# Chi square statistic

Subjects: Statistics & Probability

Contributor: Lorentz Jäntschi , Richard Gill

The Chi-Square test is based on a series of assumptions frequently used in the statistical analysis of experimental data. The main weakness of the chi-square test is that is very accurate only in convergence (in large size samples), and for small sample sizes is exposed to errors of both types (type I and type II). On two scenarios of use - goodness of fit and contingencies assessment (2x2 tables of contingency) - here are discussed different aspects involving it. Further knowledge on the regard of the type of the error in contingencies assessment push further the analysis of the data, while in the same time opens the opportunity to devise a method for filling the gaps in contingencies (e.g. censored data), both scenarios being discussed here in detail. A program designed to fill the gaps in the assumption of the association is provided.

contingency tables multiplicative effects agreement maximization

## 1. Introduction

The χ2 test was introduced by Pearson in 1900[1]. The $\chi^2$ statistic were originally devised to measure the departure in a system of n (random normally distributed) variables $\{y_1, \ldots, y_n\}$ having each a series of undefined number of observations ($y_1 = \{y_{1,1}, \ldots\}, \ldots, y_n = \{y_{n,1}, \ldots\}$) for which $\{x_1, \ldots, x_n\}$ are the means and $\{\sigma_1, \ldots, \sigma_n\}$ are the deviations. In this context the formula for the $\chi^2$ statistic is derived and it have a known distribution function.

$$PDF_{\chi^2}(x;n) = \frac{x^{n/2-1}e^{-x/2}}{2^{n-2}\Gamma(n/2)}, CDF_{\chi^2}(x;n) = \frac{1}{\Gamma(n/2)}\int_0^{x/2} t^{n/2-1}e^{-t}dt$$

It should be noted that the context in which the $\chi^2$ statistic is slightly different than the one in which is currently in use. This is the reason for which, during time a series of patches have been made to the statistic in order to make it usable.

For instance, in the contingency of two factors f1 = $\{f1_1, f1_2\}$ and f2 = $\{f2_1, f2_2\}$, under the assumption that $x_{i,j} \sim f1_i f2_j$:

| Factors | $f1_1$ | $f1_2$ |
|---|---|---|
| $f2_1$ | $x_{1,1}$ | $x_{1,2}$ |
| $f2_2$ | $x_{2,1}$ | $x_{2,2}$ |

the use of the $\chi^2$ assumes that it exists also a series of observations behind each of the $\{x_{1,1}, x_{1,2}, x_{2,1}, x_{2,2}\}$ entries in the table above, and someone may calculate the $X^2$ statistic (e.g. $X^2$ is the sample based statistic, while $\chi^2$ is the population based statistic or distribution) for the contingency as:

$$X^2 = \frac{(x_{1,1}x_{2,2}-x_{1,2}x_{2,1})^2(x_{1,1}+x_{1,2}+x_{2,1}+x_{2,2})}{(x_{1,1}+x_{1,2})(x_{1,1}+x_{2,1})(x_{1,2}+x_{2,2})(x_{2,1}+x_{2,2})}$$

Since the estimation of expected frequencies in the contingency of the factors uses rows totals $\{x_{1,1} + x_{1,2}, x_{2,1} + x_{2,2}\}$, columns totals $\{x_{1,2} + x_{2,1}, x_{2,1} + x_{2,2}\}$ and overall total $\{x_{1,1} + x_{1,2} + x_{2,1} + x_{2,2}\}$ the variation in the $X^2$ is constrained. Are exactly three independent constrains (if $x_{1,1} + x_{1,2} + x_{2,1} + x_{2,2} = n_0$, $x_{1,1} + x_{1,2} = n_1$, $x_{1,2} + x_{2,1} = n_2$ then the rows totals are $\{n_1, n_0 - n_1\}$, the columns totals are $\{n_2, n_0 - n_2\}$, and the overall total is $\{n_0\}$) and it is exactly one degree of freedom for $X^2$ (if $x_{1,1} = x$, then $x_{1,2} = n_1 - x$, $x_{2,1} = n_2 - x$, $x_{2,2} = n_0 - n_1 - n_2 + x$) and therefore the probability associated with the $X^2$ value should be taken from $\chi^2$ distribution with 1 degree of freedom. More, it can be verified by induction that for a m·n contingency table are m + n - 1 independent constrains, and thus are m·n - m - n +1 degrees of freedom.

