Received: 7 August 2023 | Accepted: 9 March 2024

DOI: 10.1111/2041-210X.14318

Methods in Ecology and Evolution

RESEARCH ARTICLE

Classical tests, linear models and their extensions for the analysis of 2 × 2 contingency tables

Rebecca Nagel | Graeme D. Ruxton | Michael B. Morrissey

Centre for Biological Diversity, School of Biology, University of St Andrews, St Andrews, UK

Correspondence
Rebecca Nagel
Email: rn71@st-andrews.ac.uk

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: 515410943; Royal Society University Research Fellowship

Handling Editor: Timo Adam

Abstract

1. Ecologists and evolutionary biologists are regularly tasked with the comparison of binary data across groups. There is, however, some discussion in the biostatistics literature about the best methodology for the analysis of data comprising binary explanatory and response variables forming a 2 × 2 contingency table.
2. We assess several methodologies for the analysis of 2 × 2 contingency tables using a simulation scheme of different sample sizes with outcomes evenly or unevenly distributed between groups. Specifically, we assess the commonly recommended logistic (generalised linear model [GLM]) regression analysis, the classical Pearson chi-squared test and four conventional alternatives (Yates' correction, Fisher's exact, exact unconditional and mid-p), as well as the widely discouraged linear model (LM) regression.
3. We found that both LM and GLM analyses provided unbiased estimates of the difference in proportions between groups. LM and GLM analyses also provided accurate standard errors and confidence intervals when the experimental design was balanced. When the experimental design was unbalanced, sample size was small, and one of the two groups had a probability close to 1 or 0, LM analysis could substantially over- or under-represent statistical uncertainty. For null hypothesis significance testing, the performance of the chi-squared test and LM analysis were almost identical. Across all scenarios, both had high power to detect non-null effects and reject false positives. By contrast, the GLM analysis was underpowered when using z-based p-values, in particular when one of the two groups had a probability near 1 or 0. The GLM using the LRT had better power to detect non-null results.
4. Our simulation results suggest that, wherever a chi-squared test would be recommended, a linear regression is a suitable alternative for the analysis of 2 × 2 contingency table data. When researchers opt for more sophisticated procedures, we provide R functions to calculate the standard error of a difference between two probabilities from a Bernoulli GLM output using the delta method. We also explore approaches to compliment GLM analysis of 2 × 2 contingency tables with credible intervals on the probability scale. These additional operations should

support researchers to make valid assessments of both statistical and practical significances.

KEYWORDS
2×2 contingency table, chi-squared test, linear models, logistic GLMs, uncertainty estimates

1 | INTRODUCTION

From sex differences in survival rates (e.g. Teder & Kaasik, 2023) to taxonomic bias in publications (e.g. Rosenthal et al., 2017), assessing categorical differences in discrete variables is central to many questions in ecology and evolution. A legitimate concern among biologists is how to correctly analyse this type of data. In particular, what statistical tests produce an unbiased estimate of group differences, what assumptions need to be satisfied, and how should one interpret and present the results?

Here, we consider a simple but commonly encountered situation where two samples of binary data are compared. Such data are often presented in a 2×2 table, also called a contingency table since the key question is whether there is a contingency between the row and column variables. For example, given a set of sampled rodents, an individual's sex and whether they showed symptoms of some specific disease may be recorded. The natural set of questions to then ask are whether sex and disease status are linked and, if so, what is the strength and direction of the effect. We would also generally like to quantify how confident we can be in our answers to these questions.

Given such a 2×2 contingency table of count data, the Pearson chi-squared test (Pearson, 1900) is widely viewed as a reasonable way to test for differences between groups (Albert, 2017; Altman & Krzywinski, 2017; Crawley, 2012; Dytham, 2011; Fagerland et al., 2017; Seltman, 2018; Whitlock & Schluter, 2009). Variations on the chi-squared test depending on data structure are also commonly discussed. For example, to account for small sample sizes, Yates' correction for continuity (Yates, 1934) or Fisher's exact test (Fisher, 1934) are sometimes recommended (Altman & Krzywinski, 2017; Crawley, 2012) although not without some dissention (Fagerland et al., 2017; Ruxton & Neuhäuser, 2010). The use of Fisher's exact test, in particular, is controversial given that both margins of the contingency table are rarely naturally fixed (see e.g. Agresti, 1992; Berkson, 1978; Kempthorne, 1979), leading to unnecessarily conservative estimates. The Fisher mid-p (Lancaster, 1961) test is therefore often suggested to better represent evidence against the null hypothesis (Fagerland et al., 2017; Hwang & Yang, 2001; Routledge, 1994). Alternatively, Lydersen et al. (2009) recommend the exact unconditional test (Barnard, 1945; Boschloo, 1970) with a Berger and Boos correction (Berger & Boos, 1994) as the 'gold standard' for testing association in 2×2 tables.

There are, however, two major disadvantages of using the chi-squared test (or its alternatives and refinements discussed above) to analyse contingency tables. First, these test statistics only assess the distribution of the response variable under the null hypothesis that there is no difference between groups. They do not provide any information on the strength or direction of association among variables. This is a particularly relevant criticism as ecologists and evolutionary biologists shift away from principally relying on null hypothesis significance testing (NHST) and *p*-values (Stephens et al., 2007). Second, these test statistics are somewhat limited in application. When testing for an association between two categorical variables, both the explanatory and response variables must be dichotomous. While this type of data is very common in ecology and evolution, it limits the expansion of the analysis to more complex datasets. For example, if the hypothetical rodent study above also recorded the mass of each individual, this continuously distributed factor cannot be easily incorporated into any of the methods discussed above.

Both these limitations of the chi-squared test can be overcome by instead adopting regression analyses. Not only does a regression analysis provide more informative insights into the relationship between explanatory and response variables (describing strength and direction of effects), the precision of an estimate is also easily assessed. Results can thus be more intuitive, that is, offering an effect size and uncertainty (mean±SE). A regression analysis is also more versatile and can be expanded beyond the 2×2 contingency table. Using this single methodology, a biologist can investigate combinations of quantitative and categorical explanatory variables hypothesised to have an important influence on the response variable. Despite these obvious advantages, there seems to be some confusion and discussion in the literature about the appropriateness of regression analyses to assess the association between binary variables.

Foremost, biologists are widely discouraged from using linear regressions to analyse contingency tables. The most commonly cited cause of concern is that, for the analysis of data comprising binary response and categorical explanatory variables, the assumptions of normality and homogeneity of variance are violated (Kaplan, 2017; Seltman, 2018; Tutz, 2012). Using a linear regression is thus regarded by some as 'unsatisfactory' (Tutz, 2012) at best and 'completely unreliable' (Seltman, 2018) at worst. By contrast, the logistic regression is largely accepted as an appropriate method to estimate probabilities of categorical explanatory variables, compensating for the aforementioned problems with linear regressions (Dunn & Smyth, 2018; Fagerland et al., 2017; Lever et al., 2016; Orme & Combs-Orme, 2009; Ramos et al., 2015; Tutz, 2012). Nonetheless, a clear advantage of linear over logistic regression is the often more straightforward interpretation of model coefficients (without requiring conversion).

When assessing the possible differences between groups in dichotomous variables, biologists tend to ask about differences in *probabilities* between groups. In the linear regression of a binomial outcome on a group variable, a one-unit increase in the explanatory *x* increases the conditional expectation of the response *y* by β units. By contrast, the coefficients of logistic regressions are given in log-odds, which will often be a less intuitive unit for researchers asking biological questions. Even after converting the coefficient for the explanatory variable back to an ordinary scale, this odds ratio does not represent a constant increase or decrease in the response *y* given the explanatory variable *x*. This can make results difficult to contextualise (Gallis & Turner, 2019; Halvorson et al., 2021), even more so by non-experts who might be unfamiliar with *odds* as a statistical measure of the probability of one outcome versus the other (Grant, 2014; Schwartz et al., 1999).

Here, we assess the appropriateness of different statistical tests for the analysis of data comprising binary explanatory and response variables (both coded as 0,1). In contrast to many previous works on this subject, we focus on different goals of statistical analysis beyond *p*-values, namely estimation and statements about uncertainty in differences in probabilities. Odds, odds ratios and log-odds ratios are all valid ways of presenting results and are well treated in many works. We focus specifically on differences in probabilities, in extension of typical considerations, because this is the scale most relevant and intuitive to biologists and the scale that will frequently map onto how biologists formulate research questions. Using a simulation scheme of different sample sizes with *n* outcomes evenly or unevenly distributed between groups, we assess the biasedness of inferences of (a) the difference in proportions between groups and (b) estimates of statistical uncertainty (standard errors and confidence intervals) when using linear and logistic regressions. We also assessed and compared (c) the false-positive rate and (d) the consistency of *p*-values for the Pearson chi-squared test and its alternatives, linear regression and logistic regression.

