An efficient method to generate data and compute exact P-values in goodness-of-fit testing

<table>
<thead>
<tr>
<th>Journal:</th>
<th><em>Communications in Statistics - Simulation and Computation</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID:</td>
<td>LSSP-2007-0065.R2</td>
</tr>
<tr>
<td>Manuscript Type:</td>
<td>Review Article</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>13-Sep-2007</td>
</tr>
<tr>
<td>Complete List of Authors:</td>
<td>Magis, David; University of Liege, Mathematics</td>
</tr>
<tr>
<td>Keywords:</td>
<td>Goodness of Fit, Pearson's Statistic, Exact Test, Exact P-Value, Data Generation</td>
</tr>
</tbody>
</table>

**Abstract:**
In this paper we use a characterization of the set of sample counts that do not match with the null hypothesis of the test of goodness of fit. Two direct applications arise: first, to instantaneously generate data sets whose corresponding asymptotic P-values belong to a certain pre-defined range; and second, to compute exact P-values for this test in an efficient way. We present both issues before illustrating them by analyzing a couple of data sets. Method's efficiency is also assessed by means of simulations. We focus on Pearson's X2 statistic but the case of likelihood-ratio statistic is also discussed.

Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.

LSSP-2007-0065.R2 Manuscript - David Magis.tex
AN EFFICIENT METHOD TO GENERATE DATA AND COMPUTE EXACT $P$-VALUES IN GOODNESS-OF-FIT TESTING

David Magis

Department of Mathematics
University of Liège
12 Grande Traverse, B-4000 Liège, Belgium
David.Magis@ulg.ac.be

RUNNING TITLE: Computing exact $P$-values in goodness-of-fit testing

KEY WORDS: Goodness of Fit; Pearson’s Statistic; Exact Test; Exact $P$-Value; Data Generation.
AN EFFICIENT METHOD TO GENERATE DATA AND COMPUTE EXACT P-VALUES IN GOODNESS-OF-FIT TESTING

ABSTRACT

In this paper we use a characterization of the set of sample counts that do not match with the null hypothesis of the test of goodness of fit. Two direct applications arise: first, to instantaneously generate data sets whose corresponding asymptotic P-values belong to a certain pre-defined range; and second, to compute exact P-values for this test in an efficient way. We present both issues before illustrating them by analyzing a couple of data sets. Method’s efficiency is also assessed by means of simulations. We focus on Pearson’s $X^2$ statistic but the case of likelihood-ratio statistic is also discussed.

KEY WORDS

Goodness of Fit; Pearson’s Statistic; Exact Test; Exact P-Value; Data Generation.

1. INTRODUCTION

The main goal of this paper is to focus on a very famous statistical test for discrete data, which aims at comparing some observed discrete distribution to a theoretical one, set by the practitioner and completely known (i.e. not depending on unknown parameters). Let us set some preliminary notations; the interested reader can find additional information about that test in e.g. Snedecor and Cochran (1967). Assume that a population is subdivided into $K$ groups or categories, and let $\pi_i$ ($i = 1, ..., K$) be the (population) probability of belonging to the $i$-th category (with $\sum_{i=1}^{K} \pi_i = 1$). In goodness-of-fit testing, the null and alternative
The hypotheses are

\[ H_0 : \pi_i = \pi_i^* \ \forall \ i = 1, \ldots, K \quad \text{vs} \quad H_1 : \exists \ j \in \{1, \ldots, K\} : \pi_j \neq \pi_j^* \]

where \( \pi_i^* (i = 1, \ldots, K) \) are the null probabilities, set by the practitioner and such that
\[ \sum_{i=1}^{K} \pi_i^* = 1. \]
To perform the test, one draws a random sample of size \( N \) from the population and one counts the number of subjects that belong to each category; let \( X_i (i = 1, \ldots, K) \) be the (random) count of the \( i \)-th category (with \( \sum_{i=1}^{K} X_i = N \)), and set \( \mathbf{X} = (X_1, \ldots, X_K) \) as the observed random sample. Then, provided that sample size is sufficiently large (Agresti, 1990), the statistics

\[ X^2(\mathbf{X}) = \sum_{i=1}^{K} \frac{(X_i - N \pi_i)^2}{N \pi_i} \quad \text{and} \quad G^2(\mathbf{X}) = 2 \sum_{i=1}^{K} X_i \log \left( \frac{X_i}{N \pi_i} \right) \]

have an asymptotic null chi-square distribution with \((K - 1)\) degrees of freedom. Statistic \( X^2 \) is also referred to as Pearson’s statistic (Pearson, 1900), whereas \( G^2 \) is often called the deviance or likelihood-ratio statistic (Wilks, 1935). Note that Cressie and Read (1984) proposed a general family of power divergence statistics holding both \( X^2 \) and \( G^2 \) as well as other alternative goodness-of-fit statistics. However in this paper we only focus on Pearson’s \( X^2 \) statistic; some comments about the deviance statistic are provided when it is appropriate.

