**New Revelations About the Titanic: Loglinear Model Analysis of the Passenger Data**

Harry J. Khamis

Professor and Director Emeritus

Statistical Consulting Center

Wright State University

Dayton, OH 45435

Mobile: 937-546-6092

Email: [harry.khamis@wright.edu](mailto:harry.khamis@wright.edu)

**New Revelations About the Titanic: Loglinear Model Analysis of the Passenger Data**

Abstract

A loglinear model analysis of the Titanic passenger data is conducted. Such an analysis method is more comprehensive and leads to more accurate results than previous analyses of these data. The principal finding is that the conventional wisdom about the sociodemographic predictors of survival is false. Specifically, it is not true that, *in general*, (i) women and children survived at a higher rate than men, nor that (ii) the survival rates for passengers differed according to economic status. In fact, the data do not strongly support a gender differential in survival rate among (i) Scandinavians or (ii) Europeans in 1st or 3rd classes. In terms of economic status, the data do not strongly support a 1st–versus–2nd class differential in survival rate among passengers from the Americas or Scandinavia, nor among women and children from Britainia/Ireland or Europe. Further, the data do not strongly support a 2nd-versus-3rd class differential in survival rate among men from Britainia/Ireland, Europe, or Scandinavia. Previous analyses of the Titanic data have relied on marginal contingency tables that have been formed in violation of the collapsibility theorem, thereby producing distorted association measures. The estimated risk ratios and corresponding confidence intervals deriving from the loglinear model analysis are the most accurate to date.

Key words: survival rate, class, region, women and children first, association, risk ratio, sociodemographic factors

1. Introduction

There have been innumerable films, documentaries, articles, and books discussing the tragic fate of the White Star Line RMS Titanic, the largest and most luxurious ocean liner of its day. After striking an iceberg on April 14, 1912, the Titanic, said to be unsinkable, foundered, and more than 1,500 lives were lost. Over 100 years later interest in this maritime disaster remains high due to three relatively recent events.

* Through advances in oceanic research techniques, including the use of undersea robots, wreckage of the Titanic was discovered south of Newfoundland more than 1200 feet below the surface of the North Atlantic in 1985 by American and French researchers. The leader of the joint expedition was Dr. Robert D. Ballard of the Woods Hole Oceanographic Institution in Massachusetts. Since 1985, hundreds of artifacts have been recovered from the Titanic and appear in museums and traveling exhibitions.
* In 1997, James Cameron directed and wrote the movie, *Titanic*, starring Leonardo DeCaprio and Kate Winslet as two passengers in far different social classes who fall in love on the ocean liner during its ill-fated maiden voyage. Again, the most modern technology available was used to create realism in the film, including scale models and computer-generated imagery. *Titanic* was both a critical and commercial success, winning 11 academy awards and tied with Ben Hur as the most awarded film in Oscar history. A 3-D version of the film was made in 2012 to commemorate the centennial of the sinking.
* On April 15, 2012 the centennial of the sinking of the Titanic was observed. Commemorative and remembrance events were in abundance around the world. In the U.S. there are permanent Titanic museums located in Pigeon Forge, TN and Branson, MO. In addition to museums, there were several cruises to the site of Titanic’s demise in memory of the lives lost.

Since the disaster, there have been many investigations in an attempt to understand how

such a tragedy could occur and how it might be prevented in the future. In addition there has been research carried out in an attempt to understand the sociological aspects of the tragic event. Specifically, it is of interest to know which sociodemographic factors influenced the outcome of death or survival. Such research is of no value unless there is accurate information about the passengers. A great amount of effort has been devoted to this very issue, see, e.g., Beavis (2002), Soldner (2000), and Dawson (1995).

The data in Table 3 of Gleicher and Stevans (2004) are based on Soldner’s (2000) compilation, regarded as the most accurate passenger list available. These data include only passengers, not crew. The data are presented in Appendix I with two modifications:

1. the ‘REGION’ categories have been rearranged in order to (1) provide a natural partition of geographic regions and (2) to produce a contingency table having sufficient cell frequencies for reliable statistical analysis; see Appendix III of Gleicher and Stevans (2004) for the groupings of the nationalities, and
2. the nine passengers for whom there is incomplete data are omitted from the data set.

Note that the factor ‘GENDER’ corresponds to the two categories ‘Men’ and ‘Women

and Children (W&C)’. This dichotomy is formed for two reasons:

1. it is intended to reflect the first principle of sea disasters: ‘women and children first’, and
2. one of the principal research questions in this study is: to what degree was this principle adhered to?

2. Methods

The statistical techniques used by previous researchers to analyze the Titanic data range from simple descriptive cross-tabulations of factors (contingency tables) to sophisticated models such as logistic regression. However, these analyses all suffer from a major flaw: conclusions are based in part on marginal (collapsed) contingency tables without having checked to determine if such tables produce distorted association measures.

In this paper, the Titanic passenger data are analyzed using a *loglinear model* (LLM). A LLM is a *comprehensive* assessment of the structural associations among *all* of the factors in a contingency table. Based on the LLM derived from the data, a *collapsibility theorem* identifies which variables may be collapsed over without distorting relationships, and which variables may *not* be collapsed over without distorting relationships. Such an analysis avoids the pitfall of drawing conclusions from a marginal table for which the association measure is distorted.

When all of the factors in a study are categorical, as in the case of the Titanic data, the LLM provides a more comprehensive analysis of the data and results in more accurate conclusions than does any other statistical model. The full spectrum of statistical tools used in the following analyses, and the reason for each, are given below. As a general reference for these techniques, see Agresti (2013).

