Exact correction factor for estimating the OR in the presence of sparse data with a zero cell in 2x2 tables

Background

Many research studies deal with sparse data in a contingency table(1). Contingency tables are referred to as sparse when many cells have small frequencies, or when some of them have empty cells. Sparsity is not restricted to the tables with smaller sample sizes alone, but could also occur with large sample sizes due to high concentration of frequencies in certain cells and small or empty counts in other cells(2). The impact of sparsity is expected to cause computational instability while estimating effect measures such as relative risk (RR) or odds ratio (OR). Collapsing cells in a variable is usually done in r x c contingency tables if there is a natural way of combining them and the loss of information is limited(3). However, in many situations this may not be meaningful or possible.

Empty cell or sparse data can cause bias in estimators of OR in 2x2 tables(4). When zero events occur in either or both groups of a study, the log-odds ratio and their standard errors become undefined. To overcome this problem, a continuity correction factor 'ε' is often added to each cell of the 2×2 table for the studies with zero events in either group(4). The traditional remedy is to add a constant 0.5 to each cell count, usually referred as continuity correction or Yates' correction(5).

However, more recent investigators have raised the question against the use of continuity correction and some of the researchers are not recommending it at all, because the effects of continuity correction on the study findings may be unfavorable(4). Agresti (1998) also suggested adding a very small constant (ε = 2) to all cells of the 2x2 table, when there is sparse data(6). Another suggestion from Subbiah and Srinivasan (2008) is that the nature of sparse data be classified as severe or moderate, and then a suitable correction factor can be added(7).

The availability of many methods but no clear mention of the situation where each of the correction factors can be applied is a concern to researchers as conclusions drawn may be different from each method. Moreover, classifying the problem of sparsity as severe or moderate and then adding a suitable correction factor is a cumbersome procedure.

Lyles et al (2012) have pointed out that Ratio estimators of effect such as relative risk and OR are ordinarily obtained by exponentiating maximum-likelihood estimators (MLEs) of log-linear or logistic regression coefficients. As these estimators can provide positive finite-sample bias, they have proposed a simple correction that removes a substantial portion of the bias due to exponentiation(8).

Bayesian methods have also been proposed to handle sparse data(9–12). Researchers frequently deal with non-informative priors that are subjective. For example, in stomach ulcer data(13,14), by selecting a non-informative prior for the regression coefficient mean as 0 with precision 0.001, the OR becomes 0.003 (95%CI: 0.0, 0.024) or mean as 0 with precision 0.01, the OR becomes 0.011. (0.0, 0.109). These two estimations differ significantly. Similarly, the OR is 0.003(0.0002, 0.024) if the mean is set to 0.5 and the accuracy is 0.001. There is not much difference in the OR and the 95% CI when there is a difference in choosing the mean. However, there is a considerable change in the OR and 95% CI by selecting various precision of non-informative prior values. Similarly, the OR and 95% CI for hyponatremia data(1) are 0.0002 (0.0, 0.0018) when the precision is 0.001, and 0.0007 (0.0, 0.0069) while the precision is 0.01. The same difference is observed when the mean is 0.5 with the precision 0.001 and 0.01 (APPENDIX III).

Therefore, it is evident how to select the precision of non-informative priors. Although the general recommendation is to choose a larger variance (less precision), there is no specific recommendation in this section. Consequently, there is scope for variability in the posterior estimates based on the selection of prior values.

Due to those afore mentioned issues, we propose to develop an algorithm that would be able to estimate exactly an epsilon ‘ε’ (correction factor), because applying different correction factors in the same 2x2 contingency table can lead to different conclusions drawn(7). Therefore, the objective of this study is to present an iterative procedure that could estimate ε where the root mean square error (RMSE) in the estimation of the OR in 2x2 tables is minimal and the coverage probability (CP) is about 95%. Also, we presented a linear regression model that used sample size and proportions to identify the optimal correction factor ‘ε’.