2 | METHODS

2.1 | Simulation scheme

Each simulation generated data for two groups. These groups could represent control vs. treatment, male vs. female, etc. Sample size in each group was determined according to four schemes. In the first, the total sample size of n_t was divided evenly between the two groups ($n_t/2$). In the second, n_t was apportioned between groups following a binomial distribution with probability 0.5. In the third scenario, the groups had sample sizes 0.8 n_t and 0.2 n_t. Finally, the total of n_t samples was partitioned between groups following a binomial distribution with probability 0.2 for group $x=0$. When data were distributed randomly between groups, if a group had fewer than two observations, we modified the values of *x* so that there were at least two cases in each group. This was done on the grounds that most researchers would not look to explore contingency between groups when one of the groups had such a small sample size.

We then simulated data to represent a Bernoulli response variable for all the observations in both groups. This variable *y* took values of 1 and 0, which could represent outcomes such as survived vs. died, mated vs. unmated, etc. Observations from group $x=0$ were assigned a probability p_0 of success (i.e. of $y=1$, rather than $y=0$), and the *n* values for *y* were drawn as independent samples from a Bernoulli distribution with probability p_0. The probability of success for observations in group $x=1$ was defined by $p_1 = p_0 + \delta$, and values of *y* for individuals in group $x=1$ were drawn from a Bernoulli distribution with probability of success of p_1. Formally, the simulation of the *y* data can be described according to

$$y_i \sim \text{Bernoulli}(p_0 + \delta x_i), \tag{1}$$

where y_i represents success or failure (e.g. survived vs. died) coded as 1 and 0, respectively, with x_i coding the group membership (e.g. male vs. female; 0, 1) of individual *i*. p_0 is the probability of success in the first group ($x=0$). δ is the difference in success probability between the two groups.

2.2 | Variable ranges and replication

For each simulation scheme, we simulated a range of sample sizes of n_t (the total number of samples across both groups) from [10, 20, 30, 50, 70, 100]. We considered two sets of values of p_0 in combination with δ. First, we considered a value of p_0 of 0.5 and values of δ between −0.5 and +0.5 in increments of 0.1. Second, we considered a value of p_0 of 0.1 and values of δ from −0.1 to +0.9, also in increments of 0.1.

2.3 | Analyses

For every simulated dataset of values of *x* and *y*, we conducted a range of analyses focused on estimation of the parameter δ, that is, of the difference in the underlying probability of success (of $y=1$) in the two groups coded $x=0$ and $x=1$. First, we conducted linear model (LM) analyses of the dependence of *y* on *x*. We fitted the model

$$y_i = \alpha_{LM} + \beta_{LM} x_i + e_i, \tag{2}$$

where y_i is the outcome (0, 1) for individual *i*, from group x_i (0, 1). α_{LM} is the intercept and β_{LM} is the slope of the linear regression of *y* on *x*, which estimate p_0 and δ, respectively. We use the standard ordinary least squares (OLS) standard error (SE) of β_{LM} as the standard error of δ, and use a Wald-type confidence interval (estimate ± 1.96 SEs) as our LM-based confidence interval (CI) for δ. We use the *t*-test-based *p*-value for β_{LM} from the summary.lm() function in base R as our LM-based *p*-value for δ.

Secondly, we conducted a generalised linear model (GLM) analysis with a logit link function and a Bernoulli response, also known as a logistic regression. This model was constructed as

$$\eta_i = \alpha_{\text{GLM}} + \beta_{\text{GLM}} x_i \tag{3a}$$

$$y_i \sim \text{Bernoulli}\left(g^{-1}(\eta_i)\right), \tag{3b}$$

where the data are modelled as a linear function on the scale of a latent variable η, related to expected probabilities via the inverse link function $g^{-1}()$ (i.e. the inverse of the logit link function), defined $g^{-1}(a) = \frac{e^a}{1+e^a}$. Bernoulli sampling of the data y_i was performed given the probabilities $g^{-1}(\eta_i)$. As such a GLM-based estimator of δ is

$$\hat{\delta}_{\text{GLM}} = g^{-1}\left(\alpha_{\text{GLM}} + \beta_{\text{GLM}}\right) - g^{-1}\left(\alpha_{\text{GLM}}\right). \tag{4}$$

$\hat{\delta}_{\text{GLM}}$ is thus a somewhat complex quantity, depending not only on the GLM parameter that describes the difference between the groups (β_{GLM}) but also on the model intercept (α_{GLM}). This renders its interpretation more difficult than the estimate from the LM analysis. While biological inferences on the logit data scale are possible, it would be useful if estimates of differences in probabilities (as in Equation 4) were used more widely. Being able to put uncertainty on the probability scale would be useful as well. A standard error for the $\hat{\delta}_{\text{GLM}}$ estimator, constructed by the delta method (Lynch & Walsh, 1998; Ver Hoef, 2012), is

$$\text{SE}\left[\hat{\delta}_{\text{GLM}}\right] \approx \sqrt{\begin{bmatrix} \frac{\partial \hat{\delta}_{\text{GLM}}}{\partial \hat{\alpha}_{\text{GLM}}} \\ \frac{\partial \hat{\delta}_{\text{GLM}}}{\partial \hat{\beta}_{\text{GLM}}} \end{bmatrix}^T \Sigma_{\alpha\beta,\text{GLM}} \begin{bmatrix} \frac{\partial \hat{\delta}_{\text{GLM}}}{\partial \hat{\alpha}_{\text{GLM}}} \\ \frac{\partial \hat{\delta}_{\text{GLM}}}{\partial \hat{\beta}_{\text{GLM}}} \end{bmatrix}}, \tag{5}$$

where $\frac{\partial \hat{\delta}_{\text{GLM}}}{\partial \hat{\alpha}_{\text{GLM}}}$ and $\frac{\partial \hat{\delta}_{\text{GLM}}}{\partial \hat{\beta}_{\text{GLM}}}$ are the derivatives of $\hat{\delta}_{\text{GLM}}$ in Equation 4 with respect to the logistic intercept and contrast estimated by the GLM (as specified in Equation 3a,b), evaluated at their estimated values. T is the transpose operator. $\Sigma_{\alpha\beta,\text{GLM}}$ is the covariance matrix of the GLM's parameter estimates. The square roots of the diagonal elements are the standard errors of the logistic intercept and contrast terms.

It is relatively straightforward to generate a CI from a binomial GLM on the probability scale for a conditional probability (e.g. for one group or the other, in the present context). One would typically generate a Wald-type CI on the linear predictor scale and transform its upper and lower limits to the probability scale with the inverse logit function. However, this strategy is not generally possible for *differences* in conditional probabilities. Consequently, similarly to how SEs for differences on the probability scale are not generally considered—useful as they would be—CIs for differences or data-scale effects in GLMs are not routinely generated by biologists.

There are many methods for setting CIs, and our aim was not to provide a comprehensive overview. We did, however, investigate the utility of the three most commonly accepted ways of constructing CIs for statistical models (Alan, 2013; Fagerland et al., 2017; Newcombe, 1998). First, we tested the simple probability scale Wald-type CIs constructed as 1.96 times the delta-SE (Equation 5) above and below the estimated difference (Equation 4). Second, we assessed the Wilson or score CIs (Wilson, 1927), making use of an adapted z2stat function from the R package PropCIs for our calculations (Scherer, 2018). Finally, we used profile likelihood-based CIs (Venzon & Moolgavkar, 1988) of δ. To generate the profile likelihood CI, we fixed δ to a range of values between [−0.99 and 0.99] and found the value of p_0 that maximised the likelihood of observing the data, given the fixed value of δ. We then recorded the upper and lower values of δ for which twice the log-likelihood difference from the unconstrained model was less than the 95% quantile of a chi-squared distribution with one degree of freedom. We assessed significance of the β_{GLM} using both the standard z-test output from the summary.glm() function in base R and the likelihood ratio test (LRT). To implement the LRT, we generated a p-value under the assumption that twice the difference in the log-likelihood of the GLM in Equation 3a,b with the log-likelihood of an intercept-only GLM is chi-squared distributed with one degree of freedom.