* **Loglinear model**. The LLM uses the contingency table data (Appendix I) to identify the structural associations among the factors REGION, CLASS, GENDER, and SURVIVED. See Appendix II for a brief description of the development of the LLM and how it works.
* **Collapsibility theorem**. This theorem is used to determine which factors may be collapsed over without distorting the associations in the resulting marginal tables. See Appendix III for a brief description of the collapsibility theorem.
* **Association graph and multigraph**. These are advanced procedures that use mathematical graphs to aid in the analysis and interpretation of the LLM and the collapsibility theorem. See Khamis (2011).
* **Breslow-Day test**. This test determines if there is strong evidence in the data to conclude that the association for a set of stratified two-way contingency tables is nonhomogeneous.
* **Cochran-Mantel-Haenszel estimate**. This is the estimate of the *common* association measure for a set of stratified two-way contingency tables having homogeneous association.
* **Fisher’s Exact Test**. This test is used to determine the statistical significance of association in two-way contingency tables that are sparse (i.e., for which the traditional chi-squared test is not reliable).

In the following sections, an introduction to contingency tables is given, then the LLM is fitted to the Titanic passenger data, analyzed, and the results interpreted. The collapsibility theorem is applied to the LLM to determine which marginal tables (if any) may safely be analyzed for associations. New revealing information about the sociodemographic predictors of survival will result from this comprehensive analysis.

Both SAS (Statistical Analysis System, 2013) version 9.4 and SPSS (Statistical Package for the Social Sciences, 2013) version 22 are used for statistical computations. All inferences are made at the 0.05 level of significance.

3. Contingency Tables: Preliminaries

A *contingency table* is a cross-classification of observations for two or more discrete factors. Notationally, a contingency table factor will be presented in capital letters. As an example, the cross-classification of passenger survival, denoted by ‘SURVIVED’, and GENDER, derived from Appendix I by collapsing over the levels of REGION and CLASS, is given by the following contingency table.

**Table 1. Cross-classification of Appendix I data according to SURVIVED and GENDER.**

SURVIVED

Yes No  **Total**

---------------------------------------------------------------------------------------------

W&C\* 367 160 527

GENDER

Men 131 651 782

----------------------------------------------------------------------------------------------

**Total** 498 811 1309

\*W&C: Women & Children

From Table 1 it can easily be seen that = 69.6% of the W&C survived while = 16.8% of the men survived. The *risk ratio* is defined as the ratio of these two survival rates: Interpretation: the survival rate of W&C is approximately four times that of men. However, let us ask the following question: how does this relationship change when we factor in CLASS and REGION? The above table is a *marginal* (or *collapsed*) table because it was obtained by collapsing (or adding) over the levels of CLASS and REGION. Was the SURVIVEDGENDER relationship distorted by this collapsing? These are the questions addressed by the LLM.

More generally, we measure the association between the ROW and COLUMN factors in a 2 2 contingency table with the *risk ratio*. Consider the following contingency table, Table 2.

**Table 2. Generic 22 ROW by COLUMN contingency table with cell frequencies a, b, c, and d.**

COLUMN

|  |  |  |  |
| --- | --- | --- | --- |
| ROW | 1 | 2 | **Total** |
| 1 | a | b | **a+b** |
| 2 | c | d | **c+d** |
| **Total** | **a+c** | **b+d** | **a+b+c+d** |

For this contingency table, the *risk ratio* of COLUMN 1 (comparing ROW 1 to ROW 2) is defined as: . Interpretation:

[the *risk* of COLUMN 1 for ROW 1] = (RR)·[the *risk* of COLUMN 1 for ROW 2].

That is, for those in ROW 1, the risk of being in COLUMN 1 is RR times that of those in ROW 2. When RR = 1, then the two factors, ROW and COLUMN, are said to be *statistically independent*. In this case, the risk of being in COLUMN 1 is the same for both ROWs.

It is easy to show that an association between two factors can change upon collapsing over the levels of a third factor. Consider the three-way contingency table for ROW COLUMN LAYER in Table 3.

**Table 3. Three-way contingency table for ROW COLUMN LAYER; fictitious data.**

COLUMN

|  |  |  |  |
| --- | --- | --- | --- |
| LAYER | ROW | 1 | 2 |
| 1 | 1 | 100 | 50 |
|  | 2 | 50 | 100 |
|  |  |  |  |
| 2 | 1 | 50 | 100 |
|  | 2 | 100 | 50 |

After collapsing over the two LAYERs in Table 3 we get Table 4.

**Table 4. Two-way marginal contingency table for ROW COLUMN collapsed over LAYER in**

**Table 3.**

COLUMN

|  |  |  |
| --- | --- | --- |
| ROW | 1 | 2 |
| 1 | 150 | 150 |
| 2 | 150 | 150 |

Here we have and . In the collapsed table, . So ROW and COLUMN are associated in each of the two LAYERs in Table 3, but they are independent in the collapsed table, Table 4.

In some cases it is safe to collapse over the levels of a third variable because the association is *preserved* under the collapsing. Consider Table 5 below.

**Table 5. Three-way contingency table for ROW COLUMN LAYER; fictitious data.**

COLUMN

|  |  |  |  |
| --- | --- | --- | --- |
| LAYER | ROW | 1 | 2 |
| 1 | 1 | 5 | 10 |
|  | 2 | 10 | 20 |
|  |  |  |  |
| 2 | 1 | 10 | 15 |
|  | 2 | 20 | 30 |

After collapsing over the two LAYERs in Table 5 we get Table 6.

