Hypothesis Testing Methods For Relational Data

Submitted for Blind Review

Abstract

The new tasks and methods addressed by knowledge discovery algorithms often require the development of new methods for testing statistical hypotheses. Knowledge discovery in relational data has previously been shown to pose unique challenges for conventional hypothesis tests because of the interdependence among data instances. Conventional tests systematically underestimate p-values, resulting in large numbers of erroneous dependencies. We formally describe when these errors occur, and we develop two specialized methods of testing for statistical independence in relational data. We show that these methods completely remove the bias of conventional tests and provide practical advice for future algorithm development.

1. Introduction

Many of the key algorithms for knowledge discovery quantify and characterize associations among variables in data. For example, algorithms for learning the structure of Bayesian networks infer the existence of conditional dependencies among variables, algorithms for learning association rules discover relationships among purchases in massive market basket data sets, and algorithms for learning classification trees identify conditional associations between the class label and other variables. Although the precise data types and tasks differ, all of these algorithms have association learning at their core.

Any algorithm that quantifies assocations does so with some statistic. One common choice is the chi-square statistic χ^2, which approximates the multinomial distribution function for categorical data.[1] The statistic calculates the normalized squared deviation of observed frequencies from their expected values:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

χ^2 is widely applicable, and it is robust to changes in data size, cardinality, and class distribution.[14]

Many algorithms attempt to infer whether a given correlation is due to systematic association in the underlying population rather than mere random variation in the data sample. For example, Bayesian network learners infer which edges to include in the network, association rule algorithms attempt to filter out spurious rules, and classification tree algorithms decide when to stop adding structure. Since even modest amounts of variation in a data sample will produce a non-zero χ^2 value, these algorithms require a method for separating signal from noise.

The most common mechanism for doing so is a statistical hypothesis test, which estimates the probability p that a value at least as high as the observed value of the statistic would occur under a given null hypothesis. This null hypothesis most often specifies no association between the given variables (although other null hypotheses can be specified). If the p-value is smaller than a given level (commonly 0.05, or 0.01), the algorithm rejects the null hypothesis and concludes that the observed correlation is due to a systematic association in the population.

Innovation in knowledge discovery tasks and methods has frequently required new methods for hypothesis testing because the tasks or methods violate the assumptions of existing tests. For example, most knowledge discovery algorithms seach vast spaces of possible models, and this violates the assumption made by many tests that only a single hypothesis will be evaluated [8, 4]. In other cases, algorithms address new types of data that require fundamentally new types of hypothesis tests. For example, several recent papers have examined specialized methods for testing hypotheses about association rules [15, 6].

Many of the most recently developed algorithms for knowledge discovery focus on analyzing *relational* data. Such data sets explicitly represent multiple types of entities and the relations among those entities. One key insight of this work is that entities are not statistically independent, and that their interdependence can be exploited to improve statistical modeling. However, this interdependence has also raised concerns, because it violates a key assumption of nearly all hypothesis tests—that data instances are independent and identically distributed (IID).

In this work, we develop a systematic description of the relational data structures for which conventional tests systematically underestimate p-values. We devise specialized permutation tests and sampling algorithms that produce un-

[1] For the remainder of this paper, we will us χ^2 to refer to the test statistic, and X^2 to refer to its theoretical distribution

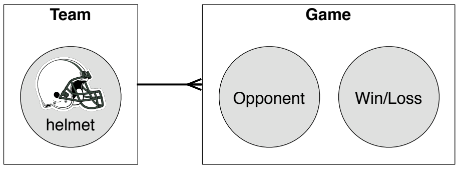


Figure 1. Entity-relationship diagram for a simple NFL data set.

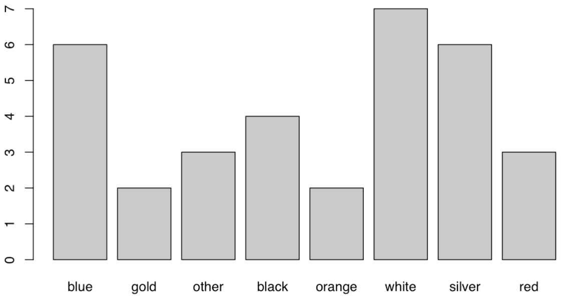


Figure 2. Distribution of helmet colors for NFL teams.

biased results. Finally, we characterize the tradeoffs among accuracy and statistical power of the different hypothesis testing methods.

