IV.6. FITTING METHODS FOR INFORMETRIC LAWS

All of the following fitting methods will work, provided the IPP is not too small.

IV.6.1. Fitting of Bradford's law

IV.6.1.1. Methodology

The ingredients to be determined are p, y_0, r_0 and k. In principle, p can be chosen freely, although some limitation is necessary since the data are finite. We have the following equations :

$$y_0 = \frac{A}{p} \quad , \qquad [IV.6.1]$$

$$k = (e^{\gamma}\, y_m)^{\frac{1}{p}} \approx (1.781\, y_m)^{\frac{1}{p}} \quad . \qquad [IV.5.3]$$

Furthermore, since $r_0 + r_0 k + \ldots + r_0 k^{p-1} = T$ we get

$$r_0 = \frac{T(k-1)}{k^p - 1} \quad . \qquad [IV.6.2]$$

IV.6.1.2. Example

We apply this method to the bibliography 'Lubrication', the data for which can be found in Bradford (1934). We consider p = 3 since Bradford himself considered this. We then form Bradford's law for p = 7, just to show that p can be chosen more or less freely.

Table IV.6.1. Lubrication, 1931 - June 1933 (L)

# journals	corresponding # articles	r	R(r) (observed)
1	22	1	22
1	18	2	40
1	15	3	55
2	13	5	81
2	10	7	101
1	9	8	110
3	8	11	134
3	7	14	155
1	6	15	161
7	5	22	196
2	4	24	204
13	3	37	243
25	2	62	293
102	1	164	395

p = 3

We have $k = (1.781 \times 22)^{1/3} = 3.40$, $y_0 = \frac{395}{3} = 131.67 \approx 132$ and $r_0 = \frac{164(k-1)}{k^3-1} = 10.30$. Hence we use $[r_0] = 10$. The groups are :

Table IV.6.2. Bradford's law for L, p = 3

	# journals	# articles	k
1st group	$r_0 = 10.30 \approx 10$	126	-
2nd group	$r_0 k = 35.02 \approx 35$	133	3.50
3rd group	$r_0 k^2 \approx 119$, which is exactly the last rank in the bibliography	136	3.40

This is better than Bradford's original example (Bradford (1934)) : he gets 8/29/127 journals yielding respectively 110/133/152 articles.

p = 7

For p = 7 we find k = 1.69, $y_0 = 56$ and $r_0 = 2.95 \approx 3$. The Bradford groups are :

Table IV.6.3. Bradford's law for L, p = 7

	# journals	# articles	k
1st group	$r_0 = 2.95 \approx 3$	55	-
2nd group	$r_0 k = 4.98 \approx 5$	55	1.67
3rd group	$r_0 k^2 = 8.42 \approx 8$	56	1.60
4th group	$r_0 k^3 = 14.23 \approx 14$	56	1.75
5th group	$r_0 k^4 = 24.05 \approx 24$	55	1.71
6th group	$r_0 k^5 = 40.64 \approx 41$	49	1.71
7th group	$r_0 k^6 = 68.68 \approx 69$ which is exactly the last existing rank	69	1.68

Other examples can be found in Egghe (1989a or 1989e).

With the above methodology one can also show that the Bradford groups found by Goffmann and Warren (1969,1980) are wrong.

IV.6.2. Fitting Leimkuhler's function R(r) = a log (1 + br)

IV.6.2.1. Methodology

The Leimkuhler function

$$R(r) = a \log (1 + br) \qquad [IV.4.2]$$

can be deduced from Bradford's law (choose any reasonable p) by using the following exact equations :

$$a = \frac{y_0}{\log k} \quad , \qquad [IV.4.5]$$

$$b = \frac{k - 1}{r_0} \quad . \qquad [IV.4.6]$$

In view of the method developed above (cf. equations [IV.6.1], [IV.5.3] and [IV.6.2]), Leimkuhler's function [IV.4.2] can easily be calculated.

We present two examples.

IV.6.2.2. Examples

1. Lubrication

When choosing p = 3, we have $y_0 = 131.67$, $k = 3.40$ and $r_0 = 10.30$. Hence a = 107.7 and b = 0.233. We leave it to the reader to verify that only negligible differences in these values will occur when another p is chosen. We obtain the following function :

$$R(r) = 107.7 \log (1 + 0.233\, r) \; . \qquad [IV.6.3]$$

We have the following fits :

Table IV.6.4. Fit of Leimkuhler's function for 'Lubrication'

r	R(r) (observed)	R(r) (calculated)
1	22	22.6
2	40	41.2
3	55	57.1
5	81	83.2
7	101	104.2
8	110	113.3
11	134	136.8
14	155	156.1
15	161	161.9
22	196	195.2
24	204	203.1
37	243	243.8
62	293	294.8
164	395	395.1

The reader can verify with a Kolmogorov-Smirnov test that the function [IV.6.3] fits very well.

2. Pope's bibliography

Pope (1975) introduces the following data with respect to a bibliography on information science :

Table IV.6.5. Pope's bibliography

# journals	corresponding # articles	r	R(r) (observed)
1	261	1	261
1	259	2	520
1	220	3	740
1	211	4	951
1	205	5	1156
1	176	6	1332
1	168	7	1500
1	164	8	1664
1	155	9	1819
1	134	10	1953
2	120	12	2193
1	115	13	2308
1	105	14	2413
1	102	15	2515
1	96	16	2611
1	85	17	2696

cont.

Table IV.6.5 (continued)

# journals	corresponding # articles	r	R(r) (observed)
cont.			
1	80	18	2776
2	79	20	2934
1	78	21	3012
1	74	22	3086
1	64	23	3150
1	63	24	3213
2	60	26	3333
1	59	27	3392
1	53	28	3445
1	52	29	3497
2	51	31	3599
1	45	32	3644
1	44	33	3688
2	42	35	3772
1	40	36	3812
2	38	38	3888
1	36	39	3924
2	33	41	3990
1	32	42	4022
5	31	47	4177
1	30	48	4207
1	29	49	4236
1	28	50	4264
1	27	51	4291
1	25	52	4316
3	24	55	4388
1	23	56	4411
6	22	62	4543
2	21	64	4585
5	20	69	4685
4	19	73	4761
8	18	81	4905
5	17	86	4990
3	16	89	5038
4	15	93	5098
7	14	100	5196
10	13	110	5326
9	12	119	5434
9	11	128	5533
7	10	135	5603
8	9	143	5675
12	8	155	5771
20	7	175	5911
14	6	189	5995
35	5	224	6170
45	4	269	6350
68	3	337	6554
140	2	477	6834
534	1	1011	7368

