

## Deriving Generalized Means as Least Squares and Maximum Likelihood Estimates



Roger L. Berger; George Casella

*The American Statistician*, Vol. 46, No. 4 (Nov., 1992), 279-282.

Stable URL:

<http://links.jstor.org/sici?sici=0003-1305%28199211%2946%3A4%3C279%3ADGMALS%3E2.0.CO%3B2-N>

*The American Statistician* is currently published by American Statistical Association.

---

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/astata.html>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

---

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact [support@jstor.org](mailto:support@jstor.org).

# Deriving Generalized Means as Least Squares and Maximum Likelihood Estimates

ROGER L. BERGER and GEORGE CASELLA\*

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:** Arithmetic mean; Exponential family; Geometric mean; Harmonic mean.

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

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

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

$$\begin{aligned} g_0(x_1, \dots, x_n) &= \lim_{\lambda \rightarrow 0} g_\lambda(x_1, \dots, x_n) \\ &= \exp \left( \frac{1}{n} \sum_{i=1}^n \log x_i \right) = \left( \prod_{i=1}^n x_i \right)^{1/n}, \end{aligned}$$

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) leads us to inquire about the conditions that would yield (1) as a measure of center (a mean), or the models that would yield (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) can, in fact, be generalized since each  $g_\lambda$  is of the form  $h^{-1}((1/n) \sum_{i=1}^n h(x_i))$ , where  $h(x)$  is either  $x^\lambda$  or  $\log x$ . This suggests an even more general class of means defined by

$$g_h(x_1, \dots, x_n) = h^{-1} \left( \frac{1}{n} \sum_{i=1}^n h(x_i) \right), \quad (2)$$

where  $h(x)$  is any continuous, monotone function. We shall call any function of the form (2) a *generalized mean*. Letting  $\mathbf{x}$  denote the vector  $(x_1, \dots, x_n)$ , we

shall use the notation  $g_h(\mathbf{x}) = h^{-1}(\bar{h}(\mathbf{x}))$  for the function in (2). Hardy, Littlewood, and Pólya (1934, chap. III) consider properties of these generalized means. They show that these have fundamental properties one would expect from a mean, such as satisfying

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

In this article we show how generalized means can be derived in two different ways. Section 1 shows that generalized means are least squares estimates from transformed data and, in Section 2, we see how generalized means can arise as maximum likelihood estimates, illustrating these results with examples using the arithmetic, geometric, and harmonic means. Section 3 addresses the questions of standard errors and confidence intervals, and Section 4 contains a short discussion.

## 1. LEAST SQUARES ESTIMATES

Suppose  $x_1, \dots, x_n$  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

$$\sum_{i=1}^n (x_i - a)^2, \quad (3)$$

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  $h(x_1), \dots, h(x_n)$ , 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

$$\sum_{i=1}^n (h(x_i) - h(a))^2. \quad (4)$$

But this minimizing value is  $h(a) = (1/n) \sum_{i=1}^n h(x_i) = \bar{h}(\mathbf{x})$ , so transforming back to the original scale yields  $a = h^{-1}(\bar{h}(\mathbf{x}))$ , the generalized mean of (2). Thus, the least squares estimate of center, based on the distance function in (4), 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

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

More generally, after transformation to  $h(x) = x^\lambda$ , the

\*Roger L. Berger is Professor, Statistics Department, North Carolina State University, Raleigh, NC 27695-8203. George Casella is Professor, Biometrics Unit, Cornell University, Ithaca, NY 14853-7801. This research was supported by National Science Foundation Grant DMS91-00839 and National Security Agency Grant 90F-073. Part of this work was performed at the Cornell Workshop on Conditional Inference, sponsored by the Army Mathematical Sciences Institute and the Cornell Statistics Center, Ithaca, NY, June 2-14, 1991.

least squares estimate is the generalized mean  $g_\lambda$  defined in (1).