1.1. Example

An example drawn from American football can demonstrate the biases in conventional tests when applied to relational data. Figure 1 depicts a simple relational database schema for the National Football League (NFL), the governing body for American football franchises. The league consists of 32 teams, with each team participating in 16 regular season games per year. In our small example, teams are characterized by a single attribute: the color of their protective helmets, which stays constant throughout the season. Games have two attributes: the opponent and the game outcome (won or lost). While "opponent" is actually a foreign key in database terminology, to avoid replication all games are associated with the home team only.

Table 1 is a contingency table tabulating the home wins and losses of each NFL teams according to the color of their helmets over a ten year span from 1998-2007. At seven degrees of freedom ((number of outcomes - 1) x (number of colors - 1) = 1 x 7 = 7), a conventional hypothesis test would infer that a χ^2 score of 33.81 is highly significant (p0.00001). To those who follow American football, this result may seem surprising, as most sports enthusiasts would consider helmet color to be independent of on-field performance. Changing helmet color would be expected to have little effect on performance of an existing team, and a new team's performance would not be expected to be predicted by their choice of helmet color.

Table 1. Contingency table for helmet color and wins/losses in the NFL, 1998–2007

helmet	wins	losses	*total*
white	467	493	960
red	252	228	480
orange	193	111	304
gold	180	140	320
blue	434	462	896
silver	494	466	960
black	295	345	640
other	205	275	480
total	2520	2520	5040

To better characterize the problem with a conventional test, consider the contingency table in Table 2. This table was generated by tabulating a variable that was assigned randomly with the same distribution of values as helmet color (a *permutation* of helmet color). Here, the permuted variable has an even more significant p-value than helmet color, which itself was highly significant. Note that the extreme p-values say nothing about the strength of effect; the the out-of-sample predictive power of this association is non-existent. Even so, how can this be?

Most analysts (and sports fans) would agree that neither the original nor the permuted variable is actually correlated with wins and losses. Something must be wrong with the statistic itself or with the hypothesis test. The p-values shown are calculated from the theoretical reference distribution of χ^2 at seven degrees of freedom. For our example,

Table 2. Contingency table for random team attribute and wins/loses in the NFL, 1998–2007

helmet	wins	losses	*total*
white	488	472	960
red	271	209	480
orange	144	112	256
gold	141	179	320
blue	522	438	960
silver	378	566	944
black	325	315	640
other	251	229	480
total	2520	2520	5040

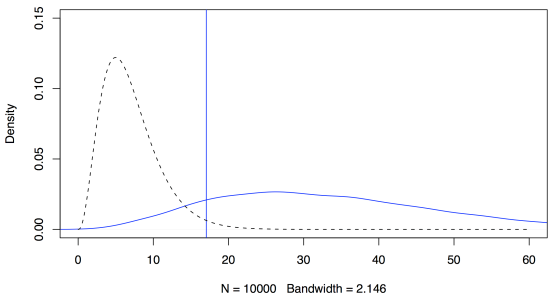


Figure 3. Empirical distribution for χ^2 statistic on NFL team attributes. The vertical line represents the value of the χ^2 statistic as calculated for the attribute helmet color. When compared against the theoretical reference distribution X^2 (dashed line), this value is significant with a p-value < 0.000001. When compared with the empirical distribution, the value is not significant.

though, the traditional X^2 reference distribution turns out to be inappropriate.

The curve in Figure 3 was constructed by creating thousands of permutations of helmet color, assessing their χ^2 values, and plotting the frequency of those values. This procedure for calculating a p-value empirically is called a permutation test and will be discussed below. The vertical line represents the χ^2 value of the helmet color attribute. When compared against this distribution (rather than the theoretical X^2), a χ^2 value of 33.81 is not at all unexpected under the null hypothesis of no association.

As we will see below, this effect is an example of a general problem that is *inherent* in essentially all relational data sets. In this work, we detail the precise nature of the dangers of using conventional hypothesis tests for knowledge discovery in relational data. Through simulation, we are able to explain the precise cause of issues raised in previous work on hypothesis testing [9, 10, 11], and identify the precise situations where conventional hypothesis tests are inadequate. Finally, we offer the first general-purpose solutions for conducting effective hypothesis tests using relational data sets.

