# Information Value Statistic and Predictors for Logistic Regression

Bruce Lund

Data Mining Consultant, Novi, MI

## Abstract

In preparing predictor variables for binary logistic regression model it is common to collapse the levels of a nominal or discrete-valued predictor X to achieve parsimony while maintaining predictive power. Once the levels have been binned, the binned predictor is transformed to weight-of-evidence (WOE) coding for usage as a predictor in the model.

In the first section of the paper an algorithm is given for collapsing the levels of a nominal or discrete-valued predictor X for predicting binary Y so that information value (IV) is maximized at each step in the collapsing. The algorithm allows the ordering of X to be maintained during the collapsing if X is ordinal. This algorithm is coded in SAS®.

In the second section a process is given to simulate the probability distribution of IV under the assumption of no association between X and Y. Since, in practice, IV does not have a parametric probability distribution, this simulation provides a tool to reject non-significant IV.

**Key Words:** information value statistic, weight-of-evidence, logistic regression, optimal binning, log likelihood.

## 1. Introduction

The information value (IV) of a predictor X and the binary target Y can be given as a formula involving an X-Y frequency table as shown in Table 1. The notation "G" and "B" is taken from credit scoring where "G" is "good" (paid as agreed) and "B" is "bad" (default). $G_k$ refers to the count of "goods" corresponding to $X = X_k$. In contrast, $g_k$ refers to the percent of total goods corresponding to $X = X_k$. Likewise for "bads", $B_k$ and $b_k$ .

**Table 1: How to compute the Information Value Statistic**

| X | Y = 0 "B" | Y = 1 "G" | b: Col % Y = 0 | g: Col % Y = 1 | Log(g/b) (base e) | g - b | (g - b) Log(g/b) |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 1 | 0.400 | 0.333 | -0.1823 | -0.067 | 0.0122 |
| 2 | 1 | 1 | 0.200 | 0.333 | 0.5108 | 0.133 | 0.0681 |
| 3 | 2 | 1 | 0.400 | 0.333 | -0.1823 | -0.067 | 0.0122 |
| Sum | 5 | 3 | | | | IV = | 0.0924 |

The IV formula is written as

$$IV = \sum_{k=1}^{K}(g_k - b_k)\log(\frac{g_k}{b_k})$$

where the number of levels of X is $K \geq 2$ and $g_k$ and $b_k$ are positive for all k = 1,…,K.

The formula for weight-of-evidence (WOE) coding of X is given as X_woe where:

$$\text{If } X = X_k \text{ then } X\_woe = \log(g_k/b_k) \text{ for } k = 1 \text{ to } K$$

and $g_k$ and $b_k$ are positive for all $k = 1,\ldots, K$

The greater the IV of a predictor X, the greater will be the strength of X, after WOE coding, in the logistics model. But when X has more than a few levels, it is appropriate to explore the collapsing of levels of X to achieve parsimony before proceeding to WOE coding.[1],[2] However, the collapsing comes at the price of decreased IV. The IV is non-increasing as levels are collapsed and strictly decreases whenever the $i^{th}$ and $j^{th}$ levels are collapsed where $g_i/b_i \neq g_j/b_j$.

The first section of the paper presents an algorithm for iteratively collapsing pairs of levels of X so that IV is maximized at each step. This algorithm also allows the ordering of X to be maintained during the collapsing process if X is ordinal.[3]

Once collapsing is complete the collapsed predictor is transformed by WOE coding and considered for usage in the model.

The second section of the paper gives a procedure for simulating the distribution of the IV under the assumption of no association between X and Y. Here, the row and column totals of the X-Y table are taken as parameters. This simulation procedure provides a method to identify non-significant IV statistics by comparing an observed IV to the upper $95^{th}$ or other cut-off point of the distribution from the simulation.

## 2. Preliminaries

The c-statistic is commonly used to evaluate the strength of a numeric (or ordered) predictor X for potential usage in a logistic regression model with binary target Y.