Finally, we conducted four classical contingency table analyses on the 2×2 contingency tables of x and y values. These analyses do not estimate the value of δ, but can be interpreted as tests of statistical significance of δ, that is, of whether its value differs from a (null) hypothetical value of zero (no difference between p_0 and p_1). The first two such tests were chi-squared tests, without (Pearson, 1900) and with Yates' continuity correction (Yates, 1934). The third test was a Fisher's exact test (Fisher, 1934). In addition to the exact test's ordinary p-value, we calculated the mid-p (Lancaster, 1961) using the R package epitools (Aragon, 2020). The final test was an exact unconditional test (Barnard, 1945; Boschloo, 1970) with the Berger and Boos correction (Berger & Boos, 1994), implemented using the R package exact2x2 (Fay & Hunsberger, 2021).

2.4 | Evaluation of estimates and uncertainty statements

We first assessed results of linear and logistic regression analyses in terms of biasedness. Bias is the difference between the average value of an estimate and the true value of the quantity it is estimating. Formally,

$$\text{bias}\left[\hat{\delta}\right] = E\left[\hat{\delta}\right] - \delta, \tag{6}$$

where δ is the true value of the difference in probabilities and $E\left[\hat{\delta}\right]$ is the expected value of the estimator of the difference. As such, in unbiased analyses, the average value of the estimator δ applied to replicate simulated datasets would not differ from the true value in those simulations.

While bias is typically defined as a difference between the expected value of an estimator and the true value of the estimand (as in Equation 6), we expressed (un)biasedness in a proportional sense for our assessment of different methods for generating standard errors of differences in probability. Specifically, we divided the average SE for any given method by the empirical SD of the estimator,

$$\text{bias}_{\text{proportional}}\left[\hat{\delta}\right] = \frac{E\left[\text{SE}\left[\hat{\delta}\right]\right]}{\text{SD}\left[\hat{\delta}\right]}, \qquad (7)$$

where $E\left[\text{SE}\left[\hat{\delta}\right]\right]$ is the average standard error across replicate simulations, and $\text{SD}\left[\hat{\delta}\right]$ is the standard deviation of the estimated differences in probabilities across replicate simulations in any given scenario. As such, a value of one for the proportional bias measure for the standard error indicates ideal behaviour.

Similarly, we assessed confidence intervals by considering their coverage properties. For a given scenario, a 95% CI has the correct or unbiased (nominal) coverage when estimates fall within its bounds in 95% of replicate simulations.

All analyses were done in R (R Core Team, 2023). Code for the simulations, analyses, and figures presented in this manuscript can be found on GitHub and Zenodo (Nagel et al., 2024a; https://github.com/rebebba/ProbUncertainty_MSCode).

3 | RESULTS

All results were very similar among the four data-generating schemes. In other words, whether the distribution of the predictor variable was fixed or random, or whether the data were balanced or not between the two groups, most results were very similar. Therefore, except when notable differences occur, we report on results for the simulations with fixed and equal proportions of the data between the two groups. Results for the other three scenarios are presented in the Supplementary Materials.

3.1 | (Un)biasedness of direct estimators of δ

The two main analyses that generate estimates of δ are the LM analysis, which estimates δ directly, and the GLM analysis, from which an estimator of δ can be recovered using Equation 4. Both the LM and the GLM returned unbiased estimates of δ in the sense that the average value of the estimator was equal to the true value (Figure 1; also see Figures A1–A3) across all parameter value combinations that we considered.

3.2 | Performance of standard errors of LM- and GLM-based estimators of δ

Providing that observations were reasonably evenly divided among groups and that sample sizes were not extremely small (i.e. >5 observations per group and $n_t > 10$) standard errors of δ from LM and GLM analyses were good reflections of the uncertainty in the estimation process (Figure 2; also see Figure A4). However, when observations were extremely unevenly distributed between groups, standard errors from the LM analysis could substantially over- or under-represent statistical uncertainty (Figure 3; also see Figure A5).

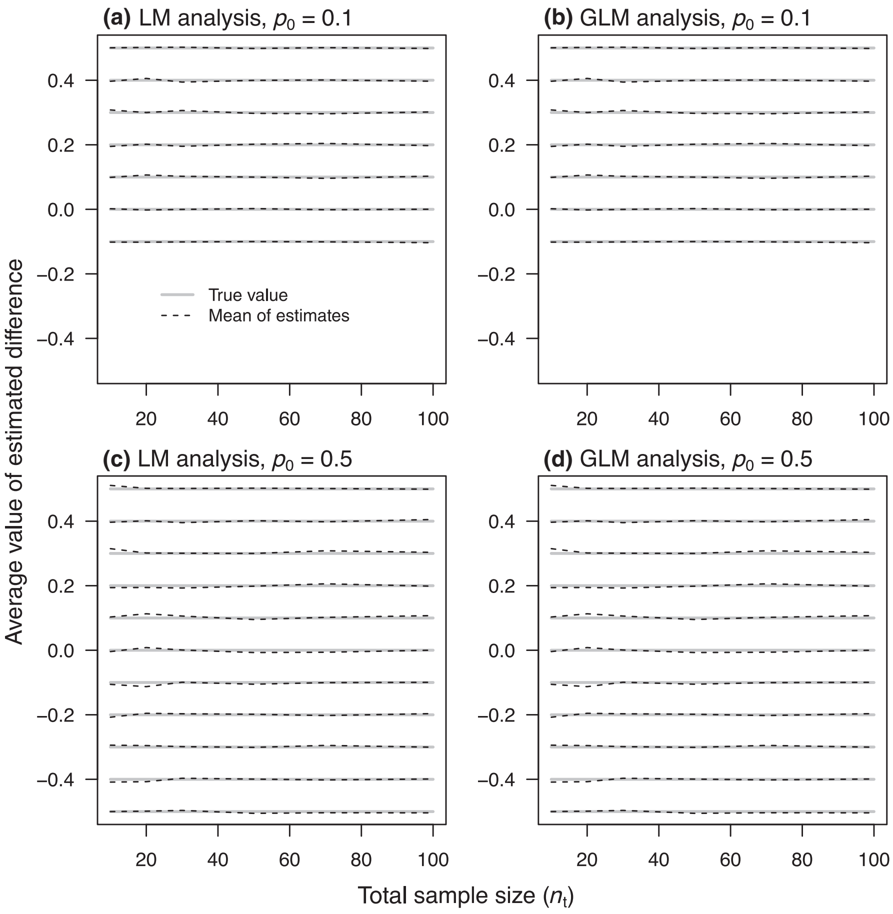

FIGURE 1 Unbiasedness of linear model (LM)- and generalised linear model (GLM)-based inferences of the difference in proportions between two groups (δ). In each simulation, one group has a true probability of p_0 such that in (a) and (b), $p_0 = 0.1$, while in (c) and (d), $p_0 = 0.5$. The other group has a probability of $p_1 = p_0 + \delta$, where δ takes values such that p_1 lies between 0 and 1. Simulated differences (black dotted lines) and true differences (grey solid lines) match closely for all simulation scenarios and all values of δ . Each parameter combination was simulated 1000 times and results given are means across replicate simulations. Simulations had fixed and equal proportions of the data between the two groups.

