THE INSTITUTE
OF STATISTICS

DERIVING GENERALIZED MEANS AS LEAST SQUARES

AND MAXIMUM LIKELIHOOD ESTIMATES

Roger L. Berger and George Casella

September, 1991

NORTH CAROLINA STATE UNIVERSITY
Raleigh, North Carolina

DERIVING GENERALIZED MEANS AS LEAST SQUARES
AND MAXIMUM LIKELIHOOD ESTIMATES

Roger L. Berger
Statistics Department
Box 8203
North Carolina State University
Raleigh, NC 27695-8203

and

George Casella[1]
Biometrics Unit
337 Warren Hall
Cornell University
Ithaca, NY 14853-7801

BU-1133-M   September 1991

ABSTRACT

Functions called generalized means are of interest in statistics because they are simple to compute, have intuitive appeal, and can serve as reasonable parameter estimates. The well-known arithmetic, geometric, and harmonic means are all examples of generalized means. We show how generalized means can be derived in a unified way, as least squares estimates for a transformed data set. We also investigate models that have generalized means as their maximum likelihood estimates.

Key words: Exponential family, arithmetic mean, geometric mean, harmonic mean.

1. INTRODUCTION

Abramowitz and Stegun (1965) define a generalized mean to be a function of n positive variables of the form

$$g_\lambda(x_1,\ldots,x_n) = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^\lambda\right)^{1/\lambda}, \quad \lambda \neq 0. \tag{1.1}$$

For $\lambda = 0$, the generalized mean is defined by continuity to be

$$g_0(x_1,\ldots,x_n) = \lim_{\lambda \to 0} g_\lambda(x_1,\ldots,x_n) = \exp\left(\frac{1}{n} \sum_{i=1}^{n} \log x_i\right) = \left(\prod_{i=1}^{n} x_i\right)^{1/n},$$

which is known as the geometric mean. Two other well-known special cases are the arithmetic mean ($\lambda = 1$) and the harmonic mean ($\lambda = -1$). The form of (1.1) leads us to inquire about the conditions that would yield (1.1) as a measure of center (a mean), or the models that would yield (1.1) as an estimate of a parameter. These are the two questions we investigate here.

We begin by noting that the generalized mean of (1.1) can, in fact, be generalized since each g_λ is of the form $h^{-1}\left(\frac{1}{n} \sum_{i=1}^{n} h(x_i)\right)$, where $h(x)$ is either x^λ or $\log x$. This suggests an even more general class of means defined by

$$g_h(x_1,\ldots,x_n) = h^{-1}\left(\frac{1}{n} \sum_{i=1}^{n} h(x_i)\right), \tag{1.2}$$

where $h(x)$ is any continuous, monotone function. We shall call any function of the form (1.2) a *generalized mean*. Letting x denote the vector $(x_1,\ldots,x_n)$, we shall use the notation $g_h(x) = h^{-1}(\bar{h}(x))$ for the function in (1.2). Hardy, Littlewood and Pólya (1934, Chapter III) consider properties of these generalized means. They show that these have fundamental properties one would expect from a mean, such as satisfying

$$\min_{1 \leq i \leq n} x_i \leq h^{-1}(\bar{h}(x)) \leq \max_{1 \leq i \leq n} x_i.$$

In this paper we show how generalized means can be derived in two different ways. In Section 2 we show that generalized means are least squares estimates from transformed data and, in Section 3, we see how generalized means can arise as maximum likelihood estimates, illustrating these results with examples using the arithmetic, geometric and harmonic means. Section 4 addresses the questions of standard errors and confidence intervals, and Section 5 contains a short discussion.

