Computational Statistics and Data Analysis 152 (2020) 107043

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics

M. Baak [a,*], R. Koopman [a], H. Snoek [b,c], S. Klous [a,d]

[a] KPMG Advisory N.V., Laan van Langerhuize 1, 1186 DS, The Netherlands
[b] Nikhef National Institute for Subatomic Physics, Science Park 105, 1098 XG Amsterdam, The Netherlands
[c] Institute of Physics, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands
[d] Informatics Institute, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 28 November 2018
Received in revised form 23 June 2020
Accepted 24 June 2020
Available online 27 June 2020

Keywords:
Data analysis
Correlation
Contingency test
Significance
Simulation

ABSTRACT

A prescription is presented for a new and practical correlation coefficient, ϕ_K, based on several refinements to Pearson's hypothesis test of independence of two variables. The combined features of ϕ_K form an advantage over existing coefficients. Primarily, it works consistently between categorical, ordinal and interval variables, in essence by treating each variable as categorical, and can therefore be used to calculate correlations between variables of mixed type. Second, it captures nonlinear dependency. The strength of ϕ_K is similar to Pearson's correlation coefficient, and is equivalent in case of a bivariate normal input distribution. These are useful properties when studying the correlations between variables with mixed types, where some are categorical. Two more innovations are presented: to the proper evaluation of statistical significance of correlations, and to the interpretation of variable relationships in a contingency table, in particular in case of sparse or low statistics samples and significant dependencies. Two practical applications are discussed. The presented algorithms are easy to use and available through a public Python library.[1]

1. Introduction

The calculation of correlation coefficients between paired data variables is a standard tool of analysis for every data analyst. Pearson's correlation coefficient (Pearson, 1895) is a *de facto* standard in most fields, but by construction only works for interval variables. In practice it is inconvenient to mix and compare different correlation coefficients when dealing with mixed variable types. While many coefficients of association exist, each with different strengths, we have not been able to identify a correlation coefficient with Pearson-like strength and a sound statistical interpretation that can be used to calculate correlations between variables of mixed type. (The convention adopted here is that a correlation coefficient is bound, *e.g.* in the range [0, 1] or [−1, 1], and that a coefficient of association is not.)

This paper describes among others a novel correlation coefficient, ϕ_K, with properties that – taken together – form an advantage over existing methods. More broadly, it covers innovations to three related topics typically encountered in data analysis.

The calculation of the correlation coefficient ϕ_K***, for each variable-pair of interest.*** The correlation ϕ_K follows a uniform treatment for interval, ordinal and categorical variables, because its definition is invariant under the ordering of

* Corresponding author.
E-mail addresses: maxbaak@gmail.com (M. Baak), rfkoopman@gmail.com (R. Koopman).

1 https://github.com/KaveIO/PhiK.

https://doi.org/10.1016/j.csda.2020.107043
0167-9473/

the values of each variable. Essentially it is treating each variable as if its type is categorical. This is particularly useful in modern-day analysis when studying the dependencies between a set of variables with mixed types, where some variables are categorical.

The correlation ϕ_K is derived from Pearson's χ^2 contingency test (Barnard, 1992), *i.e.* the hypothesis test of independence between two (or more) variables in a contingency table, henceforth called the Factorization Assumption. In a contingency table each row is the category of one variable and each column the category of a second variable. Each cell describes the number of records occurring in both categories at the same time. The values for levels of correlation are bound in the range [0, 1], with 0 for no association and +1 for complete association. By construction, the strength of ϕ_K is the same as Pearson's correlation coefficient in case of a bivariate normal input distribution. Unlike Pearson, which describes the average linear dependency between two variables, ϕ_K also captures nonlinear relations.

The evaluation of the statistical significance of each correlation. Presented is a new and robust statistical prescription for the significance evaluation of the level of variable association, based on a hybrid method of Monte Carlo sample simulation and adjustments of the χ^2 distribution when using the G-test statistic (Sokal and Rohlf, 2012). The asymptotic approximation commonly advertised to evaluate the statistical significance of the hypothesis test, *e.g.* by statistics libraries such as `R` (Anon, 0000b) and `scipy` (Anon, 0000a), makes particular assumptions on the number of degrees of freedom and the shape of the χ^2 distribution. This approach is unusable for sparse data samples, which may happen for two variables with a strong correlation and for low- to medium-statistics data samples, and leads to incorrect p-values. (Examples follow in Section 5.)

Insights in the correlation of each variable-pair, by studying outliers and their significances. To help interpret any relationship found, we provide a newly refined method for the detection of significant excesses or deficits of records with respect to the expected values in a contingency table, so-called outliers, using a statistically independent evaluation for expected frequency of records. We evaluate the significance of each outlier frequency, putting particular emphasis on the statistical uncertainty on the expected number of records and on the scenario of low statistics data samples.

The methods presented in this work can be applied to many analysis problems. Insights in variable dependencies serve as useful input to all forms of model building, be it classification or regression based, such as the identification of customer groups, outlier detection for predictive maintenance or fraud analytics, and decision making engines. More generally, they can be used to find correlations across (big) data sets, and correlations over time (in correlograms). Two use-cases are discussed: the study of numbers of insurance claims and of survey responses.

In summary, this paper covers three innovations. We present a new correlation coefficient that is useful when dealing with data sets with mixed variable types, including categorical ones. We introduced a novel "hybrid" approach for the significance evaluation of contingency tables, that works in particular for sparse and low-statistics data sets. And we discuss a new, statistically refined method for inspecting the relationship between two dependent variables, one that also works for any two mixed variable types.

This document is organized as follows. A brief overview of existing correlation coefficients is provided in Section 2. Section 3 describes the contingency test, which serves as input for Section 4, detailing the derivation of the correlation coefficient ϕ_K. The statistical significance evaluation of the contingency test is discussed in Section 5. In Section 6 we zoom in on the interpretation of the dependency between a specific pair of variables, where the significance evaluation of outlier frequencies in a contingency table is presented. Two practical applications can be found in Section 7. Section 8 describes the implementation of the presented algorithms in publicly available Python code, before concluding in Section 9.

2. Measures of variable association

A correlation coefficient quantifies the level of mutual, statistical dependence between two variables. Multiple types of correlation coefficients exist in probability theory, each with its own definition and features. Some focus on linear relationships where others are sensitive to any dependency, some are robust against outliers, etc. In general, different correlation coefficients are used to describe dependencies between interval, ordinal, and categorical variables separately. Typically their values range from −1 to +1 or 0 to +1, where 0 means no statistical association, +1 means the strongest possible association (a one-on-one dependency), and −1 means the strongest negative relation. While these particular two or three correlation points are well-defined statistically, the various correlation coefficients follow different approaches to navigate between them. When working with different correlation coefficients for different types of variables, their comparison can thus be complex. (Some correlation constants (Yoo et al., 2020) have an axiomatic foundation, which in this context forms an advantage of other correlation coefficients.)

In this paper, we are interested in correlation constants that work with interval, ordinal, and categorical variables alike. This section briefly discusses existing correlations coefficients and other measures of variable association. This is done separately for interval, ordinal, and categorical variables, as measures of association are typically designed to work with one specific variable type. In addition, several related concepts used throughout this work are presented.

An ***interval variable***, sometimes called continuous or real-valued variable, has well-defined intervals between the values of the variable. Examples are distance or temperature measurements. The Pearson correlation coefficient is a *de* *facto* standard to quantify the level of association between two interval variables. For a sample of size N with variables x and y, it is defined as the covariance of the two variables divided by the product of their standard deviations:

$$\rho = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}\,, \tag{1}$$

where $\bar{x}$ and $\bar{y}$ are the sample means. Notably, ρ is symmetric in x and y, and $\rho \in [-1, 1]$.

Extending this to a set of input variables, Pearson's correlation matrix C, containing the ρ values of all variable pairs, is obtained from the covariance matrix V as:

$$C_{ij} = \frac{V_{ij}}{\sqrt{V_{ii}V_{jj}}}\,, \tag{2}$$

where ij are the indices of a variable pair. Equivalently one can write:

$$V = E^{\frac{1}{2}}\,C\,E^{\frac{1}{2}} \quad ; \quad V^{-1} = E^{-\frac{1}{2}}\,C^{-1}\,E^{-\frac{1}{2}}\,, \tag{3}$$

where $E = \mathrm{diag}(V)$.

The Pearson correlation coefficient measures the strength and direction of the linear relationship between two interval variables; a well-known limitation is therefore that nonlinear dependencies are not (well) captured. In addition, ρ is known to be to sensitive to outlier records. Pearson's correlation coefficient requires interval variables as input, which can be unbinned or binned. It cannot be evaluated for categorical variables, and ordinal variables can only be used when ranked (see below).

A direct relationship exists between ρ and a bivariate normal distribution:

$$\begin{aligned} &f_{\text{b.n.}}(x, y\,|\,\bar{x}, \bar{y}, \sigma_x, \sigma_y, \rho) = \\ &\qquad \frac{1}{2\pi\,\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\bar{x})^2}{\sigma_x^2} + \frac{(y-\bar{y})^2}{\sigma_y^2} - \frac{2\rho(x-\bar{x})(y-\bar{y})}{\sigma_x\sigma_y}\right]\right), \end{aligned} \tag{4}$$

where σ_x (σ_y) is the width of the probability distribution in x (y), and the correlation parameter ρ signifies the linear tilt between x and y. We use this relation in Section 4 to derive the correlation coefficient ϕ_K.

Another useful measure is the global correlation coefficient (James and Roos, 1975), which is a number between zero and one obtained from the correlation matrix C that gives the highest possible correlation between variable k and the linear combination of all other variables:

$$\begin{aligned} g_k &= \sqrt{1 - \left[V_{kk} * (V^{-1})_{kk}\right]^{-1}} \\ &= \sqrt{1 - \left[(C^{-1})_{kk}\right]^{-1}}\,. \end{aligned} \tag{5}$$

An ***ordinal variable*** has two or more categories with a clear ordering of these categories. For example, take the variable "level of education" with six categories: no education, elementary school graduate, high school graduate, college and university graduate, PhD. A rank correlation measures the statistical relationship between two variables that can be ordered; the rank of a variable is its index in the ordered sequence of values. For ordinal variables a numbering is assigned to the categories, *e.g.* 0, 1, 2, 3. Note the equidistant spacing between the categorical values.

Examples of rank correlation coefficients are Spearman's ρ (Spearman, 1904), Kendall's τ (Kendall, 1938), Goodman–Krustall's γ (Goodman and Krustal, 1954, 1959, 1963, 1972), and the polychoric correlation (Drasgow, 2006). The definition of Spearman's ρ is simply Eq. (1), using the ranks of x_i and y_i as inputs, essentially treating the ranks as interval variables. This makes Spearman's ρ very robust against outliers. Note that Goodman–Krustall's γ is dependent on the order of the two input variables, resulting in an asymmetric correlation matrix. See Ref. Yoo et al. (2020) for the theoretical properties of a generalization of Kendall's τ that deals with sparse data, which is also one of the characteristics of this work.

Although ranking is regular practice, the assumption of equidistant intervals – often made implicitly – can sometimes be hard to justify. For example, adding the category of "MBA" to the above example increases the distance between "PhD" and "no education", where one could argue that this distance should be independent of the number of educational categories.

A ***categorical variable***, sometimes called a nominal or class variable, has two or more categories which have no intrinsic ordering. An example is the variable gender, with two categories: male and female. Multiple measures of association exist that quantify the mutual dependence between two (or more) categorical variables, including Pearson's χ^2 contingency test (Barnard, 1992), the G-test statistic (Sokal and Rohlf, 2012), mutual information (Cover and Thomas, 2006), and for 2×2 contingency tables Fisher's exact test (Fisher, 1922, 1970) and Barnard's test (Barnard, 1945, 1947). For an overview see Ref. Agresti (1992). These measures determine how similar the joint distribution $p(x, y)$ is to the product of the factorized marginal distributions $p(x)p(y)$. Each measure of association consists of a sum of contributions, one from each cell of the contingency table, and therefore does not depend on the intrinsic ordering of the cells.

Though typically limited to categorical variables, these test statistics can also be applied to interval and ordinal type variables. However, their values are not bound in the range [0, 1], and can become quite large. Moreover, their interpretation is often non-trivial, as their values not only depend on the level of association, but also on the number of categories or intervals and the number of records.

There are few correlation coefficients for categorical variables. Most comparable to this work is Cramér's ϕ (Harald, 1999), a correlation coefficient meant for two categorical variables, denoted as ϕ_C, based on Pearson's χ^2 test statistic, and with values between 0 (no association) and +1 (complete association):

$$\phi_C = \sqrt{\frac{\chi^2}{N \min(r-1, k-1)}}, \tag{6}$$

where r (k) is the number of rows (columns) in a contingency table. Notably, with a relatively small number of records, comparable with the number of cells, statistical fluctuations can result in large values of ϕ_C without strong evidence of a meaningful correlation. (An example of this follows in Fig. 6a.)

Cramér's ϕ can also be used for ordinal and binned interval variables. For interval variables a binning needs to be chosen. Fig. 1a shows ϕ_C for a high-statistics, binned bivariate normal input distribution with correlation parameter ρ. Compared to Pearson's ρ, ϕ_C shows relatively low values for most values of ρ, and only shoots up to one for values of ρ close to one. Moreover, the value found for ϕ_C is quite dependent on the sample size, as observed in Fig. 1b, and on the binning chosen per variable, as shown in Fig. 1c and d. There are two reasons behind the binning dependency. With a symmetric binning of both interval variables, from Eq. (6) one has $\phi_C \propto 1/\sqrt{r-1}$, pushing ϕ_C down as r increases. This is seen in Fig. 1c, where the average number of entries per bin is kept constant. In addition, yet smaller in impact, with a higher total number of bins the relative impact of statistical noise goes up, likewise bringing up ϕ_C. This is seen most clearly in Fig. 1d, where the sample size is kept constant, in the scenario $\rho_{\text{true}} = 0$. Also note in Fig. 1b, for the scenario $\rho_{\text{true}} = 0$ with no true variable dependency, that ϕ_C is biased and gives significant nonzero values. These effects make ϕ_C hard to interpret and unsuitable for interval variables, as – in our opinion – a proper correlation coefficient for interval variables should be (relatively) insensitive to both the sample size and the selected binning.

Another alternative is the contingency coefficient C_p, which has the disadvantage that its maximum value depends on the number of categories r and k, and does not reach a maximum of one. The recommendation (Smith, 2009) is not to use C_p to compare correlations in tables with variables that have different numbers of categories (*i.e.* when $r \neq k$).