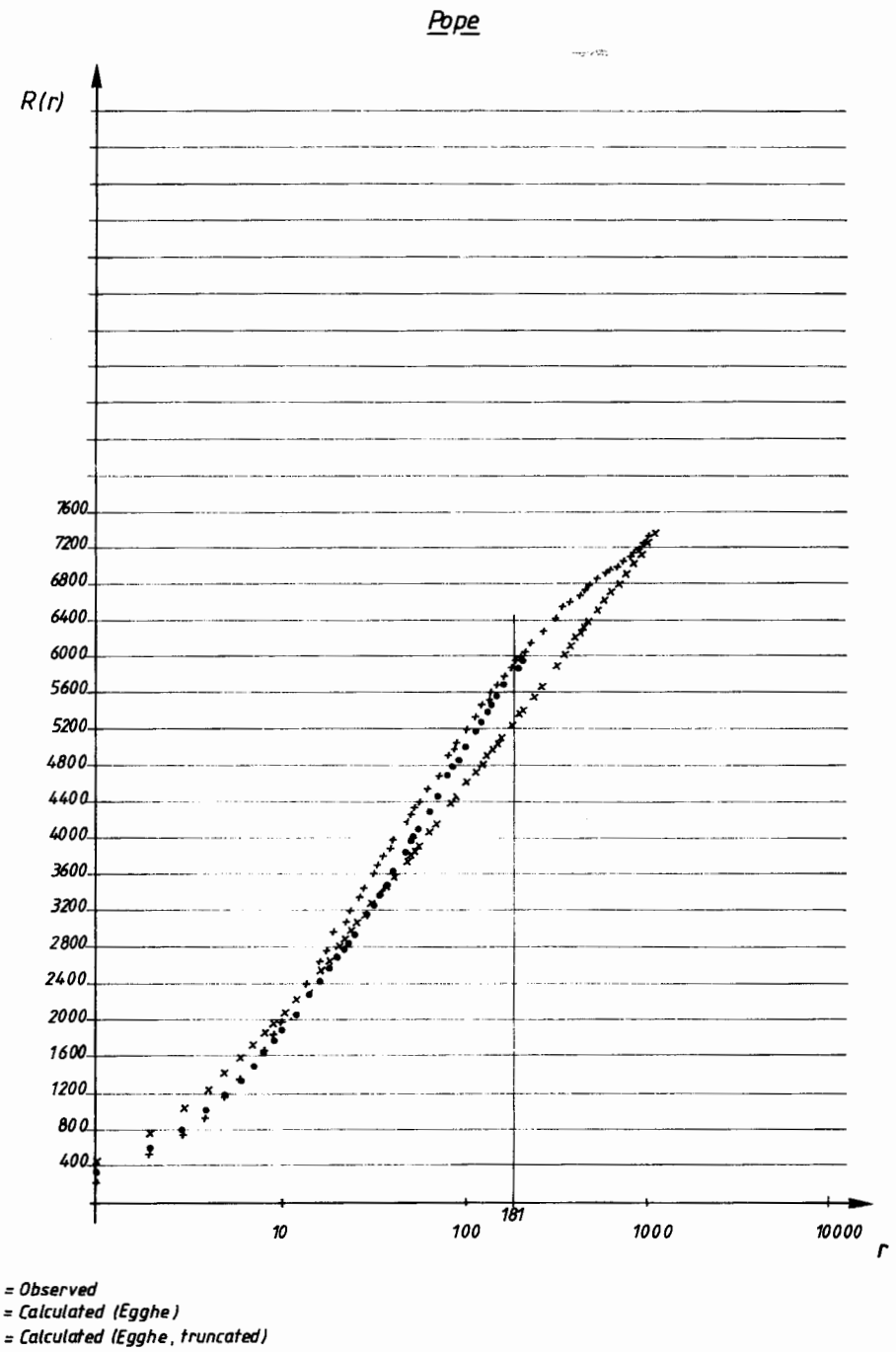

Fig.IV.6.1 Fittings of Pope's bibliography

In calculating Leimkuhler's function for Pope's bibliography, we take for instance p = 5. Then one has y_0 = 1473.6, k = 3.415, r_0 = 5.27, a = 1199.8 and b = 0.4584. This gives the function

$$R(r) = 1199.8 \log (1 + 0.4584\ r)\ . \qquad [IV.6.4]$$

See Fig.IV.6.1 (disregard the dotted curve • for the moment) for a comparison of the observed and calculated data. The fit is reasonable, taking into account the large Groos droop; cf. Groos (1967). How to 'cut off' this droop is the subject of the next section.

See also Brookes (1985) for other, more 'ad hoc' fitting methods for Leimkuhler's law.

IV.6.3. Fitting the first part of Leimkuhler's function

IV.6.3.1. Comments on the Groos droop

The above method for calculating Leimkuhler's function is very good, at least for IPP's not showing any 'Groos droop'. This is logical since Leimkuhler's function

$$R(r) = a \log (1 + br) \qquad [IV.4.2]$$

does not involve a Groos droop. Indeed, the graph of [IV.4.2] on a semi-logarithmic scale (log r, R(r)) looks like Fig.IV.6.2.

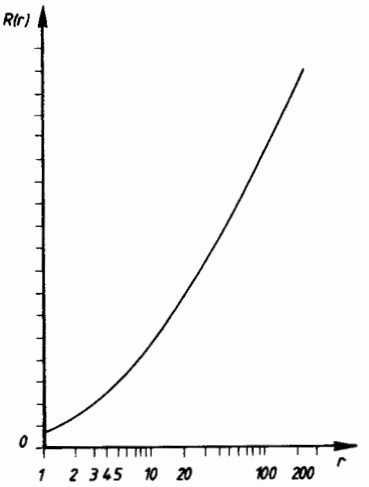

Fig.IV.6.2 Leimkuhler's function expressed graphically

We have :

$$\lim_{r\to\infty} \frac{dR(r)}{d\log r} = \lim_{r\to\infty} \frac{abr}{1+br} = a \quad ,$$

a constant.

It is a well-known fact that most practical examples of IPP's show a Groos droop (small or large); see, for example, Pope's bibliography. Other examples can be found in Aiyepeku (1977), Brookes (1969), Brookes (1973), Brookes (1977b), Brown (1977), Drott et al. (1979), Egghe (1985), Groos (1967) (although the term 'droop' was coined by B.C. Brookes), Lipatov and Denisenko (1986), Praunlich and Kroll (1978), Saracevic and Perk (1973), Singleton (1976a), Asai (1981), Avramescu (1980a), Brookes (1980a) and Haspers (1976).

A Groos droop can be defined exactly, as the occurrence of an *inflection point* r_d in the curve of the function R(r) on a semi-logarithmic scale :

$$\frac{d^2r}{(d\log r)^2}(r_d) = 0 \; . \qquad \text{[IV.6.5]}$$

This results in the graph shown in Fig.IV.6.3.

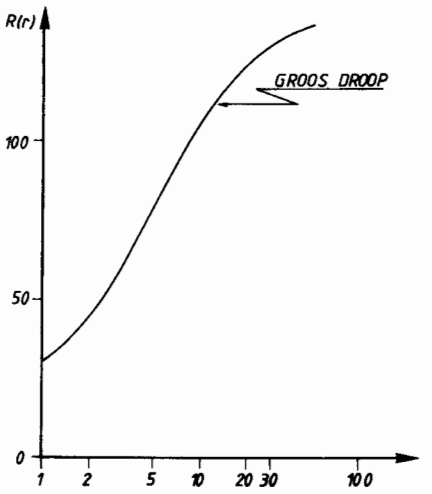

Fig.IV.6.3 The Groos droop

This model is not included in Leimkuhler's function [IV.4.2] (the case in which $\alpha = 2$) but is included in the generalised Leimkuhler function [IV.4.43], for $\alpha < 2$.

Two approaches are possible.

1. Try to model the Groos droop with function [IV.4.43] (or any other well-fitting function). This droop is then explained in so far as [IV.4.43] is explained. This reduces to the explanation of Lotka's law

$$f(j) = \frac{C}{j^{\alpha}} \quad ,$$

$j \in [\rho(0),\rho(A)] = [1,y_m]$. This has been given in Section IV.2.2. The fitting of function [IV.4.43] will be done in the last chapter of this Part IV.

2. Accept Leimkuhler's function [IV.4.2] as a good law for modelling certain 'pure' informetric phenomena and then try to explain *deviations* from it, due to several causes. In Egghe and Rousseau (1988b), this last approach has been taken. There we encountered the following possible explanations for the Groos droop (from the 'deviations' point of view) :
   - *Incompleteness* of the IPP.
   - *Merging* of IPP's (see also Egghe (1989a) for an exact definition of merging of IPP's and Rousseau (1989e) for a merging model). Interpretations of merging are : *interdisciplinarity* of the subject (cf. Pope's bibliography) or bibliographies ranging over a very *long time period* ('osmotic' merging).

Especially in the case of incomplete IPP's, one might be interested in having the completed (unknown) IPP. The main idea behind the solution to this problem is that all the important sources (lower ranks), are certainly known in the incomplete IPP. Consequently, we say that the beginning of the Leimkuhler curve R is correct and that the incompleteness occurs where the Groos droop starts. Therefore, in Egghe (1989e and 1989a) we invented the following '*cutting-off*' method.

IV.6.3.2. Methodology of 'cutting-off'

- Choose a preliminary cutt-off rank ρ_0 at which the Groos droop becomes apparent and check the production (the number of items) of the source at this rank, say n.
- Choose a number p of Bradford groups for the unknown IPP without a Groos droop. Take p high enough (e.g. p = 10) so that 'interpolation' can take place until rank $r = \rho_0$ is reached (an explanation follows below).
- The Bradford factor for the complete IPP is determined as before :

$$k = (1.781\ y_m)^{1/p} \quad . \qquad [IV.5.3]$$

- Based on equation [IV.5.4], calculate the (decimal) number of groups q that are linked to n :

$$n = \frac{k^q}{e^\gamma} \quad .$$

Hence

$$q = \frac{\gamma + \log n}{\log k} \quad . \qquad [IV.6.6]$$

- Thus, the source on rank $r = \rho_0$ belongs to the $([q]+1)^{th}$-last Bradford group.

- Since we later need a whole number of groups, we will take our cut-off point a little lower in rank (not larger, in order to exclude the Groos droop). This means that we take the source with the highest rank in the $([q]+1)^{th}$-group. This is calculated by again using equation [IV.5.4] :

$$n' = \frac{k^{[q]+1}}{e^\gamma} \quad . \qquad [IV.6.7]$$

The number n' determines the final cut-off rank r'.

- What is left after truncation at rank r' contains p - [q] - 1 Bradford groups. See Fig.IV.6.4.

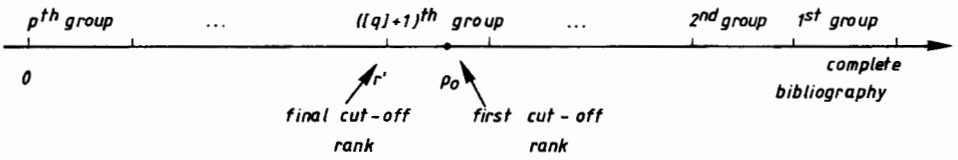

Fig.IV.6.4 Geometry of Bradford groups

- We are now in a position to calculate all parameters of Leimkuhler's function for the complete Bradford distribution, based on our truncated one. The number of sources $r' = \hat{T}$ and the number $\hat{A}$ of items in the truncated IPP are of course known directly from the table of observed data.