The main trouble in the use of the $\chi^2$ statistic is the fact that its distribution is asserted and the probability is derived under the assumption that behind $\{x_{1,1}, x_{1,2}, x_{2,1}, x_{2,2}\}$ values are a infinite number of observations ($\{y_{1,1,1}, \ldots\}, \{y_{,1,2,1}, \ldots\}, \{y_{2,1,1}, \ldots\}, \{y_{2,2,1}, \ldots\}$), which actually is true never. Even if are accessible some population means, $\{\mu_{1,1}, \mu_{1,2}, \mu_{2,1}, \mu_{2,2}\}$ those means comes actually from some estimations made from some (finite) samples of the populations, which passes sampling noises (or errors) to the statistic itself. A clear shot to this matter has been made by Fisher which patches the $\chi^2$ statistic in the case of the 2x2 contingency above given for the frequency tables (Fisher exact test, see [2]).

Despite of its use in goodness-of fit tests, actually the $\chi^2$ statistic is not suitable in this instance. Having a series of observations $\{o_1, \ldots, o_n\}$ one split the domain of the observations (or having a series of cumulative probabilities $\{q_1, \ldots, q_n\}$ associated with the series of sorted observations, see [3]) in an arbitrary number of subintervals (usually taken as 1 + ) in order to construct the series of observed and expected frequencies. By this fact alone, when it is used to test the goodness-of, the $\chi^2$ statistic is itself exposed to the risk of being in error. Turning back to the general assumption in which its formula were derived - namely the assumption that behind each observed frequency is an infinite number of observations - one may realize that this kind of state of facts is never meet in practice. Therefore, only for very large samples, in which the observed frequencies are 'large enough' too, the $\chi^2$ statistic may become of use.

Someone may ask now: Which are the subjects in which the $\chi^2$ statistic is of use? - And indeed are, but the answer is not detailed here.

## 2. Measures of agreement in contingencies of multiplicative effects

Something else is of interest in close connection with the use of the $\chi^2$ statistic for contingency tables, as first pointed out by Fisher in [4], namely its dependence on the experimental design to which it is applied. In [5] is revised and generalized this perspective, and here is given a summary of it.

Let's consider a series of observations made in a contingency of two factors (see the table below; the experimental design can be generalized to any number of factors) which affect the observation in a multiplicative manner (e.g. $u_{i,j}$ $fr_i \cdot fc_j$).

| $(u_{i,j})_{1\le i\le r,1\le j\le c}$ | $fc_1$ | … | $fc_c$ | Σ |
|---|---|---|---|---|
| $fr_1$ | $u_{1,1}$ | … | $u_{1,c}$ | $sr_1$ |
| … | … | … | … | … |
| $fr_r$ | $u_{r,1}$ | … | $u_{r,c}$ | $sr_r$ |
| Σ | $sc_1$ | … | $sc_c$ | ss |

Above $(u_{i,j})_{1\le i\le r,1\le j\le c}$ are the observations made under different constrains of the factors, where $(fr_i)_{1\le i\le r}$ are the levels of the fr factor and $(fc_j)_{1\le j\le c}$ are the levels of the fc factor. For convenience of the later use, also the sums $sr_1$ ($sr_1 = u_{1,1} + \ldots + u_{1,c}$), …, $sr_r$ ($sr_c = u_{r,1} + \ldots + u_{r,c}$), $sc_1$ ($sc_1 = u_{1,1} + \ldots + u_{r,1}$), …, $sc_c$ ($sc_c = u_{1,c} + \ldots + u_{r,c}$), and ss = $\Sigma_{1\le i\le r,1\le j\le c} u_{i,j}$ were assigned.

It can be proved [where?] that under the presence of the multiplicative effect of the two (fr and fc) factors a very good estimate of the expected values would be given [is given?] by the formula $v_{i,j} = sr_i \cdot sc_j / ss$ (see the table below).

| $(v_{i,j})_{1\le i\le r,1\le j\le c}$ | $fc_1$ | … | $fc_c$ | Σ |
|---|---|---|---|---|
| $fr_1$ | $v_{1,1}$ | … | $v_{1,c}$ | $sr_1$ |
| … | … | … | … | … |

| | | | | |
|---|---|---|---|---|
| $fr_r$ | $v_{r,1}$ | … | $v_{r,c}$ | $sr_r$ |
| Σ | $sc_1$ | … | $sc_c$ | ss |

The values of the factors ($(fr_i)_{1\le i\le r}$ and $(fc_j)_{1\le j\le c}$) are in general unknown, but even if are known, doesn't help too much in the analysis since also their values may be affected by errors, as the observations $(u_{i,j})_{1\le i\le r,1\le j\le c}$ are supposed to be.

