An Entropy Estimator of Population Variability in Nominal Data

Mickie Vanhoy
University of Central Oklahoma

Abstract

Entropy is an established measure of variability in nominal data. The present paper addresses the problem of directly estimating population entropy from an empirical sample. Thirty artificial, nominal, population distributions were subjected to Monte Carlo analysis. Comparison of sample entropy values to the known population entropy values showed that entropy is a consistent measure of nominal variability. Raw sample entropy is a biased estimator that underestimates the population value. This bias was virtually eliminated through bootstrap resampling from the samples. Bootstrap corrected sample entropy is a sufficient, consistent, minimally biased, population estimator of nominal variability that can be used in further statistical analyses.

Imagine that a school psychologist was asked to identify an intervention for a student who was making unsatisfactory progress in reading. Imagine that the student over-relies on basic spelling-sound relationships to pronounce words with the result that the student reads irregular English words (words that are exceptions to the basic spelling-to-pronunciation rules) according to basic rules while neglecting specific exceptions to the rules. For example, the student attempts to pronounce HAVE but produces a long "A" sound instead— the pronunciation rhymes with CAVE. That mis-pronunciation is actually a more reasonable phonological production, sharing the vowel sound with many other words (e.g., GAVE, SAVE, PAVE, WAVE, etc.). The student does seem somewhat sensitive to the regularities in English and has just over-generalized the silent-E rule in this case. Consider a different student who, in response to HAVE, produced a pronunciation that rhymes with SAVVY. That student's pronunciation seems less constrained by the regularities in the language than the first student's does. Perhaps students who generate pronunciations that reflect the statistical regularity of the language are more sensitive to spelling-sound mappings than students who generate pronunciations that do not reflect the pronunciation variability in the language. If so, then the two groups of students may benefit from different interventions. Resolving this question and creating a method of ensuring intervention fidelity requires comparing the pronunciation variability produced by the student (sample) to the pronunciation variability in the language (population). Unfortunately, there is currently no population estimator of the variability in such nominal data.

The present concern is with situations where such estimates are of primary importance. For example, some English words have spelling "bodies" like _EAD as in BEAD, READ, LEAD, HEAD, etc. and _OUGH as in DOUGH, COUGH, TOUGH, etc., that are variable in their pronunciations. The distribution of the pronunciations is a nominal variable. A school psychologist may want to know whether the sample distribution of the pronunciations produced by students with identified reading problems reflects (estimates) the variability in the population so that she can devise relevant reading intervention strategies. Perhaps students with reading problems come from different population distributions. Resolving this question requires a comparison of the nominal variability in the population to the nominal variability in the sample. What method allows that?

What is variability?

If a sample distribution contains three possible responses on a ratio scale, the usual practice is to use the corrected sample variance as an estimate of the population variance. In Figure l (top), responses 1 and 3 each occur once and response 2 occurs twice. In Figure l (top), the variance is .5. Yet what is the variance when the dependent variable lies on a *nominal* scale, as in Figure l (bottom)? In this case, order and even distance are meaningless.

Nominal distributions do display variability, however. Consider the example in Figure 2. English words with the spelling body _OBE are always pronounced to rhyme with ROBE and never to rhyme

A National Institutes of Health National Research Service Award (5 F32 HD08076-02) to the author supported this work. The author warmly thanks Greg Stone for his inspiration and patient mentoring on this project. Thanks also to Tom Shepherd, Mary Sweet-Darter, and two anonymous reviewers for their expert input on previous versions of this manuscript. Correspondence regarding this manuscript may be directed to Mickie Vanhoy, PhD, Department of Psychology, University of Central Oklahoma, Edmond OK 73118, USA.
mvanhoy@ucok.edu

with ROB (Figure 2, top) but words with the spelling body _EAD sometimes rhyme with BEAD and sometimes rhyme with HEAD (Figure 2, bottom). A population containing a single nominal category (Figure 2, top) is clearly less variable than a distribution with two equally probable categories (Figure 2, bottom), but traditional techniques do not estimate this population variability. Measures of variability in nominal data, like the index of diversity (or its standardized version, the index of qualitative variation), do not have the same statistical usefulness as the variance of quantitative data or the entropy of nominal data (Magidson, 1981; Reynolds, 1984). Furthermore, quantitative techniques often require assumptions about the shape of the population distribution (e.g., normality). The question is whether nominal variability is empirically estimable when the form of the population distribution is unknown.