- Since $\hat{A}$ items are divided over $p - [q] - 1$ groups and since all groups (even for the complete IPP) contain y_0 items, we have

$$y_0 = \frac{\hat{A}}{p - [q] - 1} \quad . \qquad \text{[IV.6.8]}$$

- Since y_0 and k for the complete IPP are now known, we already have

$$a = \frac{y_0}{\log k} \quad . \qquad \text{[IV.4.5]}$$

- Since every Bradford group contains $r_0, r_0k, r_0k^2, \ldots, r_0k^{p-1}$ sources respectively, the truncated IPP contains

$$\hat{T} = r_0 + r_0k + \ldots + r_0k^{p - [q] - 2}$$

sources (because in the truncated IPP there are $p - [q] - 1$ groups). Hence r_0 is also found :

$$r_0 = \frac{\hat{T}}{1 + k + \ldots + k^{p - [q] - 2}} \quad ,$$

$$r_0 = \frac{\hat{T}(k-1)}{k^{p - [q] - 1} - 1} \quad . \qquad \text{[IV.6.9]}$$

- From this we finally derive

$$b = \frac{k - 1}{r_0} \qquad \text{[IV.4.6]}$$

and

$$R(r) = a \log (1 + br) \quad , \qquad \text{[IV.4.2]}$$

representing Leimkuhler's function for the unknown IPP without a Groos droop.

IV.6.3.3. Example

We once again make use of the example of Pope's bibliography in which a Groos droop is very apparent (see Fig.IV.6.1). We propose cutting at about rank r = 185 (although better fits might be obtained when cutting at rank 50 or so; we leave this exercise to the reader). Here n = 6. Take p = 10, in which case k = 1.848. Now q = 3.86, so that [q] = 3 and [q]+1 = 4, n' = 6.55.

Lastly, we take rank r' = 189 - 0.55 (189-175) ≈ 181. Hence $\hat{T}$ = 181 and so, according to table IV.6.5, $\hat{A}$ = 5947. Thus y_0 = 991.2, a = 1614.0, r_0 = 3.95 and b = 0.215. This yields the function

$$R(r) = 1614 \log (1 + 0.215\ r) . \qquad [IV.6.10]$$

We see from Fig.IV.6.1 (dotted line) that the fit is much better now. Stated earlier, better fits are possible when cutting off earlier. Indeed, as we can see in Fig.IV.6.1, the Groos droop is also present in the ranks before 181.

This method has been applied in Rousseau (1987a). In the same paper a good definition of the nuclear zone of a Leimkuhler curve is proposed. A p-nucleus is defined within which the slope of the curve is less than the proportion p of its maximum value. This definition is scale invariant. A 0.75-nucleus is proposed for practical applications. This nucleus consists of the first $\left[\frac{3}{b}\right]$ sources. For Pope's bibliography this yields a core consisting of $\left[\frac{3}{b}\right] = \left[\frac{3}{0.215}\right] \approx 14$ journals.

Other examples and extensions of the above cutting-off method have been given in Egghe (1989e) and Egghe (1989a).

In Ravichandra Rao (1989) this 'cutting-off' method has been used in the modelling of journal productivity in economics.

In so far as a Groos droop is caused by incompleteness, the above references also include methods to estimate the upper bound on the size of the completed (unknown) bibliography.

IV.6.3.4. Note on the arcs near the end of a Leimkuhler curve

There are frequently several high-ranking sources that provide the same number of items : there might be a rather large number of sources yielding three items, a larger number yielding two items and an even larger number of sources yielding only one item each.

Since the increase of R(r) at these ranks is linear in r (per group of equal productivity), the graph of R versus log r is exponential (per group of equal productivity). These exponential graphs get more visible as the groups of sources with equal productivity get longer. This explains the *arcs near the end of a Leimkuhler curve* frequently encountered in practice; see, for example, Warren and Newill (1967), Brookes (1973), Praunlich and Kroll (1978), Wilkinson (1973), Summers (1983) and Fig.IV.6.5.

This phenomenon is a purely mathematical consequence and has nothing to do with the above-described Groos droop.

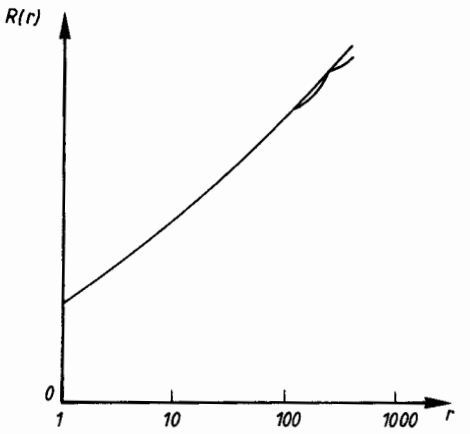

Fig.IV.6.5 The arcs near the end of a Leimkuhler curve

IV.6.4. Fitting of the generalised Leimkuhler and Lotka functions

IV.6.4.1. Methodology

This final section is devoted to the fitting of the general Leimkuhler function

$$R(r) = \frac{C}{2-\alpha}\left[y_m^{2-\alpha} - \left(y_m^{1-\alpha} - \frac{1-\alpha}{C}\, r\right)^{\frac{2-\alpha}{1-\alpha}}\right], \qquad [IV.5.2\ ]$$

where $r = 1,2,\ldots,T$.
Of course this immediately gives rise to the problem of fitting Lotka's function ($\alpha > 1$) :

$$f(j) = \frac{C}{j^{\alpha}}, \qquad [IV.3.17]$$

where $j = 1,2,\ldots,y_m$.

Note that if $\alpha < 2$, the function [IV.5.2] has a Groos droop. Other functions have been proposed to fit this phenomenon, but their distributions are usually complicated and left unexplained; see, for example, Griffith

(1988) and Sichel (1986).

Several papers have been devoted to fitting Lotka's function (for general α) : Nicholls (1986), Nicholls (1987), Pao (1982), Pao (1985), Pao (1986) and Tague and Nicholls (1987). They have devised a few methods for deriving a good α, some better than others.

The whole problem of fitting [IV.5.2] and [IV.3.17] can actually be reduced to find the 'best' α. Once α determined, C follows as indicated below :

$$T = \sum_{j=1}^{y_m} f(j) ,$$

$$T = C \sum_{j=1}^{y_m} \frac{1}{j^{\alpha}} ,$$

where $\alpha > 1$. Since $\sum_{j=1}^{\infty} \frac{1}{j^{\alpha}}$ converges ($\alpha > 1$) we have

$$C \approx \frac{T}{\zeta(\alpha)} , \qquad \text{[IV.6.11]}$$

where $\zeta(\alpha)$ denotes the classical zeta function. Since T is known, C can be determined from a table of $\zeta(\alpha)^{-1}$, as given, for instance, in Nicholls (1987), but extended and reproduced here since we need it further on : see Table IV.6.6.

Since y_m is also known, we now see that, once α is known, all parameters in [IV.5.2] and [IV.3.17] are known.

In what follows, we will suffice to investigate whether some α and C that yield a well fitting Lotka function [IV.3.17] (i.e. they fit the practical data well) will also yield a well fitting general Leimkuhler function [IV.5.2]. We will give two examples.

IV.6.4.2. Examples

Example 1 : The Murphy data

These examples can be found in Murphy (1973) as well as in Pao (1986) or Rao (1980); see Table IV.6.7. For these data, the least squares method (Nicholls (1986)) yields $\alpha = 2.104$ and $C/T = 0.6424$. Lotka's function fits well. With this α and C, equation [IV.5.2] also has a good fit.

Here $D_{max} = 0.0665$, but the 5 % critical value is approximately $\frac{1.36}{\sqrt{238}} = 0.0882$. We can accept our general Leimkuhler function :

Table IV.6.6. Table of $\frac{C}{T} = \frac{1}{\zeta(\alpha)}$ for $\alpha \in [1.11, 3.49]$ with increments of 0.01