Note that the chi-square approximation of goodness-of-fit statistics is commonly accepted as soon as the (null) expected category counts \( N \pi_i^* \) are sufficiently large. The minimum threshold is often chosen equal to 5 (Agresti, 1996), but some authors investigated that issue in more general frameworks (Cochran (1954) provides some interesting rules of thumb for various issues in discrete data analysis). When these asymptotic conditions are not fulfilled,
one has to consider the exact distribution of \( X \), usually the multinomial one, and to compute exact \( P \)-values for the test; this issue will be discussed later on.

First we display (in Section 2) a useful characterization of the set of random samples (of fixed size) that has some interesting properties. We directly make use of it by displaying an easy and rapid method to generate data sets \( X \) such that related Pearson’s statistic \( X^2(X) \) belongs to some pre-defined range of positive values. This application is particularly useful when one wishes for instance to quickly generate specific data sets for simulation studies. Section 3 is devoted to the computation of exact \( P \)-values for this test (based on the multinomial distribution). In particular, we propose a procedure to significantly reduce the number of computations needed to get the exact \( P \)-value, with respect to a classical “complete screening” approach. In Section 4 we present and discuss some simulation results that assess the efficiency of the latter method, especially when asymptotic framework can be considered as valid. We end up by illustrating our method in Section 5, with the analysis of some data sets coming from real studies.

2. DATA SET GENERATION

Before displaying the applications of interest, let us develop first some useful notions. Assume that \( K \), \( N \) and the vector of null probabilities \( \pi^* = (\pi_1^*, \ldots, \pi_K^*) \) are fixed and known, and let \( Q \) denote some positive (fixed) value. The key tool of this section is to determine and characterize the set \( S_Q \) of all samples \( X \) of size \( N \) such that statistic \( X^2(X) \) is smaller than or equal to \( Q \). Of course, if \( Q \) is the quantile of a chi-square distribution with \( (K-1) \) degrees of freedom (related to some fixed \( \alpha \) level), then \( S_Q \) is nothing else than
the set of all samples of size $N$ that do not lead to the rejection of the null hypothesis at significance level $\alpha$ (provided the asymptotic conditions are met).

The characterization one will consider here was first proposed by Magis (2007), in the framework of detecting decisive subsets of categories (i.e. minimal groups of population categories that surely yield the rejection of $H_0$). The set $S_Q$ can be written as a set of “nested” intervals:

$$S_Q = I_1 \times I_2 (X_1) \times \ldots \times I_K (X_1, \ldots, X_{K-1}),$$

such that

$$X \in S_Q \quad \text{if and only if} \quad X_1 \in I_1, \ X_2 \in I_2 (X_1), \ \ldots, \ X_K \in I_K (X_1, \ldots, X_{K-1}).$$

In other words, a data set will provoke the rejection of $H_0$ if and only if at least one of its components $X_i$ does not belong to interval $I_i (X_1, \ldots, X_{i-1})$. Moreover, interval $I_1$ depends on $N$, $Q$ and $\pi_1^*$; for any fixed $X_1 \in I_1$, interval $I_2 (X_1)$ depends on $N$, $Q$, $\pi_1^*$, $\pi_2^*$ and $X_1$; and so forth. Eventually interval $I_K (X_1, \ldots, X_{K-1})$ reduces to $\left\{ N - \sum_{i=1}^{K-1} X_i \right\}$ since the knowledge of the first $(K-1)$ counts completely determines the value of the last one.

For Pearson’s $X^2$, algebraic formulas for the intervals $I_1$, $I_2 (X_1)$ etc. can be provided, whereas for the deviance $G^2$ these intervals have to be determined numerically by computing the real roots of a second-order polynomial (see Magis (2007) for further details). So, although we discard $G^2$ statistic from our talk, the forthcoming methods could be adapted to that criterion without major problems.

The previous characterization of set $S_Q$ will be used first to randomly generate a data set $X$ such that the related statistic value $X^2 (X)$ lies between two pre-defined values, say $Q_1$
and \( Q_2 \) (with \( Q_1 < Q_2 \)). The practical situation of interest is the case where \( Q_1 \) (resp. \( Q_2 \)) corresponds to the quantile of the chi-square distribution with \((K - 1)\) degrees of freedom and related to some \( \alpha_1 \) (resp. \( \alpha_2 \)) level (with obviously \( \alpha_2 < \alpha_1 \)): in that case the asymptotic \( P \)-value of the test (obtained with the generated data set) lies between \( \alpha_1 \) and \( \alpha_2 \). The generation of data set \( X \) is easy: it is sufficient to ensure that \( X \in S_{Q_2} \) and \( X \not\in S_{Q_1} \).

