WP. 39
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Geneva, Switzerland, 9-11 November 2005)

Topic (vi): Software for statistical disclosure control

THE "JACKKNIFE" METHOD: CONFIDENTIALITY PROTECTION FOR COMPLEX STATISTICAL ANALYSES

Invited Paper

Submitted by the Federal Statistical Office, Germany[1]

[1] Prepared by Jobst Heitzig (jobst.heitzig@destatis.de).

The "Jackknife" Method: Confidentiality Protection For Complex Statistical Analyses

Jobst Heitzig

Federal Statistical Office Germany, IT User Service / Statistical and Geo-Information Systems, Gustav-Stresemann-Ring 11, 65189 Wiesbaden, Germany.
(jobst.heitzig@destatis.de)

Abstract. The "jackknife" method of confidentiality protection is a kind of compromise between protection methods based on the data (anonymisation) and protection methods based on results (like those used for tabular data). Like the former, it allows to perform all kinds of statistical analyses with the confidential micro-data, but like the latter, it allows us to release results of higher accuracy than can be computed from traditionally anonymised micro-data.

The idea is to publish not the precise analysis result but a small interval containing it, where the interval's width is chosen as small as possible but still large enough to ensure confidentiality. More precisely, this protection width is chosen so that a potential data snooper cannot distinguish between his sought target value and some random replacement value.

Depending on the actual analysis, the protection width can be determined in different ways: For robust statistics with bounded influence functions (e.g., the median), that function can be used. For non-robust or fairly robust statistics (e.g., the mean or trimmed mean), the effect of replacing each micro-data cell one at a time by a random replacement is determined or estimated, respectively. At the moment, efficient algorithms exist to protect most classical univariate and bivariate statistics and some non-linear model fitting algorithms. Prototypical implementations in SAS® are available for this.

1 Introduction

The sciences' growing demand for all kinds of statistical analyses with confidential micro-data can be answered in several ways. Recently, there seems to be some shift from releasing anonymised micro-data files to providing remote access facilities. This paper describes a new method for confidentiality protection which can in principle be applied to arbitrarily complex statistical analyses of confidential micro-data, and presents some prototypical software tools which implement the method for certain important kinds of analysis.

The proposed method achieves confidentiality protection by publishing analysis results only with some imprecision. An essential feature is that the imprecision is kept as small as possible but still large enough to ensure confidentiality even when the potential data snooper has large amounts of additional information. The basic idea of this "jackknife" method is that the necessary amount of imprecision can in

principle be determined in a way similar to the jackknife estimation of standard errors: compute a set of approximate analysis results, each based on a slightly modified micro-data file which coincides with the original data in all but one position. The published result is then an interval containing all these approximate results. Section 2 of this paper motivates and justifies the basic principle.

In practice, these approximate results can often be efficiently determined or estimated by adjusting the true result after replacing a single value in the micro-data. In case of a robust statistic (such as the median) whose influence function is bounded, it is even more efficient to use this bound directly to compute an interval that can be published safely. Section 3 gives various examples of how the jackknife method can be used to publish all kinds of descriptive statistics and test statistics, showing also that the relative imprecision introduced by the jackknife method is usually of order $O(1/N)$ or $O(\ln(N)/N)$, whereas relative standard errors are usually of the larger order $O(1/\sqrt{N})$. Section 4 then describes two corresponding prototypical SAS® macros, `%jk_means` and `%jk_freq`, which have been developed by the Federal Statistical Office Germany.

As an example of how the method works also with advanced statistical analyses, Section 5 presents a macro `%jk_nlin` providing jackknife protection for the least-squares parameter estimators of non-linear regression models.

The current status of our research is described in Section 6.

2 Motivation, basic principle and rationale

2.1 Disadvantages of anonymised micro-data files

Releasing anonymised micro-data has certain well-known problems. First of all, anonymisation as it is currently performed must often be tailored to each single data file and is almost always based on certain assumptions as to which variables are sensitive, which are possible key variables, what amount of additional knowledge the snooper might have, which observations the snooper might be interested in, which degree of uncertainty would render the disclosed data useless for the snooper, and so on. Furthermore, their level of protection is often measured by some aggregate measure of risk (e.g., the estimated percentage of re-identifiable units in certain subgroups), implying that there can easily remain a small percentage of observations which could still be at a high *individual* risk of disclosure.