To address the aforementioned issues, in this paper we define the coefficient of correlation ϕ_K, derived from Pearson's χ^2 contingency test in Section 4, and its statistical significance, derived using the G-test in Section 5. Our interest is in correlation constants that can deal with all three variable types. Pearson's ρ is used in this work as parameter in the bivariate normal distribution to derive the correlation coefficient ϕ_K. We use Cramér's ϕ for comparison purposes, because it can also be used for categorical, ordinal and binned interval variables. We do not use any of the rank correlation coefficients as they cannot be used with categorical variables.

3. Test of variable independence

The contingency test, also called the test of variable independence, determines if a significant relationship exists between two (or more) categorical variables. Though usually performed on two categorical variables, the test can equally be applied to ordinal and binned interval variables, and can be extended to an arbitrary number of variables. Specifically, the contingency test indicates how well the joint data distribution $p(x, y)$ of variables x and y is described by the product of its factorized marginal distributions $p(x)p(y)$.

Throughout this paper we employ two contingency tests, where each compares the observed frequency of each category for one variable with the expectation across the categories of the second variable:

1. Pearson's χ^2 test:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \tag{7}$$

which is used to define the correlation coefficient ϕ_K in Section 4. Pearson's χ^2 test is the standard test for variable independence.

2. The G-test, sometimes called log-likelihood ratio test:

$$G = 2 \sum_{i,j} O_{ij} \log(O_{ij}/E_{ij}), \tag{8}$$

which is used to evaluate the significance of the contingency test in Section 5. The sum is taken over all non-empty cells.

In both formulas, O_{ij} (E_{ij}) is the observed (expected) frequency of records for row i and column j of the contingency table. The stronger the dependency between x and y, the less well modeled is their distribution by the factorized distribution $p(x)p(y)$, and the larger each test statistic value.

Under the factorization assumption, the expected frequencies can be obtained in two ways: statistically dependent and independent.

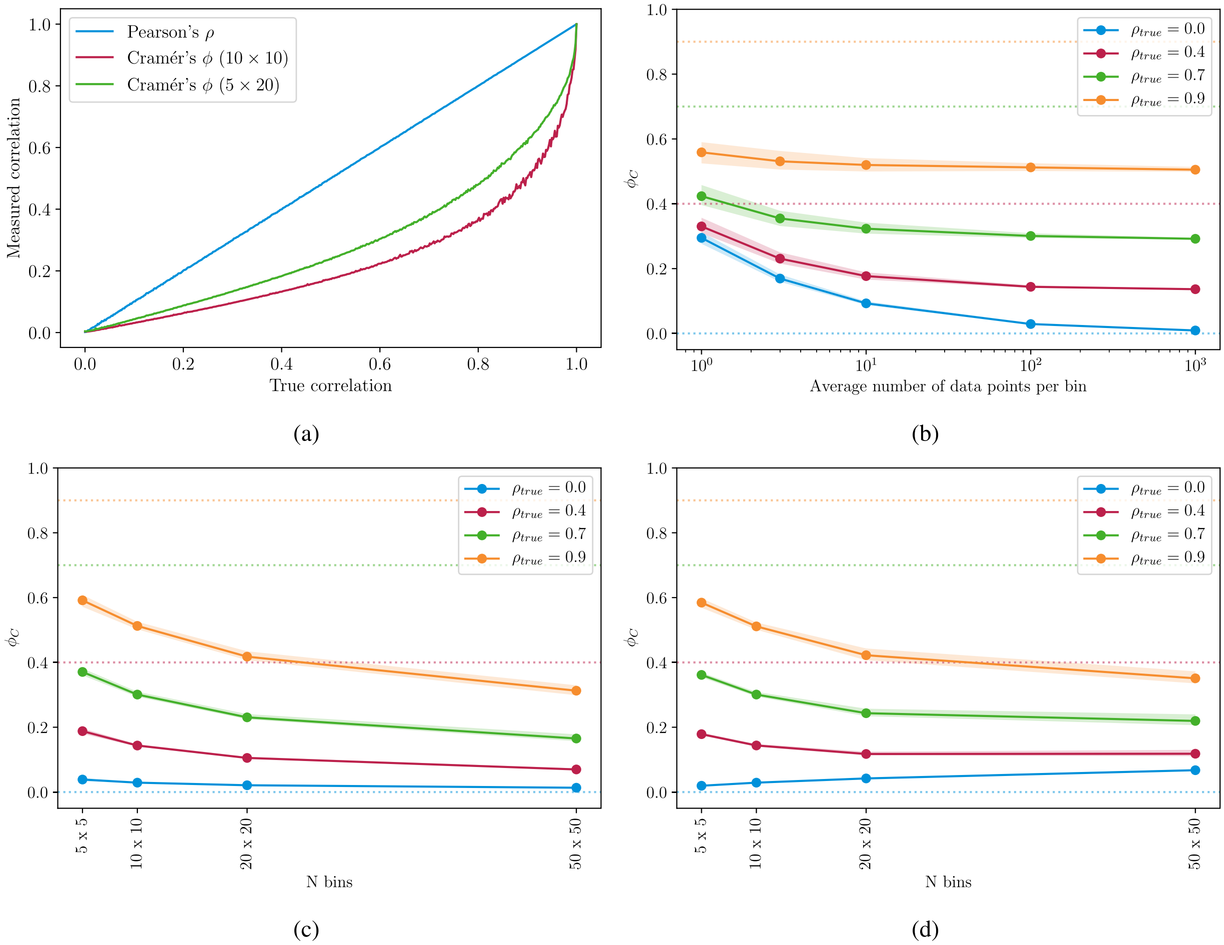


Fig. 1. (a) Cramér's ϕ versus Pearson's ρ. Each point along the curves is computed from a bivariate normal distribution generated with one million entries. The two curves for Cramér's ϕ have been evaluated with different numbers of rows r and columns k: 10×10 and 5×20 bins. (b–d) Cramér's ϕ obtained from bivariate normal distributions with different symmetric binnings and values for $\rho_{\text{true}} = 0.0, 0.4, 0.7, 0.9$. Each point shows the median value measured in 1000 simulations with ten thousand entries each, while the error band indicates the 25th to 75th percentiles. In (b) the sample size is varied, with 10×10 bins. In (c) the average number of entries per bin is fixed to 100, where in (d) the sample size is fixed to 10 000, both with symmetric varying number of bins.

3.1. Dependent frequency estimates

The default method of frequency estimation for row i and column j includes O_{ij}, so E_{ij} is statistically dependent on the observed frequency of its bin. The expected value of the two nominal variables is calculated as:

$$E_{ij} = N\, p_r(i)\, p_k(j) = \frac{(\sum_{n=1}^{k} O_{in})(\sum_{m=1}^{r} O_{mj})}{N}\,, \tag{9}$$

where $p_r(i)$ ($p_k(j)$) is the ith (jth) bin of the row-projected (column-projected) marginal probability mass function (p.m.f.) and N is the number of records. The statistical dependency between E_{ij} and O_{ij} arises as the expectation E_{ij} for cell ij includes the cell's observation O_{ij} in both the sum over columns and rows, and as part of N. The formula can be easily extended to an arbitrary number of variables. We use Eq. (9) for the definition of ϕ_K in Section 4 and for the calculation of its significance in Section 5, as this distribution matches the observed frequencies most closely.

3.2. Independent frequency estimates

The second method of estimation of E_{ij} excludes O_{ij}, *i.e.* is statistically independent of the observed frequency of records for row i and column j. This estimate, known in high energy physics as the ABCD formula (Aaboud et al., 2018), is given by:

$$E_{ij} = \frac{B_{ij}\,C_{ij}}{D_{ij}} = \frac{(\sum_{n\neq j} O_{in})(\sum_{m\neq i} O_{mj})}{\sum_{m\neq i}\sum_{n\neq j} O_{mn}}\,, \tag{10}$$

where by construction O_{ij} is not part of E_{ij}, which allows for an objective comparison between observed and expected frequencies per bin. This formula can also be extended to more variables, except that the denominator of Eq. (10), which is different for each pair of indices, can easily become zero for low statistics samples.

Note that B_{ij}, C_{ij}, and D_{ij} are sums of frequencies, each obeying Poisson statistics, and are statistically independent. Consequently, the statistical uncertainty on E_{ij} is evaluated with straight-forward error propagation (Ku, 1965) as:

$$\sigma^2_{E_{ij}} = \frac{\sigma^2_{B_{ij}} C^2_{ij}}{D^2_{ij}} + \frac{\sigma^2_{C_{ij}} B^2_{ij}}{D^2_{ij}} + \frac{\sigma^2_{D_{ij}} E^2_{ij}}{D^2_{ij}}\,. \tag{11}$$

For an observed frequency of Q records, $\sigma_Q = \sqrt{Q}$, except when $Q = 0$, in which case we set $\sigma_Q = 1$. By doing so, when B_{ij} or C_{ij} is zero, and thus $E_{ij} = 0$, this approach results in a nonzero error on E_{ij}. The statistical uncertainty on the expected frequency, $\sigma_{E_{ij}}$, is only zero when both B_{ij} and C_{ij} are zero.

The expectation from Eq. (10) is built with fewer statistics than Eq. (9) and thus is slightly less accurate. Another difference is that the ABCD formula is not a true product of two (or more) factorized marginal distributions, *i.e.* the relative predictions for one row are not identical to those for another row, as is the case for dependent frequency estimates. We use the independent frequency estimates of Eq. (10) for the detection of significant excesses or deficits of records over expected values in a contingency table in Section 6, for reasons described there in more detail.

4. Definition of ϕ_K

The correlation coefficient ϕ_K is obtained by taking a contingency table of two variables, with any combination of type, and inverting the χ^2 contingency test statistic through the steps outlined below. In short, we interpret the χ^2 value in data to come from a bivariate normal distribution with correlation parameter ϕ_K. Although the procedure could be extended to more variables, the method is described with two variables for simplicity.

We define the bivariate normal distribution of Eq. (4) with correlation parameter ρ and unit widths, centered around the origin, and in the range $[-5, 5]$ for both variables. Using uniform binning for the two interval variables, with r rows and k columns, results in a corresponding bivariate p.m.f. With N records, the observed frequencies, O_{ij}, are set equal to the probability per bin multiplied by N. The expected frequencies E_{ij}, are set to the predictions from the bivariate normal distribution with $\rho = 0$, with N records and the same binning. We then evaluate the χ^2 value of Eq. (7).

Let us define this function explicitly. First, we perform the integral of the bivariate normal distribution over the area of bin ij

$$F_{ij}(\rho) = \int_{\text{area}\,ij} f_{\text{b.n.}}(x, y \mid \rho)\,\mathrm{d}x\mathrm{d}y\,, \tag{12}$$

leading to the sum over bins:

$$\chi^2_{\text{b.n.}}(\rho, N, r, k) = N \sum_{i,j}^{k,r} \frac{\left(F_{ij}\,(\rho = \rho) - F_{ij}\,(\rho = 0)\right)^2}{F_{ij}\,(\rho = 0)}\,. \tag{13}$$

This χ^2 value explicitly ignores statistical fluctuations in observed frequencies, and is a function of the numbers of rows and columns, N, and the value of ρ.

To account for statistical noise, we introduce a sample-specific pedestal related to a simple estimate of the effective number of degrees of freedom of the bivariate sample, n_{sdof}:

$$n_{\text{sdof}} = (r-1)(k-1) - n_{\text{empty}}(\text{expected})\,, \tag{14}$$

with number of rows r and columns k, and where $n_{\text{empty}}(\text{expected})$ is the number of empty bins of the dependent frequency estimates of the sample. The pedestal is defined as:

$$\chi^2_{\text{ped}} = n_{\text{sdof}} + c \cdot \sqrt{2 n_{\text{sdof}}}\,. \tag{15}$$

The noise pedestal is configurable through parameter c, and by default $c = 0$. See Section 4.4 for the impact of the noise pedestal on ϕ_K and Section 5 for a discussion on the effective number of degrees of freedom.

The maximum possible χ^2 value (Harald, 1999) of the contingency test is:

$$\chi^2_{\max}(N, r, k) = N \min(r-1, k-1)\,, \tag{16}$$

which depends only the number of records N, rows r, and columns k, and is reached when there is a one-on-one dependency between the two variables, *i.e.* 100% correlation. Specifically note that $\chi^2_{\max}$ is independent of the shape of distribution $p(x, y)$. (The G-test does not have this useful feature, making the G-test unsuitable for the calculation of ϕ_K.)

We scale Eq. (13) to ensure it equals χ^2_{ped} for $\rho = 0$ and $\chi^2_{\max}$ for $\rho = 1$.

$$X^2_{\text{b.n.}}(\rho, N, r, k) = \chi^2_{\text{ped}} + \left\{ \frac{\chi^2_{\max}(N, r, k) - \chi^2_{\text{ped}}}{\chi^2_{\text{b.n.}}(1, N, r, k)} \right\} \cdot \chi^2_{\text{b.n.}}(\rho, N, r, k)\,. \tag{17}$$

This function is symmetric in ρ, and increases monotonically from χ^2_{ped} to $\chi^2_{\max}$ as ρ goes from zero to one.

We can now perform the necessary steps to obtain the correlation coefficient ϕ_K (see Brent, 1973).

Procedure description 1: the calculation of ϕ_K

1. In case of unbinned interval variables, apply a binning to each one. A reasonable binning is generally use-case specific, and needs to be chosen such that the bin width is small enough to capture the variations observed in the data. As a default setting we take 10 uniform bins per variable.
2. Fill the contingency table for a chosen variable pair, which contains N records, r rows and k columns.
3. Evaluate the χ^2 contingency test using the Pearson's χ^2 test statistic (Eq. (7)) and the statistically dependent frequency estimates, as detailed in Section 3.1.
4. Interpret the χ^2 value as coming from a bivariate normal distribution without statistical fluctuations, using Eq. (17).

   i. If $\chi^2 < \chi^2_{\text{ped}}$, set ϕ_K to zero.
   ii. Else, with fixed N, r, k, invert the $X^2_{\text{b.n.}}$ function, *e.g.* using Brent's method (Brent, 1973), and numerically solve for ρ in the range [0, 1].
   iii. The solution for ρ defines the correlation coefficient ϕ_K.

In summary, the correlation constant ϕ_K interprets the χ^2 value found in data as coming from a bivariate normal distribution with a fixed amount of statistical noise and with correlation parameter ϕ_K. Because of the sum in the χ^2 definition (see Eq. (7)), ϕ_K is invariant under the ordering of the (binned) values of each input variable. Stated differently, ϕ_K is evaluated identically for (binned) interval, ordinal and categorical variables, and ϕ_K is directionless: it cannot differentiate between positive and negative association values. (For example, when switching the sign of the values of one interval variable, Pearson's ρ changes sign, but ϕ_K is unchanged.) Nonlinear relations are captured by ϕ_K through the χ^2 test of variable independence. The correlation ϕ_K reverts to the Pearson correlation coefficient in case of a bivariate normal input distribution, with uniformly binned interval variables. Interval variables first get binned, and the distance between bins is not used. This makes ϕ_K insensitive to outliers in interval variables. Unlike Cramér's ϕ, the value of ϕ_K is stable against the number of bins chosen per interval variable, making it unambiguous to interpret. Like Cramér's ϕ, ϕ_K is affected by statistical fluctuations, which is relevant when the number of records is comparable with the number of cells (or lower); however, unlike Cramér's ϕ, ϕ_K has a correction for the statistical noise. Note that ϕ_K is independent of the order of the two input variables, and that the procedure can be extended to more than two variables if desired. (For more than two variables, follow the same procedure and assume a common correlation for each variable pair of a multivariate normal input distribution.)