Statistical Methods

Consider a 2x2 contingency table from a case-control study

A	B
C	D

Assume the cell count of A is zero. Without loss of generality, assume the cell count A is zero. The interpretation of odds ratio (OR) can be reversed as appropriate later on. The estimation of odds ratio for the 2x2 table is given by

$OR = \frac{AD}{BC}$ ------------------------------ (1)

and the asymptotic standard error of the natural logarithm of the odds ratio is

$SE(ln\, OR) = \sqrt{\frac{1}{A}+\frac{1}{B}+\frac{1}{C}+\frac{1}{D}}$ ---------------------------- (2)

If any of the cell counts is zero, then, in order to estimate the OR, a correction factor ‘ε’ is added to all the cells. That is,

$$OR_{\varepsilon} = \frac{(A+\varepsilon)(D+\varepsilon)}{(B+\varepsilon)(C+\varepsilon)} \quad \text{-------------------------------- (3)}$$

Thus, the asymptotic standard error (ASE) for $ln(OR_{\varepsilon})$ is

$$SE(ln\, OR_{\varepsilon}) = \sqrt{\frac{1}{A+\varepsilon} + \frac{1}{B+\varepsilon} + \frac{1}{C+\varepsilon} + \frac{1}{D+\varepsilon}} \quad \text{----- (4)}$$

As a result, the asymptotic confidence interval *100(1-α)%* for OR estimate from exponentiating the following equation:

$$(ln\, OR_{\varepsilon}) \pm Z\alpha/_2 SE(ln\, OR_{\varepsilon}) \quad \text{------------------------- (5)}$$

Bootstrap SE in case of small sizes

In order to estimate the SE in case of small numbers such as 10, 15 etc. and to validate the correction factor that is chosen based on small number with asymptotic SE we have used Bootstrap method of estimating SE. The scenarios are with which the bootstrap method was done using sample size (n=10,15 and 20) and epsilon (e= 0.1,0.2,0.3,0.4, 0.5 and 2).

Evaluation of optimal correction factor (ε) using simulation data

In order to identify the optimal correction factor, the commonly used statistics are bias, or Root Mean Square Error (RMSE). Walter and Cook (1991) used bias, AAE and RMSE to compare three different estimators for OR(15).

On the basis of simulation data, the optimal correction factor was evaluated with known OR, and varying values of proportion and sample sizes based on case-control study design with ratio 1:1. Sparse data were simulated with various proportions such as p_1 (0.001, 0.002, 0.003, 0.004, 0.005, 0.006, 0.007, 0.008, 0.009 and 0.010) , p_2 (0.30, 0.35, 0.40, 0.45, 0.50, 0.55 and 0.60) with varying sample sizes (20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300 and 500) and various correction factors (ε) such as 0.005, 0.01, 0.03, 0.06, 0.08, 0.1, 0.2, 0.3, 0.4, 0.5 (Yate's) and 2 (Agresti and Coull) were used. Throughout our simulation study we assume that the column wise marginal distributions are independent binomial distributions. The column totals n_1 and n_2 are fixed and the success probability P_1 is assumed to be too small so that first cell count is either zero or small. P_2 is assumed to be large. This kind of simulation study was carried out by Walter (1975) in the paper titled "The distribution of Levin's measure of attributable risk"(16).

The cell counts were estimated using the following equation based on the above mentioned parameters:

$$Actual\ OR = \frac{p_1 * (1-p_2)}{p_2 * (1-p_1)} \qquad \text{------------------------------ (6)}$$

The cell value A = $p_1 * n_1$; B = $(1-p_1) * n_1$; C=$p_2 * n_2$ and D= $(1-p_2) * n_2$.

Following that, the adjusted OR is estimated using each correction factors based on the equation 3.

Estimation of Bias

The difference from the actual OR without ε that was used for simulation (assumed to be a parameter) to the adjusted OR with ε is the bias, i.e.

$$Bias(ln\ OR_\varepsilon) = ln\ (\text{Actual OR}) - ln\ (\text{OR}_\varepsilon).$$

For each simulated sparse data in the traditional 2x2 table, ε was added to each cell, and then ASE and Bias were estimated.

