STATISTICS IN TRANSITION new series, March 2023
Vol. 24, No. 2, pp. 113–128, DOI 10.59170/stattrans-2023-023
Received – 03.01.2022; accepted – 06.12.2022

A new confidence interval for the odds ratio

Zofia Zielińska-Kolasińska,[1] Wojciech Zieliński[2]

ABSTRACT

We consider the problem of interval estimation of the odds ratio. An asymptotic confidence interval is widely applied in economics, medicine, sociology, etc. Unfortunately, this confidence interval has a poor coverage probability, significantly smaller than the nominal confidence level. In this paper, a new confidence interval is proposed. Its construction requires only information on the sizes of samples and the sample odds ratio. The coverage probability of the proposed confidence interval is at least the nominal confidence level.

Key words: confidence interval, odds ratio.

1. Introduction

In many practical sciences such as economy, medicine, sociology, etc. dichotomous variate is observed. Such variate is to be compared in two independent groups. Commonly used is the difference of two fractions (the risk difference), the ratio of two proportions (the relative risk) and the odds ratio. The relative risk and the odds ratio are relative measurements for comparison of two variates, while the risk difference is an absolute measurement.

The odds ratio is one of the parameters commonly used in such comparisons, especially in two-arm binomial experiments. This indicator was firstly applied by Cornfield (1951). The literature devoted to the analysis of odds ratio and its estimators is very rich, see, e.g. Encyclopedia of Statistical Sciences (2006) Volume 9, pp. 5722–5726 and the literature therein.

However, the problem is in the interval estimation. In general, there are two approaches to the problem. The first one consists in the analysis of 2×2 tables (Edwards (1963), Gart (1971), Thomas (1971)). The second approach is based on logistic model in which the odds ratio has a direct relationship with the regression coefficient (Gart (1971), McCullagh (1980), Morris (1988)). Wang, Shan (2015) constructed exact confidence interval for the odds ratio based on another approach. Namely, they applied the so-called rank function. A very exhaustive review of different confidence intervals for the odds ratio may be found in Andrés et.al (2020). Unfortunately, all those confidence intervals have one very important disadvantage: their real probability of coverage is significantly smaller than the nominal one. It is in contradiction to the Neyman (1934 p. 562) definition of a confidence interval. Hence, the risk of a wrong conclusion (i.e. overestimation or underestimation) is greater than the assumed one and unluckily remains unknown.

[1]Poland. E-mail: zzk@egrp.pl, ORCID: https://orcid.org/000-0001-8845-758X.
[2]Department of Econometrics and Statistics, Warsaw University of Life Sciences, Warsaw, Poland. E-mail: wojciech_zielinski@sggw.edu.pl, ORCID: https://orcid.org/0000-0003-0749-8764.

The most commonly used in applications is an asymptotic interval for odds ratio derived from logistic model (formula (4) in Section 3). This asymptotic interval is widely used in different statistical packages. There are also many internet scripts for calculating the asymptotic confidence interval (see, e.g. http://www.hutchon.net/ConfidOR.htm). Unfortunately, this confidence interval has some statistical disadvantages discussed in Section 3. To avoid those disadvantages a new confidence interval is proposed. The idea of construction is similar to the idea of construction of the confidence interval for the difference of two probabilities of success (the risk difference) proposed by Zieliński (2020a). It is based on the exact distribution of the sample odds ratio, hence it works for large as well as for small samples. The coverage probability of that confidence interval is at least the nominal confidence level.

In Section 2 a new confidence interval is constructed. In Section 3 some disadvantages of the asymptotic confidence interval are discussed. An example of application is given in Section 4. Final conclusions are given in Section 5.

2. A new confidence interval

Consider two independent r.v.'s ξ_A and ξ_B distributed as $Bin(n_A, p_A)$ and $Bin(n_B, p_B)$, respectively. The problem is in estimating the odds ratio:

$$OR = \frac{(p_A/(1-p_A))}{(p_B/(1-p_B))} = \frac{p_A}{(1-p_A)} \cdot \frac{(1-p_B)}{p_B}.$$

Let n_{A1} and n_{B1} be observed numbers of successes. The data are usually organized in a 2×2 table:

	success	failure	
Group A	n_{A1}	n_{A0}	n_A
Group B	n_{B1}	n_{B0}	n_B
	n_1	n_0	n

The standard estimator of OR is as follows:

$$\widetilde{OR} = \frac{n_{A1}}{n_A - n_{A1}} \cdot \frac{n_B - n_{B1}}{n_{B1}}. \tag{1}$$

The estimator $\widetilde{OR}$ is undefined for $n_{A1} = n_A$ or $n_{B1} = 0$. The probability of the nonexistence of $\widetilde{OR}$ equals

$$P_{p_A,p_B}\{\xi_A = n_A \text{ or } \xi_B = 0\} = p_A^{n_A} + (1-p_B)^{n_B} - p_A^{n_A}(1-p_B)^{n_B}.$$