Either way (the values of the factors are known or not) in the calculation of the expected values $(v_{i,j})_{1\le i\le r,1\le j\le c}$ are used the sums of the observations ($sr_i = \Sigma_{1\le j\le c}u_{i,j}$, $sc_j = \Sigma_{1\le i\le r}u_{i,j}$, $ss = \Sigma_{1\le i\le r,1\le j\le c}u_{i,j}$) as the estimates for the factor levels (or values):

$$fr_1 : \dots : fr_i : \dots fr_r \;\; sr_1 : \dots : sr_i : \dots sr_r \;\&\; fc_1 : \dots : fc_i : \dots fc_c \;\; sc_1 : \dots : sc_i : \dots sc_c$$

Three alternative assumptions may push further the analysis and are on the regard of the type of the error (possibly, probably) made in the process of observation. Since the whole process of observation under the influence of the factors is a part of a whole, it is safe to assume that the error keeps its type during the process of observation, it is random and it is accidentally (having a low occurrence) and the three alternatives along with their consequences as usable formulas are listed in the table below.

| Alternative constraint formula | Consequenced usable formulas |
|---|---|
| $S^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(u_{i,j} - fr_i \cdot fc_j)^2}{1} \to \min.$ | $fr_i = \left(\sum_{j=1}^{c} fc_j u_{i,j}\right) \Big/ \left(\sum_{j=1}^{c} fc_j^{\,2}\right), i = 1..r$<br>$fc_j = \left(\sum_{i=1}^{r} fr_i u_{i,j}\right) \Big/ \left(\sum_{i=1}^{r} fr_i^{\,2}\right), j = 1..c$ |
| $V^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(u_{i,j} - fr_i \cdot fc_j)^2}{(fr_i \cdot fc_j)^2} \to \min.$ | $fr_i = \left(\sum_{j=1}^{c} \frac{u_{i,j}^{\,2}}{fc_j^{\,2}}\right) \Big/ \left(\sum_{j=1}^{c} \frac{u_{i,j}}{fc_j}\right), i = 1..r$<br>$fc_j = \left(\sum_{i=1}^{r} \frac{u_{i,j}^{\,2}}{fr_i^{\,2}}\right) \Big/ \left(\sum_{i=1}^{r} \frac{u_{i,j}}{fc_i}\right), j = 1..c$ |
| $X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(u_{i,j} - fr_i \cdot fc_j)^2}{(fr_i \cdot fc_j)} \to \min.$ | $fr_i^{\,2} = \left(\sum_{j=1}^{c} \frac{u_{i,j}^{\,2}}{fc_j}\right) \Big/ \left(\sum_{j=1}^{c} fc_j\right), i = 1..r$<br>$fc_j^{\,2} = \left(\sum_{i=1}^{r} \frac{u_{i,j}^{\,2}}{fr_i}\right) \Big/ \left(\sum_{i=1}^{r} fr_i\right), j = 1..c$ |

It should be noted that $X^2$ formula is usable only in the assumption that $fr_i$, $fc_j > 0$.

The $S^2$ formula minimizes the absolute errors, $V^2$ formula minimizes the relative errors, while $X^2$ formula minimizes the $X^2$ value for the contingency.

The algorithm is:

- for($1 \le i \le r$ and $1 \le j \le c$) $v_{i,j} \leftarrow sr_i \cdot sc_j/ss$
- $fr_1 = fc_1 = (v_{1,1})^{1/2}$; for($2 \le i \le r$) $sr_i \leftarrow v_{i,1}/fc_1$; for($2 \le j \le c$) $sc_j \leftarrow v_{1,j}/fr_1$
- Repeat
  - for($1 \le i \le r$) $fr_i \leftarrow$ Corresponding "Consequenced usable formula" (from $S^2$, $V^2$, or $X^2$)
  - for($1 \le j \le c$) $fc_j \leftarrow$ Corresponding "Consequenced usable formula" (from $S^2$, $V^2$, or $X^2$)
- Until convergence criteria is meet.

It has been shown in [5] that consecutive using of equations given as usable formulas converges fast to the minimum (constraint formula) and thus provides better estimates of the factor levels $(fr_i)_{1\le i\le r}$ and $(fc_j)_{1\le j\le c}$ which can be further used to improve the expected estimates (population means) of the factorial experiment: $v_{i,j} \leftarrow fr_i \cdot fc_j$. Using the data given in [5] the number of steps for a change less than 0.1% in the objective function ($S^2$, $V^2$, and $X^2$ respectively) are: 2 steps for $S^2$ and $X^2$ and 3 steps for $V^2$.

## 3. Exploiting the agreement in contingencies of multiplicative effects

One of the uses of the $\chi^2$ statistic is in the presence of censored data (see for instance [6] and [7]).

