MODELS OF RANK-FREQUENCY DISTRIBUTIONS IN LANGUAGE AND MUSIC

GEJZA WIMMER*, GABRIEL ALTMANN**

ABSTRACT. One possible way to discover law-like hypotheses in musicology or in linguistics is to model variables such as length (of tones, of words, etc.), strength of voice, complexity, number of repetitions, etc. by means of discrete probability distributions. Many ranking problems could be modelled by partial-sums distributions. Four simple schemes for generating partial-sums distributions are presented. Some models and their fitting to empirical data are also demonstrated.

In musicology as well as in linguistics discrete probability distributions (p.d.) arise in two ways:

(i) By means of a generating mechanism that developed historically and brings about the genesis of data;

(ii) by means of ranking of elements of a class ordered according to their frequency of occurrence.

The variable in (i) is so to say "natural" (measurable, countable, e.g. p.d. of the word length, of the semantical productivity of words, etc.), that of (ii) is "artificial" (mathematical) if it is at all necessary to use these terms to differentiate them since both have been constructed by us. It is evident that in every class of musical (or linguistic) entities constructed by us an "artificial" order can be established if one is able to define some pertinent criteria. If, in adition, some lawfull inter-relations of ranking can be discovered, then it can serve as a criterion of "naturalness" or "correctness" or "closeness to reality". The difference between (i) and (ii) can, however, be founded also genetically and the modelling techniques can be appropriately adapted.

With "natural" variables such as length, strength of voice, complexity, number of repetitions, etc. one can assume that originally only the simplest stage existed (e.g. in the language – monosyllabic words, sentences consisting of one clause, etc.). More complex forms developed because of reduction of redundancy, on the basis of the coding requirement, expression requirement, etc. (cf. [4]), but in dependence on the properties of the simplest classes. Since it is impossible to reconstruct the elementary state of music or language, we restrict ourselves here to the dependence in the domain of frequency which is discrete or may be made discrete and displays

Supported by VEGA, grant No. 1/4196/97, grant No. 2/5126/98 and grant No. 1/3171/96.

1991 *Mathematics Subject Classification.* 60E05,62E10.

Key words and phrases. discrete probability distributions, partial-sums distributions, negative hypergeometric distribution.

conspicuous regularities. It has already been shown in many problems of quantitative linguistics that the frequency in class x is proportional to that in class $x-1$, or even to those in all lower classes (c.f. [1],[5],[7],[8]). The complex approach is

$$P_x = g(x)\sum_{j=1}^{x} h(j)P_{x-j}, \tag{1}$$

where $g(.)$ is a proportionality function, $h(.)$ is a weighting function. Thus the summation concerns the classes 0 to $x-1$, i.e. the frequency in class x turns to be a weighted sum of the frequencies in all lower classes. The simplest form of (1) is

$$P_x = g(x)P_{x-1} \tag{2}$$

and this approach has been successfully used in hundreds of cases (for word length see e.g. [2]). This approach already yields an elementary foundation or explanation but if necessary, one can go a step deeper and consider equations (1) or (2) as steady-state solutions of stochastic processes, e.g. of the birth-and-death process. This fact enables us to embed this approach into a more general theory and, at the same time, to use it as an instrument for capturing the variability of languages, texts, compositions etc.

With ranking, i.e. with variable of type (ii) the argumentation can be quite the other way round. We have here three points of departure

(a) the lowest rank is 1 but conventionally it can be set to 0,

(b) the p.d. is monotone decreasing,

(c) we observe the actual state, not a genesis of the considered set.

From (c) it follows that the frequency at rank 1 (the most frequent element of the considered set) depends on the number of the other elements and their frequencies. This fact can easily be illustrated e.g. in the case of the rank-frequency distribution of phonemes of a language. The more phonemes there are in the inventory (investigated set) the more even is the curve, the smaller is the relative frequency of the phoneme with rank 1. Thus we can assume that P_1 can be considered as a function of the sum of frequencies of other phonemes with rank greater than (or equal to) 1, P_2 as a function of the sum of frequencies of phonemes with rank greater than (or equal to) 2, etc. However, it is evident that the summed variable, called *parent* and marked as P_j^*, is different. Below we shall show four simple possibilities (schemes).