A formula for the c-statistic is given below:

$$\text{c-stat} = \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} g_i b_j + (0.5) \sum_{i=1}^{K} g_i b_i$$

The c-statistic's range is 0 to 1. It is customary to require c-statistic $\geq 0.5$ by taking $\max(\text{c-stat}, 1 - \text{c-stat})$.

The "c" or model concordance that is given in the output of logistic regression is the c-statistic of P, the probability from the logistic model, and the target Y.

---

[1] Predictors with "continuous" value ranges (e.g. dollars, distances) must first undergo preliminary binning of the levels of X. This preliminary binning is often based on formation of bins with roughly equal counts or of roughly equal width.

[2] But when a predictor X is "continuous" it may be better to search for a function (transformation) of X which provides a good fit to the log-odds of Y. See Royston and Sauerbrei (2008) for their function selection procedure (FSP).

[3] This algorithm and SAS code were introduced by Lund and Brotherton (2013). See also Lund (2013). The SAS code which is provided in Lund and Brotherton (2013) is version v6a which has less functionality than the current version v8f.

The term "x-statistic of X and Y" (or x-stat) will refer to the "c" from logistic regression of Y against X_woe.[4] There are two equivalent characterizations of the x-statistic of X and Y given by (a) and (b):

a) x-statistic equals the "c" from logistic regression of Y against X where X appears as a class variable (each level of X is a dummy variable)
b) x-statistic = $0.5 (1 + \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} | g_i b_j - g_j b_i |)$

Of particular significance is that (b) gives a way to compute the x-statistic within computer programming calculations such as a SAS DATA step.

Several properties of the x-stat and c-stat include:

a) The x-statistic is non-increasing as the levels of X are collapsed.
b) When X is numeric, the c-stat is defined and
   - The x-stat is always greater than or equal to the c-stat.
   - When x-stat equals c-stat, then X is monotonic versus Y. That is, $G_k / (G_k + B_k)$ is non-decreasing (or non-increasing) with respect to the ordering of X.

### 3. An Algorithm for Collapsing the Levels predictor X for binary Y

A common practice in preparing a predictor X for use in a logistic model is to bin the levels of X to remove outliers and reveal a trend. But IV decreases when two levels of X are collapsed with equality occurring only when the odds-ratios from the two levels is equal. In some applications the statistician will employ business knowledge when forming the bins. Alternatively, the statistician may wish to rely on an algorithm to perform the collapsing into bins which maximizes IV at each iteration.

The algorithm finds the two levels (call these levels i and j) when combined together decrease IV the least. This is equivalent to finding i and j so that D is minimized where:

$$D = D_i + D_j - D_{i,j}$$

and

$$D_i = (g_i - b_i) * \log(g_i / b_i)$$
$$D_j = (g_j - b_j) * \log(g_j / b_j)$$
$$D_{i,j} = (g_i + g_j - b_i - b_j) * \log( (g_i + g_j ) / (b_i + b_j) ).$$

The pair (i, j) which minimizes D are collapsed. The algorithm continues in this manner until a decision is made by the user to stop the collapsing. If the predictive variable X is ordered, the algorithm provides the option of maintaining the ordering by collapsing only adjacent levels of X.

The log likelihood of X and Y is given by

$$LL= \sum_{k=1}^{K}(G_k \log \frac{G_k}{(G_k+B_k)} + B_k \log \frac{B_k}{(G_k+ B_k)})$$

---

[4] The term x-statistic is preferable because the x-statistic can be computed without reference to PROC LOGISTIC as shown by (b).

The algorithm also provides the option of collapsing levels of X so as to maximum log likelihood of X as a predictor of Y (as in logistic regression). The two methods IV and LL do not necessarily result in the same sequence of collapsing.[5]

A SAS macro called %BEST_COLLAPSE implements the collapsing algorithm. Some of the macro parameters are listed:

- DATASET:    A data set name
- X:          Character or numeric variable which can have MISSING values
- Y:          Binary Target which is numeric and must have values 0 and 1 without MISSING values
- W:          Numeric frequency variable which has values which are positive integer values
- METHOD:     = IV or LL
              For METHOD = IV the criterion for selecting two eligible levels to collapse is to maximize the IV. The levels that are eligible for collapse are determined by the MODE parameter.
              For METHOD = LL the criterion for selecting two eligible levels to collapse is to maximize the Log Likelihood. The levels that are eligible for collapse are determined by the MODE parameter.
- MODE:       = A or J
              For MODE = A all pairs of levels are compared when collapsing.
              For MODE = J only adjacent pairs of levels are compared when collapsing. See interaction of MODE and MISS in the documentation for MISS.
- MISS:       = MISS
              Missing values for X are treated as a level. If MODE = J, then the "missing" level is not allowed to collapse with any other level. The "missing" level will be one of the 2 levels appearing in the final collapse at k = 2. If X is character, then the "missing" level appears as "!" in the Reports. If X is numeric, then the "missing" level appears as "." in the Reports.
- WOE:        = WOE
              WOE is used to print the WOE SAS-coded transform of X for each iteration of collapsing. WOE code is automatically saved to SAS data sets.

%Best_Collapse uses only SAS DATA STEP's, PROC MEAN's, and PROC APPEND.

There are four reports that can be produced by %BEST_COLLAPSE:
1.  The **Collapse Step** Report shows the detail of the collapsing of the predictor X.
2.  The **Summary** Report gives summary statistics for each step of the collapsing.
3.  **Binary Splits** Report is created only for MODE = J. This report checks that the collapsing process has remained optimal through the final collapse to 2 levels.
4.  **Confidence Interval for the Log-Odds Ratio** report assists in determining a Stopping Point.

---

[5] This is the case for the data of Table 3.

## 3.1. An Example of %BEST_COLLAPSE

Table 3 has counts of coded income levels called INCOME_C versus a binary response Y. INCOME_C is regarded as ordered.

**Table 3: Income Data Set**

| | INCOME_C | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Y** | 01 | 02 | 03 | 04 | 05 | 06 | 07 | 08 | 09 | 10 | 11 | 12 | **Total** |
| 0 | 1393 | 6009 | 5083 | 4519 | 8319 | 4841 | 2689 | 2090 | 729 | 292 | 253 | 294 | 36511 |
| 1 | 218 | 890 | 932 | 1035 | 2284 | 1593 | 1053 | 872 | 311 | 136 | 120 | 142 | 9586 |
| **Total** | 1611 | 6899 | 6015 | 5554 | 10603 | 6434 | 3742 | 2962 | 1040 | 428 | 373 | 436 | 46097 |

%BEST_COLLAPSE is run on the data from table with METHOD = IV and MODE = J. There are no missing values.

Table 4 shows a partial listing of the Summary Report. When the collapsing process reached k = 8 the x-stat equaled the c-stat (0.59783). Therefore, the collapsed X has a monotonic relationship to Y starting with k = 8.

**Table 4: Summary Report of %Best_Collapse of the Income Data Set**

Dataset= IV_Test_Income, Predictor= Income_C, Target= Y, Method= IV, Mode= J, Miss=

Summary Report

| k | IV | X_STAT | C_STAT | L1 | L2 | L3 | L4 to L12 omitted |
|---|---|---|---|---|---|---|---|
| | | | | | | | |
| 12 | 0.12145 | 0.59795 | 0.59775 | 01 | 02 | 03 | |
| 11 | 0.12145 | 0.59795 | 0.59775 | 01 | 02 | 03 | |
| 10 | 0.12144 | 0.59795 | 0.59775 | 01 | 02 | 03 | |
| 9 | 0.12143 | 0.59793 | 0.59773 | 01 | 02 | 03 | |
| **8** | 0.12136 | **0.59783** | **0.59783** | 01+02 | 03 | 04 | |
| 7 | 0.12113 | 0.59753 | 0.59753 | 01+02 | 03 | 04 | |
| 6 | 0.12046 | 0.59707 | 0.59707 | 01+02 | 03 | 04 | |
| **5** | 0.11792 | 0.59463 | 0.59463 | 01+02 | 03 | **04+05** | |
| 4 | 0.11513 | 0.59282 | 0.59282 | 01+02+03 | 04+05 | 06 | |
| 3 | 0.11029 | 0.58905 | 0.58905 | 01+02+03 | 04+05 | 06+07+08+09+10+11+12 | |
| 2 | **0.08439** | 0.56457 | 0.56457 | 01+02+03 | 04+05+06+07+08+09+10+11+12 | | |