2. Relational data

In recent years, statistical relational learning (SRL) has represented an exciting frontier in the data mining and knowledge discovery communities. Relational data consist of entities connected by links, where links encode relationships. Relational data sets are often represented in the form

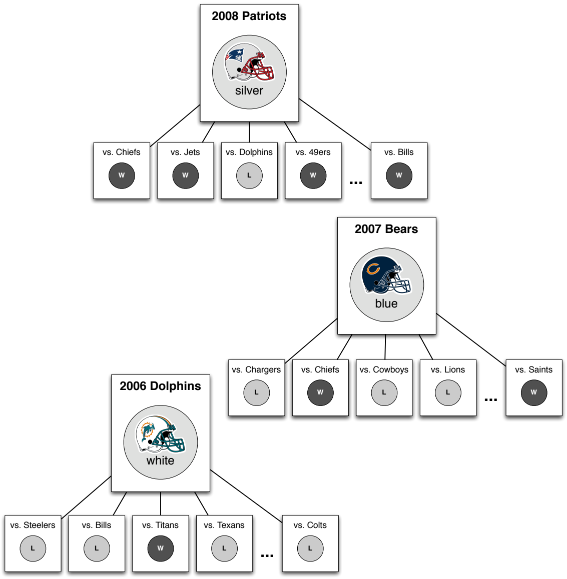
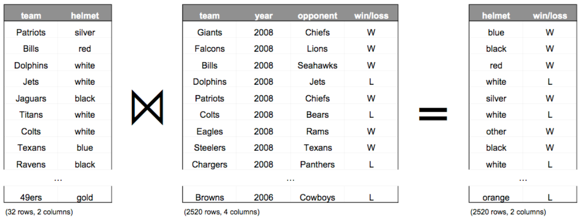


Figure 4. Relational representation of NFL data.



team	helmet
Patriots	silver
Bills	red
Dolphins	white
Jets	white
Jaguars	black
Titans	white
Colts	white
Texans	blue
Ravens	black
...	
49ers	gold

(32 rows, 2 columns)

⋈

team	year	opponent	win/loss
Giants	2008	Chiefs	W
Falcons	2008	Lions	W
Bills	2008	Seahawks	W
Dolphins	2008	Jets	L
Patriots	2008	Chiefs	W
Colts	2008	Bears	L
Eagles	2008	Rams	W
Steelers	2008	Texans	W
Chargers	2008	Panthers	L
...			
Browns	2006	Cowboys	L

(2520 rows, 4 columns)

=

helmet	win/loss
blue	W
black	W
red	W
white	L
silver	W
white	L
other	W
black	W
white	L
...	
orange	L

(2520 rows, 2 columns)

Figure 5. RDBMS representation of NFL data storage. When data from different tables is joined for processing, the relational structure (in this case, the one-to-many relationship between teams and games) is lost. The resulting table resembles an IID table of propositional data, but treating it as such can lead to erroneous significance estimates.

of a directed or undirected graph, where graph vertices and edges represent entities and links, respectively. Figure 4 depicts a small portion of the data graph for the NFL data set.

Traditionally, contingency tables were constructed from independent, identically distributed (IID) or propositional data sets. Hypothesis tests utilizing χ^2 were developed for this type of data. When we create a contingency table from relational data, we take the attributes associated with the endpoints of each link. This process can create dependence between instances in two ways. First, nodes that have a degree greater than one will source more than one entry in the table. By definition, instances born from the same node will have dependence.[5] We refer to this effect as attribute replication. For instance, in the NFL example, a single team's helmet color get replicated over one hundred times, leading to perfect dependence among those instances.

A second source of instance dependence stems from autocorrelation, also known as homophily. Autocorrelation is an association between the attribute values of nodes that share links with a common node or set of nodes. For example, game outcome is autocorrelated among the games played by a single team. The existence of an autocorrelated attribute signifies a dependence between the link structure and that attribute; as a result, the attribute values of neighboring nodes are not independent.

Propositional data sets may also exhibit instance dependence. Relational database management systems (RDBMSs) often store data in normalized form to ensure data integrity and compactness. When data is exported for analysis, the relevant tables are joined together form a record set that is then fed to a learning algorithm.

Figure 5 illustrates the join process for the NFL data set. In this example, the "helmet table" (which stores the helmet color attribute for each team in its 32 rows) is joined to a "game table." As a results of the join, the rows of the helmet table are replicated for each game table row with a matching "team" field. The effect of this replication on the resulting record set is the same instance dependence described above for the relational case. This form of dependence can be quite insidious, as the new table contains no record of the relational structure that was used to create it. When tables such as the one in Figure 5 are fed to propositional algorithms, the data is mistakenly treated as IID.

The inaccuracy of using a conventional sampling distribution with non-independent events has been known for decades.[12, 2] In addition, recent work in relational learning has illustrated the effects of autocorrelation on the accuracy of χ^2 tests. [9, 11] Given the history in the literature, it may seem obvious that naively applying χ^2 tests to data with instance dependence can result in biases. Nevertheless, many algorithms ignore the issue of independence violations while conducting hypothesis tests to determine correlation.