A recent application of the method above described has been reported in [8]. The method has been used to fill the gaps of missing data in contingencies of multiplicative effects of factors influencing the observations. The algorithm above given must be adapted in order to meet the requirements to be used and also the gaps may be filled if exists at least one observation in each row (e.g. $sr_i \ne 0$ for $1 \le i \le r$) and at least one observation in each column (e.g. $sc_j \ne 0$ for $1 \le j \le c$).

In the most general case, a recursive procedure can be designed to fill the gaps. Later on, the procedure of minimizing the residuals (e.g. $S^2$, $V^2$ or $X^2$) goes smoothly. The full program (PHP source code) is in the appendix (as a coeditor noted, an encyclopedia is not a place to put computer code). Here is given (an example of) raw data and data filled with gaps followed by the optimization of the expectances, of the $S^2$, $V^2$ and $X^2$ respectively.

Raw data ("data.txt" input data for the program given in the appendix)

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 25.3 | 28.0 | 23.3 | 20.0 | 22.9 | 20.8 | 22.3 | 21.9 | 18.3 | 14.7 | 13.8 | 10.0 |
| 26.0 | 27.0 | 24.4 | 19.0 | 20.6 | 24.4 | 16.8 | 20.9 | 20.3 | 15.6 | 11.0 | 11.8 |
| 26.5 | 23.8 | 14.2 | 20.0 | 20.1 | 21.8 | 21.7 | 20.6 | 16.0 | 14.3 | 11.1 | 13.3 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 23.0 | 20.4 | 18.2 | 20.2 | 15.8 | 15.8 | 12.7 | 12.8 | 11.8 | 12.5 | 12.5 | 8.2 |
| 18.5 | 17.0 | 20.8 | 18.1 | 17.5 | 14.4 | 19.6 | 13.7 | 13.0 | 12.0 | 12.7 | 8.3 |
| 9.5 | 6.5 | 4.9 | 7.7 | 4.4 | 2.3 | 4.2 | 6.6 | 1.6 | 2.2 | 2.2 | 1.6 |

Data with gaps (18 randomly filled values from raw data)

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 28.0 | | | | 20.8 | | | | | | |
| 26.0 | | | | 20.6 | | 16.8 | | | | | 11.8 |
| | | | | 20.1 | | | | 16.0 | 14.3 | | |
| | | 18.2 | 20.2 | | | 12.7 | | | | | |
| 18.5 | 17.0 | | | 17.5 | | | 13.7 | | | 12.7 | |
| | | | 7.7 | | | | | | | | |

Expected values $((\Sigma_i o_{i,j})\cdot(\Sigma_j o_{i,j})/(\Sigma_{i,j} o_{i,j}))$ calculated from data with gaps

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 31.387 | 27.926 | 28.595 | 31.738 | 27.230 | 20.819 | 20.472 | 22.417 | 22.339 | 19.966 | 20.781 | 14.600 |
| 25.239 | 22.457 | 22.994 | 25.521 | 21.896 | 16.741 | 16.463 | 18.026 | 17.964 | 16.055 | 16.711 | 11.741 |
| 22.653 | 20.156 | 20.638 | 22.906 | 19.653 | 15.026 | 14.776 | 16.180 | 16.123 | 14.410 | 14.999 | 10.538 |
| 19.895 | 17.702 | 18.126 | 20.118 | 17.261 | 13.197 | 12.977 | 14.210 | 14.161 | 12.656 | 13.173 | 9.255 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 19.208 | 17.090 | 17.499 | 19.422 | 16.664 | 12.741 | 12.529 | 13.719 | 13.671 | 12.218 | 12.717 | 8.935 |
| 7.633 | 6.791 | 6.954 | 7.718 | 6.622 | 5.063 | 4.979 | 5.452 | 5.433 | 4.855 | 5.054 | 3.551 |

Optimized expectances for $S^2 \rightarrow min.$

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 31.759 | 27.975 | 28.896 | 32.059 | 27.067 | 20.801 | 20.726 | 22.472 | 21.883 | 19.558 | 20.832 | 14.774 | 5.617 |
| 25.338 | 22.320 | 23.054 | 25.578 | 21.595 | 16.596 | 16.536 | 17.930 | 17.459 | 15.604 | 16.621 | 11.787 | 4.482 |
| 23.318 | 20.540 | 21.216 | 23.539 | 19.874 | 15.273 | 15.218 | 16.500 | 16.067 | 14.360 | 15.296 | 10.848 | 4.124 |
| 19.929 | 17.555 | 18.132 | 20.117 | 16.985 | 13.053 | 13.006 | 14.102 | 13.732 | 12.273 | 13.072 | 9.271 | 3.525 |
| 19.344 | 17.039 | 17.600 | 19.526 | 16.486 | 12.670 | 12.624 | 13.688 | 13.329 | 11.912 | 12.688 | 8.999 | 3.421 |
| 7.649 | 6.738 | 6.960 | 7.722 | 6.520 | 5.010 | 4.992 | 5.413 | 5.271 | 4.711 | 5.018 | 3.559 | 1.353 |
| 5.654 | 4.980 | 5.144 | 5.707 | 4.819 | 3.703 | 3.690 | 4.001 | 3.896 | 3.482 | 3.709 | 2.630 | fc\fr |

