On Dividing an Empirical Distribution into Optimal Segments

Jan W. Owsiński[1]

Abstract An approach is presented to the division of a unidimensional empirical distribution into "categories" or "classes". It is based on the use of an objective function, called bi-partial, which balances the "exactness of approximation" of the distribution by the categories determined and the "distinctness of the categories". Thereby, the optimum division, including the number of categories, can be obtained. The paper shows also how some of the existing distributions can hardly be treated with the approach and discusses reasons and consequences of such cases.

Key words empirical distribution, optimum division, classes, objective function, bi-partial approach, substantive criteria

1. The Setting

Assume a univariate empirical distribution of quantity x with values from $\boldsymbol{R}_+$. The distribution consists of n observations, indexed i, $i = 1,\ldots,n$. Denote this set of indices I. Without any loss to any sort of reasoning, we assume that the values taken by x in this distribution, denoted x_i, are ordered in a non-decreasing sequence, i.e. $x_{i+1} \geq x_i$ for all i.

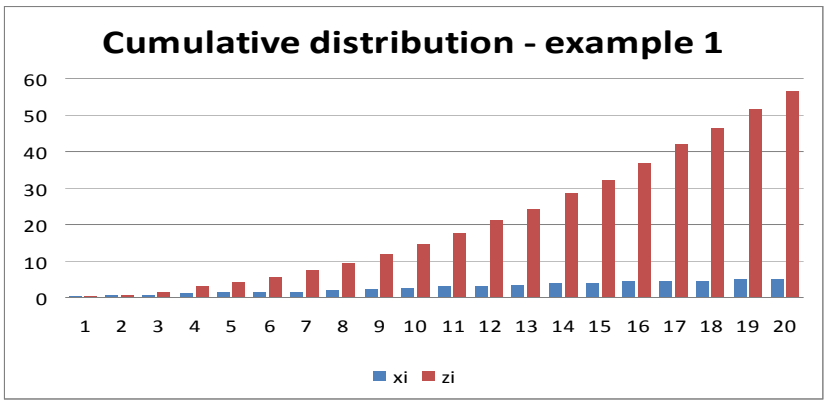
Next, assume we consider, instead of the sequence $\{x_i\}_I$, the corresponding sequence, formed by the respective cumulative distribution, i.e. the values z_i defined as $z_i = \Sigma_{j=1,\ldots,i} x_j$. So, we deal with a sequence $\{z_i\}_I$ that is increasing and also convex. This means that a straight line, joining any two points of the sequence $\{z_i\}_I$, say z_i and $z_{i+\Delta I}$, where Δi is any integer number contained in the interval $[2,n-i]$, has values above those of the corresponding z_i, i.e. $z_{i+1},\ldots,z_{i+\Delta i-1}$ (see the example of Fig. 1).

For such Lorenz-curve-like data we would like to construct a piece-wise linear approximation that is in some sense "optimal". Namely, we would like to determine a set of line segments such that the resulting error (sum of absolute differences between the actual values of z_i and the corresponding values of the approximating function) is possibly low, while the number of segments distinguished is also kept reasonably low.

---

[1] Jan W. Owsiński, Systems Research Institute, Polish Academy of Sciences, Newelska 6, 01-447 Warszawa, Poland, e-mail: owsinski@ibspan.waw.pl

Figure 1: A simple academic example of a convex cumulative distribution



2. The Interpretation

This problem, even if only roughly defined, applies not just to "approximation theory", but to a wide variety of concrete domains. The one we mean here is the distribution of social and/or economic indicator values among units indexed *i* (countries, regions, municipalities,…), for which we would like to obtain not an "approximation", but a set of "classes" or "types", among which the units can be assigned. This is the case of some development indices, for which we seek an appropriate classification of, say, countries, into groups referred to as "highly developed", "developed", …, "dramatically lagging", without assuming an arbitrary division of index values *i* or thresholds in terms of x_i. Even more – we would not like to have the number of such classes defined beforehand, but, rather, obtained as the output from the procedure.

If we obtained such a "more objective" division, based only on the shape of the sequence $\{z_i\}$, then the assignment of labels, such as "highly developed" etc., would be done a posteriori on the basis of characteristics of the classes obtained, rather than from a largely subjective perspective on how the classes "should" be defined or named.

