Bernoulli **13**(1), 2007, 169–174
DOI: 10.3150/07-BEJ5166

On Gauss's characterization of the normal distribution

ADELCHI AZZALINI[1] and MARC G. GENTON[2]

[1]*Department of Statistical Sciences, University of Padua, 35121 Padova, Italy.*
E-mail: azzalini@stat.unipd.it
[2]*Department of Statistics, Texas A&M University, College Station TX 77843-3143, USA.*
E-mail: genton@stat.tamu.edu

Consider the following problem: if the maximum likelihood estimate of a location parameter of a population is given by the sample mean, is it true that the distribution is of normal type? The answer is positive and the proof was provided by Gauss, albeit without using the likelihood terminology. We revisit this result in a modern context and present a simple and rigorous proof. We also consider extensions to a p-dimensional population and to the case with a parameter additional to that of location.

Keywords: characterization property; Cauchy functional equation; location family; maximum likelihood; normal distribution; sample mean vector

1. Background and discussion

It is a commonplace that, if $x_1, \ldots, x_n$ is a random sample from a p-dimensional normal population with mean vector μ and covariance matrix Σ, $N_p(\mu, \Sigma)$ say, then the maximum likelihood estimate (MLE) of μ is given by the sample mean vector $\bar{x} = \sum_{i=1}^n x_i/n$, irrespective of the fact that Σ is known or must be estimated. The converse statement is far less obvious and familiar: if the sample mean is known to be the MLE $\hat{\mu}$ for the location parameter μ of a given parametric family of distributions, can we conclude that the population has a normal density? To put it in mathematical notation, the problem is as follows: if $x_1, \ldots, x_n$ is a random sample from a location family of p-dimensional densities $f(x-\mu)$ such that the MLE $\hat{\mu}$ coincides with the sample mean $\bar{x}$, can we conclude that $f(x)$ coincides with $\phi_p(x; \Sigma)$, the density of $N_p(0, \Sigma)$ evaluated at x, for some covariance matrix Σ? A second layer of complication arises if an additional parameter, θ say, is introduced. Typically θ regulates the covariance matrix, so that $\Sigma = \Sigma(\theta)$; other cases are possible in principle, but ruled out by the results presented later.

The origins of the problem can be traced back to Gauss (1809) himself (see also Gauss 1963), and relate directly to the argument which led to the current prominent role of the univariate normal distribution. Gauss (1809: § 177) showed that, if the sample mean is the solution to the likelihood equation (although, of course, this terminology was not used by him) for all possible samples and possible values of n, and the score function is continuous, then the parent distribution is the normal. See also the accounts of Hald (1998: 354–355)

1350–7265 © 2007 ISI/BS

and Chatterjee (2003: 225–227) that illustrate and complement the original argument; in particular, Chatterjee shows that the argument goes through if the required property holds for sample sizes $n = 1, 2, 3$ only.

Among the few papers that have explored problems of this sort further, there is the classical work of Teicher (1961) which considers characterizations of distributions via MLE when $p = 1$; the normal case with a location parameter is clearly of most interest. Teicher weakened the conditions required by Gauss in two ways: (i) differentiability of f is not required and can be replaced by lower semi-continuity; (ii) $\bar{x}$ must be the MLE for samples of size $n = 2$ and 3 only. Teicher's (1961) result was extended to the p-dimensional setting by Marshall and Olkin (1993), although the additional condition that $\log f(x)$ admits a gradient and that this gradient must satisfy some minimal regularity condition (in order to invoke a linear solution of the Cauchy functional equation involved in the proof) would seem necessary for a rigorous statement of the theorem. Stadje (1993) studied this characterization problem as well, but with the above requirement for $n = 2, 3, 4$ simultaneously.

In addition to the papers quoted above, characterizations of distributions via MLE were considered by Hürlimann (1998) following a very different sort of argument. His approach is in a sense more general since it applies to any location–scale family, but is limited to the case $p = 1$.