α	C/T	α	C/T	α	C/T	α	C/T	α	C/T	α	C/T
		1.50	0.3828	1.90	0.5715	2.30	0.6981	2.70	0.7848	3.10	0.8450
1.11	0.1033	1.51	0.3885	1.91	0.5753	2.31	0.7007	2.71	0.7866	3.11	0.8463
1.12	0.1121	1.52	0.3942	1.92	0.5791	2.32	0.7033	2.72	0.7883	3.12	0.8475
1.13	0.1208	1.53	0.3998	1.93	0.5828	2.33	0.7058	2.73	0.7901	3.13	0.8488
1.14	0.1294	1.54	0.4054	1.94	0.5865	2.34	0.7083	2.74	0.7918	3.14	0.8500
1.15	0.1378	1.55	0.4109	1.95	0.5902	2.35	0.7108	2.75	0.7935	3.15	0.8512
1.16	0.1462	1.56	0.4163	1.96	0.5938	2.36	0.7133	2.76	0.7952	3.16	0.8524
1.17	0.1545	1.57	0.4217	1.97	0.5974	2.37	0.7157	2.77	0.7969	3.17	0.8536
1.18	0.1627	1.58	0.4270	1.98	0.6009	2.38	0.7181	2.78	0.7986	3.18	0.8547
1.19	0.1708	1.59	0.4323	1.99	0.6044	2.39	0.7205	2.79	0.8003	3.19	0.8559
1.20	0.1788	1.60	0.4375	2.00	0.6079	2.40	0.7229	2.80	0.8019	3.20	0.8571
1.21	0.1868	1.61	0.4427	2.01	0.6114	2.41	0.7252	2.81	0.8035	3.21	0.8582
1.22	0.1946	1.62	0.4478	2.02	0.6148	2.42	0.7276	2.82	0.8052	3.22	0.8593
1.23	0.2024	1.63	0.4528	2.03	0.6182	2.43	0.7299	2.83	0.8068	3.23	0.8605
1.24	0.2100	1.64	0.4578	2.04	0.6215	2.44	0.7322	2.84	0.8083	3.24	0.8616
1.25	0.2176	1.65	0.4628	2.05	0.6249	2.45	0.7344	2.85	0.8099	3.25	0.8627
1.26	0.2251	1.66	0.4677	2.06	0.6281	2.46	0.7367	2.86	0.8115	3.26	0.8638
1.27	0.2325	1.67	0.4725	2.07	0.6314	2.47	0.7389	2.87	0.8130	3.27	0.8649
1.28	0.2399	1.68	0.4773	2.08	0.6346	2.48	0.7411	2.88	0.8145	3.28	0.8660
1.29	0.2471	1.69	0.4821	2.09	0.6378	2.49	0.7433	2.89	0.8161	3.29	0.8670
1.30	0.2543	1.70	0.4868	2.10	0.6409	2.50	0.7454	2.90	0.8176	3.30	0.8681
1.31	0.2614	1.71	0.4914	2.11	0.6441	2.51	0.7476	2.91	0.8191	3.31	0.8691
1.32	0.2685	1.72	0.4961	2.12	0.6472	2.52	0.7497	2.92	0.8205	3.32	0.8702
1.33	0.2754	1.73	0.5006	2.13	0.6502	2.53	0.7518	2.93	0.8220	3.33	0.8712
1.34	0.2823	1.74	0.5051	2.14	0.6533	2.54	0.7539	2.94	0.8235	3.34	0.8723
1.35	0.2891	1.75	0.5096	2.15	0.6563	2.55	0.7560	2.95	0.8249	3.35	0.8733
1.36	0.2958	1.76	0.5140	2.16	0.6593	2.56	0.7580	2.96	0.8263	3.36	0.8743
1.37	0.3025	1.77	0.5184	2.17	0.6622	2.57	0.7600	2.97	0.8277	3.37	0.8753
1.38	0.3090	1.78	0.5227	2.18	0.6651	2.58	0.7620	2.98	0.8291	3.38	0.8763
1.39	0.3156	1.79	0.5270	2.19	0.6680	2.59	0.7640	2.99	0.8305	3.39	0.8772
1.40	0.3220	1.80	0.5313	2.20	0.6709	2.60	0.7660	3.00	0.8319	3.40	0.8782
1.41	0.3284	1.81	0.5355	2.21	0.6737	2.61	0.7680	3.01	0.8333	3.41	0.8792
1.42	0.3347	1.82	0.5397	2.22	0.6766	2.62	0.7699	3.02	0.8346	3.42	0.8801
1.43	0.3409	1.83	0.5438	2.23	0.6793	2.63	0.7718	3.03	0.8360	3.43	0.8811
1.44	0.3471	1.84	0.5479	2.24	0.6821	2.64	0.7737	3.04	0.8373	3.44	0.8820
1.45	0.3532	1.85	0.5519	2.25	0.6848	2.65	0.7756	3.05	0.8386	3.45	0.8830
1.46	0.3592	1.86	0.5559	2.26	0.6875	2.66	0.7775	3.06	0.8399	3.46	0.8839
1.47	0.3652	1.87	0.5599	2.27	0.6902	2.67	0.7793	3.07	0.8412	3.47	0.8848
1.48	0.3711	1.88	0.5638	2.28	0.6929	2.68	0.7811	3.08	0.8425	3.48	0.8857
1.49	0.3770	1.89	0.5677	2.29	0.6955	2.69	0.7830	3.09	0.8438	3.49	0.8866

$$R(r) = \frac{0.6424 \times 170}{-0.1047} [5^{-0.1047} - (5^{-1.1047} + \frac{1.1047}{0.6424 \times 170} r)^{0.0948}], \quad [IV.6.12]$$

$$R(r) \approx -1043.052 [0.8449 - (0.1690 + 0.0101 r)^{0.0948}] .$$

Table IV.6.7. The Murphy data

r	R(r) observed	R(r) calculated via [IV.6.12]
1	5	4.9
2	9	9.5
3	13	13.9
4	17	18.1
5	21	22.1
6	25	26.0
7	29	29.7
8	33	33.3
9	37	36.7
10	40	40.0
11	43	43.3
12	46	46.4
13	49	49.4
14	52	52.3
15	55	55.2
16	58	57.9
17	61	60.6
18	64	63.2
19	66	65.8
20	68	68.3
30	88	90.2
40	108	108.2
50	118	123.6
70	138	148.9
90	158	169.3
110	178	186.6
130	198	201.5
150	218	314.7
170	238	226.5

Example 2 : The Radhakrishnan-Kerdizan data

These examples can be found in Radhakrishnan and Kerdizan (1979); see also Pao (1986) and Table IV.6.8.

In this case the Nicholls least-squares method yields $\alpha = 3.4880$ and $C/T = 0.8864$. The maximum likelihood method (Nicholls (1986)) gives $\alpha = 3.4000$ and $C/T = 0.8782$. Both methods give a fit to Lotka's function [IV.3.17] (although not a splendid one), but a very bad fit to Leimkuhler's function [IV.5.2]. In this case, we propose using another simple method : Estimate C

by

$$f(1) = C \ . \qquad \text{[IV.6.13]}$$

Here this is 250.

Table IV.6.8. The Radhakrishnan-Kerdizan data

r	R(r) observed	R(r) calculated via [IV.6.15]
1	7	6.4
2	13	12.0
3	18	17.0
4	22	21.6
5	26	25.8
6	30	29.8
7	34	33.5
8	38	37.1
9	41	40.4
10	44	43.7
11	47	46.8
12	50	49.8
13	53	52.7
14	56	55.5
15	59	58.2
20	69	70.8
30	89	92.3
40	109	110.7
50	129	124.1
51	131	125.6
52	132	127.0
100	180	191.5
200	280	283.5
300	380	354.2
301	381	354.3

This produces

$$\frac{C}{T} = \frac{250}{T} = \frac{250}{301} = 0.8306 \ .$$

Using this and Table IV.6.6, we get

$$\alpha = 2.9907$$

These values not only results in a good fit of Leimkuhler's function, but the fitted Lotka function ($\varphi = \frac{f}{T}$),

$$\varphi(j) = \frac{0.8306}{j^{2.9907}} \qquad [IV.6.14]$$

is better than the least squares (LS) or maximum likelihood (ML) methods in Nicholls (1986). For Lotka's fitting we obtain $D_{max} = 0.0151$, which is smaller than Nicholls' fits :

$$LS : D_{max} = 0.0367 \ ,$$

$$ML : D_{max} = 0.0285 \ .$$

For Leimkuhler's general function [IV.5.2] we obtain $D_{max} = 0.086$ (much better than Nicholls'), which is at about the 1 % level. Consequently, at the 1 % level, we have a fit (unlike Nicholls). We have here the function

$$R(r) = \frac{0.8378 \times 301}{-1.0442} \left[7^{-1.0442} - \left(7^{-2.0442} + \frac{2.0442}{0.8378 \times 301} r\right)^{0.5108}\right]$$

$$R(r) \approx -241.4923 \left[0.1311 - (0.0187 + 0.0081\ r)^{0.5108}\right] . \qquad [IV.6.15]$$

This shows that the above simple method deserves to be investigated more closely.

IV.7. APPLICATIONS

What are applications of the informetric laws? First of all, we have seen that these laws have an explanatory function, raising the status of informetrics from a technique to a scientific theory. This is a very important (theoretical) application indeed. Next, once an informetric law has been accepted, one can deduce new properties from this law to be discussed in this Section. In these cases it is sufficient to determine only the parameters (such as C and α in Lotka's law). This saves time and does not require too much data.

IV.7.1. Aspects of concentration theory, 80/20-rule, Price's law, concentration measures

In a manner of speaking, IPP's represent very elite situations : the distribution functions are very skew in the sense that many sources have a few items and a few sources have many items (the latter sources being the geniuses where the sources are authors, or top journals when the sources are journals). These sources form two groups, divided by a 'middle' group of sources with an 'average' production.