Figure 1. A distribution of ratio data with a known variance (a) and a distribution of nominal data with unknown variability (b).

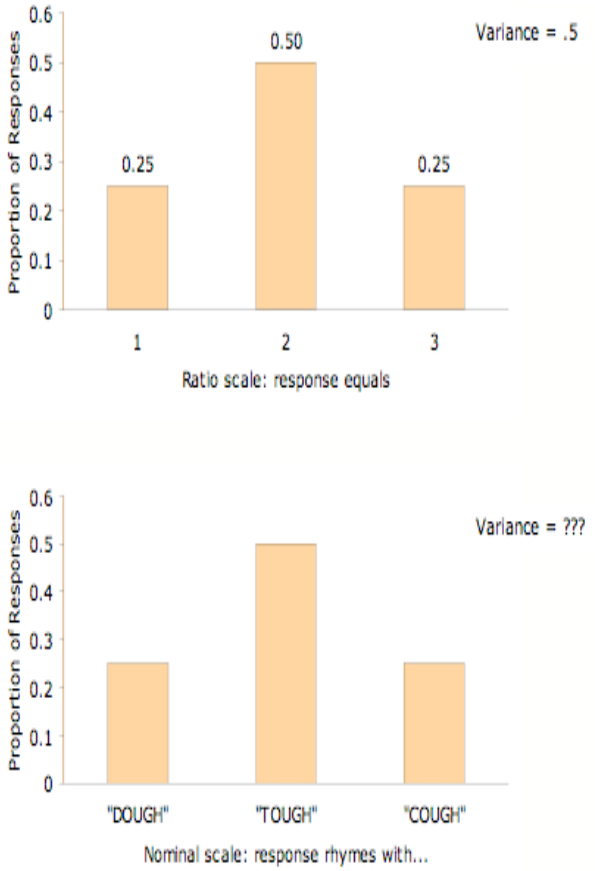
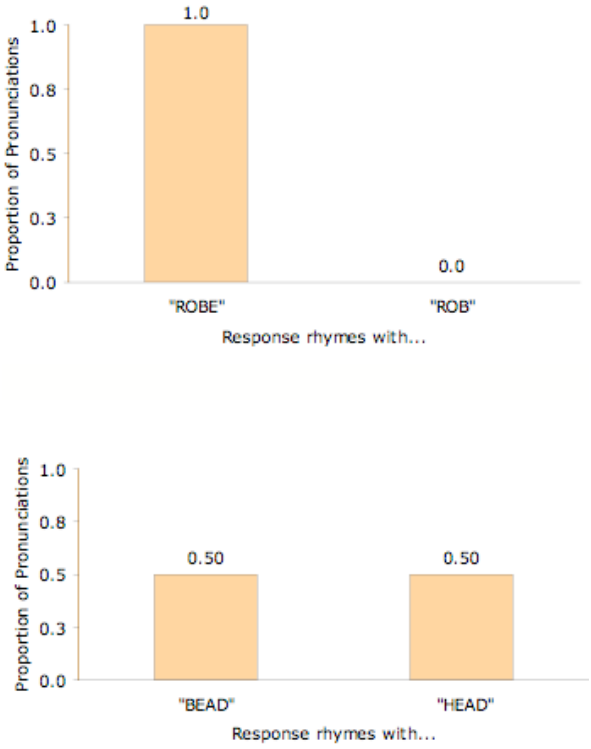


Figure 2. A nominal distribution with minimum entropy (a) and a nominal distribution with maximum entropy for three categories (b).



Entropy is a measure of uncertainty derived from information theory (e.g., Shannon, 1950). Entropy is also a measure of nominal population variability and it is the foundation for an entropy measure of association in contingency table analysis (Fano, 1961, pp. 21-61; Gokhale & Kullback, 1978; Kullback, 1968; Press, Teukolsky, Vetterling, & Flannery, 1992, p. 629). It is less familiar to most psychologists than chi-square but has similar utility in contingency table analysis. Entropy (H) equals the negative of the sum of the category probabilities times the logarithms of the category probabilities:

$$H = -\Sigma p \ln p. \quad (1)$$

For example, a single category distribution yields minimum entropy of 0:

$$H = +1 * \ln(1) = 0. \quad (2)$$

A sample of equi-probable categories yields the maximum entropy for three categories[1]:

$$H = -1/3 * \ln(1/3) - 1/3 * \ln(1/3) - 1/3 * \ln(1/3) \approx 1.0986. \quad (3)$$

Using entropy in contingency table analysis does not require investigation of sample entropy as an estimator of population entropy in univariate samples (Gokhale & Kullback, 1978). There is a need, however, for just such an estimator in many areas of psychology. Returning to the school psychology example, recall the observation that some English words have spelling "bodies" like _EAD (as in BEAD, READ, LEAD, HEAD, etc.) that are variable in their pronunciations and that students may vary in the extent to which their pronunciations reflect this variability. This is not the only research question in psychology that requires comparing the nominal variability in the population to the nominal variability in the sample.