The optimal correction factor was chosen based on bias and variance trade-off principle, either graphically or based on the root mean squared error (RMSE). The RMSE is a fit statistic, which is the summary of the trade-off between bias and variance. i.e.

$$RMSE = \sqrt{Bias(ln\ OR_\varepsilon)^2 + SE(ln\ OR_\varepsilon)^2} \qquad \text{-------------------------- (7)}$$

We determined the optimal correction factor (ε), when the SE and bias values are "crossing each other".

Coverage Probability (CP)

Coverage Probability was calculated using Monte Carlo simulation method, for every ε, that is ranging from 0.005 to 2. ε equal to 2 is nothing but the Agresti & Coull correction factor. For the simulation data, the CP was calculated using nominal method for the correction factors that were identified as the optimal based on the sample sizes. Similarly, parametric bootstrap method was used to compute 95% confidence interval(CI) and CP for real data and simulated data respectively(17). However, the Agresti and Coull method ε was not included in the real data evaluation. Based on the concept of 95% CI, the coverage probability for a parameter has to be ideally 95%. Conservatively, any CP lower or greater than 95±1 is not considered as good(17).

Fitting a model for finding optimal correction factor (ε)

The simulation was done to find optimal correction factor which ranges from 0.1 to 0.5, and 2. Moreover, the range of value from p1 and p2 are from 0.001 to 0.01 and 0.30 to 0.60 used in the simulation. Also, that, the sample size could be anywhere between 10 and 500. However, it may be impossible to identify the combination of these three and best epsilon. Therefore, the motivation is to establish the regression equation based on these parameters to estimate the optimal epsilon.

The optimal correction factor (ε) was determined using simulated data with low RMSE and when both bias and ASE intersected. This was simplified using a function of the parameters such as sample size (N) and proportion of two groups (p_1 and p_2). The natural logarithm of the mentioned parameters was used to fit this in linear regression equation as:

$$ln\ (\varepsilon) = \alpha + \beta_1 * ln\ N + \beta_2 * ln\ p_1 + \beta_3 * ln\ p_2$$

After estimating the regression coefficient, the optimal correction factor can be estimated for a given N, P_1, and P_2 using the following equation,

$$\varepsilon = exp(\widehat{\alpha}) * N^{\widehat{\beta_1}} * \mathrm{P}_1^{\widehat{\beta_2}} * \mathrm{P}_2^{\widehat{\beta_3}} \text{-------------------------} (8)$$

Other methods of estimating OR with minimum bias:

Lyles Method:

Consider a generalized linear model of the form

$$g[E(Y|X = x)] = \beta_0 + \sum_{j=1}^{k} \beta_j x_j$$

Where g(.) is a strictly increasing link function. Maximum likelihood estimator is

$$\widehat{\beta_j} = N\left(\beta_j, \sigma^2{}_j\right)$$

$\sigma^2{}_j$ is asymptotic variance of $\widehat{\beta_j}$. Thus, $\widehat{\beta_j}$ is asymptotic median unbiased. The link function g is logit or logarithm, and the target parameter is $OR = e^{\beta_j}$ with estimator $\widehat{OR} = e^{\widehat{\beta_j}}$ $(j = 1,2,3, \ldots . k)$. The first order limiting distribution of $\widehat{OR}$ is also normal with mean OR; the log-scale normal approximation is far more accurate for typical sample size. Then $\widehat{OR}$ will be more closely log normal with mean

$$E\left(\widehat{OR}\right) = e^{\beta_j + \left(\frac{\sigma^2{}_j}{2}\right)} = e^{\beta_j} * e^{\left(\frac{\sigma^2{}_j}{2}\right)} = OR * e^{\left(\frac{\sigma^2{}_j}{2}\right)}$$

However, $\widehat{OR}$ is subject to large overestimation error unless $\sigma^2{}_j$ is small. The bias factor $e^{\left(\frac{\sigma^2{}_j}{2}\right)}$ is negligible for small but increase rapidly with σ_j. However, the approximate expectation of $\widehat{OR}$ overestimates OR by more than 50% for $\sigma^2{}_j \geq 0.9$. In order to reduce the bias, the following estimator of $\widehat{OR}$ was suggested by Lyles et al(8).

$$\widehat{\mathrm{OR}}_{\mathrm{BR}} = e^{\left(\frac{-\sigma^{2}{}_{j}}{2}\right)} * \widehat{\mathrm{OR}}\,. \qquad \text{---------} \ (9)$$