Bradford himself must have shared these ideas since he always used $p = 3$ groups (cf. Bradford (1934)). The most important sources (in the first group) form the 'nucleus'. If we think of sources as journals, these sources will be bought in any case. The middle group will be bought if there are sufficient funds. The last group can be skipped. Nevertheless this simplistic reasoning is not an application of Bradford's law : one merely has to divide the articles (items) in the bibliography into 3 equal parts. One only has to look at the corresponding sources to obtain the required division, without applying Bradford's law.

One IPP is more concentrated than another. This is expressed by the fact that one IPP has a few sources with a very high number of items. This could be expressed by using the Bradford factor k (where a high k represents a very concentrated situation). However, k is not a good indicator of concentration since there is no basis for comparison and since k is p-dependent. The following three Subsections describe good ways of dealing with concentration.

IV.7.1.1. The 80/20-rule

We briefly repeat here the 80/20-rule (as was introduced earlier in Section II.6.2).

Take an arbitrary discrete IPP (e.g. a bibliography). Order, as usual, the sources in decreasing order according to the number of items they contain.

The *80/20-rule* states that 20 % of the most important sources will contain 80 % of all the items. Of course this is only a historic formulation; one can generalise this rule as follows :

Arithmetic expression of concentration :

100 x % of the sources will produce 100 θ % of the items, and we look for the function

$$x = x(\theta) \quad . \qquad [IV.7.1]$$

The following equation is given in Egghe (1986b), supposing a Bradfordian IPP :

$$x = \frac{6}{\pi^2} e^{\gamma - \frac{\pi^2}{6}\mu(1-\theta)} \quad , \qquad [IV.7.2]$$

where μ is the average number of items per source (i.e. $\mu = \frac{A}{T}$) and $e \approx 2.7183$, $\pi \approx 3.1416$, $\gamma \approx 0.5772$. We refer to Egghe (1986b) for a discussion and application of this equation (cf. also Burrell (1985a)).

IV.7.1.2. Price's law

Price (1971,1976) states (see also Allison et al. (1976)) that, if there are N sources in the IPP, then $\sqrt{N}$ of the top sources will have 50 % of the items. Stated otherwise, $N^{1/2}$ sources yield a fraction $\frac{1}{2}$ of the items. This phenomenon is associated with the occurrence of invisible colleges (see also Price and Beaver (1966)), i.e. the hierarchical elites in fields (or subfields) of science. As the 80/20-rule, this '*Price's law*' is too simple, as was already shown in Allison et al. (1976) and Nicholls (1988). We therefore formulate the following

Geometric expression of concentration

Let there be N sources. Then the N^{α} ($0 < \alpha < 1$) top sources will have 100 θ % of the items, and we look for the function

$$\alpha = \alpha(\theta) \quad . \qquad [IV.7.3]$$

Note that even $\alpha = \theta$ would generalise the above Price's law (which states in addition that $\alpha = \theta = \frac{1}{2}$). Egghe (1987b) and Egghe and Rousseau (1986) investigated this function, concluding that in most cases

$$\theta \leq \alpha \leq \frac{1 + \theta}{2} \qquad [IV.7.4]$$

For instance, for $\theta = 0.8$ one has $0.8 \leq \alpha \leq 0.9$ and certainly for θ high (i.e. close to 1) : $\alpha \approx \theta$.

IV.7.1.3. Concentration measures

So far, 'concentration' has been expressed by two parameters : arithmetically via (x,θ) and geometrically via (α,θ). One might also wonder what the requirements are for a single measure to be a good *concentration measure*. This has been investigated in Egghe and Rousseau (1988) mainly as a consequence of Allison (1978).

IV.7.1.3.1. Axioms of concentration measures

In the general situation, there are N 'boxes' each containing x_i ($i = 1,\ldots,N$) 'balls'. To cite an example, N = the number of sources (i.e. $N = T$) and x_i = the number of items in the i^{th} source (assuming the sources are numbered). While more general interpretations of the sequence $x_1,x_2,\ldots,x_N$ are possible, this interpretation suffices for this section.

A concentration measure is then a function of the N variables $x_1,\ldots,x_N$:

$$f : (x_1,x_2,\ldots,x_N) \to f(x_1,x_2,\ldots,x_N) \quad . \qquad [IV.7.5]$$

We formulate the following requirements for a function f to be a good concentration measure. Each requirement (axiom) is followed by an interpretation in econometrics since the 'elite' terminology 'rich' and 'poor' greatly facilitates our imagination.

(C1) If all x_i are *equal*, say to $c \neq 0$, then $f(x_1,\ldots,x_N)$ attains its minimum value, equal to 0.

This is a perfectly natural condition, since there is no concentration. Note also that (C1) implies that a concentration measure is never negative.

(C2) For every $(x_1,\ldots,x_N)$ and every *permutation* $\pi : \{1,\ldots,N\} \to \{1,\ldots,N\}$ we require that

$$f(x_1,\ldots,x_N) = f(x_{\pi(1)},\ldots,x_{\pi(N)}) \quad . \qquad [IV.7.6]$$

This principle expresses the fact that, to use our example, the poverty (or wealth) of a nation is not a labelled property. It is only determined by the overall configuration.

(C3) *Scale invariance.* This principle states that for every $(x_1,\ldots,x_N)$ and $c > 0$:

$$f(cx_1,\ldots,cx_N) = f(x_1,\ldots,x_N) \ . \qquad [IV.7.7]$$

It expresses the requirement that a good concentration measure should not be influenced by the units. Returning to the case of income distributions, this means that there may not be a difference, regardless of whether the income is calculated in dollars, yen or rupees.

(C4) '*When the richest source gets richer, inequality rises*'. This principle is a very natural one. It has two requirements : the first is the one mentioned above. The second is its dual : when the poorest source gets poorer, inequality also increases. In a mathematical formulation this becomes the following :

(C4a) If $x_i = \max \{x_1,\ldots,x_N\}$ and if there exists a $k \neq i$ such that $x_k \neq 0$, then, for $h > 0$,

$$f(x_1,\ldots,x_i+h,\ldots,x_N) > f(x_1,\ldots,x_N) \ . \qquad [IV.7.8a]$$

(C4b) If $x_j = \min \{x_1,\ldots,x_N\}$ and $0 < h \leq x_j$ then

$$f(x_1,\ldots,x_j-h,\ldots,x_N) < f(x_1,\ldots,x_N) \ . \qquad [IV.7.8b]$$

(C5) *The principle of nominal increase.* This principle requires that an equal, nominal increase in each source should strictly decrease the global inequality. Stated more formally, this becomes : for every $(x_1,\ldots,x_N)$, where not all x_i are equal and $h > 0$:

$$f(x_1+h,\ldots,x_N+h) < f(x_1,\ldots,x_N) \ . \qquad [IV.7.9]$$

(C6) *The transfer principle.* This principle, postulated by Dalton (1920), states that if we make a strictly positive transfer from a poorer source to a richer one, this must lead to a strictly positive increase in the index of inequality. Formulated in a precise mathematical way, this becomes : if $x_i \leq x_j$ and $0 < h \leq x_i$, then

$$f(x_1,\ldots,x_i,\ldots,x_j,\ldots,x_N) < f(x_1,\ldots,x_i-h,\ldots,x_j+h,\ldots,x_N) \ . \qquad [IV.7.10]$$

We note that such a transfer leaves the arithmetic mean unchanged. In Egghe and Rousseau (1990) it is shown that (C6) implies (C5) and (C4), assuming

(C3).

It could also be argued that a good measure of concentration should vary between 0 and 1 :

Principle (B)

For all $(x_1,\ldots,x_N)$

$$0 \leq f(x_1,\ldots,x_N) \leq 1 \quad . \qquad \text{[IV.7.11]}$$

However, this principle is only a mathematical convenience. It should not imply any preference, as simple transformations can produce any desired bounds. If a measure f is positive and does not satisfy the requirement that $f \leq 1$, then we can use the transformation

$$f \to \frac{f}{1+f} \quad .$$

This yields an increasing function of f with values in the interval [0,1]. The transformed function satisfies (C1) to (C6) if f does.

IV.7.1.3.2. Examples of good and bad concentration measures

1. It is intuitively clear that the classical notions of *standard deviation* (σ) and *variance* (σ^2), where

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 = \frac{1}{N}\sum_{i=1}^{N}x_i^2 - \mu^2$$

$$= \frac{1}{2N^2}\sum_{k=1}^{N}\sum_{\ell=1}^{N}(x_k-x_\ell)^2 \qquad \text{[IV.7.12]}$$

(and μ denotes the mean of the distribution), bear some relation to the concept of concentration.