Consider another example— the distribution of emotions in psychotherapy clients (e.g., anxious, comfortable, angry). Does a particular client's emotional landscape more closely resemble the population after a program of psychotherapy? That distribution too lies on a nominal scale. So does the distribution of bird song types (e.g., Brunton & Li, 2006) — does the behavior of the local sample reflect the behavior of the species? Having a readily available estimator of nominal variability would have supported the answer to that research question. Examples like this are too numerous to list. Nominal data may be the most plentiful kind of data in psychology and, as the emphasis on evidenced-based techniques grows (e.g., Fuchs & Fuchs, 2006; Hawley & Weisz, 2002), a population estimator of variability in nominal data becomes more important. Entropy is an established measure of variability in nominal data: is it also a suitable estimator? What are the desirable characteristics of a population estimator?

An estimator of a population parameter should ideally satisfy four criteria (Schumacker & Akers, 2001, p. 166). The primary concern is that the estimator be *consistent*. That is, as the sample size increases, the estimator converges on the population value. To be useful as a dependent variable for further statistical analysis, an estimator should be *sufficient*. That is, the estimator should reflect all the observed data. The standard variance measure for interval data is sufficient in that each observed value contributes to the sum of squared deviations. Inter-quartile range is not sufficient because values away from the quartile values do not affect the quantitative value of the estimator (the inter-quartile range provides advantages in special situations that outweigh lack of sufficiency). A population estimator should also be *efficient* relative to other estimators, that is, sample values should cluster closely around the population values.

Finally, an estimator should be *unbiased*[ii]. That is, the expected value of sample estimate should equal the population value. For example, standard variance is a biased estimator; it tends to underestimate the population variance. The expected value of the sample variance is:

$$E\left[\frac{\sum(x-\bar{x})^2}{N}\right] = \frac{(N-1)\sigma^2}{N} .$$

Multiplying by the correction factor N/N-1 corrects the expected value to σ^2. Entropy is a *de facto* sufficient estimator of nominal variability (see equation 1). Showing that it is also consistent and unbiased requires a known population distribution. Fortunately, Monte Carlo data simulation allows this analysis of samples from known populations.

Monte Carlo simulation

Imagine that a sample, $\mathbf{x_0}$, is drawn from a population and some statistic, $\mathbf{S_0}$, is calculated and used to estimate the population value, **P**. The sample, $\mathbf{x_0}$, is only one of many samples, $\mathbf{x_i}$, that could have been obtained. The set of samples, $\mathbf{x_i}$, of a given size that might have been drawn is the sampling distribution of the statistic, ($\mathbf{S_i}$ - **P**). A good approximation of ($\mathbf{S_i}$ - **P**) can often be obtained with less than an infinite number of samples (Hammersley & Handscomb, 1964; Kalos & Whitlock, 1985). Monte Carlo simulation can yield an empirical sampling distribution for any statistic that is a function of data independently and randomly drawn from a specified population. A Monte Carlo simulation begins by modeling an artificial population to obtain a parameter, **P**. A random number generator[iii] is then used to construct many synthetic data sets ($\mathbf{x_0}$, $\mathbf{x_1}$, $\mathbf{x_2}\cdots$ $\mathbf{x_i}$) for which the statistic of interest ($\mathbf{S_0}$, $\mathbf{S_1}$, $\mathbf{S_2}\cdots\mathbf{S_i}$) is calculated and compared to **P** (Noreen, 1989; Press et al., 1992, p.689). Some nonparametric statistics are examples of obtaining empirical sampling distributions. The only constraint on the population distribution is that the relative frequencies of any two different elements be specified (Noreen, 1989).

The simulations reported here included specifying artificial populations in terms of the probabilities of nominal categories, inputting the probabilities to the entropy equation to yield the population entropy values, and comparing those to values obtained by repeated sampling from the artificial populations. Because the population values were known, the sample values could then be assessed for consistency and bias.

