ON THE USE OF RELATIVE LIKELIHOOD RATIOS IN STATISTICAL INFERENCE

Charles G. Boncelet, Lisa M. Marvel**, Michael E. Picollelli**

*University of Delaware, Newark DE USA
**US Army Research Laboratory, APG Aberdeen MD USA

ABSTRACT

We ask the question, "How good are parameter estimates?," and offer criticism of confidence intervals as an answer. Instead we suggest the engineering community adopt a little known idea, that of defining plausible intervals from the relative likelihood ratio. Plausible intervals answer the question, "What range of values could plausibly have given rise to the data we have seen?" We develop a simple theorem for computing plausible intervals for a wide variety of common distributions, including the Gaussian, exponential, and Poisson, among others.

1. INTRODUCTION

We consider a basic question of statistical analysis, "How good are the parameter estimates?". Most often this question is answered with confidence intervals. We argue that confidence intervals are flawed. We propose that a little known alternative based on relative likelihood ratios be used. We term these "plausible intervals." We present a theorem on computing plausible intervals that applies to a wide range of distributions, including the Gaussian, exponential, and Chi-square. A similar result holds for the Poisson. Lastly, we show how to compute plausible regions for the Gaussian when both location and scale parameters are estimated.

This work is part of a larger effort to understand why statistical and signal processing procedures do not work as well in practice as theory indicates they should. Furthermore, we strive to develop better and more robust procedures. In this work, we begin to understand why "statistically significant" results are, upon further evaluation, not as reliable as thought. One reason is that confidence intervals are too small. They do not accurately reflect the range of possible inputs that could have given rise to the observed data.

This work is not defense specific, though there are numerous defense applications of statistical inference and signal processing. We expect this work to influence a wide range of procedures beyond simple parameter estimation. For instance, Kalman filtering, spectral estimation, and hypothesis testing are potential applications.

CGB can be reached at boncelet@udel.edu, LMM at marvel@arl.army.mil, and MEP at mpicolle@udel.edu.

Standard statistical procedures including maximum likelihood estimates and confidence intervals can by found in many textbooks, including Bickel and Docksum [1] and Kendall and Stuart [4]. Relative likelihoods have received some attention in the statistics and epidemiological literature, but little attention in the engineering literature. The best reference on relative likelihood methods is the text by Sprott [7]. One engineering reference is a recent paper by Sander and Beyerer [6].

In this paper, we adopt a "frequentist" interpretation of probability and statistical inference. Bayesian statisticians adopt a different view, one with which we have some sympathy, but that view is not explored herein. For a recent discussion of Bayesian statistics, see Jaynes [3].

2. PRELIMINARIES: LIKELIHOOD FUNCTIONS, MAXIMUM LIKELIHOOD ESTIMATES, AND CONFIDENCE INTERVALS

Consider a common estimation problem: estimating the mean of a Gaussian distribution. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be IID (independent and identically distributed) Gaussian random variables with mean μ and variance σ^2, i.e., $\boldsymbol{X}_i \sim N(\mu, \sigma^2)$.

For the moment, we assume we know σ^2 and seek to estimate μ. The density of each $\boldsymbol{X}_i$ is

$$f(x_i;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right)$$

The likelihood function of μ is

$$\begin{aligned} L(\mu; x_1^n) &= f(x_1;\mu) f(x_2;\mu) \cdots f(x_n;\mu) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(\frac{-\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}\right) \end{aligned}$$

where we use the notation $x_1^n = x_1, x_2, \ldots, x_n$.

The *maximum likelihood estimate* (MLE) of μ is found by setting the first derivative of $L(\mu; x_1^n)$ to 0,

$$0 = \frac{d}{d\mu} L(\mu; x_1^n)\bigg|_{\mu=\hat{\boldsymbol{\mu}}}$$

The calculations are somewhat easier if we find the maximum of the log-likelihood function,

$$\begin{aligned} 0 &= \frac{d}{d\mu} \log L(\mu; x_1^n)\Big|_{\mu=\hat{\boldsymbol{\mu}}} \\ &= \frac{d}{d\mu}\left(-\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \right)\Bigg|_{\mu=\hat{\boldsymbol{\mu}}} \\ &= \frac{n}{\sigma^2}\left(\frac{1}{n}\sum_{i=1}^{n} x_i - \hat{\boldsymbol{\mu}} \right) \end{aligned}$$

