IAC-08-A4.2.04
CORRECTING FOR INTERDEPENDENCE OF TERMS IN THE SAN MARINO SCALE

Prof. H. Paul Shuch
Executive Director Emeritus,
The SETI League, Inc.
121 Florence Drive
Cogan Station PA 17728 USA
paul@setileague.org

Jamie Riggs
Senior Staff Statistician,
Sun Microsystems
Vice President,
Deep Space Exploration Society
jamiedses@spannedsolutions.com

ABSTRACT

The San Marino Scale, an analytical tool for assessing the significance of transmissions from Earth, was adopted by the IAA SETI Permanent Study Group in 2007. This additive model, which encompasses estimates of both signal strength and signal characteristics, remains a work in progress. It is statistically valid only to the extent that the two terms are assumed to be wholly independent discrete random variables. We question this assumption of independence, and, in fact, find a strong negative correlation between the two terms, though we further suspect that the interaction is nonlinear. Thus, we propose to amend the San Marino Scale, making it more statistically robust by capturing, and compensating for, this interdependence between detectability and information content.

KEYWORDS

San Marino Scale, Active SETI, METI, transmission

INTRODUCTION

The San Marino Scale for quantifying the potential impact of transmissions from Earth, as first proposed by Iván Almár (Almár, 2005) at a SETI conference in the small European republic whose name it shares, consists of two presumably independent analytical terms. The Intensity ('**I**') term, treated extensively in Shuch and Almár (2006), quantifies signal strength relative to the background solar flux in the same frequency range, and at the same modulation bandwidth, as the signal of interest. Signal Characteristic (the '**C**' term) relates to information content, as described in Almár and Shuch (2007a). The resulting additive model was validated by applying it to several historical transmissions from Earth (Shuch and Almár, 2007b). This paper seeks to improve quantitative aspects of the San Marino Scale, by testing the validity of one key assumption implicit in the additive model: that the two terms are indeed independent.

INTERDEPENDENCE OF TERMS

One important aspect of the '**I**' term is that it captures the spectral density of any transmission from Earth. As such, it incorporates the modulation bandwidth of the signal, as that parameter establishes the minimum bandwidth required of an extraterrestrial receiver attempting to recover the transmission. The '**C**' term

encompasses information content. Information theory (Shannon and Weaver, 1949) states that transmission bandwidth is correlated to information content. Thus, we expect a certain degree of interdependence between the '**I**' and '**C**' terms.

INDEPENDENCE SAMPLE

In order to test for possible interaction between the '**I**' and the '**C**' terms of the San Marino Scale, we draw a sample from the population of possible interstellar transmissions from Earth. We start by characterizing six different potential Active SETI transmitters: a typical terrestrial UHF television broadcast station, a communications satellite uplink terminal, an amateur radio microwave moonbounce (EME) facility, NASA's Goldstone deep space network station, and the Evpatoria and Arecibo planetary radar transmitters. For each of these various transmitters, we contemplate the results of applying up to four possible modulation schemes: narrowband unmodulated continuous wave (CW), narrowband SSB or FM voice, wideband analog video programming, and ultra-wideband digital data. The resulting combinations of effective isotropic radiated power (EIRP) and information bandwidth allow us to quantify the spectral intensity of 24 different classes of signal, by invoking the '**I**' term of the San Marino Scale. The 24 signals in our sample span the entire 0 to 5 range of '**I**' term possibilities.

Next, we analyze the potential information content of all of our candidate modulation schemes, assigning a '**C**' term to each of the fourteen signals in our transmission sample. On the 1 to 5 range of possible '**C**' term values, our sample spans 1 to 4 (a 5 being reserved for responses to actual SETI detections, none of which has yet been confirmed).

Summing the '**I**' and '**C**' values of our sample, we see that we have described a collection of terrestrial transmissions which score between 1 and 7 on the 1-to-10 San Marino Scale. The results are summarized in Figure 1, below.

ANALYSIS METHOD

Counts of the various SMI levels are generated from **I** and **C** frequency (number of occurrences) data. With these data, we count how many times a level of **I** or **C** occurs, but we have no way of knowing how often these levels did not occur thereby giving us an incomplete, truncated event space. This contrasts with data in which we count the number of specific event occurrences, and also the count of the number of times the specified event does not occur: e.g., rolling a six-sided die and counting the number of ones and counting the number of times a one does not occur. These two counts completely specify the event space.

The usual linear regression methods which assume constant variance, normal error structures, and complete event space, are not appropriate for our count data for three main reasons: (1) the linear model might lead to the prediction of negative counts; (2) the variance of the response variable is likely to increase with the mean; and (3) the model errors will not be normally distributed.

To determine the probability of occurrence of a specified **I** and **C** combination, consider that the terms **I** and **C** are bivariate random variables described by a joint distribution. This joint distribution determines both the marginal and conditional distributions of **I** and **C**. The major thrust of the following analysis is characterizing these distributions.