The present analysis was motivated by exactly such a proposal by Nielsen, [4], concerning the country development levels. Another domain of interest with similar features is the one of distribution of wealth within a society, with the *i*'s corresponding to somehow defined wealth classes. Nielsen's [4] proposal is analysed and extended on the basis of the "bi-partial" approach, as described, e.g., by Owsiński, [5, 6].

3. Some Properties and the Initial Objective Function

It appears that solving the problem consists in minimising the error for subsequent numbers of classes (segments) and finding the "most appropriate" solution in terms of the error value and the number of classes. The weak point of such a procedure consists in finding a "proper" trade-off between the error and the number of segments.

Namely, obviously, the error for the optimum approximation decreases as a function of the number of classes. Given this, it would appear "natural" to look for a different form of the objective function we try to optimise (minimise), rather than, in very general terms, the "total error + number of classes".

For this purpose we shall introduce further notation. Denote with q the index of the subsequent classes, $q = 1,\ldots,p$, p being the overall number of classes distinguished. Denote with A_q the set of indices i of observations x_i (and so also z_i), classified in class q. We shall denote by $z^{q\min}$ and $z^{q\max}$, respectively, the minimum and maximum values of z_i, corresponding to the set A_q. These values, in turn, correspond to indices $i^{q\min}$ and $i^{q\max}$, respectively. Let us denote the set of i values, defining the partition of the sequence $1,\ldots,n$ into the subsets A_q, i.e. the sequence composed of $1=i^{1\min}, i^{1\max}, i^{2\min}, i^{2\max}, i^{3\min}, \ldots, i^{p\max}=n$, by $\boldsymbol{iq}$. By specifying $\boldsymbol{iq}$, we define $\{A_q\}$ and the entire solution. When referring to the explicit set of subsets $\{A_q\}$ we may also use the notation P, for partition of the set of observations.

For the assumed piece-wise linear approximation the general form of the q^{th} piece is

$$z^q(i) = a^q i + b^q, \tag{1}$$

where we can no longer care whether i is discrete or continuous, but we observe the values only for natural i. The values of a^q and b^q are determined in a natural manner from the standard formulae, where we assume, formally, that each segment is composed of at least two consecutive observations, i.e. $i^{qmax}>i^{qmin}$:

$$a^q = \frac{z^{q\max} - z^{q\min}}{i^{q\max} - i^{q\min}}, \tag{2}$$

$$b^q = z^{q\min} - \frac{z^{q\max} - z^{q\min}}{i^{q\max} - i^{q\min}}\, i^{q\min}. \tag{3}$$

Note that after differentiating $z^q(i)$ as in (1) we obtain the increasing sequence of levels a^q, corresponding to classes in terms of values of x_i.

In view of the convexity of the sequence of z_i, the sequence of a^q is non-decreasing, while the sequence of b^q is non-increasing.

We can now formulate the "minimum approximation error" problem, with the respective objective function, denoted $C_D(\{A_q\})$, as follows:

$$\min_{\boldsymbol{iq}} (C_D(\{A_q\}) = \sum_q \sum_{i\in Aq} (z^q(i) - z_i)), \tag{4}$$

where minimisation is performed with respect to the sequence $\boldsymbol{iq}$. We shall denote the optimum sequence, corresponding to the minimum in (4), by $\boldsymbol{iq}^*$.

As mentioned, the optimum value of this objective function is non-increasing in the number of segments, p (see the examples in Table 1, derived from the data shown in Fig. 1; although the assumptions differ among the examples, and explicit optimisation has not been carried out, the interpretation of the results appears to be obvious).

Since under convexity there is one optimum $\boldsymbol{iq}^*$ for each consecutive value of p (quite in line with Nielsen, 2011), we can denote the minimum value of $C_D(\{A_q\})$ for a given p by $C_D^*(p)$, so that $C_D^*(p) \geq C_D^*(p+1)$. Equality can only occur when sequences $x_i = x_{i+1} = \ldots$ exist, so that corresponding $z_i, z_{i+1}, \ldots$ lie on a straight line. Otherwise, any increase of p leads to a decrease of $C_D^*(p)$. One could go in this manner to the extreme of $p=n$, when $C_D^*(n) = 0$, an "ideal approximation"! Each observation would then constitute a separate "class" with one representative.

