A Class of Distributions Which Includes the Normal Ones
Author(s): A. Azzalini
Source: *Scandinavian Journal of Statistics*, Vol. 12, No. 2 (1985), pp. 171-178
Published by: Wiley on behalf of Board of the Foundation of the Scandinavian Journal of Statistics
Stable URL: https://www.jstor.org/stable/4615982
Accessed: 24-05-2019 17:04 UTC

A Class of Distributions which Includes the Normal Ones

A. AZZALINI
University of Padua

ABSTRACT. A new class of density functions depending on a shape parameter λ is introduced, such that $\lambda=0$ corresponds to the standard normal density. The properties of this class of density functions are studied.

Key words: normal distribution, skewness, chi-square, Bayes theorem, Kalman filter

1.0. Introduction

Many families of density functions approach the normal one as a certain parameter tends to an appropriate value. However, there are only a few parametric classes of distributions which include the normal one as a proper member, and not only as a limit case, while in various situations it is useful to deal with a parametric class which allows a "continuous" variation from normality to non-normality.

Moreover, among the classes which own this "strict inclusion" property, some are not easily mathematically tractable, while some others appear as "not natural". For instance, Tukey's g- and h-distributions (for a review, see Hoaglin, 1982) do not have a simple expression for the density. On the other hand, the Edgeworth and the Gram–Charlier expansions appear as approximations to some "real" densities, rather than being densities of their own.

Among the families which enjoy the "strict inclusion" property we mention the so-called power distribution, whose density is proportional to

$$\exp(-|y|^r/r) \tag{1}$$

(Box, 1953; Turner, 1960; Vianelli, 1963). The parameter r manoeuvres the kurtosis of the density, but the skewness is always 0.

Prentice (1975) has studied a wide parametric class of densities which, under suitable parametrization, contains the $N(0, 1)$ one. Although the standard normal density is again a boundary point, this class behaves fairly regularly in the neighbourhood of the $N(0, 1)$ density, at least for the purpose of discrimination among parametric families. Therefore, in some respects, Prentice's class "may be regarded as a natural generalization of the normal model". However, there are some features of the normal distribution which Prentice's class does not reproduce, while they are included in the class considered in the present paper.

It would be ideal to have at hand a class of densities with the following properties:

(i) "strict inclusion" of the normal density,
(ii) mathematical tractability,
(iii) wide range of the indices of skewness and kurtosis.

In section 2, we study the mathematical properties of a one-parameter class of density functions, which fulfils the first two of the above requirements. Section 3 deals with its statistical aspects, and in subsection 3.3 an extension of the previous class is considered. Since the class of subsection 3.3 has two shape parameters, it meets to some extent requirement (iii). This two-parameter class has been considered by O'Hagan and Leonard (1976), of which the author became aware only afterwards; however, only a limited number of properties have been given by O'Hagan and Leonard. Section 4 hints at other possible generalizations.

2.0. The class of skew-normal density functions

2.1. Definition and some simple properties

Lemma 1. *Let f be a density function symmetric about 0, and G an absolutely continuous distribution function such that G′ is symmetric about 0. Then*

$$2G(\lambda y)\,f(y) \qquad (-\infty<y<\infty) \tag{2}$$

is a density function for any real λ.

Proof. Let Y and X be independent random variables with density f and G', respectively. Then

$$½=P\{X-\lambda Y<0\}=E_Y(P\{X<\lambda y\,|\,Y=y\})=\int_{-\infty}^{\infty} G(\lambda y)\,f(y)\,dy.$$

An acceptance–rejection technique which generates a random variable Z with density (2) is the following one. Sample X and Y from G' and f, respectively. If $X<\lambda Y$, then put $Z=Y$, otherwise restart sampling a new pair of variables X and Y, until the inequality $X<\lambda Y$ is satisfied. On average, two pairs (X, Y) are necessary to produce Z.

Definition. If a random variable Z has density function

$$\phi(z;\lambda)=2\phi(z)\,\Phi(\lambda z) \qquad (-\infty<z<\infty) \tag{3}$$

where ϕ and Φ are the standard normal density and distribution function, respectively, then we say that Z is a *skew-normal* random variable with parameter λ; for brevity we, shall also say that Z is SN(λ).