The final collapse to k=2 levels gave a binary split of the values of X into [01 to 03] and [04 to 12]. The Binary Split report (Table 5) shows that this IV collapsing process became sub-optimal. Specifically, the split [01 to 04] and [05 to 12] gave the highest binary split with IV = **0.0888** which is greater than the final IV in Table 4 of **0.0844**.

A "wrong path" occurred when the level "04" was joined to "05" instead of to "01+02+03" at k = 5. As a practical matter in this example the user would certainly stop the collapsing process before k=4 due to the large drop-offs in both IV and x-stat at k=6 and further down.[6]

---

[6] See Lund and Brotherton (2013) and Lund (2013) for more discussion of sub-optimal collapsing.

**Table 5: Binary Split Report of %Best_Collapse of the Income Data Set**

| Final Step Binary Splits for MODE = J | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| IV split1 | IV split2 | IV split3 | IV split4 | IV split5 | IV split6 | IV split7 | IV split8 | IV split9 | IV split10 | IV split11 |
| | | | | | | | | | | |
| 0.0082 | 0.0580 | **0.0844** | **0.0888** | 0.0780 | 0.0594 | 0.0371 | 0.0179 | 0.0113 | 0.0076 | 0.0042 |

If levels i and j are selected to be collapsed, then their log-odds ratio is:

$$LO = \log( (G_i / B_i) \ / (G_j / B_j) ).$$

The approximate standard deviation of the LO is

$$LO\_SD = SQRT (1/G_i + 1/B_i + 1/G_j + 1/B_j ).^{7}$$

Assuming cell counts in rows i and j are large, then LO is normally distributed and an approximate 95% confidence interval (CI) is:

$$LO \ +/- \ 2 * LO\_SD \ \text{(approximate 95\% confidence interval for true LO).}$$

If LO = 0, then $g_i / b_i = g_j / b_j$ and collapsing of i and j is a good decision. Roughly, the more LO deviates from 0, the greater will be the decrease in IV from collapsing. A guideline is to stop the collapsing process if LO +/- 2* LO_SD does not include 0.

In Table 6 the 95% CI for the log-odds at level 6 omits zero. This suggests stopping at k = 7. This conclusion is reinforced by examining the change in the IV and x-stat when going from 7 to 6 levels. For each statistic there is a noticeable drop between k = 7 and k = 6. (For example, IV drops from 0.12113 at k = 7 to 0.12046 at k = 6.)

**Table 6: Log-Odds CI Report of %Best_Collapse of the Income Data Set**

| Log-odds with 95% CI | | | | | | | |
|---|---|---|---|---|---|---|---|
| K | IV | x-stat | Collapsing to | LO | LO_SD | LOminus2SD | LOplus2SD |
| | | | | | | | |
| 12 | 0.12145 | 0.59795 | 11 | -0.01820 | 0.15187 | -0.32193 | 0.28553 |
| 11 | 0.12145 | 0.59795 | 10 | -0.02786 | 0.12722 | -0.28229 | 0.22658 |
| 10 | 0.12144 | 0.59795 | 9 | -0.02225 | 0.07882 | -0.17989 | 0.13539 |
| 9 | 0.12143 | 0.59793 | 8 | 0.05507 | 0.08121 | -0.10735 | 0.21749 |
| 8 | 0.12136 | 0.59783 | 7 | -0.06920 | 0.05022 | -0.16963 | 0.03123 |
| 7 | **0.12113** | **0.59753** | 6 | -0.15575 | 0.06583 | **-0.28741** | **-0.02410** |
| 6 | **0.12046** | **0.59707** | 5 | -0.18128 | 0.04178 | -0.26483 | -0.09772 |
| 5 | 0.11792 | 0.59463 | 4 | -0.20287 | 0.04803 | -0.29894 | -0.10680 |
| 4 | 0.11513 | 0.59282 | 3 | -0.23202 | 0.03703 | -0.30609 | -0.15796 |
| 3 | 0.11029 | 0.58905 | 2 | -0.37940 | 0.02655 | -0.43251 | -0.32629 |
| 2 | 0.08439 | 0.56457 | | | | | |

---

[7] Since the standard deviation is given in terms of cell counts instead of column percentages, there is the question of how to adjust the formula for stratified sampling of goods and bads. My suggestion is to make no adjustment since the formula reflects the true degrees of freedom.