2. The *coefficient of variation*

$$V = \frac{\sigma}{\mu} \qquad \text{[IV.7.13]}$$

was introduced to deal with relative instead of absolute values. Similarly we can consider

$$V^2 = \frac{\sigma^2}{\mu^2} \qquad \text{[IV.7.14]}$$

or *Gaston's measure* (Gaston (1978))

$$Ga = \frac{\sigma^2}{\mu} \qquad [IV.7.15]$$

or *Allison's modified squared coefficient* (Allison (1980))

$$A = \frac{\sigma^2 - \mu}{\mu^2} \quad . \qquad [IV.7.16]$$

3. From linguistics we consider the *Yule characteristic* defined as

$$K = \frac{\sigma^2}{\mu^2 N} = \frac{V^2}{N} \quad . \qquad [IV.7.17]$$

Johnson (1979) advocates the use of *Simpson's index* (Simpson (1949)) in stylistic studies :

$$J = \frac{\sum_{i=1}^{n} i(i-1)\, x_i}{n(n-1)} \quad , \qquad [IV.7.18]$$

where x_i is the number of words that occur i times and n is the total number of words that occur in the text being investigated. Simpson's index is nothing but the number of identical pairs divided by the number of all possible pairs.

4. The *Schutz coefficient* (relative mean deviation) (Schutz (1951)) is

$$D = \frac{\frac{1}{N} \sum_{i=1}^{N} |x_i - \mu|}{2\mu} \quad . \qquad [IV.7.19]$$

According to Gastwirth (1972), this measure was first proposed by Yntema (1933) and Pietra in the 1930's.

5. *Pratt's measure* and the *Gini index*. In order to define Pratt's measure we first assume that the x_i's are ordered in decreasing order. Taking

$$a_i = \frac{x_i}{\sum_{i=1}^{N} x_i}$$

and

$$q = \sum_{i=1}^{N} i\, a_i \quad , \qquad \text{[IV.7.20]}$$

Pratt's measure C is defined as (Pratt (1977)) :

$$C = \frac{2\,(\frac{N+1}{2} - q)}{N-1} \quad ; \qquad \text{[IV.7.21]}$$

Gini's index is then

$$G = \frac{N-1}{N}\, C \; . \qquad \text{[IV.7.22]}$$

Bear in mind, however, that the Gini index was introduced in econometrics (Gini (1909)) long before Pratt's measure was defined. Relation [IV.7.22] was established in 1979 by Carpenter (1979). The usual definition of Gini's index uses the so-called Lorenz curve; see Egghe and Rousseau (1989).

6. *Theil's measure* (Theil (1967)). This inequality measure is defined as :

$$Th = \frac{1}{N} \sum_{i=1}^{N} (\frac{x_i}{\mu}) \log (\frac{x_i}{\mu}) \qquad \text{[IV.7.23]}$$

(cf. the notion of entropy in information theory).
Note further that in this formula one sets $0.\log(0) = 0$.

7. The *variance of logarithms*.

$$L = \frac{1}{N} \sum_{i=1}^{N} (\log(x_i) - \sum_{j=1}^{N} \frac{\log(x_j)}{N})^2 \; ,$$

$$= \frac{1}{2N^2} \sum_{k=1}^{N} \sum_{\ell=1}^{N} (\log x_k - \log x_\ell)^2 \quad , \qquad \text{[IV.7.24]}$$

which is only defined if all $x_i \neq 0$.

8. *Atkinson's index*. Atkinson (1970) introduced a family of concentration measures defined as :

$$A(e) = 1 - (\frac{1}{N} \sum_{i=1}^{N} (\frac{x_i}{\mu})^{1-e})^{\frac{1}{1-e}} \quad , \qquad \text{[IV.7.25]}$$

where $e > 0$ and $e < 1$.
If all $x_i \neq 0$, $A(1)$ is defined as $\lim_{e \to 1} A(e)$, which is nothing but

$$\frac{\mu - GM(x_i)}{\mu} \quad , \qquad \text{[IV.7.26]}$$

as can easily be seen.
Here $GM(x_i)$ denotes the geometric mean of the x_i, i = 1,...,N (see also I.1.4.3). Note that $\lim_{e \to 0} A(e) = 0$. The equation for the Atkinson index as presented in Allison (1978), p.873, seems to be in error.

9. The *CON-index* (Ray and Singer (1973)). In the authors' own words, this index is the standard deviation of the percentage shares divided by the maximum possible standard deviation in a system of size N.
This yields :

$$CON = \sqrt{\frac{\sum_{i=1}^{N} a_i^2 - \frac{1}{N}}{1 - \frac{1}{N}}} \; .$$

Since $a_i = \frac{x_i}{\mu N}$, this formula can be rewritten as follows :

$$CON = \frac{1}{\mu} \sqrt{\frac{\sum_{i=1}^{N} x_i^2 - \mu^2 N}{N(N-1)}} = \frac{1}{\mu} \frac{\sigma}{\sqrt{N-1}} = \frac{V}{\sqrt{N-1}} \qquad \text{[IV.7.27]}$$

(using equations [IV.7.12] and [IV.7.13]).
Equation [IV.7.27] shows that CON is only a variant of the coefficient of variation.

10. *Lotka's* α. Lastly, Rao (1988) pointed out that when data follow Lotka's distribution :

$$f(j) = \frac{C}{j^{\alpha}} \quad , \qquad \text{[IV.3.17]}$$

the exponent α could be used as a measure of concentration.

Egghe and Rousseau (1989) present the proof of the validity of the following table (where Y = has the property and N = does not have the property). Some proofs are long and based on Hardy-Littlewood-Polýa inequalities for convex functions (Hardy, Littlewood and Polýa (1988)).

Table IV.7.1. Properties of concentration measures

	(C1)	(C2)	(C3)	(C4)	(C5)	(C6)
σ,σ^2	Y	Y	N	———	———	———
V,V^2	Y	Y	Y	Y	Y	Y
Ga	Y	Y	N	———	———	———
A	Y	Y	N	———	———	———
K,CON	Y	Y	Y	Y	Y	Y
J	Y	Y	N	———	———	———
D	Y	Y	Y	Y	Y	N
C	Y	Y	Y	Y	Y	Y
G	Y	Y	Y	Y	Y	Y
Th	Y	Y	Y	Y	Y	Y
L	Y	Y	Y	Y	Y	N
A(e) $(0<e<1)$	Y	Y	Y	Y	Y	Y
α	Y	Y	Y	N	———	———

In conclusion, we can state that the following groups of similar measures are good concentration measures :

1) V, V^2, K, CON

2) C, G

3) Th

4) A(e), e > 0.

Egghe and Rousseau (1989) define a new property, the *extended transfer principle*, dealing with transfers over the whole population rather than between two elements, as in (C6). V, C and G are shown to have this strong (C6) property, while Th and A(e) do not. So 'excellent' measures are the ones in the above groups 1 and 2. As shown in Egghe and Rousseau (1989), these two groups can be reformulated in one formula : the measures

$$P(r) = \frac{\left(\frac{1}{2N(N-1)} \sum_{k=1}^{N} \sum_{\ell=1}^{N} |x_k - x_\ell|^r\right)^{\frac{1}{r}}}{\mu} \qquad \text{[IV.7.28]}$$

for $r > 0$. Then it can be seen that $P(1) = C$ and $P(2) = \sqrt{\frac{N}{N-1}}\, V$. The measure P is called the '*generalised Pratt measure*'. All these measures P are 'excellent', as described above. The relation of P(2) with the variance [IV.7.12] and another

property of P(2) (and derivatives, such as V, V^2, K, CON) makes these measures the 'ultimate best' ones (as shown in Egghe and Rousseau (1990)). Furthermore, CON is normalised.

Note :

The aim of *dispersion measures*, such as those studied in Heine (1978), is opposite to concentration measures : they measure the degree of dispersion of a situation $x_1, x_2, \ldots, x_N$. In general, for concentration measures f satisfying $0 \leq f \leq 1$:

$$g = 1 - f$$

is a dispersion measure (and vice versa). Therefore no other theory of dispersion measures is needed.

This remark includes, for instance, Singleton's index; see Heine (1978) or Singleton (1976b). In Egghe and Rousseau (1989) it is shown that this index is nothing but 1 - C.

IV.7.2. Compression of databases

A trivial conclusion is that, when searching codes for words in texts, it is best to assign the shortest codes to the most frequently used words, as then they occupy the least space. Not every coding can be used. After all codes must be able to be decoded. If decodability was not a requirement, the problem would have a trivial solution : emphasising binary codes; i.e. take Ø for the most frequently used word W1 (rank 2), 1 for the word W2 at rank 2, ØØ for the word W3 at rank 3, etc. But this would not be useful, since ØØ can mean W3 as well as W1 W1 and hence ØØ cannot be decoded.

Nowadays different decodable *compression* techniques exist. The optimality depends on the different way the words are used in the text. Knowing, for example, that a text is Zipfian gives us concrete information about selecting the best compression technique. The effect is quite considerable. The interested reader is advised to consult Heaps (1978) and Jones (1979) for the details.

IV.7.3. Style and authorship

Zipf's law, or better, the deviations from this law, can be used to quantitatively determine the *stylistic* properties of texts. Statistical (mainly χ^2-) tests, in particular, can be used to :

- deny an alledged authorship
- put a sequence of texts in chronological order.

The basic idea here is that an individual author has a proper style, which is determined by word choice and word use, and that it more or less deviates from a certain Zipfian pattern. In most cases one uses a derived equation rather than the Zipf function g [IV.1.5] itself.