Where, $\widehat{\sigma^{2}}_{J}$ is variance of $\widehat{\beta}_{J}$

Datasets

***Hyponatremia* Data**

The original data was obtained from a case-control study conducted at the Christian Medical College in Vellore, Tamil Nadu, India, to find the association between hyponatremia and hiccups. This dataset consists of 50 subjects with hiccups (cases) and 50 subjects without hiccups (controls). The hiccups groups were categorized according to the severity of the disease so that among 50 subjects with hiccups, 23 subjects had mild, 12 subjects had moderate, and 15 subjects had a severe kind of hiccups. For the illustrative purpose, we considered the cases that had severe hiccups and 50 controls (1).

Tuberculomas Data

The data was obtained from a retrospective study conducted in a neurosurgical unit of a tertiary-care center from January 2000 to December 2015.In this study, they attempted to find the differences in the characteristics of the tuberculomas in the standard (ATT ≤ 2 years) and prolonged therapy (ATT >2 years) groups. Of 67 patients who received standard therapy, 24 had the tuberculoma in infratentorial site and 43 had in Supratentorial/þinfratentorial. Similarly, of 19 patients who received prolonged therapy, all of them had the tuberculoma in Supratentorial/þinfratentorial(18).

Results

Optimal correction factor (ε)

Table 1 shows the results of asymptotic standard error (ASE), bias and corresponding RMSE value of the simulated data for various values of ε and sample sizes respectively.

When the sample size is 20, the ASE and the bias cross each other at the epsilon of 0.2 with RMSE of 3.94. The epsilon of 0.2 is similar for the sample sizes from 30 to 80 with lower RMSE. The similar findings are presented pictorially in figure 1 with the intersection of ASE and bias values which determined the optimal correction factor.

If the total sample size of a study is 40, we need to look at the RMSE row in Table 1 and find the minimal RMSE. The minimal RMSE in this situation is 3.336. Then check above the row of epsilon and find the one that equals 3.336, that is 0.2 in this case. This should be considered as the optimal correction factor and its corresponding coverage probability is 97.9%. We chose 0.2 as the optimal correction factor since RMSE is the initial selection criteria. Despite the fact that the coverage probability for 0.3 is about 0.96 better than the CP for epsilon 0.2, we still suggest RMSE-based selection. Similarly, for sample sizes of 90 to 200, the optimal epsilon was indicated as 0.3 by the lowest RMSE. The epsilon is 0.4 when the sample size is 300 and 0.5 when the sample size is 500 or more. In many situations with epsilon 0.2 and various sample sizes, the coverage probability is higher than the nominal threshold of 95% (by 2%). Appendix 1 provides exact correct factor for varying proportions and varying sample size.

Coverage Probability

Optimal correction factors for each sample size with fixed cell probabilities ranging from 0.001 to 0.4 were obtained in Table 1. Based on equation 5, the nominal method's coverage probability (CP) was calculated and presented in Table 2. The parametric bootstrap technique was used to calculate the coverage probability for the optimal correction factor (epsilon) which was taken from Table 1. When the sample size is 20 and the epsilon is 0.2, then the CP of nominal method is 0.985. This was 0.980 using parametric bootstrap method. The CP of the nominal method was 0.984, 0.979, and 0.975 when the sample size was 30, 40 and 50 respectively, with epsilon value 0.2. However, the CP of the parametric bootstrap method was

0.870, 0.900 and 0.940 for those situations. When the sample size was between 60 to100, the CP of the nominal method was around 0.95 at corresponding epsilon values of 0.2 and 0.3. Similarly, when the sample size was 200 and 300, the CP of the nominal method was about 0.907 and 0.864 at epsilon 0.3 and 0.4, respectively.

