REVISED VERSION

WHEN IS A DISTRIBUTION DETERMINED BY ITS LETTER VALUES?

Rameshwar D. Gupta*
Division of Mathematics,
Computer Science & Engineering
University of New Brunswick
St. John, New Brunswick, Canada

Donald St. P. Richards**
Department of Statistics
University of North Carolina
Chapel Hill, N.C. 27514

ABSTRACT

Suppose that the differences $x_{k+1}-x_k$ between the successive letter values of a continuous symmetric distribution form a geometric progression. Under very simple regularity assumptions it is shown that the distribution is uniquely determined. As an application, a statistic based only on the letter values is constructed for testing the hypothesis that a data set is drawn from any given, completely specified distribution.

AMS 1980 Subject Classification: Primary 62G10, Secondary 62G25.

Key words and phrases: Letter values, hypothesis testing exploratory data analysis, Laplace distribution, uniform distribution.

*Partially supported by the National Scientific and Engineering Research Council, grant no. A-4850.
**Partially supported by a grant from the Research Council of the University of North Carolina.

1. INTRODUCTION

In recent years, exploratory methods for analyzing data have become widespread among statisticians. Witnesses to this trend include the books [1], [2] and the large number of references therein. There are several reasons for the popularity of the new methods and we shall mention one of the most important. For exploratory purposes, many of the simple summary statistics for a batch of data are computed using sorting and counting rules. Consequently, these summaries can be quite robust; that is, a large variation in a small portion of the data set causes only a small change in the value of the summary statistic. In general, exploratory summaries are particularly useful for investigators who need to quickly ascertain important information about data sets, such as the presence of outliers.

On the other hand, classical summaries such as the sample mean and variance, while markedly nonrobust, can be used to identify the distribution underlying sets of data. For example, if the mean of a sample of independent observations is normally distributed then it is well known that the parent population is necessarily normal. Similar results for exploratory summaries are rare. Indeed, the purpose of this paper is to show that one class of exploratory statistics, the letter values, may be used to identify the parent population for certain continuous distributions. As an important consequence, we show that the letter values may also be used to test the hypothesis that a data set is drawn from a given, completely specified distribution.

2. LETTER VALUES

Table: Relationship between letter values and tail areas for continuous distributions

Label	Tag	Tail Area
x_0	M	$1/2$
x_1	F	$1/2^2$
x_2	E	$1/2^3$
x_3	D	$1/2^4$
x_4	C	$1/2^5$

If F(x) is an absolutely continuous distribution function, then the k-th upper letter value x_k is the unique solution of the equation

$$1 - F(x_k) = 2^{-(k+1)} , \quad k=0,1,2,... \tag{1}$$

Thus, x_0 is the median, x_1 is called the upper fourth, x_2 the upper eighth, etc. Similarly the k-th lower letter value x_{-k} is the unique solution of the equation

$$F(x_{-k}) = 2^{-(k+1)} , \quad k=0,1,2,... \tag{2}$$

In analogy with the upper fourth, x_{-1} is called the lower fourth, etc. The tags M (denoting median), F (denoting fourths), E (denoting eighths), etc., are the classical labels for the letter values, and are due to Tukey (cf. [2]).

The original motivation for our work is a problem posed in [1; p. 56]. There, the reader is asked to find the probability density function of a continuous distribution whose letter values are equally spaced; that is, the differences $x_{k+1}-x_k$

are independent of k. Using (1) and (2), it can be shown that the Laplace density function

$$f(x) = \frac{1}{2\sigma} \exp(-\frac{|x-\mu|}{\sigma}) \ , \ -\infty<x<\infty, \tag{3}$$

where $\sigma>0$, $-\infty<\mu<\infty$, has equally spaced letter values. Another distribution with nicely spaced letter values is the uniform distribution on (-1,1); here, the spacings $x_{k+1}-x_k$ form a geometric progression with common ratio $r = \frac{1}{2}$.

In view of these two examples, it is natural to ask whether the function (3) is the only probability density having equally spaced letter values. More generally, if the spacings $x_{k+1}-x_k$ form a geometric progression, what is the underlying distribution?

Clearly some symmetry assumptions are needed; indeed, the standard exponential density

$$f(x) = e^{-x}, \ x>0, \tag{4}$$

has equally spaced <u>upper</u> letter values, while its lower letter values are not even spaced in geometric progression. Below we prove that if the density f(x) is symmetric about the median x_0, and has letter values which are spaced according to a geometric progression, then with minimal regularity assumptions, f(x) is uniquely determined. Further, we shall derive the explicit form of f(x).

