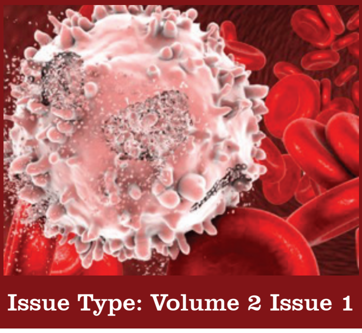
RAS ONCOLOGY & THERAPY

ISSN : 2766-2586

Research Article : A Note on the Exact Relation Between Mixture Likelihood and Entropy



Issue Type: Volume 2 Issue 1

Author Name:
Anthony M. Orlando* & Rahul Dhanda**
* President, ANOMX LLC
**Adjunct Associate Professor, UTHSC

Corresponding Author:
Anthony M. Orlando

Citation: Anthony M. Orlando, A Note on the Exact Relation Between Mixture Likelihood and Entropy

Received Date: 9th March 2021

Published Date: 18th March 2021

Abstract:
It is interesting to note that the expected value of the log likelihood function is entropy. This note shows that there is an exact relationship between the mixture log likelihood function (ln L_M) and the sum of the mixing distribution entropy (H_M) and the mixture density entropy (H_D). *Ln* L_M is seen as a function exactly of four Shannon entropies, each a unique measure of uncertainty.

This method, known as mixtures of linear models (MLM), is a form of empirical Bayes which uses a non-informative uniform prior and generates both confidence intervals and p-values which clinicians and regulatory agencies can use to evaluate scientific evidence.

An example based on allergic rhinitis symptoms scores are given and show how easy it is to assess the fit of the model and evaluate the results of the trial.

Key words: Expectation; Interaction; Uncertainty; Uniqueness.

Introduction:

Inference when conducting Bayesian analyses is derived from the posterior probability which is the result of two precursors: a prior probability and a likelihood function. The likelihood function is derived from the statistical modeling of the observed data, and inference derived by using the Bayes' theorem. In essence, Bayes theorem or factor can be written as the product of the likelihood and the prior probability (posterior probability = likelihood * prior probability). This relationship allows us to update our earlier theory (hypothesis) with a subsequent theory based on the observed data. (Faulkenberry, 2018).

A consequence of using evidence-based methods (which combines clinical expertise, patient's values and best available scientific evidence) to assist in making treatment decisions for their patients by clinicians is their reliance on p-values when evaluating scientific evidence, (Masic et al, 2008). A key feature of evidence based medicine is meta-analyses and systematic literature reviews. However, often the correct interpretation of the p-value eludes those most interested in evaluating the scientific evidence in front of them. The suggestion has been to use the Bayesian inference method to draw inferences, (Cohen, 2011). Indeed, the push within the statistical community has been to push aside classical methods that place heavy emphasis on using p-values to alternative methods that have less reliance on the use of p-value, such as the Bayesian inference, (Wagenmakers, 2007).

There are some considerations to think about prior to shifting away from the standard century old method. These considerations include among others:

- The selection of priors; and
- The incorporation of prior distributions into meta-analytic framework

Perhaps a method that incorporates the best of both the standard approaches to hypothesis testing and the rigor of Bayesian framework can be used to overcome these limitations, and enables an easy interpretation of scientific interpretation by clinicians and regulatory agencies.

It is interesting to note that the expected value of the log likelihood function is entropy. This note shows that there is an exact relationship between the mixture log likelihood function (ln L_M) and the sum of the mixing distribution entropy (H_M) and the mixture density entropy (H_D). *Ln* L_M is seen as a function exactly of four Shannon entropies, each a unique measure of uncertainty.

This method, known as mixtures of linear models (MLM), is a form of empirical Bayes which uses a non-informative uniform prior and generates both confidence intervals and p-values which clinicians and regulatory agencies can use to evaluate scientific evidence.