From which we conclude the MLE of μ is the sample mean

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}_n$$

Just how good is the sample mean as an estimator of μ? A commonly used measure of goodness is the *confidence interval*. Find $u(\boldsymbol{X}_1^n)$ and $v(\boldsymbol{X}_1^n)$ such that

$$\Pr\big[u(\boldsymbol{X}_1^n) \le \mu \le v(\boldsymbol{X}_1^n)\big] \ge 1-\alpha$$

for some $\alpha > 0$. Normally, the confidence interval is selected as the minimal range $v(\boldsymbol{X}_1^n) - u(\boldsymbol{X}_1^n)$. For the Gaussian example, the confidence interval is

$$\begin{aligned} \Pr\big[u(\boldsymbol{X}_1^n) \le \mu \le v(\boldsymbol{X}_1^n)\big] &\ge 1-\alpha \\ u(\boldsymbol{X}_1^n) &= \hat{\boldsymbol{\mu}} - \frac{c\sigma}{\sqrt{n}} \\ v(\boldsymbol{X}_1^n) &= \hat{\boldsymbol{\mu}} + \frac{c\sigma}{\sqrt{n}} \end{aligned}$$

where $c = \Phi^{-1}(1-\alpha/2)$ and $\Phi(\cdot)$ is the Normal distribution function. In the usual case where $\alpha = 0.05$, $c = 1.96$.

3. CRITICISMS OF CONFIDENCE INTERVALS

Many criticisms of confidence intervals have been raised. Here we list a few of them:

There is considerable confusion as to what a confidence interval actually represents. The standard interpretation goes something like this: Before the experiment is done, we agree to compute a sample average and a confidence interval (u, v) as above. Then the probability the interval will *cover* μ is $1-\alpha$.

Note, however, after the experiment is done, the $\boldsymbol{X}_i$ have values, x_i. Then, $\hat{\boldsymbol{\mu}}$, u, and v are numbers. Since μ is not considered to be random, we cannot even ask the question, "What is the probability μ is in the interval (u, v)?" μ is either in the interval or not, but the question is not within the realm of probability.

The confidence interval has probability about the true mean of $1-\alpha$ if $\mu = \hat{\boldsymbol{\mu}}$. In general $\mu \neq \hat{\boldsymbol{\mu}}$, and the interval contains less mass,

$$\Phi\big((v-\mu)/\sigma\big) - \Phi\big((u-\mu)/\sigma\big) < 1-\alpha$$

The confidence interval is not too helpful at predicting future values either. For instance, consider the following change to the experiment: After observing n measurements we will compute a sample average and a confidence interval. Then we will make another n measurements (independent of the first) and ask what is the probability the second sample mean is in the confidence interval? Let the second sample mean be denoted $\hat{\boldsymbol{\mu}}'$. Then,

$$\begin{aligned} \Pr\big[u \le \hat{\boldsymbol{\mu}}' \le v\big] &= \Pr\big[\hat{\boldsymbol{\mu}} - \frac{c\sigma}{\sqrt{n}} \le \hat{\boldsymbol{\mu}}' \le \hat{\boldsymbol{\mu}} + \frac{c\sigma}{\sqrt{n}}\big] \\ &= \Pr\big[-\frac{c\sigma}{\sqrt{n}} \le \hat{\boldsymbol{\mu}}' - \hat{\boldsymbol{\mu}} \le \frac{c\sigma}{\sqrt{n}}\big] \end{aligned}$$

Since both $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\mu}}'$ are independent $N(\mu, \sigma^2/n)$ random variables, the probability is

$$\Phi(c/\sqrt{2}) - \Phi(-c/\sqrt{2}) = 2\Phi(c/\sqrt{2}) - 1$$

For example, when $\alpha = 0.05$, $c = 1.96$ and the probability evaluates to only 0.834.

We regard the latter criticism as particularly damning. The usual reason to perform statistical analysis is to determine something about future observations. After all, the current observations are already known. Parameter estimates are often useful to the extent they help inform us about future observations. Confidence intervals are misleading indicators of future values.

To guarantee that $\hat{\boldsymbol{\mu}}'$ is in the confidence interval with probability $1-\alpha$, c must increase to $1.96\sqrt{2} = 2.78$.

4. RELATIVE LIKELIHOOD RATIO INTERVALS

We hope to revive an older idea that has received little attention in the engineering literature, *relative likelihood ratios*. A good reference is the text by Sprott [7]. The relative likelihood ratio is the following:

$$R(\theta; x_1^n) = \frac{L(\theta; x_1^n)}{\sup_\theta L(\theta; x_1^n)} = \frac{L(\theta; x_1^n)}{L(\hat{\boldsymbol{\theta}}; x_1^n)}$$

where θ represents the unknown parameter or parameters and $\hat{\boldsymbol{\theta}}$ is the MLE of θ.