To simplify our purpose, let \( S_{Q_1} \) and \( S_{Q_2} \) be written as in (1) but wherein each particular interval gets superscript 1 and 2 respectively. We suggest then to proceed as follows.

1. Generate count \( X_1 \) as a random integer value belonging to \( I_1^2 \) but not to \( I_1^1 \) (which is possible, provided the gap between \( Q_1 \) and \( Q_2 \) is sufficiently large). With this first step, one ensures that the generated data set \( X \) will not belong to \( S_{Q_1} \).

2. Then, for any \( j \in \{2, ..., K - 1\} \), successively generate count \( X_j \) as a random integer value belonging to \( I_j^2 (X_1, ..., X_{j-1}) \) (this will ensure that \( X \in S_{Q_2} \)).

3. End up by setting \( X_K = N - \sum_{i=1}^{K-1} X_i \) and \( X = (X_1, ..., X_K) \).

So, the first generated count \( X_1 \) ensures a \( X^2 \) value that is larger than \( Q_1 \), and the next steps successively set category counts \( X_2 \) to \( X_K \) to ensure finally that Pearson’s statistic is smaller than \( Q_2 \), as expected. Of course, if \( Q_1 \) and \( Q_2 \) are chosen too close together, then it is possible that no integer value \( t \) can be found such that \( t \in I_1^1 \) and \( t \not\in I_1^1 \); in that case it is impossible to generate such a data set.

Let us provide a simple illustration of that method. Assume that \( K = 4 \), \( N = 100 \) and \( \pi^* = (0.4, 0.3, 0.2, 0.1) \). Set \( \alpha_2 = 0.1 \) and \( \alpha_1 = 0.2 \), leading to \( Q_1 = 4.64 \) and \( Q_2 = 6.25 \). 

URL: http://mc.manuscriptcentral.com/lssp E-mail: comstat@univmail.cis.mcmaster.ca
(as the quantiles of the chi-square distribution with 3 degrees of freedom). One gets then
$I_1 = [28.45; 50.55]$ and $I_2 = [27.75; 52.25]$; thus any integer value in $\{28, 51, 52\}$ is allowed
as first category count. Set for instance $X_1 = 51$. Next, $I_2^2 (X_1) = [20.24; 28.76]$ and the
random draw provides $X_2 = 23$ for instance. In a third step, $I_3^2 (X_1, X_2) = [14.68; 19.99]$
and one selects for instance $X_3 = 15$. Eventually, $X_4 = 11$ and the random sample is
$X = (51, 23, 15, 11)$; it is direct to see that $X^2 (X) = 6.008$ and lies between $Q_1$ and $Q_2$, as
expected. Equivalently, the $P$-value of the test is equal to 0.11 and belongs to $[\alpha_2; \alpha_1]$.

This method is particularly efficient, easy to implement and useful when one wishes to
quickly generate lots of data sets within some ranges of quantile (or $P$-) values, for instance
to conduct simulation studies (as we present in Section 4).

3. COMPUTING EXACT $P$-VALUES

When the null expected count values are not sufficiently large, then chi-square asymptotic
approximation of statistics’ distribution is not assessed. One may nevertheless perform the
test of goodness of fit by computing now the exact $P$-value of that test. First, one makes
use of the multinomial distribution to characterize a particular random draw $X$ of size $N$:

$$P_{mult} (X) = \frac{N! \prod_{i=1}^{K} \pi_i^{x_i} x_i!}{\prod_{i=1}^{K} X_i!}.$$  \hspace{1cm} \hspace{1cm} (2)

This discrete statistical law provides in fact the probability that the realization $X$ occurs by
random drawing. Now, let $X^*$ be the current data set; then, using (2), the $P$-value of the
test can be computed by

$$p = \sum_{X \in S} P_{mult} (X).$$  \hspace{1cm} \hspace{1cm} (3)
where $S$ is the set of all samples $X$ such that $X^2(X) > X^2(X^*)$. In other words, it is sufficient to determine, among all samples of size $N$, those that yield a $X^2$ statistic value larger than the one obtained with original data $X^*$, and to sum the corresponding probabilities of occurrence by using (2).

In a classical “complete screening” approach, one should consider all possible random samplings of size $N$ and to keep those belonging to $S$, but this implies to examine $N = C(N + K - 1, N)$ different cases (with $C(a, b) = a!/[b! (a - b)!]$). This is of course not at all thinkable for large $K$ or $N$ values. In order to significantly decrease the number of cases to consider, we suggest to make use of the following procedure.

1. Determine first the set $S_{X^2(X^*)}$ of all samples $X$ leading to a Pearson’s statistic $X^2(X)$ smaller than or equal to current statistic value $X^2(X^*)$. With the developments of Section 2, this is an easy task.

