Songklanakarin J. Sci. Technol.
35 (2), 235-241, Mar. - Apr. 2013

SONGKLANAKARIN JOURNAL OF SCIENCE AND TECHNOLOGY
http://www.sjst.psu.ac.th

Original Article

Democratic Confidence Intervals for comparing two Proportions

Phattrawan Tongkumchum[1,2]*, Rattikan Saelim[1,2], Nifatamah Makaje[1,2], and Aniruth Phon-On[1,2]

[1] *Department of Mathematics and Computer Science; Faculty of Science and Technology,*
Prince of Songkla University, Mueang, Pattani, 94000 Thailand.

[2] *Centre of Excellence in Mathematics CHE, Si Ayutthaya RD., Bangkok, 10400 Thailand.*

Received 12 September 2012; Accepted 18 February 2013

Abstract

We offer methods to create an accurate graph for comparing confidence intervals for two proportions. The first method involves using weighted sum contrasts when fitting logistic regression with a binary explanatory variable. Since it is complicated to estimate standard errors of the logit function, the second method is thus based on applying a delta method to the logit function of the proportions to get their individual estimated standard errors and shrink them to get their comparison standard errors. The two methods give an accurate approximation of confidence intervals for comparing proportions that is consistent with the p-values near 0.05 provided by logistic regression.

Keywords: weighted sum contrasts, delta method, unbalanced design, proportion

1. Introduction

When comparing population parameters based on estimates from samples, it is important to have a measure of the accuracy of the estimate. Graphing confidence intervals is recommended to present the results. We offer methods to create comparing confidence intervals for two proportions. The rationale for these methods is that confidence intervals for comparisons are often misinterpreted. For example, suppose we wish to compare the mean blood lead levels between boys and girls at a school in a village where residents have been exposed to contamination from previous mining activity (Geater *et al.*, 2000). Each panel of Figure 1 shows means and 95% confidence intervals. The confidence intervals in the left panel are individual confidence intervals, meaning that each interval contains the mean of the corresponding population with 95% certainty. For the population of boys, it is thus 95% probable that the mean is between 16.0 and 19.2 mg/dl, whereas the corresponding range for the girls is between 12.9 and 16.8 mg/dl.

Now suppose we ask the question “do boys and girls have different means?” It would appear from the plot in the left panel that there is insufficient evidence, because the confidence intervals overlap. However, if we use a t-test to compare the population means, the p-value is 0.03, indicating (using the conventional rule that a p-value below 0.05 indicates a difference) that the mean for the boys is higher than that for the girls. This apparent anomaly can lead to a mistaken conclusion. The non-agreement between individual confidence intervals and p-values has been raised and

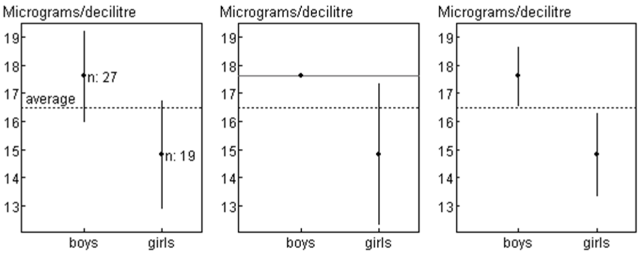

Figure 1. Confidence intervals for blood lead levels by gender: (a) individual (left panel), (b) based on treatment contrasts (centre panel), (c) based on weighted sum contrasts (right panel).

* Corresponding author.
Email address: tphattra@bunga.pn.psu.ac.th

discussed in the literature (Schenker and Gentleman, 2001; Austin and Hux, 2002; Ryan and Leadbetter, 2002; Wolfe and Hanley, 2002; Payton *et al.*, 2003; Cumming and Finch, 2005; Wheeler *et al.*, 2006; Cumming *et al.*, 2007; Cumming, 2009; Knol and Pestman, 2011).

The explanation is that the confidence intervals shown in the left panel are separate intervals for each sex, whereas what is needed is a confidence interval for comparing the means of the two sexes, as shown in middle panel of Figure 1. This interval, centered at the mean for the girls, gives the range for the difference between the two population means (the mean for girls minus the mean for boys), i.e., (–5.3 to –0.3). Adding the mean for the boys (17.6), this interval is plotted as (12.3–17.3), and is entirely below the mean for the boys, so we can conclude that the mean for the girls is lower than that for the boys, in agreement with the p-value.