Transmitter	Modulation	I	C	SMI
CommSat	CW	0	1	1
CommSat	voice	0	2	2
CommSat	video	0	3	3
CommSat	data	0	4	4
UHF TV	CW	4	1	5
UHF TV	voice	3	2	5
UHF TV	video	1	3	4
UHF TV	data	0	4	4
23 cm EME	CW	3	1	4
23 cm EME	voice	2	2	4
23 cm EME	video	1	3	4
23 cm EME	data	0	4	4
Goldstone DSN	CW	5	1	6
Goldstone DSN	voice	4	2	6
Goldstone DSN	video	4	3	7
Goldstone DSN	data	3	4	7
Evpatoria Radar	CW	5	1	6
Evpatoria Radar	voice	4	2	6
Evpatoria Radar	video	3	3	6
Evpatoria Radar	data	2	4	6
Arecibo Radar	CW	5	1	6
Arecibo Radar	voice	5	2	7
Arecibo Radar	video	4	3	7
Arecibo Radar	data	3	4	7

Figure 1
Sample of possible interstellar transmissions from Earth

Intensity **I** has 6 levels (0-5) and Content **C** has 5 levels (1-5) which gives 30 possible combinations of **I** and **C** classifications. Both the **I** and **C** scales have an ordered structure as, though the difference of any two adjacent levels differ by one, this difference may not scale linearly throughout the respective **I** and **C** ranges. Further, the message content factor **C** may be better represented as a nominal scale, in which the level differences have no significance beyond indicating a particular level is somehow greater or lessor than another. We perform a test of the viability of Content represented as ordinal versus nominal in the Parametric Analysis section below.

A randomly chosen combination from a population of combinations has a probability distribution. If we represent the **I**x**C** combinations in a table with 6 rows of intensity categories and 5 columns of message content categories, we have a contingency table with the frequencies of the randomly chosen **I**x**C** combinations contained in each of the table cells. The probability of occurrence of each **I**x**C** combination can be formed from the marginal sums, and is the joint probability of the **I**x**C** combinations. Also, we can form the conditional probabilities, thereby gaining an understanding of how, say, the probabilities of **C** change as the level of **I** changes.

To determine the independence of **I** and **C** we equate the cell probabilities with the product of the cell's marginal (the cell row and the cell column) probabilities. If the equality holds, then **I** and **C** are independent. Thus, if our sample is an unbiased sample of combinations of intensity and message content, we may determine the independence of **I** and **C**. Once independence (or lack thereof) has been established, we can ask several other questions such as, for our SMI situation, given a message is rated benign, what is the probability of it having a low intensity as well, or the converse, given a case has a high intensity, what is probability it has a compromising message?

For a 6x5 table, it is not typically possible to summarize measures of association with a single number (e.g., correlation) without loss of information. While single such numbers can represent specific features of association, e.g., the Pearson correlation is a measure of the monotonicity between **I** and **C**, it can be used only if our data are ordinal. We investigate the viability of correlation measures in the Parametric Analysis section below.

Ordinal scales have an ordered structure with the potential for factor association and interaction, including monotone trends. We now propose the use of log-linear models using the expected frequencies **I**x**C** rather than the cell probabilities so we can use Poisson-distributed errors, which are often appropriate for frequency. Log-linear models allow us to recover, in addition to the presence of trending, cell weights (i.e., **I**x**C** weights) in the form of probabilities, odds ratios for adjacent cells, and tests for independence. Log-linear modeling is data-based, hence we must use a completely random sample of **I** and **C** combinations. Although it is unlikely that our sample is random, more exhaustive sampling is not currently tractable, so we make the assumption our data are from a random sample.

As stated previously, there are 30 combinations of **I** and **C** possible from the 6 levels of **I** and the five levels of **C**. This means that there are 30 ways these combinations can sum **I** and **C** to 10: there is one way to get a sum of 1 and a sum of 10, there are two ways to sum to 2 and 9, 3 ways to sum to 3 and 8, 4 ways to sum to 4 and 7,

and 5 ways to sum to 5 and 6. Thus each possible combination of the sums must be weighted differently - the weighting for, or probability of, obtaining a sum of 5 when **I**=4 and **C**=1 versus when **I**=1 and **C**=4 are quite likely very different. Therefore, the opportunities for finding specific, single weightings of **I** and **C** (such as through a regression relationship) are limited, whereas weighting on the **I**x**C** cells occurs naturally in contingency tables. Thus, another reason for adopting a log-linear modeling approach is it allows for assigning individual cell weights.

The log-linear model is one of the specialized cases of general linear models, and is often used with Poisson-distributed (counts) data. Our assumed randomly selected cases of combinations of **I** and **C** are counts data and hence our choice of a log-linear model. Now, for two-way tables, an independence (no interaction between **I** and **C**) model rarely gives an adequate fit. As we have more than two levels of **I** and **C**, we won't have a saturated model (i.e., we have fewer parameters to estimate than we have cell combinations) leaving us sufficient degrees of freedom for parameter estimation. The form of the log-linear model we use is (Agresti, 1990):

$$\ln\left(e_{ic}\right)=\mu+\alpha_i+\beta_c+\gamma_{ic}, \quad i=0,1,\ldots,5, \quad c=1,2,\ldots,5 \quad (1)$$

Where $\ln(e_{ij})$ is the naperian log of the expected frequencies of the **I** and **C** cell combinations, μ is the overall mean of the log of the expected frequencies, α, β, and γ are the parameters to be estimated for **I** , **C**, and **I×C** , respectively, and i and j are the category levels for **I** and **C**.