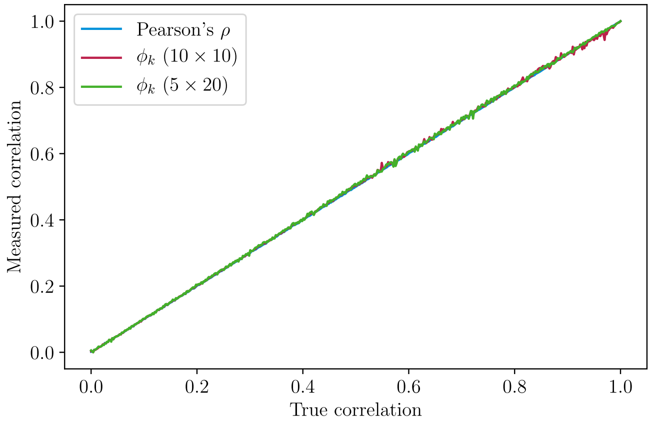
All coefficients presented in Section 2 are computationally inexpensive to evaluate. The calculation of ϕ_K is computationally expensive because of the integrals of correlated bivariate normal distributions evaluated in Eq. (17), but is feasible on any modern laptop, typically taking only a fraction of a second.

4.1. Performance on benchmark samples

When plotting the ϕ_K values obtained from high-statistics, bivariate normal distributions, evaluated both with symmetric or asymmetric binning (see Fig. 2, which uses the same setup as Fig. 1a.), we obtain indistinguishable diagonal lines. In this high-statistics scenario, we observe that the measured correlation is quite stable against (uniform) rebinning and that the strength of the correlation overlaps very well with Pearson's ρ.

A comparison of ϕ_K with the correlation coefficients Pearson's ρ and Cramér ϕ based on benchmark samples (Boigelot, 0000) is given in Fig. 3. By construction, the interpretation of ϕ_K is similar to that of Pearson's correlation coefficient, in particular for the bivariate normal input distributions and the linear shapes, shown in the left and middle columns. Unlike Pearson, however, ϕ_K also captures nonlinear relations as shown in the right column. Moreover, ϕ_K can be determined for categorical, ordinal, and interval variables alike. Note that Cramér's ϕ gives relatively low values for all samples.

The finite sample bias, convergence rate, and dependency of choice of binning have been studied for ϕ_K for the bivariate normal input distributions of Fig. 3 left column and the benchmark shapes of Fig. 3 right column. The results of these are shown in Fig. 4. In low statistics samples with weak true correlation the uncertainty on ϕ_K can be large and asymmetric, see Fig. 4a–b. Compared with Pearson's ρ, on average there is a small bias towards higher ϕ_K values for the bivariate normal input distribution. The value of ϕ_K converges rapidly to the true correlation when the sample size increases. The convergence rate of ϕ_K depends highly on the shape of the underlying distribution. Although, some variation in ϕ_K value



(a)

Fig. 2. ϕ_K versus Pearson's ρ. Each point along the curves is computed from a bivariate normal distribution generated with 1 million entries. The two curves for ϕ_K have been evaluated with different numbers of rows r and columns k: 10 × 10 and 5 × 20 bins. The measured correlation does not depend on the number of bins and overlaps with Pearson's ρ.

is observed in low statistics samples, the value of ϕ_K converges rapidly for high statistics samples. Note that for the shapes on the right no true correlation is defined. The measured value of ϕ_K can be affected by the binning, as shown for (c) the bivariate normal input distribution and (d) the benchmark shapes. In both figures the average number of entries per bin is kept fixed at 100. In particular the choice of binning affects the calculation of ϕ_K when the number of bins is too low to capture the variation observed in the distribution of a variable. Once sufficient number of bins have been chosen, increasing the number does not significantly affect the measured value of ϕ_K for the shapes under study. Although overall ϕ_K stays relatively stable, Fig. 4e–f show that, with a fixed sample size of 10 000 and large number of bins for both variables, going towards an average of only 4 entries per bin, the contribution of statistical noise to the χ^2 becomes sizeable, thereby increasing the value of ϕ_K, albeit slowly.

The choice of binning for interval variables is use-case specific, and care should be taken to select a binning with the minimum resolution required to capture the variation observed in each interval variable. The total number of bins should also not be too high, in order to ensure that the contribution of statistical noise to ϕ_K stays marginal. The correction term for statistical fluctuations to ϕ_K is discussed further in Section 4.4.

A comparison of Figs. 1 and 4 leads to several interesting insights. Compared with Cramér's ϕ:

1. For bivariate normal distributions the strength of ϕ_K is very similar to Pearson's ρ.
2. The median value of ϕ_K is relatively insensitive to sample size, although the uncertainty on ϕ_K can be large.
3. For binned interval variables, ϕ_K is relatively robust against the selected number of bins per variable.

In our view the last two items are key, practical requirements for a good correlation constant.

4.2. Example correlation matrix

When studying the dependencies of a set of variables with a mixture of types, one can now calculate the correlation matrix for all variable pairs, filled with ϕ_K values, which is a useful overview to have for a data analyst. For illustration purposes a synthetic data set with car insurance data has been created. The data set consists of 2000 records. Each record contains 5 (correlated) variables of mixed variable types, see Table 1. These data are used throughout the paper to provide insights in the practical application of the methods introduced in this work. The ϕ_K correlation matrix measured on the car insurance data set is shown in Fig. 5a.

4.3. Global correlation coefficients

Besides the variable-pair information available from the correlation matrix C, it is also interesting to evaluate per variable the global correlation coefficient, g_k, defined in Eq. (5). Strictly speaking g_k is only defined for interval variables, but here we extend the definition to the full correlation matrix available. Example global correlations measured in the car insurance data are shown in Fig. 5b. They give a tenable estimate of how well each variable can be modeled in terms of all other variables, irrespective of variable type.

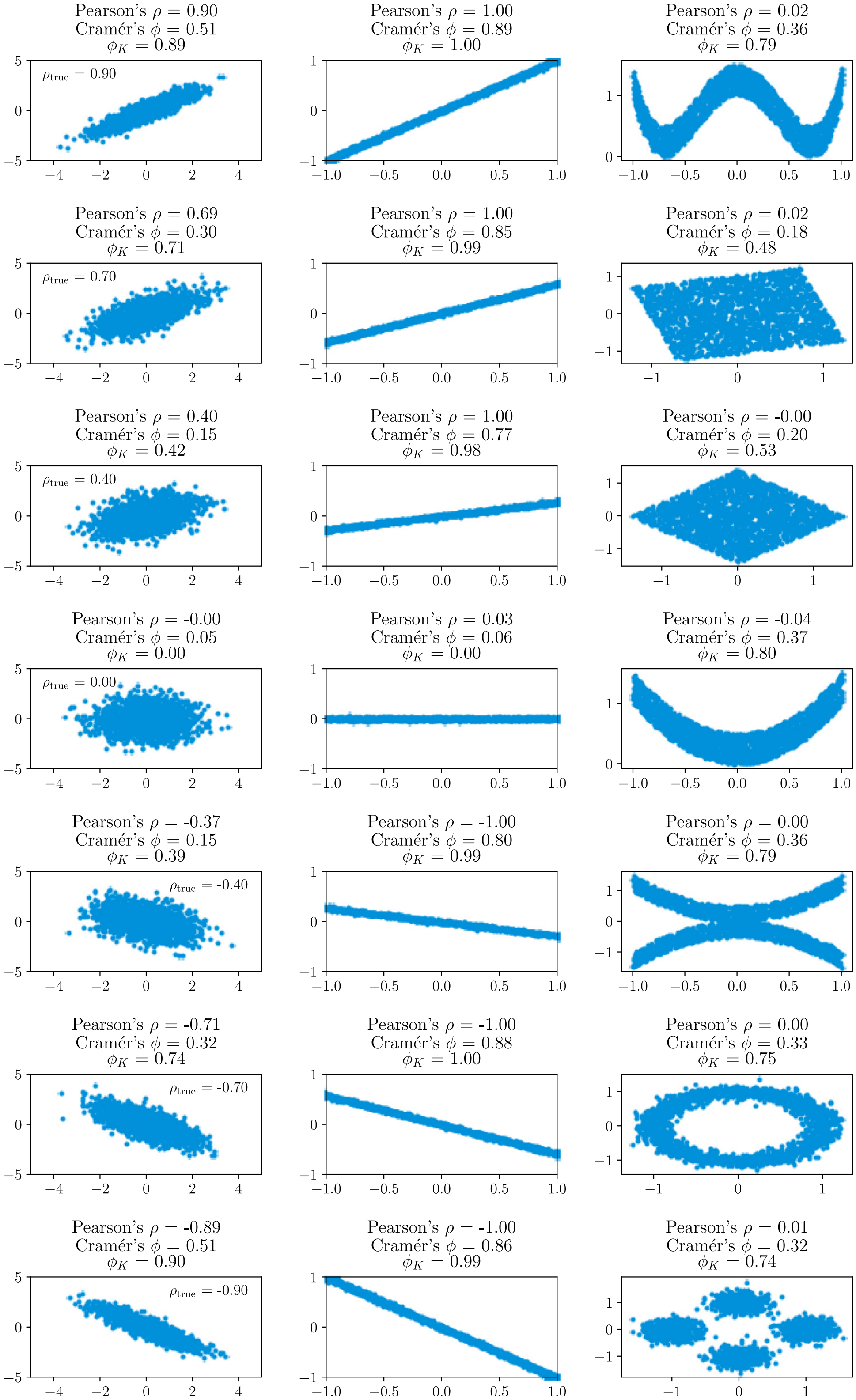


Fig. 3. Benchmark sample results for ϕ_K. Each synthetic data set contains 2000 data points. For the leftmost column, from top to bottom, the bivariate normal distributions have been generated with true correlations: $\{0.9, 0.7, 0.4, 0, -0.4, -0.7, -0.9\}$. For the middle column a linear data set is generated which is rotated around the origin. In the rightmost column various data sets are generated with nonlinear correlations. Note that these nonlinear correlations are well-captured by ϕ_K, while Pearson's ρ is close to zero for all cases.

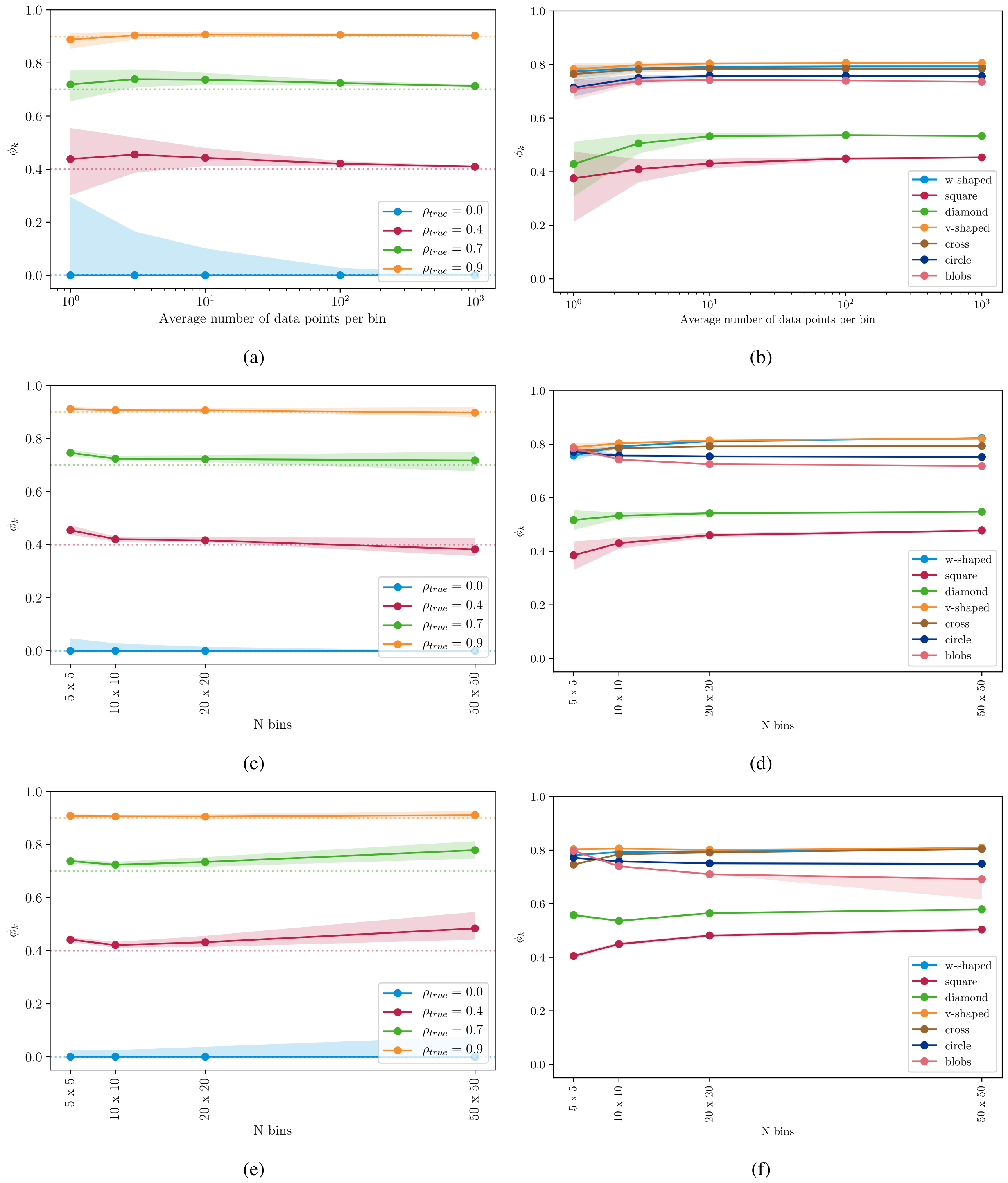


Fig. 4. Convergence rates for: (left) bivariate normal distributions with $\rho_{\text{true}} = 0.0, 0.4, 0.7, 0.9$ and (right) the benchmark shapes introduced in Fig. 3. Shown is the median value measured in 1000 simulations, while the error band indicates the 25th to 75th percentiles. For the bivariate normal distribution the true correlation is shown as dashed horizontal lines. In (a–b) the sample size is varied, with 10 × 10 bins. In (c–d) the average number of entries per bin is fixed to 100, and in (e–f) the sample size is fixed to 10 000. (c-f) have symmetric varying number of bins.