The lack of attention paid to independence violations may be due to the fact that not all types of dependence will skew results. Different data sets encode different relationships between entities, and the severity of the errors associated with non-independence is a function of these relationships and how they interact with attribute values. In the following section, we identify the situations where replication and autocorrelation affect the sampling distribution of χ^2 in bipartite data.

3. Link structure, autocorrelation, and biased p-values

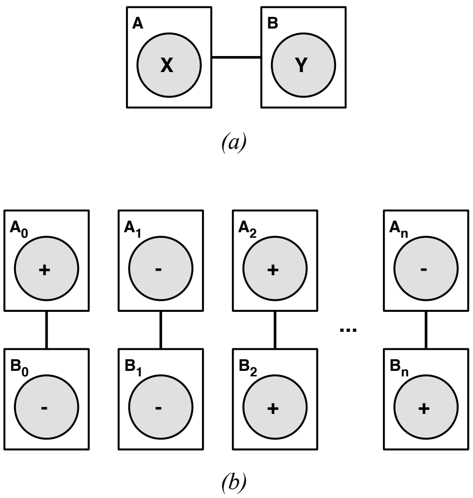
In propositional data, the x,y pairs that populate a contingency table come from a single data table, perhaps stored in the form of rows in a relational database. The attributes x and y are associated with the same experimental unit (e.g., a person's height and weight), and each unit is associated with a single row. Since the units (and rows representing them) are IID, and there is no hidden dependency encoded in the contingency table. Relational data is more complicated in terms of structure and dependence. Below, we examine the effects of different structural motifs on the distribution of the χ^2 statistic.

To isolate the effects of structure and dependence, we utilize a synthetic data generator. The generator works as follows: first, the graph structure is created in accordance with the degree distributions supplied as inputs. Once the structure is in place, attribute values are randomly assigned to each node according to a class distribution supplied as input. To create autocorrelated attributes on the entities of type B, a "latent" attribute is generated for each A object. The autocorrelated attributes are then drawn randomly for the B objects, conditioned on the values of the latent attribute for the A objects they connect.

3.1. One-to-one subgraphs

The first case we consider consists of simple object pairs. Figure 6a shows a schema for the data indicating a one-to-one relationship between entities A and B. The variables X and Y are statistically independent, indicated by the lack of an arrow connecting the two variables. Figure 6b gives an example data set corresponding to the schema.

While technically relational, each isolated subgraph contains only a single value of X and Y, and thus pairs of values are IID, satisfying the independence assumptions of the conventional χ^2 test. As a result, p-values supplied by the theoretical X^2 distribution for the appropriate degrees of freedom will be accurate.



(a)

(b)

Figure 6. Entity-relationship diagram for one-to-one data along with sample subgraphs.

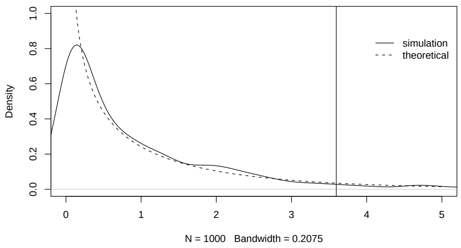


Figure 7. Distribution of χ^2 for synthetic one-to-one data.

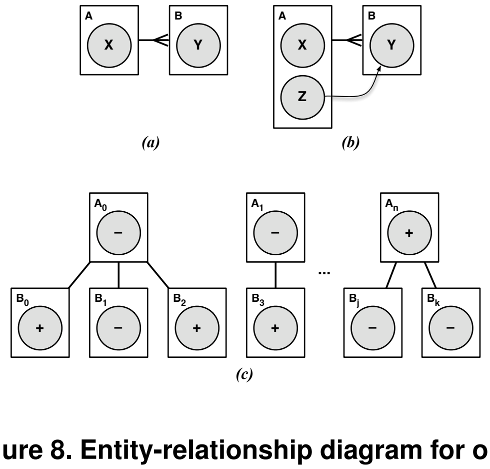


Figure 8. Entity-relationship diagram for one-to-many data along with sample subgraphs.

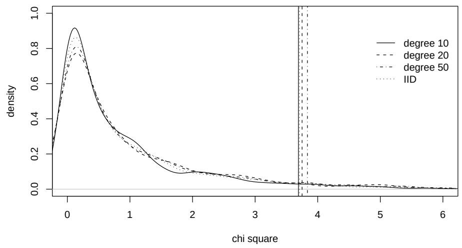


Figure 9. Distribution of χ^2 for synthetic one-to-many data with no autocorrelation.