From Equation (1) we have four parameters to estimate. The first step to building a log-linear model is to calculate the expected cell frequencies from the cell data resulting from a sample, as shown in Table 1. The basic calculation for each expected cell count e_{ic} from each respective sample (observed) count o_{ic} is the row sum times the column sum divided by the grand sum:

$$e_{ic}=\frac{o_{i.}\times o_{.c}}{o_{..}}, \quad i=1,\ldots,6, \quad c=1,\ldots,5 \quad (2)$$

where

$$o_{i.}=\sum_{c=1}^{5} o_{ic}, \quad i=1,\ldots,6,$$

$$o_{.c}=\sum_{i=1}^{6} o_{ic}, \quad c=1,\ldots,5, \text{ and } \quad o_{..}=\sum_{i=1}^{6}\sum_{c=1}^{5} o_{ic}.$$

Data supplied by H. Paul Shuch (2008) populates a contingency table (Table 2) showing the observed counts and the cell expectation as calculated from (2) by cell along with the row and column marginal sums. Table 3 converts the counts and expectations data into, respectively, the observed proportions and their respective expectations. With the expected frequencies calculated from Equation (2), we now know the expected probabilities (weights) of each **I** and **C** combination from the proportions in Table 3, and we can estimate the parameters α, β, and γ. The parameters are related to the association measure known commonly as the odds ratio. Odds essentially are the ratios between the frequency of being in one category (or level) by the frequency of not being in that category. For example, the odds of being in **I** = 4 versus **I** ≠4 is

$$\frac{o_{5.}}{o_{..}-o_{5.}}.$$

The odds ratio is the cell count of one or more levels by the cell count of one or more other levels. For example, the odds ratio of a

message being in level 2 rather than level 1 given the intensity is 4 is

$$\frac{o_{52}}{o_{51}},$$

and the odds ratio of a message being in level 2 given the intensity level is 4 is

$$\frac{o_{52}}{o_{.2}-o_{52}}.$$

DESCRIPTIVE ANALYSIS

Before determining the parameters of the log-linear model in Equation (1), we first explore the behavior of the sample data (size 24) used to estimate μ, α, β, and γ. We begin by examining the statistical properties of SMI. Figure 2 is a histogram of SMI overlaid with a kernel density plot (Venables, 2002). We see that these data have a skewed probability distribution, which is also indicated by the Box plot in Figure 3. The statistical test for identifying a normal distribution is the Shapiro-Wilk test (Shapiro and Wilk, 1965) which indicates that SMI is roughly normal (W = 0.9185, p-value = 0.03185, we assume a W of 0.95 or higher with a p-value of 0.05 or less constitutes adequate acceptance that SMI is normal for sample sizes of from 3 to 2,000), which is contrary to the graphical representations.

We suspect that the absence of **C**=5 data cause the skewed appearance of SMI. The SMI median is 6, the interquartile range (iqr) is 2.24, the mean is 5.21, and the standard deviation is 1.69. Additional data may help resolve this apparent discrepancy.

Next we examine the properties of Intensity **I** and Content **C**. Figures 4 and 5 show the respective histograms. Notice that neither is remotely normal, which is expected of, particularly, small samples of counts data. These distributions appear roughly uniform.

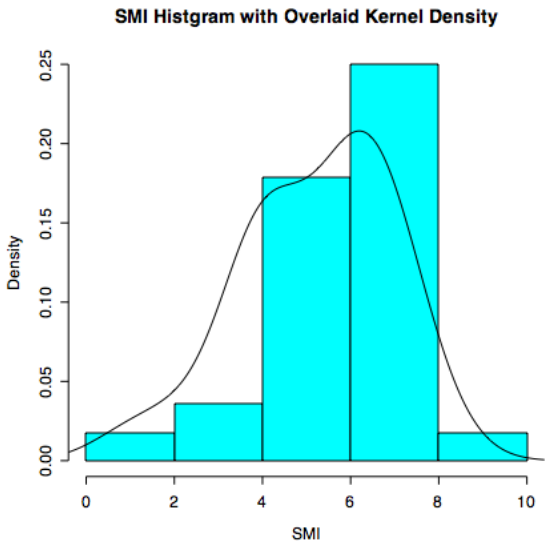


Figure 2
A histogram of SMI including a kernel distribution overlay. This plot leads us to suspect the sample is not normally distributed.

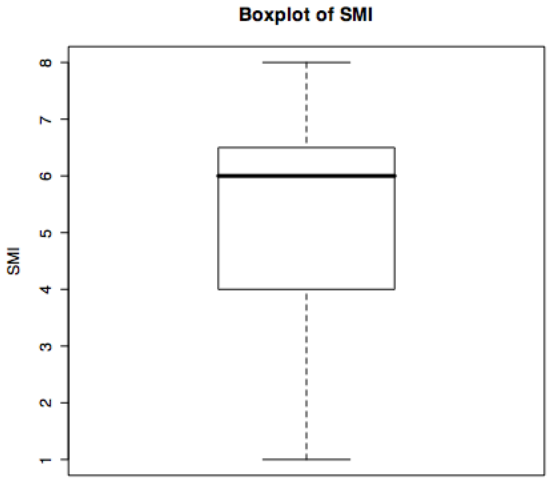


Figure 3
Box plot of SMI showing a median of 6, iqr of 2.25, mean of 5.21, and standard deviation of 1.69.

It is useful to examine the coverage of SMI versus **I** and **C**. We see in Figure 6 that as **I** increases, the range of coverage of SMI by each level of **I** decreases even though the SMI median increases. The range (interquartile) coverage of SMI by **C** is nearly constant, though the SMI median values at each level of **C** vary. These characteristics of **I** and **C** contribute to making the usual modeling methods such as regression inadequate for determining the properties of the **I**x**C** interaction.

