple tables are for the same X_1. Let's assume that there are k tables of X_1. In this case, probabilities for each table are shown as follows,

$$Pr\left\{\psi_j = u\right\} = p_j(u), j = 1, \cdots, k.$$

The distribution of all tables will be as follows,

$$Pr\{\psi = u\} = \frac{1}{k}\sum_u p_j(u), j = 1, \cdots, k.$$

Since the mean E is derived from the k sample selected from the mass, it will be able to represent the mass ideally. Finally, the following sample about combined odds ratio are presented.

The following example table was taken from Afshari et al.(6). This meta-analysis study by Mahdi investigates the effect of opium and smoking on bladder cancer. Table 7 was created by considering only opium use. The distribution and expected value of odds ratio were obtained for each study. At the end of the table is the expected value of the combined odds ratio. The matlab program used in the calculation is attached.

Conclusion

In general, when we look at the studies in the field of biostatistics, a comprehensive and technically rich literature is emerging. This is due to the fact that many scientific techniques are combined with medical data gathered under biostatistics. A scientific technique needs not only an opinion, but also an interpretation. The interpretation to be made is usually attributed to the data. However, this interpretation is the common point of data and technique, which increases the scientific value of results

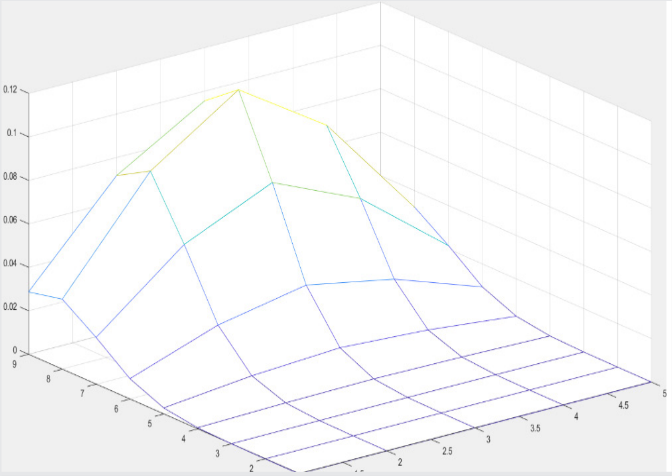

Figure 2. Graph of the multiplication probabilities in Table 5

Table 4. Sample data

	Positive effect	Negative effect	Total
Case	X_1 = a = 8	b = 2	m = 10
Control	X_2 = c = 3	d = 3	n = 6
Total	r = 11	s = 5	N = 16

Table 5. Multiplication probabilities

	$Pr\{X_2 = k\|X_2 > 0\}$						
	k	1	2	3	4	5	6
$Pr\{X_1 = k\|X_1 < 10\}$	0	1.09×10^{-8}	2.73×10^{-8}	3.64×10^{-8}	2.73×10^{-8}	1.09×10^{-8}	1.82×10^{-9}
	1	4.37×10^{-7}	1.09×10^{-6}	1.45×10^{-6}	1.09×10^{-6}	4.37×10^{-7}	7.28×10^{-8}
	2	7.86×10^{-6}	1.96×10^{-5}	2.62×10^{-5}	1.96×10^{-5}	7.86×10^{-6}	1.31×10^{-6}
	3	8.39×10^{-5}	2.09×10^{-4}	2.79×10^{-4}	2.09×10^{-4}	8.39×10^{-5}	1.39×10^{-5}
	4	5.87×10^{-4}	1.46×10^{-3}	1.95×10^{-3}	1.46×10^{-3}	5.87×10^{-4}	9.78×10^{-5}
	5	2.81×10^{-3}	6.04×10^{-3}	9.39×10^{-3}	7.04×10^{-3}	2.81×10^{-3}	4.69×10^{-4}
	6	9.39×10^{-3}	2.35×10^{-2}	3.13×10^{-2}	2.34×10^{-2}	9.39×10^{-3}	1.56×10^{-3}
	7	2.14×10^{-2}	5.37×10^{-2}	7.16×10^{-2}	5.37×10^{-2}	2.14×10^{-2}	3.58×10^{-3}
	8	3.22×10^{-2}	8.05×10^{-2}	0.1074	8.05×10^{-2}	3.22×10^{-2}	5.37×10^{-3}
	9	2.86×10^{-2}	7.16×10^{-2}	9.54×10^{-2}	7.16×10^{-2}	2.86×10^{-2}	4.77×10^{-3}