The right panel plots comparison confidence intervals. These use weighted sum contrasts to give confidence intervals for each group that overlap if and only if the p-value based on the two sample t-test exceeds 0.05 (Tongkumchum and McNeil, 2009). In this plot, the confidence interval for each sex compares its mean with the overall mean. We call these comparison confidence intervals *democratic* because they are applied equitably to each group, whereas the interval for the difference is just one confidence interval measured from a reference group that is taken to be fixed and thus does not have a confidence interval. For two groups, the lengths of the democratic confidence intervals are each shorter than the corresponding individual confidence intervals by the "shrinkage" factors $\sqrt{1-r_1}$ (for sample 1) and $\sqrt{1-r_2}$ (sample 2), where $r_1 = n_1/(n_1+n_2)$, $r_2 = n_2/(n_1+n_2)$. In this case the respective shrinkage factors for boys and girls are $\sqrt{1-27/(27+19)} = 0.643$ and $\sqrt{1-19/(27+19)} = 0.766$.

We can compare two proportions using a similar method. But in this case there is no exact theory, only asymptotic theory for large sample sizes. Logistic regression gives a p-value and a corresponding standard error for the logarithm of the odds ratio that can be modified for comparing the two proportions. If, for example, we are comparing two treatments (A and B) for a disease, where 10 of 21 patients on treatment A improved, compared with only 3 of 19 patients on treatment B, the p-value is 0.0393. The graph in the left panel of Figure 2 shows the analogous plot to Figure 1(b), using treatment contrasts based on logistic regression with B as a reference group. This interval is entirely above the success proportion for patients on treatment B, so we can conclude that the proportion of improvement for patients on treatment A is higher than that on treatment B. The middle panel shows that the conclusion is still the same when using treatment A as a reference group.

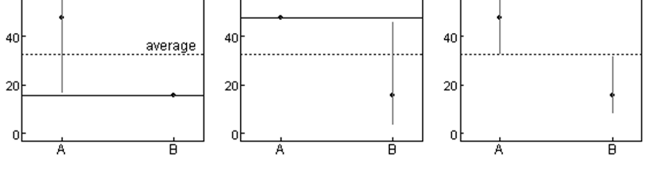

Figure 2. Confidence intervals for comparing two proportions: (a) B as reference (left panel), (b) A as reference (middle panel), (c) using weighted sum contrasts.

The right panel shows the analogous plot to Figure 1(c), using the method described in Tongkumchum and McNeil (2009) and Kongchouy and Sampantarak (2010). The democratic confidence intervals correctly show that the proportions are evidently different.

This study aims to compare the democratic confidence intervals using the weighted sum contrasts method with those using the more exact delta method. The delta method has already been used to get standard errors for individual proportions (e.g. Oehlert, 1992; Papke and Wooldridge, 2005; Xu and Long, 2005). The idea is applying a delta method to the logit transformation function of the proportions to get their individual standard errors and then shrinking them to get comparison standard errors. The validation of the methods will be assessed in terms of its agreement with the p-value given by logistic regression.

2. Method

The method involves the process of estimating standard errors for comparison of two proportions in a 2 by 2 contingency table. The notations used are described as follows. For j = 1 or 2, let $p_j = s_j/n_j$ denotes the proportion of adverse outcomes and $r_j = n_j/n$ denotes the proportion of cases in category j, where the number of successes is s_j, the sample size is n_j, $n = \sum_{j=1}^{2} n_j$ and the observed overall proportion $p = \sum_{j=1}^{2} s_j /n$. The logit of a proportion p_j takes the form $f(p_j) = \ln(p_j/(1\text{-}p_j))$. This function will transform the data range from (0, 1) to $(-\infty,\infty)$.