To complete our examination of the Intensity and Content sample data, we now see how the frequencies of **I** and **C** populate the 6x5 contingency table. We do this using an association plot and a mosaic plot. These plots are shown in Figures 7 and 8. The association plot in Figure 7 shows the departures from expectations of the observed frequencies in the contingency table. Note that we must subtract one from the levels of Intensity **I** in the plot to obtain the correct sample **I** value. The black bars rising above any specified level **I** show the excess of counts over the expected counts for a particular **I**x**C** cell. The largest excesses occur for cells (**I**=0)x(**C**=4), (**I**=1)x(**C**=3), and (**I**=5)x(**C**=1). Although the smaller than expected departures (red bars dropping below any specified level of **I**) are less significant than the excesses, we observe that the two largest are (**I**=4)x(**C**=4) and (**I**=5)x(**C**=4).

The mosaic plot shows that there are significantly more (**I**=5)x(**C**=1) (blue rectangle) counts than expected, which is shown in the association plot as well. The mosaic displays the standardized residuals of the log-linear model of the counts by the color and outline of the mosaic's tiles.

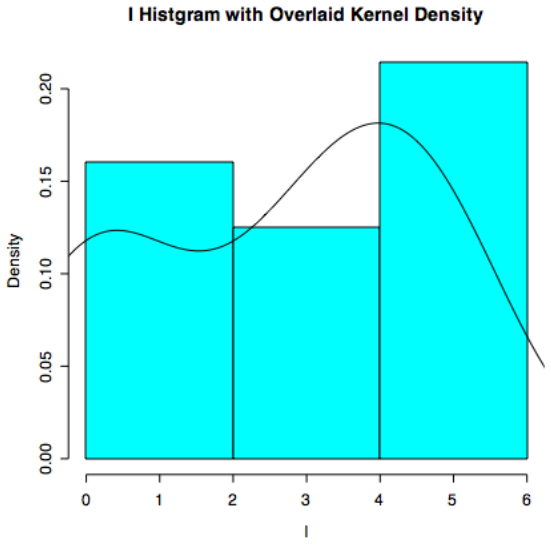


Figure 4
A histogram of Signal Intensity I including a kernel distribution overlay. This implies a uniform distribution.

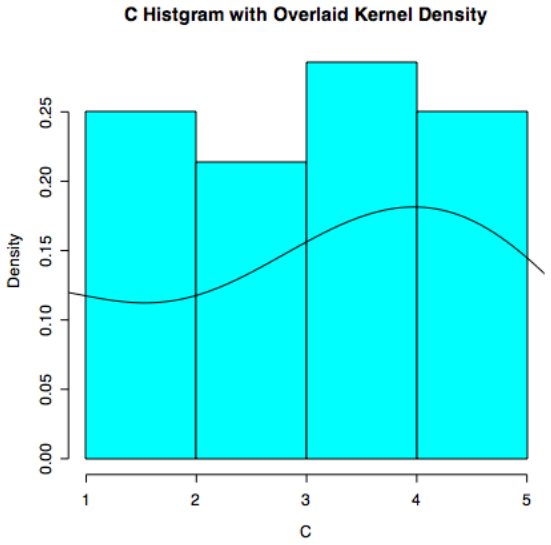
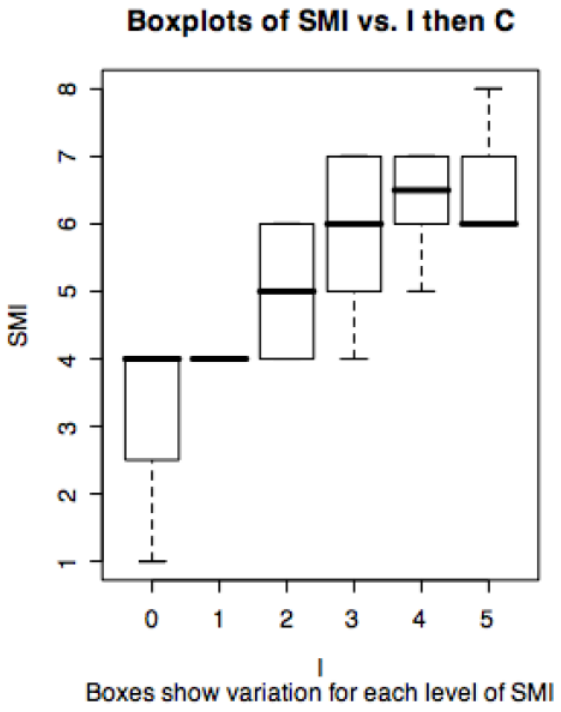


Figure 5
A histogram of Message Content C including a kernel distribution overlay. This implies a uniform distribution.



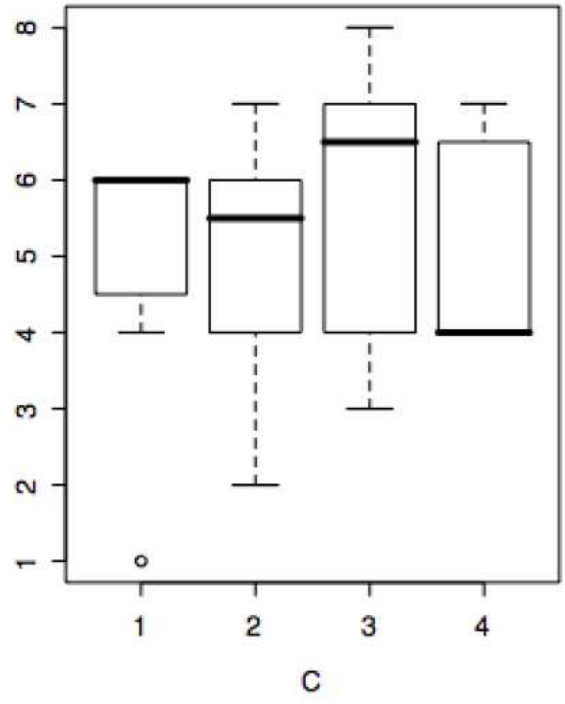


Figure 6

Box plots of SMI by first Intensity I and then Content C. Notice the range change through the levels of I over increasing SMI versus the small range change of SMI by C. The medians of SMI by C fluctuate without apparent pattern.