At the same time, raising this level of protection costs a lot. For one thing, global anonymisation methods (like sub-sampling, global recoding, additive or multiplicative perturbation, etc.) increase the error of most analysis results by a constant factor, independently of the actual number of observations entering the specific analysis, and independently of how problematic these results are with respect to confidentiality. For example, relative standard errors of statistics computed from a 70% subsample are usually about 20% higher than in the whole sample. And recoding both the row and column variables for a χ^2-test from $R = C = 9$ to $R = C = 3$ categories increases the relative standard error ($\sqrt{2/(R-1)(C-1)}$) of the test statistic even by 300%. When variables are removed or recoded to too coarse a level, their quantitative analysis becomes impossible. In addition, anonymisation

often introduces a bias into many multivariate and/or non-linear analyses. Even a simple estimation of the sum gets biased when top-coding is used.

All these disadvantages can be justified when the goal is to hand out micro-data to researchers so that they can look at them or run simulations, for example. But when the goal is to provide researchers with a remote access facility for complex but somewhat standardised statistical analyses, they seem to be a great price.

2.2 Difficulties in judging by the number of observations

One approach to protect confidential data in a remote access facility could be to publish precise results when the number of used observations seems large enough, and to suppress the output altogether when it is not – just as it is often done in case of frequency tables. However, already simple examples show that it is not at all obvious what number of observations should be considered safe, even when we only want to provide the researcher with some standard descriptive statistics.

Assume there were seven persons with two variables X and Y, and we were to publish the (sample) mean, variance, skewness and kurtosis of both, together with their covariance. Then any two persons in that group could easily compute the Y-value of any third person in the group of whom they know the X-value. They only need to subtract their own values and then solve the nine equations for the nine unknown values. Or assume there were even 20 persons and we only wanted to publish the (sample) means m_X and m_Y, variances s_X^2 and s_Y^2, and the covariance of X and Y. Then it might turn out that in this particular group the Pearson correlation between X and Y is 0.99, which would allow anyone who knows the X-value of any of the persons to infer with certainty that the corresponding Y-value is in an interval with centre $m_Y + 0.99 s_Y \frac{x - m_X}{s_X}$ and a width $\leqslant 1.2 s_Y$ (note that this is not a confidentiality region but certain). This interval becomes even narrower the more x deviates from m_X, that is, the more unusual the target person is (see [Heitzig 2004]).

The existing literature on "statistical databases" shows that this approach of judging by the number of observations is complicated enough already when one wants to publish only sums.

2.3 Definition of the jackknife method

From a mathematical point of view, almost every statistical method of analysis can be formalised as a function f defined on the set D of possible micro-data files, whose value $f(M)$ (often a real number or vector) is in some set W (e.g., the set of real numbers).

Now, in analogy to the jackknife estimation of standard errors, the jackknife method for confidentiality protection is based on the principal idea to calculate a number of *approximate results* $f(M_i)$ instead of the true result $f(M)$, where for each approximate result we use a slightly *modified micro-data file* M_i instead of the true file M. The file M_i is produced from M by replacing an individual value e_i at exactly one position i of the original file by a *replacement value* z_i (drawn with a pseudo-random number generator, for instance) from which e_i cannot be

determined. The index i runs through all positions of the individual *values* of the original file M (that is, i is not a row but a cell index!), and the distribution of z_i is independent of e_i and sufficiently widespread. For example, this *replacement distribution* could be a uniform distribution on the domain of the corresponding variable, or a normal distribution centred at the mean and with twice the standard deviation of the variable, or a conditional distribution from some regression model, or derived by some imputation method, etc. When the variable can contain missing values in principle, a missing value should also be used for z_i with some probability. In addition, it is necessary in some cases to use not only one but several replacements. For instance, one could use three replacements in case of "dummy" 0-1-variables and two replacements in case of variables with three to seven possible values, so that in each case the probability that all these replacements equal the true value is at most $\frac{1}{8}$.