2. LEAST SQUARES ESTIMATES

Suppose $x_1, \ldots, x_n$ are data values for which a measure of center is desired. One way of deriving such a measure is to find the value a that minimizes

$$\sum_{i=1}^{n} (x_i - a)^2 \, , \tag{2.1}$$

for which the answer is $a = \bar{x}$. This is a least squares derivation of the arithmetic mean as given, for example, in Hogg and Craig (1978). If the data are now transformed to $h(x_1), \ldots, h(x_n)$, where h is a specified monotone, continuous function, a natural measure of center of the transformed data values is the value $h(a)$ that minimizes

$$\sum_{i=1}^{n} \left(h(x_i) - h(a)\right)^2 \tag{2.2}$$

But this minimizing value is $h(a) = \frac{1}{n} \sum_{i=1}^{n} h(x_i) = \bar{h}(x)$, so transforming back to the original scale

yields $a = h^{-1}(\bar{h}(x))$, the generalized mean of (1.2). Thus, the least squares estimate of center, based on the distance function in (2.2), is the generalized mean based on the function h.

Example 1. The arithmetic mean is the least squares estimate after the data have been transformed to $h(x) = x$ (no transformation), and the geometric mean is the least squares estimate after the data have been transformed to $h(x) = \log x$. If the data are transformed to $h(x) = 1/x$, the harmonic mean is the least squares estimate since

$$h^{-1}(\bar{h}(x)) = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i}\right)^{-1} .$$

More generally, after transformation to $h(x) = x^{\lambda}$, the least squares estimate is the generalized mean g_{λ} defined in (1.1).

Example 2. A popular family of data transformations is the Box-Cox (1964) family defined by

$$h_{\lambda}(x) = \begin{cases} \frac{x^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ \\ \log x & \lambda = 0 \end{cases} . \tag{2.3}$$

After transformation by a member of the Box-Cox family, the least squares estimate of center is $h_{\lambda}^{-1}(\bar{h}_{\lambda}(x))$. Furthermore, for any monotone, continuous function h, it is easy to verify that if $g(x) = ah(x) + b$, where a and b are constants that do not depend on x and $a \neq 0$, then $h^{-1}(\bar{h}(x)) =$

$g^{-1}(\overline{g}(x))$. (In fact, two generalized means, defined in terms of functions g and h, are the same if and only if g is a linear function of h. See Hardy, Littlewood and Pólya (1934), Theorem 83.) So, $h_\lambda^{-1}(\overline{h}_\lambda(x)) = h^{-1}(\overline{h}(x))$ where $h(x) = x^\lambda$. Thus, the generalized mean g_λ from (1.1) is the least squares estimate after transformation by h_λ of (2.3).

3. MAXIMUM LIKELIHOOD ESTIMATES

Generalized means defined as in (1.2) can also be derived as maximum likelihood estimates.

Example 3. Let $X_1,\ldots,X_n$ be a random sample from a lognormal population with density

$$f(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma} \frac{1}{x} e^{-(\log x - \log \theta)^2/(2\sigma^2)} , \quad x > 0 ,$$

where $\theta > 0$ is the unknown parameter to be estimated. Maximizing the likelihood function, as a function of θ, amounts to minimizing $\sum_{i=1}^{n} (\log x_i - \log \theta)^2$. This is the least squares problem that was discussed in Section 2, and the maximum likelihood estimate of θ is the geometric mean of $x_1,\ldots,x_n$.

When sampling from certain exponential families, generalized means arise as maximum likelihood estimators. Suppose now that $X_1,\ldots,X_n$ is a random sample from a population with a one-parameter, exponential-family density or probability mass function given by

$$f(x|\theta) = e^{\theta h(x) - H(\theta)} g(x) , \tag{3.1}$$

where $h(x)$ is monotone increasing. (Without loss of generality we assume $h(x)$ is increasing, otherwise we reparametrize in terms of $-\theta$, rather than θ.) Further, suppose that h and H are related by $h(x) = \frac{d}{dx} H(x)$. The log likelihood function is then

$$\ell(\theta|x) = \theta \sum_{i=1}^{n} h(x_i) - nH(\theta) + \sum_{i=1}^{n} \log g(x_i) .$$

Setting the derivative of $\ell(\theta|x)$ equal to zero yields $\sum_{i=1}^{n} h(x_i) = nh(\theta)$, and solving for θ gives the generalized mean $\hat{\theta} = h^{-1}(\overline{h}(x))$. Taking the second derivative and using the fact that $\frac{d}{d\theta} h(\theta) > 0$ (since h is increasing) for all θ, shows that $\hat{\theta}$ is the maximizing value.