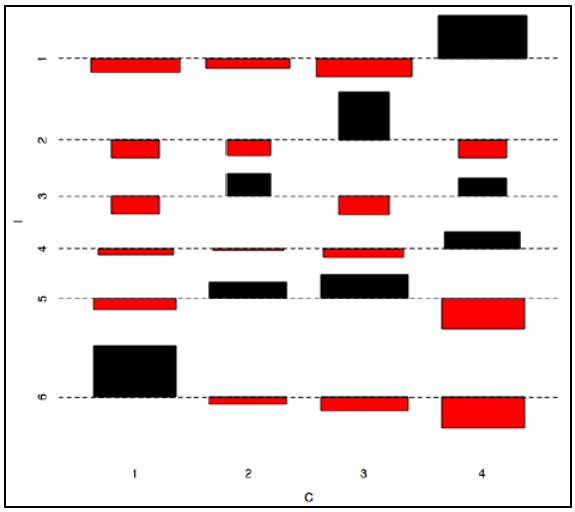


Figure 7

The association plot indicates deviations from the expected counts for each I×C cell. Red bars dropping below an I level indicate less than the expected counts, and the black bars rising above an I level indicate counts that exceed the expected numbers. Note: subtract one from the levels of I in the plot to obtain the correct sample I values.

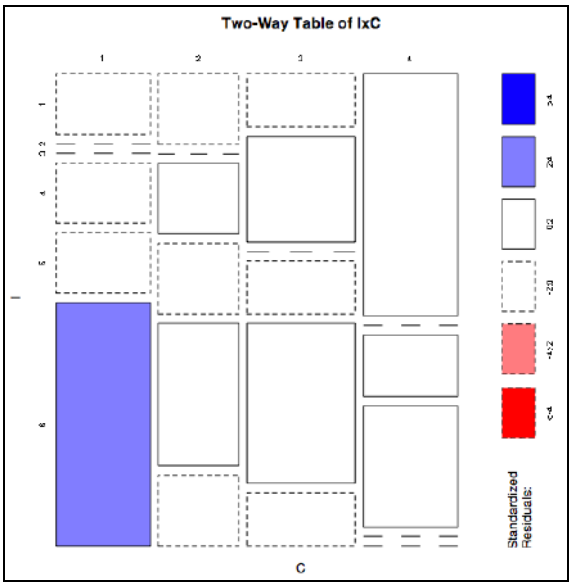


Figure 8

The mosaic plot shows that there are significantly more (I=5)×(C=1) (blue rectangle) counts than expected. The mosaic tiles display the standardized residuals of the log-linear model of the counts by the color and outline of the mosaic tiles. Note: subtract one from the levels of I in the plot to obtain correct sample I values.

This mosaic plot shows that we can expect a good fit to the log-linear model introduced above, the details of which we explore following a discussion of sample size. Recall that we must subtract one from the levels of Intensity **I** in the plot to obtain the correct sample **I** value.

SAMPLE SIZE

A contingency table is considered sparse when many cells have small frequencies. This is certainly the case with the SMI **I**x**C** table. An index measuring sparseness is formed as the ratio of the sample size *n* to the number of cells *N* in the table as *n/N*=24/30=0.80. Small ratio values indicate sparse tables.

The size *n/N* that gives measures of adequate contingency table models tend to decrease as *N* increases. Koehler and Larntz (1980) suggest

$$\frac{n}{N} > \left(\frac{10}{N}\right)^{1/2} \Rightarrow n > (10N)^{1/2} = (10 \cdot 30)^{1/2} = 17.32 \approx 18.$$

Clearly, our sample size of 24 satisfies this minimum number of **I** and **C** counts. However, as we have a sparse table - i.e., the counts data are not distributed across all combinations of **I** and **C** - this estimation may be compromised.

Garwood (1936) suggested a 100(1−α)% confidence interval for the parameter θ of a Poisson distribution (which is how the counts data are distributed, as we shall see in the Parametric Analysis section) as:

$$R(\alpha, \omega; X) = \left[\frac{\chi^2_{2df;\omega_L}}{2n}, \frac{\chi^2_{2(df+1);\omega_U}}{2n}\right]$$

where *df* is the degrees of freedom derived from the number of rows *(r)* and columns *(s)* of the contingency table as:

df=(r-1)(s-1)=(6-1)(5-1)=20,

ω is a shape parameter for the confidence interval such that

$\omega_U - \omega_L = \alpha, \omega_U, \omega_L \in [0, 1]$,

α is the desired confidence level such that $\alpha \in (0, 1)$,

and $\chi^2_{2df;\omega}$ denotes a chi-square distribution with 2*df* degrees of freedom and shape ω. The confidence interval for the Poisson parameter θ is dependent on the contingency table size, the desired confidence level, and the associated quantiles from a χ^2 distribution. We use this information to determine a minimum sample size.