**$S^2$ = 3.4052**, $V^2$ = 0.0095, $X^2$ = 0.1776

Optimized expectances for $V^2 \rightarrow min.$

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 31.437 | 27.910 | 29.069 | 32.123 | 27.428 | 20.828 | 20.676 | 22.416 | 22.125 | 19.775 | 20.780 | 14.760 | 5.620 |
| 25.114 | 22.296 | 23.222 | 25.662 | 21.911 | 16.639 | 16.517 | 17.907 | 17.675 | 15.797 | 16.600 | 11.791 | 4.490 |
| 22.803 | 20.244 | 21.085 | 23.300 | 19.894 | 15.108 | 14.997 | 16.259 | 16.048 | 14.343 | 15.073 | 10.706 | 4.077 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 19.627 | 17.424 | 18.148 | 20.055 | 17.123 | 13.003 | 12.908 | 13.995 | 13.813 | 12.345 | 12.973 | 9.215 | 3.509 |
| 19.235 | 17.076 | 17.786 | 19.655 | 16.781 | 12.744 | 12.651 | 13.715 | 13.537 | 12.099 | 12.714 | 9.031 | 3.439 |
| 7.573 | 6.723 | 7.003 | 7.739 | 6.607 | 5.018 | 4.981 | 5.400 | 5.330 | 4.764 | 5.006 | 3.556 | 1.354 |
| 5.593 | 4.966 | 5.172 | 5.715 | 4.880 | 3.706 | 3.679 | 3.988 | 3.937 | 3.518 | 3.697 | 2.626 | fc\fr |

$S^2$ = 3.7697, **$V^2$ = 0.0089,** $X^2$ = 0.1812

Optimized expectances for $X^2$ → min.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 31.617 | 27.946 | 29.014 | 32.154 | 27.258 | 20.811 | 20.722 | 22.442 | 22.005 | 19.667 | 20.804 | 14.783 | 5.621 |
| 25.214 | 22.286 | 23.137 | 25.641 | 21.737 | 16.596 | 16.525 | 17.897 | 17.548 | 15.684 | 16.590 | 11.789 | 4.482 |
| 23.068 | 20.389 | 21.168 | 23.459 | 19.887 | 15.184 | 15.119 | 16.373 | 16.055 | 14.349 | 15.178 | 10.785 | 4.101 |
| 19.764 | 17.469 | 18.137 | 20.099 | 17.039 | 13.009 | 12.954 | 14.029 | 13.756 | 12.294 | 13.005 | 9.241 | 3.514 |
| 19.307 | 17.065 | 17.717 | 19.635 | 16.645 | 12.709 | 12.654 | 13.704 | 13.438 | 12.010 | 12.704 | 9.027 | 3.432 |
| 7.605 | 6.722 | 6.979 | 7.734 | 6.556 | 5.006 | 4.984 | 5.398 | 5.293 | 4.731 | 5.004 | 3.556 | 1.352 |
| 5.625 | 4.972 | 5.162 | 5.720 | 4.849 | 3.703 | 3.687 | 3.993 | 3.915 | 3.499 | 3.701 | 2.630 | fc\fr |

$S^2$ = 3.5072, $V^2$ = 0.0090, **$X^2$ = 0.1751**

The program given in the appendix can be feed on any data with gaps (like the example below) to fill the gaps (like the filled gaps) doing this by optimizing the expectances (three alternatives, $S^2$, $V^2$ and $X^2$ respectively).