Method

Fisher information is the limiting form of several different measures of entropy & has been referred to as a "mother" information by Frieden (1998). Physical entropy is the volume of the region of the energy of the system, (Einstein 1904). Molecules that collide in a non-equilibrium system produce entropy. Statistical entropy is related to the volume of the likelihood function parameter space, where the likelihood function is the energy of the data generating system. This is demonstrated by the product of the date densities. Fisher information is related to the surface (Cover and Thomas, 1992) and is the limit of entropy. Shannon entropy (statistical information theory) for a discrete distribution with probabilities $p_1, p_2, \ldots, p_k$ is

$$H = -\sum_{j=1}^{k} p_j \ln(p_j), \qquad (1)$$

A unique measure of uncertainty for a proof, see Applebaum (1996). Entropy for the mixing distribution of a MLM as derived by Orlando and Allen (2001) is

$$H_M = -\sum_{i=1}^{n}\sum_{j=1}^{k} \frac{P(j|i)}{n} \ln\left(\frac{\hat{\rho}_j}{P(j|i)}\right)$$

where P $(j|i)$ is the Bayes posterior conditional probability of the i^{th}observation in the j^{th} component and $\rho\hat{}j$ is the estimated mixture proportion for the jthcomponent. H_M can be expressed as

$$H_M = H_B - H_W = H_B - H_{WC} + \ln n, \qquad (2)$$

where H_B is the Shannon entropy for the population, H_{WC} for the interaction, and ln n for the correction, each a unique measure of uncertainty. H_W is the average within component entropy. In this note we show that an exact relationship exists between the log likelihood function for a MLM (ln L_M) and entropy.

LIKELIHOOD AND ENTROPY

The exact relationship between the likelihood function and entropy for a uni-component probability distribution has been established (see, e.g. Kapur and Kesavan (1992); Lavenda (1991)). For the unicomponent normal density we define

$$-2\ln L = 2nH_N. \text{ Here, } H_N \text{ is} \qquad (3)$$

the entropy for the normal density,

$$H_N = \frac{1}{2}(\ln\hat{\sigma}^2 + \ln 2\pi + 1), \qquad (4)$$

where $\hat{\sigma}^2$ is the estimated residual mean square (Kapur and Kapur 1990); (Orlando and Allen 2000).

The main result of this note is the following theorem:

Theorem 1 *Consider the mixture of linear models*

$$y_i = \begin{cases} x_i^T\beta_1 + \epsilon_{i1} \text{ with probability } \rho_1 \\ x_i^T\beta_2 + \epsilon_{i2} \text{ with probability } \rho_2 \\ \qquad\vdots \\ x_i^T\beta_k + \epsilon_{ik} \text{ with probability } \rho_k \end{cases}$$

where the error terms are independent with mean 0 and variance σ_j^2, $j = 1, 2, \ldots, k$ and the mixing proportions satisfy $\sum_{j=1}^{k}\rho_j = 1, \rho_j > 0$. With each quantity evaluated at the ML estimates we have that the log likelihood function is given by

$$-\ln L_M = nH_M + nH_D \qquad (5)$$

where H_D is the mixture density entropy,
Proof: See the Appendix.

Substituting for H_M from Equation 2, we have that the average -ln L_M is a function of four unique measures of uncertainty and is itself a unique and complete measure of information. The total entropy is the expected value of -ln L_M and represents the total volume of the parameter space region of L_M. Equation 5 holds only at the solution for max -ln L_M where the difference in -ln L_M for two hypotheses is better approximated asymptotically by χ^2.

For a mixture of normal densities we have

$$-2\ln L_M = 2nH_M + n\sum_{j=1}^{k}\hat{\rho}_j \ln\hat{\sigma}_j^{\,2} + n\ln 2\pi + n. \qquad (6)$$

The following example provides an illustration of a normal MLM application and the computing of -2ln L_M from the Equation 6 formula.

Example 1. Figure 1 shows the treatment cumulative distribution functions (cdf) for allergic rhinitis total symptom scores from an actual double-blind clinical trial with 149 patients comparing three doses of a test drug to a positive reference and placebo. The crossing of the cdf's, in particular the placebo cdf at the median, indicates a mixed response (Oja 1981). Also, there is marked non-normality and unequal treatment variances. Consequently, the uni-component model analysis results do not yield suitable residuals.

As an alternative, we use a two-component normal mixture of linear models (MLM) with treatment and site as main effects. MLM is a proprietary software program that has been validated based on the comparison of the results obtained from the algebraic expressions with the algorithm computations. Further, the Assay validity of the study was established by the p-value of 0.06 for the established reference drug, compared to placebo, as computed by the MLM software.