**Table 6. Two-way marginal contingency table for ROW COLUMN collapsed over LAYER in**

**Table 5.**

COLUMN

|  |  |  |
| --- | --- | --- |
| ROW | 1 | 2 |
| 1 | 15 | 25 |
| 2 | 30 | 50 |

The RR for ROWCOLUMN in each LAYER in Table 5 is 1.0 and the RR for the collapsed table, Table 6, is 1.0. In this case, the independence of ROW and COLUMN is *preserved* under the collapsing of the two LAYERs.

From these illustrations it should be clear that the RR measuring the association between two factors can change upon collapsing over the levels of a third factor. For further discussion of this phenomenon see Khamis, 2011, chapter 5. When the association between two factors actually *reverses* upon collapsing over the levels of a third factor, then the phenomenon is called *Simpson’s Paradox*; see Simpson (1951). For example, it is possible that > 1 for each LAYER, but < 1 in the table collapsed over the levels of LAYER. There are many real-data examples exhibiting this phenomenon; see, e.g., Agresti (2013), Appleton et al. (1996), Deming (1975), and Wagner (1982).

Technically, the LLM uses contingency table data to model the natural logarithm of an expected cell frequency in terms of contrasts of logarithms of marginal expected cell frequencies (see Appendix II for details). The analysis begins by fitting different LLMs to the data to determine which model fits best. When a ‘best-fitting’ model is obtained from the data, it is used to identify the structural associations among all of the factors in the contingency table. The ‘best-fitting’ LLM is the one having the fewest terms but for which all statistically significant terms are included. The collapsibility theorem is then applied to this model to determine which factors may be collapsed over without distorting associations (see Appendix III for details of the collapsibility theorem). When collapsing is prohibited by the collapsibility theorem, then alternative analysis strategies must be employed, such as using the Cochran-Mantel-Haenszel estimator or stratifying the analysis according to the levels of a third factor.

4. Analysis Results

Because there is only one 1st CLASS Asian and twelve 2nd CLASS Asians, the data for Asia are considered to be too sparse in these categories to insure reliable results. Therefore, this ‘REGION’ is analyzed separately from the comprehensive analysis using the LLM (see the summary section below). Only the four main REGIONs, Americas, Britainia/Ireland, Europe, and Scandinavia, will be included in the LLM analysis. The four-way REGION SURVIVED GENDER CLASS contingency table in Appendix I, omitting Asia, contains 48 cells with total sample size n = 1185. Consequently, there is an average frequency of 24.7 per cell. This sample size is entirely adequate to produce reliable results from the LLM analysis (see, e.g., Roscoe and Byars, 1971).

A backward elimination LLM selection procedure with entry criterion P < 0.05 (SPSS, 2013) is applied to the Titanic passenger data (i.e., the data in Appendix I omitting Asia) in order to identify the best-fitting LLM. The LLM resulting from this procedure is the one for which the four-factor interaction, REGION SURVIVED GENDER CLASS, is statistically nonsignificant (P = 0.438) and the three-factor interaction REGION SURVIVED CLASS, is statistically nonsignificant (P = 0.148), but for which the three remaining three-factor interactions are highly statistically significant. The resulting best-fitting LLM contains three three-factor interactions,

* REGION GENDER CLASS (P < 0.000),
* REGION SURVIVED GENDER (P < 0.000), and
* SURVIVED GENDER CLASS (P < 0.000),

This best-fitting LLM fits the data well (P = 0.222, based on the goodness of fit likelihood ratio chi-squared statistic).

We now apply the collapsibility theorem to this LLM. According to the collapsibility theorem, for a LLM this complex *no variable may be collapsed over without potentially distorting the associations of the factors in the resulting marginal tables*. Consequently, drawing conclusions from *any* table formed by collapsing over one or more variables can produce incorrect results. For example, the relationship between SURVIVED and GENDER as measured from the contingency table collapsed over REGION and CLASS is given by = 4.2. However, the results of the LLM analysis show that this number depends on REGION and CLASS. But how? In order to answer this question we break down the analyses according to REGION. That is, we analyze the structural associations among SURVIVED, GENDER, and CLASS for each REGION separately.

AMERICAS

For the Americas the best-fitting LLM for the SURVIVED GENDER CLASS contingency table is one for which the three-factor interaction is not significant (P = 0.2885) but all three two-factor interactions are significant (P 0.0021 in each case). Such a model is called the *homogeneous association model* because the association between any two factors is homogeneous (not significantly different) across the levels of the third factor. In particular, the association between SURVIVAL and GENDER is homogeneous across the three CLASSes, and the association between SURVIVAL and CLASS is homogeneous for the two GENDERs.

Because the SURVIVEDGENDER association is homogeneous across CLASS levels according to the LLM, a *common* SURVIVEDGENDER association can be obtained using the Cochran-Mantel-Haenszel (CMH) estimate. The CMH common SURVIVEDGENDER estimated risk ratio across CLASSes with 95% CI is: 3.4; [2.6, 4.4]. Interpretation: the survival rate for W&C is 3.4 times that for Men, regardless of CLASS.

The CMH common SURVIVEDCLASS estimated risk ratios with 95% CIs are given in Table 7.

**Table 7. Americas: CMH common risk ratio point estimates across GENDERs and 95%**

**confidence intervals for the SURVIVEDCLASS associations**

CLASSes

1st vs 2nd 1st vs 3rd 2nd vs 3rd

1.3; [1.0, 1.6] 2.5; [1.6, 3.8] 2.0; [1.2, 3.1]

Regardless of GENDER, 1st CLASS passengers survived at a rate of 2.5 times that of 3rd CLASS passengers and 2nd CLASS passengers survived at a rate of 2.0 times that of 3rd CLASS passengers. Because the CI for 1st vs 2nd CLASSes contains 1.0 (barely), we conclude that there is not quite strong enough statistical evidence at the 0.05 level of significance to claim that the survival rate differs between these two CLASSes.