The relative likelihood ratio helps answer the question, "What values of θ could plausibly have given the data x_1^n that we observed?" The relative likelihood is useful after the experiment is run, while probabilities are most useful before the experiment is run.

As an example, we consider the Gaussian example above. The unknown parameter is $\theta = \mu$ and the MLE is $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\mu}}$. Compare the relative likelihood ratio to a threshold,

$$R(\theta; x_1^n) = \frac{\exp\left(\dfrac{-\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right)}{\exp\left(\dfrac{-\sum_{i=1}^{n}(x_i-\hat{\boldsymbol{\mu}})^2}{2\sigma^2}\right)} \ge \alpha \qquad (1)$$

After taking logs and simplifying, the relation becomes

$$(\mu - \hat{\boldsymbol{\mu}})^2 \leq \frac{2\sigma^2}{n} \log 1/\alpha \tag{2}$$

Solving for μ gives a relative likelihood ratio interval, which we shall refer to as a *plausible interval*.

$$\begin{aligned} \hat{\boldsymbol{\mu}} - \sqrt{\frac{2\sigma^2 \log 1/\alpha}{n}} \leq \mu \leq \hat{\boldsymbol{\mu}} + \sqrt{\frac{2\sigma^2 \log 1/\alpha}{n}} \\ \hat{\boldsymbol{\mu}} - c\frac{\sigma}{\sqrt{n}} \leq \mu \leq \hat{\boldsymbol{\mu}} + c\frac{\sigma}{\sqrt{n}} \end{aligned} \tag{3}$$

When $\alpha = 0.05$, $c = 2.45$.

We see the plausible interval is bigger than the confidence interval. It is a more conservative measure. To reiterate, the plausible interval gives all values of θ that could have plausibly given rise to the data observed, where plausibly is measured by the ratio of the likelihood at θ to the maximum likelihood.

As another example, consider estimating the parameter in an exponential distribution. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be IID exponential with density $f(x) = \lambda e^{-\lambda x}$.

$$\begin{aligned} L(\lambda; x_1^n) &= \lambda^n \exp(-\lambda \sum_{i=1}^n x_i) \\ \hat{\lambda} &= \frac{n}{\sum_{i=1}^n x_i} = \frac{1}{\overline{x}_n} \\ R(\lambda; x_1^n) &= \left(\frac{\lambda}{\hat{\lambda}}\right)^n \exp\left(n(1 - \lambda/\hat{\lambda})\right) \end{aligned}$$

If we let $\gamma = \lambda/\hat{\lambda}$ and compare the relative likelihood to α, we obtain an interesting result:

$$\gamma e^{1-\gamma} \geq \alpha^{1/n} \tag{4}$$

We get a semi-graphical way of determining the plausible interval, as demonstrated in Figure 1. Compute $\alpha^{1/n}$ and find graphically or numerically the upper and lower values, u and v. An essential feature of the exponential inference problem is the asymmetry of the upper and lower values of the plausible interval. The upper limit is much farther from $\gamma = 1$ than is the lower limit.

As a third example of the utility of relative likelihoods, consider estimating the parameters in a multinomial distribution. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be n IID multinomial random variables with parameters $p_1, p_2, \ldots, p_k$ ($p_1 + p_2 + \cdots + p_k = 1$).

$$\begin{aligned} L(p_1, p_2, \ldots, p_k; x_1^n) &= p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k} \\ \hat{p}_j &= \frac{n_j}{n} \quad \text{for } j = 1, 2, \ldots, k \\ R(p_1, p_2, \ldots, p_k; x_1^n) &= \left(\left(\frac{p_1}{\hat{p}_1}\right)^{\hat{p}_1} \left(\frac{p_2}{\hat{p}_2}\right)^{\hat{p}_2} \cdots \left(\frac{p_k}{\hat{p}_k}\right)^{\hat{p}_k}\right)^n \end{aligned}$$

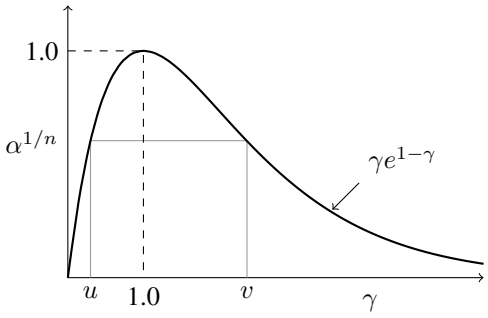

Fig. 1: Graphical representation of the plausible interval for an exponential distribution.