The -2ln L_M value as computed by the MLM software using a differential evolution search algorithm is 1056.19, where -2ln L_M is computed as

$$-2\ln L_M = -\sum_{i=1}^{n}\ln\sum_{j=1}^{k}\hat{\rho}_j f_j(y_i).$$

The Equation 6 formula also gives 1056.19,
$-2 \ln L_M = 65.550 + 567.796 + 273.844 + 149 = 1056.19$.
An analysis of mixtures (ANOMX) comparison of this model with the treatment only model shows an increase in -2ln L_M of 15.28 for eight additional parameters, but a decrease in 2nH_M of 28.37 (30.2%). For the treatment only model, we have
$-2 \ln L_M = 93.918 + 554.706 + 273.844 + 149 = 1071.47$
This is due primarily to a decrease in H_W from 0.471 to 0.114 as a consequence of a small estimated variance of 0.189 for the smaller component, indicating a spurious local maximizer solution. An entropy-penalized log likelihood in the treatment and site model was used increasing -2ln L_M to 1048.38 and 2nH_M to 102.69 with a ten-fold increase in the smaller variance estimate. The increase in -2ln L_M of 23.09 is compared to the critical value of 10.31 for χ^2 at the 5% level, supporting the inclusion of site in the model, data given in Table 1, below.

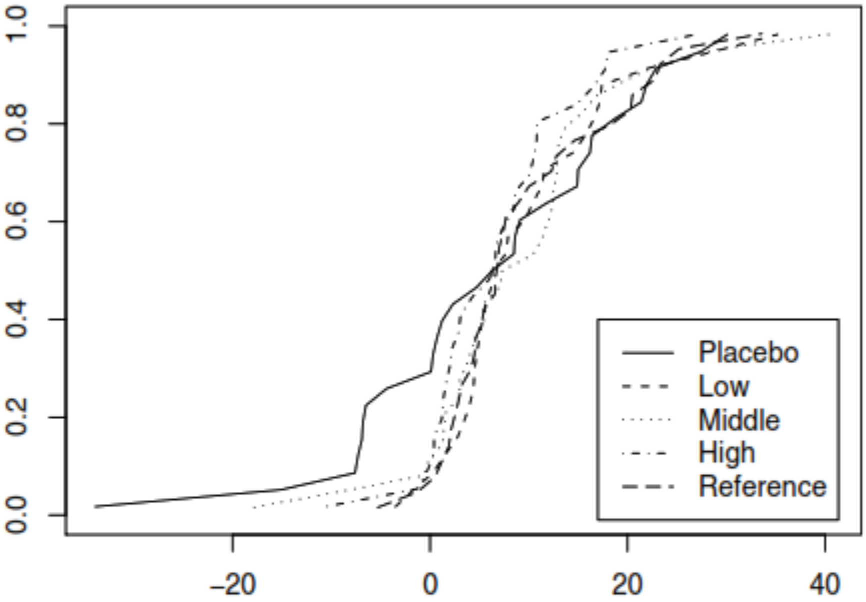


Figure 1: Treatment cumulative distribution function for allergic rhinitis total symptom scores. The observations are from an actual double-blind clinical trial with 149 patients comparing three doses of a test drug to a positive reference and placebo.

Table 1: TSS Data Pooling the Two Normal Subpopulations and Resulting Treatment Comparisons of the Means

Treatment	Placebo (P)	High (H)	Middle (M)	Reference (R)	Low (L)
N	29	28	31	32	29
Mean	4.6	6.9	8.7	9.5	9.7
Comparison	R-P	L-P	M-P		
Mean (SE)	4.9 (2.6)	5.1 (2.6)	4.0 (2.5)		
95% CI	-0.196-9.996	0.004-10.196	-0.90-8.9		
P-value	0.06	0.05	0.11		

Conclusion

We have shown that there is an exact relationship between entropy and the log mixture likelihood function, and that this relationship allows us to derive confidence intervals and p-values that can be interpreted easily. Entropy is a valuable tool for finding the optimal solution and supporting the use of mixture models for inference.