3.2. One-to-many without autocorrelation

Bipartite data grouped in a one-to-many manner is the simplest form of truly relational data. Examples of data of this type include movie studios and the films they produce, journals and the articles they contain, etc. Figure 8a shows the schema with a one-to-many relationship between A and B and independent variables X and Y. Figure 8c shows a representative data set corresponding to this schema.

As noted previously, when data of this form are flattened to populate a contingency table, the values used to calculate frequencies for χ^2 are dependent. Whether this violation of the independence assumption affects the probability distribution depends on the relationship between links and the attributes on their endpoints. Problematic dependencies are reflected in the autocorrelation of the ”many” objects within individual subgraphs.

When autocorrelation is absent from the values of Y, then pairs of values can be considered independent. In this situation, again, the conventional χ^2 test is accurate.

3.3. One-to-many with autocorrelation

When the Y variable does exhibit autocorrelation, assessing correlation between X and Y becomes more complicated. Figure 8b shows the schema with a one-to-many relationship between A and B and autocorrelation among the values of Y produced by the latent variable Z on A. Note that X and Z are independent, as are X and Y. The relational structure of the data would look similar to Figure 8c, although the values of Y would be autocorrelated.

Here, the link structure is not independent of the attribute values that we care about, as a result, the units in the contingency table are not independent, and the distribution of the χ^2 statistic changes. This case matches the helmet color example detailed earlier, as well as the issues explored in by previous investigators.[9] Here, the null hypothesis H_0 dictates that any perceived correlation between X and Y (as

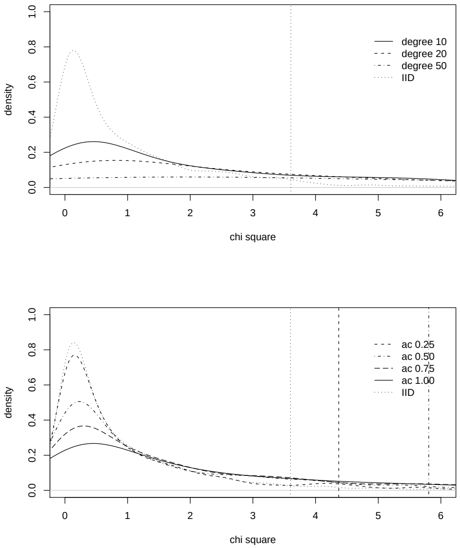


Figure 10. Distribution of χ^2 for synthetic one-to-many data for various degree distributions (top) and levels of autocorrelation (bottom). As degree and autocorrelation increase, the sampling distribution of χ^2 is shifted.

reflected by a high χ^2 value) is actually a statistical artifact of the same dependence that produces autocorrelation.

Depending on the degree distribution and level of autocorrelation, the distribution of the χ^2 statistic can be equivalent to one produced by multiplying contingency a table by a constant factor. As first detailed by Jensen and Neville, the higher the degree ("concentrated linkage", in their language) and autocorrelation, the more the sampling distribution is shifted. Figure 10 illustrates the effects of different levels of each on the empirically-derived distribution of the statistic.

It should be noted that many propositional data sets are actually "flattened" versions of data sets with this structure. Furthermore, depending on representation choices, it's not always obvious whether or not autocorrelation-producing dependencies exist.

3.4. Many-to-many without autocorrelation

Figure 11a shows the schema with a many-to-many relationship between A and B and independent variables X and Y. Figure 11d shows a representative data set corresponding to this schema.

This case is similar to the one-to-many without autocorrelation case. While a contingency table drawn from this data will certainly not be composed of independent observations, the lack of autocorrelation among the values of Y indicates that the distribution of χ^2 will not be affected; as a result, a conventional test is adequate.

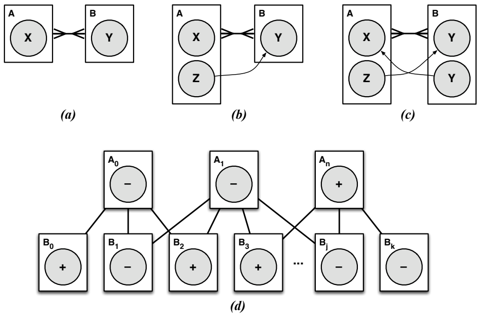


Figure 11. Entity-relationship diagram for many-to-many data along with sample subgraphs.