## Appendix - program filling the data with gaps and computing the expected values in the gaps (PHP source code)

[A long uncommented computer code listing does not belong in an encyclopedia! RDG]

```
function get_all(&$o){ //get the data from data.txt file
$a = explode("\r\n",file_get_contents("data.txt"));
$o = array(); for($i = 0; $i < count($a); $i++){$o[$i] = explode("\t",$a[$i]); }
}
function gen_mat($plus){ //censor the data from data.txt file

$a = explode("\r\n",file_get_contents("data.txt"));
$o = array(); for($i = 0; $i < count($a); $i++){$o[$i] = explode("\t",$a[$i]); }
$n = count($o); $m = count($o[0]);
for($i = 0; $i < $n; $i++)for($j = 0; $j < $m; $j++)$q[$i][$j] = ""; //gaps everywhere here

for($i = 0; $i < $n; $i++){$j = rand(0,$m-1); $q[$i][$j] = $o[$i][$j]; } //some gaps filled with values in lines (n)

for($j = 0; $j < $m; $j++){$i = rand(0,$n-1); $q[$i][$j] = $o[$i][$j]; } //some gaps filled with values in columns (m)

//gapped data file should have at least one nonempty value on each line and column
for($k = 0; $k < $plus; $k++){
for($add_plus = 0; $add_plus < $plus - $m - $n; ){
$i = rand(0,$n-1); $j = rand(0,$m-1);
if(($q[$i][$j] = = = ""){$q[$i][$j] = $o[$i][$j]; $add_plus++; }
}
}
$r = array(); for($i = 0; $i < $n; $i++)$r[$i] = implode("\t",$q[$i]);
file_put_contents("data_censored.txt",implode("\r\n",$r));
}
function get_mat(&$o){ //get censored data
$a = explode("\r\n",file_get_contents("data_censored.txt"));
$o = array(); for($i = 0; $i < count($a); $i++){$o[$i] = explode("\t",$a[$i]); }
}
function set_mat(&$o,&$q){ //copy o into q
$q = array();
for($i = 0; $i < count($o); $i++)for($j = 0; $j < count($o[0]); $j++)$q[$i][$j] = $o[$i][$j];
}
function set1mat(&$o,&$q){ //copy only nonempty values from o into q
for($i = 0; $i < count($o); $i++)for($j = 0; $j < count($o[0]); $j++)if(!($o[$i][$j] = = = ""))$q[$i][$j] = $o[$i][$j];
}
function expect(&$a,&$b){$ss = 0.0; $sr = array(); $sc = array(); $b = array();
for($i = 0; $i < count($a); $i++){
```

```
$sr[$i] = 0.0; for($j = 0; $j < count($a[$i]); $j++)$sr[$i]+ = $a[$i][$j];
} //calculate row sums
for($j = 0; $j < count($a[0]); $j++){
$sc[$j] = 0.0; for($i = 0; $i < count($a); $i++)$sc[$j]+ = $a[$i][$j];
} //calculate column sums
for($i = 0; $i < count($a); $i++)for($j = 0; $j < count($a[$i]); $j++)$ss+ = $a[$i][$j];
for($i = 0; $i < count($a); $i++)for($j = 0; $j < count($a[$i]); $j++)$b[$i][$j] = $sr[$i]*$sc[$j]/$ss;
} //calculate expectances
function estim1(&$b,&$r,&$c){
$r = array(); $r[0] = sqrt($b[0][0]);
$c = array(); $c[0] = sqrt($b[0][0]);
for($i = 1; $i < count($b); $i++)$r[$i] = $b[$i][0]/$c[0];
for($i = 1; $i < count($b[0]); $i++)$c[$i] = $b[0][$i]/$r[0];
} //iterate expectances
function af_mat(&$a){$r = array(); $t = array();
for($i = 0; $i < count($a); $i++){
for($j = 0; $j < count($a[0]); $j++)
$r[$i][$j] = trim(sprintf("%.3f",$a[$i][$j]));
$t[$i] = implode("\t",$r[$i]);
} //display predicted data
file_put_contents("data_pred.txt",implode("\r\n",$t));
}
function val2S(&$a,&$r,&$c){
$s = 0.0; for($i = 0; $i < count($r); $i++)for($j = 0; $j < count($c); $j++)if(!($a[$i][$j] = = = ""))$s+ = pow($a[$i]
[$j]-$r[$i]*$c[$j],2); return($s);
} //iterate S^2 alternative
function val2V(&$a,&$r,&$c){
$s = 0.0; for($i = 0; $i < count($r); $i++)for($j = 0; $j < count($c); $j++)if(!($a[$i][$j] = = = ""))$s+ = pow($a[$i]
[$j]-$r[$i]*$c[$j],2)/pow($r[$i]*$c[$j],2); return($s);
} //iterate V^2 alternative
function val2X(&$a,&$r,&$c){
$s = 0.0; for($i = 0; $i < count($r); $i++)for($j = 0; $j < count($c); $j++)if(!($a[$i][$j] = = = ""))$s+ = pow($a[$i]
[$j]-$r[$i]*$c[$j],2)/($r[$i]*$c[$j]); return($s);
} //iterate X^2 alternative
function af1mat($s,&$o,&$a,&$rf,&$cf){$r = array(); $t = array();
for($i = 0; $i < count($a); $i++){
for($j = 0; $j < count($a[0]); $j++)
$r[$i][$j] = trim(sprintf("%.3f",$a[$i][$j]));
$t[$i] = implode("\t",$r[$i]);
```

```
} //display a resulted matrix
$u = implode("\r\n",$t);
$v = array(); for($i = 0; $i < count($rf); $i++)$v[$i] = trim(sprintf("%.3f",$rf[$i]));
$u. = "\r\nRows factors:\r\n".implode("\t",$v);
$v = array(); for($j = 0; $j < count($cf); $j++)$v[$j] = trim(sprintf("%.3f",$cf[$j]));
$u. = "\r\nCols factors:\r\n".implode("\t",$v);
$u. = "\r\nS2 = ".sprintf("%.4f",val2S($o,$rf,$cf));
$u. = "\r\nV2 = ".sprintf("%.4f",val2V($o,$rf,$cf));