First, to be 100(1−α) sure that the error does not exceed the amount *d* (using the normal distribution approximation for large sample sizes) of 100(1-α) we have

$$d = Z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n_t}} = 1.96\frac{(1.45/2)}{\sqrt{20}} = 0.3177, \quad \alpha = 0.05,$$

where the estimated standard deviation σ is from the residual deviance derived in the Parametric Analysis section and n_t is the *df* of the contingency table. This, then, gives an expression for the sample size *n* (using $d = R[\alpha, \omega, \chi]$)

$$n = \frac{\chi^2_{2(30);0.025}}{4d} = \frac{59.34}{(4)(0.3177)} = 46.69 \approx 47.$$

To be 95% sure that our contingency table observation population will give a sufficient log-linear model, we need a sample size of at least 47 distributed across all contingency table cells. Our non-random sample size of 24 thus gives us an at-best confidence level of:

$$\chi^2_{2(20);\alpha/2} = 4\,dn = 4(0.3177)(24) = 30.49$$
$$\Rightarrow \alpha = 0.28,$$

which gives us a confidence level of 72.22%. It is desirable, then, to re-run all our analyses with a sample size of at least twice the sample size used for modeling in the Parametric Analysis section below, and the cases must be randomly selected.

PARAMETRIC ANALYSIS

In this section, we address the specifics of the parameter estimation of the log-linear model that we have suggested will best represent the frequencies of the **I**x**C** cell combinations. Two assumptions about log-linear models need testing relative to our data set prior to evaluating the significance and weight of the model parameters. The two assumptions are the viability of using ordinal scales for **I** and **C**, and whether the errors are Poisson-distributed. As it turns out, we can check both these assumptions using the same model-building process.

To assess the fitness of the Poisson error structure for the ordinal scale log-linear model, we first fit the independence model (i.e., Equation (1)) without the γ parameter. From the output of the log-linear model fit without the interaction (Table 4), we see that the residual deviance of 28.123 on 21 degrees of freedom approaches the ideal of equal residual deviance to the number of degrees of freedom, indicating that the Poisson error model is quite promising. We see also, that when we construct an ordinal-by-nominal model (**I** ordinal and **C** nominal, Table 5), which assumes that Content levels are without the relationship of linear level-to-level differences, and then compare these two models, we see the ordinal-by-ordinal model is superior. The Analysis of Variance (ANOVA) comparing the ordinal-by-ordinal model with the ordinal-by-nominal model shows (Table 6) that these two models are not significantly different (Deviance = 0.2581, Prob(>|χ|=0.0789), but we note comparing the Akaike Information Criterion (AIC, Akaike, 1969) of the ordinal-by-ordinal independence model versus the ordinal-by-nominal independence model, 73.487 versus 77.229, has the ordinal-by-ordinal model superior (smaller AIC is better). Also, we note that the ordinal-by-nominal model residual deviance of 27.865 on 19 degrees of freedom still leads us to support a Poisson error distribution model.

Next we test the interaction term in the ordinal-by-ordinal model. Table 4 shows that output of the independence model. We see that neither **I** nor **C** make significant contribution (neither **I** nor **C** has Prob(>|z|)$\leq$0.05) to the frequencies of the contingency table cells. This alone has us question the viability of the independence model. We construct the Equation (1) model whose output is in Table 7, then use ANOVA to compare the independence model with the interactive model. Indeed, the ANOVA output (Table 8) gives a Deviance of 8.4570 with Prob(>|χ|)=0.0036 indicates the interaction model is superior to the independence model. In addition, the AIC of the interaction model versus the independence model of 67.03 versus 73.487 shows that the interaction model is superior.

Two other model characteristics recommend the interaction model. The first is that the Poisson error structure is sufficiently supported with a residual deviance of 19.666 on 20 degrees of freedom – the equality is clear with this model. The second is that all three parameters that are estimated from the data are statistically significant, with the Prob(>|z|)$\leq$0.00895 of the **C** term being the largest of the three.

Therefore, we have that the log-linear interaction model constructed from

the Shuch data is superior to all other models considered.

SMI LEVEL WEIGHTING

Using either the log-linear model or the contingency table probabilities, weights can be assigned to any SMI risk level.

SMI	Weight (%)
1	6
2	7
3	10
4	13
5	13
6	16
7	14
8	11
9	5
10	0

Due to rounding error in the contingency table probabilities, the total weight is 95% and not the expected 100%.

Interestingly, we would expect the lowest SMI levels to have the highest probabilities of occurrence, as it is suspected that there are far more broadcasts of low intensity, innocuous message content than is present in the frequencies in Table 2. The weightings above are suspected of being the result of non-random case selection. This situation can be corrected in two ways:

(1) use a random sample, or

(2) use expert opinion to establish the expected values for each **I×C** cell.

CONCLUSIONS

We set out to provide the added dimension of SMI level weight (probability) to the SMI risk scale. We began by examining the statistical properties of SMI, Intensity, and message Content. These properties gave us insight into the distributions of our sample, across their respective levels. We found that Intensity and Content have roughly uniform distributions, and SMI has a roughly normal distribution. From a purely combinatorics position, SMI will be nearly normal as the center levels of 5 and 6 have five ways to arrive at these sums, with decreasing counts for sums less than 5 and greater than 6.

We showed that our sample size of 24 is half what is needed to achieve a 95% confidence level in our modeling outcomes. The confidence level associated with the sample size of 24 is approximately 72%. This means that we could arrive at the same log-linear model just by chance 28% of the time.