For a given odds ratio equal to $r > 0$

$$p_B = \frac{p_A}{p_A + r(1-p_A)} \text{ and } 1 - p_B = \frac{r(1-p_A)}{p_A + r(1-p_A)}$$

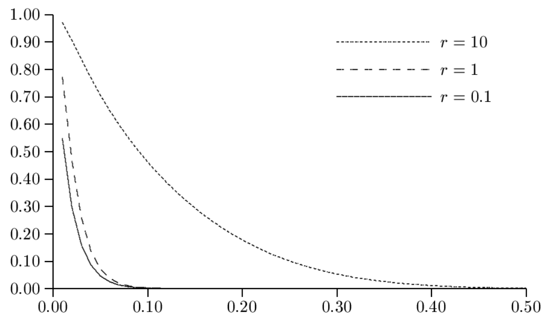

Figure 1: The probability $P_{r,p_A}\{\xi_A = n_A \text{ or } \xi_B = 0\}$

and

$$P_{r,p_A}\{\xi_A = n_A \text{ or } \xi_B = 0\} = p_A^{n_A} + (1 - p_A^{n_A})\left(\frac{r(1-p_A)}{p_A + r(1-p_A)}\right)^{n_B}.$$

In Figure 1 the above probability (y axis) is shown for different values of true odds ratio r with respect to probability p_A (x axis). In Figure 1 $n_A = 60$ and $n_B = 70$ were taken.

It is seen that the probability of nonexistence of $\widetilde{OR}$ is quite high, especially for small values of the probability p_A. To eliminate that phenomena another approach is needed.

Usually, the problem of estimating an odds ratio is considered in the following statistical model:

$$(\{0,1,\ldots,n_A\} \times \{0,1,\ldots,n_B\}, \{Bin(n_A,p_A) \cdot Bin(n_B,p_B), (p_A,p_B) \in (0,1) \times (0,1)\}).$$

Since we are interested in estimating the odds ratio OR, consider now a new statistical model. This model is the one-parameter model: the odds ratio is an unknown parameter

$$(\mathscr{X}, \{F_r, 0 \le r \le +\infty\}),$$

where

$$\mathscr{X} = \left\{\frac{n_{A1}}{n_A - n_{A1}} \cdot \frac{n_B - n_{B1}}{n_{B1}} : n_{A1} \in \{0,1,\ldots,n_A\}, n_{B1} \in \{0,1,\ldots,n_B\}\right\}.$$

The cumulative distribution functions (CDF) $F_r(\cdot)$ are defined as follows.

Since the estimator $\widetilde{OR}$ given by formula (1) is undefined for $n_{A1} = n_A$ or $n_{B1} = 0$ we extend the definition of the estimator of the odds ratio in the following way:

$$\widehat{OR} = \begin{cases} 0, & \text{for } (n_{A1} = 0, n_{B1} \geq 1) \text{ or } (n_{A1} \leq n_A - 1, n_{B1} = n_B), \\ +\infty, & \text{for } (n_{A1} = n_A, 1 \leq n_{B1} \leq n_B - 1) \text{ or } (n_{A1} \geq 1, n_{B1} = 0), \\ 1, & \text{for } (n_{A1} = 0, n_{B1} = 0) \text{ or } (n_{A1} = n_A, n_{B1} = n_B), \\ \text{formula } (1), & \text{elsewhere.} \end{cases} \tag{2}$$

The probability of observing $\xi_A = n_{A1}$ and $\xi_B = n_{B1}$ equals

$$P_{p_A,p_B}\{n_{A1}, n_{B1}\} = \binom{n_A}{n_{A1}} p_A^{n_{A1}} (1-p_A)^{n_A - n_{A1}} \binom{n_B}{n_{B1}} p_B^{n_{B1}} (1-p_B)^{n_B - n_{B1}}.$$

Equivalently

$$P_{r,p_A}\{n_{A1}, n_{B1}\} = r^{n_B - n_{B1}} \binom{n_A}{n_{A1}} \binom{n_B}{n_{B1}} \frac{p_A^{n_{A1}+n_{B1}} (1-p_A)^{n_A + n_B - n_{A1} - n_{B1}}}{(p_A + r(1-p_A))^{n_B}}.$$

Let

$$F_{r,p_A}(t) = P_{r,p_A}\left\{\widehat{OR} \leq t\right\} = \sum_{n_{A1}=0}^{n_A} \sum_{n_{B1}=0}^{n_B} P_{r,p_A}\{n_{A1}, n_{B1}\} \mathbf{1}\left(\widehat{OR}(n_{A1}, n_{B1}) \leq t\right),$$

where $\mathbf{1}(q) = 1$ when q is true and $= 0$ elsewhere. For any given $p_A \in (0,1)$ the family $\{F_{r,p_A}, r > 0\}$ is stochastically ordered, i.e.

$$F_{r_1,p_A}(\cdot) \geq F_{r_2,p_A}(\cdot) \text{ for } r_1 \leq r_2.$$

It follows from the fact that for a given n_{A1}, n_{B1} and p_A the probability $P_{r,p_A}\{n_{A1}, n_{B1}\}$ is the decreasing function of odds ratio r.