2. Consider each sample $X \in S_{X^2(X^*)}$ by successively fixing $X_1 \in I_1, X_2 \in I_2(X_1)$, etc. until $X_{K-1} \in I_{K-1}(X_1, ..., X_{K-2})$. Compute then $P = \sum_{X \in S_{X^2(X^*)}} P_{\text{mult}}(X)$ using the multinomial distribution given in (2).

3. Set then $p = 1 - P$.

Indeed, it is obvious that $S = N \setminus S_{X^2(X^*)}$, which implies the above link between $p$ and $P$. Nevertheless, our method permits to focus on a restricted number of settings (those that belong to $S_{X^2(X^*)}$), which of course permits to diminish both computational time and effort. Furthermore, a slight modification of the method can also provide the number of cases to
consider, say \#\(p\): one has indeed \#\(p\) = \(\sum_{x \in S} x^2(x^*)\). Note that for efficiency reasons, both \(P\) and \#\(p\) should be computed altogether in step 2 of the procedure above.

The definition of set \(S\) implies that the exact \(P\)-value is obtained by considering samples \(X\) leading to Pearson’s statistic values \(X^2(X)\) strictly larger than \(X^2(X^*)\). In practice one could also be interested in computing the same \(P\)-value with samples \(X\) such that \(X^2(X) \geq X^2(X^*)\); it is sufficient to modify the method by considering open intervals \(I_1\), \(I_2(X_1)\), etc. and to proceed in the same way. Moreover, as the sample size increases, the gap between both definitions will reduce significantly.

Note that computing exact \(P\)-values is particularly useful in practice when the sample size is small; otherwise both exact and asymptotic \(P\)-values should be quite close (the larger the sample size, the closer the \(P\)-values). However the next section will highlight the gain that our approach keeps on the classical “complete screening” one, with either small or large samples.

4. SIMULATIONS

Although the method displayed in Section 3 is built up in order to reduce the number of settings to investigate (\#\(p\) instead of \(C(N + K - 1, N)\)), there may happen some cases where this reduction is not that important with respect to the total number of settings. To better characterize this issue we conducted some simulation study. We display first the setting of this study, before presenting the results assessing the efficiency of our method.
4.1. Simulation Settings

First we considered number of categories $K$ in $\{3, 4, 5, 6\}$ and in each case we considered several samples sizes $N$, respectively equal to 10, 20, 50 and 100. For each simulation setting $(K, N)$, 1000 pairs of data sets and vectors of null probabilities were randomly generated. Vectors of probabilities were generated first, and the generation process was such that no single value $\pi_i^* (i = 1, ..., K)$ was smaller than five percent. In that way, taking $N$ equal to 100 is sufficient to meet usual conditions about chi-square asymptotic distribution of the goodness-of-fit statistics (whatever the value of $K$), and taking lower $N$ values yields cases where these asymptotic rules are not fulfilled. Then, for each vector of null probabilities, a data set was randomly created using the method of Section 2. Moreover, the first (resp. second, ..., tenth) 100 data sets were generated such that their corresponding asymptotic $P$-values (computed with $X^2$ statistic) belong to the range $]0; 0.1[$ (resp. $[0.1; 0.2[$, ..., $[0.9; 1[$). This allows to consider the whole range of possible observable $P$-values, and thus to investigate in the largest panel of plausible situations.

For each setting $(K, N)$ the size of the set $S_{X^2(X^*)}$ (i.e. $\#P$) was computed and then divided by $N = C(N + K - 1, N)$ to keep the corresponding proportion of samples to consider in the computational task. The elapsed time for computing the exact $P$-values (related to all the generated settings) was also estimated to enhance the appreciation of our method’s efficiency. Note that this elapsed time is very dependent on the power of the computer in use; so these time values will be provided as indicators of the computational effort to be produced with our method.
All computations were performed using the R (2006) software.

4.2. Efficiency of the Method

We summarized the collected information by means of the minimum, first quantile, median, third quantile and maximum values of the proportion of settings (POS), i.e. \( \#p/N \), and the computational elapsed time (CET). These results are summarized in Table 1.

**APPROXIMATE POSITION OF TABLE 1**

Several main conclusions can be drawn from this study. Let us discuss first the POS results. For any \( K \), the medians (resp. third quantiles) are always smaller than 11\% (resp. 20\%), which means that one half (resp. two thirds) of the exact \( P \)-values can be computed by considering a very small POS of the same size. Moreover, the summary statistics decrease as \( N \) increases (for any fixed \( K \)). Hence our method permits to significantly reduce the number of settings to consider, at least for a large majority of cases we simulated. There remains however some situations where the POS is quite large, as the last column of Table 1 indicates. Fortunately these large maximal proportions often appear when \( N \) is quite small, and they decrease quite rapidly with \( N \).