2.1 The weighted sum contrasts method

The proportions of adverse outcomes and their corresponding standard errors can be estimated by fitting a logistic regression model, and it is appropriate to use weighted sum contrasts to obtain the standard errors underlying for comparing two proportions. Logistic regression provides a straightforward method for estimating a proportion that varies with a determinant of interest (Hosmer and Lemeshow, 2000; Kleinbaum and Klein, 2002). The sum contrasts (Venables and Rilpey, 2002) are available in commonly used software packages such as R (R Development Core Team, 2009) but the weighted sum contrasts need a specific contrast matrix (Tongkumchum and McNeil, 2009).

Suppose that x is a binary factor used as an explanatory variable in a logistic regression model being fitted to two grouped data. The equation expressing one of the two contrasts in terms of the individual logit of proportions takes the form $\alpha^* = D_1\alpha$, where α is the column vector containing the two classes of logit of proportions. Solving these equations gives $\alpha = C_1\alpha^*$ where C_1 is the inverse of the matrix D_1. We omit the first column of C_1 to obtain the desired contrast matrix C, which is then specified when fitting the logistic regression model.

Let $f(p)$ denote the logit of overall proportion p. The equations we use are as follows.

$$\alpha_1^* = r_1 f(p_1) + r_2 f(p_2) \neq f(p) \qquad (1)$$

$$\alpha_2^* = f(p_2) - f(p) \qquad (2)$$

The matrix D_1 comprises Equation 1 and 2. The matrix C then takes the form $\begin{bmatrix} 1 \\ -r_1/r_2 \end{bmatrix}$ for group 1 and $\begin{bmatrix} 1 \\ -r_2/r_1 \end{bmatrix}$ for group 2. The standard errors that result when a logistic regression model is fitted using C as the contrast matrix are used to obtain confidence intervals for the logit proportions used in the contrasts after adjusted for intercept bias. Finally, we obtain the confidence interval for the omitted group by repeating the procedure with this group included and another omitted.

This enables us to construct a graph showing confidence intervals for each of the two proportions being compared, by transforming the confidence intervals for the logits back to confidence intervals for the proportions, using the inverse of the logit function, it follows that $p_j = 1/(1+\exp(-f(p_j)))$ for group j. The simplest logistic model with the binary factor takes the additive form $\ln(p_j/(1+p_j)) = a+b_j$, where a and b_j are coefficients from the model and the proportion itself is thus expressed as $p_j = 1/(1+\exp(-a-b_j))$, for $j = 1, 2$. The confidence intervals for comparing proportions of group j are obtained from $1/(1+\exp(-\{(a^*+b_j \pm 1.96 \times \mathrm{SE}(b_j))\}))$, where a^* is $f(p)$ and $\mathrm{SE}(b_j)$ is the standard error of b_j. The constant a is replaced by a^* to adjusted for bias due to logit of the overall proportion is not the same as the mean of the logit p_1 and logit p_2.

2.2 The delta method

The concept of the delta method is that it takes a function of a random variable for which the variance is not analytically computable, creates a linear approximation of that function and then computes the variance of the simpler linear function that can be used for large sample inference.

To estimate the variance of $f(\hat{\theta})$ using the delta method, let $\hat{\theta}$ be a random variable with mean θ and variance σ^2/n, and f be a smooth function. Then using first-order Taylor series expansions, we get $f(\hat{\theta})$ approximately equal to $f(\theta) + (\hat{\theta} - \theta) f'(\theta)$. The mean and variance of $f(\hat{\theta})$ are approximately equal to $f(\theta)$ and $f'(\theta)^2\sigma^2/n$, respectively. The approximate standard deviation of $f(\hat{\theta})$ is thus $f'(\theta)\sigma/\sqrt{n}$.

We now apply the delta method to the logit function $f(p_j) = \ln(p_j/(1-p_j))$ of a proportion p_j in a sample of size n_j, which has mean π_j and standard deviation $\sqrt{\pi_j(1-\pi_j)/n_j}$. Since the derivative of $\ln(\pi_j/(1-\pi_j))$ with respect to π_j is $1/(\pi_j/(1-\pi_j))$ to the first approximation in the Taylor series, the standard deviation is $1/\sqrt{\pi_j(1-\pi_j)n}$. In general the proportion parameter π_j can be estimated by the proportion estimator p_j. Therefore the estimated standard deviation of $f(p_j)$ is $1/\sqrt{p_j(1-p_j)n_j}$. In order to get the standard error of the difference between the logit function $f(p_j)$ of the individual proportion and the logit function $f(p)$ of the average proportion we apply the shrinkage factors $\sqrt{1-r_j}$ as in comparing the means (Tongkumchum and McNeil, 2009). We hence obtain the standard error $\sqrt{(1-r_j)/p_j(1-p_j)n_j}$ of $f(p_j)-f(p)$.