BRITAINIA/IRELAND

For Britainia/Ireland, the SURVIVED CLASS GENDER three-factor interaction is significant (P < 0.0001). Consequently, if any factor is collapsed then the association in the resulting two-way marginal table for the other two factors may be distorted. The SURVIVEDGENDER association will be analyzed for each CLASS separately, and the SURVIVEDCLASS association will be analyzed for each GENDER separately in Tables 8 and 9 respectively.

**Table 8. Britainia/Ireland: risk ratio point estimates and 95% confidence intervals for the**

**SURVIVEDGENDER association for each CLASS.**

CLASS

1st 2nd 3rd

---------------------------------------------------------------------------------------------------------------

4.2; [2.3, 7.6] 15.8; [7.3, 34.5] 5.4; [3.1, 9.4]

The imbalance in survival rate of W&C over Men is statistically significant in every CLASS, but it’s far larger in 2nd CLASS than in 1st or 3rd CLASSes.

**Table 9. Britainia/Ireland: risk ratio point estimates and 95% confidence intervals for the**

**SURVIVEDCLASS association for each GENDER.**

CLASS comparisons: W&C Men

----------------------------------------------------------------------------------------------

1.0; [0.9, 1.2] 3.9; [1.4, 10.3]

1.9; [1.5, 2.5] 2.5; [1.1, 5.5]

1.9; [1.5, 2.3] 0.6; [0.3, 1.6]

The survival rate does not differ significantly between 1st and 2nd CLASSes for W&C, nor between 2nd and 3rd CLASSes for Men (note that these CIs contain 1.0). For all other comparisons, the survival rate in the higher CLASS is between 2 and 4 times that in the lower CLASS. For 1st versus 3rd CLASSes, the values for the two GENDERs (1.9 and 2.5) do not differ significantly (P = 0.0733; Breslow-Day test). The CMH common risk ratio across GENDERs for 1st versus 3rd CLASSes is 2.1; [1.6, 2.9].

EUROPE

For Europe the SURVIVED CLASS GENDER three-factor interaction is significant (P = 0.0078). Consequently, if any factor is collapsed then the association in the resulting two-way marginal table for the other two factors may be distorted. The SURVIVEDGENDER association will be analyzed for each CLASS separately and the SURVIVEDCLASS association will be analyzed for each GENDER separately; the results are given in Tables 10 and 11 respectively.

**Table 10. Europe: risk ratio point estimates and 95% confidence intervals for the**

**SURVIVEDGENDER association for each CLASS.**

CLASS

1st 2nd 3rd

-----------------------------------------------------------------------------------------------------------------------

1.5; [1.0, 2.2] 8.8; [2.4, 33.0] 2.6; [1.0, 7.0]

There is not strong statistical evidence of a differential survival rate between Men and W&C in 1st and 3rd CLASSes. The survival rate for W&C is higher than for Men (by a factor of almost nine) in 2nd CLASS.

**Table 11. Europe: risk ratio point estimates and 95% confidence intervals for the SURVIVED**

**CLASS association for each GENDER.**

CLASS comparisons: W&C Men

--------------------------------------------------------------------------------------------

1.1; [0.9, 1.3] 6.3; [1.6, 24.9]

3.5; [1.7, 7.2] 6.7; [3.2, 13.9]

3.5; [1.6, 7.6] 1.1; [0.2, 4.5]

The survival rate does not differ significantly between 1st and 2nd CLASSes for W&C, nor between 2nd and 3rd CLASSes for Men. For all other comparisons, the survival rate in the higher CLASS is about 3.5 to 7 times that in the lower CLASS. For 1st versus 3rd CLASSes, the values for the two GENDERs (3.5 and 6.7) do not differ significantly (P = 0.2882; Breslow-Day test). The CMH common risk ratio across GENDERs for 1st versus 3rd CLASSes is 4.4; [2.6, 7.4].

SCANDINAVIA

For Scandinavia the SURVIVED CLASS GENDER three-factor interaction is significant (P = 0.0075). Consequently, if any factor is collapsed then the association in the resulting two-way marginal table for the other two factors may be distorted. The SURVIVEDGENDER association will be analyzed for each CLASS separately and the SURVIVEDCLASS association will be analyzed for each GENDER separately. Results are given in Tables 12 and 13 respectively.

**Table 12. Scandinavia: risk ratio point estimates and 95% confidence intervals for the**

**SURVIVEDGENDER association for each CLASS.**

CLASS

1st 2nd 3rd

---------------------------------------------------------------------------------------------------------------

1.4; [0.6, 3.3] 15.0; [1.0, 232.5] 1.4; [0.9, 2.3]

There is not strong evidence from the data that the survival rate for W&C differs from that of Men in any of the three CLASSes, though the result is nearly statistically significant for 2nd CLASS passengers. The lack of statistical significance may be due to the sparseness of the data in 1st and 2nd CLASSes.

**Table 13. Scandinavia: risk ratio point estimates and 95% confidence intervals for the**

**SURVIVEDCLASS association for each GENDER.**

CLASS comparisons: W&C Men

-------------------------------------------------------------------------------------------------

1.2; [0.6, 2.1] 12.5; [0.8, 206.8]

2.7; [1.7, 4.5] 3.0; [1.3, 7.3]

2.5; [1.5, 4.4] 0.2; [0.0, 3.4]

The survival rate does not differ significantly between 1st and 2nd CLASSes for either W&C or Men, nor between 2nd and 3rd CLASSes for Men. For all other comparisons, the survival rate in the higher CLASS is between 2.5 and 3 times that in the lower CLASS. For 1st versus 3rd CLASSes, the values for the two GENDERs (2.7 and 3.0) do not differ significantly (P = 0.6004; Breslow-Day test). The CMH common risk ratio across GENDERs for 1st versus 3rd CLASSes is 2.8; [1.8, 4.4].