About the elapsed time for computing exact \( P \)-values, it appears that for fixed \( K \) (resp. \( N \)) value, CET increases with \( N \) (resp. \( K \)). However the displayed values remain quite small in general, thus indicating that the method is not particularly computer-intensive (except for some settings i.e. whenever \( K = 6 \) and \( N = 100 \)).
When comparing the number of settings $#_p$ to CET, one gets (among the 16,000 generated data sets) an estimated correlation coefficient of around 0.999. Moreover usual linear regression of CET with respect to $#_p$ provides the following relationship:

$$CET = 0.0761 + 9.435 \times 10^{-5} \times (#_p).$$

(4)

Coefficients of (4) were obtained by ordinary least squares regression. The quality of a linear adjustment in that case is almost perfect since $R^2 \approx 99.9\%$. Hence the needed time for computing an exact $P$-value seems linearly increasing with the number of settings $#_p$ to consider. Linear equation (4) will be used in the next section to predict the needed time for getting exact results. Recall finally that these time values strongly depend on the type of computer that is used; the main interest of this analysis was therefore to highlight the strong evidence of a linear link between CET and $#_p$.

To eventually better describe our results, for any simulation setting $(K, N)$, generated data sets were subdivided into groups with corresponding exact $P$-values belonging to the ranges $]0, 0.5]$, $[0.5, 0.15]$, ..., $[0.85, 0.95]$, $[0.95, 1]$. Inside each interval we computed the median POS (just as in Table 1) and we summarized them on plots; they are displayed in Figure 1. One can clearly see that, for any $K$ and $N$, the median proportion is far more larger for small exact $P$-values, whereas it keeps decreasing as $P$-values increase. Moreover, for fixed $K$, these median curves hold smaller values for larger $N$, which illustrates the fact that median proportions become smaller as the sample size increases.

**APPROXIMATE POSITION OF FIGURE 1**
The decreasing trend of median proportions (as the exact $P$-value increases) can be intuitively explained as follows: when a data set yields a very small (exact) $P$-value, then both sets $N$ and $S_{X^2}(X^*)$ are almost equal, which implies that a large proportion of settings (among all the observable ones) must be considered in the computation. As the $P$-value increases, fewer cases must be taken into account, which makes the median proportion drop.

In conclusion, our method for computing exact $P$-values helps in reducing the computational work with respect to a “complete screening” approach and is not in general computationally intensive. It is particularly useful when large $P$-values are obtained, but less efficient for smaller results.

4.3. Efficiency of Exact vs Approximate Framework

An important asset of this method is that it validates the use of the test of goodness of fit even when asymptotic conditions are not satisfied. If however the sample size is sufficiently large, then both exact and asymptotic frameworks may be considered and their corresponding results should be quite close together. It would therefore be useful to determine whether the exact method is still efficient (in terms of POS and CET) in the cases of an acceptable asymptotic framework.

To perform this study, we considered once again the 16,000 generated settings of the previous simulations, and we retained all those that satisfied the commonly accepted rule of having all null expected values $N \pi_i^*$ larger than 5. Then, among those selected settings, we computed the same summary statistics as in Table 1; they are summarized in Table 2. An additional column (denoted by “Prop.”) was included to display the number of data
sets (among the 1,000 in each choice of $K$ and $N$) that ensured the hypothesis of a valid asymptotic chi-square approximation.

**APPROXIMATE POSITION OF TABLE 2**

Two important conclusions may be drawn. First, for any $K$, a small sample size $N$ will never lead to a valid asymptotic framework (summary statistics in Table 1 were therefore replaced by NA symbols); in that case computing exact $P$-values is the only way to get reliable statistical results, and it was shown in the previous section that our method was particularly efficient in these settings. Second, when comparing both tables, one can see that summary statistics (either about POS or CET) are a bit larger when one discards settings not leading to a valid asymptotic framework (i.e. those in Table 2), but the increase in proportions of settings and in elapsed times is not that large.

Hence one can conclude that our method is very efficient with respect to the classical asymptotic chi-square test. The latter being not always valid, one can make use of the exact computation of $P$-values to perform the test; and even if the null expected counts are sufficiently large, our exact method still performs efficiently and quickly (for most of the settings).