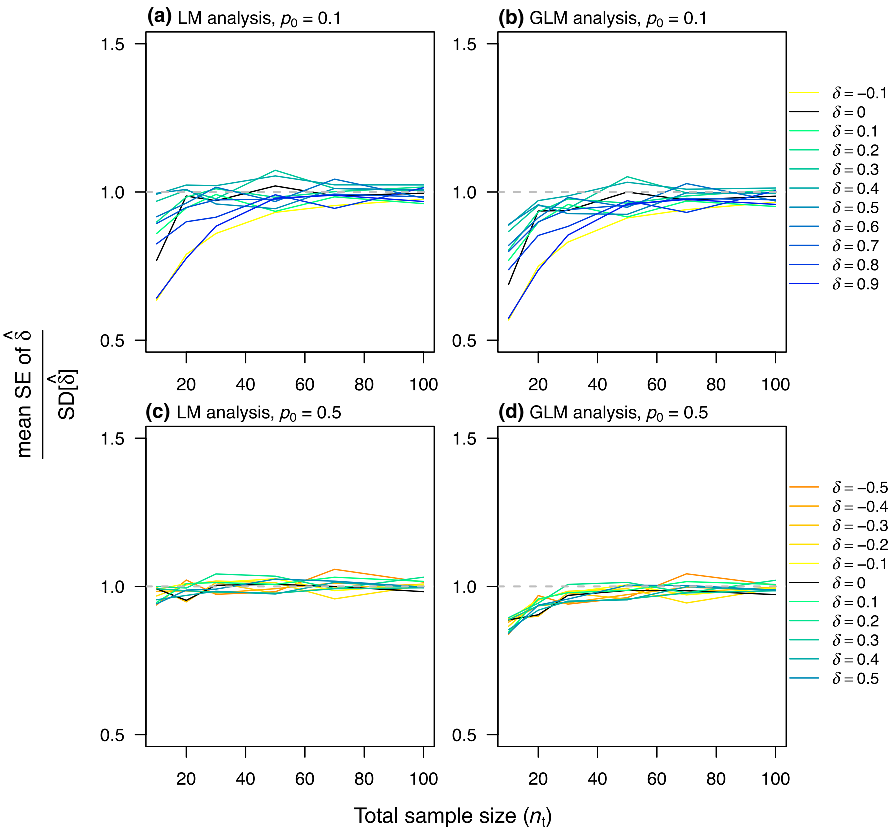

FIGURE 2 Validity of standard errors of linear model (LM)- and generalised linear model (GLM)-based inferences of the difference in proportions between groups (δ) when data are evenly distributed between the two groups. The standard error for the $\hat{\delta}_{GLM}$ estimator was constructed using the delta method. In (a) and (b), at least one group always has a probability of p_0=0.1, while in (c) and (d) at least one group always has a probability of p_0=0.5.

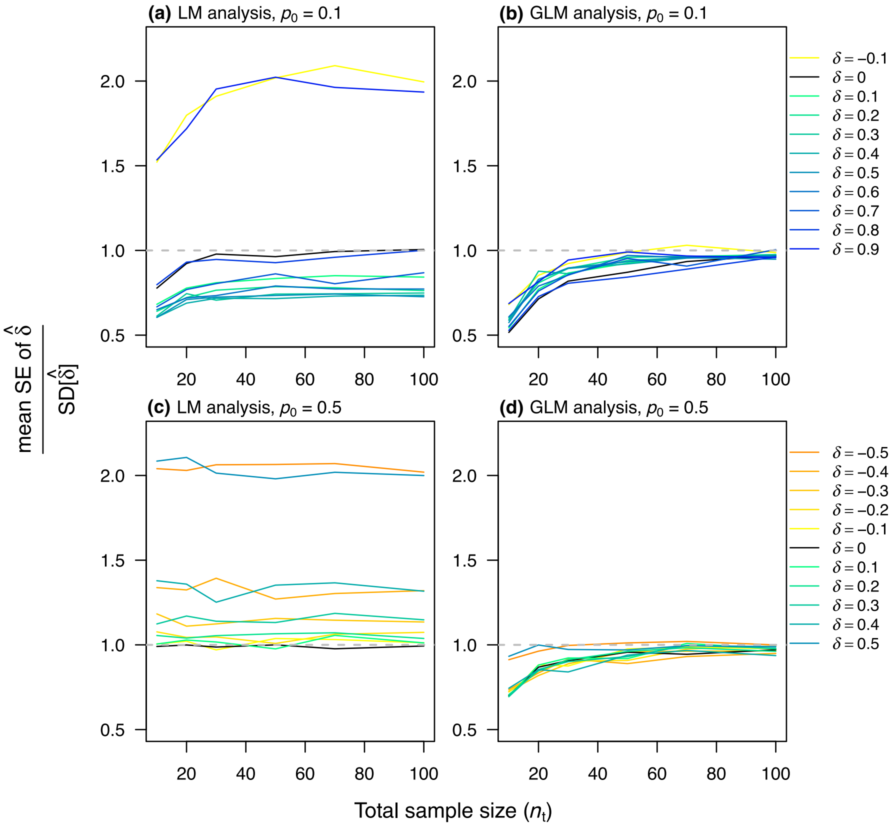

FIGURE 3 Validity of standard errors of linear model (LM)- and generalised linear model (GLM)-based inferences of the difference in proportions between groups (δ) when data are unevenly distributed between the two groups. The standard error for the $\hat{\delta}_{GLM}$ estimator was constructed using the delta method. In (a) and (b), the group with 20% of the observations always has a probability of p_0=0.1, while in (c) and (d) the group with 20% of the observations always has a probability of p_0=0.5.

3.3 | Coverage properties of confidence intervals for δ

For the LM analysis, coverage properties largely reflected the behaviour of standard errors. Performance was good providing a reasonable sample size ($n_t \geq 20$), and only deteriorated appreciably when data were highly unequal across groups and there were large differences in the true probabilities between groups (Figure 4a,e; also see corresponding plots in Figures A6–A8). The Wald-type CIs we generated based on the delta-SEs from the GLM analysis were generally conservative (Figure 4b,f), indicating greater uncertainty in the GLM estimates of differences between groups than was actually achieved. The score (Figure 4c,g) and profile likelihood (Figure 4d,h) methods for generating CIs for δ performed reasonably well across all scenarios that we simulated (also see corresponding plots in Figures A6–A8).

3.4 | Control of type 1 error rate

Under a balanced experimental design, the chi-squared test, mid-p, exact unconditional and LM analysis had a type 1 error rate very close to the conventionally accepted 5% threshold (Figure 5 at $\delta = 0$; also see corresponding plots in A10). Yates' correction, Fisher's exact and the GLM analysis with z-test p-values were generally most conservative. When $p_0 = 0.1$, observations were very unevenly distributed between groups, and sample sizes were small ($n_t \leq 20$) the type 1 error rate for both the chi-squared test and LM analysis was closer to 10% (Figure A9). Across most scenarios, the highest type 1 error rate was realised using a GLM analysis with p-values generated using the LRT (Figure 5 at $\delta = 0$; also see corresponding plots in Figures A9–A11).

3.5 | Power of null hypothesis statistical tests for δ

Across the range of scenarios considered, the chi-squared test, LM analysis and GLM analysis with p-values generated using the LRT generally gave the best performance, with the highest power to detect non-null effects (Figure 5; also see corresponding plots in Figures A9–A11). The poorest performance was realised in the GLM analysis with default p-values (z-test based) from the summary.glm() function in base R.

3.6 | Comparison of p-values for LM analyses and chi-squared tests

The underlying p-values generated from the linear models and chi-squared tests were nearly equivalent, both across the full dataset and at the $\alpha = 0.05$ level (Figure 6; also see Figures A12–A14).