Let γ be the given confidence level and let $\hat{r}$ be the observed odds ratio. For a given p_A the confidence interval for r takes on the form

$$(Left(\hat{r}, p_A), \; Right(\hat{r}, p_A)), \tag{3}$$

where

$$\begin{cases} Left(\hat{r}, p_A) = \max\left\{r : G_{r,p_A}(\hat{r}) \geq (1+\gamma)/2\right\}, \\ Right(\hat{r}, p_A) = \min\left\{r : F_{r,p_A}(\hat{r}) \leq (1-\gamma)/2\right\}. \end{cases}$$

Here, $G_{r,p_A}(t)$ denotes the probability $P_{r,p_A}\left\{\widehat{OR} < t\right\}$.

The coverage probability by the construction is at least a given confidence level γ. In Figure 2 the coverage probability for $p_A = 0.5$ and $n_A = 50$, $n_B = 10$ is presented ($\gamma = 0.95$). For other values of $p_A \in (0,1)$ the graphs are similar. On the x-axis the value r of the odds ratio is given and on the y-axis the probability of coverage is shown.

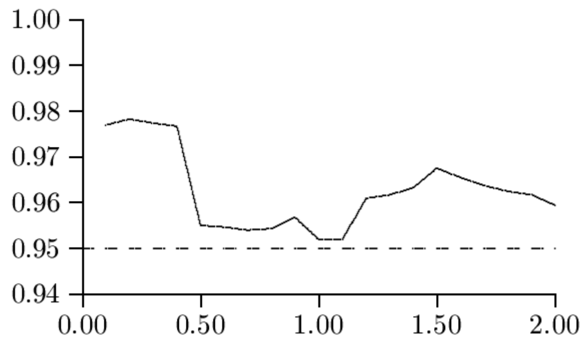

Figure 2: Coverage probability of (3)

Since the probability p_A is unknown it should be treated as a nuisance parameter. In statistics two methods of eliminating a nuisance parameter are common: estimation of such parameter or integration over nuisance parameter. In what follows we choose the second method, i.e. appropriate integration:

$$P_r\{n_{A1}, n_{B1}\} = \int_0^1 P_{r,p_A}\{n_{A1}, n_{B1}\} w(p_A) dp_A,$$

where $w : (0,1) \to R^+$ is a weighting function such that $\int_0^1 w(u)du = 1$. The function w may be chosen quite arbitrary. The choice is interpreted as an *a priori* knowledge of probability p_A. The most common is the choice of function w proportional to $(u-a)^{\alpha-1}(b-u)^{\beta-1}$ for positive α and β and $0 \le a < b \le 1$. In what follows $\alpha = \beta = 1$, $a = 0$ and $b = 1$ is taken. So, it is assumed that the probability p_A may be any number from the interval $(0,1)$.

We obtain

$$\begin{aligned} P_r\{n_{A1}, n_{B1}\} &= \int_0^1 P_{r,p_A}\{n_{A1}, n_{B1}\} dp_A \\ &= (n_A + n_B)! \frac{\binom{n_A}{n_{A1}}\binom{n_B}{n_{B1}}}{\binom{n_A+n_B}{n_{A1}+n_{B1}}} \left(\frac{1}{r}\right)^{n_{B1}} {}_2\widetilde{F}_1\left[n_B, n_{A1}+n_{B1}+1; n_A+n_B+2; 1-\frac{1}{r}\right], \end{aligned}$$

where

$${}_2\widetilde{F}_1[x,y;z;t] = \frac{1}{\Gamma(z-y)\Gamma(y)} \int_0^1 u^{y-1}(1-u)^{z-y-1}(1-ut)^{-x} du \text{ (for } z > y > 0)$$

is the regularized confluent hypergeometric function. The CDF of $\widehat{OR}$ equals (for $t \ge 0$)

$$F_r(t) = P_r\left\{\widehat{OR} \le t\right\} = \sum_{n_{A1}=0}^{n_A} \sum_{n_{B1}=0}^{n_B} P_r\{n_{A1}, n_{B1}\} \mathbf{1}\left(\widehat{OR}(n_{A1}, n_{B1}) \le t\right),$$

where $\mathbf{1}(q) = 1$ when q is true and $= 0$ elsewhere.

Since $\widehat{OR}$ is given by formula (2) the CDF may be written as

$$F_r(t) = \sum_{n_{A1}=0}^{n_A-1} \sum_{n_{B1}=h(n_{A1})}^{n_B} P_r\{n_{A1}, n_{B1}\}$$
$$= (n_A+n_B)! \sum_{n_{A1}=0}^{n_A-1} \sum_{n_{B1}=h(n_{A1})}^{n_B} \frac{\binom{n_A}{n_{A1}}\binom{n_B}{n_{B1}}}{\binom{n_A+n_B}{n_{A1}+n_{B1}}} \left(\frac{1}{r}\right)^{n_{B1}} {}_2\widetilde{F}_1\left[n_B, n_{A1}+n_{B1}+1; n_A+n_B+2; 1-\frac{1}{r}\right],$$

where

$$h(n_{A1}) = \begin{cases} \left\lceil \frac{n_B}{t\left(\frac{n_A}{n_{A1}}-1\right)+1} \right\rceil, & \text{for } n_{A1} \geq 1, \\ 0, & \text{for } n_{A1} = 0, \end{cases}$$

(here $\lceil x \rceil$ denotes the smallest integer not less than x).