After taking logs, multiplying by -1, and comparing to α, the relation reduces to

$$-\sum_{j=1}^k \hat{p}_j \log\left(\frac{p_j}{\hat{p}_j}\right) = KL(\hat{p}||p) \leq -\frac{\log \alpha}{n} \tag{5}$$

where $KL(\hat{p}||p)$ is the Kullback-Leibler divergence between $\hat{p}$ and p (Kullback and Leibler [5]). In words, the set of plausible p's are those with a Kullback-Leibler divergence from $\hat{p}$ less than or equal to $-\log(\alpha)/n$.

Below we present a theorem on the calculation of plausible intervals for a wide class of exponential-type distributions.

Theorem 1 *If $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ are IID with common density*

$$f(x) = C\lambda^k x^{k-1} e^{-d\lambda^l x^l} \tag{6}$$

where k and l are shape parameters, d is a convenience constant (e.g., for the Gaussian, $d = 0.5$), and C is a normalizing constant. Let $x_1, x_2, \ldots, x_n$ denote the corresponding observations. The maximum likelihood estimate of λ is

$$\hat{\lambda} = \frac{nk}{dl \sum_{i=1}^n x_i^l} \tag{7}$$

The relative likelihood reduces to

$$\gamma e^{1-\gamma} \geq \alpha^{l/nk} \tag{8}$$

where $\gamma = (\lambda/\hat{\lambda})^l$ and α is the threshold level.

This theorem applies to a wide variety of distributions. Some are listed below in Table 1. For example, the exponential distribution has $k = 1$, $l = 1$, and $d = 1$ and (8) reduces to (4), since $l/k = 1$.

This theorem is easy to apply:

1. Compute the MLE of λ using (7).
2. Compute $\alpha^{l/nk}$.
3. Using graphical (Figure 1) or numerical means, solve (8) for u and v.

Exponential	$k = 1, l = 1, d = 1$
Weibull	$k = l, d = 1$
Erlang	$l = 1, d = 1$
Gaussian (known mean)	$k = 0, l = 2, d = 0.5$
Chi-Square	$k \leftarrow (k/2) - 1, l = 1, d = 0.5$
Rayleigh	$k = 1, l = 2, d = 0.5$

Table 1: Some distributions that meet the conditions of Theorem 1

4. Solve for upper and lower bounds on λ using $\lambda = \hat{\lambda} u^{1/l}$ and $\lambda = \hat{\lambda} v^{1/l}$.

While the Poisson does not fit the conditions of the theorem, its probability mass function is sufficiently close the same general conclusions apply. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be n IID Poisson random variables with parameter λ. Then,

$$\Pr\left[\boldsymbol{X}_i = x_i\right] = \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}$$

$$L(\lambda; x_1^n) = \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} e^{-n\lambda}$$

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x}_n$$

$$R(\lambda; x_1^n) = \left(\frac{\lambda}{\hat{\lambda}}\right)^{n\hat{\lambda}} e^{n\hat{\lambda}(1-\lambda/\hat{\lambda})}$$

Letting $\gamma = \lambda/\hat{\lambda}$ and comparing to α results in

$$\gamma e^{1-\gamma} \geq \alpha^{1/(n\hat{\lambda})} \tag{9}$$

The only difference for the Poisson is that α is raised to a data dependent power, $1/(n\hat{\lambda}) = (x_1 + x_2 + \ldots + x_n)^{-1}$.

5. LINEAR REGRESSION

In this section, we consider the linear regression problem with additive Gaussian noise of unknown variance. Let the regression problem be $y = X\beta + \epsilon$ where $\epsilon \sim N(0, \sigma^2 I)$. The MLE of β is

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

The MLE of σ^2 is

$$\widehat{\sigma^2} = \frac{(y - X\hat{\beta})^T (y - X\hat{\beta})}{n} = \frac{1}{n} \|y - X\hat{\beta}\|_2^2$$

Define two derived quantities:

$$\delta^2 = \frac{(\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta})}{n\widehat{\sigma^2}} \tag{10}$$