The argument developed in the next section deals with the problem described above. While the proof shares some features with others already developed by the aforementioned authors, it has the advantage of requiring that $\hat{\mu} = \bar{x}$ for only one value of n, provided $n \geqslant 3$, in place of making this requirement for $n = 2$ and $n = 3$ simultaneously. Compared to Teicher's (1961) argument, the present proof has more stringent regularity conditions, but it relaxes the condition on the sample size, and it seems to us somewhat simpler; in addition, the present proof holds for general dimension p and allows the presence of a parameter in addition to μ.

To preserve the simplicity of the proof in its essential form, we first present the argument for the basic case, and then add levels of complication in subsequent formulations of increasing generality.

2. Characterization

2.1. Univariate location family

Theorem 1. *Consider a parametric location family for a one-dimensional continuous random variable, such that for any choice of $\mu \in \mathbb{R}$ the corresponding probability density function at the point $x \in \mathbb{R}$ is $f(x - \mu)$. Assume that a random sample of size $n \geqslant 3$ is drawn from a member of this parametric family, and that the following conditions hold:*

- (i) *$f(x)$ is a differentiable function of x and its derivative $f'(x)$ is continuous at least at one point $x \in \mathbb{R}$;*
- (ii) *for each set of sample values, $x_1, \ldots, x_n$, the sample mean $\bar{x} = \sum_{i=1}^{n} x_i / n$ is a solution of the likelihood equation for the location parameter μ.*

Then $f(x)$ is the one-dimensional normal density $\phi(x;\, \sigma^2)$ for some positive σ^2.

Proof. Equating to 0 the score function for μ derived from the log-likelihood function associated with the sample $x_1, \ldots, x_n$, we obtain

$$\sum_{i=1}^{n} g(x_i - \mu) = 0, \tag{1}$$

where

$$g(x) = \frac{\mathrm{d}}{\mathrm{d}x} \log f(x). \tag{2}$$

Notice that we allow $f(x) = 0$ and in that case adopt the common convention that $\log 0 = -\infty$. However, since $f(x) > 0$ must hold true for a range of x values, then (2) must correspondingly be finite for a range of values of μ and the solution to (2) is to be searched for in this set. Under condition (ii) of the theorem, we have

$$\sum_{i=1}^{n} g(x_i - \bar{x}) = 0$$

for all possible choices of the sample values.

Consider a sample with $x_1 = \ldots = x_n = u$ for some constant u. Then (1) becomes

$$n\, g(u - \mu) = 0,$$

whose solution is $\hat{\mu} = u$ under the assumption of the theorem; hence $g(0) = 0$.

Consider now the sample $2u$, 0, $u, \ldots, u$. Then again $\hat{\mu} = u$ and

$$g(u) + g(-u) + (n-2)g(0) = 0,$$

which implies $g(-u) = -g(u)$, that is, that g is antisymmetric.

For any two points u and v, consider further the sample u, v, $-(u+v)$, $0, \ldots, 0$ such that $\hat{\mu} = 0$; therefore

$$g(u) + g(v) = g(u + v). \tag{3}$$

This is the celebrated Cauchy functional equation which has been discussed extensively in the mathematical literature; see, in particular, Aczél and Dhombres (1989) under the assumption of continuity at a point for $g(\cdot)$. The solution to equation (3) is of the type

$$g(x) = -cx,$$

for some $c \in \mathbb{R}$ not depending on x; the minus sign is inserted for mere notational convenience. From (2), the corresponding integral function is of the form

$$\log f(x) = d - \frac{1}{2}\, c\, x^2,$$

for some real constant d. It is implicit that $c > 0$, otherwise $f(x)$ would not be integrable over the real line. After the constant d is suitably adjusted so that f integrates to 1 over $\mathbb{R}$, we obtain the normal density $\phi_1(x;\, 1/c)$ for f, where $c > 0$. This completes the proof. □

Remark 1. If the requirements of Theorem 1 refer to the case $n = 2$ instead of some $n \geqslant 3$, then the conclusion does not hold. A counter-example is given by the density function

$$f(x) = \text{constant} \times \mathrm{e}^{-x^2/2+w(x)}, \qquad x \in \mathbb{R},$$

where $w : \mathbb{R} \to \mathbb{R}$ is an even function. Equation (1) becomes

$$\sum_{i=1}^{n}(x_i - \mu) = \sum_{i=1}^{n} w'(x_i - \mu),$$

where the derivative w' is an odd function. If $n = 2$, the value $\mu = \bar{x}$ is always a solution to this equation, even if $f(x)$ is not normal. The choice $w(x) = \cos(x)$ is particularly appropriate because it is then obvious that the log-likelihood function is log-concave.