This enables us to construct a graph showing confidence intervals for each of the two proportions being compared, by transforming the confidence intervals for the logits back to confidence intervals for the proportions, using the inverse of the logit function. The confidence intervals for comparing proportions are obtained from $1/(1+\exp(-\{f(p_j) \pm 1.96 \times \mathrm{SE}\}))$, where SE is the standard error of $f(p_j)-f(p)$.

2.3 The p-value

The most common method for getting p-value for a 2 by 2 table is Pearson's (1900) chi-squared statistic. However, it cannot be extended to test an association with more than one explanatory variable. For such a general situation logistic regression has been widely used in practice, especially for health science research. Therefore, we use the p-value from logistic regression for assessing our confidence intervals.

3. Simple Illustrations

To illustrate the methods, consider data for comparing proportions of improvement of patients using treatments

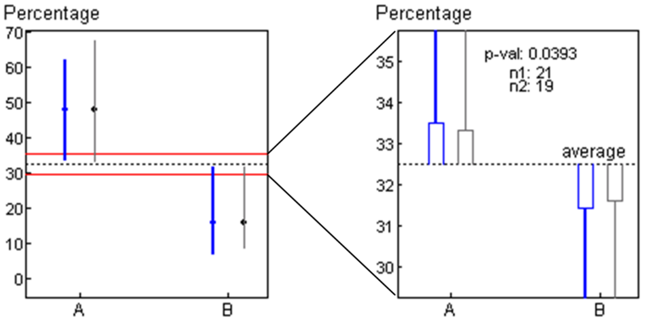

Figure 3. Comparing confidence intervals using weighted sum contrasts (light) and more exact delta method (thick).

A and B. The left panel of Figure 3 shows results with the thicker confidence intervals getting from the more exact delta method and the lighter confidence intervals getting from the weighted sum contrasts based on logistic regression. The two methods are consistent with the p-value, showing that treatment A is adjudged superior to treatment B. The confidence interval for treatment A is entirely above the mean, and the confidence interval for treatment B is entirely below the mean. For treatment A (a larger proportion, $p_1 = 0.476$) the weighted sum contrasts method gives a wider confidence interval but for treatment B (a smaller proportion $p_2 = 0.158$) the weighted sum contrasts method gives a shorter confidence interval compared to those from the more exact delta method. However, the right panel of Figure 3 shows that the more exact delta method gives larger gaps from the overall average (dotted line) for both proportions.

Figure 4 compares confidence intervals using the more exact delta method and those using weighted sum contrasts for the data from Figure 3 and other five data sets with increasing sample sizes to check asymptotic results. For these data sets the logistic models give p-values close to 0.04. The dotted line represents the mean overall proportion. The thicker confidence intervals are from the more exact delta method and the lighter confidence intervals are from weighted sum contrasts based on logistic regression. Specific values for the lower and the upper bounds of the confidence intervals are listed in Appendix A1. We can see that both methods give consistent results with the p-values, showing that treatment A is adjudged superior to treatment B. The two methods give confidence interval for treatment A is entirely above the mean, and the confidence interval for treatment B is entirely below the mean.

For situations with p-values greater than 0.05, we choose cases with p-values close to 0.06. Figure 5 shows such results. The confidence intervals from the two methods give consistent results with the p-values, showing that treatment A and treatment B are not different. The two methods give the confidence interval for treatment A with one arm is crossing the mean, and the confidence interval for treatment B is also one arm crossing the mean. Specific values for the lower and the upper bounds of the confidence intervals are listed in Appendix A2.