Although (3.1) is a form of a one-parameter, exponential family, the question arises whether any exponential families exist with $h(x) = \frac{d}{dx} H(x)$. They do and two examples follow.

Example 4. Suppose $X_1,...,X_n$ is a random sample from a normal population with unknown mean θ and known variance σ^2. Factoring the normal density,

$$f(x|\theta) = (2\pi\sigma^2)^{-1/2}\exp\{-(x-\theta)^2/(2\sigma^2)\} ,$$

as in (3.1) yields $h(x) = x/\sigma^2$, $H(\theta) = \theta^2/(2\sigma^2)$ and $g(x) = (2\pi\sigma^2)^{-1/2}\exp\{-x^2/(2\sigma^2)\}$. Clearly, $\frac{d}{dx} H(x) = h(x)$ and $h(x)$ is increasing, so $\hat{\theta} = h^{-1}(\bar{h}(x))$ is the maximum likelihood estimate of θ. Since the generalized mean is unchanged by multiplication by a constant (see Example 2), we can use $h(x) = x$ rather than $h(x) = x/\sigma^2$ to define $\hat{\theta}$. Thus the arithmetic mean is the maximum likelihood estimate of θ.

Example 5. Suppose $X_1,...,X_n$ is a random sample from an inverted gamma distribution with density $f(x|\theta) = \theta x^{-2}e^{-\theta/x}$, $x > 0$. This is an exponential family of the form (3.1) with $h(x) = -1/x$, $H(\theta) = -\log\theta$ and $g(x) = x^{-2}$, $x > 0$. Clearly, h is increasing and $\frac{d}{dx} H(x) = h(x)$. So the maximum likelihood estimate of θ is $\hat{\theta} = h^{-1}(\bar{h}(x))$. As in Example 4, we get the same generalized mean if we multiply h by -1 and define $\hat{\theta}$ in terms of $h(x) = 1/x$. Thus, for this model, the maximum likelihood estimate is the harmonic mean.

We have seen three densities that yield the arithmetic, geometric and harmonic means as maximum likelihood estimates. Other exponential families of the form (3.1), which will have generalized means as maximum likelihood estimates, might be constructed in the following way. Let $H(t)$ be the cumulant generating function (log of the moment generating function) of a nondegenerate probability distribution. $H(t)$ is defined on some interval of t values for which $t = 0$ is an interior point in the interval, and $H(t)$ is strictly convex (see Brown (1986)). For simplicity we assume $H(t)$ is twice differentiable so the maximization argument following Example 3 holds. Let Y be a random variable whose distribution has cumulant generating function $H(t)$, and let $g(x)$ be the density of $X = h^{-1}(Y)$. (Since $H(t)$ is strictly convex, $h(t) = \frac{d}{dt} H(t)$ is increasing. Here we also have to assume that the range of $h(t)$ contains the support of Y so that $h^{-1}(Y)$ is well defined.) Then

$$\int_{-\infty}^{\infty} e^{\theta h(x)}g(x) \, dx = Ee^{\theta h(X)} = Ee^{\theta Y}$$

is the moment generating function of Y. It is defined for θ in the same interval as $H(\theta)$ is defined, and,

in fact, by definition,

$$\int_{-\infty}^{\infty} e^{\theta h(x)} \, g(x) \, dx = e^{H(\theta)} \, .$$

Thus $\exp\{\theta h(x) - H(\theta)\}g(x)$ integrates to 1 and defines an exponential family. The generalized mean $h^{-1}(\bar{h}(x))$ is the maximum likelihood estimate of θ for this model.