5. ANALYSIS OF SOME EXAMPLES

Let us end up this paper by analyzing two real data sets; only results corresponding to Pearson’s $X^2$ are presented here. The first one comes from Mendel’s experiments about segregation of plants and genetic inheritance; this study is briefly presented in Dagnelie (1998).
Mendel observed a total of 556 plants, having either round (R) or angular (A) seed, and either yellow (Y) or green (G) albumen. According to Mendel’s theory, categories (R,Y), (R,G), (A,Y) and (A,G) should be displayed with the corresponding (9,3,3,1) proportions. The vector of respective category counts is $X^* = (315, 108, 101, 32)$, leading to a smallest expected count equal to 34.75 and therefore validating the use of the asymptotic chi-square approximation. The asymptotic $P$-value equals 0.925, leading to the conclusion that Mendel’s hypotheses about plant segregation match very well with the experiment. Now to determine the exact $P$-value, the total number of observable samples $N$ is equal to $C(559, 556)$, which is larger than 28 millions and thus makes a complete screening approach inefficient. With our technique however, only $#_P = 615$ settings have to be considered (which represents 0.002 % of the total number of samples); the related (exact) $P$-value equals 0.926. The gap between exact and approximate results is quite small since all null expected counts are large; but the main interest of this example was to illustrate the particular efficiency of our computational method. The elapsed time to obtain this result was 0.14 seconds, whereas the predicted value (using (4) and the value $#_P$ above) is 0.134; in that example the predicted and elapsed times are quite close together.

The second example is quite similar to the previous one; it is called the Lindstrom’s plant data set and is displayed in Snedecor and Cochran (1967). It consists in the observation of crosses of two types of maize, leading to four distinct types of plants. With a sample of 1301 plants, the vector of respective counts is $X = (773, 231, 238, 59)$. The null hypothesis is that such a display follows Mendel’s rule of genetic inheritance, as given in the previous
example. The asymptotic approximation of $P$-value (with chi-square distribution) can be assessed since the smallest expected count is larger than 80; furthermore this asymptotic $P$-value is equal to 0.026. Hence the observed population of crossed plants does not seem to follow classical genetic theory. Now the exact $P$-value is almost equal to the asymptotic one (absolute difference being smaller than $10^4$); but to compute it a total of $#_P = 195,062$ samples were retained among the $C(1304, 1301)$ observable ones (the latter corresponding to more than 368 million samples), which corresponds to a ratio of around 0.05%. In this example, although this ratio is small, the number of settings to consider remains rather large.

Using linear prediction (4) of elapsed time, one gets an estimated value of 18.48 seconds, but in fact only 15.44 seconds were necessary to obtain the exact displayed result.

6. SUMMARY

The main aim of this paper was to provide some methods to either quickly generate data sets or to compute exact $P$-values by reducing the number of settings to consider. Some examples highlighted that these techniques can be very efficient. Furthermore, the methods were fully described with Pearson’s $X^2$ statistic; but as mentioned several times, an easy adaptation to the use of likelihood-ratio statistic can be performed without requiring much more computational effort. Current research also aims at generalizing this approach to other goodness-of-fit statistics such as those displayed in Cressie and Read (1984) for instance.

Note that even when it is possible to diminish the number of cases to consider in a convincing way (as the simulations showed), the total number of settings to compute can remain very large, especially when the number of categories increases; this is the main
drawback of our method. However, as already discussed, our approach is useful and mainly applicable when sample sizes are small, the most often case of rejecting the usual asymptotic chi-square approximation for goodness-of-fit statistics.

Recall also that the proposed method is based on the determination of the set $S_Q$ which implicitly depends on the (known) null expected values $N \pi^*_i$. In this paper it was assumed that these values were given in advance; but in practice they may be directly determined by estimating them from the data. If parameters are estimated from the data, then the set $S_Q$ will become data-dependent, and the validity of our approach is not yet clear. This issue is till now under consideration.

Some further research tries to extend this methodology to other discrete tests, such as the test of homogeneity of (vectors of) proportions among several populations. The task is far more difficult since expected count values implicitly depend on the data set, which was not the case here. This problem is related to the one we discussed above. It is nevertheless hoped that some adapted results will be available soon.

ACKNOWLEDGMENTS.

The author wishes to thank a referee whose accurate advice helped in improving the quality of this paper.

BIBLIOGRAPHY


Pearson, K. (1900). On a 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. London, Edinburgh and Dublin Philosophical Magazine and Journal of Science 50: 157–175.


List of Table Captions:

Table 1: Summary of simulation results for the proportion of settings (POS) and the computational elapsed time (CET) in exact $P$-value computation, for various choices of $K$ and $N$ parameters. POS values are expressed in percentage; CET values are expressed in seconds and displayed in parentheses.

Table 2: Summary of simulation results for the POS and the CET in exact $P$-value computation (for settings having null expected values larger than 5), for various choices of $K$ and $N$ parameters. POS values are expressed in percentage; CET values are expressed in seconds and displayed in parentheses.