The family $\{F_r, r \geq 0\}$ is stochastically ordered, i.e. for a given $t > 0$

$$F_{r_1}(t) \geq F_{r_2}(t) \text{ for } r_1 \leq r_2.$$

It follows from the fact that for a given n_{A1}, n_{B1} and p_A the probability $P_{r,p_A}\{n_{A1}, n_{B1}\}$ is the decreasing function of odds ratio r and hence $P_r\{n_{A1}, n_{B1}\}$ is also decreasing in r.

Let $G_r(t)$ denote the probability $P_r\left\{\widehat{OR} < t\right\}$. Let γ be the given confidence level and let $\hat{r}$ be the observed odds ratio. The confidence interval for r takes on the form

$$(Left(\hat{r}), Right(\hat{r})), \tag{4}$$

where

$$Left(\hat{r}) = \begin{cases} 0, & \hat{r} = 0, \\ 0, & \text{if } \lim_{r\to 0} G_r(\hat{r}) < (1+\gamma)/2, \\ r_*, & r_* = \max\{r : G_r(\hat{r}) \geq (1+\gamma)/2\}, \end{cases}$$

and

$$Right(\hat{r}) = \begin{cases} \infty, & \hat{r} = \infty, \\ \infty, & \text{if } \lim_{r\to\infty} F_r(\hat{r}) > (1-\gamma)/2, \\ r^*, & r^* = \min\{r : F_r(\hat{r}) \leq (1-\gamma)/2\}. \end{cases}$$

Theorem. For $n_A > \frac{2}{1-\gamma} - 1$ the confidence interval for the odds ratio is two-sided and is one-sided otherwise.

For the proof see Appendix 1.

If $\hat{r}$ is the observed odds ratio then the confidence interval for r takes on the following form:

$$\text{for } \hat{r} \in [0,1): \begin{cases} \langle 0, r^*), & \text{for } n_A \le \frac{2}{1-\gamma} - 1, \\ (r_*, r^*), & \text{for } n_A > \frac{2}{1-\gamma} - 1, \end{cases}$$

$$\text{for } \hat{r} \in [1,+\infty): \begin{cases} (r_*, +\infty), & \text{for } n_A \le \frac{2}{1-\gamma} - 1, \\ (r_*, r^*), & \text{for } n_A > \frac{2}{1-\gamma} - 1, \end{cases}$$

where r_* and r^* are given by formula (4). Minimal sample sizes n_A for which two-sided confidence interval exists are given in Table 1.

Table 1: Minimal sample size

γ	0.9	0.95	0.99	0.999
n_A	20	40	200	2000

For a given $r > 0$ the coverage probability, by construction, equals at least γ. Figure 3 shows the coverage probability for $n_A = 60$, $n_B = 70$ and $\gamma = 0.95$. On the x-axis the value r of the odds ratio is given and on the y-axis the probability of coverage is shown. The coverage probabilities are calculated, not simulated.

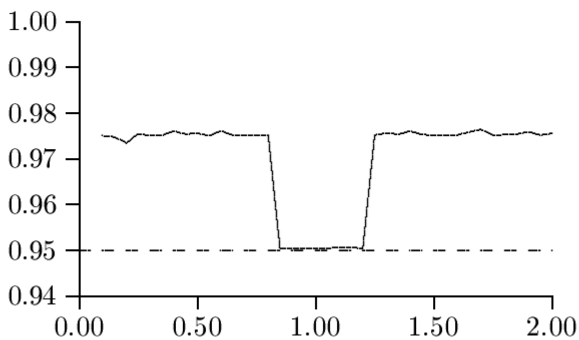

Figure 3: Coverage probability of (4)

Remark. The above considerations are made for A versus B. It is obvious that

$$OR(A \text{ vs } B) = \frac{1}{OR(B \text{ vs } A)}.$$

It is easily seen that the new confidence interval has the following natural property:

$$Left(A \text{ vs } B) = \frac{1}{Right(B \text{ vs } A)} \quad \text{and} \quad Right(A \text{ vs } B) = \frac{1}{Left(B \text{ vs } A)}.$$

In the case of considering B versus A in the Theorem, the sample size n_A should be changed to n_B.

