ELSEVIER

Journal of Clinical Epidemiology 61 (2008) 295–300

Journal of Clinical Epidemiology

Pisces did not have increased heart failure: data-driven comparisons of binary proportions between levels of a categorical variable can result in incorrect statistical significance levels

Peter C. Austin[a,b,c,*], Meredith A. Goldwasser[a]

[a]*Institute for Clinical Evaluative Sciences, Toronto, Ontario, Canada*
[b]*Department of Public Health Sciences, University of Toronto, Toronto, Ontario, Canada*
[c]*Department of Health Policy, Management and Evaluation, University of Toronto, Toronto, Ontario, Canada*

Accepted 28 May 2007

Abstract

Objective: We examined the impact on statistical inference when a χ^2 test is used to compare the proportion of successes in the level of a categorical variable that has the highest observed proportion of successes with the proportion of successes in all other levels of the categorical variable combined.

Study Design and Setting: Monte Carlo simulations and a case study examining the association between astrological sign and hospitalization for heart failure.

Results: A standard χ^2 test results in an inflation of the type I error rate, with the type I error rate increasing as the number of levels of the categorical variable increases. Using a standard χ^2 test, the hospitalization rate for Pisces was statistically significantly different from that of the other 11 astrological signs combined ($P = 0.026$). After accounting for the fact that the selection of Pisces was based on it having the highest observed proportion of heart failure hospitalizations, subjects born under the sign of Pisces no longer had a significantly higher rate of heart failure hospitalization compared to the other residents of Ontario ($P = 0.152$).

Conclusions: Post hoc comparisons of the proportions of successes across different levels of a categorical variable can result in incorrect inferences.

Keywords: Chi-squared test; Maximal proportion; Type I error rate; Contingency table; Astrology; Significance testing

1. Introduction

Chi-square (χ^2) tests are frequently used to assess the association between two categorical variables forming the rows and columns of a contingency table when the rows (columns) consist of independent groups. Given two categorical variables, one with r levels and the other with c levels, the χ^2 statistic for the associated contingency table is defined as:

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \left(O_{ij} - E_{ij}\right)^2 / E_{ij}$$

where O_{ij} denotes the observed frequency and E_{ij} denotes the expected frequency under the null hypothesis. Under the null hypothesis of no association, for large samples, the X^2 statistic has approximately a χ^2 distribution with $(r-1)(c-1)$ degrees of freedom [1].

Frequently, one of the variables is dichotomous, for example, success versus failure, and one is interested in comparing the proportion of successes across the levels of the other categorical variable. In this setting, when the other categorical variable has c levels, the X^2 statistic follows a χ^2 distribution with $(c-1)$ degrees of freedom under the null hypothesis. Several authors have studied the problem in which a continuous outcome variable is dichotomized and the prevalence of the resulting binary variable is compared across different levels of a categorical variable. It has been shown that if the threshold for dichotomizing the continuous variable is chosen to maximize the resulting X^2 statistic, then the nominal χ^2 reference distribution is incorrect [2–5]. Asymptotic distributions and small sample approximations of the distributions of maximally selected X^2 statistics under the null hypothesis have been provided for this situation [2–5]. The related problem of a maximally

* Corresponding author. Institute for Clinical Evaluative Sciences, G1 06, 2075 Bayview Avenue, Toronto, Ontario, M4N 3M5, Canada. Tel.: 416-480-6131; fax: 416-480-6048.

E-mail address: peter.austin@ices.on.ca (P.C. Austin).

0895-4356/08/$ – see front matter
doi: 10.1016/j.jclinepi.2007.05.007

What is new?

Key finding

- Using a standard χ^2 test to compare the prevalence of a dichotomous variable between the level of a categorical variable that has the highest observed prevalence of that variable with the prevalence of that dichotomous variable in the remaining levels of the categorical variable can result in misleading inferences.
- Treating post hoc comparisons in $k \times 2$ contingency tables as though they were specified a priori can result in inflated type I error rates.

What this adds to what was known?