List of Figure Captions:

Figure 1: Curves of median proportions of settings with simulated data and for various values of parameters $K$ and $N$. 
<table>
<thead>
<tr>
<th>K</th>
<th>N</th>
<th>Minimum</th>
<th>1\textsuperscript{st} Quartile</th>
<th>Median</th>
<th>3\textsuperscript{rd} Quartile</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>10</td>
<td>&lt; 0.01 (&lt; 0.01)</td>
<td>0.05 (&lt; 0.01)</td>
<td>0.11 (&lt; 0.01)</td>
<td>0.20 (&lt; 0.01)</td>
<td>1.00 (0.02)</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>&lt; 0.01 (&lt; 0.01)</td>
<td>0.02 (&lt; 0.01)</td>
<td>0.05 (&lt; 0.01)</td>
<td>0.11 (&lt; 0.01)</td>
<td>1.00 (0.04)</td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td>&lt; 0.01 (&lt; 0.01)</td>
<td>0.01 (&lt; 0.01)</td>
<td>0.03 (&lt; 0.01)</td>
<td>0.05 (0.01)</td>
<td>1.00 (0.11)</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>&lt; 0.01 (&lt; 0.01)</td>
<td>0.01 (&lt; 0.01)</td>
<td>0.01 (&lt; 0.01)</td>
<td>0.03 (0.02)</td>
<td>1.00 (0.39)</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>&lt; 0.01 (&lt; 0.01)</td>
<td>0.03 (&lt; 0.01)</td>
<td>0.08 (&lt; 0.01)</td>
<td>0.17 (0.01)</td>
<td>1.00 (0.50)</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>&lt; 0.01 (&lt; 0.01)</td>
<td>0.01 (&lt; 0.01)</td>
<td>0.04 (0.01)</td>
<td>0.08 (0.02)</td>
<td>1.00 (0.18)</td>
</tr>
<tr>
<td>4</td>
<td>50</td>
<td>&lt; 0.01 (&lt; 0.01)</td>
<td>&lt; 0.01 (&lt; 0.01)</td>
<td>0.01 (0.03)</td>
<td>0.02 (0.06)</td>
<td>1.00 (1.97)</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>&lt; 0.01 (&lt; 0.01)</td>
<td>&lt; 0.01 (&lt; 0.01)</td>
<td>&lt; 0.01 (0.07)</td>
<td>&lt; 0.01 (0.16)</td>
<td>1.00 (14.38)</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>&lt; 0.01 (&lt; 0.01)</td>
<td>0.03 (&lt; 0.01)</td>
<td>0.07 (0.01)</td>
<td>0.17 (0.03)</td>
<td>0.87 (0.14)</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>&lt; 0.01 (&lt; 0.01)</td>
<td>&lt; 0.01 (0.02)</td>
<td>0.03 (0.04)</td>
<td>0.06 (0.09)</td>
<td>0.56 (0.61)</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td>&lt; 0.01 (&lt; 0.01)</td>
<td>&lt; 0.01 (0.08)</td>
<td>&lt; 0.01 (0.20)</td>
<td>0.01 (0.47)</td>
<td>0.18 (5.25)</td>
</tr>
<tr>
<td>5</td>
<td>100</td>
<td>&lt; 0.01 (&lt; 0.01)</td>
<td>&lt; 0.01 (0.28)</td>
<td>&lt; 0.01 (0.74)</td>
<td>&lt; 0.01 (1.81)</td>
<td>0.06 (22.99)</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>&lt; 0.01 (&lt; 0.01)</td>
<td>0.03 (0.02)</td>
<td>0.08 (0.04)</td>
<td>0.18 (0.09)</td>
<td>0.86 (0.39)</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>&lt; 0.01 (&lt; 0.01)</td>
<td>&lt; 0.01 (0.07)</td>
<td>0.02 (0.18)</td>
<td>0.06 (0.43)</td>
<td>0.54 (3.45)</td>
</tr>
<tr>
<td>6</td>
<td>50</td>
<td>&lt; 0.01 (&lt; 0.01)</td>
<td>&lt; 0.01 (0.50)</td>
<td>&lt; 0.01 (1.42)</td>
<td>0.01 (4.02)</td>
<td>0.16 (55.81)</td>
</tr>
<tr>
<td>6</td>
<td>100</td>
<td>&lt; 0.01 (0.02)</td>
<td>&lt; 0.01 (2.38)</td>
<td>&lt; 0.01 (7.40)</td>
<td>&lt; 0.01 (20.50)</td>
<td>0.04 (358.14)</td>
</tr>
<tr>
<td>K</td>
<td>N</td>
<td>Prop.</td>
<td>Minimum</td>
<td>1\textsuperscript{st} Quartile</td>
<td>Median</td>
<td>3\textsuperscript{rd} Quartile</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>------</td>
<td>---------</td>
<td>-------------------------------</td>
<td>--------</td>
<td>-------------------------------</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>0%</td>
<td>NA (NA)</td>
<td>NA (NA)</td>
<td>NA (NA)</td>
<td>NA (NA)</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>11%</td>
<td>&lt; 0.01</td>
<td>0.03 (&lt; 0.01)</td>
<td>0.08 (&lt; 0.01)</td>
<td>0.16 (&lt; 0.01)</td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td>73%</td>
<td>&lt; 0.01</td>
<td>0.01 (&lt; 0.01)</td>
<td>0.03 (&lt; 0.01)</td>
<td>0.05 (0.01)</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>100%</td>
<td>&lt; 0.01</td>
<td>0.01 (&lt; 0.01)</td>
<td>0.01 (&lt; 0.01)</td>
<td>0.03 (0.02)</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>0%</td>
<td>NA (NA)</td>
<td>NA (NA)</td>
<td>NA (NA)</td>
<td>NA (NA)</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>0%</td>
<td>NA (NA)</td>
<td>NA (NA)</td>
<td>NA (NA)</td>
<td>NA (NA)</td>
</tr>
<tr>
<td>4</td>
<td>50</td>
<td>50%</td>
<td>&lt; 0.01</td>
<td>0.01 (0.02)</td>
<td>0.01 (0.03)</td>
<td>0.03 (0.06)</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>100%</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01 (0.03)</td>
<td>&lt; 0.01 (0.07)</td>
<td>0.01 (0.16)</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>0%</td>
<td>NA (NA)</td>
<td>NA (NA)</td>
<td>NA (NA)</td>
<td>NA (NA)</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>0%</td>
<td>NA (NA)</td>
<td>NA (NA)</td>
<td>NA (NA)</td>
<td>NA (NA)</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td>24.8%</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01 (0.09)</td>
<td>0.01 (0.25)</td>
<td>0.02 (0.59)</td>
</tr>
<tr>
<td>5</td>
<td>100</td>
<td>100%</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01 (0.28)</td>
<td>&lt; 0.01 (0.74)</td>
<td>&lt; 0.01 (1.81)</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>0%</td>
<td>NA (NA)</td>
<td>NA (NA)</td>
<td>NA (NA)</td>
<td>NA (NA)</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>0%</td>
<td>NA (NA)</td>
<td>NA (NA)</td>
<td>NA (NA)</td>
<td>NA (NA)</td>
</tr>
<tr>
<td>6</td>
<td>50</td>
<td>9.5%</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01 (0.74)</td>
<td>0.01 (2.46)</td>
<td>0.01 (5.65)</td>
</tr>
<tr>
<td>6</td>
<td>100</td>
<td>100%</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01 (2.38)</td>
<td>&lt; 0.01 (7.40)</td>
<td>&lt; 0.01 (20.50)</td>
</tr>
</tbody>
</table>
David MAGIS  
Department of Mathematics (B37)  
University of Liège  
12, Grande Traverse  
B-4000 LIEGE  
BELGIUM  
Phone : (+32) 4 366 94 24  
Email: David.Magis@ulg.ac.be  