The challenging situation is when p-values are very close to 0.05. Figure 6 shows such results. The confidence intervals from the two methods are again consistent with p-values. Specific values for the lower and the upper bounds

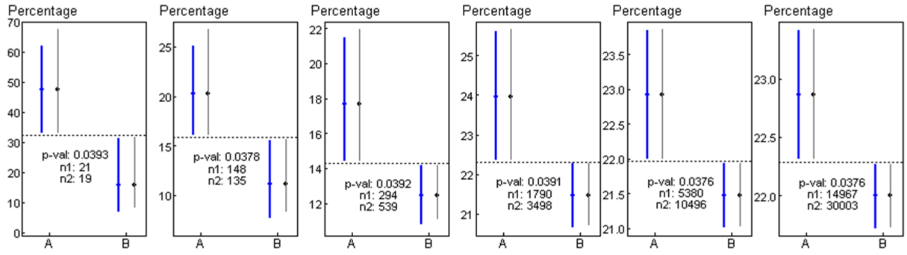

Figure 4. Examples with p-value less than 0.05.

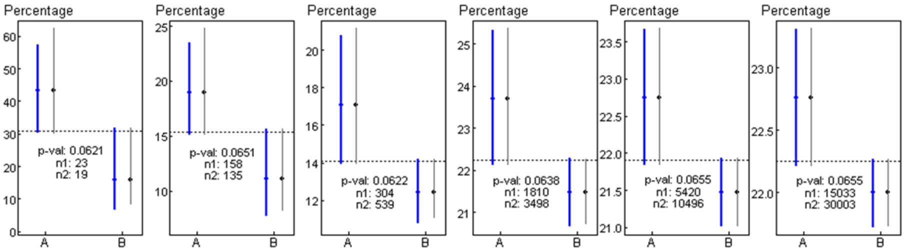

Figure 5. Examples with p-value greater than 0.05.

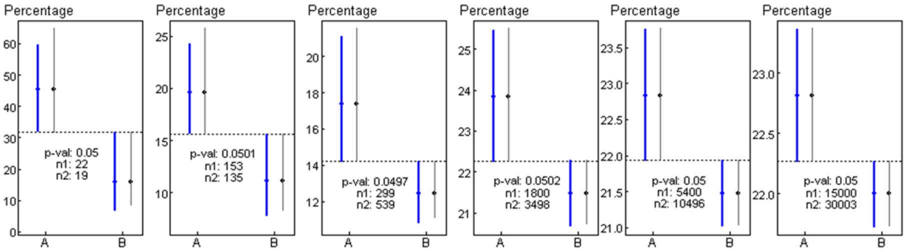

Figure 6. Examples with p-value close to 0.05.

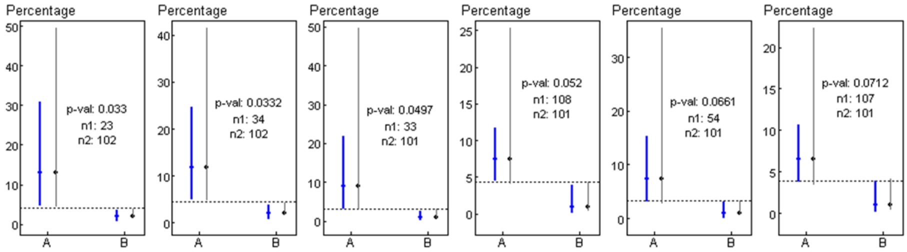

Figure 7. Examples of data with one proportion close to 0.

of the confidence intervals are listed in Appendix A3. In this case the confidence intervals from the weighted sum contrasts method are closer to the p-value than those from the delta method.

Next we use the two methods for situations when one group with proportion less than 0.01. Figure 7 shows such results. Specific values for the lower and the upper bounds of the confidence intervals are listed in Appendix A4. For these situations the two methods are consistent with p-values shown only for the graphs on the three left most panels. On the three right most panels, the p-values are greater than 0.05 but the more exact confidence intervals do not overlap. The method performs better when the p-value is larger with one arm across the mean as the right most panel shows. However, for situations with one group having a small value proportion the p-value from logistic regression is also unstable. The situation with one proportion close to 0 or 1 is a special case and different methods are needed for analysis of such data (Irala *et al.*, 1997).