- Prior research has shown that collapsing a continuous variable into a dichotomous variable so as to maximize the X^2 statistic from a resultant 2×2 contingency table can result in misleading inferences if a standard χ^2 test is used. We demonstrated that similar misleading conclusions can arise when a categorical variable is collapsed into a dichotomous variable based upon the observed prevalence of a dichotomous variable.

What is the implication, what should change now?

- Data-driven or post hoc comparisons should not be analyzed statistically as though they were a priori specified analyses.

selected X^2 statistic in the case when one variable is a binary variable and the variable that one wants to dichotomize is either a nominal or ordinal categorical variable has been examined by Boulesteix [6–7].

A related problem that has not, to the best of our knowledge, been examined in the statistical literature is the impact on significance testing when an observed $2 \times k$ contingency table is collapsed to a 2×2 contingency table based upon the level of the nominal variable that had the highest observed proportion of successes. For instance, assume that both a dichotomous variable denoting success versus failure and a k-level nominal variable are measured on each subject. The observed proportion of successes is computed for each of the k levels of the categorical variable. The initial $2 \times k$ contingency table is then collapsed to a 2×2 contingency table in which the proportion of successes is compared between that level with the highest observed proportion of successes and the remaining $k - 1$ levels considered in aggregate. In the context of either a significant or a nonsignificant overall χ^2 test, researchers may be motivated to make this type of post hoc comparison to identify a group of subjects that has a higher probability of success or failure compared to other subjects. However, the test of this comparison is usually conducted assuming a χ^2_1 distribution under the null hypothesis. The impact of this practice on statistical inference has not been examined in the statistical literature.

The objective of the current study was to determine the impact on statistical inference of collapsing a $2 \times k$ contingency table to a 2×2 contingency table when it is assumed that the resultant X^2 statistic follows a χ^2_1 distribution under the null hypothesis. We restrict our attention to the setting in which the k-level categorical variable is a nominal variable. The current paper is divided into three sections. First, we present the results of a simulation study in which we illustrate the inflation of the type I error rate when a $2 \times k$ contingency table is collapsed to a 2×2 contingency table. Second, we present data examining the association between astrological sign of birth and health outcomes. We propose that Monte Carlo simulations be used to obtain the empirical sampling distribution of the resultant X^2 statistic under the null hypothesis for the situation when a $2 \times k$ contingency table is collapsed to a 2×2 contingency table. Finally, we summarize our findings and discuss them within their context in the statistical literature.

2. Inflation of type I error when a $2 \times k$ contingency table is collapsed

In this section, we describe a set of Monte Carlo simulations that illustrate the inflation of the type I error when a $2 \times k$ contingency table is collapsed to a 2×2 contingency table, in which the proportion of successes is compared between the category with the highest observed proportion of successes and the remaining $k - 1$ categories. In the $2 \times k$ contingency table, the two rows correspond to the two levels of the dichotomous response variable indicating success or failure, whereas the k columns refer to the k levels of a nominal categorical variable. We consider only the case of a nominal, and not an ordinal, variable. Thus, there is no ordering inherent in the categorical variable. Furthermore, we do not consider the case of genetic data, in which a structured approach to dichotomization might be appropriate.

2.1. Monte Carlo simulations: design

We assumed that there was a k-level nominal variable and a binary variable, with 1 denoting success and 0 denoting failure. We allowed the following factors to vary in the design of our Monte Carlo simulations: the number of subjects, the number of levels in the nominal variable, the probability of success under the null hypothesis, and the distribution of the subjects across the k levels of the nominal variable. We now describe the choices for the different levels of these factors.

Because we are examining the inflation of the type I error rate, the data were generated such that the underlying probability of success (the event rate) was P_{success} in each of the k levels of the categorical variable. We allowed P_{success} to take on the following three values: 0.1, 0.3, and 0.5. We examined scenarios with three different sample sizes: 100, 1,000, and 10,000. We allowed the number of levels of the nominal variable to range from three to 10 in increments of one. Finally, we examined two different scenarios for the distribution of subjects across the k levels of the nominal variable. First, subjects were allocated to the k levels with equal probability (referred to as the ''equiprobable'' case). In this scenario, each subject had probability $1/k$ of being allocated to each of the k levels. Second, given k levels, subjects were allocated to the first $\lfloor k/2 \rfloor$ levels with probability $(1/k) - (1/2k)$, and to the second $\lfloor k/2 \rfloor$ levels with probability $(1/k) + (1/2k)$ (referred to as the ''nonequiprobable'' case) (here $\lfloor \rfloor$ denotes the floor function). If k was odd, then subjects were allocated to the kth level with probability $1/k$. The probabilities of allocation to different levels are described further in Table 1. We thus examined 144 different scenarios: three sample sizes × three probabilities of success × eight levels of the nominal variable × two cases for probabilities of assignment to the different levels of the nominal variable.