3. Standard confidence interval

Estimating the odds ratio is one of the crucial problems in medicine, biometrics, etc. The most widely used confidence interval at the confidence level γ is of the form

$$\left(\widetilde{OR}\cdot\exp\left(u_{\frac{1-\gamma}{2}}\sqrt{\frac{1}{n_{A1}}+\frac{1}{n_{A0}}+\frac{1}{n_{B1}}+\frac{1}{n_{B0}}}\right),\widetilde{OR}\cdot\exp\left(u_{\frac{1+\gamma}{2}}\sqrt{\frac{1}{n_{A1}}+\frac{1}{n_{A0}}+\frac{1}{n_{B1}}+\frac{1}{n_{B0}}}\right)\right), \tag{5}$$

where u_δ denotes the δ quantile of $N(0,1)$ distribution. In the above formula the estimator $\widetilde{OR}$ is given by (1). Unfortunately, this confidence interval has at least three disadvantages. They are as follows.

1. Confidence interval (5) does not exist if at least one of n_{A0}, n_{A1}, n_{B0} or n_{B1} equals zero or $\widetilde{OR}$ does not exist. The probability of such an event may be quite large, so in many real experiments it may happen (cf. Figure 1) that the confidence interval is undefined.

2. The coverage probability of c.i. (5) is less than the nominal one. In Figure 4 the coverage probability is shown for $n_A = 60$, $n_B = 70$ and $\gamma = 0.95$ (the value r of odds ratio is given on the x-axis and the coverage probability is given on the y-axis). The probability of wrong conclusion, i.e. of overestimation or underestimation is greater than the assumed 0.05. It means that the true value of odds ratio may be smaller than the left end of the confidence interval (4) or greater than its right end. The risk of such an event is greater than the nominal 0.05 and unfortunately remains unknown. Note that this is in contradiction to Neyman (1934, p. 562) definition of a confidence interval.

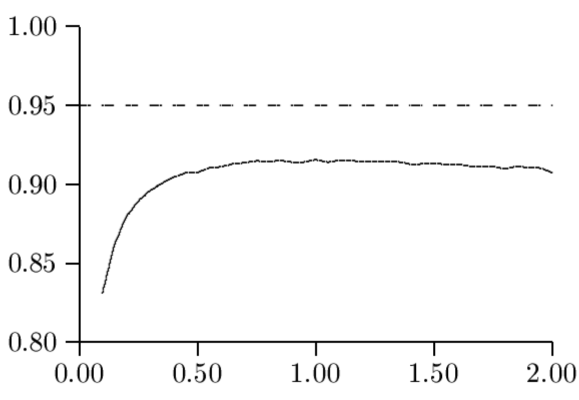

Figure 4: Coverage probability of (5)

3. The standard asymptotic confidence interval requires the knowledge of sample sizes as well as sample proportions in each sample. Unfortunately, it may lead to misunderstandings. Namely, suppose that six experiments were conducted. In each experiment two samples of sizes sixty and seventy, respectively, were drawn ($n_1 = 60$, $n_2 = 70$). The resulting numbers of successes are shown in Table 2 (the first two columns).

Table 2: Confidence intervals in six experiments

n_{A1}	n_{B1}	$\widetilde{OR}$	$left$	$right$
6	14	0.4444	0.1592	1.2410
8	18	0.4444	0.1776	1.1122
15	30	0.4444	0.2095	0.9428
24	42	0.4444	0.2199	0.8985
36	54	0.4444	0.2078	0.9506
48	63	0.4444	0.1627	1.2141

It is seen that the sample odds ratio (the third column) is the same in all experiments, but the confidence intervals are quite different. Moreover, for example, in the first experiment it may be claimed that the population odds in groups A and B may be treated as equal, while in the fourth one such a conclusion should not be drawn.

4. An example of application

The aim of the study was to compare the chances of survival of trading companies in Mazowieckie voivodship versus Warsaw (Poland). The question was about the chances of surviving during the first ten years of activity (Zieliński 2020b).

Let p_A denote the probability of surviving the first ten years of activity of a firm established in Mazowieckie voivodship, and let p_B denote the appropriate probability for a firm established in Warsaw. We are interested in the estimation of the odds ratio, i.e. $(p_A/(1-p_A))/(p_B/(1-p_B))$.

From the REGON (*National Business Registry Number*) registry it is known that 32760 firms started their activity in 2007. Among them 17130 were established in Mazowieckie voivodship, while 15630 were established in Warsaw. Among firms established in 2007 the random sample of size 320 was taken and it was observed how many of those firms were still active in 2017. The data are given in Table 3.

Table 3: Random sample of firms

	Active	Nonactive	
Mazowieckie	96	74	170
Warsaw	85	65	150

On the basis of those data the odds ratio would be estimated.

Note that the estimator of the odds ratio is defined for random variables distributed as binomial. In our investigation we deal with random variables distributed as hypergeometric. It is well known that hypergeometric distribution may be approximated by an appropriate binomial distribution. Some remarks on consequences of such approximation may be found in Zieliński (2011). In what follows, it is assumed that binomial approximation to the hypergeometric one is fairly enough.

The estimate of odds for Mazowieckie voivodship equals $(96/170)/(74/170) = 1.297$. It means that almost 30% more of the firms established in 2007 were still working than were nonactive. A similar indicator for Warsaw equals 1.308.