Example :

The *entropy measure*

$$H = \frac{-\sum_{r=1}^{T} p(r) \log_2 p(r)}{\log_2 T} \quad , \qquad \text{[IV.7.29]}$$

where $p(r) = \frac{g(r)}{T}$, and T is the total number of words in the text. Equation [IV.7.29] is independent of the length of the text, a logical requirement. For more information about these aspects of quantitative linguistics, we refer the reader to the basic works of Herdan (1960) and especially Herdan (1964), where extensive and non-trivial methods for style distinctions are described.

IV.7.4. Storage and text retrieval in a computer

Let us suppose that the words of a text we wish to *store* in a computer comprise T word types (or 'sources' in our terminology).

Let us further suppose that these words (or their code numbers) are stored randomly. Clearly, on the average, the system must check $\frac{T}{2}$ entries in order to find a specific word.

Let us also suppose that the text is Zipfian : to determine the ideas, we take $\beta = 1$ in [IV.1.5] :

$$g(r) = \frac{F}{r} \quad , \qquad \text{[IV.1.5]}$$

where r is the rank of the word. The words are ranked, as usual, in decreasing order of their use in the text. If the words in the text are stored in this order, the system must check r entries in order to search for the word at rank r. The probability for this is expressed as

$$\frac{F}{rA} \ ,$$

where A is the length of the text (or 'items' in our terminology). On the average the system has to check

$$\sum_{r=1}^{T} \frac{F}{rA} \cdot r = \frac{FT}{A} \quad . \qquad \text{[IV.7.30]}$$

But, again using [IV.1.5], we get

$$A = \sum_{r=1}^{T} \frac{F}{r} \approx F \log T \ . \qquad [IV.7.31]$$

Equations [IV.7.30] and [IV.7.31] now yield the average number of entries to be checked, which is

$$\frac{FT}{F \log T} = \frac{T}{\log T} \ , \qquad [IV.7.32]$$

or significantly less than $\frac{T}{2}$ as $T \gg e^2$ for every text. Purely mathematically, it is interesting to note that, for large T, expression [IV.7.32] equals the number of prime numbers which are smaller than or equal to T (see e.g. Rosen (1988)), a remarkable coincidence.

IV.7.5. Bradford's law and sampling

If a bibliography is completely known, it is very simple to determine Bradford's law (see e.g. Section IV.6.1). Of course, Bradford's law itself cannot be used for purposes of collection management. If we know the complete bibliography, it is much simpler to use it directly - without having recourse to any laws - in order to determine the number of sources needed to have, say, a certain percentage of the items.

We are faced with a different problem in the case (encountered frequently in practice) in which the librarian does not have the complete bibliography on a certain topic but only has a *sample* (for time and budgetary reasons!). How can we construct a nucleus of, for example, journals to have, say, 80 % of all the articles on this topic?

This problem has been solved by Tague (1988). To come up with the solution, one needs :

- The Bradford form of the sample.
- The Bradford model of the complete (unknown) bibliography.
- The knowledge of the most important journal (a very natural supposition).

Armed with these ingredients and using an extended 'rule of three', one can estimate the parameters r_0 and k of Bradford's law for the complete bibliography and hence determine the desired nucleus.

In general we can say that *informetric laws are useful for modelling complete (unknown) IPP's based on knowledge of the sampled IPP's* (cf. also Section IV.6.3).

IV.8. NOTES AND COMMENTS

IV.8.1. History

The reader is referred to Cole and Eales (1917) and Hulme (1923) for the first recorded informetric (avant la lettre) studies and to Egghe (1988f), Schmidmaier (1984) and Bonitz (1982) for more details on the history of informetrics. Petruszewyc (1978) deals with the historical developments leading to Pareto's and Zipf's laws (also called 'Estoup's law').

IV.8.2. Explanations

1. Avramescu (1973,1975,1980b) studied information transfer from a physical point of view. He considered the transfer as a diffusion process like heat conduction.

2. Naranan (1970) constructed an informetric model based on the (doubtful) principle (see IV.8.8) of the exponential growth of both the sources and the items. See also Hubert (1976) for a review of this paper. In any case this 'explanation' is based on 'unexplained' assumptions.

3. Karmeshu, Lind and Cano (1984) presented a rationale for Lotka's law based on the random cutting of a square (or of the random crushing of rocks). The different sizes in the end conform to a log-normal distribution.

4. Schubert and Glänzel (1984b) gave an explanation of the so-called Waring distribution. Let $\varphi(j)$ denote the fraction of authors with j publications. Then

$$\varphi(j) = \frac{a(a+k)(a+k-1)\ldots(k+1)}{(a+k+j+1)(a+k+j)\ldots(k+j+1)} \qquad [IV.8.1]$$

Their explanation is based on :
1) a self-reproducing property,
2) a cumulative advantage ('success breeds succes', cf. Section IV.2.1),
3) a uniform leakage.
However, the three proposed functions that describe the above principles are not explained, although, they are admittedly given in the simplest possible forms. They do obtain reasonable fits with practical data. The Waring distribution also appears in Herdan (1960 and 1964) and Irwin (1962) in respectively linguistical and biological contexts; see also Telcs, Glänzel and Schubert (1985) and Schubert and Glänzel (1984a).

5. The principle of least effort (Zipf (1949)) states that a human being will tend to solve problems in such a way that the total work is minimised. However, no clear link between this principle and Zipf's law exists.

6. For a (not very recent) review of explanations of informetric laws we refer the reader to Fedorowicz (1982). This article also cites further evidence for Lotka-type laws. (This includes a derivation from Bose-Einstein's cell occupancy model; see Subsection I.2.5.2). The Bose-Einstein model was already used before by Woodroofe and Hill (1975) in connection with Lotka's law; see also some references in this article.

7. In Tague (1981) a generalisation of the success breeds success principle (generalising an urn-model of Price linked to this principle) yielded the derivation of the negative binomial distribution and the Mandelbrot function.

8. Rao (1980) also refined the negative binomial distribution, based on a success breeds success argument. Bookstein (1979) reviewed explanations of bibliometric laws.

9. The 'Matthew effect' is a socio-psychological phenomenon related to the success breeds success principle and the cumulative advantage effect. It was first described by Merton (1968) (see also Merton (1988) and Bensman (1985)) who gave it its name by referring to the Gospel according to St. Matthew :

'For unto everyone that hath shall be
given, and he shall have abundance;
but from him that hath not shall be
taken away even that which he hath'.

The Matthew effect as it was understood by Merton refers to the habit people have of giving credit (e.g. for scientific discoveries) to already famous people and minimising or withholding recognition for scientists who have not (yet?) made their mark.

10. The informetric distributions discussed in Part II on library circulations are other models, comparable to Lotka's distribution. These models fit library circulation better than the Lotka functions and are derived in a statistical sense. However, no model-theoretical rationales, as developed in this part for Lotka's law, exist for these circulation models. The same can be said about Sichel's Generalised Inverse Gaussian-Poisson (GIGP) distribution (Sichel (1985,1986) and Sichel's earlier work). Especially the Sichel functions are extremely intricate ('if it does not fit well, add another parameter : then the new fit cannot be worse than the old one'), unexplained, and do not really serve the working librarian. Burrell is more conscious about this last aspect (see e.g. Burrell (1980,1988c) and other articles).

For a review of several distributions in relation to document use,

the reader is advised to consult Rao (1982).

11. In Chen and Leimkuhler (1986 and 1987) the laws of Lotka, Bradford and Zipf (or better, the functions they represent) are studied from a discrete index approach and functional relationships are derived. In a manner of speaking, they are discrete analogues of the relationships developed here. Morse and Leimkuhler (1979) deal witj analogues, although this work contains a few speculative relations.

IV.8.3. Zipf - Pareto

Zipf's law (Pareto's law) is included in the equivalent set of informetric laws determined in theorem IV.4.1.1 (the case in which $\alpha = 2$) and theorem IV.4.2.1 (the case in which $\alpha \neq 2$) by taking $H = 1$. Note that in this case, for every $r \in [0,T]$, one has :

$$g(r) = \frac{G}{(1+r)^{\beta}} \qquad [IV.8.2]$$

rather than

$$g(r) = \frac{G}{r^{\beta}} \quad . \qquad [IV.8.3]$$

These two laws do not differ much in practice, but we stress the fact that only the form [IV.8.2] can be used (as form [IV.8.3] is not even defined in $r = 0$). Egghe (1989a and 1989f) also investigated what other laws are equivalent to [IV.8.2]. The well-known law of Brookes (or Weber-Fechner) and the graphical version of Bradford's law (see Wilkinson (1973)) are appearing in this context.

Let us determine the case in which $\alpha = 2$. Taking $H = 1$ for Zipf's law (Pareto's law) implies (see [IV.4.8]) $b = 1$. Then [IV.4.6] gives

$$k = 1 + r_0 \; . \qquad [IV.8.4]$$

Equation [IV.8.4] shows that the cases of IPP's where Zipf's law (Pareto's law) is valid are very concentrated (cf. situations as described in Section IV.7.1). Indeed : either r_0 is small, which is a way of saying (when $p = 3$ is taken to determine the ideas) that the core group of frequently produced sources is small, or r_0 is large. In the latter case, according to [IV.8.4], k must be large and hence the core group of r_0 sources will be small with respect to the other groups r_0k, r_0k^2, etc.