$$\gamma = \frac{\widehat{\sigma^2}}{\sigma^2} \tag{11}$$

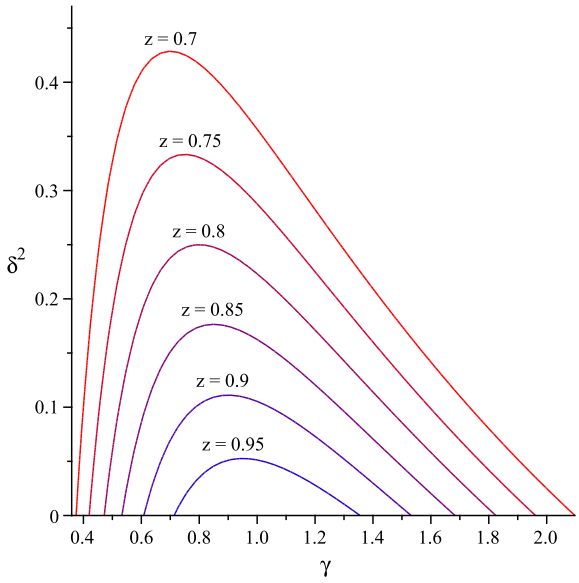

Fig. 2: A contour plot of location error, δ^2, versus scale error, γ for the Gaussian, with $z = \gamma \exp(1 - \gamma - \gamma\delta^2)$.

Using γ and δ^2, the relative likelihood can be written as

$$R(\beta, \sigma; y, X) = \gamma^{n/2} \exp\left(\frac{n}{2}(1 - \gamma - \gamma\delta^2)\right)$$

Comparing this to α and simplifying results in the following:

$$\gamma e^{(1-\gamma-\gamma\delta^2)} \geq \alpha^{2/n} \tag{12}$$

Since the Gaussian depends on two parameters, the relative likelihood depends on two parameters: δ^2 represents the error in the location estimate and γ the error in the scale estimate. When compared to α, we obtain *plausible regions*, a generalization to higher dimensions of plausible intervals.

The plausible region is defined by (12). When $\delta^2 = 0$, (12) reduces to (8); when $\gamma = 1$, (12) reduces to (2).

The function $z = \gamma \exp(1-\gamma-\gamma\delta^2)$ is shown in Figure 2. The location error is large when δ^2 is large. This happens when $\gamma < 1$, i.e., when $\widehat{\sigma^2} < \sigma^2$.

When $\gamma = 1$, $\delta^2 = -\log(\alpha^{(2/n)})$. It is interesting to ask what is the maximum value of δ^2 when γ is allowed to vary (i.e., when the variance is also unknown). One can solve the following maximization problem:

$$\max_{\delta^2, \gamma} \delta^2 \quad \text{such that} \quad \gamma e^{1-\gamma-\gamma\delta^2} \geq \alpha^{(2/n)}$$

The solution is $\gamma = \alpha^{(2/n)}$ and $\delta^2 = \alpha^{-(2/n)} - 1$. That the maximum value of δ^2 is larger when γ is allowed to vary is a restatement of the familiar log inequality, $x - 1 \geq \log(x)$.

6. CONCLUSION

We argue that confidence intervals are flawed and do not accurately represent the range of possible parameter values. Better we argue are plausible intervals based on relative likelihood ratios. Plausible intervals answer the question, "What range of parameter values could plausibly have given rise to the data we have observed?" Unlike probabilities, likelihood ratios are informative after the experiment is run.

We have presented a theorem that provides a simple sequence of steps to compute the plausible interval for a wide range of distributions, including the Gaussian, exponential, Weibull, Chi-squared, Erlang, and Poisson. We have extended the theorem to the Gaussian when both location and scale parameters are unknown.

The relative likelihood for the multinomial distribution reduces to the Kullback-Leibler divergence.

This work is part of a larger effort to understand why statistical and signal processing procedures often do not work as well in practice as the theory indicates and to develop better, more robust, procedures. Specific work to follow includes more analysis of the Gaussian case, including applications such as Kalman filtering, autoregressive fitting, and spectral analysis. Also, we need to extend Theorem 1 to a wider class of distributions.

We believe that plausible intervals can replace the need for resampling techniques such as the bootstrap and the jackknife (Efron [2]). Rather than creating new data, we can use plausible intervals to explore the space of plausible inputs.

7. REFERENCES

[1] P. J. Bickel and K. A. Doksum. *Mathematical Statistics*. Holden-Day Series in Probability and Statistics. Holden-Day, Inc., San Francisco, 1977.

[2] B. Efron. *The Jackknife, the Bootstrap and Other Resampling Plans.* Siam, Philadelphia, PA, 1982.

[3] Edwin T. Jaynes. *Probability Theory The Logic of Science*. Cambridge University Press, 2003.

[4] M. Kendall and A. Stuart. *The Advanced Theory of Statistics*, volume 1. Charles Griffin & Company Limited, London, 1977.

[5] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[6] Jennifer Sander and Jurgen Beyerer. Decreased complexity and increased problem specificity of bayesian fusion by local approaches. In *11th Internatonal Conference on Information Fusion*, pages 1036–1042, June 2008.

[7] D. A. Sprott. *Statistical Inference in Science*. Springer Series in Statistics. Springer, 2000.