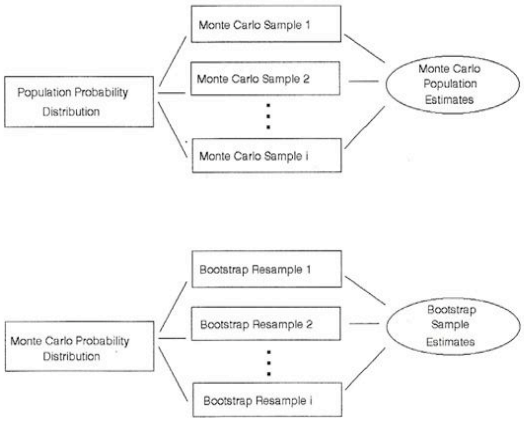
All simulations and data analyses were conducted on personal computers and entailed two steps. The first step was randomly sampling from thirty experimenter-constructed population distributions containing two, three, four, or five possible nominal responses. The populations represented a range of low to high nominal variability and thus, entropy. They were represented as sets of response category probabilities as shown in Table 1. The second step was to calculate the mean entropy of each sample and the "true" entropy of each population, as well as the squared difference between the sample and population values.

Table 1. The 30 population distributions and the probability of each category's occurrence. Maximum variability (equiprobability) is presented in bold red font.

Probability (%) of category's occurrence				
A	B	C	D	E
95	5			
90	10			
80	20			
70	30			
60	40			
50	**50**			
90	5	5		
80	10	10		
70	20	10		
60	30	10		
50	40	10		
40	30	30		
33.33	**33.33**	**33.33**		
85	5	5	5	
70	10	10	10	
60	20	10	10	
50	30	10	10	
40	40	10	10	
40	30	20	10	
40	30	15	15	
30	25	25	20	
25	**25**	**25**	**25**	
80	5	5	5	5
70	15	5	5	5
60	10	10	10	10
60	20	10	5	5
50	20	10	10	10
40	30	10	10	10
30	30	20	10	10
20	**20**	**20**	**20**	**20**

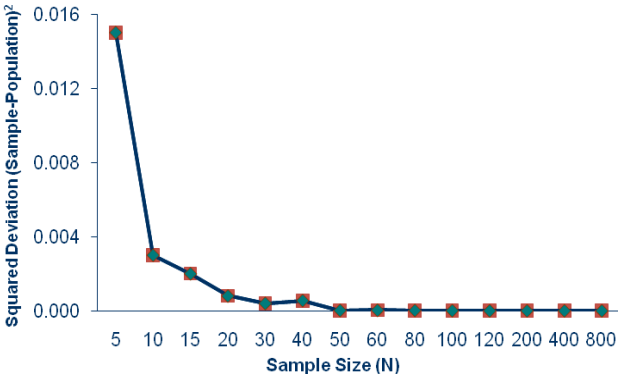
The inputs to the sampling algorithm included the number of samples drawn from each artificial population, the sample size, the number of possible categories, and the probability of each category's occurrence in the artificial population. The output of the sampling algorithm included the population category probabilities and the sample category probabilities. Twelve-hundred Monte Carlo samples were drawn from each of the 30 population distributions and this procedure was repeated over 14 levels of sample size (5, 10, 15, 20, 30, 40, 50, 60, 80, 100, 120, 200, 400, and 800). Figure 3 illustrates the simulations.

Figure 3. The Monte Carlo simulations entailed sampling many times from a known population probability distribution. The bootstrap bias-correction simulations entailed using the observed sample probabilities from the Monte Carlo simulations as the "population" probability distribution and sampling to obtain many bootstrap resamples.



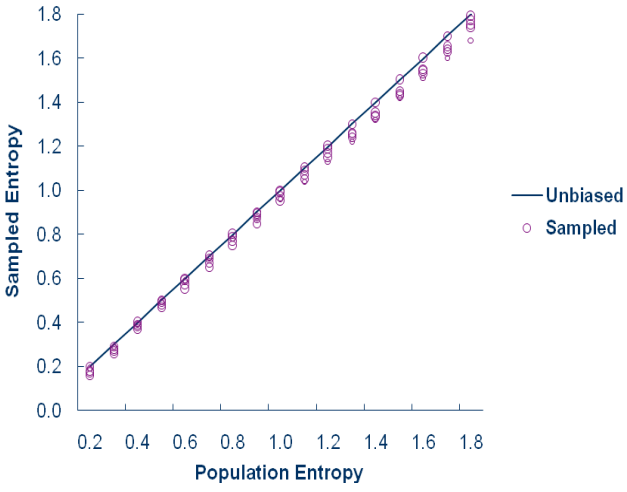
The Monte Carlo simulations showed that entropy is a consistent estimator. The squared deviation of the sample entropy from the population entropy decreased as the sample size increased, as seen in the data plotted in Figure 4.