The contingency table populated with our sample observations provided individual cell (Intensity by Content) weights (probabilities). These weights allow us to characterize any SMI level with a probability of occurrence.

Finally, the log-linear model generated from the sample showed the viability of an interactive, linear model as given in Equation (1). The sufficiency of the log-linear model allows us to predict the weights of any SMI level, even those that are missing or marginally represented in the sample.

We have shown that SMI can be enhanced by the addition of weighting the levels of SMI. The enhancement can be combined with the SMI risk for form a tuple of risk and probability of occurrence. Thus, we would specify SMI as SMI(6, 5%) indicating a risk of 6 with a 5% chance of occurring.

Despite our preference for a log-linear model, we hesitate to recommend any immediate changes to the San Marino Scale, as currently adopted by the SETI Permanent Study Group of the International Academy of Astronautics. Our complete analysis is based upon a convenience sample of

hypothetical transmissions from Earth, the size of which we have shown to be inadequate for high levels of confidence.

While we would desire a larger sample of historical transmissions with which to further enhance our model, the fact is that deliberate transmissions from Earth to extraterrestrial intelligence are exceedingly rare events. In fact, it can be argued, the number of facilities on this planet currently engaged in significant METI (Messaging to Extra-Terrestrial Intelligence) experiments can be counted on the thumbs of one hand.

REFERENCES

Agresti, A., (1990), *Categorical Data Analysis*, John Wiley & Sons, Inc., New York.

Akaike, H., (1969), "Fitting Autoregressive Model for Prediction," *Annuals of the institute of Statistical Mathematics*, Tokyo, 21, 243-247.

Almár, Iván (2005), "Quantifying Consequences Through Scales" paper presented at the 6th World Symposium on the Exploration of Space and Life in the Universe, Republic of San Marino, March.

Almár, Iván, and H. Paul Shuch (2007a), "The San Marino Scale: A New Analytical Tool for Assessing Transmission Risk," Academy Transaction Note, *Acta Astronautica* 60(1): 57-59, January.

Garwood, F., (1936), "Fiducial Limits for the Poisson Distribution," *Biometrika*, 28, 437-442.

Koehler, K. and Larntz, K., (1980), "An Empirical Investigation of Goodness-of-Fit Statistics for Sparse Multinomials," *J. Amer. Statist. Assoc.*, 75, 336-344.

Shannon, Claude and Warren Weaver (1949), *The Mathematical Theory of Communication*. University of Illinois Press.

Shapiro, S. and Wilk, M., (1965). "An analysis of variance test for normality (complete samples)," Biometrika, 52, 3 and 4, 591-611.

Shuch, H. Paul and Iván Almár (2006), "Updating The San Marino Scale," Paper IAC06-A4.1.01, 57th International Astronautical Congress, Valencia, Spain, October.

Shuch, H. Paul and Iván Almár (2007b), "Quantifying Past Transmissions Using the San Marino Scale," Paper IAC07-A4.2.04, 58th International Astronautical Congress, Hyderabad, India, September.

Shuch, H. Paul and Iván Almár (2007c), "Shouting in the jungle: the SETI transmission debate." *Journal of the British Interplanetary Society* 60(4):142-146, April.

Venables, W.N., and Ripley, B.D., (2002), *Modern Applied Statistics with S*, 4th ed., New York: Springer Science + Business Media, Inc.

ACKNOWLEDGMENTS

The authors are indebted to:

- Prof. Iván Almár, for conceiving the San Marino Scale
- Prof. Allen Tough, for supporting this research
- Prof. Doug Vakoch and Dr. Jill Tarter, for pointing out weaknesses in the San Marino Scale as originally proposed, and suggesting ways to strengthen it.

DEDICATION

This paper is fondly dedicated to the memory of Prof. Ward D. Edwards (1928 - 2005), Bayesian extraordinaire.

TABLES

Table 1: Observed o_{ij} and expected e_{ij} cell values with row sums $o_{i.}$ and $e_{i.}$, column sums $o_{.j}$ and $e_{.j}$, and the grand sums $o_{..}$ and $e_{..}$.

	Message Level (C)					
Intensity (I)	**1**	**2**	**3**	**4**	**5**	**Column Sums**
0	o_{11}	o_{12}	o_{13}	o_{14}	o_{15}	$o_{1.}$
	e_{11}	e_{12}	e_{13}	e_{14}	e_{15}	$e_{1.}$
1	o_{21}	o_{22}	o_{23}	o_{24}	o_{25}	$o_{2.}$
	e_{21}	e_{22}	e_{23}	e_{24}	e_{25}	$e_{2.}$
2	o_{31}	o_{32}	o_{33}	o_{34}	o_{35}	$o_{3.}$
	e_{31}	e_{32}	e_{33}	e_{34}	e_{35}	$e_{3.}$
3	o_{41}	o_{42}	o_{43}	o_{44}	o_{45}	$o_{4.}$
	e_{41}	e_{42}	e_{43}	e_{44}	e_{45}	$e_{4.}$
4	o_{51}	o_{52}	o_{53}	o_{54}	o_{55}	$o_{5.}$
	e_{51}	e_{52}	e_{53}	e_{54}	e_{55}	$e_{5.}$
5	o_{61}	o_{62}	o_{63}	o_{64}	o_{65}	$o_{6.}$
	e_{61}	e_{62}	e_{63}	e_{64}	e_{65}	$e_{6.}$
Row Sums	$o_{.1}$	$o_{.2}$	$o_{.3}$	$o_{.4}$	$o_{.5}$	$o_{..}$
	$e_{.1}$	$e_{.2}$	$e_{.3}$	$e_{.4}$	$e_{.5}$	$e_{..}$

Table 2: Observed counts and expected values (calculated from Equation 2) per cell with observed and expected row and column sums, and the grand sums for both the observed counts and the expected values.