Comparison of Bootstrap and Asymptotic SE in identifying epsilon

Table 3 presents the parametric Bootstrap and Asymptotic SE, with various values of epsilon and sample sizes. The SEs based on asymptotic method overestimated the SE about 3 to 4 times more as compared to Bootstrap methods when the sample size is <=100. When the sample size was 10 and the epsilon was 0.2 then the ratio between Asymptotic SE vs Bootstrap method was 2.7532. As the epsilon increased the ratio decreased for a given sample size. Also, that as the sample size increased above 1000 the ratio started approaching 1. Especially when the sample size is 1500 or more the ratio is about 1. Thus, the SEs become similar as sample size increased over 1500. Therefore, in order to facilitate the readers to estimate correct SE using Bootstrap method for small sizes, we have provided the Bootstrap code as well.

Fitting a Regression model to find the optimal correction factor (ε)

If a researcher wants to find out the optimal epsilon for their data, they have to substitute the values of sample size, proportions p_1 and p_2 in equation 8. This will provide approximately the optimal correction factor for their data.

Based on equation (8), the estimated values of the parameters such as $\alpha, \beta_1, \beta_2, \beta_3$ are -0.399, 0.333, 0.350, and 0.034 respectively.

For example, N=20, p_1=0.001, p_2 = 0.40, then

$\varepsilon = exp\ (-0.399) * 20^{0.333} * 0.001^{0.350} * 0.40^{0.034}$

$\varepsilon = 0.157 \sim 0.2.$

For various combinations of p_1, p_2 and sample size the epsilon is presented in Appendix II. For example, in appendix II, if p_1=0.001, p_2=0.30 and with sample size 100, the epsilon based on equation (8) is 0.27 while this is 0.3 by simulation with asymptotic method. Similarly, if p_1=0.005, p_2=0.40 and with sample size 40 the best epsilon value from the suggested model is 0.35 and from the simulation with asymptotic method is 0.30. Based on the above two example, the impact in the quality is 0.03 and 0.05 units. Thus, the regression is as good as simulation with asymptotic method.

Other methods of estimating OR with minimum bias (Lyles Method):

Table 4 presents the mean bias and RMSE based on the Regression Method and simulation with asymptotic method for p_1=0.001 and various values for p_2 and various sample sizes. In order to compare the two methods, we have used RMSE statistics. When the sample size is small (n=20) and the p_2=.3 with p_1=0.001, the asymptotic method provided epsilon as 0.2 while the Lyles method provided this as 0.3. When the sample size is 40 for the same proportions both the methods provided the epsilon was 0.3. However, when the sample size is 100 then the epsilon was 0.5 and 0.4 for Lyles method and asymptotic method respectively. The same trend was observed when the p_2 was 0.4, or 0.5 with various sample sizes. When the sample size was 500 then both methods suggested epsilon as 0.5. The asymptotic method provided epsilon which is about 0.1 units lower than the Lyles method. Throughout, Lyles method gives very low RMSE. As sample size increases RMSE of both methods were close to each other.

Real time data:

The distribution and results of real time sparse data were presented in Table 5. In hyponaetremia data with sample size of 65, we have compared severe hiccups with the subjects who did not have hiccups (Control). According to Table 1 with sample size 70, and the RMSE row, the minimum RMSE was 2.908 and the corresponding epsilon was 0.2. Thus, the optimal correction factor for this study is 0.2 with coverage probability 96.6% and this was evaluated using Monte Carlo simulation method. The rows were interchanged in order to get the zero value into 'A' cell. After that, OR was computed and then inverted into the original form of the OR. Based on Yates correction, the inverted OR was 320 with 95 % CI: 15, 6687, when the epsilon was 0.5. According to this study's suggestion, the optimal epsilon is 0.2, with an OR and 95% confidence interval is 836 (9.1, 77013.9). Based on the bootstrap method, the CI for corresponding OR is 325.7, 3720.2. This approach is much narrower than the 95% CI based on normal method. Likewise, for the tuberculomas study, the optimal epsilon is 0.2 that provided the OR 53.8 (0.6, 4529.4) and 95% Bootstrap CI (26.2, 99.04).