Figure 4. Entropy is a consistent estimator, i.e., the squared deviation of the sample entropy from the population entropy decreased with increasing sample size.



Entropy is a biased estimator; it tends to underestimate the population value, as seen in the data plotted in Figure 5. One way to reduce the bias is through *bootstrap* data resampling. The bootstrap procedure simply treats the sample as the population and resamples from it accordingly. The bootstrap bias correction simulations replicated the Monte Carlo method described above except that the obtained sample response probabilities from the first simulations were input as the "population" probabilities and twelve-hundred bootstrap resamples were drawn from each "population" (recall Figure 3 above).

Figure 5. Entropy tends to underestimate the population value—raw sample entropy is a biased estimator of population entropy.



The bootstrap procedure yielded a correction factor for the biased sample entropy. The relationship between the population and sample values is:

$$P = mS, \quad (4)$$

where P is the population value, S is the sample value, and m is the bias function, which is unknown. The relationship between S and R (the resample value) can estimate m by:

$$S = mR \quad \text{and} \quad (5)$$

$$\tilde{m} = k\,(R/S), \quad (6)$$

where **k** is a constant. Substituting into equation 4 gives:

$$P = k(R/S)\,(S) = kR \quad \text{and} \quad (7)$$

$$k = P/R. \quad (8)$$

The values of **k** (mean = 1.056) for each of the 30 distributions were computed based on the population and mean resample entropy; **k** was then applied to each mean sample entropy to yield a bias-corrected entropy estimate (R^2 = 1.0, slope = 1.0). Similar results were obtained when the sample probabilities were treated as the population distribution, the resample entropy (Rl) was treated as the sample estimate, resamples (R2) were taken from the original resamples (Rl), and **k** was computed as the ratio of the sample entropy to the second resample (R2) entropy (mean **k** = 1.056). Similar results were also obtained when the original resample (Rl) probabilities were treated as the population distribution, the second resample (R2) entropy was treated as the sample estimate, resamples were drawn from R2 to yield R3, and **k** was computed as the ratio of Rl to R3 (mean **k** = 1.055). The mean value of **k** across all three situations (population as population, sample as population, and resamples as populations) was 1.056. In all cases, applying **k** to the "sample" value yielded the "population" entropy. Bootstrap simulations, which sampled from the Monte Carlo sample probability distributions as though they were population distributions, allowed the calculation of a correction factor, **k**, which virtually eliminated the bias. The value of **k** varied little as a function of whether the true population, the sample, the original resample (R1), or the resample from the original resample (R2) was treated as the population. These results suggest that bias-corrected entropy is an efficient estimator of population nominal variability. That is, bias-corrected sample entropy values cluster closely around the population entropy values.

Estimating population variability

Sample entropy satisfies all four criteria for a population estimator. First, sample entropy is a sufficient estimator—its calculation includes all the data. Two computer-intensive resampling techniques showed that bias-corrected sample entropy is a consistent and efficient estimator of the variability in nominal population distributions. Monte Carlo simulations established the consistency of the entropy estimator by repeatedly sampling from artificial and, therefore, known population distributions. The sample entropy values were compared to the population values to establish that, as the sample size increased, the estimator converged on the population values. The bootstrap resampling technique capitalized on the similarity between the population:sample relationship and the sample:resample relationship and produced a bias correction factor. Applying the correction factor makes sample entropy an efficient estimator, adjusting the sample values to be very close to the population values just as the correction factor N/N-1 corrects the expected sample variance value to σ^2 in ratio data.

Entropy is a well-established measure of nominal variability already used in contingency table analysis. The present work extends the usefulness of the measure to estimating nominal variability in populations from nominal variability in samples. Sample entropy is now available as the dependent variable in standard techniques like the analysis of variance. Constructing a confidence interval for a sample mean requires knowing the variability of the sample mean, which requires knowing how to compute the variability of the sampling distribution. That information was previously inaccessible in nominal data, but the empirical sampling distribution described here means that the variability is known. Why do we need to measure the variability? Because a confidence interval is the upper and lower bound between which we can be (measurably) confident the population mean falls. Confidence intervals quantify the uncertainty of an estimate and its precision by specifying an interval between which the true population value can be said to lie with specified probability. The entropy estimator is potentially useful in many areas where only nominal data are available.