Table 1
A synthetic data set with car insurance data. The data set consists of 2000 records and is used to illustrate the calculations of ϕ_k, statistical significance (in Section 5) and outlier significance (in Section 6).

Car color	Driver age	Area	Mileage	Car size
Blue	60.4	Suburbs	3 339	XS
Blue	30.9	Suburbs	53 370	XL
Blue	18.5	Suburbs	112 557	XL
Green	40.9	Downtown	29 605	L
Gray	23.7	Downtown	15 506	M
Multi-color	60.3	Downtown	33 148	L
White	66.7	Suburbs	91 132	XL
Red	69.2	Downtown	152 445	XXL
Metallic	43.5	Hills	147 275	S
...	...	...	...	...

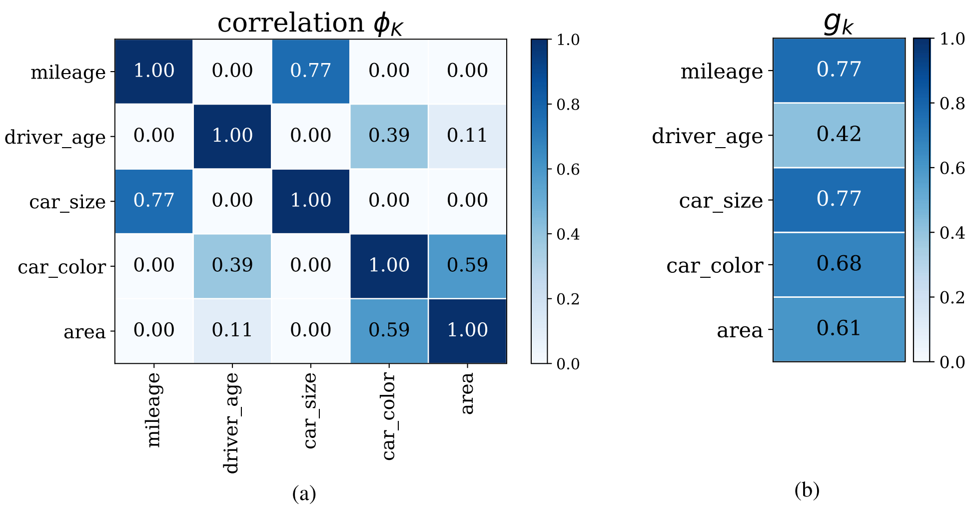


Fig. 5. Correlation coefficients calculated on the synthetic car insurance data set (Table 1) containing mixed variables types. (a) The ϕ_K correlation matrix. (b) The global correlations g_k.

4.4. Statistical noise correction

The calculation of ϕ_K contains a correction for statistical fluctuations: for any χ^2 value below the sample-specific noise threshold χ^2_{ped} of Eq. (15), indicating that no meaningful correlation can be determined, ϕ_K is set to 0 by construction. The impact of the noise correction is seen in Fig. 6a, showing the absolute value of Pearson's ρ, Cramér's ϕ, and ϕ_K measured for 1000 synthetic data sets with only 500 records each, simulated from a bivariate normal distribution with no correlation, and each binned in a 10×10 grid. Without absolute function applied, the distribution of Pearson's ρ values is centered around zero, as expected. The calculation of Cramér's ϕ results in a seemingly significant bump at 0.2. This cannot be interpreted as a meaningful correlation, but results from the statistical noise contributing to each sample's χ^2 value. ϕ_K is only evaluated when $\chi^2 > \chi^2_{\text{ped}}$. The noise threshold is set such that about 50% of all simulated zero correlation samples gets assigned $\phi_K = 0$. The remaining samples result in a wide distribution of ϕ_K values. (Without noise correction, the ϕ_K distribution shows a similar peak as Cramér's ϕ, at value 0.5.)

Fig. 6b shows ϕ_K as a function of true correlation, where ϕ_K is obtained from the median χ^2 value of 1000 synthetic data sets with 500 data points each. The median gives the most representative, single synthetic data sample. In the calculation of ϕ_K three configurations for the noise pedestal of Eq. (15) are tested: no pedestal, and $c \in \{0, 1\}$. No pedestal gives ϕ_K values that overshoot the true correlation significantly at low values. Configuration $c = 1$ undershoots: the calculation of ϕ_K turns on too late. Configuration $c = 0$ follows the true correlation most closely. The residual differences disappear for larger data samples, and we deem this acceptable for this level of statistics (with on average only 5 records per bin).

The sample-specific noise threshold χ^2_{ped} depends mostly on the number of filled cells, and stabilizes for larger sample sizes. Consequently, its impact is rather limited for large samples with a meaningful, nonzero correlation, typically having $\chi^2 \gg \chi^2_{\text{ped}}$. As also clear from Fig. 6a, for small sample sizes any correlation coefficient value found should always be held up against the significance of the hypothesis test of variable independence – the topic of Section 5 – before further interpretation.

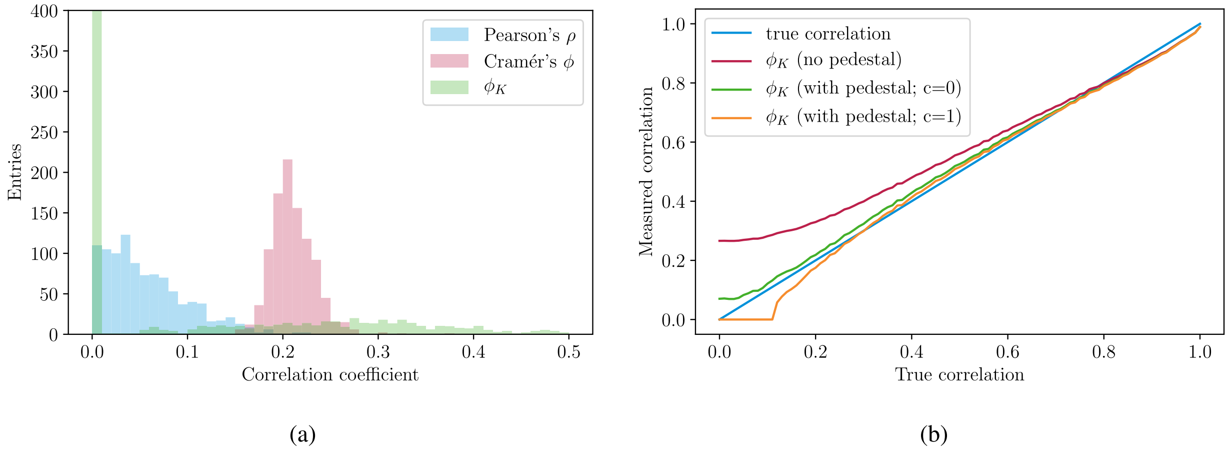


Fig. 6. (a) The correlation coefficients Pearson's ρ, Cramér's ϕ and ϕ_K, measured for 1000 synthetic data sets with 500 data points each, which are simulated using a bivariate normal with parameter $\rho = 0$. The absolute value of the Pearson's ρ is taken as the measured ρ can also take on negative values. (b) The median ϕ_K value measured in 1000 synthetic data sets containing 500 data points each, simulated using a bivariate normal distribution, as a function of true correlation ρ. The value of ϕ_K is evaluated using three different configurations of the noise pedestal parameter c (see Eq. (15)).

5. Statistical significance

In practice, when exploring a data set for variable dependencies, the studies of correlations and their significances are equally relevant: a large correlation may be statistically insignificant, and vice versa a small correlation may be very significant. The aim of this section is to describe a novel method of significance evaluation for contingency tables that works for any number of rows and columns, and in particular for sparse and low-statistics tables. In short, our method for p-value determination is a hybrid approach between brute force Monte Carlo simulation and the analytical formulas valid in the asymptotic (high-statistics) regime, valid over a large range of Z-scores, yet only requires several thousand simulated data samples to obtain high-precision Z-scores.

For the significance analysis of contingency tables several test statistics have been mentioned in Section 2. Fisher's exact test and Barnard's test are meant for 2×2 contingency tables only, and are ignored here for that reason. And for a contingency table, the G-test statistic is equal to the level of mutual information multiplied with two times the number of entries in the table. In the remainder of this work we therefore focus on Pearson's χ^2 contingency test and the G-test statistic.

Both Pearson's χ^2 test and the G-test asymptotically approach the χ^2 distribution (Barnard, 1992). For samples of a reasonable size (Cochran's rule on what defines "reasonable size" follows below), the default approach to obtain the p-value for the hypothesis test of variable independence is to integrate the χ^2 probability density function $g(x|k)$ over all values equal to or greater than the observed test statistic value t_{obs}:

$$p = \int_{t_{\text{obs}}}^{\infty} g(x|k)\,\mathrm{d}x\,, \tag{18}$$

with the p.d.f. of the χ^2 distribution:

$$g(x|k) = \frac{1}{2^{\mu}\,\Gamma(\mu)} \cdot x^{\mu-1} \cdot e^{-x/2}\,, \tag{19}$$

where $\mu = k/2$, $\Gamma(\mu)$ is the gamma function, and k is set to the number of degrees of freedom n_{dof}. The solution of this integral is expressed as the regularized gamma function. (The integral runs up to infinity, even though the contingency test has a maximum test statistic value. In practice the difference is negligible.) This approach holds for samples of a reasonable size, and when using the χ^2 test statistic or G-test.

For the independence test of n_{dim} variables, the number of degrees of freedom is normally presented (Bock et al., 2007) as the difference between the number of bins n_{bins} and model parameters n_{pars}

$$\begin{aligned} n_{\text{dof}} &= n_{\text{bins}} - n_{\text{pars}} \\ &= \left[\prod_{i=1}^{n_{\text{dim}}} n_i\right] - \left[\sum_{i=1}^{n_{\text{dim}}} (n_i - 1) \,+\, 1\right]. \end{aligned} \tag{20}$$

where n_i is the number of categories of variable i. Explained using Eq. (9), each dimension requires $(n_i - 1)$ parameters to model its p.m.f., which is normalized to one, and the p.m.f. product is scaled to the total number of events, which requires one more parameter. For just two variables this reduces to:

$$n_{\mathrm{dof}} = (r-1)(k-1)\,. \tag{21}$$

In practice Eq. (20) does not hold for many data sets, in particular for distributions with unevenly filled or unfilled bins. For example, in the case of two (binned) interval variables with a strong dependency. The *effective* number of degrees of freedom, n_{edof}, is often smaller than the advocated value, n_{dof}, and can even take on floating point values, because the number of available bins is effectively reduced.

The asymptotic approximation, Eqs. (18)–(19), breaks down for sparse data sets, for example for two (interval) variables with a strong correlation, and for low-statistics data sets. The literature on evaluating the quality of this approximation is extensive; for an overview see Refs. Agresti (2001) and Kroonenberg and Verbeek (2018). Cochran's rule of thumb is that at least 80% of the expected cell frequencies is 5 counts or more, and that no expected cell frequency is less than 1 count. For a 2×2 contingency table, Cochran recommends (Cochran, 1952, 1954) that the test should be used only if the expected frequency in each cell is at least 5 counts.

How to properly evaluate the p-value when the test statistic does not follow the χ^2 distribution, and Eq. (18) cannot be safely applied? A reasonable approach is to evaluate Eq. (18) directly with Monte Carlo data sets, sampled randomly from the distribution of expected frequencies. However, this approach quickly becomes cumbersome for p-values smaller than 0.1%, *i.e.* once more than 1000 simulations are needed for a decent p-value estimate, and practically impossible when needing at least a million simulations. Given that variable dependencies can be very significant, we prefer a common approach that works for both strong and weak dependencies and both low- and high-statistics samples.

In this section we propose another option: a hybrid approach where a limited number of Monte Carlo simulations is used to fit an analytical, empirical description of the χ^2 distribution. Specifically, we describe two corrections to Eqs. (18)–(19):

1. A procedure to evaluate the effective number of degrees of freedom for a contingency test;
2. A correction to Eq. (19) for low statistics samples, when using the G-test statistic.

We conclude the section with a prescription to evaluate the statistical significance of the hypothesis test of variable independence, and a brief overview of sampling methods to help evaluate the p-value.

5.1. Effective number of degrees of freedom

To obtain the effective number of degrees of freedom of any sample, we use the property of Eq. (19) that, for a test statistic distribution obeying $g(x|k)$, to good approximation the average value of $g(x|k)$ equals k. The effective number of degrees of freedom for any sample is obtained as follows:

Procedure description 2: effective number of degrees of freedom

1. For the two variables under study, the dependent frequency estimates form the factorized distribution most accurately describing the observed data. Using Monte Carlo sampling techniques, this distribution is used to randomly generate 500 independent synthetic data sets with the same number of records as in the observed data set.
   Optionally, sampling with fixed row and/or column totals may be chosen. A short discussion of sampling methods is held in Section 5.6.
2. For each synthetic data set, evaluate the G-test statistic using the statistically dependent frequency estimates, as detailed in Section 3. The choice of G-test statistic is motivated in Section 5.2.
3. The effective number of degrees of freedom, n_{edof}, is taken as the average value of the G-test distribution of all generated Monte Carlo samples.

As an example application of this process, Fig. 7 shows a "smiley" data set of two interval variables, consisting of two blobs and a wide parabola, which are binned into a 20×20 histogram, for which we can generate an arbitrary number of records. We use this as an example in the following sections, although the procedures described work with any data set.

The bottom two curves in Fig. 8 show n_{edof} obtained for this sample, as a function of the number of records in the data set, N, and evaluated using the G-test and χ^2 test statistic. Using Eq. (21), the advocated number of degrees of freedom of this sample equals 361. For both test statistics this number is only reached for very large sample sizes ($\geq 10^6$). and drops significantly for smaller values of N, where the drop is slightly steeper for the G-test statistic. The top two curves show the same data set on top of a uniform background of 1 record per cell, ensuring that each is always filled, again evaluated using the G-test or χ^2 test statistic. Now the G-test overshoots, and the χ^2 test statistic happens to level out at the expected value.

To understand the behavior of under- and overshooting, realize that n_{edof} relates directly to the distribution of dependent frequency estimates. By construction, the dependent frequency estimates of Eq. (9) make nonzero predictions

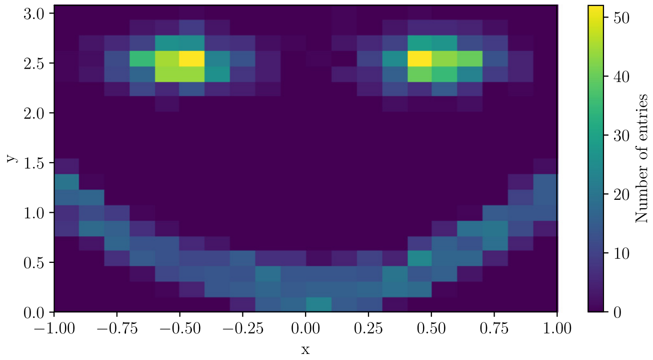


Fig. 7. Example "smiley" data set of two interval variables binned in 20 bins in the x and y directions.