Scheme I.

$$P_1 = C_1\{P_1^* + P_2^* + P_3^* + P_4^* + \ldots\}$$

$$P_2 = C_1\{P_2^* + P_3^* + P_4^* + \ldots\}$$

$$P_3 = C_1\{P_3^* + P_4^* + \ldots\}$$

$$\vdots$$

As $\sum_{x\geqq 1} P_x = 1$, we obtain

$$P_x = \frac{1}{\mu'^*_1} \sum_{j\geqq x} P^*_j, \quad x = 1, 2, \ldots,$$

μ'^*_1 being the mean of the parent p.d.
The recurrence formula for probabilities is

$$P_x = P_{x-1} - \frac{P^*_{x-1}}{\mu'^*_1}, \quad x = 2, 3, \ldots .$$

Scheme II.

$$P_1 = C_2\{P^*_1 + \frac{P^*_2}{2} + \frac{P^*_3}{3} + \frac{P^*_4}{4} + \ldots\}$$

$$P_2 = C_2\{\frac{P^*_2}{2} + \frac{P^*_3}{3} + \frac{P^*_4}{4} + \ldots\}$$

$$P_3 = C_2\{\frac{P^*_3}{3} + \frac{P^*_4}{4} + \ldots\}$$

$$\vdots$$

Now we obtain

$$P_x = \sum_{j\geqq x} \frac{P^*_j}{j}, \quad x = 1, 2, \ldots,$$

and the recurrence formula is

$$P_x = P_{x-1} - \frac{P^*_{x-1}}{x-1}, \quad x = 2, 3, \ldots .$$

Scheme III.

$$P_1 = C_3\{P^*_2 + P^*_3 + P^*_4 + P^*_5 + \ldots\}$$

$$P_2 = C_3\{P^*_3 + P^*_4 + P^*_5 + \ldots\}$$

$$P_3 = C_3\{P^*_4 + P^*_5 + \ldots\}$$

$$\vdots$$

For the probabilities we obtain

$$P_x = \frac{1}{\mu'^*_1 - 1} \sum_{j\geqq x+1} P^*_j, \quad x = 1, 2, \ldots,$$

and for the recurrence formula

$$P_x = P_{x-1} - \frac{P_x^*}{\mu_1'^* - 1}, \quad x = 2, 3, \ldots .$$

And finally

Scheme IV.

$$P_1 = C_4\{P_2^* + \frac{P_3^*}{2} + \frac{P_4^*}{3} + \frac{P_5^*}{4} + \ldots\}$$

$$P_2 = C_4\{\frac{P_3^*}{2} + \frac{P_4^*}{3} + \frac{P_5^*}{4} + \ldots\}$$

$$P_3 = C_4\{\frac{P_4^*}{3} + \frac{P_5^*}{4} + \ldots\}$$

$$\vdots$$

Where

$$P_x = \frac{1}{1 - P_1^*} \sum_{j \geqq x+1} \frac{P_j^*}{j-1}, \quad x = 1, 2, \ldots,$$

and the recurrence formula is

$$P_x = P_{x-1} - \frac{P_x^*}{(x-1)(1 - P_1^*)}, \quad x = 2, 3, \ldots .$$

Remarks.

(a) If the parent and the resulting partial sums p.d. begin with 0, the formulas must be slightly modified.

(b) From a particular distribution one can construct several other ones by partial summation.

We believe that many ranking problems in musicology and linguistics could be captured in this way.

Example.

(1) Let the parent p.d. be zero-truncated (= positive) Poisson p.d. with probability mass function (p.m.f.)

$$P_j^* = \frac{e^{-a} a^j}{j!(1 - e^{-a})}, \quad j = 1, 2, \ldots, \ a > 0.$$

According to Schemes I – IV we obtain four partial sums p.d. as follows

(1.I)
$$P_x = e^{-a} \sum_{j \geqq x} \frac{a^{j-1}}{j!}, \quad x = 1, 2, \ldots,$$

(1.II) $$P_x = \frac{e^{-a}}{1-e^{-a}} \sum_{j \geqq x} \frac{a^j}{j!j}, \quad x = 1, 2, \ldots,$$

(1.III) $$P_x = \frac{e^{-a}}{e^{-a}+a-1} \sum_{j \geqq x+1} \frac{a^j}{j!}, \quad x = 1, 2, \ldots,$$