The estimate of odds ratio for Mazowieckie voivodship versus Warsaw equals $1.292/1.308 = 0.992$. The confidence interval (4) at 95% confidence level is $(0.437, 2.049)$. Since this confidence interval covers 1, it may be expected that for the firms established in 2007 the chances of surviving the first ten years of activity for Mazowieckie voivodship and for Warsaw are similar.

The above conclusion may of course be wrong. It must be stressed that the risk of over- or under-estimation is at most 5%, in contradiction to the standard confidence interval.

Simple calculations show that the standard confidence interval (5) at 95% confidence level for odds ratio is $(0.989, 1.544)$. This confidence interval is narrower than (4), but unfortunately the risk of not covering the true value of the odds ratio is greater than assumed 5% and remains unknown.

Table 4: Number of firms in REGON registry in 2007

	Active	Nonactive	
Mazowieckie	9448	7682	17130
Warsaw	9607	6023	15630

In the presented example we are very lucky since we have full information about the number of firms established in 2007 which survived until 2017. Hence, we may calculate the exact value of odds ratio for that population. Those data are presented in Table 4 (data comes from the REGON registry).

The exact value of odds ratio in that population equals $(9448/7682)/(9607/6023) = 0.771$. Note that the new confidence interval (4) covers this value, while the standard asymptotic confidence interval does not.

5. Conclusions

In this paper a new confidence interval for the odds ratio is proposed. The confidence interval is based on the exact distribution of the sample odds ratio, hence it works for large as well as for small samples. The coverage probability of that confidence interval is at least the nominal confidence level, in contrast to the asymptotic confidence intervals known in the literature. It must be noted that the information on the sample sizes and the sample odds ratio is sufficient for constructing the new confidence interval. Unfortunately, no closed formulae for the ends of the confidence interval are available. However, for given n_A, n_B and observed $\widehat{OR}$ the ends may be easily numerically computed with the aid of the standard software such as R, Mathematica, etc. (see Appendix 2).

Since the proposed confidence interval may be applied for small as well as for large sample sizes, it may be recommended for practical use.

References

Baumol, W. J., (2015). *Macroeconomics: Principles and policy*, Cengage Learning, Inc.

Andrés, M.A., Mato, S.A., Tejedor, H.I., (2020). Pseudo-Bayesian test for the comparison of two proportions, *Metron*, 49 (1-4), pp. 151–162.

García-Pérez M.A, Núñez-Antón V. (2020) Asymptotic versus exact methods in the analysis of contingency tables: Evidence-based practical recommendations, *Stat Methods Med Res.*, 29(9), pp. 2569–2582.

Cornfield, J., (1951). A Method of Estimating Comparative Rates from Clinical Data. Applications to Cancer of the Lung, Breast, and Cervix, *JNCI: Journal of the National Cancer Institute*, 11, pp. 1269–1275, DOI: 10.1093/jnci/11.6.1269.

Edwards, A.W.F., (1963). The Measure of Association in a 2×2 Table. *Journal of the Royal Statistical Society*, Ser. A. 126, pp. 109–114, DOI: 10.2307/2982448.

Encyclopedia of Statistical Sciences, (2006). Wiley & Sons.

García-Pérez M.A., Núñez-Antón V., (2020) Asymptotic versus exact methods in the analysis of contingency tables: Evidence-based practical recommendations. *Stat Methods Med Res.*, 29(9), pp. 2569–2582.

Gart, J.J., (1971). The comparison of proportions: a review of significance tests, confidence intervals, and adjustments for stratification. *Review of the International Statistical Institute*, 39, pp. 148–169.

Lawson, R., (2004). Small Sample Confidence Intervals for the Odds Ratio. *Communications in Statistics - Simulation and Computation*, 33, pp. 1095–1113, DOI: 10.1081/SAC-200040691.

Morris, J.A., Gardner M.J., (1988). Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates. *British Medical Journal*, 296, pp. 1313–6, DOI: 10.1136/bmj.296.6632.1313.

McCullagh, P., (1980). Regression Models for Ordinal Data. *Journal of the Royal Statistical Society*, Ser. B. 42, pp. 109–142.

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, 97, pp. 558–625.

Thomas, D.G., (1971). Algorithm AS-36: exact confidence limits for the odds ratio in a 2×2 table. *Applied Statistics*, 20, pp. 105–110.

Wang, W., Shan G., (2015) Exact Confidence Intervals for the Relative Risk and the Odds Ratio. *Biometrics*, 71, pp. 985-995, DOI: 10.1111/biom.12360.

Zieliński, W., (2011) Comparison of confidence intervals for fraction in finite populations. *Quantitative Methods in Economics*, XII, pp. 177–182.

Zieliński, W., (2020a). A new exact confidence interval for the difference of two binomial proportions. *REVSTAT-Statistical Journal*, 18, pp. 521–530.