3. MAIN RESULT

Theorem: Let X be a random variable with density function f(x). Assume that f(x) is (i) continuous and symmetric about the median x_0, (ii) differentiable everywhere on the range of X, except possibly at x_0, (iii) log-concave and monotonic decreasing for $x>x_0$. If the spacings between the letter values of X form a geometric progression,

$$x_{k+1}-x_k = ar^k, \quad k=0,1,2,... \tag{5}$$

where $0<r<1$, $a>0$, then up to location and scale,

$$f(x) = \begin{cases} \frac{1}{2}\alpha(1-|x|)^{\alpha-1}, & -1<x<1 \\ 0, & \text{elsewhere} \end{cases} \tag{6}$$

where $r = 1/2^{1/\alpha}$, $1\leq\alpha<\infty$.

Before proving this result, let us note some of its implications. First, the case $r=\frac{1}{2}$ or $\alpha=1$ corresponds to the uniform distribution on (-1,1); in this case, the theorem remains valid if assumption (iii) is replaced by the weaker hypothesis (iii)' f(x) is concave for $x>x_0$. To recover the case r=1, let $Y = \alpha X$; then the density function of Y converges to the Laplace density (3) (with $\mu=0$, $\sigma=1$) as $\alpha \to \infty$ or as $r \to 1$.

Proof of the Theorem: Since $0<r<1$, then the sum

$$\sum_{k=0}^{\infty} (x_{k+1}-x_k) = a \sum_{k=0}^{\infty} r^k = a/(1-r).$$

Hence, the range of X is a finite interval. By shifting and scaling X, we may assume without loss of generality that the

range of X is the interval (-1,1). Then, $x_{-k} = -x_k$ for all k and

$$\sum_{k=0}^{\infty} (x_{k+1}-x_k) = 1$$

so that a = 1-r.

Next, define for $k \geq 0$,

$$G(x) = \int_{x_k}^{x} f(t)dt = F(x)-F(x_k), \quad x_k<x<x_{k+1}$$

where F(x) is the cdf of X. By the mean value theorem, there exists u_k in the interval (x_k, x_{k+1}) such that

$$f(u_k) = G'(u_k) = \frac{G(x_{k+1})-G(x_k)}{x_{k+1}-x_k}$$

$$= \frac{F(x_{k+1})-F(x_k)}{x_{k+1}-x_k} = \frac{b}{(2r)^k}, \quad b = \frac{1}{4a},$$

the last equality following from (1) and (5). In particular, $u_k \to 1$ as $k \to \infty$. Let $u_k = 1-e^{-v_k}$, $k \geq 0$. Choose and fix an integer $n \geq 0$ and let y = px+q be the straight line which passes through the points $(v_n, \ln f(u_n))$ and $(v_{n+1}, \ln f(u_{n+1}))$. Define

$$g(x) = \ln f(1-e^{-|x|}) - (p|x|+q), \quad -\infty<x<\infty. \tag{7}$$

Then, g(x) is symmetric about x=0, satisfies $g(v_n) = g(v_{n+1}) = 0$, and is differentiable everywhere except

possibly at x=0.

Clearly, the function $h_1(x) = 1-e^{-x}$ is concave on $(0,\infty)$. Since $h_2(x) = \ln f(x)$ is concave and monotonic decreasing then the composite function $(h_2 \circ h_1)(x) = \ln f(1-e^{-x})$ is concave on $(0,\infty)$. Since

$$g'(x) = \begin{cases} [\ln f(1-e^{-x})]' - p, & x>0 \\ [\ln f(1-e^{x})]' + p, & x<0 \end{cases}$$

then $g(x)$ is also concave on $(0,\infty)$.

Now suppose that $g'(v_n) < 0$. Since $g(x)$ is concave then $g'(x) \leq g'(v_n) < 0$ for all $x > v_n$. Thus $g(x)$ is strictly decreasing for $x > v_n$; in particular, $g(v_{n+1}) < g(v_n) = 0$, contradicting $g(v_{n+1}) = 0$. Therefore $g'(v_n) \geq 0$ for all $n=0,1,2,\ldots$ Next, Rolle's theorem and the concavity of $g(x)$ show that $g'(x)$ has exactly one zero, w, in (v_n,v_{n+1}). Hence,

$$0 \leq g'(v_{n+1}) \leq g'(w) = 0$$

so that $g'(v_{n+1}) = 0$ for all $n \geq 0$. Again by concavity, we have $g'(x) = 0$ for $|x| \geq v_1$; that is, $g(x) = g(v_1) = 0$ for $|x| \geq v_1$. Hence, (7) implies

$$f(x) = c(1-|x|)^{\alpha-1}, \quad u_1 \leq |x| < 1, \tag{8}$$

where c is constant, and the log-concavity of $f(x)$ entails $\alpha \geq 1$. Integrating (8) over $(x_2,1)$ and $(x_3,1)$, applying (1) and simplifying, we even find that $r = 1/2^{1/\alpha}$ and $c = \alpha/2$.