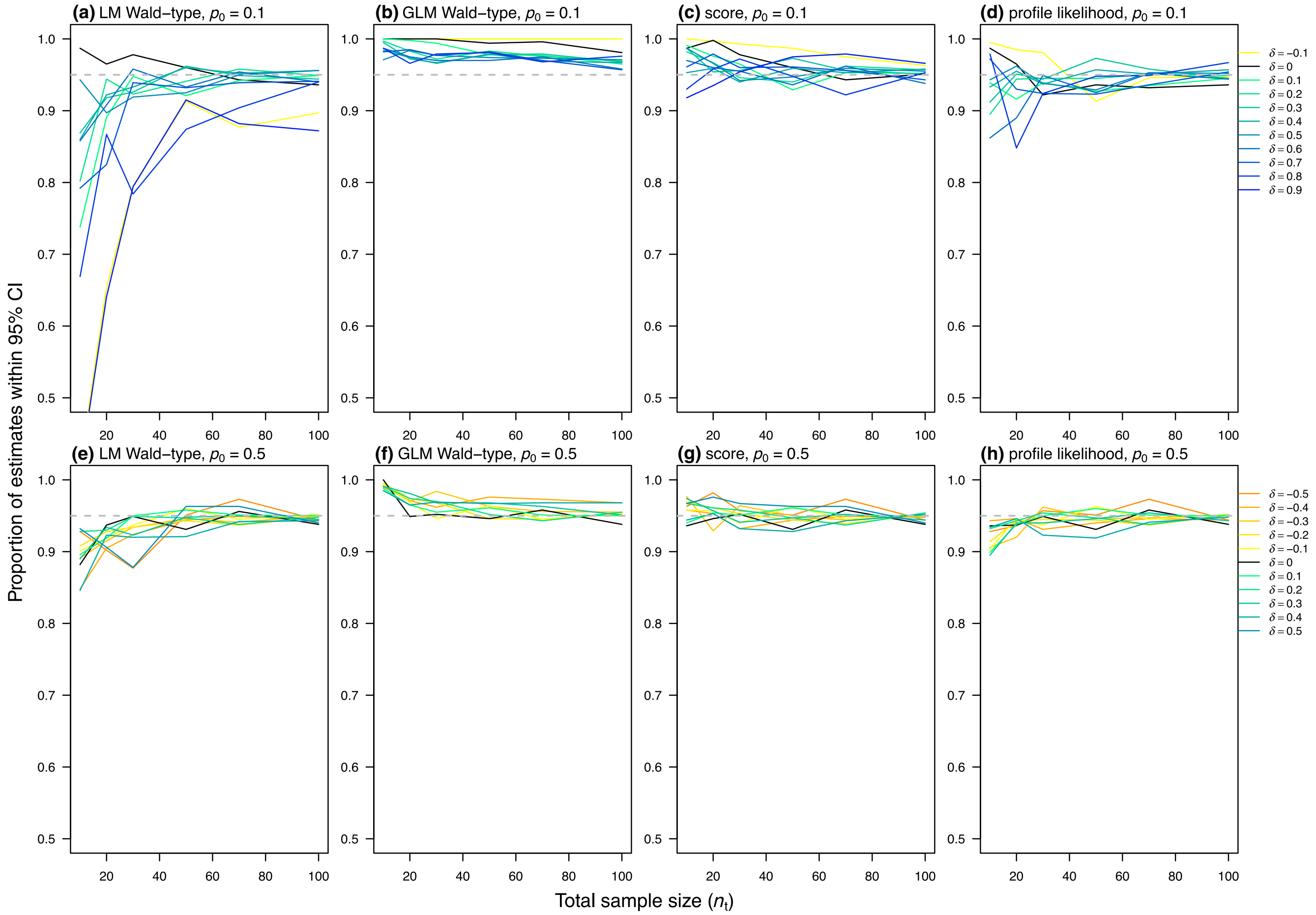

FIGURE 4 Coverage properties for 95% confidence intervals of differences between proportions in two groups for a range of analyses and true differences between groups. In the top row (a–h), at least one group has a probability of $p_0 = 0.1$ and in the bottom row (e–h), at least one group has a probability of $p_0 = 0.5$. Note that in (a), the proportion of estimates within 95% CI for $\delta = -0.1$ (yellow) and $\delta = 0.9$ (blue) at $n_t = 10$ are outside the y-axis range; the respective values are 0.404 and 0.398.

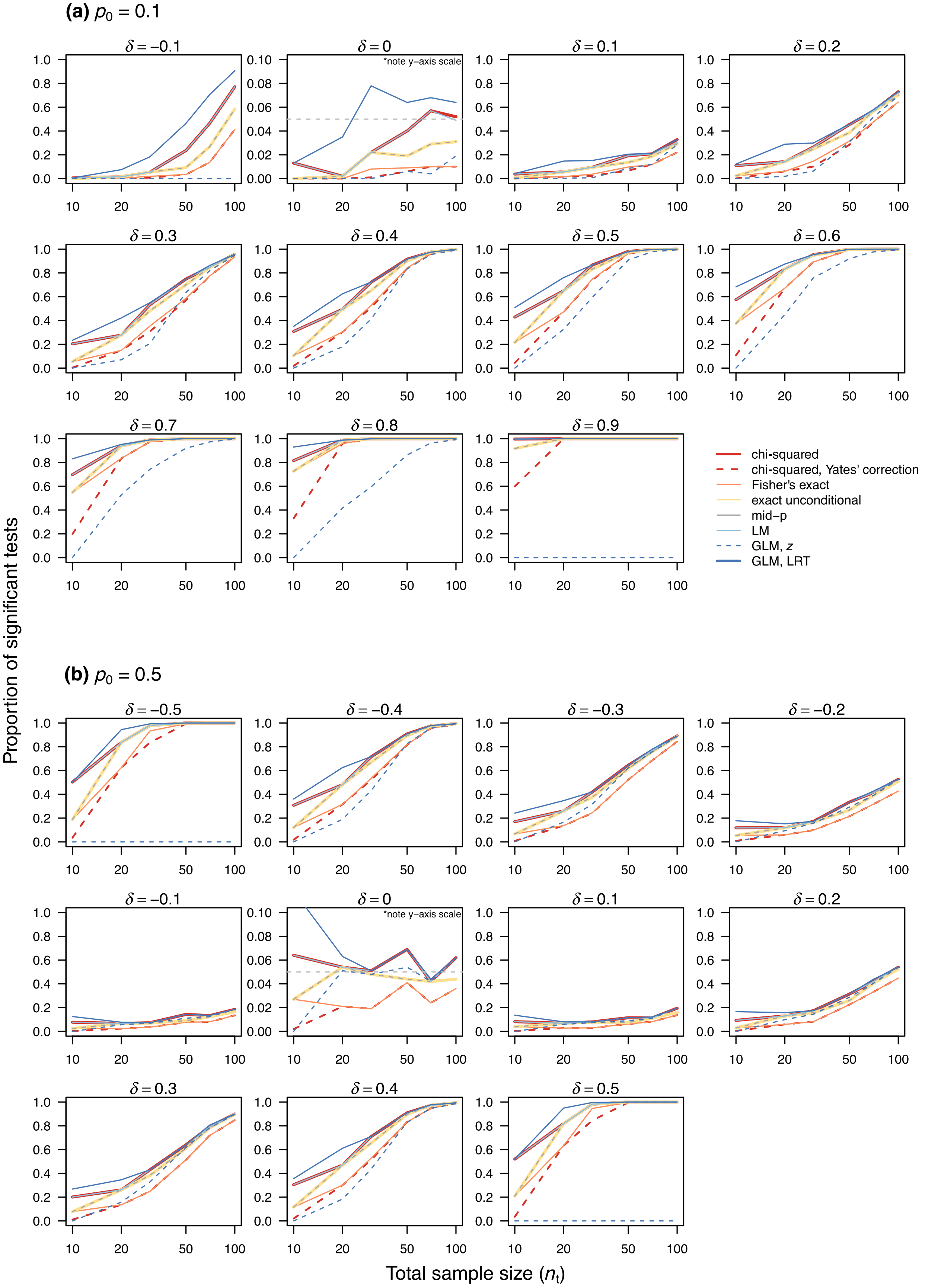

FIGURE 5 Power to reject false null hypotheses and false-positive rates in the case of no effect (δ = 0) of different tests. All lines depict the proportion of significant results at the α=0.05 level. In (a) at least one group always has a probability of p_0=0.1, and in (b) at least one group has a probability of p_0=0.5. Note the different y-axis scale in plots with no effect (δ=0), where we scale from [0:0.1] rather than [0:1] to focus on behaviour relative to the idealised false-positive rate of 0.05 (grey dotted line). Note that in (b), plots with equal values of | δ | should be identical, noise from finite numbers of simulations notwithstanding.