SURVIVAL RATE COMPARISONS AMONG THE FOUR REGIONS

Table 14 shows the estimated survival rates among the four REGIONs for each GENDER-CLASS combination of levels.

**Table 14. Survival rate point estimates for the four REGIONs broken down by GENDER and**

**CLASS\*.**

REGION

GENDER CLASS Americas Brit/Ireland Europe Scandinavia

Men 1 .31a .22a .67b .67b 2 .16 .06 .11 .00 3 .08b .09b .10b .22a

W&C 1 .96 1.00 1.00 1.00

2 .85 .91 .93 .80

3 .44 .49 .26 .32

\*Estimated survival rates in a given row having different letters in the superscript are statistically significantly different at the 0.05 level of significance; otherwise, rates are not significantly different. (Bonferroni correction, SPSS, 2013.)

Table 14 reveals that only for Men in 1st and 3rd CLASSes do the survival rates differ significantly among the four regions. Specific comparisons are made below.

* For Men in 1st class, the survival rate is significantly higher for Europe and Scandinavia (67%) than in the Americas or Britainia/Ireland (average 29.3%).
* For Men in 3rd class, the survival rate is significantly higher in Scandinavia (22%) than in the Americas, Britainia/Ireland, or Europe (average 9.2%).
* For all other GENDER-CLASS levels, the survival rate does not differ significantly among the four REGIONs.

5. Summary

Trying to condense all of the results presented above into just a few lines would be difficult at best. One way to make results more ‘digestible’ is to present them through simple graphics. The two figures below display the survival risk ratio point estimates along with the 95% CIs in the four main REGIONs for Men versus W&C (Figure 1) and for comparisons among the three classes (Figure 2). Only those comparisons that are statistically significant are presented in Figures 1 and 2. Consequently, none of the CIs presented in these two figures cross over the vertical line at RR = 1 (dotted line). The width of the CI is an indication of the degree of reliability of the point estimate. Generally, the lower the sample size the wider the CI.

**Figure 1. Point estimates and 95% confidence intervals for .**

C:\Users\w001hjk\AppData\Local\Microsoft\Windows\Temporary Internet Files\Content.IE5\COKBI967\fig1.tiff

**Figure 2. Point estimates and 95% confidence intervals for .** C:\Users\w001hjk\AppData\Local\Microsoft\Windows\Temporary Internet Files\Content.IE5\EYZPJ6YS\fig2.tiff

If we declare that a survival differential exists between GENDERs or CLASSes *only* if the passenger data strongly (α = 0.05) support such a declaration, then the results of this LLM analysis clearly indicate that

(i) adherence to the principle “women and children first” and

(ii) commonly held beliefs that those in higher CLASSes survived at higher rates

are not always true, that they depend on other factors.

With this in mind, general conclusions concerning the survival rate comparison

(i) between Men and W&C, (ii) among the three CLASSes, and (iii) among the four REGIONs will be made separately.

SURVIVAL COMPARISONS BETWEEN MEN AND W&C

* Adherence to the ‘W&C first’ principle (i.e., > 1.0) is not strongly supported for (1) Scandinavian passengers in any CLASS or (2) European passengers in 1st or 3rd classes.
* For Britainia/Ireland, the is higher for 2nd class than for 1st or 3rd classes.
* For Europe, the is significant for 2nd class only.
* There appears to be no class distinction in the gender survival differential for the Americas; i.e., does not differ significantly among the three CLASSes.
* Adherence to the ‘W&C first’ principle appears to be most extreme for 2nd class passengers from Britainia/Ireland and Europe with 15.8 and 8.8 respectively.

SURVIVAL COMPARISONS AMONG THE THREE CLASSES

There is not strong evidence in the passenger data to conclude that the survival rate differs:

* between 1st and 2nd CLASS passengers from the Americas or Scandinavia
* between 1st and 2nd CLASS W&C from Britainia/Ireland or Europe
* between 2nd and 3rd CLASS Men from Britainia/Ireland, Europe, or Scandinavia

For other CLASS comparisons, there is a significant difference in survival rates, as follows.

* the only significant differences in survival rate between 1st and 2nd CLASSes is for Men in Britainia/Ireland and Europe with = 3.9 and 6.3 respectively,
* there is a significant difference in survival rate between 1st and 3rd CLASSes for every REGION and both GENDERs, with ranging from 2.1 to 4.4,
* there is a significant difference in survival rate between 2nd and 3rd CLASS passengers from the Americas for both GENDERs and for W&C from Britainia/Ireland, Europe, and Scandinavia with ranging between 1.9 and 3.5.

SURVIVAL COMPARISONS AMONG THE FOUR REGIONS

* the survival rate for 1st CLASS Men from Scandinavia and Europe is higher than for Men from the Americas or Britainia/Ireland by a factor of 2.3,
* the survival rate for 3rd CLASS Men from Scandinavia is higher than for Men from the Americas, Britainia/Ireland, or Europe by a factor of 2.4,
* for W&C, regardless of CLASS, the survival rate did not differ among the four REGIONs
* for 2nd CLASS Men, the survival rate did not differ among the four REGIONS.

For Asia the SURVIVED CLASS GENDER structural associations could not be determined reliably through a LLM analysis due to sparseness in the data (lack of convergence by the algorithm). Note, for example, that there was only one Asian in 1st CLASS. Therefore, nothing at all can be said about the SURVIVEDGENDER association in 1st CLASS, except that the one man in 1st CLASS survived.