The set $F(M)$ of all thus computed approximate results $f(M_i)$ is the basis for what we publish. In some cases, the true result can be estimated from $F(M)$ with large certainty (e.g., if $f(M)$ is a frequency, then $F(M) = [f(M) - 1, f(M) + 1]$) with large probability, hence $F(M)$ cannot be published directly without revealing $f(M)$. Therefore, $F(M)$ gets moderately enlarged in some random way, giving a *publishing set* $V(M)$ which is finally published instead of the accurate result $f(M)$.

In case of metric or ordinal result values, $V(M)$ would be an *interval* $V(M) \supset F(M)$. For a metric result, a good choice for $V(M)$ seems to be the interval

$$f(M) + 4(b-a)\delta \quad \pm \quad \max\{2, 1 + 4a + 4b)\}\delta$$

where $\delta := \max_i |f(M_i) - f(M)|$ is the maximal error of the approximate results, and a and b are drawn independently from a Beta$(2, 3)$ distribution. In this way, the published interval's centre $f(M) + 4(b-a)\delta$ has a nearly normal distribution with mean $f(M)$ (so that unbiasedness is preserved), standard deviation 1.13δ and slightly non-normal kurtosis 2.68, but its maximal deviation from $f(M)$ is 4δ. In the special case of frequency tables, this is roughly comparable to adding a Normal$(0, 1.13^2)$-distributed noise. Note that, even if $a = b = 0$, the published interval's width is at least 4δ, so that the snooper cannot infer $f(M)$ even if he guesses δ.

Although most analysis results are real numbers or vectors, or are at least on an ordinal scale, the method also works for non-ordinal results (e.g., the mode of a categorical variable). In this case, one could publish a set $V(M) \supset F(M)$ of less than thrice the cardinality of $F(M)$, by adding a number of $2|F(M)| - 2$ values which are drawn independently and with replacement from the uniform or marginal distribution on all possible result values.

2.4 Mechanism of confidentiality protection

The rationale behind using $F(M)$ to estimate how much imprecision is sufficient for confidentiality protection is that, in this way, the published result is not only compatible with the true data, but also with data in which just the single value the snooper might be interested in was replaced by a random value. Thus the snooper should not be able to distinguish between the true and the replaced value.

Let us assume that the data snooper is interested in some individual *target value* e_t in the micro-data. We can formalise his additional knowledge by assuming he knows that $M \in A$, where A is the (usually infinite) set of all possible micro-data files which are not in conflict with his additional knowledge. If $f(M)$ was published, the snooper could try to calculate e_t from $f(M)$ using his additional knowledge. In principle, this corresponds to determining the pre-image $U := f^{-1}(f(M))$, then computing the intersection $S := U \cap A$, and finally determining the set E_t of all values e'_t occurring at the *target position* t in some of the micro-data files $M' \in S$. The snooper would then know at best that $e_t \in E_t$. The risk of (attribute) disclosure consists in the fact that, given sufficiently detailed additional knowledge, that is, given that A is sufficiently small, the set S may contain only files M' in which $e'_t = e_t$, so that the set E_t would only contain the true value e_t.

Now assume that, following the jackknife method, we publish the set $V(M)$ instead of $f(M)$, and that the snooper tries an attribute disclosure as above. Then he gets a considerably larger pre-image $U := f^{-1}(V(M))$ which contains all the files M_i (but might not contain M). Even in the extreme case where the snooper would already know *all* individual values of M other than e_t, the only file in U compatible to this knowledge is M_t, from which he can at best determine the replacement value z_t, but this tells him nothing about e_t.

Using an alternative strategy, the snooper might also try to determine a set of micro-data files not containing M but some specific replacement file M_j with $j \neq t$, since this file *would* contain the true value e_t instead of z_t. In order to study this strategy, we can again formulate the additional knowledge of the snooper as $M_j \in \tilde{A}$. But because M_t and M_j only differ in the two positions j and t, and since the snooper has no knowledge about the randomly chosen values z_t and z_j, we can conclude that $M_t \in \tilde{A}$, i.e., the snooper cannot distinguish M_t from M_j. Again he cannot determine whether e_t or z_t is the true value, even when he already knows all values except e_t.