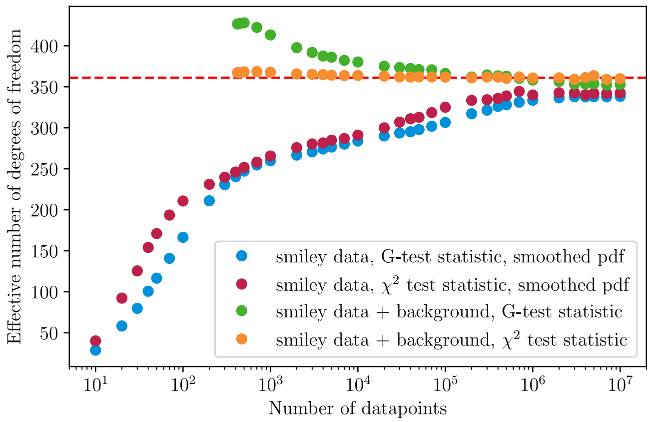


Fig. 8. The effective number of degrees of freedom as a function of the number of data points in the input data set (Fig. 7). The theoretical number of degrees of freedom, $n_{dof} = 361$, is indicated with the dashed line.

for each bin in the distribution, as long as the input data set contains at least one record per row and column. Under the assumption of variable independence, each bin in the distribution is expected to be filled.

Consider the bottom two curves of Fig. 8. For an uneven input distribution, for example two strongly correlated interval variables, one may expect many bins with low frequency estimates. A data set sampled randomly from a distribution with very low frequency estimates, such as the data set in Fig. 7, is likely to contain empty bins. On average, high-statistics bins contribute $n_{\mathrm{dof}}/n_{\mathrm{bins}}$ ($\lesssim 1$) to the G-test or χ^2 test statistic, but the low-statistics bins do not obey this regime. As an example, let us focus on the empty bins. By construction their contribution to the G-test is zero. The contribution to the χ^2 test statistic is nonzero: $\sum_i E_i$, where the sum runs over all empty bins. It is clear however, when $E_i \ll 1$, that this sum is relatively small and contributes only marginally. Taken over many randomly sampled data sets, this effect reduces the average value of the G-test or χ^2 test statistic distribution to lower values, and likewise decreases n_{edof} compared with n_{dof}. For the top two curves, by construction $E_i > 1$ for each bin, bringing them closer to the nominal regime and increasing the G-test and χ^2 test statistics. For a discussion of the contribution of low-statistics contingency table cells to the χ^2 test statistic, see Ref. Yates (1934).

In summary, depending on the shape and statistics of the input data set, the effective number of degrees of freedom of a contingency table can differ significantly from the advocated value of n_{dof} (Eq. (20)). To be certain of the effective value to use, we derive it as the average value of the test statistic distribution, which is obtained with Monte Carlo simulations of the expected frequency distribution.

5.2. Modified χ^2 distribution

Given a large enough data sample, and given the hypothesis that the observed frequencies result from a random sampling from the distribution of expected frequencies, the G-test statistic can be approximated by Pearson's χ^2. (The approximation is obtained with a second-order Taylor expansion of the logarithm around 1.) In this scenario both the G-test and χ^2 value are described by the χ^2 distribution of Eq. (19), with the same number of degrees of freedom, and applying any one test leads to the same conclusion.

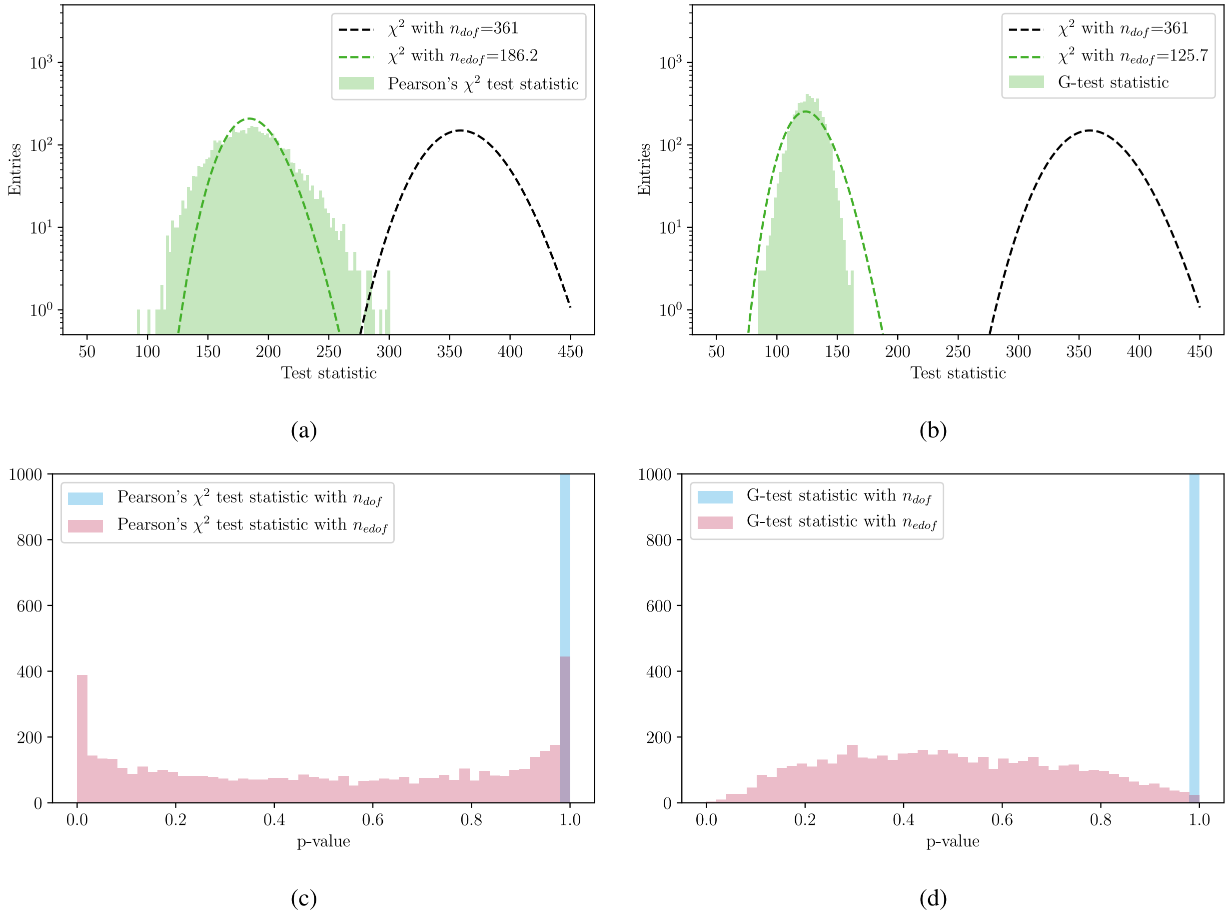


Fig. 9. The simulated distribution of χ^2 and G-test statistics using the smiley data set as input (Fig. 7). (a) The χ^2 distribution (green) is wider than the expected distribution (dashed lines), and (b) the G-test distribution (green) is narrower. The corresponding p-value distributions are shown in panels (c) and (d), both when using n_{dof} (blue) and n_{edof} (pink). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

For low statistics samples – to be more specific, samples with many bins of low expected and observed frequencies – the distributions of G and χ^2 start to differ, and both distributions diverge from the nominal χ^2 distribution. This can be seen in Fig. 9, which uses the smiley data set of Fig. 7 as input. The simulated distribution of test statistics is wider than the χ^2-distribution in case of the Pearson χ^2-test statistic (Fig. 9a) and narrower than the χ^2-distribution in case of the G-test statistic (Fig. 9b). This results in p-value distributions with elevated frequencies around zero and one for the Pearson χ^2-test statistic (Fig. 9c) and lower frequencies near zero and one for the G-test statistic (Fig. 9d). Note that here the effective number of degrees of freedom is much lower than the theoretical value; using n_{dof} in the p-value calculation results in an uneven distributions peaked towards one.

This section addresses the question whether the test statistic distribution for the contingency test can be modeled for all sample sizes, knowing that Eq. (19) cannot be safely used for low statistics data sets. In particular we are interested in assessing the p-values of large test statistic values, coming from possibly strong variable dependencies. To evaluate these correctly, it is important to properly model the high-end tail of the test statistic distribution.

We observe empirically that for low-statistics samples the G-test statistic distribution converges towards a Gaussian distribution $G(x|\mu,\sigma)$, with mean $\mu = n_{\mathrm{edof}}$ and width $\sigma = \sqrt{n_{\mathrm{edof}}}$. For high-statistics samples the distribution is modeled by $g(x|k)$, with $k = n_{\mathrm{edof}}$ degrees of freedom. Experimentally we find that, for any sample size, the G-test statistic distribution can be well described by the combined probability density function $h(x|f)$:

$$h(x|f) = f \cdot g(x|n_{\mathrm{edof}}) + (1-f) \cdot G(x|n_{\mathrm{edof}}, \sqrt{n_{\mathrm{edof}}}), \tag{22}$$

where the parameters of $g(x|k)$ and $G(x|\mu,\sigma)$ are fixed as above, and f is a floating fraction parameter between [0, 1]. The presence of a Gaussian term we explain as follows. For low-statistics samples the relative amount of statistical noise per cell becomes large, up to 100% for cell counts of 0 or 1. The G-test in Eq. (8) sums over all cells in the contingency table. With large noise levels the terms in the sum become sufficiently independent, and the sum behaves approximately like a

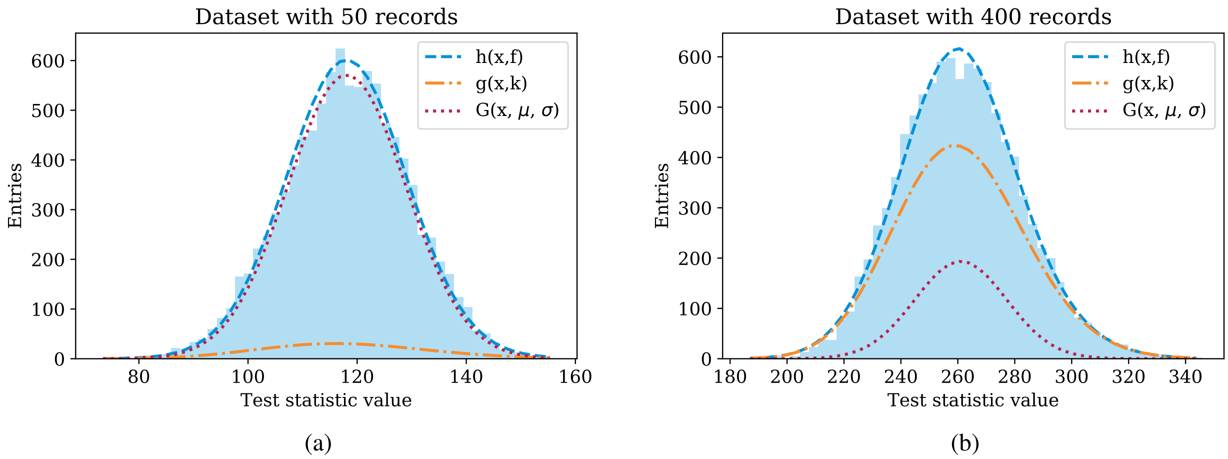


Fig. 10. The G-test statistic distribution for two smiley data sets containing (a) 50 and (b) 400 data points. The distribution is modeled with the $h(x|f)$ distribution. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

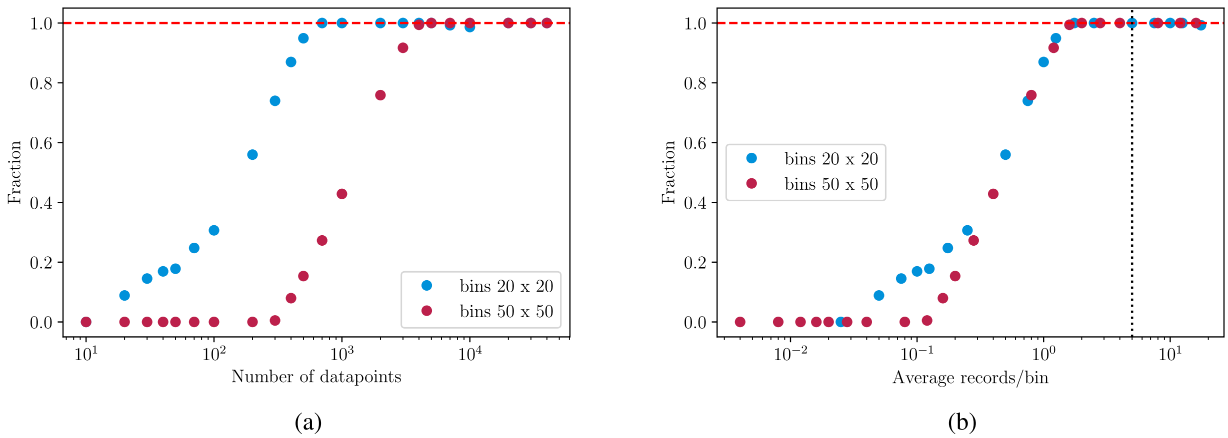


Fig. 11. (a) The fit fraction f as a function of the number of records per simulated data set, N. (b) The same data points, but here f is shown as a function of the average number of records per bin.

Gaussian distribution, as established by the central limit theorem. The argument for using Eq. (22) is that this distribution can morph smoothly between the asymptotic and low-statistics regimes. Below we use $h(x|f)$ as the modified χ^2 p.d.f. to model the G-test statistic distribution for any data set.

Fig. 10 shows the results of binned log-likelihood fits of $h(x|f)$ to two G-test statistic distributions, each with 10k entries generated with the procedure of Section 5.1, using the smiley data set with 20×20 bins with: (a) $N = 50$ and (b) $N = 400$ records for the simulated data sets. Clearly, these distributions are not well modeled using $g(x|n_{\text{edof}})$ or $G(x|n_{\text{edof}}, \sqrt{n_{\text{edof}}})$ alone. The fit of $h(x|f)$ can separate the two components of the p.d.f. given that the RMS-value of $g(x|n_{\text{edof}})$ is $\sqrt{2n_{\text{edof}}}$ and the width of the Gaussian is fixed to $\sqrt{n_{\text{edof}}}$. For $N = 50$, the distribution is dominated by the Gaussian, and for $N = 400$ by the theoretical χ^2 distribution. Note that $G(x|n_{\text{edof}}, \sqrt{n_{\text{edof}}})$, when present, contributes to the core of the distribution while $g(x|n_{\text{edof}})$ dominates in the tails.