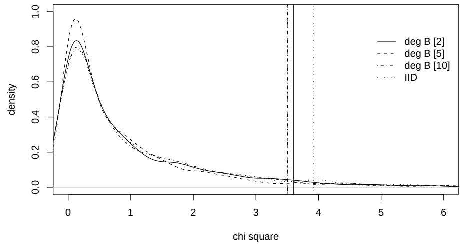


Figure 12. Distribution of χ^2 for synthetic many-to-many data.

3.5. Many-to-many with one-sided autocorrelation

Figure 11b shows the schema with a many-to-many relationship between A and B and autocorrelated values of Y.

This case is similar to the one-to-many case with autocorrelation, in that dependencies alter the distribution of χ^2. As seen in Figure 13 that the degree distribution and of the A entities and the dependency between attributes on the B entities are associated with changes to the reference distribution, but the degree distribution of the B entities does not affect the value distribution of the statistic.

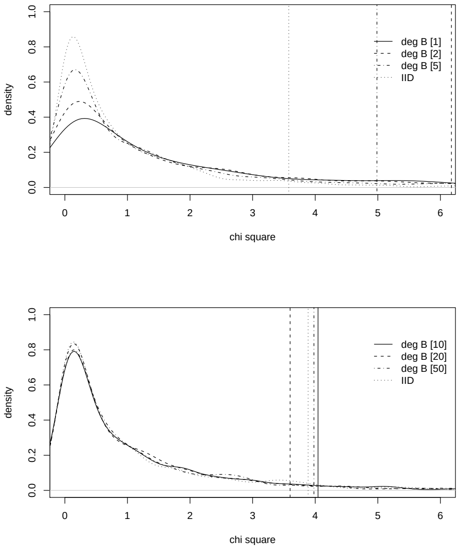


Figure 13. Distribution of χ^2 for synthetic many-to-many data for various degree distributions. For small degrees (top), the distribution of χ^2 resembles the one-to-many case shown in Figure 10. For large degrees, however, the distribution of χ^2 shifts back toward the propositional case.

3.6. Many-to-many with two-sided autocorrelation

Figure 11c shows the schema with a many-to-many relationship between A and B and autocorrelated values of both X and Y.

In this scenario, the two entity types are linked in a many-to-many relationship, with autocorrelation on both attributes. While less interesting in terms of its effect on the χ^2 distribution, this case illustrates an important point regarding the nature of autocorrelation among entity types and its relationship to correlation between entity types. Simply put, autocorrelation can exist on either side of a bipartite graph only in cases where there is correlation.

4. Solutions for hypothesis testing

In the previous section, we outlined several scenarios in which the use of a conventional χ^2 test to assess statistical significance will produce biased results. Here, we consider several alternative methods for conducting valid hypothesis tests to determine association in relational data sets. Each of our proposed methods modify the conventional χ^2 test in one of two ways: altering the reference distribution to which our test statistic is compared, or modifying the calculation of the test statistic itself and comparing it to a the theoretical X^2 probability distribution.

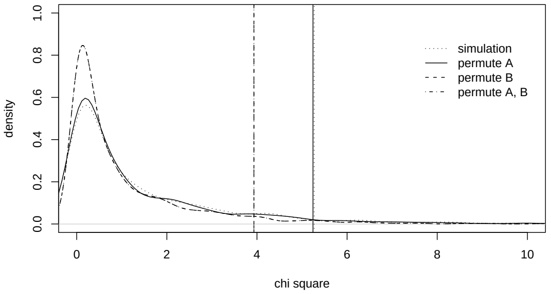


Figure 14. Using a permutation test, the distribution of χ^2 can be corrected under autocorrelation.

4.1. Permutation tests

Permutation tests (sometimes also called randomization tests) are a type of computationally-intensive method for conducting accurate hypothesis tests. [3, 7, 13]

Permutation tests work by creating several near-replicates of the data set, called pseudosamples. By calculating the test statistic on each pseudosample, an empirically-derived estimate of the probability distribution can be estimated. The pseudosamples are generating by randomly permuting the values of a given attribute among the entities in the data. Jensen, Neville, and Rattigan demonstrated that such a technique is effective for performing hypothesis tests on single linkage (one-to-many) data that exhibit autocorrelation, as in the case above.

The test works by modelling the distribution of the test statistic under the null hypothesis that any measurable correlation is an artifact of dependencies between instances (which are, in turn, reflected in the form of autocorrelation). By permuting the attributes on the A objects while holding those on the B objects constant, the level of autocorrelation among y attributes is preserved. In contrast, the randomization procedure effectively destroys any systematic association between connected x and y variables (as dictated by H_0). The distribution curve in figure x was generated using a permutation test. Specifically, since teams are connected to games in a one-to-many fashion, the values of the team attribute "helmet color" were permuted, while holding the "win/loss" attribute on games constant.