For a metric result $f(M)$ and $V(M) = f(M) + 4(b-a)\delta \pm \max\{2, 1+4a+4b)\}\delta$ with $a, b \sim \text{Beta}(2,3)$, as it was suggested above, one can show that, with at least 91% probability, the same interval $V(M)$ had been published if z_t instead of e_t had been the true value and if δ would not be affected by this change. For the usually vast majority of target positions t for which $|f(M_t) - f(M)| < \frac{\delta}{3}$, this probability increases to at least 99%.

Despite these considerations, a formal proof of the sketched protection mechanism remains to be found.

3 Performance and simple examples

The effort needed to apply the described jackknife method to some specific kind of analysis depends on how much the computation of all the $f(M_i)$ costs. Many statistical analyses can be implemented in a way which makes the effort of determining the difference $\delta_i := f(M_i) - f(M)$ independent of N, the number of observations, so that the total effort is some small constant multiple of that needed to compute $f(M)$. For example, this is the case for statistics based on moments or power-sums, like fre-

quencies, sums, square sums, means, standard deviations, (co-)variances, (partial) product-moment correlations, skewness, kurtosis, Cronbach's alpha, test statistics of t- and F-tests, all kinds of table statistics for (fixed size) contingency tables, etc.

Alternatively, if the δ_i are known to fulfil some upper bound, this bound can also be used directly to compute an interval $V(M)$. This is closely related to basic concepts from the theory of robustness (see [Hampel 1986]). If the gross error sensitivity γ^* of f is finite, then δ_i is asymptotically bounded by $2\gamma^*/N$. Otherwise, it usually has asymptotic upper confidence limits of order $\ln(N)/N$. This indicates that the width of $V(M)$ will be of smaller order than the standard error. Consequently, the jackknife method can be expected to clearly out-perform anonymisation methods for large N.

For example, let $f(M)$ be the sample mean of a sample of N values, drawn independently from the standard normal, and let the replacement values also come from that distribution. Then $\delta < \ln(N)\frac{\sqrt{2}}{N}$ with probability at least 95% for all $N \geqslant 15$, and some further calculation shows that $V(M)$'s centre $f(M) + 4(b-a)\delta$ differs from $f(M)$ by at most $16\ln(N)/5N$ with probability at least 90% whenever $N \geqslant 15$. When we compare this to the 90% confidence limit of the sampling error of $f(M)$, which is about $2/\sqrt{N}$, we find that for $N \geqslant 30$, the confidence limit of the additional imprecision is smaller than that of the sampling error, while for $N < 30$, the former is still at most 1.18 times the latter. This shows that even for the extremely non-robust sample mean, the imprecision that is introduced additionally by the jackknife method is acceptable when compared to the sampling error.

Here are some examples in which an upper bound for δ can be used to construct $V(M)$ without actually computing the $f(M_i)$:

- Order statistics: $f(M) = x_{(k)}$, $\delta \leqslant \max\{x_{(k)} - x_{(k-1)}, x_{(k+1)} - x_{(k)}\}$.
- k-times trimmed mean: $\delta \leqslant \max\{x_{(N-k+1)} - x_{(k+1)}, x_{(N-k)} - x_{(k)}\}/N$.
- Kendall's and Spearman's rank correlation: $\delta \leqslant \frac{6}{N-3}$ resp. $\delta \leqslant \frac{6}{N-1}$.
- Sign test: $f(M) = (N^+ - N^-)/2$ (which is of order N), $\delta \leqslant 1$.
- Wilcoxon's signed rank test: $f(M) = \sum\{\text{rank}(|x_i - \mu_0|) : x_i > \mu_0\}$ (of order N^2), $\delta \leqslant N$.
- Wilcoxon's test for two samples: $f(M) = \sum\{\text{rank}(x_i) : x_i \in \text{sample } 1\}$ (of order N^2), $\delta \leqslant \max\{N_1, N_2\}$.
- Kolmogorov–Smirnov test of fit: $f(M) = \sup_x |F_N(x) - F(x)|$ (F_N being the empirical distribution function), $\delta \leqslant 1/N$.
- Bowker's test for $R \times R$-tables: $f(M) = \sum\sum_{i<j}(n_{ij} - n_{ji})^2/(n_{ij} + n_{ji})$ (of order N), $\delta \leqslant 4(2R-3)$.
- Entropy: $f(M) = \sum_i \frac{n_i}{N}\log_2\frac{n_i}{N}$, $\delta \leqslant 2\frac{\log_2 N}{N}$.
- Greenwood's G (Sum of Squares of Spacings): $f(M) = \sum_i (x_{(i)} - x_{(i-1)})^2$, $\delta \leqslant \max\{\max_i(x_{(i)} - x_{(i-1)})^2/2, 2\max_i(x_{(i+1)} - x_{(i)})(x_{(i)} - x_{(i-1)})\}$.

4 Implementation for univariate statistics and contingency table statistics

For most of the above-mentioned statistics, the Federal Statistical Office Germany has implemented the jackknife method prototypically as SAS® macros. These macros `%jk_means` and `%jk_freq` provide essentially the same functionality (and similar syntax) as the original SAS procedures `means` and `freq`, with some additional robust statistics. Their basic syntax is

```
%jk_means ( data = dataset, where = optional condition,
            by = optional classifying variables,
            var = analysis variables,
            weight = optional weight variable,
            stats = requested statistics,
            jk_cntl = control dataset )

%jk_freq  ( data = dataset, where = optional condition,
            by = optional classifying variables,
            row = row variable, col = column variable,
            jk_cntl = control dataset )
```

with some additional advanced options. The *control dataset* specifies the replacement distributions for the variables in a certain way. `%jk_means` currently reports intervals for these statistics:

`N, SumWgt`	No. of observations and sum of weights
`Mean, StdErr, LCLM, UCLM`	Mean with standard error and confidence limits
`Sum, USS, CSS`	Sum and [un]corrected square sum
`StdDev, LCLStd, UCLStd`	Standard deviation with confidence limits
`Var, CV`	Variance, coefficient of variation
`T, ProbT`	t-test for $mean = \mu_o$, with p-value
`Skew, Kurt`	Skewness and kurtosis
`Min, Max, Range`	Extremes and their difference
`Q1, Q3, QRange`	Quartiles and their difference
`P1, P5, P10, P90, P95, P99`	Further percentiles
`Median, Biweight, Trimean`	Robust location estimators
`MAD`	Median absolute deviation from the median
`QSkew, MSkew`	Bowley's and Pearson's measures of skewness
`H10Skew, H5Skew, H1Skew`	Hinkley's robust measures of skewness
`KurtB, M5Kurt, CS5Kurt`	Some robust measures of kurtosis (see [Blest 2003])

`%jk_freq` currently computes:

χ^2-tests (classical, likelihood-ratio, continuity-adjusted, Mantel-Haenszel)
Derived statistics (Phi, contingency coefficient, Cramer's V)
Measures of association with asymptotic tests (Gamma, τ_b, τ_c, Somers' D)
Measures of association with asymptotic confidence limits
(Spearman, [a]symmetric Lambda and uncertainty coefficients)
Tests for agreement or trend (Bowker, Kappa coefficient, Cochran-Armitage)
p-values for all tests (one- and two-sided)

For all reported values, the published interval's centre is an unbiased and consistent estimator of the precise sample value of the statistic. Note that anonymisation methods, in contrast, do not usually guarantee unbiasedness.