Within a given scenario, subjects were randomly allocated to the k levels of the nominal variable. Successes and failures were randomly generated for each subject from a Bernoulli distribution with parameter P_{success}. Thus, data were randomly generated so that the proportion of successes was the same across all k categories of the categorical variable. This allowed us to generate data such that the null hypothesis was true. The observed proportion of successes was then computed within each of the k levels of the categorical variables. The $2 \times k$ contingency table was collapsed to a 2×2 contingency table in which the proportion of successes in the category with the highest proportion of successes was compared with the proportion of successes in all other categories combined. In the event of ties, in which more than one category simultaneously had the highest proportion of successes, the proportion of successes in these categories combined was compared with the proportion of successes in the remaining categories considered in aggregate. We computed the X^2 statistic associated with this collapsed 2×2 contingency table. The statistical significance of the resulting X^2 statistic was evaluated using the χ^2_1 distribution. The null hypothesis was rejected when the significance level was less than 0.05. The above process was repeated 10,000 times, and the estimated type I error rate was the proportion of times that the null hypothesis was rejected over the 10,000 simulations.

Table 1
Distribution of subjects across the levels of the categorical variable under the nonequiprobable distribution

Number of levels for the categorical variable (k)	Probabilities of allocation of subjects to the different levels of the categorical variable
3	Level 1: 1/6; level 2: 3/6 Level 3: 2/6
4	Levels 1–2: 1/8; levels 3–4: 3/8
5	Levels 1–2: 1/10; levels 3–4: 3/10 Level 5: 2/10
6	Levels 1–3: 1/12; levels 4–6: 3/12
7	Levels 1–3: 1/14; levels 4–6: 3/14 Level 7: 2/14
8	Levels 1–4: 1/16; levels 5–8: 3/16
9	Levels 1–4: 1/18; levels 5–8: 3/18 Level 9: 2/18
10	Levels 1–5: 1/20; levels 6–10: 3/20

2.2. Monte Carlo simulations: results

The relationship between k and the empirical type I error rate that occurred when a χ^2_1 distribution was used to test the statistical significance of a collapsed 2×2 table is described in Fig. 1, with one panel for each of the three sample sizes. A horizontal line has been added to each panel, denoting a type I error rate of 0.05. Several observations are apparent from examining Fig. 1. First, one observes that the inflation in type I error rates tended to increase as k increased. Second, one observes that the inflation of the type I error rate tended to be smaller when the overall event rate was 0.3, compared to when it was 0.1. Similarly, the inflation in the type I error rate tended to be lower when the overall event rate was 0.5, compared to when it was 0.3. Third, that for a given k and for a fixed event rate, the equiprobable distribution of subjects in the levels of the nominal variable tended to result in greater inflation of the type I error rate than did the nonequiprobable distribution. Fourth, one observes that the rate of increase in the inflation of the type I error rate with increasing k tended to be greater when assignment probabilities were equiprobable compared to when they were nonequiprobable. We repeated the above simulations using a likelihood ratio χ^2 statistic [1] and Fisher's Exact Test [1] and observed qualitatively similar inflation of the type I error rates.