Obviously, when the already mentioned sequences $x_i = x_{i+1} = \ldots$, occur, so that the $z_i, z_{i+1}, \ldots$, are situated on a straight line, $C_D^*(p)$ shall remain at 0 also for $p < n$, down to the value, determined by the total length of such uniform sequences.

While construction of approximating segments is not a question, the issue that we address here is related to finding a way to tell "how different the successive observations have to be in order to assign them to different segments (classes)?".

Table 1 Examples of division of the cumulative distribution from Fig. 1 and the corresponding values of $C_D(\{A_q\})$. In this distribution $i = 1,\ldots,20$.

Number of segments:	**3**	**4**	**5**	**6**
Subsets of indices forming the division:	{1-3} {4-15} {16-20}	{1-5} {5-10} {10-15} {15-20}	{1-3} {4-7} {8-10} {11-15} {16-20}	{1-3} {4-7} {8-10} {11-13} {14-15} {16-20}
$C_D(\{A_q\})$	30.75	7.6	2.4	1.05

4. Construction of a Bi-Partial Objective Function

If the formulation were "minimise the error with as low number of segments as possible", the following formulation would result:

$$\min (C_D(\{A_q\}) + w(p)) \qquad (5)$$

where $w(p)$ is the weight attached to the number of segments. For consecutive values of p the minimum of $C_D(\{A_q\})$ would be found, and then the minimum of the function from (5) determined. Although this procedure might seem cumbersome, but, as we expect not too many segments to correspond to optimum, it would be numerically feasible. We shall, though, not go into the technical details, for reasons given below.

Namely, now, the essence of the problem is transferred to determination of the function $w(.)$. In cases when p has a concrete interpretation, like cost, while error minimisation leads to definite benefits (in technical applications or in operational research), then determination of $w(.)$ is feasible, even if difficult. This is not, however, the case with our problem, where we look for some possibly "natural" division of the distribution, and no cost / benefit, except for the facility of use of appropriate linguistic labels ("very highly developed", "highly developed",…), is involved.

As we try to find the "natural" division of the cumulative distribution (provided it exists, and the method we aim at ought to tell us whether it does), therefore, we should refer to some "counterweight", analogous to that of $w(p)$ in (5), but having the same sort of meaning and kind of measurement as $C_D(\{A_q\})$. In this way we might be able to try to define the proper p and at the same time the $\boldsymbol{iq}$, or, otherwise, the $\{A_q\} = P$.

Thus, similarly as in (5), we would like to add to $C_D(\{A_q\})$ a component that would penalize, in this case, for the division into segments that are in some way "too similar", especially in terms of subsequent a^q. In general terms the respective bi-partial objective function and the corresponding problem would look like

$$\min (C_D(\{A_q\}) + C^S(\{A_q\})), \qquad (6)$$

where $C^S(\{A_q\})$ corresponds to aggregate similarity between the consecutive segments, based primarily on differences of consecutive a^q. A concrete form of $C^S(\{A_q\})$ might be constructed as follows:

-- first, a kind of difference between two consecutive segments, q-1 and q, is measured, from the point of view of the succeeding segment, q, as, for instance,

$$z_{iq\min} - a^{q-1} i^{q\min} - b^{q-1} \qquad (7)$$

i.e. the difference between the actual value of z_i at the beginning of the next, q^{th} subset of observations, and the "approximation" of the same, resulting from the previous segment. This difference is always non-negative, due to convexity of $\{z_i\}$, and can be interpreted as a "distance" between the two consecutive segments in the approximation;

-- as we wish to penalize with $C^S(.)$ *similarity*, not distance (difference), in order to convert (7) into similarity, we subtract it from an upper bound, which might be constituted by the maximum of a similar difference for a given data set, namely the biggest difference of tangents along the curve of z_i, i.e. between its beginning and end; the two extreme tangents, $a^{(1)}$ and $a^{(n)}$, are defined as:

$$a^{(1)} = z_1/i_1; \quad \text{and} \quad a^{(n)} = (z_n - z_{n-1})/(i_n - i_{n-1}); \tag{8}$$