4. STANDARD ERRORS AND CONFIDENCE INTERVALS

The value of a point estimator like $h^{-1}(\bar{h}(\mathbf{X}))$ is greatly increased if its sampling distribution is known or, at least, some estimate of its precision is available. Fortunately, the functional form of $h^{-1}(\bar{h}(\mathbf{X}))$ is convenient enough that measures of precision can be provided. In simple cases, the exact distribution of $h^{-1}(\bar{h}(\mathbf{X}))$ can be derived. This is the case for the maximum likelihood estimators in Examples 3, 4 and 5. If $X_1,\ldots,X_n$ is a sample from a normal population, then the arithmetic mean also has a normal distribution. In the lognormal case, the geometric mean also has a lognormal distribution, and, in the inverted gamma case, the harmonic mean has an inverted gamma distribution. In these cases the standard error of the estimator can be calculated, or the exact distribution can be used to construct confidence intervals.

In more complicated situations, the exact distribution of $h^{-1}(\bar{h}(\mathbf{X}))$ may be hard to derive. But two simple approximation methods may be used to obtain confidence intervals based on the generalized mean. Both approximations are based on the realization that $\bar{h}(\mathbf{X})$ is just the sample mean of the random sample $h(X_1),\ldots,h(X_n)$. So by the Central Limit Theorem, $\bar{h}(\mathbf{X})$ is asymptotically normally distributed with mean $E_\theta h(X_1)$ and variance σ_h^2/n where $\sigma_h^2 = \text{Var}_\theta h(X_1)$. There is a difficulty here in that, typically, $E_\theta h(X_1) \neq h(\theta)$ since the transformation introduces a bias. But we shall assume that this bias is negligible and that we can replace $E_\theta h(X_1)$ by $h(\theta)$.

Using the asymptotic normality of $\bar{h}(\mathbf{X})$, an approximate $100(1-\alpha)\%$ confidence interval for $h(\theta)$ is given by $\bar{h}(\mathbf{X}) \pm t_{\alpha/2,n-1} s_h/\sqrt{n}$, where $t_{\alpha/2,n-1}$ is the upper $100(1-\alpha)$ percentile of a Student's t distribution with n-1 degrees of freedom and s_h^2 is the sample variance calculated from $h(X_1),\ldots,h(X_n)$,

$$s_h^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(h(x_i) - \bar{h}(x) \right)^2 \, . \tag{4.1}$$

Now, since h is monotone (we can assume it is increasing), this interval can be inverted to obtain

$$h^{-1}\left(\overline{h}(X) - t_{\alpha/2,n-1}\frac{s_h}{\sqrt{n}}\right) < \theta < h^{-1}\left(\overline{h}(X) + t_{\alpha/2,n-1}\frac{s_h}{\sqrt{n}}\right) \tag{4.2}$$

as an approximate confidence interval for θ. This confidence interval contains the point estimator $h^{-1}(\overline{h}(X))$, but, typically, the interval will not be centered at the point estimator.

Care must be taken that the endpoints, $\overline{h}(X) \pm t_{\alpha/2,n-1}s_h/\sqrt{n}$, are in the domain of h^{-1}. For example, consider the inverted gamma problem in Example 5 with $h(x) = -1/x$ (increasing). Since $0 < \theta < \infty$, we have $-\infty < h(\theta) < 0$. So the proper domain of h^{-1} is $-\infty < y < 0$. If the right endpoint, $\overline{h}(X) + t_{\alpha/2,n-1}s_h/\sqrt{n}$, is positive, it should be replaced by 0. Then $h^{-1}(0) = +\infty$ gives the correct endpoint for the confidence interval for θ. Blindly applying h^{-1} to a positive value will give an incorrect negative value for the endpoint.