Because of lemma 1, (3) is a proper density function. Strictly speaking, we are using the symbol ϕ with two different meanings; however, this is justified by Property A below. For the rest of the paper, Z will denote a random variable with density function (3). Fig. 1 shows the shape of (3) for two values of λ. The following properties follow immediately from the definition.

Property A. *The SN(0) density is the N (0, 1) density.*

Property B. *As $\lambda\to\infty$, $\phi(z;\lambda)$ tends to the half-normal density.*

Property C. *If Z is a SN(λ) random variable, then $-Z$ is a SN($-\lambda$) random variable.*

Property D. *The density (3) is strongly unimodal, i.e. log $\phi(z;\lambda)$ is a concave function of z.*

2.2. Distribution function

Denote by $\Phi(z;\lambda)$ the distribution function of (3), i.e.

$$\Phi(z;\lambda)=2\int_{-\infty}^{z}\int_{-\infty}^{\lambda t} \phi(t)\,\phi(u)\,du\,dt.$$

The region of integration is denoted by A in Fig. 2, which refers to positive z and λ. We recall the function $T(h, a)$ studied by Owen (1956). For positive h and a, $T(h, a)$ gives the integral of the standard normal bivariate density over region bounded by the lines $x=h$, $y=0$, $y=ax$ in the

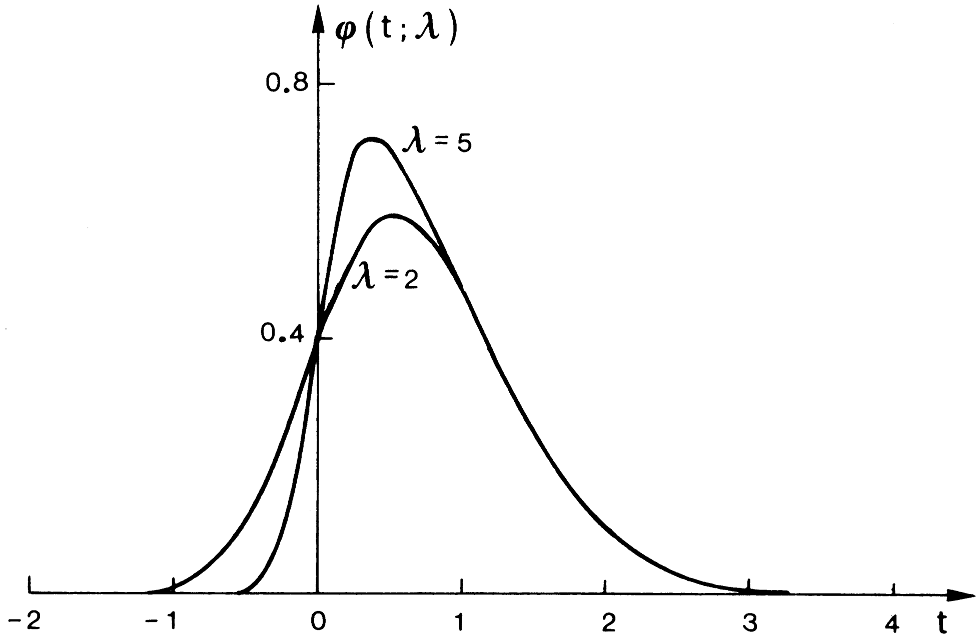

Fig. 1. The density functions SN(2) and SN(5).

(x, y) plane. Then

$$\Phi(z;\lambda)=\Phi(z)-2T(z,\lambda), \tag{4}$$

since $T(z, \lambda)$ is the integral over region B of Fig. 2.