(1.IV) $$P_x = \frac{e^{-a}}{1-e^{-a}-ae^{-a}} \sum_{j \geqq x+1} \frac{a^j}{j!(j-1)}, \quad x = 1, 2, \ldots .$$

(2) Using the 1-displaced Poisson p.d. with p.m.f.

$$P_j^* = \frac{e^{-a}a^{j-1}}{(j-1)!}, \quad j = 1, 2, \ldots, \ a > 0$$

we obtain merely one new p.d., namely

(2.I) $$P_x = \frac{e^{-a}}{a+1} \sum_{j \geqq x} \frac{a^{j-1}}{(j-1)!}, \quad x = 1, 2, \ldots .$$

(Type (2.II) and (2.III) are identical with type (1.I); Type (2.IV) is identical with Type (1.II).)

In order to make fitting easier we bring some other partial sums p.d.

(3) From the 1-displaced geometric p.d. with p.m.f.

$$P_j^* = pq^{j-1}, \quad j = 1, 2, \ldots, \quad 0 < p < 1, \ q = 1-p$$

we obtain only one new p.d.

(3.II) $$P_x = \frac{p}{q} \sum_{j \geqq x} \frac{q^j}{j}, \quad x = 1, 2, \ldots$$

identical with (3.IV).

(4) From the zero-truncated binomial p.d. with p.m.f.

$$P_j^* = \binom{n}{j} \frac{p^j q^{n-j}}{1-q^n}, \quad j = 1, 2, \ldots, n, \quad 0 < p < 1, \ q = 1-p$$

we obtain

(4.I) $$P_x = \frac{1}{np} \sum_{j=x}^{n} \binom{n}{j} p^j q^{n-j}, \quad x = 1, 2, \ldots, n,$$

(4.II) $$P_x = \frac{1}{1-q^n}\sum_{j=x}^{n}\binom{n}{j}\frac{p^j q^{n-j}}{j}, \quad x = 1,2,\ldots,n,$$

(4.III) $$P_x = \frac{1}{np+q^n-1}\sum_{j=x+1}^{n}\binom{n}{j}p^j q^{n-j}, \quad x = 1,2,\ldots,n,$$

(4.IV) $$P_x = \frac{1}{1-q^n-npq^{n-1}}\sum_{j=x+1}^{n}\binom{n}{j}\frac{p^j q^{n-j}}{j-1}, \quad x = 1,2,\ldots,n.$$

(5) From the zero-truncated negative binomial p.d. with p.m.f.

$$P_j^* = \binom{k+j-1}{j}\frac{p^k q^j}{1-p^k}, \quad j = 1,2,\ldots, \quad 0<k,\ 0<p<1,\ q=1-p$$

we obtain

(5.I) $$P_x = \frac{p^{k+1}}{kq}\sum_{j\geqq x}\binom{k+j-1}{j}p^k q^j, \quad x = 1,2,\ldots,$$

(5.II) $$P_x = \frac{p^k}{1-p^k}\sum_{j\geqq x}\binom{k+j-1}{j}\frac{q^j}{j}, \quad x = 1,2,\ldots,$$

(5.III) $$P_x = \frac{p^{k+1}}{qk-p+p^{k+1}}\sum_{j\geqq x+1}\binom{k+j-1}{j}q^j, \quad x = 1,2,\ldots,$$

(5.IV) $$P_x = \frac{p^k}{1-p^k-kp^k q}\sum_{j\geqq x+1}\binom{k+j-1}{j}\frac{q^j}{j-1}, \quad x = 1,2,\ldots.$$

(6) Finally from the logarithmic p.d. with p.m.f.

$$P_j^* = \frac{q^j}{-j\log_e(1-q)}, \quad j = 1,2,\ldots, \quad 0<q<1$$

we obtain

(6.I) $$P_x = (1-q)\sum_{j\geqq x}\frac{q^{j-1}}{j}, \quad x = 1,2,\ldots,$$

(6.II) $$P_x = \frac{1}{-\log_e(1-q)} \sum_{j \geqq x} \frac{q^j}{j^2}, \quad x = 1, 2, \ldots,$$

(6.III) $$P_x = \frac{-(1-q)}{q + (1-q)\log_e(1-q)} \sum_{j \geqq x+1} \frac{q^j}{j}, \quad x = 1, 2, \ldots,$$

(6.IV) $$P_x = \frac{-1}{q + \log_e(1-q)} \sum_{j \geqq x+1} \frac{q^j}{(j-1)j}, \quad x = 1, 2, \ldots .$$

Some examples from linguistics.