5 Example: non-linear OLS regression

A number of statistical analyses, such as methods of model fitting, are based on parameter estimation by numerical optimisation. Although the effect of replacing a single e_t by z_t on the estimators found by such algorithms cannot in general be determined exactly without repeating the optimisation, it (or some upper bound for it) can still often be estimated quite accurately, for example by using the first and second derivatives of the objective function at the found optimum. Assume that ϑ is a k-dimensional vector of real parameters, $L(M;\vartheta)$ is the objective function (e.g., least-squares loss), and $\frac{\partial}{\partial\vartheta}L(M;\vartheta_{\text{opt}}) = 0$. Now, under certain smoothness assumptions, the Theorem on implicit functions implies that, when $\tilde{M}$ is sufficiently near M, then $\frac{\partial}{\partial\vartheta}L(\tilde{M};\tilde{\vartheta}_{\text{opt}}) = 0$ for some $\tilde{\vartheta}_{\text{opt}}$ which fulfils

$$\tilde{\vartheta}_{\text{opt}} - \vartheta_{\text{opt}} \approx \left(\frac{\partial^2}{\partial\vartheta^2}L(M;\vartheta_{\text{opt}})\right)^{-1} \frac{\partial}{\partial\vartheta}L(\tilde{M};\vartheta_{\text{opt}}).$$

That is, the change of the estimated parameters due to a small change in the data is approximately the inverse Hessian matrix of the objective function times the gradient of the objective function at the original estimators but with the new data. Using this approximation, δ can be estimated quickly in $O(Nk^2)$ time. Although this estimation of δ is somewhat less thorough than the exact computation, additional confidentiality protection arises from the fact that from such numerical optimisation results $f(M)$, it is even more difficult to determine the pre-image $U = f^{-1}(f(M))$ than in case of other kinds of analysis.

The above technique is used in the macro `%jk_nlin` which reports parameter estimates for non-linear least-squares regression:

```
%jk_nlin ( data = dataset, where = optional condition,
           model = model equation without error term,
           parms = parameters with start values,
           jk_cntl = control dataset )
```

As in the SAS procedure `nlin`, also probit models (and analogously logit and complementary log-log models) can be estimated by specifying as model equation

$$0 = \sqrt{-2\ln\begin{cases} \Phi(h) & \text{if } y = 0 \\ 1 - \Phi(h) & \text{if } y = 1 \end{cases}},$$

where Φ is the standard normal distribution function, y is the dependent variable and h is some function of the predictors. This is because minimising the least-squares loss of this model corresponds to maximising the (log-)likelihood of the actual probit model.

6 Status and further steps

At the moment, the Research Data Centres of the Federal Statistical Office Germany and the Statistical Offices of the Länder offer researchers who want to analyse confidential data three ways of access: they can use Scientific Use Files, or come to one of the offices and work at so-called "safe scientific workstations", or submit program code for manual execution ("controlled remote data processing"). For the latter two ways of access, confidentiality protection is performed manually on a per-case basis, using traditional protection methods as far as possible.

The jackknife method of confidentiality protection is *not* yet used for any requests from researchers. Currently, the Federal Statistical Office is trying to evaluate the practical quality of results produced with the jackknife method and compare them with results from anonymised business-data files from the project "De-facto anonymisation of business micro-data" (see [Ronning et al. 2005]), and later on also with anonymised household-data files, both for small and large N.

We plan to proceed to prototypically implement the method for further types of analyses like (partial) correlations, principal components analysis, forecasting, plots, ANOVA, etc. For evaluation purposes, all prototypes are/will be available to experts on confidentiality protection, upon request.

The main task, however, will be to find a thorough proof of the conjectured level of protection. We would be grateful for any helpful comments in this direction.

Our hope is that, eventually, the method could be integrated into a remote access facility for micro-data from German official statistics, so that researchers would be able to comfortably perform many kinds of statistical analyses with confidential data at their home office and still get high-quality results.

References

Blest, D. C. (2003), "A new measure of kurtosis adjusted for skewness", Australian and New Zealand Journal of Statistics, 45 (2) 175–179.

Hampel et al. (1986), Robust Statistics, Wiley.

Heitzig, J. (2004) "Protection of Confidential Data when Publishing Correlation Matrices", in: Proceedings in Computational Statistics (16th COMPSTAT Symposium), 1163–1170.

Ronning et al. (2005), Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten (Statistik und Wissenschaft, Vol. 4), Statistisches Bundesamt, Wiesbaden.