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

$$h_\lambda(x) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log x & \lambda = 0 \end{cases} \quad (5)$$

After transformation by a member of the Box-Cox family, the least squares estimate of center is  $h_\lambda^{-1}(\bar{h}_\lambda(\mathbf{x}))$ . 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  $h^{-1}(\bar{h}(\mathbf{x})) = g^{-1}(\bar{g}(\mathbf{x}))$ . [In fact, two generalized means, defined in terms of functions  $g$  and  $h$ , are the same iff  $g$  is a linear function of  $h$ . See Hardy et al. (1934), Theorem 83.] So,  $h_\lambda^{-1}(\bar{h}_\lambda(\mathbf{x})) = h^{-1}(\bar{h}(\mathbf{x}))$  where  $h(x) = x^\lambda$ . Thus, the generalized mean  $g_\lambda$  from (1) is the least squares estimate after transformation by  $h_\lambda$  of (5).

## 2. MAXIMUM LIKELIHOOD ESTIMATES

Generalized means defined as in (2) can also be derived as maximum likelihood estimates.

*Example 3.* Let  $X_1, \dots, X_n$  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  $\theta$ , amounts to minimizing  $\sum_{i=1}^n (\log x_i - \log \theta)^2$ . This is the least squares problem that was discussed in Section 1, and the maximum likelihood estimate of  $\theta$  is the geometric mean of  $x_1, \dots, x_n$ .

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

$$f(x|\theta) = e^{\theta h(x) - H(\theta)} g(x), \quad (6)$$

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  $\theta$ .) Further, suppose that  $h$  and  $H$  are related by  $h(x) = (d/dx)H(x)$ . The log likelihood function is then

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

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

Although (6) is a form of a one-parameter exponential family, the question arises whether any exponential families exist with  $h(x) = (d/dx)H(x)$ . They do and two examples follow.

*Example 4.* Suppose  $X_1, \dots, X_n$  is a random sample from a normal population with unknown mean  $\theta$  and known variance  $\sigma^2$ . Factoring the normal density,

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

as in (6) 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  $(d/dx)H(x) = h(x)$  and  $h(x)$  is increasing, so  $\hat{\theta} = h^{-1}(\bar{h}(\mathbf{x}))$  is the maximum likelihood estimate of  $\theta$ . 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  $\theta$ .

*Example 5.* Suppose  $X_1, \dots, X_n$  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 (6) with  $h(x) = -1/x$ ,  $H(\theta) = -\log \theta$  and  $g(x) = x^{-2}$ ,  $x > 0$ . Clearly,  $h$  is increasing and  $(d/dx)H(x) = h(x)$ . So the maximum likelihood estimate of  $\theta$  is  $\hat{\theta} = h^{-1}(\bar{h}(\mathbf{x}))$ . 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 (6), 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,  $h(t) = (d/dt)H(t)$  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

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

is the moment generating function of  $Y$ . It is defined for  $\theta$  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  $h^{-1}(\bar{h}(\mathbf{x}))$  is the maximum likelihood estimate of  $\theta$  for this model.

### 3. 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  $X_1, \dots, X_n$  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  $h(X_1), \dots, h(X_n)$ . So by the central limit theorem,  $\bar{h}(\mathbf{X})$  is asymptotically normally distributed with mean  $E_\theta h(X_1)$  and variance  $\sigma_h^2/n$  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  $h(X_1), \dots, h(X_n)$ ,

$$s_h^2 = \frac{1}{n-1} \sum_{i=1}^n (h(x_i) - \bar{h}(\mathbf{x}))^2. \quad (7)$$

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

$$h^{-1}\left(\bar{h}(\mathbf{X}) - t_{\alpha/2, n-1} \frac{s_h}{\sqrt{n}}\right) < \theta < h^{-1}\left(\bar{h}(\mathbf{X}) + t_{\alpha/2, n-1} \frac{s_h}{\sqrt{n}}\right) \quad (8)$$

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

Care must be taken that the endpoints,  $\bar{h}(\mathbf{X}) \pm t_{\alpha/2, n-1} s_h / \sqrt{n}$ , 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,  $\bar{h}(\mathbf{X}) + t_{\alpha/2, n-1} s_h / \sqrt{n}$ , is positive, it should be replaced by 0. Then  $h^{-1}(0) = +\infty$  gives the correct endpoint for the

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

$$h^{-1}(\bar{h}(\mathbf{X})) \pm t_{\alpha/2, n-1} |h'(h^{-1}(\bar{h}(\mathbf{X})))|^{-1} \frac{s_h}{\sqrt{n}} \quad (9)$$

as an approximate  $100(1 - \alpha)\%$  confidence interval for  $\theta$ . 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 (8) and (9) 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  $\theta$ . 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  $\theta$ . Table 1 gives the results of the simulation.