3. Case study

We provide an illustration of our findings in a setting in which most people would assume that the null hypothesis should be true: the association between astrological sign and hospitalization for a specific diagnosis. The Registered Person's Database (RPDB) contains basic demographic data on all residents of Ontario, Canada. We examined all residents of Ontario who were between the ages of 18 and 100 years in the year 2000, and who were alive on their birthday in 2000. The Canadian Institutes for Health Information (CIHI) discharge abstract database (DAD) contains information on all hospital discharges in the province of Ontario. Among the demographic and clinical information contained in the

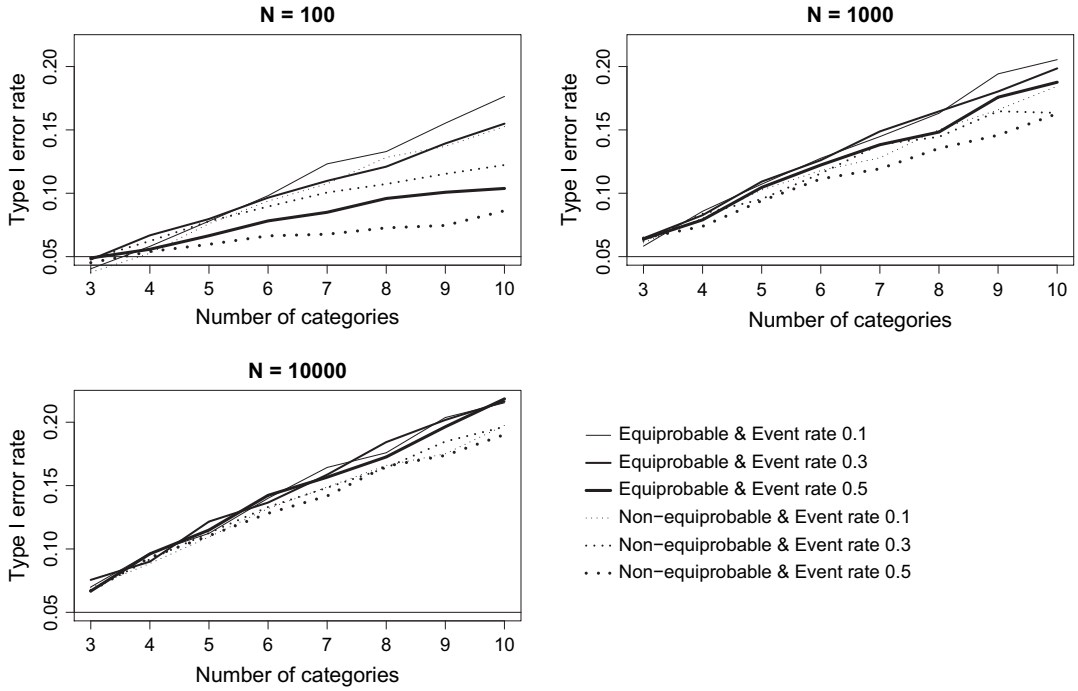

Fig. 1. Inflation of the type I error rate when a $k \times 2$ contingency table is collapsed to a 2×2 table.

DAD is the clinical diagnosis that was most responsible for the patient's hospitalization. Prior to 2002, these diagnoses were classified using the ICD-9 coding system. The RPDB and the CIHI DAD can be linked deterministically using patients' encrypted health card numbers.

Using the RPDB, we determined the astrological sign under which each resident of Ontario was born. We then determined the proportion of residents hospitalized in the year following their birthday in 2000 with a diagnosis of congestive heart failure (CHF), classified using an ICD-9 code of 428. These data have been analyzed more fully in a related paper [8]. The number and proportion of Ontario residents born under each astrological sign who were hospitalized for CHF is reported in Table 2. A χ^2 test was used to compare the proportion of residents hospitalized for CHF across the 12 astrological signs. The overall χ^2 test was significant ($P = 0.0017$).

Table 2
Proportion of residents hospitalized for congestive heart failure (CHF) according to astrological sign of birth in Ontario, Canada

		Hospitalized for CHF
Astrological sign	Number of residents	*N* (%)
Aquarius	856,301	1,433 (0.167%)
Aries	888,348	1,476 (0.166%)
Cancer	917,553	1,496 (0.163%)
Capricorn	844,635	1,343 (0.159%)
Gemini	937,615	1,553 (0.166%)
Leo	903,009	1,497 (0.166%)
Libra	897,503	1,350 (0.150%)
Pisces	893,332	1,522 (0.170%)
Sagittarius	846,813	1,277 (0.151%)
Scorpio	850,128	1,297 (0.153%)
Taurus	918,512	1,534 (0.167%)
Virgo	921,196	1,445 (0.157%)

Residents born under the astrological sign of Pisces had the highest probability of hospitalization for CHF. A χ^2 test was used to compare the probability of hospitalization for CHF for residents born under Pisces to that of residents born under all other astrological signs combined. The X^2 statistic associated with this test was 4.9384. Using a conventional χ^2_1 test, the P-value associated with this test was 0.026, indicating that residents born under the sign of Pisces were significantly more likely to be hospitalized with a diagnosis of CHF than were residents born under the remaining 11 astrological signs combined.