In addition to the one-to-many case, permutation tests can be applied to many-to-many data with one-sided autocorrelation as described in a case above. Finally, permutation tests can be effectively applied to propositional data, as an alternative to a conventional test when prior processing may have introduced latent dependencies through database joins.

4.2. Link sampling

Link sampling is a novel technique for accurate hypothesis testing in relation data. Rather than adjusting the reference distribution, link sampling works by modifying the calculation of the test statistic itself such that it will be correctly distributed with a X^2 distribution. Recall that the problem identified in the previous section stemmed for non-independence between attribute values when link endpoints are used to populate a contingency table. Using link sampling, we can "enforce" the independence assumption by constructing a contingency table out of independent links. To select the set of links for inclusion in the contingency table, we use the randomized greedy matching algorithm presented by Aronson et al.[1] This algorithm produces a *matching*, a set of edges which share no common vertices. While the algorithm as presented seeks to find a maximal matching, it can be trivially adapted to select a set of independent links of a given target size (assuming that one exists).

To gauge correlation using link sampling, the contingency table is populated with a subset of the links such that no two links in the subset share an endpoint. If the assumption holds within the data that the attribute values of any node are conditionally independent of others in the graph given those of their neighbors, the x,y pairs that fill the contingency table will be independent as well. Since the independence assumption is no longer violated, a χ^2 statistic calculated from the subset contingency table will be distributed with the theoretical X^2.

The link sampling technique is applicable to any form of relational data, regardless of link structure or attribute distribution and dependencies. However, as we will see in the following section, link sampling can drastically reduce sample size, negatively effecting power.

5. Discussion

In the previous sections, we outlined several data scenarios in which conventional χ^2 tests fail, and provided a pair of solutions. Given these choices, how should a practitioner proceed? The answer to this question depends on the structure of the data, and whether the attribute values exhibit autocorrelation.

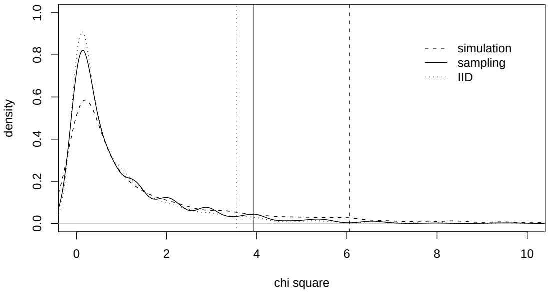


Figure 15. When generated from independent edges, the distribution of χ^2 closely matches the theoretical X^2 distribution.

5.1. Applicability

If the data are propositional, then the conventional χ^2 test often applies, assuming the absence of latent dependence as discussed in 2. Furthermore, relational data sets with one-to-many or many-to-many link structures will also work with a conventional test if the data are not autocorrelated. Of course determining whether autocorrelation exists is not a trivial task itself, and failing to assess it correctly will result in the variety of Type I error detailed in the helmet example (incorrectly rejecting the null hypothesis).

The permutation test approach produces accurate p-values for data that are not IID. Since it derives sampling distributions empirically, it is robust to autocorrelation in the data. Furthermore, permutation tests can be utilized in situations in which there is no information about the relationship between attributes and structure. The one exception occurs with many-to-many data that exhibit undetermined or two-sided autocorrelation. Since the procedure relies on preserving the autocorrelation on one entity type while destroying correlation between entity types under the null hypothesis, it cannot be applied if dependence exists amongst both entity types.

The link sampling method can be applied to any relational data set or attribute structure. However, the size of maximal matchings may be limited in data sets with dense link structures, which limits the methods usefulness for heavily-linked data.

5.2. Accuracy and power

Statistical power is the probability that a procedure will not commit a Type II error (incorrectly accepting the null hypothesis). Figure illustrates the power advantage of the permutation method over link sampling for a synthetic one-

Table 3. Summary of applicability of each hypothesis test method

method	1-to-1	1-to-m, no ac	1-to-m, ac	m-to-m, no ac	m-to-m, sing. ac	m-to-m, doub. ac
conventional X^2	X	X		X		
permutation	X	X	X	X	X	
link sampling	X	X	X	X	X	X