Liège, September 13, 2007.

Prof. N. Balakrishnan, Editor,  
Communications in Statistics - Simulation and Computation

Dear Sir,

I am pleased to propose a revised version of my previously submitted manuscript, entitled “An efficient method to generate data and compute exact $P$-values in goodness-of-fit testing”, for publication into the journal Communications in Statistics - Simulation and Computation. This manuscript was first submitted on April 18, 2007 and it was assigned reference number LSSP-2007-0065. It was then revised and submitted on August 7, 2007 and it was assigned reference number LSSP-2007-0065.R1.

In the following documents you will find the corrected version of the manuscript (wherein modifications are highlighted in red) as well as a list of detailed answers to the referee’s questions and comments.

Thanking you in advance for your attention,

Yours sincerely,

David MAGIS
Communications in Statistics - Simulation and Computation
Manuscript # LSSP-2007-0065.R1
Answers to Referee’s Questions and Comments

In this document answers to referee’s questions and comments are provided.

General misprints.

The following misprints were mentioned in the decision letter and were corrected in the revised manuscript:

- Page 11, line 1: “usig”
- Page 12, third line after equation (4): “excat”
- Page 12, sixth line after equation (4): “anbalyis”
- Page 14, line 9: “st atistics”, “ar”, “whane”
- Page 16, line -4: Please write “such as those displayed...”
- Page 16, line -3: Please write “Note that even when it is...”

Page 12, first line after equation (4): Left parenthesis is missing.

This was also a misprint; right parenthesis has been withdrawn form the revised text.

Page 16, Line -1: Is the word “important” appropriate?

This is indeed an inappropriate word. The sentence “... is important” was replaced by “increases” to clarify the meaning of the text.
Page 17, Line 6: “practitioner”.

The sentence “... by the practitioner” was replaced by “in advance” in the revised manuscript.

Page 18, Line 1 (references): Dagnelie is not in alphabetic order.

This mistake was corrected in the revised manuscript.

David MAGIS