yet, in order to calculate the proper difference, we must have full expressions for the lines corresponding to $a^{(1)}$ and $a^{(n)}$, allowing for their use for consecutive subsets A_q; we assume, namely, that all four lines involved, corresponding to a^q, a^{q-1}, $a^{(1)}$ and $a^{(n)}$ cross at the point, defined otherwise by the crossing of the lines, corresponding to A_{q-1} and A_q; from this condition we derive the values of b, to be used in conjunction with $a^{(1)}$ and $a^{(n)}$ (denoted, respectively, $b^{*(1)}$ and $b^{*(n)}$) in the appropriate expression, namely:

$$b^{*(1)} = b^q - (a^{(1)} - a^q)(b^{q-1} - b^q)/(a^q - a^{q-1}) \tag{9a}$$

$$b^{*(n)} = b^q - (a^{(n)} - a^q)(b^{q-1} - b^q)/(a^q - a^{q-1}). \tag{9b}$$

Now, the expression for $C^S(.)$ for a single q can be written down as

$$a^{(n)}i^{q\min} + b^{*(n)} - (a^{(1)}i^{q\min} + b^{*(1)}) - (z_{iq\min} - a^{q-1}i^{q\min} - b^{q-1}) \tag{10}$$

where the second term in brackets is equivalent to the difference, given by (7), while the preceding terms define the reference for the given q. The proposed $C^S(P)$ is the sum over q of (10). Altogether, the minimised objective function takes on the form:

$$C_D{}^S(\{A_q\}) = C_D(\{A_q\}) + C^S(\{A_q\})) =$$
$$\Sigma_q\Sigma_{i\in Aq}(a^q i + b^q - z_i) + \Sigma_q(a^{(n)}i^{q\min} + b^{*(n)} - (a^{(1)}i^{q\min} + b^{*(1)}) - (z_{iq\min} - a^{q-1}i^{q\min} - b^{q-1})). \tag{11}$$

where we formally assume $a^0 = 0$ (which is natural) and $b^0 = 0$ (which is a bit artificial).

For the illustrative example considered here, the results for the divisions, already referred to in Table 1, taking, additionally, into account (11), are shown in Table 2.

Table 2 Examples of division of the cumulative distribution from Fig. 1 and Table 1, and the corresponding values of $C_D(\{A_q\})$, $C^S(\{A_q\})$, and $C_D{}^S(\{A_q\})$. In this distribution $i = 1,\ldots,20$.

Number of segments:	**3**	**4**	**5**	**6**
Subsets of indices forming the division:	{1-3} {4-15} {16-20}	{1-5} {5-10} {10-15} {15-20}	{1-3} {4-7} {8-10} {11-15} {16-20}	{1-3} {4-7} {8-10} {11-13} {14-15} {16-20}
$C_D(\{A_q\})$	30.75	7.6	2.4	1.05
$C^S(\{A_q\})$	4.64	8.01	10.88	15.19
$C_D{}^S(\{A_q\})$	35.39	15.61	**13.28**	16.24

The bi-partial objective function selects among the examples provided the one with five segments, its value for six segments being also higher than that for four. Since the respective partitions are (close to) nested, i.e. the increasing number of segments corresponds to divisions of selected A_q forming the preceding partition, this example shows that indeed we might deal with a convex objective function along such nested

families of partitions. This implies the existence of a non-trivial minimum of $C_D^S(\{A_q\})$ in the set of all ***iq***, i.e. partitions, though we shall not be trying to demonstrate this here.

5. Some General Properties and Potential Algorithms

The construction of the bi-partial objective function follows only quite general prerequisites of "global' rationality, namely that we oppose two measures that individually represent a one-sided rationality (like error minimization), and that together imply a compromise, based on their joint minimization or maximization. In this, we do not enforce neither the concrete structure (value of p), nor any weight – although weights can, of course, be applied, and even may be effectively used.

In this particular case, we constructed the bi-partial objective function out of components $C_D(\{A_q\})$ corresponding to the error, resulting from the "approximation" of the sequence of z_i with a limited number of line segments, and $C^S(\{A_q\})$, corresponding to the penalty for the too small change of angle of the line between two consecutive segments. Although we have not shown this with respect to $C^S(\{A_q\})$, the two components display opposite monotonicity along the number of segments, p, that is - minimum $C_D(\{A_q\})$, or $C_D^*(p)$, decreases along p, while $C_D^S(\{A_q\})$ increases (we refer here only to the sequence of ***iq*** minimizing $C_D(\{A_q\})$).