Fig. 11 uses a similar setup, with 20×20 or 50×50 bins, where the fit fraction f is shown as a function of (a) the number of records per simulated data set, N, and (b) the average number of records per cell, $\bar{n}$. The fraction f rises as a function of sample size, such that $h(x|f)$ turns into $g(x|n_{\text{edof}})$ for large enough data sets. With 20×20 bins, for a fraction of 0.50 (0.99) the approximately sample size equals 175 (700), and the average number of entries per cell equals 0.4 (1.8). Note that the fraction reaches 1 well before n_{edof} reaches the advocated value of n_{dof} in Fig. 8.

In summary, to assess the p-value for the hypothesis test of variable independence, in this work we choose to work with the G-test statistic, and not Pearson's χ^2, for two reasons:

1. We manage to describe the G-test statistic distribution most successfully for any sample and sample size.
2. As seen from Fig. 9b, for a large observed test statistic value, corresponding to a large significance of variable dependency, applying the naive formula of Eq. (18) over-covers, *i.e.* gives a conservative p-value (the green distribution is narrower than expected).

We use the distribution $h(x|f)$ of Eq. (22) as modified χ^2 distribution in Eq. (18) to assess the p-value for the hypothesis test.

5.3. Evaluation of significance

The statistical significance of the hypothesis test of any variable independence is obtained with the following procedure (see Lin et al., 2014).

Procedure description 3: "hybrid" significance evaluation of variable-pair dependency

1. For the contingency table of the input data set calculate the average number of entries per cell, $\bar{n}$. If $\bar{n} < 4$, set $n_{\mathrm{sim}} = 2000$, else $n_{\mathrm{sim}} = 500$ samples.
2. Follow the procedure of Section 5.1 to generate n_{sim} synthetic data sets based on the dependent frequency estimates of the input data set. For each synthetic data set evaluate its G-test value. Take the average of the G-test distribution to obtain n_{edof}.
3. If $\bar{n} < 4$, to obtain f fit the probability density function $h(x|f)$ to the G-test distribution, with n_{edof} fixed. Else, skip the fit and set $f = 1$.
4. With this fraction, use Eq. (18) with $h(x|f)$ as modified χ^2 distribution to obtain the p-value for the hypothesis test, using the G-test value from data as input.
5. The p-value is converted to a normal Z-score:

$$Z = \Phi^{-1}(1-p) \; ; \quad \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} \, \mathrm{d}t \, , \tag{23}$$

   where Φ^{-1} is the quantile (inverse of the cumulative distribution) of the standard Gaussian, *e.g.* Z is the significance in 1-sided Gaussian standard deviations. (For example, the threshold p-value of 0.05 (95% confidence level) corresponds to $Z = 1.64$.)
6. When the p-value is too small to evaluate Eq. (23) numerically, at $p \lesssim 10^{-310}$, anyhow a very strong variable dependency, Z is estimated using Chernoff's bound (Lin et al., 2014) to ensure a finite value. Let $z \equiv G/n_{\mathrm{edof}}$, Chernoff states when $z > 1$:

$$p \leq f \cdot (ze^{1-z})^{n_{\mathrm{edof}}/2} \, , \tag{24}$$

   where we safely ignore the contribution from the narrow Gaussian in $h(x|f)$. This is converted to Z with the approximation (valid for large $Z > 1.5$):

$$Z = \sqrt{u - \log u}; \quad u = -2 \log(p\sqrt{2\pi}) \, . \tag{25}$$

This significance procedure is illustrated in Fig. 12, which shows the p-value distributions of the two G-test distributions of Fig. 10, with $N = 50$ and $N = 400$ records per sample. The two p-value distributions in each figure have been calculated in two ways.

1. Using the original χ^2 distribution $g(x|k)$ of Eq. (18), with the effective number of degrees of freedom, n_{edof}. This results in the blue distributions.
2. Fitting each test statistic distribution with $h(x|f)$ of Eq. (22), and using that to calculate the p-values, resulting in the red distributions.

The blue distributions drop around zero and one, in particular for the low statistics sample ($N = 50$). This is because the G-test distribution is more narrow than the χ^2 distribution, as shown in Fig. 9. The red p-value distributions, evaluated with $h(x|f)$, are uniform, as desired in both setups.

The choice of simulating 2000 synthetic data sets for the fit of $h(x|f)$ is a compromise between accuracy and speed. With this number, Z typically varies at the level of 0.04, and is calculated in just a fraction of a second. The procedure described here has been validated with all benchmark samples shown in Fig. 3. As with the smiley data set in Fig. 12, for each benchmark sample the p-value distributions evaluated with $h(x|f)$ are uniform.

To illustrate the impact of these two corrections, let us apply the statistical procedure to a low-statistics and sparse data sample. A smiley data set with 100 entries, in a histogram with 20 × 20 bins, has correlation value $\phi_K = 0.73$ and test statistic value $G = 227.4$. The Z calculation is done in three increasingly refined ways:

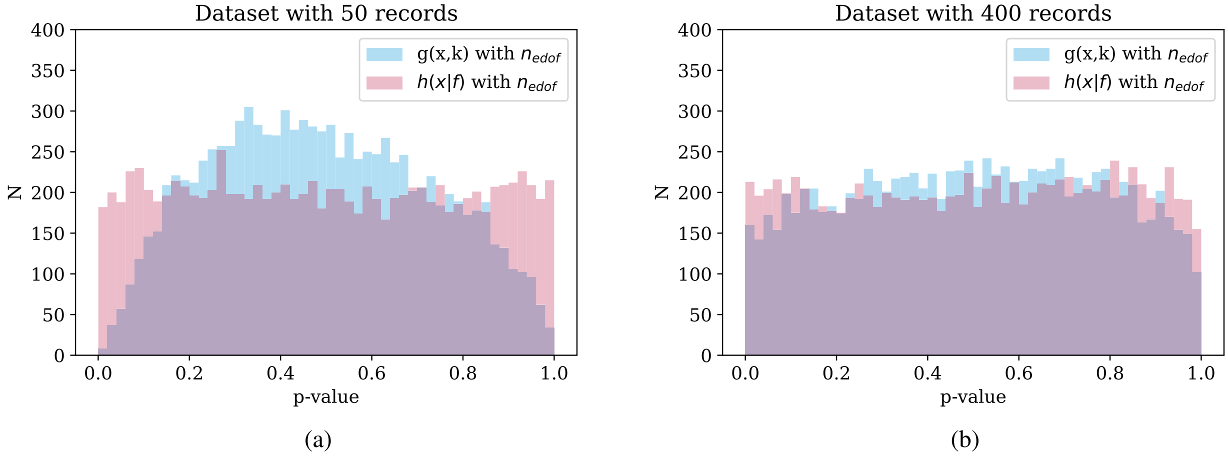


Fig. 12. The p-value distributions corresponding of the two G-test distributions of Fig. 10, with (a) $N = 50$ and (b) $N = 400$ records per sample. See the text for a description of the two p-value calculations performed. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

1. *The asymptotic approximation*: using $n_{\text{dof}} = 361$ and the asymptotic χ^2 distribution $g(x|k)$ gives: $Z = -5.7$;
2. *Effective number of degrees of freedom*: using $n_{\text{edof}} = 189.3$ and the asymptotic χ^2 distribution $g(x|k)$ results in: $Z = 1.9$;
3. *Modified χ^2 distribution*: with $n_{\text{edof}} = 189.3$, the modified χ^2 distribution $h(x|f)$, and fit fraction $f = 0.10$ one finds: $Z = 2.5$.

In this example, between the three approaches the Z-value increases by more than 8 units! Typically, using the effective number of degrees of freedom gives the largest correction to Z, and the modified χ^2 distribution only gives a small correction on top of that.

A related example is seen in Fig. 13a, which is based on a set of bivariate normal distributions with 10×10 bins and 250 entries each, with a varying set of ρ values, each resulting in a different Z-score in the range of $[-4, 4]$. The Z-scores on the x-axis are evaluated with the MC-based approach using the G-test statistic, each based on 1M simulations, which are taken here as ground truth values. (The MC-based method can be either G-test or χ^2 based, but the two results agree so well, at the level of 0.01 units in Z, that we only show the G-test results.) On the y-axis are shown the asymptotic method and our proposed hybrid solution, based on both the G-test and χ^2 test statistic. The G-test hybrid approach agrees best with the MC-based Z-scores, with an average absolute difference of 0.02 and a maximal difference of 0.1, where the χ^2 hybrid approach has a maximum difference of 0.7. In this setup the two asymptotic methods differ typically 4–6 units with the MC-based approach.

Fig. 13b–d compares the Z-scores of the hybrid method with the MC method, evaluated on a bivariate normal distribution with varying: correlation parameter ρ, sample size, and number of x- and y-bins respectively. We can compare the methods up to a Z-score of 5. Beyond that the MC method needs more than 10 million generated samples to evaluate reliably. The hybrid method does not suffer from this effect. In (b–d) each point on the curves is the Z-score corresponding to the median G-test value of 40k generated bivariate normal data samples, with a given value of ρ, sample size or binning. Again the hybrid and MC methods agree well, up to 0.1 unit in Z. As expected, the Z-score goes up as a function of increasing correlation strength and sample size. With an increasing number of bins and a fixed sample size, the relative amount of noise per bin goes up, thereby decreasing the overall statistical significance.

Based on our findings, for any sample size we recommend the p-value of the null hypothesis of no correlation to be calculated with the modified χ^2 distribution $h(x|f)$, based on the G-test statistic and using n_{edof} degrees of freedom. If not, the p-value may over-cover for strong variable dependencies and at low-statistics, resulting in a Z-value that is too small, possibly by multiple units. This is important to know, as it can lead to rather incorrect conclusions regarding the studied variable dependency.

5.4. Modeling alternate samples

We have investigated if $h(x|f)$ of Eq. (22) can also describe the G-test distribution of data samples corresponding to an arbitrary alternate model, where the G-test value of each sample is obtained with its factorized expectation values. For any alternate sample with a true variable dependency this is clearly an incorrect model, resulting in high G-test values. Asymptotically this distribution is modeled quite well by the Gaussian term in $h(x|f)$, which we understand as follows. For any alternate model with a true variable dependency the terms in the sum of the G-test formula in Eq. (8) become sufficiently independent such that their sum behaves like a normal distribution, as posed by the central limit theorem. For

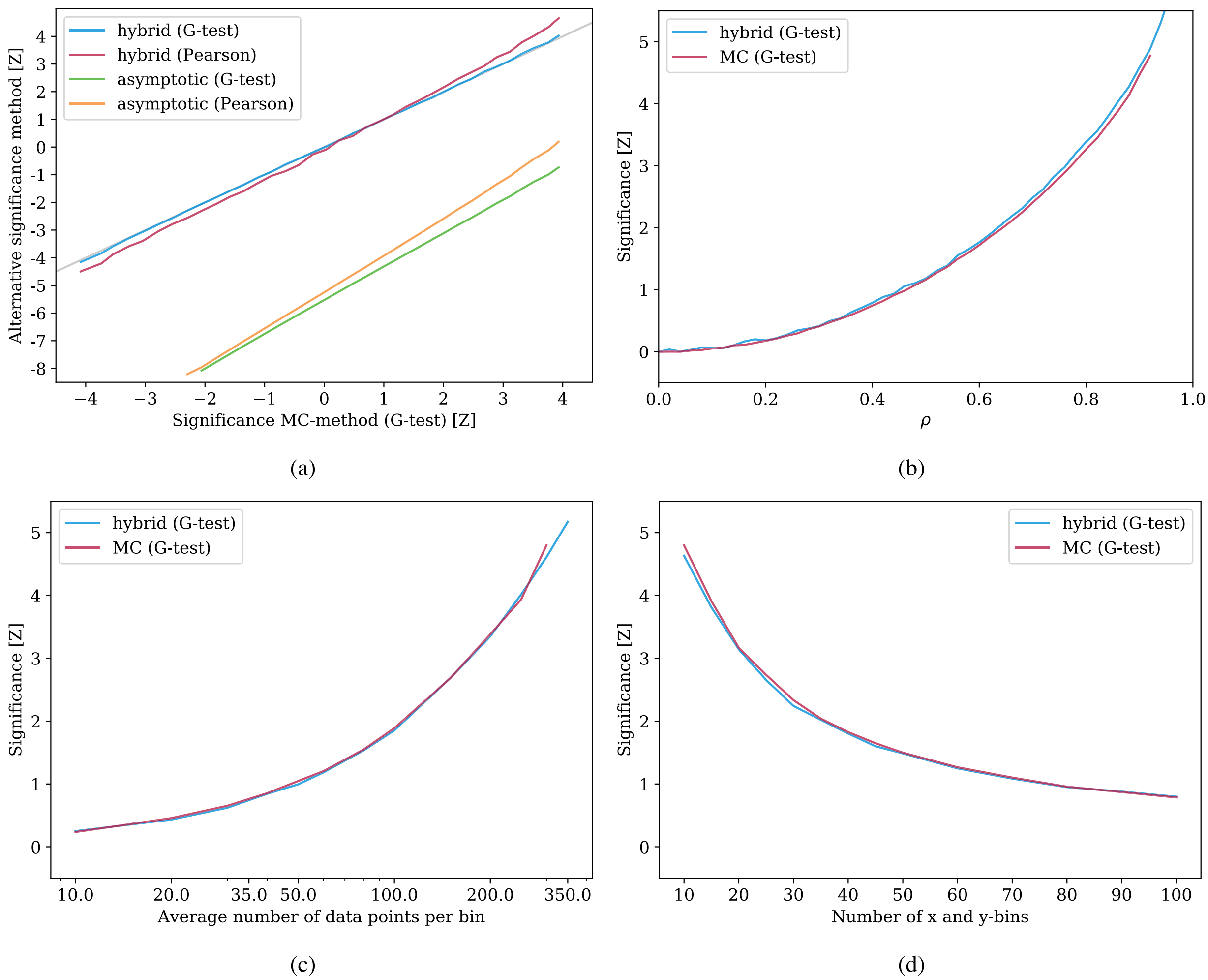


Fig. 13. (a) Comparison of Z-scores, as evaluated on a perfect bivariate normal distribution with varying correlation parameter, between the MC method (on the x-axis), interpreted as ground truth, and the asymptotic method or our proposed hybrid solution (on the y-axis). As visual aid a gray diagonal line has also been drawn, which reflects the perfectly calibrated Z-score. (b–d) Compares the Z-scores of the hybrid and MC methods for a bivariate normal sample with: (b) 50 entries and 10×10 bins, with varying correlation parameter ρ; (c) $\rho = 0.05$ and 10×10 bins, as a function of sample size; and (d) $\rho = 0.05$ and 30 000 entries, as a function of number of x- and y-bins, *i.e.* the total number of bins is the square.

low-statistics samples the G-test distribution turns asymmetric. This asymmetry is in turn modeled by the χ^2-distribution component; we have no good rationale for using this particular shape, however.