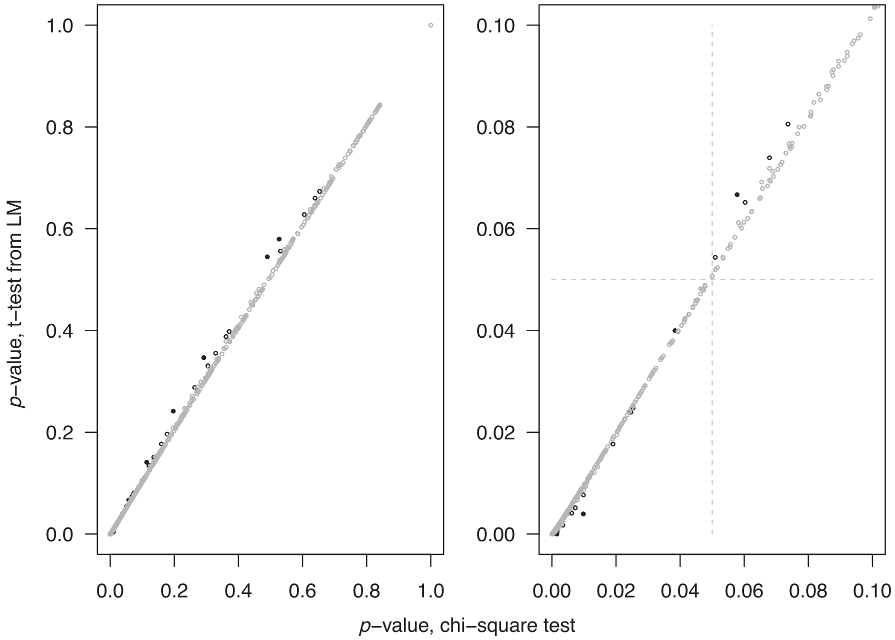

FIGURE 6 Correspondence between p-values from linear model (LM)-based inferences and the Pearson chi-squared test. For the LM analysis, response variables were regressed on groups, both coded as 0/1 data, such that the model slope estimated the difference between groups. In (a), results for all 1000 replicates of each of the 132 combinations of simulation parameters are shown. The same dataset is shown in (b), but values are restricted to $p \leq 0.1$ so that the correspondence between statistical significance at the $\alpha = 0.05$ level (grey dotted lines) can be more directly assessed. p-values from simulations with the smallest total samples sizes ($n_t = 10$) are plotted with closed black circles, sample sizes $n_t = 20$ are plotted with open black circles, and the largest sample sizes ($n_t \geq 30$) are plotted with open grey circles.

4 | DISCUSSION

Many biological phenomena have binomial outcomes: diagnostic tests may be passed or failed, individuals may survive or die, sites may or may not be occupied. Consequently, many biological questions boil down to investigating differences in probabilities. Here, we have investigated a range of approaches to characterise differences between groups. Our primary focus was on the estimation of the difference in probability between two groups and the ability to make reasonable statements about uncertainty in those differences. To this end, we first considered the performance of LMs and Bernoulli GLMs to produce unbiased estimates and meaningful standard errors and confidence intervals. Secondly, given that widely recommended approaches for the analysis of binary data (e.g. GLMs) do not directly generate standard errors or confidence intervals for differences in probabilities, we developed and tested methods to do so. Finally, we compared model-based approaches and classical tests for 2×2 contingency table analysis in terms of the performance of p-values associated with the null hypothesis of no difference between two groups. We discuss the main observations from these exercises in terms of what they mean for practicing scientists, and how analysing differences in probabilities might fit into the process of learning practical statistics.

4.1 | Estimation of differences in probability and assessment of uncertainty

A post-doc is looking over a study area with the long-term PI on their new project:

New post-doc: Do the unicorns in the valley have higher survival than those on the hills?

Experienced PI: Hmm, yes. Their log-odds of survival probably differs by 1.5.

New post-doc: As much as 1.5 on the log-odds scale?

Experienced PI: Indeed; maybe as much as 2!

A conversation that definitely never happened.

Almost certainly, the underlying interests of both participants in this fanciful conversation would have been about survival probability and the difference in those probabilities between habitats—not about differences in log-odds ratios. This is not to discount the fact that the log-odds are a statistically convenient scale, and not without biological importance. Rather, we think it likely that most biologists prefer to consider their data and results in terms of probabilities, finding them a more intuitive and illustratable statistic (Gallis & Turner, 2019; Halvorson et al., 2021). That, in the above example, the post-doc made a judgement about biological effect size and the PI made a statement about uncertainty, both on the log-odds ratio scale, is pretty improbable. We therefore consider how best to support this inclination and focus on the estimation of differences in probabilities between groups and what we can say about uncertainty in those estimates.

First, we focus on the linear regression, which gives estimated differences in probabilities and its standard errors directly as a model output. While many biostatistics texts give the impression that using a linear model to analyse binomial data would be dangerously naïve (Kaplan, 2017; Lever et al., 2016; Seltman, 2018; Tutz, 2012), we found that a linear model with a binary outcome (0, 1) regressed on a binary explanatory variable (0, 1) gave unbiased estimates of differences in probabilities. More specifically, the LM accurately estimated the true difference between the means of two groups (Figure 1). Standard errors from the LM did not, however, perfectly reflect the true uncertainty in estimated differences (Figures 2 and 3). Basic Wald-type confidence intervals (estimate ± 1.96 SEs) performed well under a broad range of circumstances but were compromised when sample size was small and data distributions were very skewed (Figure 4). Nonetheless, SEs and CIs from the LM analysis were generally of the correct order of magnitude. Furthermore, the conditions under which the LM-based SEs and CIs performed poorly are the same conditions under which a chi-squared test is generally discouraged (Crawley, 2012; Fagerland et al., 2017; Kang et al., 2006; Ruxton & Neuhäuser, 2010).

Where linear regression is discouraged for the analysis of contingency tables, logistic regression is widely recommended as an ideal means of modelling binomial data (Dunn & Smyth, 2018; Lever et al., 2016; Orme & Combs-Orme, 2009; Ramos et al., 2015; Tutz, 2012). While a binomial GLM analysis does not directly return inference of differences in probabilities, the conversion of predictions to the probability scale is reasonably commonplace. Calculating standard measures of uncertainty (SEs and CIs) on the probability scale is, however, less common. For this reason, we investigated a method that is standard in some areas of statistics, but not routinely utilised by biologists, to derive SEs of differences in probabilities: the delta method. We find that this linear approximation of the transformation from the log-odds scale to the difference in probabilities (Equation 5) performs reasonably well. It gave accurate SEs of differences in probabilities and was outperformed by the OLS-SEs only in a few circumstances involving the smallest sample sizes, and then only modestly so (Figure 2). The GLM delta-SEs also performed well in some scenarios where the LM analysis performed poorly, namely when differences between groups were large and one group had a true value very near 0 or 1 (Figure 3).

Similar to SEs of differences in probabilities, generating CIs of differences in probabilities from binomial GLMs is not standard practice in the fields of ecology and evolution. We tested three methods to do so. First, we applied the standard Wald-type method (estimate ± 1.96 SEs) using the delta-SEs. This generated CIs that were typically overly conservative, that is, that depicted greater ranges of uncertainty in the estimate of differences than actually occurred (Figure 4). This is undesirable since excessive conservatism can equate to wasting sample size, money and effort. We then tested the score and profile likelihood-based methods for generating CIs. Neither are directly linked to LM- or GLM-based inferences, but both focus on estimated differences in probabilities and make statements about uncertainty of those differences. Both methods generated CIs with approximately correct coverage properties across the range of scenarios investigated (Figure 4).

4.2 | Power and performance under H_0

While our primary focus was on the estimation of differences in probabilities and the generation of statements about uncertainty to accompany these estimates, we also considered two *p*-value-based assessments of model performance: power to reject false null hypotheses and false-positive rates in the case of no effect. Given that most classical approaches for analysing contingency tables generate *p*-values but not estimates or uncertainty statements about differences in probabilities, this also allowed us to compare regression model outputs with four commonly recommended alternatives and refinements to the chi-squared test. We offer some general conclusions drawn from this comparison.

4.2.1 | LM and chi-squared are basically identical from a NHST perspective

The power and false-positive rates from the LM analysis and the Pearson chi-squared test were nearly identical (Figures 5 and 6). This is inevitable, given that the underlying *p*-values generated by the two approaches are also nearly identical. The *p*-values from the LM analysis assume that estimation errors follow a t-distribution, which rapidly converges on a normal distribution when sample size ≥10 (Seltman, 2018). The *p*-values from the chi-squared test arise from a normal approximation to deviations from the null model of equal probabilities (Lydersen et al., 2009).

4.2.2 | GLM is good for H_0 testing, but requires care

The GLM-based approach was the most powerful NHST when H_0 was false; it was not outperformed under any combination of simulation parameters assessed (Figure 5). Also, the GLM false-positive rate most closely reflected the nominal rate at $\alpha = 0.05$, except in some cases with the smallest sample size. However, these properties are only true for GLM *p*-values generated from the LRT. The default statistical hypothesis test for GLMs, the *z*-test (i.e. the standard summary.glm() output in base R), can severely lack power when one or both groups have observed rates near 0 or 1; it can also be excessively conservative under the null hypothesis. This is a known phenomenon (Bolker et al., 2009). It is thus widely recommended to apply the LRT to binomial GLMs, especially when some cells of an experiment have very high or low probabilities (Agresti, 2007; Hauck & Donner, 1977). It is worth noting, however, that the LRT is not perfect. It gives modestly inflated type 1 error rates (i.e. *p*-values <0.05 when the true difference between groups is 0) in a range of circumstances and is sometimes excessively conservative.