We then determined the empirical sampling distribution of the maximal proportion X^2 statistic under the null hypothesis. This was done using the R program reported in Appendix A. Using the empirical sampling distribution of the maximal proportion χ^2 statistic, one obtains a P-value of 0.1523. These results are summarized in Table 3. Therefore, after accounting for the fact that one has collapsed a 2×12 contingency table to a 2×2 contingency table based on the maximal proportion, residents born under the sign of Pisces appear to have the same probability of hospitalization for CHF as do residents born under the remaining 11 astrological signs combined.

4. Discussion

The objective of the current study was to examine the impact on the type I error rate when a $2 \times k$ contingency

Table 3
Statistical significance of the association between Pisces and hospitalization for CHF

Statistical method	Assumption	*P*-value
Conventional chi-square test comparing Pisces with all other zodiac signs in aggregation.	Decision to compare Pisces to other signs combined was made a priori—this assumption was not satisfied (Pisces was selected because it had highest rate of hospitalization for CHF).	0.026
Empirical sampling distribution of chi-squared statistic for comparing astrological sign with the highest rate of hospitalization to all other astrological signs combined.	Decision to compare Pisces to other signs was made after observing the data. Pisces was selected because it had the highest probability of hospitalization for CHF.	0.1523

table is collapsed to a 2×2 table, and a X^2 statistic is used to compare the proportion of successes between subjects in the category with the highest observed proportion of successes and subjects in all other categories combined.

The current study adds to the growing body of statistical literature that demonstrates that data-driven methods of analysis can lead to misleading inferences. Given a binary exposure variable and a continuous response variable, dichotomizing the continuous variable so as to maximize the resulting X^2 statistic can result in inflated type I error rates and incorrect inference about the association if the test statistic is assumed to follow a χ^2 distribution with one degree of freedom [2–5]. Similarly, this has been shown for the case where an ordinal or nominal categorical variable is dichotomized to maximize the X^2 statistic [6–7]. A seemingly unrelated issue is the use of automated variable selection methods such as forward variable selection or backward variable elimination for selecting regression models. Using these, and other related methods, repeated significance testing is used to select the variable for inclusion in the final regression model. Automated variable selection has been shown to result in regression models in which coverage rates for confidence intervals are much lower than the nominal values [9]. Similarly, the estimates of the residual variance are biased downward [10], leading to significance levels that are biased low. Thus, our finding adds to the growing body of statistical literature that data-driven methods of analysis can lead to incorrect statistical inferences. Additionally, automated variable selection methods can result in nonreproducible regression models [11–12]. Finally, dichotomizing a continuous covariate in a logistic regression model at the sample median can result in inflation of the type I error rate [13].

In conclusion, collapsing a $2 \times k$ contingency table into a 2×2 contingency table in which the proportion of successes among subjects in the level of the category with the highest proportion of successes is compared with the proportion of successes with subjects in all other categories combined, results in an inflation of the type I error rate when it is assumed that the associated X^2 statistic follows a χ^2 distribution with one degree of freedom.

Acknowledgments

The Institute for Clinical Evaluative Sciences is supported in part by a grant from the Ontario Ministry of Health and Long Term Care. The opinions, results, and conclusions are those of the authors, and no endorsement by the Ministry of Health and Long-Term Care or by the Institute for Clinical Evaluative Sciences is intended or should be inferred. This research was supported in part by a Discovery grant from the Natural Sciences and Engineering Research Council. Dr. Austin is supported in part by a New Investigator award from the Canadian Institutes of Health Research.