In general we observe that this description works quite well for the benchmark shapes in Fig. 3. However, we do find edge cases where $h(x|f)$ is not a good fit, in particular for low statistics samples with coarse binning. In conclusion, we do not recommend to use the hybrid approach to model the G-test distribution of alternate samples, but to use Monte Carlo sampling instead.

5.5. Example significance matrix

Fig. 14 shows the significance matrix determined for the car insurance data set of Table 1, containing 2000 data points, evaluated both with the hybrid method advocated in this work and the asymptotic approach. Compared with the correlation matrix of Fig. 5, the low ϕ_K values happen to be statistically insignificant, but the higher values are very significant. The car insurance data set is not sparse, and with this sample size the two significance methods differ only little, at the level of 0.25 units in Z.

5.6. Sampling approaches

Based on the statistically dependent frequency estimates, three sampling approaches are offered to generate synthetic data sets for testing the hypothesis of no variable association:

1. *Multinomial sampling*: with only the total number of records fixed. The hypothesis of no association is independent of the row and column variables.
2. *Product-multinomial sampling*: with the row or column totals fixed in the sampling. The hypothesis of no association is also called homogeneity of proportions. This approach is commonly used in cohort and case-control studies.
3. *Hypergeometric sampling*: both the row or column totals are fixed in the sampling. This approach is also known as Fisher's exact test. We use Patefield's algorithm (Patefield, 1981) to generate the samples.

Asymptotically the three different sampling approaches lead to the same result. The default approach used in this paper is multinomial sampling. There is a long debate about sampling design for tests of variable independence; for a discussion and further references see Ref. Kim and Agresti (1997).

6. Interpretation of relation between two variables

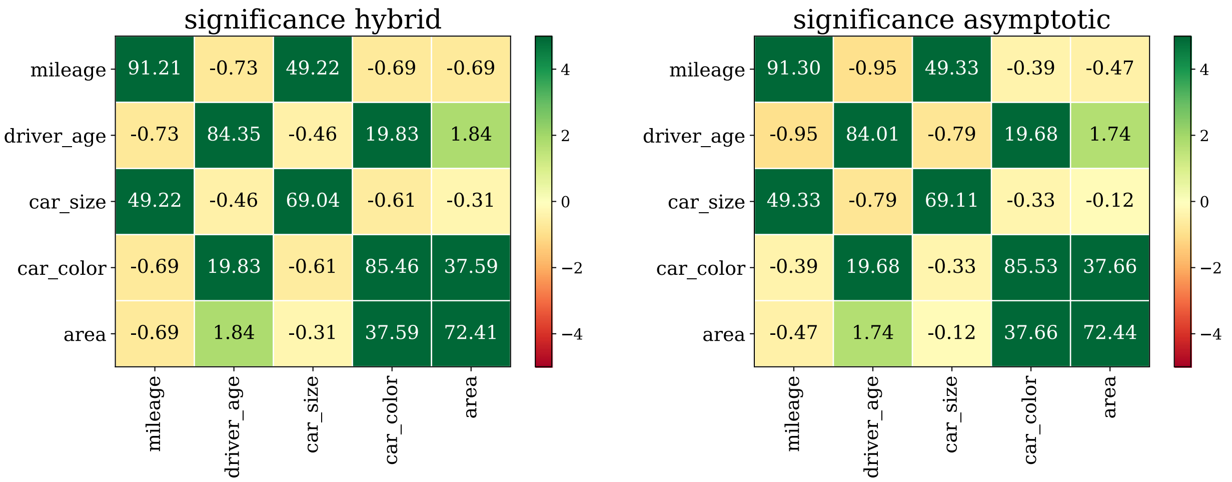
After the evaluation of ϕ_K and its significance, the specific relationship between two variables is typically inspected. Note that ϕ_K measures only the strength of the correlation, not the direction. This section presents a new, refined method for inspecting the relationship between two dependent variables. In particular we are interested in finding observed excesses or deficits of records with respect to expected values in the contingency table, as these are the cause behind any dependency found. To facilitate this interpretation, a refined significance evaluation of any such excesses or deficits is presented here, one that combines a handful of subtle statistics techniques from various fields. This evaluation method is valid also for contingency tables with very low cell counts. The related visualization technique can be used to help interpret the dependency between any two mixed variable types.

The statistical significance for each cell in the table is obtained from a hypothesis test between a background-only and signal-plus-background hypothesis for a Poisson process. Such hypothesis tests, *i.e.* for the presence of new sources of (Poisson) counts on top of known "background" processes, are frequently performed in many branches of science, for example gamma ray astronomy and high energy physics, and have been discussed extensively in the literature (Cousins et al., 2008). We employ a measure of statistical significance commonly used in both fields, one that accounts for the mean background rate having a non-negligible uncertainty. The background estimate and its uncertainty have been derived from an auxiliary or side-band measurement, typically assumed to be a Poisson counting setup, as in the case of the ABCD estimate of Section 3.2 (see Linnemann, 2003; Lancaster, 1961).

Procedure description 4: significance evaluation of observed over expected cell counts

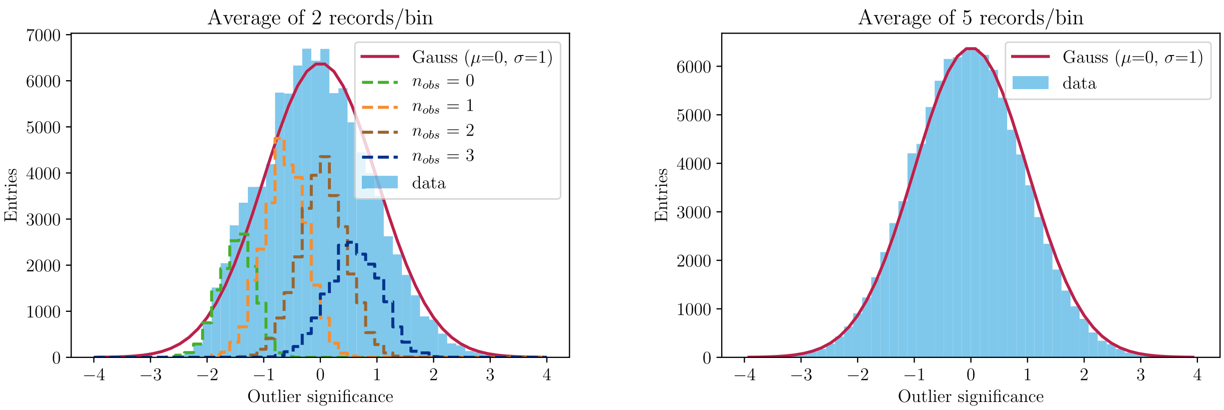
1. We use as background estimate the statistically independent frequency estimate (and related uncertainty) of Eq. (10) (11).
2. The hybrid Bayesian-Frequentist method from Linnemann (2003) is used to evaluate the probability of the hypothesis test (p-value). Per cell, Linneman's probability calculation requires the observed count n_o, the expected count n_e, and the uncertainty on the expectation σ_e:
$$p_B = B\big(1/(1+\tau),\ n_o,\ n_e\,\tau + 1\big)\,, \tag{26}$$
where B is the incomplete Beta function, and $\tau = n_e/\sigma_e^2$. We apply four corrections on top of this calculation.
3. The incomplete Beta function returns no number for $n_o = 0$, when by construction the p-value should be 1.
4. The incomplete Beta function is undefined when $\sigma_e = 0$, in which case we simply revert to the standard Poisson distribution.
5. The incomplete Beta function always returns 1 when $n_e = 0$, irrespective of n_o and σ_e. The scenarios $n_o = 0$ and $\sigma_e = 0$ are captured by the previous two fixes. In all other cases we set $n_e = \sigma_e$ before evaluating Eq. (26). In particular, this procedure prevents a significance of minus infinity for such low statistics cell.
6. As we combine an integer-valued measurement (namely the observed frequency) with a continuous expectation frequency and uncertainty, resulting in a continuous (combined) test statistic, we correct p_B to Lancaster's mid-P value (Lancaster, 1961), which is the null probability of more extreme results plus only half the probability of the observed result:
$$p = P(s = \text{observed}\,|\,\text{background})/2 + P(s > \text{observed}\,|\,\text{background})\,, \tag{27}$$
with s the integrated-over number of cell counts. (The standard p-value definition is: $p = P(s \geq \text{observed}\,|\,\text{background})$.) This p-value is then translated into the Z-value using Eq. (23). When observing the expected frequency by construction Lancaster's mid-P value (Z-value) is close to 0.5 (0), even at low statistics. Likewise, for background-only samples the Lancaster's mid-P correction centers the Z distribution around zero.

Concerning step 5 of the procedure, we wish to add the following remark. When $n_e = 0$, B or C is zero in Eq. (10), so Eq. (11) typically gives $\sigma_e < 1$. For example, for $n_e = 0$, $\sigma_e = 0.14$, and $n_o = 0$ (1), correction three to p_B results in



(a) (b)

Fig. 14. The significance matrix, showing the statistical significances of correlated and uncorrelated variable pairs. The color scale indicates the level of significance, and saturates at ± 5 standard deviations. (a) Using the hybrid method advocated in this work, and (b) using the asymptotic approach.



(a) (b)

Fig. 15. The distribution of outlier significances measured in 1000 randomly generated data sets of two variables obeying a uniform probability mass distribution for a data set containing (a) 200 and (b) 500 records, collected in a 10×10 contingency table. Normal distributions have been overlaid. In plot (a) the Z distributions from 0, 1, 2, and 3 observed entries per cell are shown as well.

$Z = -0.29$ (1.10). Varying n_e between $\sigma_e/2$ and $3\sigma_e/2$ gives a maximum absolute shift in Z of 0.05 (0.12). To do outlier detection, we deem this level of systematic uncertainty acceptable.

Fig. 15a shows the Z distribution from 1000 randomly generated samples of two variables obeying a uniform probability mass distribution, *i.e.* the samples have no variable dependency. Each sample contains only 200 records collected in a 10×10 contingency table, so on average each cell contains 2.0 records. As can be seen from the Gaussian curve, even for such low statistics samples the Z distribution is fairly consistent with a normal distribution, albeit slightly shifted towards negative values. Fig. 15b shows a similar distribution, built from samples with on average 5.0 records per contingency table cell. Clearly, with more statistics the distribution converges to the normal distribution relatively quickly.

A comment on low-statistics samples. With an average of less than 1.0 records per bin, the Z distribution gets more distorted, and breaks up into individual peaks of 0, 1, 2, etc. observed entries per cell. The distribution peaks at negative Z values, corresponding to no observations, and the tail at negative Z gets truncated. Relevant here, the mean of the distribution remains close to zero, its width is (slightly less than) one, and the positive tail is similar to that of a normal distribution.

We can now filter out significant excesses or deficits of records over expected values in the contingency table, with a simple procedure that works for nearly all sample sizes. One simply demands $|Z|$ to be greater than a specified value,

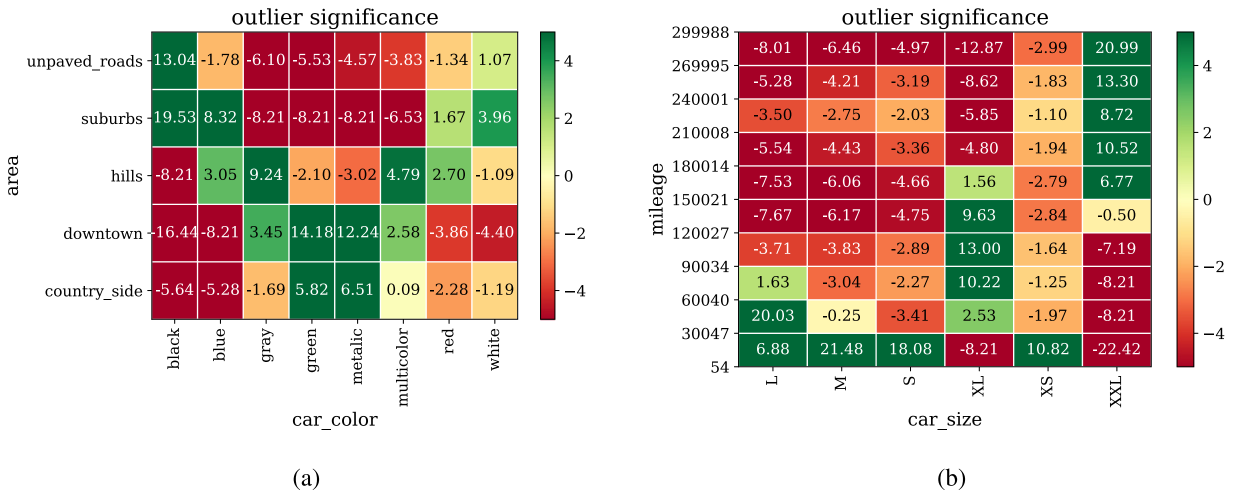


Fig. 16. Significances of excesses or deficits of records over the expected values in a contingency table for (a) the categorical variables "car color" and "area" and (b) the ordinal variable "car size" and the interval variable "mileage", measured on the synthetic data of Table 1.

e.g. 5 standard deviations. For two variables with a dependency, note that excesses and deficits always show up together, since the frequency estimates of Section 3.2 smooth the input distribution.

Two example contingency tables are shown in Fig. 16, one for a combination of categorical variables, and one for the combination of an interval and an ordinal variable, both based on the synthetic car insurance data of Table 1. Per cell each figure shows the evaluated Z-score. For example, black-colored cars occur significantly more in suburbs and significantly less downtown, and huge cars have significantly higher mileage.

In practice these turn out to be valuable plots to help interpret correlations, in particular between categorical variables. In essence, for a data sample with a dependency, the contingency table cells with large $|Z|$ values show the variable dependency. The idea of visualizing the variable dependency through the Z-scores of its contingency cells is not new. For example, R (Anon, 0000b) provides an option to show the score $(n_o - n_e)/\sqrt{n_e}$ for each cell, using dependent estimates for number of entries, and where the assumption is made that a high-statistics (Gaussian) regime applies. Our implementation, however, uses an independent estimate for the expected number of entries in each cell, accounts for the uncertainty on this estimate, and also for the possibility of low numbers of observed entries per cell, which is new for contingency tables.