In 2nd CLASS there are only twelve Asians. To determine the statistical significance of the SURVIVEDGENDER association Fisher’s Exact Test can be applied (see, e.g., Agresti, 2013). This test results in P = 0.0101. Conclusion: there is a statistically significant association between SURVIVED and GENDER; W&C survived at a significantly higher rate than Men.

In 3rd CLASS there are 111 Asians, a large enough sample size to insure the reliability of the asymptotic chi-squared test. In this case, P < 0.0001. Conclusion: there is a highly statistically significant association between SURVIVED and GENDER for 3rd CLASS Asians; the estimated RR is 4.11; [2.37, 7.11]. Thus, it appears that W&C survived at a higher rate than Men for 3rd CLASS Asians.

6. Conclusions

It is important to note that the Appendix I data do not represent a truly random sample from some well-defined population about which we are making inferences. For this reason, terms such as ‘statistical significance’ cannot be interpreted in the precise statistical sense. Instead, we might regard these data to be loosely representative of some theoretical population of interest. The resulting analyses then provide an approximation of the relationships of interest. It is in this context that the term ‘statistical significance’ is used.

The conclusions presented here are derived *exclusively* from the Titanic passenger data. A risk ratio is not declared different from 1.0 and two survival rates are not declared different unless there is *strong* evidence in the data to substantiate such a declaration, namely, a statistical test resulting in P < 0.05 or a CI that excludes 1.0 (‘statistical significance’).

The conventional wisdom concerning the sociodemographic aspects of the Titanic is that, generally, W&C survive at a higher rate than Men, and that the survival rates are generally negatively correlated with CLASS number. The ‘new revelations’ emanating from the loglinear model analysis are that (i) these beliefs are not true in all cases and that (ii) previous estimates of survival rate comparisons are inaccurate due to violations of the collapsibility theorem.

In particular, there is not strong enough evidence in the data to conclude that W&C survive at a different rate than Men for (i) Scandinavians or (ii) 1st or 3rd class Europeans. There is not strong evidence of a difference in survival rates between 1st and 2nd classes for

(i) passengers from the Americas, (ii) W&C from Britainia/Ireland or Europe, or (iii) passengers from Scandinavia. There is not strong evidence of a difference in survival rates between 2nd and 3rd CLASS Men from Britainia/Ireland, Europe, or Scandinavia.

It is interesting to note that the best-fitting LLM for three of the four main REGIONs (Britainia/Ireland, Europe, and Scandinavia) is one for which the SURVIVEDGENDERCLASS three-factor interaction is statistically significant. This indicates that the is not the same for every CLASS. However, for the Americas the three-factor interaction is not statistically significant, though all three two-factor interactions are significant. This indicates that the is homogeneous across CLASSes. The implication of these results is that the Americas are more “democratic” than other REGIONs with regard to adherence to the “W&C first” principle across CLASSes.

Broadly speaking, some of the other interesting features revealed by the LLM analysis are:

* there is not strong statistical support that the ‘W&C first’ principle was realized for Scandinavians,
* the ‘W&C first’ principle was strongly adhered to among 2nd class passengers in all REGIONs except Scandinavia, especially so for Britainia/Ireland and Europe,
* all three CLASSes in the Americas adhered to the ‘W&C first’ principle approximately equally,
* the only significant 1st vs 2nd class difference in survival rate is for Men in Britainia/Ireland and Europe,
* there is not a 2nd vs 3rd class difference in survival rate for Men in Britainia/Ireland, Europe, or Scandinavia,
* the difference in survival rate between 1st and 3rd class passengers is significant for all REGIONs and is about the same for both Men and W&C,
* the only significant differences in survival rate among the four REGIONs occur for 1st and 3rd CLASS Men.

Stating that the survival rate of W&C on the Titanic is about four times that of Men, as one might conclude from the two-way marginal contingency table, is extremely misleading. Based on the Titanic passenger data, the above analyses clearly indicate that this number depends on other factors such as REGION and CLASS. Similarly, comparisons among the CLASSes depends on REGION and GENDER. But *how* do these comparisons depend on the other factors? This question is now, for the first time, thoroughly and comprehensively answered through a LLM analysis of the passenger data.

These results give rise to a number of sociological, behavioral, and cultural questions. Why is it that only 1st and 3rd CLASS Men survival rates differ among the REGIONs? Why is it that Scandinavia fairs so well in terms of survival rates? Why is the ‘W&C first’ principle so strong for 2nd CLASS passengers from Britainia/Ireland and Europe? Why do 1st CLASS Men fair so much better than 2nd CLASS Men only in Britainia/Ireland and Europe? Why is there a difference in survival rates between 2nd and 3rd CLASSes for W&C but not for Men from Britainia/Ireland, Europe, and Scandinavia? Hopefully these questions will prompt Titanic researchers with the appropriate expertise to continue the quest for understanding the sociological predictors of survival that were at play on the Titanic on that fateful night.

Finally, while considerable analyses have been performed on these data, the author is acutely aware that every number in Appendix I is connected with passenger lives, lives that have been touched by fate in a cataclysmic and tragic way. As with any human tragedy, the hope is that we learn from the event in order to prevent its future occurrence and to acquire a better understanding of human behavior and the world we live in.

REFERENCES

Agresti, Alan, 2013. *Categorical Data Analysis*, 3rd edition. John Wiley & Sons.

Appleton, David R., French, Joyce M., and Vanderpump, Mark P.J., 1996. Ignoring a Covariate:

An Example of Simpson’s Paradox, *The American Statistical Association* 50(4): pp. 340-341.