The above remark indicates one of the fundamental principles of construction of the bi-partial objective function, namely the *opposite monotonicity* of the two components.

One might indicate, though, that the two components in this example are not quite "symmetric": there is only one element per segment in $C^S(\{A_q\})$, while there are cardA_q-2 elements per segment in $C_D(\{A_q\})$, which, definitely, introduces a bias (in this realisation of the bi-partial objective function the segments obtained cannot be too big, i.e. cardA_q high). Indeed, we have provided here only an example: the entire formulation of the problem, also involving the "error function", is arbitrary (we could use the sum of error squares).

Concerning the optimisation algorithms, the off-the-shelf choice for a single-dimensional problem is dynamic programming, like in classical categorisation problem (see, e.g., [1]). Yet, we can also consider the approach by the present author, closely associated to the idea and properties of the bi-partial objective function. We shall provide here only the basic precepts of this approach.

We shall illustrate the approach with the concrete form of $C_D^S(\{A_q\})$, considered here. Thus, assume we consider, instead of (11), a parameterised form:

$$C_D^S(\{A_q\},r) = (1\text{-}r)C_D(\{A_q\}) + rC^S(\{A_q\})) \qquad (12)$$

with $r \in [0,1]$, and we look for minimum $C_D^S(\{A_q\},r)$ over ***iq***, i.e. $C_D^{S*}(r)$.

Then, assume we start the procedure from $r = 0$. We have $C_D^S(\{A_q\},0) = C_D(\{A_q\})$, and, of course, the "optimum" partition is the one with $p = n$, or, at most (according to our form of the "error function"), p = int[n/2]+1, where int[v] is the highest integer number lower than v (due to zeroing of elements of $C_D(\{A_q\})$ at the segment endpoints). Yet, we are not, in general, interested in such a solution. As we increase r from 0, non-zero weight starts to be assigned to $C^S(\{A_q\})$, and in order to obtain $C_D^{S*}(r)$ for such r, it will "pay" at some definite value of r, say r^1, to join two segments, for which the difference of angle is the smallest, and hence the penalty in $C^S(\{A_q\})$ is the biggest. The value of r^1 can be easily determined on the basis of the formulae here provided.

As we increase the parameter r further, we find the next one, r^2, for which merging of another pair of consecutive segments "pays" in terms of $C_D^S(\{A_q\},r)$. And so on. The proper solution is found for the last r^t that is not bigger than ½ - i.e. for the equal weights of the two components of $C_D^S(\{A_q\},r)$. This, of course, is a sub-optimisation algorithm, as it does not guarantee reaching of the proper minimum of $C_D^S(\{A_q\})$. Yet, experience shows that it either actually reaches the minimum, or is very close to it.

No matter which method is used, the basic rationale consists in forming the segments A_q for sequences of possibly similar x_i, i.e. z_i approaching a straight line. An adequately pronounced "jump" over one or several consecutive i's would then correspond to a change from q to q+1 in the optimum solution.

6. The Spiteful Reality

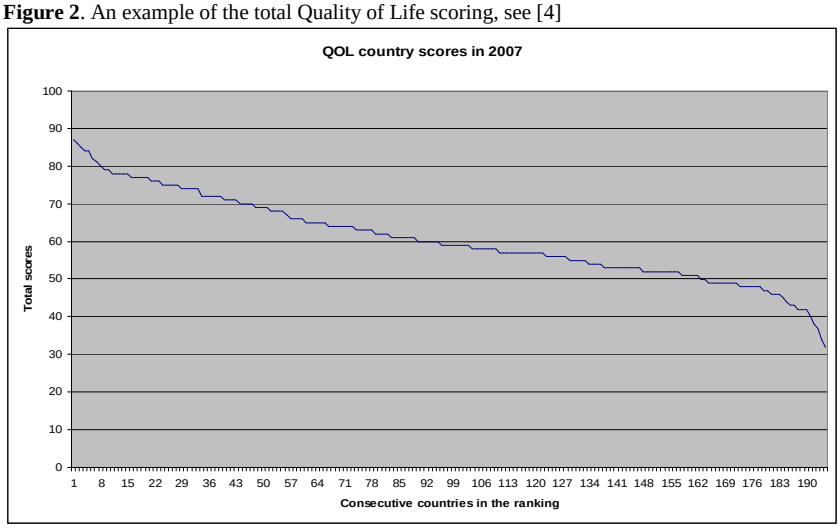
The above rationalization of the perspective on dividing the distributions of the kind we mean here, though, encounters often an essential hindrance in the actual shapes of such distributions. An example, of quite a mild character at that is provided in Fig. 2, showing ordered (from the "best" to the "worst") country scores of well known Quality of Life (QoL) ranking, [2] (here: for 2007, [3]), for close to 200 countries of the world. The total scores shown, between 0 and 100, are based on nine partial scores for broad domains, such as "living costs", "economy", "environment", "freedom", "safety", etc.