To obtain a confidence interval that is centered at $h^{-1}(\overline{h}(X))$, we can use an alternative derivation. Since $\overline{h}(X)$ is asymptotically normal, $h^{-1}(\overline{h}(X))$ is also asymptotically normal with mean $h^{-1}(E_\theta h(X_1))$ and variance $(\frac{d}{d\theta}h^{-1}(\theta))^2\sigma_h^2/n$ (using the delta method). Again we assume the bias is negligible, that is, $h^{-1}(E_\theta h(X_1)) \approx \theta$. Note that $\frac{d}{d\theta}h^{-1}(\theta) = (\frac{d}{d\theta}h(\theta))^{-1} \equiv (h'(\theta))^{-1}$. (We must assume $h'(\theta) \neq 0$ for this asymptotic normality to hold.) Replacing parametric quantities by estimators, we obtain

$$h^{-1}(\overline{h}(X)) \pm t_{\alpha/2,n-1}\left|h'(h^{-1}(\overline{h}(X)))\right|^{-1}\frac{s_h}{\sqrt{n}} \tag{4.3}$$

as an approximate $100(1-\alpha)\%$ confidence interval for θ. A slightly more liberal interval would be obtained by replacing $t_{\alpha/2,n-1}$ with the normal percentile $z_{\alpha/2}$.

To compare the intervals (4.2) and (4.3) we conducted a small simulation study for the lognormal and inverted gamma cases of Examples 3 and 5. For the inverted gamma case, it is straightforward to verify that both intervals can be written in terms of the random variables $Y_i = \theta/X_i$, and thus the coverage probabilities are independent of θ. For the lognormal case, both intervals can be written in terms of the random variables $Y_i = \log X_i - \log \theta$, and the coverage probabilities are again independent of θ. Table 1 gives the results of the simulation.

Table 1: Coverage probabilities for the intervals (4.2) and (4.3) when sampling from the lognormal and inverted gamma densities. Estimates based on 20,000 simulated samples, nominal confidence coefficient = .90. Standard errors for all estimates are approximately .0021.

| | Estimated Coverage Probability | | | |
| | lognormal | | inverted gamma | |
Sample Size	Interval (4.2)	Interval (4.3)	Interval (4.2)	Interval (4.3)
2	.901	.885	.841	.911
5	.901	.885	.827	.899
8	.901	.890	.844	.890
10	.902	.893	.852	.892
15	.896	.893	.864	.889
20	.903	.899	.871	.889
25	.899	.897	.878	.891
30	.897	.897	.883	.890
40	.901	.899	.882	.894
50	.902	.902	.886	.898

In the lognormal case, interval (4.2) is exact, not approximate, because $h(X) = \log X$ has an exact normal distribution. The estimated probabilities in Table 1 reflect this, all being within two standard errors of the true value .90. But interval (4.3) also performs well. The coverage probabilities appear to be above .88 for all sample sizes and very near the nominal .90 for sample sizes above 20.

In the inverted gamma case, we see that the interval (4.2) tends to have coverage probability well under the nominal value, while the interval (4.3) has coverage closer to the nominal value. The change in center of the interval (4.3) as well as the extra factor, based on the derivative of h, leads to an interval with coverage probability closer to (or greater than) the nominal level.

5. DISCUSSION

The generalized mean, as given in either (1.1) or (1.2) has not seen very much use in statistical applications, with the exception of the arithmetic, geometric and harmonic varieties. A possible reason for this is that models, leading to generalized means as estimates, have not been the object of much study.

The purpose of this article is not to advocate the use of (2.2) as a least squares criterion, or (3.1) as a population model, but rather to illustrate the consequences of doing so. In particular, we see how

these situations can lead naturally to the consideration of an estimate that is a generalized mean. Such concerns might lead an experimenter to consider an estimate of the form $[\frac{1}{n}\sum x_i^{1/2}]^2$ if the original data have been transformed by square roots. Moreover, models for which this estimate has reasonable properties can be easily constructed. If such models make statistical sense in the context of the problem at hand, the generalized mean may be a reasonable estimate.

REFERENCES

Abramowitz, M. and Stegun, I. A. (1965). *Handbook of Mathematical Functions.* New York: Dover.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society Series* B, **26**, 211-252.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families.* Hayward, CA: Institute of Mathematical Statistics.

Hardy, G. H., Littlewood, J. E. and Pólya, G. (1934). *Inequalities.* Cambridge: Cambridge University Press.

Hogg, R. V. and Craig, A. T. (1978). *Introduction to Mathematical Statistics.* New York: Macmillan.