Beavis, Debbie, 2002. *Who Sailed on Titanic? The Definitive Passenger Lists.* Ian Allen.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W., 1975. Discrete Multivariate Analysis. MIT

Press.

Dawson, Robert J., 1995. The ‘Unusual Episode’ Data Revisited, *Journal of Statistics Education*,

3(3).

Deming, W.E., 1975. On Probability as a Basis for Action, *The American Statistical Association*

29: pp. 146-152.

Gleicher, David and Stevans, Lonnie K., 2004. Who Survived the Titanic? A Logistic Regression

Analysis, *International Review of Maritime History*, XVI(2), pp. 61-94.

Khamis, Harry J., 2011. *The Association Graph and the Multigraph for Loglinear Models*. Sage.

Roscoe, J.T. and Byars, J.A., 1971. Sample size restraints commonly imposed on the use of the

chi-square statistic, *Journal of the American Statististical Assoc.* 66: 755-759.

Simpson, E.H., 1951. The Interpretation of Interaction in Contingency Tables, *Journal of the*

*Royal Statistical Society* B 13: pp. 238-241.

Soldner, Hermann, 2000. *RMS Titanic: Passenger and Crew List*, Awie Arger Verlag.

Statistical Analysis System (SAS), version 9.4, 2013. SAS Institute Inc., Cary, NC.

Statistical Program for the Social Sciences (SPSS), version 22, 2013. Printice Hall.

Wagner, Clifford H., 1982. Simpson’s Paradox in Real Life, *The American Statistical Association*

36(1): p. 46.

**APPENDIX I**

**Cross-classification of Titanic Passengers: REGION SURVIVED GENDER CLASS**

FIRST CLASS SECOND CLASS THIRD CLASS

REGION SURVIVED M W&C M W&C M W&C **TOTAL**

-------------------------------------------------------------------------------------------------------------------------------

Americas Yes 38 112 4 22 2 11 189 No 83 5 21 4 24 14 151

Britainia/Ireland Yes 8 13 6 63 13 48 151

No 28 0 98 6 132 51 315

Asia Yes 1 0 1 4 12 30 48

No 0 0 7 0 57 12 76

Europe Yes 8 15 2 13 9 5 52

No 4 0 17 1 81 14 117

Scandinavia Yes 2 3 0 4 25 24 58

No 1 0 9 1 89 52 152

-------------------------------------------------------------------------------------------------------------------------------

**TOTAL** 173 148 165 118 444 261 **1309**

**APPENDIX II**

**Brief Description of the Loglinear Model**

The loglinear model (LLM) is a member of the class of generalized linear models using the log link function with a Poisson response. This model is used for modeling contingency table data for the purpose of identifying the structural associations among the categorical factors. The Titanic passenger list data is a prime example of contingency table data.

By way of illustration, consider a three-way contingency table with factors X (row), Y (column), and Z (layer). Consider a null hypothesis (H0) that specifies a certain structural association among the three factors. Let denote the expected cell frequency under this H0 for the ith row, jth column, and kth layer, where i = 1, 3, …, I; j = 1, 2, …, J; and k = 1, 2, … K. Then the natural logarithm of , which we denote by , is modeled as follows:

(1)

There are so-called ‘zero-sum constraints’ imposed on the -terms appearing on the right side of (1); namely, the sum over any subscript of any -term is zero:

These constraints are necessary in order to insure unique estimates for the -terms when the model is fitted to the contingency table data. Henceforth the zero-sum constraints are assumed to be in force whenever a LLM is written.

The -terms can be expressed as *effects* of the levels of the factors X, Y, and Z, analogous to the standard ANOVA model. For example, when X, Y, and Z are mutually independent, the LLM becomes:

(2)

and

.

Similarly for and .

The different types of structural associations among X, Y, and Z are determined by which sets of -terms are set to zero in (1). For three categorical factors, there are five types of structural associations, as follows.

MUTUAL INDEPENDENCE: in (1).

The LLM is given in (2). Interpretation: X, Y, and Z are statistically mutually independent. If we denote the probability that a subject falls in the ith row, jth column, and kth layer by , then we can write , where N = population size. The probabilistic model for mutual independence can then be written as

(3)

where a ‘+’ in the subscript represents summation over the index replaced by the ‘+’ symbol. It can be shown that the probabilistic model given in (3) is mathematically equivalent to the LLM given in (2).

JOINT INDEPENDENCE: in (1).

The LLM is . Interpretation: X is statistically independent of both Y and Z, but there is an association between Y and Z. The probabilistic model is:

Note that there are two additional joint independence models:

and

.

CONDITIONAL INDEPENDENCE: in (1).

The LLM is . Interpretation: X is statistically independent of Y conditional on Z; i.e., for every level of Z, X and Y are independent. The probabilistic model is:

.

Note that there are two additional conditional independence models:

and

HOMOGENEOUS ASSOCIATION: in (1).

The LLM is . Interpretation: every pair of factors is associated, but the association is homogeneous across all levels of the third factor. For this LLM there is no closed form expression for the probabilistic model. Such a LLM is called a *nondecomposable* model; the other LLMs above are called *decomposable* models. Decomposable LLMs have special properties; see Khamis (2011), section 4.4.

SATURATED LLM: No λ-terms are set to zero in (1).

The LLM is given in (1). Interpretation: every pair of factors is associated, and the association is *not* homogeneous across all levels of the third factor. This model contains all possible λ-terms in it.

For M > 3 categorical variables, the LLM contains λ-terms in the saturated LLM. By successively setting sets of λ-terms to zero, starting with the highest order such terms, simpler and simpler models, having fewer and fewer λ-terms, are obtained. By fitting LLMs to contingency table data, we seek the simplest model (having the fewest λ-terms) but retaining all λ-terms that are important (i.e., statistically significant). Such a model is called the *best-fitting* LLM.