It is known that $T(h, a)$ is a decreasing function of h and

$$-T(h, a)=T(h, -a), \qquad T(-h, a)=T(h, a), \qquad 2T(h, 1)=\Phi(h)\,\Phi(-h).$$

Therefore, taking into account the properties of $T(h, a)$ and inspecting Fig. 2, one sees that (4) holds also for negative values of z and λ, i.e. (4) is the general expression of the distribution

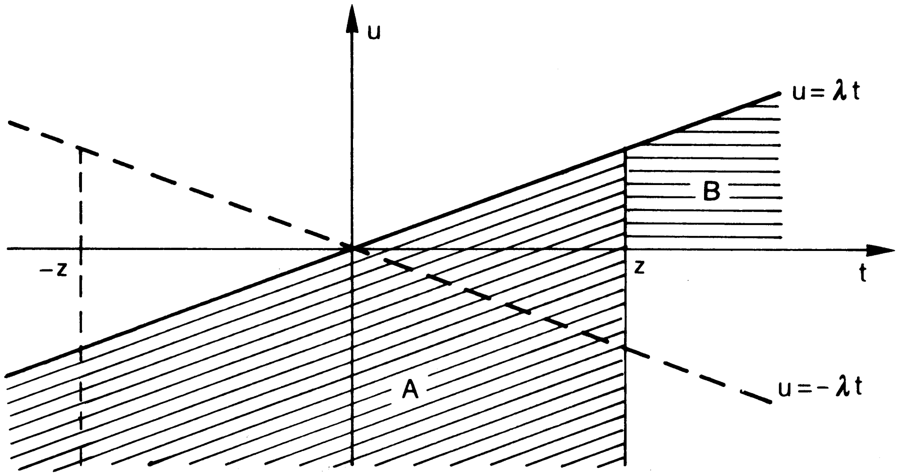

Fig. 2. The integration region for $\Phi(z, \lambda)$.

function of (3). A computer routine which evaluates $T(h, a)$ has been given by Young & Minder (1974); see also Hill (1978) and Thomas (1979).

From Property C and the above properties of the function $T(h, a)$, we get the following corollaries.

Property E. $1-\Phi(-z;\lambda)=\Phi(z;-\lambda)$.

Property F. $\Phi(z;1)=\{\Phi(z)\}^2$.

Property G. $\sup_z |\Phi(z)-\Phi(z;\lambda)|=\pi^{-1}\arctan|\lambda|$.

Property H. *If Z is SN*(λ), *then* Z^2 *is* χ^2_1.

2.3. *Moments*

Because of Property H, the even moments of Z are equal to the even moments of the standard normal distribution. For computing the odd moments, we use the next well-known result; see, for instance, Zacks (1981, pp. 53–54).

Lemma 2. *If U is a N (0, 1) random variable, then*

$$E\{\Phi(hU+k)\}=\Phi\{k/\surd(1+h^2)\}$$

for any real h, k.

Using this lemma, we have at once that the moment generating function of Z is

$$M(t)=2\exp(t^2/2)\,\Phi(\delta t)$$

where

$$\delta=\lambda/\surd(1+\lambda^2). \qquad (5)$$

Hence, after some algebra, we obtain

$$E(Z)=b\delta,$$

$$\mathrm{var}\,(Z)=1-(b\delta)^2,$$

$$\gamma_1(Z)=\tfrac{1}{2}(4-\pi)\,\mathrm{sign}\,(\lambda)\left[\frac{\{E(Z)\}^2}{\mathrm{var}\,(Z)}\right]^{3/2},$$

$$\gamma_2(Z)=2(\pi-3)\left[\frac{\{E(Z)\}^2}{\mathrm{var}\,(Z)}\right]^{2}$$

where

$$b=\sqrt{\frac{2}{\pi}}$$

and γ_1, γ_2 denote the third and fourth standardized cumulants. The maximum value of γ_1 is about 0.995, while for γ_2 it is 0.869.

It is of some interest to note that, considering log $M(t)$ and expanding log $\{2\Phi(\delta t)\}$ around 0, the rth cumulant of Z, for $r>2$, is seen to be proportional to δ^r, although it is not easy to obtain a closed form expression for the coefficient of δ^r.

If U is a $N(0, 1)$ random variable independent of Z, then the moment generating function of $U+Z$ is $2\exp(t^2)\,\Phi(\delta t)$. Hence, after rescaling, we have the following corollary.

Property I. *If Z is $SN(\lambda)$ and U is an independent $(0, 1)$ random variable, then $(Z+U)/\sqrt{2}$ is* $\mathrm{SN}(\lambda/\sqrt{(2+\lambda^2)})$.