Appendix

Proof of Theorem 1. We define for any mixture of linear models the basic probability unit, the within cell probability, as

$$\frac{P(j|i)}{n} = \frac{\hat{\rho}_j f_j(y_i)}{nL_i}, \qquad (7)$$

where $f_j(y_i)$ is the probability ordinate for the i^{th} observation in the j^{th} component and L_i is the likelihood for the i^{th} observation, the marginal density function. The $f_j(y_i)$ and L_i are evaluated at the *ML* estimates. The within cell or interaction entropy is

$$H_{WC} = -\sum_{i=1}^{n}\sum_{j=1}^{k}\frac{P(j|i)}{n}\ln\left[\frac{\hat{\rho}_j f_j(y_i)}{nL_i}\right]. \qquad (8)$$

From Equation 2, H_{WC} can be expressed as

$$H_{WC} = H_B - H_M + \ln n. \qquad (9)$$

From Equations 8 and 9 we have

$$\begin{aligned} H_B - H_M + \ln n = \\ -\sum_{i=1}^{n}\sum_{j=1}^{k}\frac{P(j|i)}{n}\ln\hat{\rho}_j - \sum_{i=1}^{n}\sum_{j=1}^{k}\frac{P(j|i)}{n}\ln f_j(y_i) \\ +\sum_{i=1}^{n}\sum_{j=1}^{k}\frac{P(j|i)}{n}\ln n + \sum_{i=1}^{n}\sum_{j=1}^{k}\frac{P(j|i)}{n}\ln L_i. \end{aligned}$$

Noting that $\hat{\rho}_j = \sum_{i=1}^{n} P(j|i)/n$ and simplifying, we get

$$\begin{aligned} H_B - H_M + \ln n = H_B - \sum_{i=1}^{n}\sum_{j=1}^{k}\hat{\rho}_j \ln f_j(y_i) + \ln n + \frac{\sum_{i=1}^{n}\ln L_i}{n} \\ -\sum_{i=1}^{n}\ln L_i = nH_M - n\sum_{i=1}^{n}\sum_{j=1}^{k}\hat{\rho}_j \ln f_j(y_i) \\ -\ln L_M = nH_M + nH_D, \end{aligned}$$

where H_D is the mixture density entropy. We then have that $-E(\ln L_M) = H_M + H_D$ and is estimated by the average of the negative log likelihood function.

References

Faulkenberry T. J. (2018). A Simple Method for Teaching Bayesian Hypothesis Testing in the Brain
and Behavioral Sciences. Journal of undergraduate neuroscience education : JUNE : a publication of FUN, Faculty for Undergraduate Neuroscience, 16(2), A126–A130.

Masic, I., Miokovic, M., &Muhamedagic, B. (2008). Evidence based medicine - new approaches and challenges. Acta informatica medica : AIM : journal of the Society for Medical Informatics of Bosnia & Herzegovina (2008): casopisDrustva za medicinskuinformatikuBiH, 16(4), 219–225. https://doi.org/10.5455/aim.16.219-225

Hillel W. Cohen (2011), P Values: Use and Misuse in Medical Literature, American Journal of Hypertension, Volume 24, Issue 1, January, Pages 18–23, https://doi.org/10.1038/ajh.2010.205

Wagenmakers EJ (2007). A practical solution to the pervasive problems of p-values. Psychon Bull Rev.;14:779–804. doi: 10.3758/bf03194105.

D. Applebaum (1996). Probability and Information. Cambridge University Press, New York.

M. T. Cover and J. A Thomas (1992). Elements of Information Theory. John Wiley and Sons, New York.

B. R. Frieden (1998). Physics from Fisher Information (A Unification). Cambridge University Press, New York.

J. N. Kapur and J. S. Kapur (1990). On the relationship between entropy and variance. Metron, 48:113–130.

A. M. Orlando and D.M. Allen (2000). The analysis of entropy and variance for a mixture model (ANOMX: The Analysis of Mixtures). Presented at the Joint Statistical Meetings in Indianapolis, August 13.

J. N. Kapur and H. K. Kesavan (1992). Entropy Optimization Principles with Applications. Academic Press, New York.

B. H. Lavenda (1991). Statistical Physics (A Probabilistic Approach). John Wiley and Sons, New York.

A. M. Orlando (1999). Entropy and classification (discovering order in a clinical trial). Presented at the Conference on Non linear Statistical Models: Implementation and Application, University of Kentucky, Lexington, November 6.

H. Oja (1981). On location, scale, skewness and kurtosis of univariate distributions. Scandanavian Journal of Statistics, 8:154–168.

A. M. Orlando and D. M. Allen (2001). The analysis of entropy and likelihood for a mixture model (ANOMX: the analysis of mixtures). ASA 2000 Proceedings of the Biopharmaceutical Section, pages 45–50.

Einstein A (1904). On the General Molecular Theory of Heat. Annalen der Physik, 14::354-362