The problem lies in the shape of the curve, corresponding to our sequence of x_i. One can easily see that – due to the nature of the scoring system – in the middle part of the curve there are numerous shorter and longer flat segments, some of them separated only by minimum jumps. For this part of the curve the methodology here outlined, and also the broader rationality referred to in the preceding section, can well be successfully applied. Yet, the two ends of the curve display a completely different character: sharp increase of the gradient towards the two extremes.

Within these two extreme parts of the curve the methodology – and the broader rationality – would have to distinguish several classes, with very few objects, indeed, in most cases – just one – in each consecutive class. This seems to bend the rationality we made use of. Also the approach of Nielsen [4] will have troubles with this shape.

This shape is not an incidental result of the methodology, adopted in creating the QoL ranking and the actual data used. It is a consistent feature – the very same shape appears in most of the partial score-based rankings. It is even much more pronounced in some of them (e.g. for "living costs", "economy", "environment", "infrastructure"). This shape appears also from year to year. The rankings do not result, though, from some statistical measurement, at least not as they are reported. They are either the immediate result of quite subjective assessments of experts on the individual variables, contributing to the particular domains, or of the data, characterising these variables. Thus, ultimately, we deal with somehow aggregated expert opinions. This fact may largely explain the character of the final output. It is, namely, so that for many of the "intermediate" countries, with respect to particular variables, expert assessments barely distinguish between them, while the extremes are easily noted. Now, since there is generally a high level of correlation between many variables (roughly +70% being typical correlation coefficient), such observations, concerning the extremes, summing up, therefore, and creating the ends of the respective curves, as observed in Fig. 2.

Figure 2. An example of the total Quality of Life scoring, see [4]



Actually, the issue is, in general, insofar more serious as many of the empirical, "objective" or "statistical" distributions behave, indeed, according to highly regular functional shapes, so that there is very little ground for dividing them in a different manner than on the basis of *substantive criteria* (e.g. the "biological minimum" or "social minimum" thresholds in the case of poverty). Application of the approach outlined here would then involve the measures of fit to / divergence from the matching functional shapes.

Hence, the following question arises in the context of the optimum distribution division problem: if the results of a division exercise indicate a similar phenomenon to that here commented upon, indicating a sort of (unexpected?) regularity, can we deduce something about the way in which the respective distribution has been constructed?

Acknowledgment: The research reported in the paper has been partly supported within the projects: TIROLS, funded by the Polish Ministry of Science and Higher Education, N N516 195237, and "Development Trends of Masovia", MBPR/W/TRM-0712-08-2011, co-financed from the EU funds.

References

1. Gan G., Ma Ch. and Wu J. (2007) *Data Clustering. Theory, Algorithms and Applications.* SIAM & ASA, Philadelphia.
2. http://www.economist.com/media/pdf/QUALITY_OF_LIFE.PDF – The Economist Intelligence Unit's quality-of-life index (of 2005) (as seen on Nov. 18th, 2011)
3. http://www.il-ireland.com/il/qofl07/ – 2007 Quality of Life Index
4. Nielsen, L. (2011) Classification of Countries Based on Their Level of Development: How it is Done and How it Could be Done. IMF Working Paper, WP/11/31, IMF.
5. Owsiński J.W. (1990) On a new naturally indexed quick clustering method with a global objective function. *Applied Stochastic Models and Data Analysis*, 6, 157-171.
6. Owsiński, J.W.: The bi-partial approach in clustering and ordering: the model and the algorithms. *Statistica & Applicazioni*, 2011, Special Issue, pp. 43-59.