3.0. Statistical aspects

3.1. Likelihood and Fisher information

In practice, one will often work with the family of distributions generated by the linear transformation

$$Y=\lambda_1+\lambda_2 Z \qquad (\lambda_2>0). \tag{6}$$

The Fisher information for the parameter $(\lambda_1, \lambda_2, \lambda)$ is easily computed, obtaining

$$I_\lambda=\begin{pmatrix} (1+\lambda^2 a_0)/\lambda_2^2 & \left(E(Z)\dfrac{1+2\lambda^2}{1+\lambda^2}+\lambda^2 a_1\right)\Big/\lambda_2^2 & \left(\dfrac{b}{(1+\lambda^2)^{3/2}}-\lambda a_1\right)\Big/\lambda_2 \\ \left(E(Z)\dfrac{1+2\lambda^2}{1+\lambda^2}+\lambda^2 a_1\right)\Big/\lambda_2^2 & (2+\lambda^2 a_2)/\lambda_2^2 & -\lambda a_2/\lambda_2 \\ \left(\dfrac{b}{(1+\lambda^2)^{3/2}}-\lambda a_1\right)\Big/\lambda_2 & -\lambda a_2/\lambda_2 & a_2 \end{pmatrix}$$

where

$$a_k=a_k(\lambda)=E\left\{Z^k\left(\frac{\phi(\lambda Z)}{\Phi(\lambda Z)}\right)^2\right\} \qquad (k=0, 1, 2),$$

which has to be evaluated numerically.

A peculiar aspect of the above information matrix is that it becomes singular as $\lambda\to 0$, although all three parameters are still identifiable. An informal explanation for this phenomenon is as follows. Near $\lambda=0$, the order of magnitude of λ is $|\gamma_1|^{1/3}$. While γ_1 can be estimated with the usual variance rate $1/n$, if n is the sample size, the variance of the estimate of λ has a higher order of magnitude. Since $\delta\approx\lambda$ near $\lambda=0$ and $E(Y)=\lambda_1+\lambda_2 b\delta$, then not even λ_1 can be estimated with variance rate $1/n$.

Because of these considerations, the singularity problem can be avoided by re-parametrizing with $(\theta_1, \theta_2, \gamma_1)$ where

$$Y=\theta_1+\theta_2\left(\frac{Z-E(Z)}{\sqrt{\mathrm{var}\,(Z)}}\right). \tag{7}$$

For these new parameters the information matrix is

$$I_\theta=D'I_\lambda D$$

where D is the matrix of derivatives of the old parameters with respect to the new ones. As $\lambda\to 0$, I_θ converges to diag $(\theta_2^2, \theta_2^2/2, 6)$.

3.2. Maximum likelihood estimation

The likelihood equations for a simple random sample of size n from Y, defined by (6), are readily written down. For any fixed λ, the equations for λ_1, λ_2 have a unique solution simply

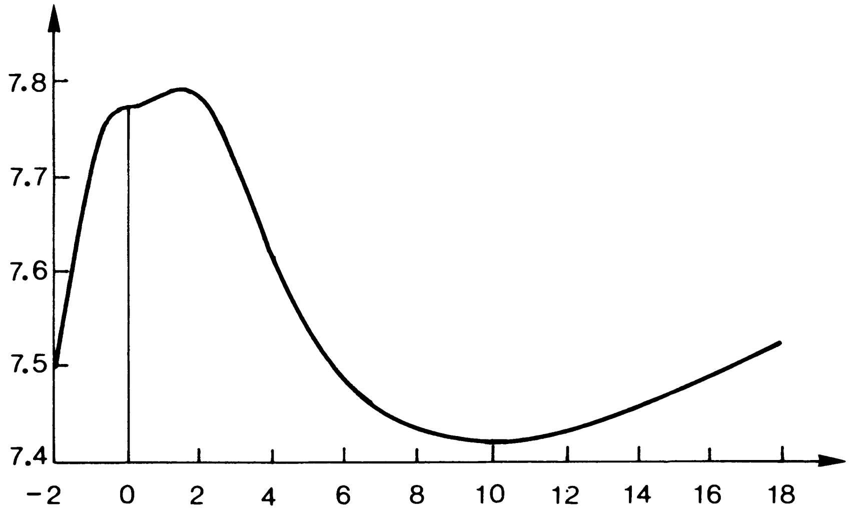

Fig. 3. The profile log likelihood for a sample of size 10.

because of the log-concavity of (3) (Property D), as remarked by Burridge (1981); in fact this uniqueness property holds for the more general model

$$y_j=x_j'\beta+\lambda_2 z_j \qquad (j=1, \ldots, n)$$

where x_j is a p-vector of covariates, β is a p-dimensional parameter and $z_1, \ldots, z_n$ are independent and identically distributed SN(λ) random variables.