Appendix A

R program for determining the empirical sampling distribution of the maximal proportion chi-squared statistic

```
# Number of subjects at each level of the categorical
# variable are
# determined by the empirical setting.
#
# Computes the chi-squared statistic associated with
# collapsing the kx2 table
# to a 2x2 table based upon the category with the highest
# observed
# proportion.
set.seed(16062006)
# Seed set for June 16, 2006. This will ensure that the
# empirical distribution is reproducible.
n.levels <- 12
# Number of levels of the categorical variable.
# In our example, there are 12 astrological signs.
N.cat <- c(856301,
  888348,
  917553,
  844635,
  937615,
  903009,
  897503,
  893332,
  846813,
  850128,
  918512,
  921196)
# Number of subjects in each of the 12 astrological
# signs.
event.rate <- 0.0016
# Event rate in each group under the null hypothesis.
```

```
# This is the overall event rate in the sample.
chisq.max <- NULL
# Vector for storing values of the Chi-squared statistic.
for (i in 1:100000){
# Use 100,000 simulated samples to determine empirical
# sampling distribution.
    outcomes <- rbinom(n.levels,N.cat,event.rate)
    # Randomly generate binomial number of events in
# each of the levels of the
    # categorical variable.
    p <- outcomes/N.cat
    # Observed success probability in each category.
    tot.outcomes <- sum(outcomes)
    # Total number of successes across the groups
    group.id <- 1:n.levels
    # Labels for the groups.
    max.group <- group.id[p==max(p)]
    # Group with the highest proportion of successes.
    outcomes.max <- sum(outcomes[max.group])
    # Number of successes in group(s) with highest
# observed proportion of successes.
    max.matrix <- matrix(0,2,2)
    # Matrix for collapsing to a 2x2 table.
    max.matrix[1,1] <- outcomes.max
    max.matrix[1,2] <- tot.outcomes - outcomes.max
    max.matrix[2,1] <- sum(N.cat[max.group]) -
outcomes.max
    # Allow for more than one group to have maximum
# proportion.
    max.matrix[2,2] <- sum(N.cat) - max.matrix[1,1] -
max.matrix[1,2] - max.matrix[2,1]
    p.max <- chisq.test(max.matrix)$p.value
    chisq.max[i] <- chisq.test(max.matrix)$statistic
}
p.value <- (length(chisq.max[chisq.max > 4.9384]))/
(length(chisq.max))
# 4.9384 was the observed Chi-squared statistic when
# the 2x12 contingency table was collapsed to a 2x2
# table.
cat("P-value for Pisces and CHF: ",p.value,file="pisces.
emp.r.out",fill=T, append=T)
```

References

[1] Agresti A. An introduction to categorical data analysis. New York, NY: John Wiley & Sons; 1996.

[2] Miller R, Siegmund D. Maximally selected chi square statistics. Biometrics 1982;38:1011–6.

[3] Halpern J. Maximally selected chi square statistics for small samples. Biometrics 1982;38:1017–23.

[4] Koziol JA. On maximally selected chi-square statistics. Biometrics 1991;47:1557–61.

[5] Betensky RA. Maximally selected χ^2statistics for $k\times2$ tables. Biometrics 1999;55:317–20.

[6] Boulesteix A-L. Maximally selected chi-square statistics for ordinal variables. Biom J 2006;48:451–62.

[7] Boulesteix A-L. Maximally selected chi-square statistics and binary splits of nominal variables. Biom J 2006;48:838–48.

[8] Austin PC, Mamdani MM, Juurlink DN, Hux JE. Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health. J Clin Epidemiol 2006;59:964–9.

[9] Hurvich CM, Tsai C-L. The impact of model selection on inference in linear regression. Am Stat 1990;44:214–7.

[10] Copas JB, Long T. Estimating the residual variance in orthogonal regression with variable selection. Statistician 1991;40:51–9.

[11] Austin PC, Tu JV. Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. J Clin Epidemiol 2004;57:1138–46.

[12] Austin PC, Tu JV. Bootstrap methods for developing predictive models in cardiovascular research. Am Stat 2004;58:131–7.

[13] Austin PC, Brunner LJ. Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses. Stat Med 2004;23:1159–78.