4. Conclusions

In this paper we have clarified the method of weighted sum contrasts for logistic regression proposed by Tongkumchum and McNeil (2009) and applied the more exact delta method commonly used for individual proportions to estimate the standard error and thus to construct confidence intervals for comparing two population proportions in the 2×2 table.

The two methods can be used to create a graph for comparing the two confidence intervals of proportions that agree with the p-value from logistic regression. They give an accurate graph for comparing confidence intervals consistent with the p-value.

The difference between the two methods is that for the larger proportion the weighted sum contrasts method gives a wider confidence interval but for the smaller proportion the weighted sum contrasts method gives a shorter confidence interval compared to those from the more exact delta method.

An advantage of the weighted sum contrasts method is that it can be extended to more general situations and also can handle covariates using the same procedure. Note that the standard errors using this weighted sum contrasts method are shorter than the standard errors of individual by a factor $\sqrt{1-r_j}$. Further research would usefully focus on situations when one of the proportions is close to 0 or 1.

Acknowledgement

This research is supported by the Centre of Excellence in Mathematics, the Commission on Higher Education, Thailand. We would like to thank Don McNeil for his helpful guidance.

References

Austin, P.C. and Hux, J.E. 2002. A brief note on overlapping confidence intervals. Journal of Vascular Surgery. 36 (1), 194-195.

Cumming, G. 2009. Inference by eye: Reading the overlap of independent confidence intervals. Statistics in Medicine. 28, 205-220.

Cumming, G. and Finch, S. 2005. Inference by eye: Confidence Intervals and How to Read Pictures of Data. American Psychologist. 60(2), 170-180.

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., Lo, J., Mcmenamin, N. and Wilson, S. 2007. Statistical reform in psychology: Is anything changing? Psychological Science. 18, 230-232.

Geater, A., Duerawee, M., Chompikul, J., Chairatanamanokorn, S., Pongsuwan, N., Chongsuvivatwong, V. and McNeil, D. 2000. Blood lead levels among school children living in the Pattani river basin: two contamination scenarios. Journal of Environmental Medicine. 2(1), 11-16.

Hosmer, D.W. and Lemeshow, S. 2000. Applied Logistic Regression. 2nd edition. John Wiley and Sons, New York, U.S.A.

Irala, J.I., Navajas, F.C. and Castillo, A.S. 1997. Abnormally wide confidence intervals in logistic regression: interpretation of statistical program results. Revista Panamericana de Salud Pública / Pan American Journal of Public Health. 2(4), 268-271.

Kleinbaum, D.G.and Klein, M. 2002. Logistic Regression: A Self-Learning Text. 2nd edition. Springer-Verlag, New York, U.S.A.

Knol, M.J. and Pestman, W.R. 2011. The (mis)use of overlap of confidence intervals to assess effect modification. European Journal of Epidemiology. 26, 253-254.

Kongchouy, N. and Sampantarak, U. 2010. Confidence Intervals for Adjusted Proportions using Logistic Regression. Modern Applied Science. 4(6), 2-7.

Oehlert, G.W. 1992. A note on the Delta Method. The American Statistician. 46(1), 27-29.

Papke, L.E. and Wooldridge, J.M. 2005. A computational trick for delta-methods errors. Economics Letters. 86, 413-417.

Payton, M.E., Greenstone, M.H. and Schenker, N. 2003. Overlapping confidence intervals or standard error intervals: What do they mean in terms of statistical significance? Journal of Insect Science. 3(34), 1-6.

Peason, K. 1900. On a critical that a given system of deviations from the probable in the case of a correlated system of variable is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine. 50(5), 157-175.

R Development Core Team. (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org [March, 16, 2013].

Ryan, G.W. and Leadbetter, S.D. 2002. On the misuse of confidence intervals for two means in testing for the significance of the difference between the means. Journal of Modern Applied Statistical Methods. 1(2), 473-478.

Schenker, N. and Gentleman, J.F. 2001. On Judging the significance of differences by examining the overlap between confidence intervals. The American Statistician. 55(3), 182-186.

Tongkumchum, P. and McNeil, D. 2009. Confidence intervals using contrasts for regression model. Songklanakarin Journal of Science and Technology. 31(2), 151-156.

Venables, W. and Ripley, B. 2002. Modern Applied Statistics with S. Springer-Verlag, New York, U.S.A.