Log Likelihood and Information Value do not always collapse in the same way. Using the Income data set (Table 3) and collapsing by MODE = J, the maximum log likelihood and the IV algorithms collapse X differently.

- LL: For k = 5 this algorithm collapsed "03" with "01+02"
- IV: For k = 5 this algorithm collapsed "04" with "05".

## 3.2 Stopping Guidelines for Collapsing

Subjective judgment by the statistician will inevitably play a large role in deciding when to stop collapsing levels when applying %BEST_COLLAPSE. This is sound and practical since the statistician will be familiar with the predictor variable. This judgment can be assisted by the statistics produced by %BEST_COLLAPSE as outlined below:

- If there is an "abrupt" change in IV and x-stat from one step to the next, this is an indication that too much predictive power was lost by a collapse.
- In the case of numeric predictors, the equality of x-stat and c-stat signals monotonicity.
- If the 95% confidence internal for the log odd-ratio omits the value of zero, then a stopping point has probably been reached.

## 3.3 Other Methods Of Binning

1. A method of collapsing nominal predictors (using any-pairs collapsing) is based on clustering of levels using SAS PROC CLUSTER. This method selects the pair for collapsing which maximizes the Pearson chi-square. A stopping criterion is defined by selecting the iteration which produces the minimum chi-square statistic probability (right tail probability) of association between the target and the collapsed predictor.

   The clustering method was illustrated by Manahan (2006, Appendix 1) who provides SAS macro code.[8]

2. A decision tree can also be used to collapse a predictor X. The predictor X can be nominal or ordinal. The leaf nodes that are the result of the splitting process define the collapsed levels. A stopping criterion must be specified.

   JMP® has a decision tree called PARTITION. In the case of a single predictor X and a nominal binary target Y, the entropy criterion (denoted by G^2 in JMP output ) is used to determine where to split. Here, the entropy criterion for splitting is equivalent to Log Likelihood criterion for collapsing.[9]

## 4. Simulation of the Distribution of IV when X and Y are not associated

Popular credit scoring text books give guidelines for evaluation of the strength of a WOE-coded predictor X for a binary target Y in terms of its IV statistic.

The following is taken from Siddiqi (2006 p. 81).

---

[8] Additional code is needed to compute the chi-square probabilities.
[9] http://www.jmp.com/support/help/

IV Guidelines for Rating the Strength of a Predictor in Logistic Regression

- Less than 0.02: unpredictive
- 0.02 to 0.1: weak
- 0.1 to 0.3: medium
- 0.3 +: strong

These guidelines are familiar and are grounded in good practice but there is no statistical measure associated with them.

The IV statistic for Table 7 is IV = 0.0924, a value which is viewed as "weak" by Siddiqi's guidelines. But can we also say that IV = 0.0924 is evidence that there is no association between X and Y?

**Table 7: Example Calculation of Information Value Statistic**

| X | Y = 0 "B" | Y = 1 "G" | b: Col % Y = 0 | g: Col % Y = 1 | Row count N | P = g / (g + b) | Log(g/b) (base e) | g - b | (g - b) * Log(g/b) |
|---|---|---|---|---|---|---|---|---|---|
| X1 | 200 | 100 | 40% | 33.3% | 300 | 0.333 | -0.1823 | -0.067 | 0.0122 |
| X2 | 100 | 100 | 20% | 33.3% | 200 | 0.500 | 0.5108 | 0.133 | 0.0681 |
| X3 | 200 | 100 | 40% | 33.3% | 300 | 0.333 | -0.1823 | -0.067 | 0.0122 |
| SUM | 500 | 300 | 100% | 100% | 800 | 1.000 | | **IV =** | **0.0924** |