For the three parameter model (6), it is easy to check that the maximum likelihood estimate $\hat\lambda_1, \hat\lambda_2$ of λ_1, λ_2 satisfy the relationship

$$\hat\lambda_2^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat\lambda_1)^2 \tag{8}$$

which matches a well-known fact for normal random variables.

Therefore, in practice, a convenient way of proceeding is the following one. For a fixed value of λ, solve the two equations for (λ_1, λ_2) taking into account (8); repeat these steps for a reasonable range of values of λ.

Some care is necessary in dealing with the profile log-likelihood of λ. A stationarity point always exists at $\lambda=0$. Moreover, for small samples, other stationarity points may occur, as illustrated by Fig. 3 which refers to a set of $n=10$ values. For very large values of λ the log-likelihood exceeds the local maximum at $\lambda=1.48$; however, this abnormality has not been observed for largish samples. As an overall conclusion on the basis of some numerical experience of the author, visual inspection of the profile likelihood is recommended, at least if n is small, say $n<30$.

3.3. *An extended class of densities and Bayes theorem*

A well-known property of the normal distribution is that, if Y is $N(W, \sigma^2)$ where *a priori* W is a normal random variable, then the *a posteriori* distribution W is still normal.

An analogous fact is true if *a priori* W has probability density function in t

$$\phi(t;\lambda_1,\lambda_2,\lambda,\xi)=\phi\left(\frac{t-\lambda_1}{\lambda_2}\right)\Phi\left\{\lambda\left(\frac{t-\lambda_1}{\lambda_2}\right)+\xi\right\}\bigg/\left\{\lambda_2\,\Phi\left(\frac{\xi}{\sqrt{(1+\lambda^2)}}\right)\right\} \tag{9}$$

which is a proper density function because of lemma 2. Then, some simple algebra shows that the *a posteriori* density function of W given that $Y=y$ is still of type (9) with $(\lambda_1, \lambda_2, \lambda, \xi)$ replaced by

$$\frac{y/\sigma^2+\lambda_1/\lambda_2^2}{1/\sigma^2+1/\lambda_2^2},\qquad (1/\sigma^2+1/\lambda_2^2)^{-1/2},$$

$$\lambda(1+\lambda_2^2/\sigma^2)^{-1/2},\qquad \xi+(y-\lambda_1)\,\frac{\lambda_2\lambda}{\sigma^2+\lambda_2^2}.$$

Note that the parameter λ shrinks towards 0, independently of y, and that the updating formulae of the first two parameters are the same of the standard normal case.

Consider now the Kalman filter setting

$$\begin{aligned} W_t&=\varrho W_{t-1}+\varepsilon_t \\ Y_t&=W_t+\eta_t \end{aligned}\qquad (t=1, 2, \ldots)$$

where $\{\varepsilon_t\}$ is white noise $N(0, \sigma_\varepsilon^2)$ and $\{\eta_t\}$ is white noise $N(0, \sigma_\eta^2)$, with $\{\varepsilon_t\}$ independent of $\{\eta_t\}$. If the initial prior of W_0 is normal, then all other conditional distributions of W_t given $(Y_1, \ldots, Y_t)$ are still normal. An analogous property holds for distribution (9): if the initial prior distributions of W_0 is of type (9), then all subsequent posterior distributions of W_t given $(Y_1, \ldots, Y_t)$ are again of type (9). This fact follows from the above conjugacy property of (9) with the normal distribution plus the following extension of Property I, which can be obtained by simple algebra.