obtained from data and importance of the technique used. Therefore, biostatistics studies are important studies that bring data and technique together. If the odds ratio value obtained from a data in Table 1 is smaller than 1, the factor decreases the risk of disease. If the odds ratio is equal to 1, the factor has no effect on the disease. If the odds ratio is bigger than 1, the factor increases the risk of the disease. Thus, X_1 is the most important variable to ensure a high odds ratio. Considering the coincidence of the value of X_1, it is more important to know the maximum probability value. For this, the distribution of X_1 and its interpretation should be made. In the study, a data table of 2 × 2 type has been shown to have hypergeometric distribution when considered unconditionally. Depending on this distribution, the variable X_1 takes the maximum probability with value $(m + 1)(n + 1)/(N + 2)$. In addition, this value is the maximum probability value of the odds ratio. This result is very important in terms of both data and theory. If data were not interpreted with the theoretical structure, then a conclusion will never be obtained. Similarly, obtaining the distribution of ψ is also important in terms of interpretation. When values obtained from different tables are combined in a probability distribution, the distribution of a single variable ψ can be obtained for all tables. This result is also very important for meta-analysis. The mean $E\psi$ for a single table is so important for combined tables. Many methods have been presented to combine odds ratios in literature; however, no such method has been presented. The reason for this is that presented methods have the ease of calculation in terms of researchers. However, using probabilistic methods is more important for more optimal results. Finally, one point that should be taken into consideration is that if the number of case and control is sufficient in a *2x2* table, parametric methods can be used easily. An example would be the Mantel-Haenszel, Peto, General Variance, and DerSimonian-Laird methods. If the number of data is quite low, it is more appropriate to use probabilistic methods.

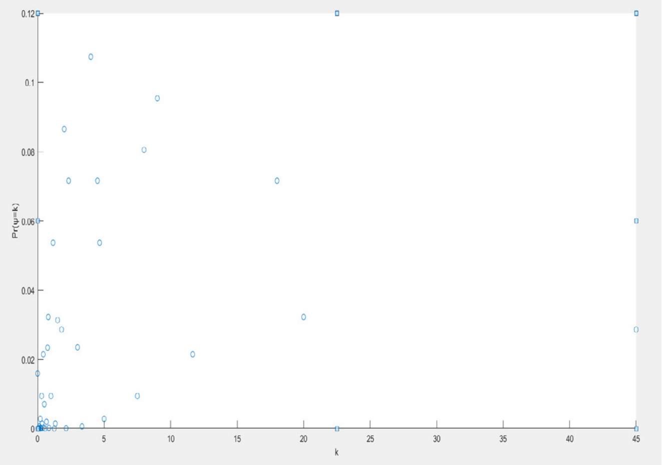

Figure 3. Graph of the probabilities in Table 6

Table 6. Probability values of ψ

k	$Pr\{\psi = k\}$	k	$Pr\{\psi = k\}$
0	0.015862197	1	9.39×10^{-3}
0.022	4.37×10^{-7}	1.16	5.37×10^{-2}
0.031	1.96×10^{-5}	1.25	7.86×10^{-6}
0.05	7.86×10^{-6}	1.333	1.46×10^{-3}
0.055	1.09×10^{-6}	1.5	3.13×10^{-2}
0.085	8.39×10^{-5}	1.8	2.86×10^{-2}
0.111	1.45×10^{-6}	2	0.08654
0.133	5.87×10^{-4}	2.142	8.39×10^{-5}
0.2	2.81×10^{-3}	2.33	7.16×10^{-2}
0.214	2.09×10^{-4}	3	2.35×10^{-2}
0.222	1.09×10^{-6}	3.333	5.87×10^{-4}
0.25	2.62×10^{-5}	4	0.1074
0.3	9.39×10^{-3}	4.5	7.16×10^{-2}
0.333	1.46×10^{-3}	4.66	5.37×10^{-2}
0.428	2.79×10^{-4}	5	2.81×10^{-3}
0.466	2.14×10^{-2}	7.5	9.39×10^{-3}
0.5	7.06×10^{-3}	8	8.05×10^{-2}
0.555	4.37×10^{-7}	9	9.54×10^{-2}
0.666	1.95×10^{-3}	11.66	2.14×10^{-2}
0.75	2.34×10^{-2}	18	7.16×10^{-2}
0.8	3.22×10^{-2}	20	3.22×10^{-2}
0.857	2.09×10^{-4}	45	2.86×10^{-2}