## 4.1 The Simulation Procedure

A simulation of the distribution of IV under the null hypothesis of no association between X and Y will allow a test of hypothesis. Specifically, if IV = 0.0924 exceeds (1 - α)% of the distribution values, then the null hypothesis of no association between X and Y could be rejected at α%. This significance would not bring with it a description such as "weakly predictive", only the determination that 0.0924 did not arise by chance from a population of IV values with no association between X and Y.

It is now necessary to make precise the notion of no association between X and Y and the sampling that underlies the simulation. Let $N_k$ for k = 1 to K denote the row totals in a K by 2 table and let $B_T$ and $G_T$ denote the column totals. Also let N be the overall table count. There are three sampling frameworks that can be considered.

1. Only the table total N is fixed. Then random samples are taken from the population of K by 2 tables with fixed table count of N.
2. Both $B_T$ and $G_T$ are fixed. Fixed $B_T$ and $G_T$ would arise from stratified sampling, a common practice in direct marketing and credit scoring. Then two random samples will be taken. One is from the population of K by 1 tables with fixed column total $G_T$ and the other with fixed column total $B_T$.
3. All row and column totals $N_k$, $B_T$, $G_T$ are fixed. Then random samples are taken of the cell counts subject to these fixed marginals.

It seems most natural to select sampling framework #3. Here the statistician accepts the results of the sample that created the training data set. Subject to the fixed marginals, how much information does X give about Y? The result 0.0924 is one instance but could this have arisen by chance if there was no association between X and Y?

Next a process to define and implement the hypothesis of "no association" is needed. The definition and the process are explained together as follows.[10]

- First a sample without replacement of size $N_1$ is taken from a population of size N with $G_T$ "goods" and $B_T$ "bads". This is hypergeometric sampling. Some $G_1$ "goods" are sampled.
- Next a sample without replacement of size $N_2$ is taken from a population of size $N - N_1$ with $G_T - G_1$ "goods". Again this is hypergeometric sampling.
- This process is continued until row K where no sampling is required since $G_K$ is equal to the remaining "goods".

The process produces a count of "goods" $G_1$ through $G_K$ which enables the calculation of an IV. A problem associated with this process occurs if any $G_k$ is 0 or $N_k$. This creates a zero cell which prevents the logarithm in IV to be computed. Such samples will be extremely rare in practical cases but must be discarded if they occur.

The SAS code in Figure 1 implements the hypergeometric sampling process and creates a simulated probability distribution for IV. For simplicity, the program is pre-coded for K=3. To run this program the user specifies values for SAMPLES (the number of IV's to be computed in the simulation), N1, N2, N3, GT, and SEED (a random number stream seed).

**Figure 1: SAS program creates distribution for IV under no association assumption.**

```
DATA SAMPLES;
    ARRAY G{*} G1 - G3; ARRAY B{*} B1 - B3; ARRAY N{*} N1 - N3;
K=3;
    * Parameters;
    SAMPLES = 50000;
    SEED = 12345;
    N1=300; N2=200; N3=300; GT=300;
    * Initializations;
    NT=N1+N2+N3;
    BT=NT-GT; * Total BAD;
    ZERO_FLAG=0;  * Samples with zero cell;
    DO S = 1 TO SAMPLES;
        GX=GT; NX=NT; IV=0;
        CALL STREAMINIT(SEED); /* Creates new random stream */
        DO I = 1 TO K;
            IF GX > 0 THEN G{I} = RAND('HYPER',NX,GX,N{I});
                ELSE G{I} = 0;
            B{I} = N{I}-G{I};
            GX = GX-G{I}; NX = NX-N{I}; * For next pass thru Loop;
            IF B{I} = 0 | G{I} = 0 THEN IV = .;
            END;
        DO I = 1 TO K;
            IF IV > . THEN IV = IV + (G(I)/GT - B(I)/BT) * LOG((G(I)/GT)/(B(I)/BT));
            END;
        IF IV = . THEN ZERO_FLAG = 1;
        OUTPUT;
        END;
RUN;
PROC MEANS DATA = SAMPLES NOPRINT;
VAR IV ZERO_FLAG;
```