1. The rank-order distribution of the suffix – e in modern German according to its meaning ([6], p.112):

p.d. Type (1.III)

rank	frequency of occurrence	fitted value
1	13	11.65
2	11	8.66
3	6	6.12
4	5	4.40
5	5	3.43
6	3	2.97
7	2	2.77
8	2	2.69
9	2	2.67
10	2	2.66
11	2	2.66
12	1	2.66
13	1	2.66
14	1	0.00

$\hat{a} = 3.3840$ P = 0.98

2. The rank-order distribution of word classes in FAZ-corpus (FAZ = Frankfurter Allgemeine Zeitung) ([3], p.228):

p.d. Type (1.IV)

rank	frequency of occurrence	fitted value
1	104	80.91
2	56	69.24
3	53	56.54
4	41	44.08
5	34	33.22
6	24	24.77
7	15	18.85
8	14	15.09
9	1	0.72

$\hat{a} = 6.5372$ P = 0.11

3. The rank-order distribution of reflexives in German according to their meaning ([6], p.113):

p.d. Type (3.II)

rank	frequency of occurrence	fitted value
1	51	44.45
2	19	24.39
3	11	15.89
4	10	11.10
5	9	8.05
6	8	5.99
7	7	4.53
8	5	3.48
9	3	2.69
10	3	2.10
11	2	1.65
12	2	1.31
13	1	0.37

$\hat{q} = 0.8469$ P = 0.44

A very good model for rank-order distribution in music seems to be the 1-displaced negative hypergeometric p.d. with p.m.f.

$$P_x = \frac{\binom{M+x}{x-1}\binom{K-M+n-x}{n-x+1}}{\binom{K+n-1}{n}}, \quad x = 1, 2, \ldots,$$

$K > M > 0, \quad n \in \{0, 1, 2, \ldots\}$.

We can demonstrate this model on L.v.Beethoven's Sonata op.27, No.2 (investigated is the ranked frequency of occurrence of tones):

rank	frequency of occurrence	fitted value	rank	frequency of occurrence	fitted value
1	106	113.17	28	11	11.07
2	89	85.34	29	11	10.22
3	84	73.19	30	10	9.42
4	79	65.26	31	10	8.66
5	68	59.28	32	5	7.95
6	66	54.43	33	5	7.27
7	58	50.31	34	5	6.64
8	50	46.72	35	5	6.04
9	44	43.52	36	4	5.47
10	42	40.63	37	4	4.94
11	42	37.99	38	3	4.45
12	34	35.56	39	3	3.98
13	33	33.32	40	3	3.55
14	28	31.23	41	3	3.15
15	25	29.27	42	2	2.77
16	21	27.44	43	2	2.43
17	20	25.72	44	2	2.11
18	19	24.09	45	2	1.82
19	18	22.56	46	1	1.55
20	17	21.11	47	1	1.31
21	17	19.74	48	1	1.09
22	16	18.44	49	1	0.90
23	14	17.21	50	1	0.73
24	14	16.05	51	1	0.57
25	14	14.94	52	1	0.44
26	14	13.90	53	1	0.33
27	13	12.90	54	1	0.86
28	11	11.96			