$u. = "\r\nX2 = ".sprintf("%.4f",val2X($o,$rf,$cf));
file_put_contents("data_".$s."_min.txt",$u);
}
function not_empty(&$a){
for($i = 0; $i < count($a); $i++)for($j = 0; $j < count($a[0]); $j++)if(!($a[$i][$j] = = = ""))return(array($i,$j));
} //return first nonempty entry
function fill_recurs($i,$j,&$a,&$r,&$c){

//fill recursive the gaps from the entry point
$i_ = array();
for($k = 0; $k < count($r); $k++)if($k < >$i){
if($a[$k][$j] = = = "")continue;
$r[$k] = $a[$k][$j]/$c[$j]; $i_[] = $k;
}
$j_ = array();
for($l = 0; $l < count($c); $l++)if($l < >$j){
if($a[$i][$l] = = = "")continue;
$c[$l] = $a[$i][$l]/$r[$i]; $j_[] = $l;
}
for($k = 0; $k < count($r); $k++)if(!($r[$k] = = = ""))
for($l = 0; $l < count($c); $l++)if(!($c[$l] = = = ""))
if(($a[$k][$l] = = = ""))$a[$k][$l] = $r[$k]*$c[$l];
for($k = 0; $k < count($r); $k++)if(!($r[$k] = = = ""))
for($l = 0; $l < count($c); $l++)if(!($a[$k][$l] = = = ""))
if(($c[$l] = = = ""))$c[$l] = $a[$k][$l]/$r[$k];
for($l = 0; $l < count($c); $l++)if(!($c[$l] = = = ""))
for($k = 0; $k < count($r); $k++)if(!($a[$k][$l] = = = ""))
if(($r[$k] = = = ""))$r[$k] = $a[$k][$l]/$c[$l];
$empty = 0;
for($k = 0; $k < count($r); $k++)
for($l = 0; $l < count($c); $l++)if(($a[$k][$l] = = = ""))$empty++;
```

```
if($empty>0){
for($k = 0; $k < count($i_); $k++)
for($l = 0; $l < count($j_); $l++)
fill_recurs($i_[$k],$j_[$l],$a,$r,$c);
}
}
function opti_iterat(&$o,&$q){ //optimize the expectances
for($i = 0; $i < 20; $i++){
set1mat($o,$q); expect($q,$e); set_mat($e,$q);
}
}
function sum_row($i,&$a,&$c,$pa,$pc){ //apply a sum function to a row
$t = 0.0;
for($j = 0; $j < count($c); $j++){
$ta = pow($a[$i][$j],$pa); $tc = pow($c[$j],$pc); $t+ = $ta*$tc;
}
return($t);
}
function sum_col($j,&$a,&$r,$pa,$pr){ //apply a sum function to a column
$t = 0.0;
for($i = 0; $i < count($r); $i++){
$ta = pow($a[$i][$j],$pa); $tr = pow($r[$i],$pr); $t+ = $ta*$tr;
}
return($t);
}
function estim2S(&$a,&$r,&$c){
for($i = 0; $i < count($r); $i++)
$r[$i] = sum_row($i,$a,$c,1,1)/sum_row($i,$a,$c,0,2);
for($j = 0; $j < count($c); $j++)
$c[$j] = sum_col($j,$a,$r,1,1)/sum_col($j,$a,$r,0,2);
for($i = 0; $i < count($r); $i++)for($j = 0; $j < count($c); $j++)$a[$i][$j] = $r[$i]*$c[$j];
} //estimate for S^2 to min
function estim2V(&$a,&$r,&$c){
for($i = 0; $i < count($r); $i++)
$r[$i] = sum_row($i,$a,$c,2,-2)/sum_row($i,$a,$c,1,-1);
for($j = 0; $j < count($c); $j++)
$c[$j] = sum_col($j,$a,$r,2,-2)/sum_col($j,$a,$r,1,-1);
for($i = 0; $i < count($r); $i++)for($j = 0; $j < count($c); $j++)$a[$i][$j] = $r[$i]*$c[$j];
} //estimate for V^2 to min
```

```
function estim2X(&$a,&$r,&$c){
for($i = 0; $i < count($r); $i++)
$r[$i] = sqrt(sum_row($i,$a,$c,2,-1)/sum_row($i,$a,$c,0,1));
for($j = 0; $j < count($c); $j++){
$c[$j] = sqrt(sum_col($j,$a,$r,2,-1)/sum_col($j,$a,$r,0,1));
}
for($i = 0; $i < count($r); $i++)for($j = 0; $j < count($c); $j++)$a[$i][$j] = $r[$i]*$c[$j];
} //estimate for X^2 to min

//main program begins here
get_all($o_all); //get data.txt
gen_mat(0); //make data_censored.txt (data with gaps)
get_mat($o); //get censored data from data_censored.txt (data with gaps)
set_mat($o,$q); //make a copy of the censored data
$r = array(); for($i = 0; $i < count($q); $i++)$r[$i] = ""; //empty row factors
$c = array(); for($j = 0; $j < count($q[0]); $j++)$c[$j] = ""; //empty column factors
list($i,$j) = not_empty($q); $r[$i] = sqrt($q[$i][$j]); $c[$j] = sqrt($q[$i][$j]); //estimated factors from first nonempty value
fill_recurs($i,$j,$q,$r,$c); //recursively filled values from nonempty values
opti_iterat($o,$q); //optimize for expectances
af_mat($q); //display the result as data_pred.txt file
set_mat($q,$qS2); //prepare for S^2 to min optimization
estim1($qS2,$rS2,$cS2);
for($k = 1; $k < 20; $k++){set1mat($o,$qS2); estim2S($qS2,$rS2,$cS2); }
af1mat("S2",$o,$qS2,$rS2,$cS2); //display S^2 to min optimization result
set_mat($q,$qV2); //prepare for V^2 to min optimization
estim1($qV2,$rV2,$cV2);
for($k = 1; $k < 20; $k++){set1mat($o,$qV2); estim2V($qV2,$rV2,$cV2); }
af1mat("V2",$o,$qV2,$rV2,$cV2); //display V^2 to min optimization result
set_mat($q,$qX2); //prepare for X^2 to min optimization
estim1($qX2,$rX2,$cX2);
for($k = 1; $k < 20; $k++){set1mat($o,$qX2); estim2X($qX2,$rX2,$cX2); }
af1mat("X2",$o,$qX2,$rX2,$cX2); //display X^2 to min optimization result
```