Zieliński, W., (2020b). A New Confidence Interval for the Odds Ratio: an Application to the Analysis of the Risk of Survival of an Enterprise. *The 14th Professor Aleksander Zeliaś International Conference on Modelling and Forecasting of Socio-Economic Phenomena*, pp. 185–191.

Appendix 1

A few remarks before the proof.

Remark 1. For $1 \leq n_{A1} \leq n_A - 1$ and $1 \leq n_{B1} \leq n_B - 1$

$$P_r\{n_{A1}, n_{B1}\} \to \begin{cases} 0, & \text{as } r \to 0 \\ 0, & \text{as } r \to +\infty \end{cases}$$

Proof of Remark 1. For $1 \leq n_{A1} \leq n_A - 1$ and $1 \leq n_{B1} \leq n_B - 1$

$$P_{r,p_A}\{n_{A1}, n_{B1}\} \propto p_A^{n_{A1}}(1-p_A)^{n_A - n_{A1}} \cdot \left(\frac{p_A}{p_A + r(1-p_A)}\right)^{n_{B1}} \left(\frac{r(1-p_A)}{p_A + r(1-p_A)}\right)^{n_B - n_{B1}}$$
$$\to \begin{cases} 0, & \text{as } r \to 0 \\ 0, & \text{as } r \to +\infty \end{cases}$$

Hence, $P_r\{n_{A1}, n_{B1}\} \to 0$ as $r \to 0$ or $r \to \infty$.

Remark 2. $P_r\{\widehat{OR} = 0\} \to \begin{cases} \frac{n_A}{n_A+1}, & \text{as } r \to 0 \\ 0, & \text{as } r \to +\infty \end{cases}$

Proof of Remark 2. Note that $\widehat{OR} = 0$ if and only if $(n_{A1} = 0$ and $n_{B1} \geq 1)$ or $(1 \leq n_{A1} \leq n_A - 1$ and $n_{B1} = n_B)$. Hence,

$$\begin{aligned} &P_{r,p_A}\{\widehat{OR} = 0\} \\ &= (1-p_A)^{n_A} \sum_{n_{B1} \geq 1} \binom{n_B}{n_{B1}} p_B^{n_{B1}}(1-p_B)^{n_B - n_{B1}} + p_B^{n_B} \sum_{n_{A1}=1}^{n_A - 1} \binom{n_A}{n_{A1}} p_A^{n_{A1}}(1-p_A)^{n_A - n_{A1}} \\ &= (1-p_A)^{n_A}\left(1 - \left(\frac{r(1-p_A)}{p_A + r(1-p_A)}\right)^{n_B}\right) + \left(\frac{p_A}{p_A + r(1-p_A)}\right)^{n_B}\left(1 - p_A^{n_A} - (1-p_A)^{n_A}\right) \\ &\to \begin{cases} (1-p_A)^{n_A} + \left(1 - p_A^{n_A} - (1-p_A)^{n_A}\right) = 1 - p_A^{n_A}, & \text{as } r \to 0 \\ 0, & \text{as } r \to +\infty \end{cases} \end{aligned}$$

We obtain

$$P_r\{\widehat{OR} = 0\} = \int_0^1 P_{r,p_A}\{\widehat{OR} = 0\} dp_A \to \begin{cases} \frac{n_A}{n_A+1}, & \text{as } r \to 0 \\ 0, & \text{as } r \to +\infty \end{cases}$$

Remark 3. $P_r\{\widehat{OR} = 1\} \to \begin{cases} \frac{1}{n_A+1}, & \text{as } r \to 0 \\ \frac{1}{n_A+1}, & \text{as } r \to +\infty \end{cases}$

Proof of Remark 3. Note that $\widehat{OR}=1$ iff $n_{A1}n_B = n_{B1}n_A$. Hence,

$$\begin{aligned}
&P_{r,p_A}\{\widehat{OR}=1\}\\
&=(1-p_A)^{n_A}(1-p_B)^{n_B}+p_A^{n_A}p_B^{n_B}+\sum_{n_{A1}=1}^{n_A-1}P_{r,p_A}\{n_{A1},n_{B1}\}\\
&=(1-p_A)^{n_A}\left(\frac{r(1-p_A)}{p_A+r(1-p_A)}\right)^{n_B}+p_A^{n_A}\left(\frac{p_A}{p_A+r(1-p_A)}\right)^{n_B}+\sum_{n_{A1}=1}^{n_A-1}P_{r,p_A}\{n_{A1},n_{B1}\}\\
&\to\begin{cases}p_A^{n_A}, & \text{as } r\to 0\\ (1-p_A)^{n_A}, & \text{as } r\to+\infty\end{cases}
\end{aligned}$$

We obtain

$$P_r\{\widehat{OR}=1\}=\int_0^1 P_{r,p_A}\{\widehat{OR}=1\}dp_A\to\begin{cases}\frac{1}{n_A+1}, & \text{as } r\to 0\\ \frac{1}{n_A+1}, & \text{as } r\to+\infty\end{cases}$$

Theorem. For $n_A > \frac{2}{1-\gamma}-1$ the confidence interval for r is two-sided and is one-sided otherwise.