Discussion

When dealing with binary type response variables, the existence of zero cell observations is unavoidable for many medical researchers. It arises most often when the sample size is small or when there is strong association hyponatremia and severe hiccups study(1). This paper investigated the optimal correction factor for estimating the OR, minimizing the RMSE, and maintaining the coverage probability. Agresti and Coull's (1998) suggestion of adding a count in all cells of the 2x2 table provides a 95% CI that is closer to the nominal confidence interval. However, this study, as well as Subbiah and Srinivasan's (2008) paper highlighted the fact that the narrow CI does not have higher coverage, but rather a lower coverage probability. Subbiah and Srinivasan's (2008) suggestion of finding the type of sparsity and adding an appropriate correction is cumbersome. Despite the fact that they have provided the algorithms for identifying the sparsity and correction factor, it is still rather not easy and practical.

In a 2x2 table, the method for determining the optimal correction factor is dependent on the total sample size. For example, when the sample size is 20, the Yates correction (epsilon=0.5) provides us a coverage probability of 16.5%, while the optimal epsilon is 0.2, provides us a coverage probability of 98.5%, which is closer to the desired value of 95%. The epsilon remains the same up to sample size 80; i.e. 0.2. However, the coverage probability improves as the sample size increases. When the sample size is 300 or more, then the CP decreases very much, while the RMSE is smaller. However, the bootstrap CIs are closer to the 95%, suggesting that the correction factors that we have suggested are robust. According to Agresti and Coull (1998); Sweeting (2004), we have also used CP as the main indicator of validation. Our Bootstrap evaluation also suggested the same using the identified optimal correction factor. Furthermore, we validated the optimal correction factor using RMSE, which is a function of bias and SE, while previous papers have used only the CP method.

Puhr et al (2017) have reported that "Firth's penalization is equivalent to maximum likelihood estimation after adding a constant of 0.5 to each cell"(19). Thus, the proposed optimal correction factor provided a narrow length of the CI, as suggested by the bootstrap method compared to Yates correction (or) Penalized logistic regression method. That is, based on the Yates correction of 0.5, the OR and 95% CI was 320.3 (16.3, 6292.9), while this was 836 (325.7, 3720.2) based on the proposed epsilon (0.2) and bootstrap CI. As a result, the proposed epsilon's CI, the length of the CI is much narrower, while the CP is also closer to 95%. We are not in a position to compare the proposed method with Agresti and Coull's (1998) method, since their epsilon is 2, and CP is around 0%, suggesting that the 95%CI is totally away from the actual parameter (OR) that come out of the data. This was pointed out by Subbiah and Srinivasan (2008). That is, the higher the value of epsilon, then the narrower the CI. However, the CP may be very low and nowhere close to the desired CP. As a limitation, since the epsilon is a fraction, traditional 2x2 calculators may not be able to accept the correction factor in the 2x2 cell, as these are assumed to be count data. We need to do this calculation manually. However, we have provided the R codes in the Appendix. Though there are good amount of work has been going on using Bayesian method we would limit our scope to frequentist method as most of the researchers are still ignorant about Bayesian methods. Also, that in the absence of prior or using non-informative prior provides varying epsilon. We would like to demonstrate the use of Bayesian method for sparse data in future communication.

In addition, this study found that the epsilon increases as the sample size increases. However, we also expected that the epsilon has to decrease as the sample size increase. One probable reason could be that the epsilon is not like MSE or variance of an estimator that needs to depend on sample size. In the literatures, the suggestion to handle sparse data is adding a constant 0.5 (Yates correction) or 2 (Agresti and Coull) have been recommended irrespective of the sample size. If severe sparsity is present, then this will be the case even when we have a moderately large sample size. For example, if 0.001 is the probability of the first cell in the 2x2 contingency table, then we cannot expect a single value in the first cell in most of the realizations. However, the reasonable question that we could ask is whether epsilon decreases as the probability of first cell becomes large. This needs to be investigated further.

Lyles et al (2012) method is easy to use as compared to simulation with asymptotic method. Based on our extensive simulation with various levels of p_1 and p_2 and sample size we have found out that Lyles method suggested the epsilon which is 0.1 unit more that the simulation with asymptotic method. However, when the sample size is about 500, both the method provided the same correction factor. Therefore, still we recommend the simulation with asymptotic method as we have provided correction factors for various values of p_1, p_2 and sample size in Appendix II. Limitation of the Lyles method is that if any one of the cell values is zero then we may not be able to use Lyles method. In this situation, someone uses regression method to find correction factor, he/she could use for examples p_1=0.0001 when the cell value A is zero (p_1=0).