Table 7. Opium exposure among cases and controls in the primary studies entered into meta-analysis

First author		+	−	*Eψ*
1. Akbari	Case	43	155	6.2324
	Control	18	378	
2. Aliramaji	Case	58	117	
	Control	27	148	2.8501
3. Asgari	Case	13	39	
	Control	5	103	9.0933
4. Ghadimi	Case	16	136	
	Control	2	150	10.2698
5. Hosseini	Case	60	119	
	Control	7	172	5.0727
6. Ketabchi	Case	80	32	
	Control	31	99	8.5306
7. Lotfi	Case	52	147	
	Control	21	179	3.1957
8. Nourbakhsh	Case	41	214	
	Control	12	243	4.2959
9. Sadeghi	Case	44	44	
	Control	7	81	14.2746
10. Tootonchi	Case	16	126	
	Control	7	135	2.9773
Combined *Eψ* = 6.6792				

Ethics

Ethics Committee Approval: There is no approval of the Ethics Committee, since there is no “animal or human element” in our study, and the study was completely conducted on hypothetical theoretical data.

Peer-review: Externally peer reviewed.

Authorship Contributions

Concept: M.G., M.O.K., Y.G., Design: M.G., M.O.K., Analysis or Interpretation: M.G., M.O.K., Y.G., Literature Search: M.G., Writing: M.G., Y.G.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

References

1. Katz D, Baptista J,Azen SP, Pike MC. Obtaining Confidence Intervals for the Risk Ratio in Cohort Studies. BIOMETRICS 1978;34:469-74.
2. Morris JA, Gardner MJ. Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates. Br Med J (Clin Res Ed) 1988;296:1313-6.
3. PATNAIK PB. The power function of the test for the difference between two proportions in a 2 X 2 table. Biometrika 1948;35:157-75.
4. Stevens WL. Mean and variance of an entry in a contingency table. Biometrika 1951;38:468-70.
5. Cornfield J. A Statistical Problem Arising from Retrospective Studies. Third Berkeley Symp. on Math. Statist. and Prob 4. 1956; p. 135-48.
6. Afshari M, Janbabaei G, Bahrami MA, Moosazadeh M. Opium and bladder cancer: A systematic review and meta-analysis of the odds ratios for opium use and the risk of bladder cancer. PLoS One 2017;12:0178527.

Appendix

```
A = input (“a=”); b = input(“b=”); c = input(“c=”); d = input(“d=”);
M = a + b; n = c + d; pv = a/m; qv = b/m; pk = c/n; qk = d/n;
V = [0:1:m]; k =[0:1:n]; A = zeros (m, 1); B = zeros (n, 1);
p1v = zeros (m + 1, 1); p2k = zeros (n + 1, 1); phi = zeros (m – 1, n – 1); Pc = zeros (m – 1, n – 1); C = zeros (m – 1, n – 1);
SC = zeros (1, n – 1); SSC = 0; for i = 1: m + 1
        p1v (i) = nchoosek (m, v[i]) × (pv^v[i]) × (qv^[m – v(i)]); end for I = 1: n + 1
        p2k(i) = nchoosek (n, k [i]) × (pk^k[i]) × (qk^[n – k(i)]);
end
tv = 1 – p1v (m + 1); tk = 1 – p2k (1);
for i = 1: m
        A(i) = p1v(i)/tv;
end
for i = 1:n
        B (i) = p2k (i + 1)/tk;
end
Phi0 = A (1) × (sum [B]) + B(n) × (sum [A] – A [1])
a1 = (1:1:m – 1);c1 = (1:1:n – 1);
for i = 1:m – 1
for j = 1:n – 1
b1 = m – a1(i);
  d1 = n – c1(j);
  phi (i, j) = a1(i) × d1/(b1 × c1[j]);

for i = 2:m
for j = 1:n – 1
Pc(i – 1, j) = A (i) × B (j);
end
end
for i = 1:m – 1
for j = 1:n – 1
C(i, j) = phi (i, j) × Pc (i, j);
end
end
phi %values of phi
Pc %multiplication table
SC = sum(C);
SSC = sum (SC)    %expected value of phi
```