In the lognormal case, interval (8) 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 (9) 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 interval (8) tends to have coverage probability well under the nominal value, while interval (9) has coverage closer to the nominal value. The change in center of interval (9) as

Table 1. Coverage Probabilities for the Intervals (8) and (9) When Sampling From the Lognormal and Inverted Gamma Densities

Sample Size	Estimated coverage probability			
	Lognormal		Inverted gamma	
	Interval (8)	Interval (9)	Interval (8)	Interval (9)
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

NOTE: Estimates based on 20,000 simulated samples, nominal confidence coefficient = .90. Standard errors for all estimates are approximately .0021.

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.

#### 4. DISCUSSION

The generalized mean, as given in either (1) or (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 (4) as a least squares criterion, or (6) 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  $[(1/n) \sum x_i^{1/2}]^2$  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.

[Received October 1991.]

#### 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*, Ser. 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, U.K. Cambridge University Press.
- Hogg, R. V., and Craig, A. T. (1978), *Introduction to Mathematical Statistics*, New York: Macmillan.

---

## A Lesson in Least Squares and $R$ Squared

WILLIAM BECKER and PETER KENNEDY\*

---

An exercise is presented that checks students' understanding of least squares and reminds them of problems with  $R^2$  when an intercept is not included in the regression.

**KEY WORDS:** Coefficient of determination; Regression; Zero intercept.

---

This note provides a simple exercise that we have found to be a useful learning experience for students, strengthening their understanding of least squares regression and reminding them that the usual way of calculating  $R^2$ , the coefficient of determination, is not meaningful when estimation is undertaken without an intercept.

Exhibit 1 reproduces a questionnaire answered by over 100 students in undergraduate statistics or econometrics courses, and graduate students in econometrics courses, at four universities. The formula used for the calculation of  $R^2$  is that typically used in commercial computer software. In Figure 1 of this questionnaire the four data points differ from the usual elliptical scatterplot in textbook examples in that they form a square.

---

\*William Becker is Professor of Economics, Indiana University, Bloomington, IN 47405. Peter Kennedy is Professor of Economics, Simon Fraser University, Burnaby, B.C., Canada V5A 1S6. We are grateful to several instructors at our own university and two other universities for having administered our questionnaire. They have requested anonymity, under the false belief that their students performed below the norm (overall, only about 2% of students gave correct answers to either of the last two questions).

A variant of this questionnaire, in which Figure 1 is replaced by Figure 2, is also of interest, although results from using it with approximately 100 additional students were similar to those obtained using Figure 1. Placement of the southwestern (SW) point at the origin as in Figure 1, however, has pedagogical advantages as explained below.

The first three questions were included as a check of student understanding of regression basics; we thought that almost all students would know or could figure out easily that the horizontal line, representing  $\bar{y}$ , is the regression line for these data, with a zero  $R^2$ . We were surprised that only slightly more than 50% of the students gave correct answers to these three questions. Their instructors speculated that they had been influenced by the fact that all the examples to which they had been exposed consisted of regression lines sloping upward through elliptical scatterplots.

The vast majority of students thought that the 45-degree line was the correct answer to the fourth question. Explaining why this is incorrect is a good classroom lesson in the logic of least squares. Because the regression line is constrained to pass through the origin, in Figure 1 the residuals associated with the two left-hand observations cannot be changed by the choice of regression line. Thus the answer is generated by minimizing the sum of squared residuals associated with the other two observations, accomplished by passing the regression line through the midpoint of the vertical line joining them.

This explanation is more awkward when Figure 1 is replaced by Figure 2, but students still see it as a rev-