In conclusion, we illustrated a proposed method using simulated and real time data and recommended a new method to find optimal correction factor for the sparse data in a conventional 2X2 table model based on sample size and proportions. The optimal correction factor does not change whether we use Bootstrap or Asymptotic SE when the total size of a table is less than or equal to 40. In addition, we have also presented a regression model to identify optimal correction factor using above mentioned parameters which is an easier method for medical researchers. However, we recommend that regression equation can be used to find out the best epsilon, and then OR can be calculated using either Lyles method or simulation with asymptotic method with the help of epsilon.

References

1. George J, Thomas K, Jeyaseelan L, Peter JV, Cherian AM. Hyponatraemia and hiccups. Natl Med J India. 1996 Jun;9(3):107–9.

2. Sangeetha U, Subbiah M, Srinivasan MR. Estimation of confidence intervals for Multinomial proportions of sparse contingency tables using Bayesian methods. 2013;3(4):7.

3. Agresti A. Introduction to Categorical Data Analysis. :394.

4. J. Sweeting M, J. Sutton A, C. Lambert P. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. Statist Med. 2004 May 15;23(9):1351–75.

5. Yates F. Contingency Tables Involving Small Numbers and the χ 2 Test. Supplement to the Journal of the Royal Statistical Society. 1934;1(2):217.

6. Haviland MG. Yates's correction for continuity and the analysis of 2 × 2 contingency tables. Statist Med. 1990 Apr;9(4):363–7.

7. Subbiah M, Srinivasan MR. Classification of 2×2 sparse data sets with zero cells. Statistics & Probability Letters. 2008 Dec;78(18):3212–5.

8. Lyles RH, Guo Y, Greenland S. Reducing Bias and Mean Squared Error Associated With Regression-Based Odds Ratio Estimators. J Stat Plan Inference. 2012 Dec 1;142(12):3235–41.

9. Agresti A, Hitchcock DB. Bayesian inference for categorical data analysis. JISS. 2005 Dec;14(3):297–330.

10. Greenland S. Bayesian perspectives for epidemiological research. II. Regression analysis. International Journal of Epidemiology. 2007 Feb;36(1):195–202.

11. Galindo-Garre F, Vermunt JK, Ato-García M. BAYESIAN APPROACHES TO THE PROBLEM OF SPARSE TABLES IN LOG- LINEAR MODELING. :16.

12. Greenland S, Schwartzbaum JA, Finkle WD. Problems due to Small Samples and Sparse Data in Conditional Logistic Regression Analysis. American Journal of Epidemiology. 2000 Mar 1;151(5):531–9.

13. Efron B. Empirical Bayes Methods for Combining Likelihoods. Journal of the American Statistical Association. 1996 Jun 1;91(434):538–50.

14. Xie M, Singh K, Strawderman WE. Confidence Distributions and a Unifying Framework for Meta-Analysis. Journal of the American Statistical Association. 2011 Mar 1;106(493):320–33.

15. Walter SD, Cook RJ. A Comparison of Several Point Estimators of the Odds Ratio in a Single 2 X 2 Contingency Table. Biometrics. 1991 Sep;47(3):795.

16. WALTER SD. The distribution of Levin’s measure of attributable risk. Biometrika. 1975 Aug 1;62(2):371–2.

17. Efron B, Tibshirani RJ. An Introduction to the Bootstrap [Internet]. Boston, MA: Springer US; 1993 [cited 2021 Apr 19]. Available from: http://link.springer.com/10.1007/978-1-4899-4541-9

18. Nair BR, Rajshekhar V. Factors Predicting the Need for Prolonged (>24 Months) Antituberculous Treatment in Patients with Brain Tuberculomas. World Neurosurg. 2019 May;125:e236–47.

19. Puhr R, Heinze G, Nold M, Lusa L, Geroldinger A. Firth’s logistic regression with rare events: accurate effect estimates and predictions?: R. PUHR *ET AL.* Statist Med [Internet]. 2017 [cited 2021 Jan 25]; Available from: http://doi.wiley.com/10.1002/sim.7273