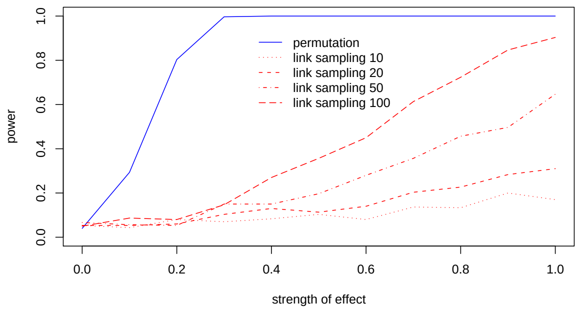


Figure 16. Power as a function of effect strength on synthetic graphs for the permutation and link sampling methods. Four variations of the latter are presented for four different levels of sample size s (10, 20, 50, 100). The higher the sample size, the better the power; however, the permutation method's performance dominates even at s=100.

to-many data set. This higher power comes at a cost, however, as permutation tests are computationally intensive.

5.3. Conclusions

Which technique works best depends on the requirements of the task and domain. Given a relational data set, an assessment of the link structure and autocorrelation among attributes can help indicate the proper test. For large data sets, the link sampling approach can be safely utilized regardless of attribute structure. For smaller data sets, or any situation where maximum statistical power is desired, the permutation technique is a good choice.

In this work we explained two hypothesis testing procedures for relational data, and detailed the scenarios where they improve performance. While this issues described in Section 2 can be present in propositional data sets, they cannot be accounted for using these techniques. The ability to conduct accurate hypothesis tests in the presence of data instance dependence is a compelling argument for the use of relational representations.

References

[1] J. Aronson, M. E. Dyer, A. M. Frieze, and S. Suen. Randomized greedy matching II. *Random Structure and Algorithms*, 6(1):55–74, 1995.

[2] K. L. Delucchi. The use and misuse of chi-square: Lewis and Burke revisited. In *Proceedings of the Sixty-fifth Annual Meeting of the American Educational Research Association*, Los Angeles, CA, April 1981.

[3] E. S. Edgington. *Randomization Tests*. Marcel Dekker, New York, 1980.

[4] E. Frank and I. H. Witten. Using a permutation test for attribute selection in decision trees. In J. W. Shavlik, editor, *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 152–160, Madison, WI, July 1998. Morgan Kaufmann, San Francisco, CA.

[5] N. Friedman, L. Getoor, D. Koller, and A. Pfeffer. Learning probabilistic relational models. In *International Joint Conference on Artificial Intelligence*, volume 16, pages 1300–1309. Citeseer, 1999.

[6] A. Gionis, H. Mannila, T. Mielikäinen, and P. Tsaparas. Assessing data mining results via swap randomization. In T. Eliassi-Rad, L. H. Ungar, M. Craven, and D. Gunopulos, editors, *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 167–176, Philadelphia, PA, August 2006. ACM Press, New York, NY.

[7] P. Good. *Permutation Tests: A Practical Guide for Testing Hypotheses*. Springer Series in Statistics. Springer-Verlag, New York, 1994.

[8] D. D. Jensen. Knowledge discovery through induction with randomization testing. In G. Piatetsky-Shapiro, editor, *Proceedings of the Knowledge Discovery in Databases Workshop*, pages 148–159, Meno Park, CA, 1991. AAAI Press.

[9] D. D. Jensen and J. Neville. Linkage and autocorrelation cause feature selection bias in relational learning. In C. Sammut and A. G. Hoffmann, editors, *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 259–266, Sydney, Australia, July 2002. Morgan Kaufmann.

[10] D. D. Jensen, J. Neville, and M. Hay. Avoiding bias when aggregating relational data with degree disparity. In T. Fawcett and N. Mishra, editors, *Proceedings of the Twentieth International Conference on Machine Learning*, pages 274–281, Washington, DC, August 2003. AAAI Press.

[11] D. D. Jensen, J. Neville, and M. J. Rattigan. Randomization tests for relational learning. Technical Report UM-CS-2003-05, Department of Computer Science, University of Massachusetts, Amherst, MA, 2003.

[12] D. Lewis and C. J. Burke. The use and misuse of the chi-square test. *Psychological Bulletin*, 46(6):433–489, November 1949.

[13] E. W. Noreen. *Computer-Intensive Methods for Testing Hypotheses: An Introduction*. Wiley-Interscience, New York, NY, April 1989.

[14] K. Pearson. On the criterion that a given system of deviations from the probable in the case of correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50:157–175, 1900.

[15] G. I. Webb. Discovering significant rules. In T. Eliassi-Rad, L. H. Ungar, M. Craven, and D. Gunopulos, editors, *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 434–443, Philadelphia, PA, August 2006. ACM Press, New York, NY.