Property I′. *If W is a random variable with density function (9) amd U is an independent $N(0, \sigma^2)$ variable, then $W+U$ has density function of type (9) with the same λ_1, and λ_2, λ, ξ replaced by*

$$(\sigma^2+\lambda_2^2)^{1/2},\qquad \lambda\left(1+(1+\lambda^2)\frac{\sigma^2}{\lambda_2^2}\right)^{-1/2},\qquad \xi\left(\frac{\sigma^2+\lambda_2^2}{\sigma^2(1+\lambda^2)+\lambda_2^2}\right)^{1/2},$$

respectively.

Using lemma 2 again, the moment generating function of (9) with $\lambda_1=0$, $\lambda_2=1$ is immediately written down, namely

$$\exp\,(t^2/2)\,\Phi\left(\frac{\lambda t+\xi}{\sqrt{(1+\lambda^2)}}\right)\bigg/\Phi\left(\frac{\xi}{\sqrt{(1+\lambda^2)}}\right).$$

From the last expression, the first four cumulants of (9) have been computed, but their expression is not reported here. However, γ_1 ranges between about -1.2 and 1.2, while γ_2 varies between 0 and about 2. Only a small subset of all (γ_1, γ_2) pairs in this rectangle is actually feasible, mainly because when $\gamma_1=0$ (i.e. $\lambda=0$) also $\gamma_2=0$.

4.0. Further generalizations

We have already mentioned an extension of (3) in section 3.3. We sketch some further possibilities.

Denote by $\phi_k(y; \mathrm{P})$ the k-dimensional normal density function with standardized marginals and correlation matrix $\mathrm{P}=((\varrho_{ij}))$. A multivariate extension of (3) is

$$c\phi_k(y;\,\mathrm{P})\prod_{j=1}^{k}\Phi(\lambda_j y_j)$$

where $y=(y_1, \ldots, y_k)' \in R^k$, $\lambda_1, \ldots, \lambda_k$ are k real numbers and $1/c$ is the orthant probability of a standardized normal random variable whose off-diagonal elements of the correlation matrix are of the form $\delta_i\delta_j\varrho_{ij}$ with δ_i defined similarly to (5). The fact that c is the correct normalization constant follows from an obvious generalization of lemma 2.

It has been suggested that the power density (1) be used as a reference density for tackling robustness problems, since the parameter r acts on the weight of the tails. Multiplication of (1) by $2G(\lambda y)$ for some symmetric distribution function G allows for different weights of the tails.

Acknowledgement

I thank the referees for various improvements to the first draft of the paper.

References

Box, G. E. P. (1953). A note on regions of kurtosis. *Biometrika* **40**, 465–468.

Burridge, J. (1981). A note on maximum likelihood estimation for regression models using grouped data. *J. R. Statist. Soc. B* **43**, 41–45.

Hill, I. D. (1978). Remark ASR 26. *Appl. Statist* **27**, 379.

Hoaglin, P. (1982). *g*- and *h*-distributions. In *Encyclopedia of statistical sciences* (ed. N. Johnson & S. Kotz), Wiley, New York.

O'Hagan, A. & Leonard, T. (1976). Bayes estimation subject to uncertainty about parameters constraints. *Biometrika* **63**, 201–202.

Owen, D. B. (1956). Tables for computing bivariate normal probabilities. *Ann. Math. Statist.* **27**, 1075–1090.

Prentice, R. L. (1975). Discrimination among some parametric models. *Biometrika* **62**, 607–614.

Thomas, G.E. (1979). Remark ASR 30. *Appl. Statist.* **28**, 113.

Turner, M. E. (1960). On heuristic estimation methods. *Biometrics* **16**, 299–401.

Vianelli, S. (1963). La misura di variabilità condizionata in uno schema generale delle curve normali di frequenza. *Statistica* **23**, 447–473.

Young, J. C. & Minder, Ch. E. (1974). Algorithm AS 76: an integral useful in computing non-central *t* and bivariate normal probabilities. *Appl. Statist.* **23**, 455–457.

Zacks, S. (1981). *Parametric statistical inference.* Pergamon Press, Oxford.

Received September 1984, in final form February 1985

A. Azzalini, Università di Padova, Dipartimento di Scienze Statistiche, Via S. Francesco 33, 35121 Padova, Italy