4.2.3 | The behaviour of classical tests

All chi-squared test refinements and alternatives (Yates' correction, Fisher's exact test, mid-P and exact unconditional; see Section 2.3) produced results either more conservative than or comparable to the LM and Pearson chi-squared analyses (Figure 5). More specifically, they were excessively conservative when the null hypothesis was true and generally less powerful when the null hypothesis was false. Being excessively conservative or aiming to never conduct a test that might have a false positive rate slightly above the stated rate may initially seem laudable. However, conservatism when the null hypothesis is true is inevitably linked to lower power when the null hypothesis is false, and it could be argued that employing tests that underestimate the power and precision of an estimate is an inefficient use of resources.

Broadly speaking, across the range of circumstances that we simulated, p-values from the LM analysis (or very nearly equivalently, from the Pearson chi-squared test) were more often near the nominal value of $\alpha = 0.05$ than any other test, including the often-recommended GLM-LRT approach. However, marginal differences between tests under most simulated circumstances suggest that adopting the GLM-LRT approach on this basis risks adding unnecessary confusion to the analysis of contingency tables. While rigorously avoiding excessive false positives is of course desirable, it can be taken to extremes. These results also highlight the problems with expressing results only in terms of p-values (Stephens et al., 2007): None of the methodologies we assessed proved perfect for NHSTing under all simulated conditions.

4.3 | General considerations

The LM analysis robustly estimated differences of proportions between groups. Measures of statistical uncertainty (SEs and associated CIs) were largely correct representations of the true uncertainty, although somewhat compromised when total sample size was small and strongly unbalanced between the groups. However, these are the same conditions under which a chi-squared test is generally discouraged (Crawley, 2012; Fagerland et al., 2017; Kang et al., 2006; Ruxton & Neuhäuser, 2010). As such, the LM analysis of 2 × 2 contingency table data is equivalent to (in power and performance under the null) or superior to (easily generated estimates and uncertainty statements) the chi-squared analysis. This suggests at least three broad benefits to more widespread application of the LM approach to 2 × 2 contingency table analysis.

First, instruction in biological statistics may be well served by shifting to a LM-based approach. However, given the ubiquity of advice that standard linear regressions are unsuited to data such as arises in a 2 × 2 contingency table analysis, biostatistics instructors face a quandary. One option is to teach LMs for everything *except* contingency tables and then cover all statistical procedures typically handled in introductory statistics (e.g. the chi-squared test and its alternatives). A second option has instructors teaching GLM analyses. This would undoubtedly benefit students, but most curricula do not afford enough time to teach GLMs well and this leads to greater problems down the road. By one estimation, 58% of GLMs in ecology and evolution are inappropriate in some way (Bolker et al., 2009). Alternatively, our results show that the typical advice is wrong and LMs are reasonable alternatives to the chi-squared test. This should allow basic LMs to be used for model-based teaching covering the full range of classical tests (chi-squared, t-test, simple and multiple regression, ANOVA, ANCOVA, etc.).

Second, our results suggest that simple LMs provide a good balance between robustness and appropriateness of the methods, and ease of application and interpretation. LMs produce direct parameter estimates and unbiased uncertainty statements of differences in probabilities without common errors in application or interpretation. While the modern biologist may opt for more sophisticated procedures (e.g. GLM), confidence that more straightforward methods are indeed quite robust is useful. Furthermore, researchers can take comfort in the fact that results from those who prefer simpler methods (e.g. LM) are likely reliable. This view may be particularly useful to meta-analysts, as inclusion of results from relatively simple analyses in synthetic works may be perfectly justified.

Third, while GLM analysis can generally be used to good effect for basic 2 × 2 contingency table analyses, it is only with substantial additional effort that one can extract the basic biological information that the LM analysis generates directly. To our surprise, no general biostatistics source that we consulted discussed how to generate a standard error for the difference in probability between two groups from a fitted GLM analysis. To this end, for those interested in implementing the linear approximation of the sampling error in the difference in probabilities (delta method SE) in the case of the 2 × 2 contingency table analysis, we provide a function in the form of a GitHub R package named ProbUncertainty (Nagel et al., 2024b; https://github.com/rebebba/ProbUncertainty). We also provide functions to calculate the score and profile likelihood CIs.

5 | CONCLUSIONS

We have confirmed the consensus in the biostatistics literature that GLMs are well suited to the analysis of 2 × 2 contingency table data. However, our results highlight that substantial care is needed for their application and interpretation. Little consideration has previously been given to the fact that the GLM does not readily output uncertainty in differences between groups in terms of probabilities, despite the fact that such information might correspond best to how biologists formulate research questions. We have illustrated ways to rectify this issue. Perhaps more surprising to many readers, we found that the direct inference of differences in probability and associated uncertainly (SEs and CIs) generated by LM analyses perform well in all circumstances where the basic chi-squared test can be recommended. This realisation, that the LM analysis is generally reasonable and directly yields the kind of information most people want to know, should be helpful when designing instruction in introductory biostatistics courses and could be broadly useful to researchers. Finally, while statisticians have been industrious in inventing alternatives to the basic Pearson chi-squared test, we found no circumstances where the commonly recommended alternatives unambiguously outperformed the classical test (or the LM analysis).

AUTHOR CONTRIBUTIONS

Michael B. Morrissey and Graeme D. Ruxton conceived the ideas and designed the methodology. Michael B. Morrissey and Rebecca Nagel drafted the simulation code. Rebecca Nagel ran the final versions of all simulations, made final versions of all graphs and led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

ACKNOWLEDGEMENTS

Rebecca Nagel was supported by a German Research Foundation grant (515410943). Michael B. Morrissey was funded by a Royal Society University Research Fellowship.

CONFLICT OF INTEREST STATEMENT

The authors declare no competing interests.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.14318.

DATA AVAILABILITY STATEMENT

All data were simulated. Code for the simulations, analyses and figures presented in this manuscript can be found on GitHub (https://github.com/rebebba/ProbUncertainty_MSCode) and Zenodo (https://zenodo.org/doi/10.5281/zenodo.10807611). The code and a worked example for the introduced R package can also be found on GitHub (https://github.com/rebebba/ProbUncertainty) and Zenodo (https://zenodo.org/doi/10.5281/zenodo.10807613).

ORCID

Rebecca Nagel https://orcid.org/0000-0002-2925-1028

Michael B. Morrissey https://orcid.org/0000-0001-6209-0177

REFERENCES

Agresti, A. (1992). A survey of exact inference for contingency tables. *Statistical Science*, *7*(1), 131–153. https://doi.org/10.1214/ss/1177011454

Agresti, A. (2007). *An introduction to categorical data analysis*. JohnWiley & Sons, Inc.

Alan, A. (2013). *Categorical data analysis* (3rd ed.). John Wiley & Sons.

Albert, A. (2017). Biostatistics: Facing the interpretation of 2 × 2 tables. *Journal of the Belgian Society of Radiology*, *101*, 1–5. https://doi.org/10.5334/jbr-btr.1399

Altman, N., & Krzywinski, M. (2017). Points of significance: Tabular data. *Nature Methods*, *14*(4), 329–330. https://doi.org/10.1038/NMETH.4239

Aragon, T. J. (2020). *epitools: Epidemiology Tools*. [Computer software version 0.5-10.1]. https://cran.r-project.org/package=epitools

Barnard, G. A. (1945). A new test for 2 × 2 tables. *Nature*, *156*(3954), 177. https://doi.org/10.1038/156177a0

Berger, R. L., & Boos, D. D. (1994). P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, *89*(427), 1012–1016. https://doi.org/10.1080/01621459.1994.10476836

Berkson, J. (1978). In dispraise of the exact test. *Journal of Statistical Planning and Inference*, *2*(1), 27–42. https://doi.org/10.1016/0378-3758(78)90019-8

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution*, *24*(3), 127–135. https://doi.org/10.1016/j.tree.2008.10.008

Boschloo, R. D. (1970). Raised conditional level of significance for the 2 × 2-table when testing the equality of two probabilities. *Statistica Neerlandica*, *24*(1), 1–9. https://doi.org/10.1111/j.1467-9574.1970.tb00104.x

Crawley, M. J. (2012). *The R book*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118448908

Dunn, P. K., & Smyth, G. K. (2018). *Generalized linear models with examples in R*. Springer Nature.