For more detailed development of the LLM, see Agresti (2013), chapters 4, 9, and 10 or Khamis (2011), chapters 2 and 3.

**APPENDIX III**

**Brief Description of Collapsibility Conditions**

It can be proven that:

* ROW and COLUMN variables may be independent in each of two LAYERs but associated in the table collapsed over LAYER,
* ROW and COLUMN variables can be associated in each of two LAYERs but independent in the table collapsed over LAYER (see Tables 3 and 4 for an example),
* ROW and COLUMN variables can be independent in each of two LAYERs and independent in the table collapsed over LAYER (see Tables 5 and 6 for an example), and
* ROW and COLUMN variables can be associated in each of two LAYERS with risk ratio

RR 1 in each LAYER, and the risk ratio in the table collapsed over LAYER is also RR.

That is, the association between ROW and COLUMN in each of two LAYERs may or may not be preserved after collapsing over the LAYERs. So, when we collapse over the levels of a factor in a contingency table, how can we know whether relationships among the remaining factors will be preserved or become distorted?

Consider the SURVIVEDGENDER contingency table for Titanic passengers from the Americas in Appendix Table 1 below.

**Appendix Table 1. Cross-classification of Appendix I data according to SURVIVED and**

**GENDER for Titanic passengers from the Americas.**

SURVIVED

Yes No  **Total**

---------------------------------------------------------------------------------------------

W&C\* 145 23 168

GENDER

Men 44 128 172

----------------------------------------------------------------------------------------------

**Total** 189 151 340

\*W&C: Women & Children.

If this table is broken up according to CLASS, we get Appendix Table 2.

**Appendix Table 2. Cross-classification of Appendix I data according to SURVIVED, GENDER,**

**and CLASS for Titanic passengers from the Americas.**

CLASS

1st 2nd 3rd

SURVIVED: Yes No Yes No Yes No

------------------------------------------------------------------------------------------------------------------------------

W&C 112 5 22 4 11 14

GENDER

Men 38 83 4 21 2 24

Note that Appendix Table 1 is obtained from Appendix Table 2 by collapsing over the three levels of CLASS. Each of the three SURVIVEDGENDER tables in Appendix Table 2 (one for each CLASS) is called a *partial table* or *cross-sectional table* because it represents the SURVIVEDGENDER cross-classification for a fixed level of CLASS. Appendix Table 1 is called a *marginal table* because it represents the SURVIVEDGENDER cross-classification after collapsing over the levels of CLASS.

One of the most important results in loglinear model theory is that the association between two factors in a marginal table may not be the same as in the partial tables. That is, in general,

**marginal association partial association.**

In fact, only under very specific conditions is it true that marginal association = partial association. Those conditions are provided in the *collapsibility theorem*, and they generally depend upon the degree of independence or conditional independence occurring among the contingency table factors. Since the loglinear model identifies the structural associations among a set of categorical factors, and hence also the independence and conditional independence structure among the factors, the collapsibility theorem is predicated upon the loglinear model.

A statement of the collapsibility theorem for multiway contingency tables is given in Bishop et al. (1975, page 47); it will not be presented here. However, for just three categorical factors, X, Y, and Z, the collapsibility theorem can be stated simply: the association between X and Y is preserved after collapsing over the levels of Z if (i) X is independent of Z conditional on Y or (ii) Y is independent of Z conditional on X.

Methodologically, the strategy is to obtain the best-fitting LLM from the contingency table data and then apply the collapsibility theorem to that model to determine which factors may be collapsed over without distorting the association in the resulting marginal table. Recent research has provided effective tools using mathematical graphs to interpret the collapsibility conditions for a given LLM; see Khamis (2011).

Consider Appendix Table 2 above. As shown in Section 4 of this paper, the best-fitting LLM for these data is the homogeneous association model. For such a complex model having no conditional independencies, the collapsibility theorem indicates that no collapsing may be done without potentially distorting associations. The SURVIVEDGENDER risk ratio point estimates are given in Appendix Table 3:

**Appendix Table 3. SURVIVEDGENDER risk ratio point estimates and 95% confidence**

**intervals in each CLASS for Titanic passengers from the Americas.**

CLASS

1st 2nd 3rd

3.0; [2.3, 4.0] 5.3; [2.1, 13.2] 5.7; [1.4, 23.3]

According to the LLM, these risk ratios are not statistically significantly different. This is confirmed by the Breslow-Day test, where P = 0.2115 for

H0: SURVIVEDGENDER RRs are homogeneous across CLASSes.

The Cochran-Mantel-Haenszel (CMH) common SURVIVEDGENDER RR estimate is = 3.4 with 95% confidence interval [2.6, 4.4]. The SURVIVEDGENDER RR estimate for the *marginal* table, Appendix Table 1, is = 5.0 with 95% confidence interval [3.4, 7.4]. The RR estimate from the marginal table is an *overestimate* of the true strength of the association by almost 50%, confirming the inappropriateness of collapsing over CLASS in order to study the SURVIVEDGENDER association.

This discussion illustrates the importance of applying the collapsibility theorem to the best-fitting LLM for a set of contingency table data in order to determine which variables can be collapsed over without distorting associations. If one or more variables are collapsed over in violation of the collapsibility theorem, then the association measure for the two variables in the resulting marginal table may be distorted. This discussion also points out the importance of using the CMH estimate of the common association when appropriate. If the association in partial tables is homogeneous but collapsing is not allowed according to the collapsibility theorem, then the overall association should be measured using the CMH estimate, not by collapsing to form the marginal table.