---

[10] See Raimi and Lund (2012) for additional discussion

```
OUTPUT OUT = MEANOUT
MEAN(IV) = IV_MEAN
P5(IV) = IV_P5  P10(IV) = IV_P10  P25(IV) = IV_P25
P50(IV) = IV_P50  P75(IV) = IV_P75
P90(IV) = IV_P90  P95(IV) = IV_P95
SUM(ZERO_FLAG) = ZERO_FLAG;
RUN;
PROC PRINT DATA = MEANOUT;
RUN;
```

The result of a simulation with SAMPLES = 50,000 is given in Table 8. The 95th percentile of the distribution is IV = **0.0324** (IV_P95) which is far less than the observed IV =**0.0924**. On that basis, the null hypothesis of no association between X and Y is rejected. Also, there were no cases of a zero-cell occurring among the 50,000 in the sample.

**Table 8: IV Distribution for Marginals of Table 7 under No-Association Assumption**

| COUNT | IV_MEAN | IV_P5 | IV_P10 | IV_P25 | IV_P50 | IV_P75 | IV_P90 | **IV_P95** | ZERO_FLAG |
|---|---|---|---|---|---|---|---|---|---|
| 50000 | 0.0107 | 0.0006 | 0.0009 | 0.0028 | 0.0075 | 0.0149 | 0.0247 | **0.0324** | 0 |

## 4.2 Is there a Close-End Formula for IV Distribution under No-Association?

For Table 7 there are 39,601 unique combinations of $G_1$, $G_2$, $G_3$ which meet the marginal requirements.[11] Many of these combinations lead to identical IV values. The task for a close-end description for the IV distribution under no-association is to compute a probability for each 39,601 and to apply these probabilities to the computation of the probability distribution of IV.

Although a close-end formula seems possible in principle, given the complexity of such a formula there is no benefit versus a simulation.

## 4.3 Does the Order in which the K rows enter the DO LOOPS matter?

The simulation begins with k=1 with $G_1$=200 and $N_1$=300. The results of the sampling without replacement for k=1 determine the parameters for sampling without replacement for k=2. Tests were performed for Table 7 and other examples where the ordering of the K rows was reshuffled. With large enough SAMPLES the distributions of IV are almost identical. A mathematical proof has not been developed, however.

## 4.4 Conclusion

The simulation of the distribution of IV is suitable for small samples (of the size of Table 7) where the guidelines of Siddiqi do not reflect sample size and randomness considerations.

## 5. Software is Available

The SAS macro %BEST_COLLAPSE will be provided by the author.
Contact: blund_data@mi.rr.com

---

[11] This count reflects the fact that some hypergeometric samples, such as $G_1$=0, $G_2$=150, $G_3$=150 must be excluded.

# References

Lund, B. (2013). "Preparing Interaction Variables for Logistic Regression", *SCSUG 2013, Proceedings,* South Central SAS Users Group, Inc.

Lund, B. and Brotherton, D. (2013). "Information Value Statistic", *MWSUG 2013, Proceedings,* Midwest SAS Users Group, Inc., paper AA-14.

Manahan, C. (2006). "Comparison of Data Preparation Methods for Use in Model Development with SAS Enterprise Miner", *Proceedings of the 31th Annual SAS Users Group International Conference*, Paper 079-31.

Raimi, S. and Lund, B. (2011). "Before Logistic Modeling - A Toolkit for Identifying and Transforming Relevant Predictors", *MWSUG 2011, Proceedings*, Midwest SAS Users Group, Inc., Paper SA-03-2011.

Royston P. and Sauerbrei W. (2008). *Multivariate Model-building,* John Wiley & Sons, Ltd, West Sussex, England.

Siddiqi, N. (2006). *Credit Risk Scorecards,* Hoboken, NJ: John Wiley & Sons, Inc.