Dytham, C. (2011). *Choosing and using statistics: A biologist's guide*. Wiley-Blackwell.

Fagerland, M. W., Lydersen, S., & Laake, P. (2017). *Statistical analysis of contingency tables*. Taylor & Francis Group, LLC.

Fay, M. P., & Hunsberger, S. A. (2021). Practical valid inferences for the two-sample binomial problem. *Statistics Surveys*, *15*, 72–110. [Computer software version 1.6.8]. https://doi.org/10.1214/21-SS131

Fisher, R. A. (1934). *Statistical methods for research workers* (5th ed.). Oliver & Boyd.

Gallis, J. A., & Turner, E. L. (2019). Relative measures of association for binary outcomes: Challenges and recommendations for the global health researcher. *Annals of Global Health*, *85*(1), 1–12. https://doi.org/10.5334/aogh.2581

Grant, R. L. (2014). Converting an odds ratio to a range of plausible relative risks for better communication of research findings. *BMJ*, *348*, f7450. https://doi.org/10.1136/bmj.f7450

Halvorson, M. A., McCabe, C. J., Kim, D. S., Cao, X., & King, K. M. (2021). Making sense of some odd ratios: A tutorial and improvements to present practices in reporting and visualizing quantities of interest for binary and count outcome models. *Psychology of Addictive Behaviors*, *36*(3), 284–295. https://doi.org/10.1037/adb0000669

Hauck, W. W., & Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association*, *72*(360a), 851–853. https://doi.org/10.1080/01621459.1977.10479969

Hwang, J. T. G., & Yang, M. C. (2001). An optimality theory for mid p-values in 2 × 2 contingency tables. *Statistica Sinica*, *11*(3), 807–826.

Kang, S.-H., Lee, Y., & Park, E. (2006). The sizes of the three popular asymptotic tests for testing homogeneity of two binomial proportions. *Computational Statistics and Data Analysis*, *51*(2), 710–722. https://doi.org/10.1016/j.csda.2006.03.006

Kaplan, D. T. (2017). *Statistical modeling: A fresh approach*. Project MOSAIC Books. https://dtkaplan.github.io/SM2-bookdown/preface-to-this-electronic-version.html

Kempthorne, O. (1979). In dispraise of the exact test: Reactions. *Journal of Statistical Planning and Inference*, *3*(3), 199–213. https://doi.org/10.1016/0378-3758(79)90012-0

Lancaster, H. O. (1961). Significance tests in discrete distributions. *Journal of the American Statistical Association*, *56*(294), 223–234. https://doi.org/10.1080/01621459.1961.10482105

Lever, J., Krzywinski, M., & Altman, N. (2016). Points of significance: Logistic regression. *Nature Methods*, *13*(7), 541–542. https://doi.org/10.1038/nmeth.3904

Lydersen, S., Fagerland, M. W., & Laake, P. (2009). Recommended tests for association in 2×2 tables. *Statistics in Medicine*, *28*, 4267–4278. https://doi.org/10.1002/sim.3531

Lynch, M., & Walsh, B. (1998). Appendix 1: Expectations, variances, and Covariances of compound variables. In *Genetics and analysis of quantitative traits* (pp. 807–822). Sinauer Associates, Inc.

Nagel, R., Ruxton, G. D., & Morrissey, M. B. (2024a). Data from: Classical tests, linear models, and their extensions for the analysis of 2x2 contingency tables. https://doi.org/10.5281/zenodo.10807612

Nagel, R., Ruxton, G. D., & Morrissey, M. B. (2024b). *Classical tests, linear models, and their extensions for the analysis of 2x2 contingency tables: ProbUncertainty R package* [Computer software version 1.0]. https://doi.org/10.5281/zenodo.10807614

Newcombe, R. G. (1998). Interval estimation for the difference between independent proportions: Comparison of eleven methods. *Statistics in Medicine*, *17*(8), 873–890. https://doi.org/10.1002/(SICI)1097-0258(19980430)17:8<873::AID-SIM779>3.0.CO;2-I

Orme, J. G., & Combs-Orme, T. (2009). Regression with a dichotomous dependent variable. In *Multiple regression with discrete dependent variables* (pp. 30–90). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195329452.003.0005

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, *50*(302), 157–175. https://doi.org/10.1080/14786440009463897

R Core Team. (2023). *R: A language and environment for statistical computing* [Computer software version 4.2.3]. R Foundation for Statistical Computing. http://www.r-project.org/

Ramos, M. R., Oliveira, M. M., Borges, J. G., & McDill, M. E. (2015). Statistical models for categorical data: Brief review for applications in ecology. *AIP Conference Proceedings*, *1648*, 840015. https://doi.org/10.1063/1.4913055

Rosenthal, M. F., Gertler, M., Hamilton, A. D., Prasad, S., & Andrade, M. C. B. (2017). Taxonomic bias in animal behaviour publications. *Animal Behaviour*, *127*, 83–89. https://doi.org/10.1016/j.anbehav.2017.02.017

Routledge, R. D. (1994). Practicing safe statistics with the mid-p. *Canadian Journal of Statistics*, *22*(1), 103–110. https://doi.org/10.2307/3315826

Ruxton, G. D., & Neuhäuser, M. (2010). Good practice in testing for an association in contingency tables. *Behavioral Ecology and Sociobiology*, *64*(9), 1505–1513. https://doi.org/10.1007/s00265-010-1014-0

Scherer, R. (2018). *PropCIs: Various Confidence Interval Methods for Proportions* [Computer software version 0.3-0]. https://cran.r-project.org/package=PropCIs

Schwartz, L. M., Woloshin, S., & Welch, H. G. (1999). Misunderstandings about the effects of race and sex on physicians' referrals for cardiac catheterization. *The New England Journal of Medicine*, *341*, 279–283. https://doi.org/10.1056/NEJM199907223410411

Seltman, H. J. (2018). *Experimental design and analysis*. Carnegie Mellon University. https://www.stat.cmu.edu/~hseltman/309/Book/

Stephens, P. A., Buskirk, S. W., & del Rio, C. M. (2007). Inference in ecology and evolution. *Trends in Ecology & Evolution*, *22*(4), 192–197. https://doi.org/10.1016/j.tree.2006.12.003

Teder, T., & Kaasik, A. (2023). Early-life food stress hits females harder than males in insects: A meta-analysis of sex differences in environmental sensitivity. *Ecology Letters*, *26*, 1419–1431. https://doi.org/10.1111/ele.14241

Tutz, G. (2012). Binary regression: The logit model. In *Regression for categorical data* (pp. 29–50). Cambridge University Press. https://doi.org/10.1017/CBO9780511842061

Venzon, D. J., & Moolgavkar, S. H. (1988). A method for computing profile-likelihood-based confidence intervals. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *37*(1), 87–94.

Ver Hoef, J. M. (2012). Who invented the delta method? *American Statistician*, *66*(2), 124–127. https://doi.org/10.1080/00031305.2012.687494

Whitlock, M. C., & Schluter, D. (2009). *The analysis of biological data*. Roberts and Company Publishers.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22*(158), 209–212. https://doi.org/10.2307/2276774

Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statistical Society*, *1*(2), 217. https://doi.org/10.2307/2983604

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Figure A1. As for Figure 1, but with an unbalanced design.

Figure A2. As for Figure 1, but with a type 3 experimental design.

Figure A3. As for Figure 1, but with an unbalanced type 3 experimental design.

Figure A4. As for Figure 2, but with a type 3 experimental design.

Figure A5. As for Figure 3, but with a type 3 experimental design.

Figure A6. As for Figure 4, but with an unbalanced design.

Figure A7. As for Figure 4, but with a type 3 experimental design.

Figure A8. As for Figure 4, but with an unbalanced type 3 experimental design.

Figure A9. As for Figure 5, but with an unbalanced design.

Figure A10. As for Figure 5, but with a type 3 experimental design.

Figure A11. As for Figure 5, but with an unbalanced type 3 experimental design.

Figure A12. As for Figure 6, but with an unbalanced design.

Figure A13. As for Figure 6, but with a type 3 experimental design.

Figure A14. As for Figure 6, but with an unbalanced type 3 experimental design.

How to cite this article: Nagel, R., Ruxton, G. D., & Morrissey, M. B. (2024). Classical tests, linear models and their extensions for the analysis of 2 × 2 contingency tables. *Methods in Ecology and Evolution*, *00*, 1–13. https://doi.org/10.1111/2041-210X.14318