2.2. Multivariate location family

Theorem 2. *Consider a parametric family for a p-dimensional continuous random variable, depending on a location parameter $\mu \in \mathbb{R}^p$, such that the corresponding probability density function at point $x \in \mathbb{R}^p$ is $f(x - \mu)$. Assume that a random sample of size $n \geqslant 3$ is drawn from a member of this parametric family, and that the following conditions hold:*

(i) *$f(x)$ admits partial derivatives with respect to the p components of x and the gradient is continuous at least at one point $x \in \mathbb{R}^p$;*

(ii) *for each set of sample p-dimensional vectors, $x_1, \ldots, x_n$, the sample mean vector $\bar{x} = \sum_{i=1}^{n} x_i/n$ is a solution of the likelihood equation for the location parameter μ.*

Then $f(x)$ is the p-dimensional normal density $\phi_p(x; \Sigma)$ for some positive definite matrix Σ.

Proof. Since the proof is largely the same as that of Theorem 1, we only highlight the differences. Once the definition (2) of the derivative $g(x)$ is replaced by the gradient formed by the p partial derivatives with respect to the components of x, the proof proceeds as before up to equation (3), where the quantities u, v are now intended as p-dimensional vectors.

Under the condition of continuity of $g(x)$ at one point, the solution to the vector equation (3) is still of linear type; see, for instance, Aczél and Dhombres (1989). Write the solution as

$$g(x) = -Cx,$$

where C is a $p \times p$ constant matrix. On integrating this expression, we have

$$\log f(x) = d - \tfrac{1}{2}x^{\mathrm{T}}Cx$$

where d is some real constant.

There is no loss of generality in assuming that C is symmetric since $x^{\mathrm{T}}Cx = (x^{\mathrm{T}}Cx)^{\mathrm{T}}$, and it is implicit that $C \geqslant 0$. Otherwise, if there were a point $x_0 \in \mathbb{R}^p$ such that $x_0^{\mathrm{T}}Cx_0 < 0$, then the same would hold for all multiples αx_0 for any real α. Combining this fact with continuity of $x^{\mathrm{T}}Cx$, there would be a cone where the density has arbitrarily high values, against the hypothesis of integrability.

After the constant d is suitably adjusted so that f integrates to 1 over $\mathbb{R}^p$, we obtain the multivariate normal density $\phi_p(x; C^-)$ for f, where C^- is a generalized inverse of C. However, if C were not of full rank, the density of the resulting singular normal distribution would lie on a linear subspace of $\mathbb{R}^p$ and would not exist over the entire space $\mathbb{R}^p$, since the probability mass outside the aforementioned linear space must be 0; see, for instance, Rao (1973: 527–528). Since this would contradict the hypothesis of continuous differentiability of f at one point at least, we conclude that C^{-1} exists, and f is the density of $\phi_p(x; \Sigma)$, with $\Sigma = C^{-1}$. This completes the proof. □

Remark 2. Here again, if the requirements of Theorem 2 refer to the case $n = 2$ instead of some $n \geqslant 3$, then the conclusion does not hold. A counter-example similar to that in Remark 1 is given by the density function

$$f(x) = \text{constant} \times \mathrm{e}^{-x^{\mathrm{T}}x/2+w(x)}, \qquad x \in \mathbb{R}^p,$$

where $w : \mathbb{R}^p \to \mathbb{R}$ satisfies $w(-x) = w(x)$ for all $x \in \mathbb{R}^p$.