Finally, it remains to be shown that (8) also holds when $|x| \leq u_1$. Since $g(x)$ is concave on $(0,v_1)$, then $g'(x) \geq g'(v_1) = 0$, $0<x<v_1$. That is, g is monotone increasing,

and we even have $g(x) \leq g(v_1) = 0$, $0<x<v_1$. By the preceding arguments, we know that

$$g(x) = \ln f(1-e^{-x}) - \ln f_\alpha(1-e^{-x}), \quad x > 0,$$

where $f_\alpha(x)$ is the function in (6), so $g(x) \leq 0$ is equivalent to

$$f(x) \leq f_\alpha(x), \quad 0<x<v_1. \tag{9}$$

If there is strict inequality in (9) at $x = t_0$, then the continuity of $f(x)$ and $f_\alpha(x)$ at t_0 guarantees that the strict inequality holds in an open neighborhood of t_0. Then by integrating (9), we get

$$\int_0^{x_2} f(x)dx < \int_0^{x_2} f_\alpha(x)dx = 5/8,$$

contradicting the definition of x_2 as the second upper letter value. Therefore $f(x) = f_\alpha(x)$ on $(0,1)$ and by symmetry and continuity on all of $(-1,1)$. □

If the function $f(x)$ is assumed to be log-convex and monotone increasing, then analogous arguments lead to the conclusion $f = f_\alpha$ where $0<\alpha\leq 1$. Even more is true, the case when the ratio $r>1$ can be analyzed using similar methods. Alternatively, note that if X is the random variable in our result, and Y is defined by

$$1 + X = \frac{1}{1+Y},$$

then Y has range $(-\infty,\infty)$ and the letter values of Y are spaced

according to a geometric progression with $r>1$. Then a theorem analogous to the main result can be obtained, characterizing the density functions

$$g_{\alpha}(x) = \frac{\alpha}{2(1+|x|)^{\alpha+1}}, \quad -\infty<x<\infty,$$

where $\alpha>0$. We leave the precise details to the reader.

4. TESTING FOR A COMPLETELY SPECIFIED DISTRIBUTION

For an application of our results, consider the problem of testing the hypothesis H_0 that a random sample $X_1, X_2, \ldots, X_n$ is drawn from a continuous population whose distribution function $F(x)$ is completely specified. The important case where $F(x) = \Phi(x)$, the standard normal distribution function, has been reviewed in [1; p. 425 ff.].

For the X_i to follow the distribution $F(x)$, it is necessary and sufficient that the transformed data $Y_1 = F(X_1), \ldots, Y_n = F(X_n)$ follow the uniform distribution on $(0,1)$. If $F(x)$ satisfies the assumptions of our main result, we can base a test of the hypothesis on the letter values $x_1, x_2, \ldots$ of the transformed data. Thus, we would fail to reject the null hypothesis if the spacings $x_{k+1}-x_k$ approximate a geometric progression with $r = \frac{1}{2}$. This is a rule of thumb procedure for testing H_0. One possible test statistic for testing H_0 is

$$R = \sum_{k=-N}^{N} \left(\frac{x_{k+1}-x_k}{x_k-x_{k-1}} - \tfrac{1}{2}\right)^2,$$

where N is chosen to equal the smallest k for which $x_{k+1}-x_k$ is negligible. H_0 is to be rejected for large values of R. In work now in progress we are performing simulation studies of R and other statistics, based only on the letter values, for testing H_0. Results will appear elsewhere.

Acknowledgement. We are grateful to David Ruppert for bringing to our attention an error in a previous version of this work.

REFERENCES

1. Hoaglin, D.C., Mosteller, F., and Tukey, J.W., eds. Understanding Robust and Exploratory Data Analysis, 1983, New York: Wiley.

2. Tukey, J.W. Exploratory Data Analysis, 1977. Reading: Addison-Wesley.