	Message Level (C)					
Intensity (I)	**1**	**2**	**3**	**4**	**5**	**Column Sums**
0	1	1	1	4	0	7
	1.75	1.50	2.00	1.75	0.00	7.00
1	0	0	2	0	0	2
	0.50	0.43	0.57	0.50	0.00	2.00
2	0	1	0	1	0	2
	0.50	0.43	0.57	0.50	0.00	2.00
3	1	1	1	2	0	5
	1.25	1.07	1.43	1.25	0.00	5.00
4	1	2	3	0	0	6
	1.50	1.29	1.71	1.50	0.00	6.00
5	4	1	1	0	0	6
	1.50	1.29	1.71	1.50	0.00	6.00
Row Sums	7	6	8	7	0	28
	7.00	7.30	7.99	7.00	0.00	29.29

Table 3: Proportions of observed counts and expected values per cell with observed and expected row and column proportion sums, and the grand proportion sums for both the observed counts and the expected values.

	Message Level (C)					
Intensity (I)	**1**	**2**	**3**	**4**	**5**	**Column Sums**
0	0.04	0.04	0.04	0.4	0	0.25
	0.06	0.05	0.07	0.06	0.00	0.25
1	0	0	0.07	0	0	0.07
	0.02	0.01	0.02	0.02	0.00	0.07
2	0	0.04	0	0.04	0	0.07
	0.02	0.01	0.02	0.0.2	0.00	0.07
3	0.04	0.04	0.04	0.07	0	0.18
	0.04	0.04	0.05	0.04	0.00	0.17
4	0.04	0.07	0.11	0	0	0.21
	0.05	0.04	0.06	0.05	0.00	0.20
5	0.14	0.04	0.04	0	0	0.21
	0.05	0.04	0.06	0.05	0.00	0.20
Row Sums	0.25	0.21	0.29	0.25	0	1.00
	0.24	0.25	0.29	0.25	0.00	1.00

Table 4: Ordinal-by-ordinal independence model.

```
glm(formula = Freq ~ I + C, family = poisson, data
= frame.IC)

Deviance Residuals:

    Min       1Q   Median       3Q      Max

-1.6800  -1.4287  -0.1670   0.1891   2.2059

Coefficients:

            Estimate Std. Error z value Pr(>|z|)

(Intercept) -0.13807    0.62161  -0.222    0.824

I            0.06137    0.11104   0.553    0.581

C            0.02858    0.16909   0.169    0.866

(Dispersion parameter for poisson family taken to
be 1)

Null deviance: 28.458  on 23  degrees of freedom

Residual deviance: 28.123  on 21  degrees of
freedom

AIC: 73.487

Number of Fisher Scoring iterations: 5
```

Table 5: Ordinal-by-nominal independence model.

```
glm(formula = Freq ~ I + C, family = poisson, data
= frame.IC)

Deviance Residuals:

    Min       1Q   Median       3Q      Max

-1.6448  -1.3739  -0.1416   0.2802   2.2625

Coefficients:

              Estimate Std. Error z value Pr(>|z|)

(Intercept) -6.612e-02  5.565e-01  -0.119    0.905

I            6.137e-02  1.110e-01   0.553    0.581

C2          -1.542e-01  5.563e-01  -0.277    0.782

C3           1.335e-01  5.175e-01   0.258    0.796

C4           1.039e-15  5.345e-01 1.94e-15   1.000

(Dispersion parameter for poisson family taken to
be 1)

Null deviance: 28.458  on 23  degrees of freedom

Residual deviance: 27.865  on 19  degrees of
freedom

AIC: 77.229

Number of Fisher Scoring iterations: 6
```

Table 6: Ordinal-by-ordinal independence model and ordinal-by-nominal independence model ANOVA.

```
Model 1: Freq ~ I + C

Model 2: Freq ~ I + C

  Resid. Df Resid. Dev Df Deviance P(>|Chi|)

1        21    28.1227

2        19    27.8645  2   0.2581    0.8789
```

Table 7: Ordinal-by-ordinal interaction model.

```
glm(formula = Freq ~ I * C, family = poisson, data
= frame.IC)

Deviance Residuals:

     Min        1Q    Median        3Q       Max

-1.96365  -0.90847  -0.08875   0.54262   1.67707

Coefficients:

            Estimate Std. Error z value Pr(>|z|)

(Intercept)  -3.0897     1.3821 -2.236  0.02538 *

I             0.8146     0.2936  2.774  0.00554 **

C             1.1127     0.4257  2.614  0.00895 **

I:C          -0.2917     0.1014 -2.877  0.00401 **

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to
be 1)

Null deviance: 28.458  on 23  degrees of freedom

Residual deviance: 19.666  on 20  degrees of
freedom

AIC: 67.03

Number of Fisher Scoring iterations: 5
```

Table 8: Ordinal-by-ordinal independence model and ordinal-by-ordinal interaction model ANOVA.

```
Model 1: Freq ~ I + C

Model 2: Freq ~ I * C

  Resid. Df Resid. Dev Df Deviance P(>|Chi|)

1        21    28.1227

2        20    19.6657  1   8.4570    0.0036
```