Now the school psychologist who is trying to ensure intervention fidelity may ask whether the sample distribution of the pronunciations produced by students with identified reading problems reflects the variability in the population. Perhaps students who generate pronunciations that reflect the pronunciation variability of the language and students who do not would benefit from different interventions. Computing entropy for those samples and using it as the dependent variable in an analysis of variance could provide an indication of whether the two samples are from different populations and guide the design of relevant reading intervention strategies. Psychology is rife with research questions that could benefit from this approach and nominal data are plentiful.

Nominal data is arguably the most readily available data in many types of psychological research. It is commonly accepted that nominal data cannot be as informatively analyzed as ordinal, interval, or ratio data.

Much information in the plentiful nominal data was inaccessible because there was no estimate of population variability as there is when computing means and variances to construct confidence intervals for interval and ratio data. This has led to information loss in the analysis of nominal data. Now, however, entropy used in conjunction with traditional techniques like analysis of variance can wring more information from nominal data and thus address new research questions in psychology.

References

Brunton, D. H., & Li, X. (2006). The song structure and seasonal patterns of vocal behavior of male and female bellbirds (*Anthornis melanura*). *Journal of Ethology*, *34*, 17–25.

Fano, R. M. (1961). *Transmission of information*. New York: MIT Press and John Wiley & Sons.

Fuchs, D., & Fuchs, L. S. (2006). Introduction to response to intervention: What, why, and how valid is it? *Reading Research Quarterly*, *41*, 92-99. Retrieved February 12, 2008, from http://www.reading.org/Library/Retrieve.cfm?D=10.1598/RRQ.41.1.4&F=RRQ-41-1-Fuchs.html

Gokhale, D. V., & Kullback, S. (1978). *The information in contingency tables*. New York: Marcel Dekker.

Hammersley, J. M., & Handscomb, D. C. (1964). *Monte Carlo methods*. New York: John Wiley & Sons.

Hardy, M. (2003). An illuminating counterexample. *The American Mathematical Monthly, 110*, 234-238.

Hawley, K. M., & Weisz, J. R. (2002). Increasing the relevance of evidence-based treatment review to practitioners and consumers. *Clinical Psychology: Science and Practice, 9*, 225–230.

Kalos, M. H., & Whitlock, P. A. (1985). *Monte Carlo methods: Volume I: Basics*. New York: John Wiley & Sons.

Kullback, S. (1968). *Information theory and statistics*. New York: Dover.

Magidson, J. (1981). Qualitative variance, entropy, and correlation ratios for nominal dependent variables. *Social Science Research*, *10*, 177-194.

Noreen, E. W. (1989). *Computer-intensive methods for testing hypotheses*. New York: John Wiley & Sons.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes in c: The art of scientific computing* (pp. 628-636). New York: Cambridge University Press.

Reynolds, H. T. (1984). *Analysis of nominal data*. Beverly Hills: Sage Publications.

Shannon, C. E. (1950). Prediction and entropy of printed English, *The Bell System Technical Journal*, *30*, 50-64.

Schumacker, R. E., & Akers, A. (2001). Understanding statistical concepts using S-plus. Mahwah, NJ: Lawrence Erlbaum.

---

[i] The choice of log base in the equation was made for convenience and is not a critical matter (e.g. Fano, 1961, p. 27). Recall the entropy equation:
$H_B = -\Sigma\, p \log_B (p)$,
and recall that
$\log_B(x) = \ln(B) * \ln(x)$,
thus,
$H_B = \ln(B)\, H$.

[ii] This is not universally true, however (e.g., Hardy, 2003).

[iii] Although it is beyond the scope of this paper to specify procedures for generating random samples from a distribution, the literature contains examples of sampling from (among others) the following types of distributions: binomial, Poisson, geometric, exponential, normal, Student's *t*, gamma, F, Weibel and Cauchy, chi-squared, and beta (Rubenstein, 1982; Cooke, Craven, and Clarke, 1982). In short, one can sample from a distribution by generating uniformly distributed, random numbers between zero and one, inverting the cumulative distribution function form of the model, and computing a statistic for a simulated random sample. The significance of the obtained value of the statistic can then be evaluated relative to the simulated samples (Noreen, 1989).