## References

1. Karl Pearson; On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from

random sampling. *Philosophical Magazine* **1900**, *50*, 157-175, 10.1080/14786440009463897.

2. R. A. Fisher; The Logic of Inductive Inference. *Journal of the Royal Statistical Society* **1935**, *98*, 39-54, 10.2307/2342435.
3. Lorentz Jäntschi; A Test Detecting the Outliers for Continuous Distributions Based on the Cumulative Distribution Function of the Data Being Tested. *Symmetry* **2019**, *11*, 835(15p), 10.3390/sym11060835.
4. R. A. Fisher; On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society* **1922**, *85*, 87-94, 10.2307/2340521.
5. Sorana D. Bolboacă; Lorentz Jäntschi; Adriana F. Sestraş; Radu E. Sestras; Doru C. Pamfil; Pearson-Fisher Chi-Square Statistic Revisited. *Information* **2011**, *2(3)*, 528-545, 10.3390/info2030528.
6. Mugur C. Bălan; Tudor P. Todoran; Sorana D. Bolboacă; Lorentz Jäntschi; Mugur C. Bălan; Sorana Bolboaca; Lorentz Jäntschi; Assessments about soil temperature variation under censored data and importance for geothermal energy applications. Illustration with Romanian data. *Journal of Renewable and Sustainable Energy* **2013**, *5(4)*, 41809(13p), 10.1063/1.4812655.
7. Lorentz Jäntschi; Radu E. Sestras; Sorana D. Bolboacă; Modeling the Antioxidant Capacity of Red Wine from Different Production Years and Sources under Censoring. *Computational and Mathematical Methods in Medicine* **2013**, *2013*, a267360(7p.), 10.1155/2013/267360.
8. Donatella Bálint; Lorentz Jäntschi; Missing Data Calculation Using the Antioxidant Activity in Selected Herbs. *Symmetry* **2019**, *11(6)*, 779(10p.), 10.3390/sym11060779.