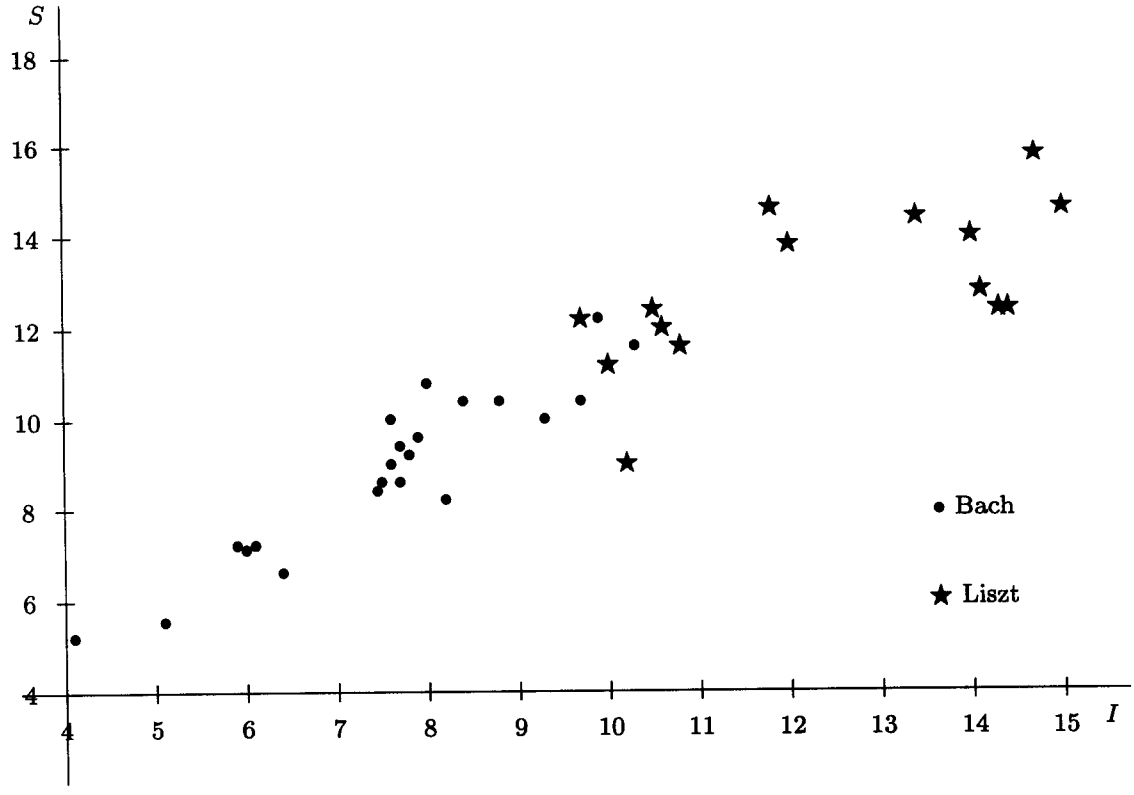
$\hat{K} = 4.0964 \quad \hat{M} = 0.7836 \quad \hat{n} = 59 \quad \mathrm{P} = 0.9945$

Using the $< I, S >$ characterization (suggested by Ord) where

$$I = \frac{\mu_2}{\mu_1'}, \qquad \mu_2 \text{ is the variance of the best fitted p.d.}$$

$$S = \frac{\mu_3}{\mu_2}, \qquad \mu_3 \text{ is the third central moment of the p.d.}$$

of the best fitted 1-displaced negative hypergeometric p.d. to the measured data (ranked frequency of occurence of tones) we obtain the next figure for some compositions of J.S.Bach and F.Liszt.



Evidently J.S.Bach and F.Liszt can be fairly well discriminated.

We hope that this way of investigation has good perspectives towards discovering law-like hypotheses in musicology as well as in linguistics or other sciences.

References

[1] Altmann, G., *Modelling diversification phenomena in language*, In: Rothe, U. (ed.), Diversification processes in language: grammar, Rottmann, Hagen, 1991.

[2] Best, K.-H., Altmann, G., *Project report*, Journal of Quantitative Linguistics **3** (1996), 85-88.

[3] Becker, H., *Die Wirtschaft in der deutschsprachigen Presse*, Lang, Frankfurt, 1995.

[4] Köhler, R., *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*, Brockmeyer, Bochum, 1986.

[5] Köhler, R., Altmann, G., *"Language Forces" and synergetic modelling of language phenomena*, Glottometrika **15** (1996), 62-67.

[6] Rothe, U., *Verteilung der Suffixe denominaler Verben nach ihren semantischen Wortbildungsmustern*, Glottometrika **15** (1990), 107-114.

[7] Wimmer, G., Altmann, G., *The theory of word length: Some results and generalizations*, Glottometrika **15** (1996), 112-133.

[8] Wimmer, G., Köhler, R., Grotjahn, R., Altmann, G., *Towards a theory of word length distribution*, Journal of Quantitative Linguistics **1** (1994), 98-106.

*Mathematical Institute, Slovak Academy of Sciences, Štefánikova 49, 814 73 Bratislava, Slovakia

**Lüdenscheid, Germany