Proof. For $0<t<1$ we have

$$P_r\left\{\widehat{OR}\le t\right\}=P_r\left\{\widehat{OR}=0\right\}+P_r\left\{0<\widehat{OR}\le t\right\}\to\begin{cases}\frac{n_A}{n_A+1}, & \text{as } r\to 0\\ 0, & \text{as } r\to+\infty\end{cases}$$

If $\frac{n_A}{n_A+1}>\frac{1+\gamma}{2}$, i.e. $n_A>\frac{2}{1-\gamma}-1$, the confidence interval is two-sided. Otherwise, the c.i. is one-sided with the left end equal to 0.
For $1\le t<+\infty$ we have

$$P_r\left\{\widehat{OR}\le t\right\}=P_r\left\{\widehat{OR}<1\right\}+P_r\left\{\widehat{OR}=1\right\}+P_r\left\{1<\widehat{OR}<+\infty\right\}\to\begin{cases}1, & \text{as } r\to 0\\ \frac{1}{n_A+1}, & \text{as } r\to+\infty\end{cases}$$

If $\frac{1}{n_A+1}<\frac{1-\gamma}{2}$, i.e. $n_A>\frac{2}{1-\gamma}-1$, the confidence interval is two-sided. Otherwise, the c.i. is one sided with the right end equal to $+\infty$.

Appendix 2

An exemplary R code for calculating the confidence interval for the odds ratio is enclosed.

```
OR=function(n,m){
ifelse(m[1]==0 & m[2]==0,0,
ifelse(m[1]==n[1] & m[2]==n[2],2*(n[1]-1)*(n[2]-1),
ifelse(m[2]==0,2*(n[1]-1)*(n[2]-1),
ifelse(m[1]==n[1],2*(n[1]-1)*(n[2]-1),m[1]*(n[2]-m[2])/(n[1]-m[1])/m[2])
)))}
f=function(rr,k1,k2,pA){dbinom(k1,n[1],pA)*dbinom(k2,n[2],pA/(pA+rr*(1-pA)))}
nieostra=function(rr,tt){
line<-0
prawd=c()
for (k1 in 0:(n[1]-1)){
RS=round(n[2]/(tt*(n[1]/k1-1)+1),2)
Niod=ifelse(k1==0,ifelse(tt<1,1,0),ceiling(RS))
for (k2 in Niod:n[2])
{mrob=c(k1,k2)
line=line+1;
prawd[line]=integrate(f,0,1,rr=rr,k1=k1,k2=k2,subdivisions = 1000L,
  stop.on.error = FALSE)$value;}}
td=sum(prawd)}
ostra=function(rr,tt){
line<-0
prawd=c()
for (k1 in 0:(n[1]-1)){
RS=round(n[2]/(tt*(n[1]/k1-1)+1),2)
Osod=ifelse(k1==0,ifelse(tt<=1,1,0),ifelse(RS==trunc(RS),RS+1,ceiling(RS)))
for (k2 in Osod:n[2])
{mrob=c(k1,k2)
line=line+1;
prawd[line]=integrate(f,0,1,rr=rr,k1=k1,k2=k2,subdivisions = 1000L,
  stop.on.error = FALSE)$value;}}
tg=sum(prawd)}
CI=function(n,m,level){
orobs<-OR(n,m)
eps=1e-4
ifelse(orobs<1,
{ifelse(n[1]<=2/(1-level)-1,
{L=0;
P=uniroot(function(t){ostra(t,orobs)-(1-level)/2}, lower = orobs,
  upper = 2*(n[1]-1)*(n[2]-1),tol = eps)$root},
{L=uniroot(function(t){nieostra(t,orobs)-(1+level)/2}, lower = 0.00000001,
  upper = orobs, tol = eps)$root;
P=uniroot(function(t){ostra(t,orobs)-(1-level)/2}, lower = orobs,
  upper = 2*(n[1]-1)*(n[2]-1), tol = eps)$root})},
{ifelse(n[1]<=2/(1-level)-1,
{L=uniroot(function(t){nieostra(t,orobs)-(1+level)/2}, lower = 0.00000001,
  upper = orobs, tol = eps)$root;
P=Inf},
{L=uniroot(function(t){nieostra(t,orobs)-(1+level)/2}, lower = 0.00000001,
  upper = orobs, tol = eps)$root;
P=uniroot(function(t){ostra(t,orobs)-(1-level)/2}, lower = orobs,
```

```
  upper = 2*(n[1]-1)*(n[2]-1), tol = eps)$root})}
)
print(paste("Confidence interval for odds ratio (",round(L,5),",",round(P,5),")
  at the confidence level ", level,sep=""),quote=FALSE)
print(paste("Sample odds ratio equals ",round(orobs,4), "; n1=",n[1],",
  n2=",n[2],sep=""),quote=FALSE)}

#Example of usage
n=c(60,70) # input n_A and n_B
m=c(7,63) # input n_A1 and n_B1
CI(n,m,level=0.95)
```