7. Two practical applications

The methods described in this paper are useful for data exploration and model building purposes, and can be applied to numerous analysis problems. Given a set of mixed-type variables and using the methods described in this work, one can: find variable pairs that have (un)expected correlations, evaluate the statistical significance of each correlation, interpret the dependency between each pair of variables. Two interesting applications are given below: modeling the frequency of insurance claims and finding unexpected answers in surveys. Although a lot of literature exists on these problems and a detailed discussion of falls outside of the scope of this paper, we do want to cover both briefly, as our methods help give refined insights that complement those of existing techniques.

7.1. Modeling the frequency of insurance claims

One interesting application is the modeling of numbers of expected insurance claims, *e.g.* car damage claims as a function of car type, type of residential area, mileage, age of driver, etc. — a set of variables with a mixture of types. The results of this work can be used to select the variables with the strongest correlation with car damage, and in particular to identify the most significant cross-terms to be included in the insurance model. In our experience this is often done in a manual, experimental manner, whereas our approach allows one to compile a list and model the terms automatically.

The aggregate loss incurred by an insurer S is the total amount paid out in claims over a fixed time period: $S = \sum_{n=1}^{N} s_n$, where s_n is an individual claim amount, known as the severity, and N is the total number of claims paid out in the time period. Traditionally it is assumed that the individual claim amounts are mutually independent, and that N does not depend on the values of the claims. The total expected severity is then expressed as a product of the expected number of claims times the average claim amount: $E(S) = E(N) \cdot E(s)$, where each term can be estimated separately. When a vector of variables $\vec{x}$ is available at the individual claim level, this information is incorporated through two independent generalized linear models (GLMs): one for the claim frequency N, and the other for the severity s. See Ref. Garrido et al. (2016) for more information.

Here we focus on the GLM for modeling the frequency of insurance claims. Suppose that claims data are available for m different classes of policy holders, and that class i has N_i claims. Assume the claim frequency for each class is Poisson distributed, $N_i \sim P(\nu_i)$, where ν_i is the expectation for N_i. Let $\vec{x}_i = (x_{i0}, \ldots, x_{ik})$ be the vector of variables for class i at claim level, with the baseline convention that $x_{i0} \equiv 1$. One writes:

$$\nu_i = E(N_i|\vec{x}_i) = g^{-1}(\vec{\alpha} \cdot \vec{x}_i)\,, \tag{28}$$

where $\vec{\alpha} = (\alpha_{i0}, \ldots, \alpha_{ik})$ is a vector of regression constants. (Sometimes the ratio of claims to no claims per class of policy holders is modeled instead.) In GLM terminology g is the link function. When the frequency GLM uses a logarithmic link function, Eq. (28) simplifies to:

$$\nu_i = e^{\vec{\alpha}\cdot\vec{x}_i}\,, \tag{29}$$

yielding a simple rating structure which ensures that $\nu_i > 0$. The logarithmic function reflects the common practice that each variable alters the baseline claim rate by a multiplicative factor.

In initial GLM risk models, no relations are typically assumed between the input variables, and each variable category j (or interval bin) is one-hot-encoded, $x_{ij} \in \{0, 1\}$, and assigned one model parameter. The number of regression parameters per variable equals its number of categories or interval bins. Note that it is common practice to merge low-statistics categories until they contain sufficient records. Take the example variables of residential area and car type, each with multiple categories. Three classes of policy holders could be: "city, small car", "city, SUV", and "countryside, SUV", where the first two share the regression parameter α_{city}, and the last two the regression parameter α_{SUV}. The predicted, factorized number of claims for class "city, SUV" simply reads: $N_0\, e^{\alpha_{\text{city}}+\alpha_{\text{SUV}}}$, where $N_0 \equiv e^{\alpha_0}$ is the nominal number of claims shared between all classes, and $x_{\text{city}} = x_{\text{SUV}} = 1$. In a refinement modeling step, to improve the factorized estimates, cross-terms between categories of variable pairs can be added to the linear sum in the power of Eq. (29). However, there is an exponentially large number of cross-terms to choose from. Practical modeling questions are: which are the most relevant terms to add? And can they be picked in an effective way that limits their number?

To help answer these, realize that the shape of Eq. (29) and the assumption of variable independence are identical to the factorization assumption of Eq. (9). A practical approach is then:

1. Use the ϕ_K values and their significances to select the variable pairs with the strongest correlations.
2. The most relevant model cross-terms for each variable pair pq, having the largest impact in the model's likelihood, can be identified by studying the outliers in the correlation plots of Section 6.
3. Cross-terms can be included in a manner that limits the number of extra regression parameters. For example, for a given variable pair pq, introduce one cross-term parameter β_{pq} that affects only the contingency table cells with a Z value greater than a predefined value (and one for those smaller). To model those outlier cells, use Eq. (11): the cross term for each selected cell ij should scale with the uncertainty on the statistically independent estimate for that cell, $\sigma_{E_{ij}}\beta_{pq}x_{p,i}x_{q,j}$.

7.2. Finding unexpected answers in questionnaires

Our refined approach for calculating the statistical significances of outliers and deficits in a contingency table can help select the most interesting answers to a survey in an automated way, and, compared with conventional methods, can also be applied more safely to smaller surveys, where statistical fluctuations may be large. (For an overview of statistical analysis techniques of contingency tables see Ref. Fagerland and Morten. Lydersen (2020).) When interpreting questionnaires one is often interested in finding all "unexpected" correlations between ordinal or categorical answers given to a set of survey questions (the definition of what constitutes an unexpected correlation is typically survey specific). For example, for a particular survey question with a fixed set of multiple-choice answers, have certain groups of people provided specific answers that are significantly different in frequency from those expected from all other groups?

The methods presented in this paper can help to filter out these specific groups and answers:

1. By selecting question-pairs that have an interesting ("unexpected") ϕ_K correlation and significance on the one hand;
2. And selecting those with relatively high $|Z|$ values in the contingency tables of their respective answers on the other hand.

This allows one to quickly compile a list with all answer-pairs significantly deviating from the norm of no correlation, of which the most unexpected pairs form a subset.

8. Public implementation

The ϕ_K correlation analyzer code is publicly available as a Python library through the PyPi server, and from GitHub at https://github.com/KaveIO/PhiK. Install it with the command:

```
pip install phik
```

The web-page https://phik.readthedocs.io contains a description of the source code, tutorials on how to set up an analysis, and working examples of how to use and run the code.

9. Conclusion

We have presented a new correlation coefficient, ϕ_K, based on the χ^2 contingency test, which is most useful for variables of mixed types (interval, ordinal, and categorical). Internally ϕ_K treats every variable as having a categorical type. This is useful when studying data sets with mixed variable types, where some of the variables are categorical. Compared to Cramér's ϕ, the only viable alternative, by construction the strength of ϕ_K is very similar to Pearson's ρ, making it easy to interpret, its calculation is relatively robust against the binning per interval variable, and ϕ_K contains a noise correction against statistical fluctuations, making it more stable against changing sample size.

To evaluate the statistical significance of the hypothesis test of variable independence, a new hybrid approach is proposed using the *G*-test statistic, where only a limited number of Monte Carlo simulations is used to determine the effective number of degrees of freedom and to fit an analytical, empirical description of the χ^2 distribution. Also, we have presented a newly refined way to evaluate the statistical significance of outlier frequencies with respect to the factorization assumption, which is a helpful technique for interpreting any dependency found, *e.g.* between categorical variables. Two practical use-cases are discussed, the study of frequency of insurance claims and of survey responses, but plenty of other applications exist. The methods described are easy to apply through a Python analysis library that is publicly available.

Acknowledgments

We thank our colleagues of KPMG's Advanced Analytics & Big Data team for fruitful discussions and in particular Ralph de Wit for carefully reading the manuscript. This work has been supported by KPMG, The Netherlands Advisory N.V.

References

Aaboud, M., et al., 2018. Search for the direct production of charginos and neutralinos in final states with tau leptons in $\sqrt{s} = 13$ TeV *pp* collisions with the ATLAS detector. Eur. Phys. J. C78 (2), 154. http://dx.doi.org/10.1140/epjc/s10052-018-5583-9, arXiv:1708.07875.
Agresti, A., 1992. A survey of exact inference for contingency tables. Statist. Sci. 7 (1), 131–153. http://dx.doi.org/10.1214/ss/1177011454.
Agresti, A., 2001. Exact inference for categorical data: recent advances and continuing controversies. Stat. Med. 20 (17–18), 2709–2722. http://dx.doi.org/10.1002/sim.738.
Anon, scipy.stats.chi2_contingency. URL https://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.stats.chi2_contingency.html.
Anon, r.stats.chisq.test. URL https://www.rdocumentation.org/packages/stats/versions/3.5.1/topics/chisq.test.
Barnard, G., 1945. A new test for 2 2 tables. Nature 156 (3954), 177. http://dx.doi.org/10.1038/156177a0.
Barnard, G.A., 1947. Significance tests for 2 2 tables. Biometrika 34 (1/2), 123. http://dx.doi.org/10.2307/2332517.
Barnard, G.A., 1992. Introduction to pearson (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. In: Springer Series in Statistics Breakthroughs in Statistics, pp. 1–10. http://dx.doi.org/10.1007/978-1-4612-4380-9_1.
Bock, D.E., Velleman, P.F., De Veaux, R.D., 2007. Stats: Modeling the World. Pearson/Addison-Wesley.
Boigelot, D., Correlation benchmark samples. URL https://en.wikipedia.org/wiki/Correlation_and_dependence#/media/File:Correlation_examples2.svg.
Brent, R.P., 1973. Algorithms for Minimization Without Derivatives. Prentice-Hall.
Cochran, W., 1952. The χ^2 test of Goodness of fit. Ann. Math. Stat. 23 (3), 315–345. http://dx.doi.org/10.1214/aoms/1177729380.
Cochran, W.G., 1954. Some methods for strengthening the common χ^2 tests. Biometrics 10 (4), 417. http://dx.doi.org/10.2307/3001616.
Cousins, R.D., Linnemann, J.T., Tucker, J., 2008. Evaluation of three methods for calculating statistical significance when incorporating a systematic uncertainty into a test of the background-only hypothesis for a Poisson process. Nucl. Instrum. Meth. A595 (2), 480–501. http://dx.doi.org/10.1016/j.nima.2008.07.086, arXiv:physics/0702156.
Cover, T.M., Thomas, J.A., 2006. Elements of Information Theory. John Wiley & Sons.
Drasgow, F., 2006. Polychoric and polyserial correlations. Encyclopedia Stat. Sci. http://dx.doi.org/10.1002/0471667196.ess2014.pub2.
Fagerland, S.L.P., Morten. Lydersen, 2020. Statistical Analysis of Contingency Tables. CRC PRESS.
Fisher, R.A., 1922. On the interpretation of 2 from contingency tables, and the calculation of p. J. R. Stat. Soc. 85 (1), 87. http://dx.doi.org/10.2307/2340521.
Fisher, R.A., 1970. Statistical Methods for Research Workers. Oliver and Boyd.
Garrido, J., Genest, C., Schulz, J., 2016. Generalized linear models for dependent frequency and severity of insurance claims. Insurance Math. Econom. 70, 205–215. http://dx.doi.org/10.1016/j.insmatheco.2016.06.006.
Goodman, L., Krustal, W., 1954. Measures of association for cross classifications. J. Amer. Statist. Assoc. 49 (268), 732–764. http://dx.doi.org/10.2307/2281536.
Goodman, L., Krustal, W., 1959. Measures of association for cross classifications. II: Further discussion and references. J. Amer. Statist. Assoc. 54 (285), 123–163. http://dx.doi.org/10.1080/01621459.1959.10501503.
Goodman, L., Krustal, W., 1963. Measures of association for cross classifications III: Approximate sampling theory. J. Amer. Statist. Assoc. 58 (302), 310–364. http://dx.doi.org/10.1080/01621459.1963.10500850.
Goodman, L., Krustal, W., 1972. Measures of association for cross classifications, IV: Simplification of asymptotic variances. J. Amer. Statist. Assoc. 67 (338), 415–421. http://dx.doi.org/10.1080/01621459.1972.10482401.
Harald, C., 1999. Mathematical Methods of Statistics. Princeton University Press.
James, F., Roos, M., 1975. Minuit - a system for function minimization and analysis of the parameter errors and correlations. Comput. Phys. Comm. 10 (6), 343–367. http://dx.doi.org/10.1016/0010-4655(75)90039-9.
Kendall, M.G., 1938. A new measure of rank correlation. Biometrika 30 (1/2), 81. http://dx.doi.org/10.2307/2332226.
Kim, D., Agresti, A., 1997. Nearly exact tests of conditional independence and marginal homogeneity for sparse contingency tables. Comput. Statist. Data Anal. 24 (1), 89–104. http://dx.doi.org/10.1016/s0167-9473(96)00038-2.
Kroonenberg, P.M., Verbeek, A., 2018. The tale of cochrans rule: My contingency table has so many expected values smaller than 5, what am i to do?. Amer. Statist. 72 (2), 175–183. http://dx.doi.org/10.1080/00031305.2017.1286260.
Ku, H.H., 1965. Notes on the use of propagation of error formulas. http://dx.doi.org/10.6028/nbs.rpt.9011.
Lancaster, H.O., 1961. Significance tests in discrete distributions. J. Amer. Statist. Assoc. 56 (294), 223. http://dx.doi.org/10.2307/2282247.

Lin, X., Genest, C., Banks, D.L., Molenberghs, G., Scott, D.W., Wang, J.-L., 2014. Past, Present, and Future of Statistical Science. CRC Press.
Linnemann, J.T., 2003. Measures of significance in HEP and astrophysics. eConf C030908, MOBT001, arXiv:physics/0312059.
Patefield, W.M., 1981. Algorithm as 159: An efficient method of generating random r c tables with given row and column totals. Appl. Stat. 30 (1), 91. http://dx.doi.org/10.2307/2346669.
Pearson, K., 1895. Note on regression and inheritance in the case of two parents. Proc. R. Soc. Lond. (1854-1905) 58 (-1), 240–242. http://dx.doi.org/10.1098/rspl.1895.0041.
Smith, S.M., 2009. Fundamentals of Marketing Research. Sage Publications.
Sokal, R.R., Rohlf, F.J., 2012. Biometry: The Principles and Practice of Statistics in Biological Research. Palgrave Macmillan.
Spearman, C., 1904. The proof and measurement of association between two things. Amer. J. Psychol. 15 (1), 72–101. http://dx.doi.org/10.2307/1412159.
Yates, F., 1934. Contingency tables involving small numbers and the 2 test. Suppl. J. R. Stat. Soc. 1 (2), 217. http://dx.doi.org/10.2307/2983604.
Yoo, Y., et al., 2020. A new correlation coefficient for comparing and aggregating non-strict and incomplete rankings. European J. Oper. Res. 285 (3), 1025–1041. http://dx.doi.org/10.1016/j.ejor.2020.02.027.