Wheeler, M.W., Park, R.M. and Bailer, A.J. 2006. Comparing median lethal concentration values using confidence interval overlap or ratio tests. Environmental Toxicology and Chemistry. 25(5), 1441-1444.

Wolfe, R. and Hanley, J. 2002. If we're so different, why do we keep overlapping? When 1 plus 1 doesn't make 2. Canadian Medical Association Journal. 166(1), 65-66.

Xu, J. and Long, J.S. 2005. Confidence intervals for predicted outcomes in regression models for categorical outcomes. The Stata Journal. 5(4), 537-559.

Appendix

A1. Lower bound (CILB) and upper bound (CIUB) of confidence intervals from Figure 4.

data set	group	Delta method		Weighted sum contrasts		Overall percentage
		CILB	CIUB	CILB	CIUB	
1	A	33.503	62.126	33.311	67.527	32.500
	B	7.126	31.422	8.723	31.616	
2	A	16.161	25.111	16.158	26.749	15.901
	B	7.816	15.561	8.415	15.623	
3	A	14.447	21.472	14.450	21.950	14.286
	B	10.868	14.182	11.134	14.197	
4	A	22.395	25.612	22.396	25.658	22.315
	B	20.688	22.272	20.728	22.273	
5	A	22.018	23.844	22.018	23.860	21.964
	B	21.021	21.936	21.035	21.936	
6	A	22.319	23.418	22.319	23.423	22.286
	B	21.729	22.269	21.734	22.270	

A2. Lower bound (CILB) and upper bound (CIUB) of confidence intervals from Figure 5.

data set	group	Delta method		Weighted sum contrasts		Overall percentage
		CILB	CIUB	CILB	CIUB	
1	A	30.643	57.252	30.263	62.417	30.952
	B	7.001	31.833	8.412	31.798	
2	A	15.178	23.488	15.124	24.799	15.358
	B	7.773	15.640	8.272	15.636	
3	A	13.981	20.760	13.969	21.164	14.116
	B	10.852	14.202	11.082	14.200	
4	A	22.148	25.329	22.147	25.368	22.231
	B	20.686	22.275	20.720	22.274	
5	A	21.856	23.668	21.855	23.681	21.909
	B	21.020	21.937	21.032	21.937	
6	A	22.221	23.315	22.221	23.319	22.253
	B	21.728	22.270	21.732	22.270	

A3. Lower bound (CILB) and upper bound (CIUB) of confidence intervals from Figure 6.

data set	group	Delta method		Weighted sum contrasts		Overall percentage
		CILB	CIUB	CILB	CIUB	
1	A	32.003	59.604	31.708	64.913	31.707
	B	7.062	31.633	8.564	31.707	
2	A	15.654	24.274	15.622	25.740	15.625
	B	7.794	15.601	8.341	15.628	
3	A	14.210	21.110	14.205	21.550	14.200
	B	10.860	14.192	11.108	14.198	
4	A	22.271	25.469	22.271	25.513	22.273
	B	20.687	22.273	20.724	22.273	
5	A	21.936	23.756	21.936	23.770	21.936
	B	21.021	21.936	21.033	21.936	
6	A	22.270	23.366	22.270	23.371	22.270
	B	21.728	22.270	21.733	22.270	

A4. Lower bound (CILB) and upper bound (CIUB) of confidence intervals from Figure 7.

data set	group	Delta method		Weighted sum contrasts		Overall percentage
		CILB	CIUB	CILB	CIUB	
1	A	4.773	30.983	4.541	49.444	4.000
	B	1.085	3.518	2.004	3.887	
2	A	5.125	24.761	4.916	41.493	4.412
	B	0.984	3.871	1.823	4.255	
3	A	3.446	21.888	2.991	49.691	2.985
	B	0.375	2.589	0.981	2.983	
4	A	4.625	11.659	4.271	25.300	4.306
	B	0.242	3.957	0.517	4.344	
5	A	3.397	15.399	2.957	35.405	3.226
	B	0.312	3.099	0.740	3.379	
6	A	3.942	10.666	3.556	22.310	3.846
	B	0.243	3.945	0.493	4.178	