2.3. Multivariate location family with additional parameters

Theorem 3. *Consider a parametric family for a p-dimensional continuous random variable, depending on a location parameter $\mu \in \mathbb{R}^p$, and possibly an additional parameter $\theta \in \Theta$, Θ an open subset of $\mathbb{R}^q$, such that the corresponding probability density function at point $x \in \mathbb{R}^p$ is $f(x - \mu; \theta)$. Assume that a random sample of size $n \geqslant 3$ is drawn from a member of this parametric family, and that the following conditions hold:*

(i) *for any fixed θ, $f(x; \theta)$ admits partial derivatives with respect to the p components of x and the gradient is continuous at least at one point $x \in \mathbb{R}^p$, and $f(x; \theta)$ admits partial derivative with respect to θ;*

(ii) *for each set of sample p-dimensional vectors, $x_1, \ldots, x_n$, and for each choice of θ, the sample mean vector $\bar{x} = \sum_{i=1}^n x_i/n$ is a solution of the likelihood equation for the location parameter μ.*

Then $f(x; \theta)$ is the p-dimensional normal density $\phi_p(x; \Sigma)$ for some positive definite matrix Σ that depends on θ.

Proof. If θ is known, then the argument is exactly as in the proof of Theorem 2. In the case where θ is not known, there is of course a second likelihood equation. However, the argument of Theorem 2 for the first equation still carries over, since it works for any fixed θ, and we still obtain that $f(x; \theta) = \phi_p(x; \Sigma)$ where $\Sigma = C^{-1}$ depends on θ. This completes the proof. □

Remark 3. Under the assumptions of Theorem 3, the MLE $\hat{\theta}$ of θ has to maximize the profile log-likelihood

$$\ell(\theta) = \text{constant} - \frac{n}{2}\log|\Sigma(\theta)| - \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^{\mathrm{T}}\Sigma(\theta)^{-1}(x_i - \bar{x}).$$

An interesting special case is when the parameter θ has dimension $q = p(p+1)/2$ and represents the non-duplicated elements of Σ, under the constraint that $\Sigma = \Sigma(\theta)$ is positive definite. It follows from Theorem 3 that $\Sigma(\hat{\theta})$ in this setting must be the sample covariance matrix. In particular, when $p = 1$ and $f(x - \mu; \theta) = f[(x - \mu)/\theta]/\theta$, a location–scale family with $\theta > 0$, the implication is that the MLE of the scale parameter θ must be the square root of the sample variance.

Acknowledgements

We are grateful to Bibhuti Bhattacharyya, Werner Ehm, Alessandro Languasco, Arthur Pewsey, Giuliana Regoli and an earlier anonymous reviewer for various insightful comments, including the reviewer's suggestion of a better counter-example than our original one. This work was started while the first author was visiting the Department of Statistics, Texas A&M University, USA, whose kind hospitality is gratefully acknowledged. The work of Genton was partially supported by National Science Foundation grant DMS-0504896.

References

Aczél, J. and Dhombres, J. (1989) *Functional Equations in Several Variables with Applications to Mathematics, Information Theory and to the Natural and Social Sciences*, Encyclopedia Math. Appl. 31. Cambridge: Cambridge University Press.

Chatterjee, S.K. (2003) *Statistical Thought: A Perspective and History.* Oxford: Oxford University Press.

Gauss, C.F. (1809) *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium.* Hamburg: Perthes et Besser. English translation by C.H. Davis, reprinted by Dover, New York (1963).

Gauss, C.F. (1963) Theory of the motion of the heavenly bodies moving about the sun in conic sections (trans. C.H. Davis). New York: Dover.

Hald, A. (1998) *A History of Mathematical Statistics from 1750 to 1930.* New York: Wiley.

Hürlimann, W. (1998) On the characterization of maximum likelihood estimators for location-scale families. *Comm. Statist. Theory Methods*, **27**, 495–508.

Marshall, A.W. and Olkin, I. (1993) Maximum likelihood characterizations of distributions. *Statist. Sinica*, **3**, 157–171.

Rao, C.R. (1973) *Linear Statistical Inference and its Applications*, 2nd edition. New York: Wiley.

Stadje, W. (1993) ML characterization of the multivariate normal distribution. *J. Multivariate Anal.*, **46**, 131–138.

Teicher, H. (1961) Maximum likelihood characterization of distributions. *Ann. Math. Statist.*, **32**, 1214–1222.

Received November 2005 and revised April 2006