We conclude that linguistics (or econometrics) can be viewed as part of

informetrics, but in practice there is a separation since :

1. In most informetric examples, $b < 1$ and even $b \ll 1$ (see Sections IV.6.2 and IV.6.3, but also many examples in Egghe (1989a,e)).
2. In linguistics and econometrics one often finds $b = 1$. In informetrics we only know one example of a bibliography where $b = 1$: the ORSA-bibliography, cf. Kendall (1960).

Incidentally, $b = 1$ coincides with the distribution called the 'law of anomalous numbers'. This is the rank distribution of occurrence of the numbers with first digit r ($r = 1,\ldots,9$) of numbers as they appear in real life (see Brookes and Griffith (1978), Brookes (1984a) or Feller (1948)).

The following philosophical explanation accounts for the fact that linguistic IPP's are more concentrated than, say, bibliographies (cf. Egghe (1988d,1989a)).

In bibliographies, the most important sources naturally tend to lower the number of items a little, i.e. the most prolific authors will not publish less important (but still publishable) work. Similarly, the most important journals in a research area will become more and more selective in accepting papers, and so on. This is not the case with texts : the most frequently used words are words such as 'the', 'a', 'and', etc. There is no limitation on these words for grammatical reasons! Synonyms are in use only for popular but not so frequently used words.

So this explains again why Zipf's law is a highly concentrated version of Mandelbrot's law.

IV.8.4. Applications

1. An application of informetric laws to the departmental allocation of funds can be found in Bookstein (1988).

2. An application of Zipf's law to the calculation of the entropy of a language is given in Yavuz (1974).

3. Pratt (1977) also introduces a measure of 'relative concentration', which measures the concentration of one situation compared to another. For example, the concentration of journal articles in a specific journal (on different topics) with respect to the articles in the whole area. Egghe (1988e) shows that this measure does not satisfy all 'axioms' that such a relative concentration measure should have and proposes another measure. Egghe (1988a) studies the time evolutions of concentrations.

4. Equation [IV.7.2] is not valid if $\alpha \neq 2$. However, in Egghe (1986b) it is shown that the following effect still holds : the higher μ is, the lower the fraction x of the sources is needed in order to have a fixed fraction θ of

the items.

5. As noted earlier, Price's law is generally not valid. Glänzel and Schubert (1985) constructed an informetric law such that Price's square root law (the most simple one) is valid. This distribution is termed the 'Price distribution', but does not seem to have many applications.

6. Marshall and Olkin (1979) introduce the so-called 'Lorenz-order' (see also Hardy, Littlewood and Pólya (1988)) on sequences $(x_1,x_2,\ldots,x_N)$. Functions that are increasing with respect to this order are shown to satisfy all the concentration principles (C1) - (C6). However, the generalised transfer property (Egghe and Rousseau (1989), Egghe and Rousseau (1990)) is an extension of this unifying model.

7. Egghe (1987a,b) and Egghe and Rousseau (1986), study concentration aspects of the laws of Lotka, Mandelbrot and Zipf as well as of the geometric distribution.

8. Hustopecký and Vlachý (1978) graphed six concentration measures, including G and CON, as a function of the mean for cases of four probability distributions.

9. Concentration theory is rooted in econometrics, as is clear from the use of the Gini index (Gini (1909)). Researchers in informetrics should generally be aware of the fact that their problems are not always original and open to solution. Despite a small change in form or terminology, a problem is often analogous to an existing problem (or even solution) in another discipline. A typical example was the introduction (in informetrics) of the Pratt measure (Pratt (1977)). Carpenter (1979) proved that Pratt's measure was basically nothing more than Gini's index (cf. equation IV.7.22). See, for example, Lambert (1985), Pfähler (1985) and Berrebi and Silber (1985)) for a few recent articles on concentration in econometrics. Still, a lot more articles have been published on this topic. For a variety of references, consult the econometric journals.

Brookes (1977a) discusses dispersion versus concentration measures and hints how to apply such measures.

10. Price's square root law (see Section IV.7.1.2) is sometimes called 'Rousseau's law' (J.J. not R.!) as J.J. Rousseau mentions this explicitly in his 'Contrat Social' concerning the size of the 'elite' ('elite' refers here to participating in the government). More information about this early sociometric statement can be found in Zipf (1949) and Rescher (1978).

11. Somewhat related (but different from) concentration measures are the so-called collaboration measures. These measures calculate the 'degree' of collaboration in research groups (as e.g. reflected in the co-authored

papers). We refer to Ajiferuke, Burrell and Tague (1988) and Englisch (1990) for some very valuable attempts to define 'good' collaboration measures. Based on these ideas Egghe (1990b) studies the necessary parameters and formulates some 'collaboration principles'. Furthermore, new measures are proposed that satisfy these principles.

IV.8.5. Non-Gaussian

The non-Gaussian nature (i.e. statistics for which the central limit theorem (see Section I.3.3) is not valid, due to the infinity of certain moments) of some informetric laws is discussed in Haitun (1982a,b,c and 1983), Yablonsky (1985) and Brookes (1984b,c), though Burrell (1988b) and Sichel (1986) apply Gaussian techniques to Zipfian distributions.

IV.8.6. n-dimensional informetrics

Informetrics can be studied in many different ways. One division of informetrics might be as follows (cf. Egghe (1989a)) :

Any study dealing with either the sources or the items separately (i.e. not linked to each other), can be said to be a 1-dimensional study.

Examples :

1. The numbers of books in a library.
2. The numbers of circulations in a library.
3. The total number of publications in the field of Geography (say in a year).
4. The total number of researchers in mathematics in Belgium.

Such data can be very interesting, especially in the connection with evolution in time. Many publications have resulted from such studies.

A 2-dimensional study, linking sources and items, was the main object of Chapter IV.3 and subsequent chapters.

A 3-dimensional approach will no doubt be devised in the future.

Examples :

1. Journals have papers and these papers are written by authors.
2. Journals have papers and papers receive (or have) citations.
3. Papers have references but do also receive citations.

These examples conform schematically with the following diagrams and graphs :

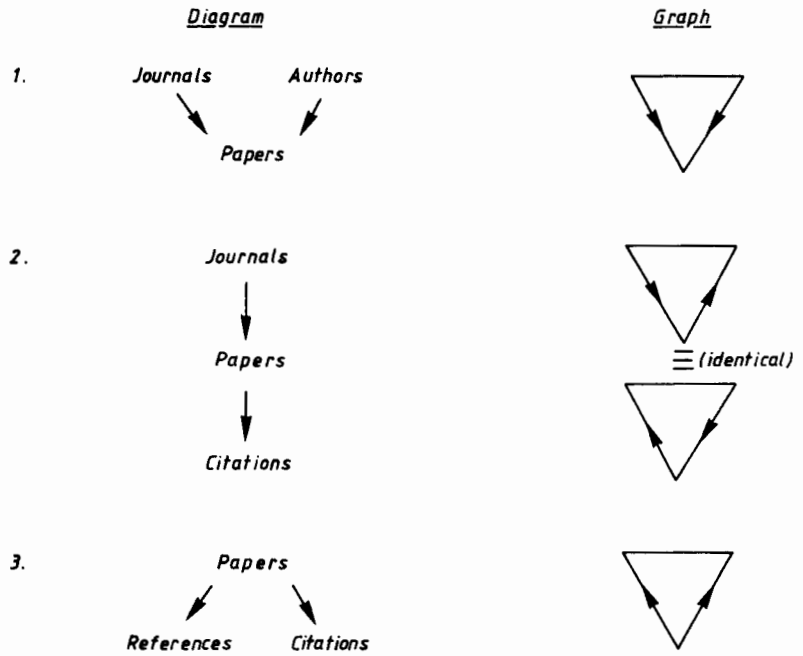

Fig.IV.8.1 Schematic representations of three-dimensional informetrics

One can even conceive of 4-dimensional (or even higher dimensional) informetrics. This is not an easy problem and may be broken down into several research projects.

Although it is not clear how to deal with the above 'triangles', one remark can be made. All the above differently oriented triangles can be unified into one model (see Fig.IV.8.2).

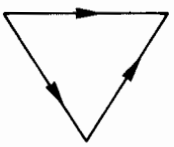

Fig.IV.8.2 Unification of three-dimensional problems

Indeed, depending on how one looks at the problem (i.e. depending on which side one concentrates on), any of the above situations will occur. But,

contrary to what Fig.IV.8.2 might suggest, the solution to the three-dimensional problem cannot be obtained by putting together two two-dimensional functions. The main reason for this is the fact that, considering real life situations, orderings in three-dimensional IPP's are different, depending on what sources are considered. Instead, a kind of 'triality' as found in geometry, might be needed here : points, lines, planes (in three-dimensional space).

IV.8.7. Many-to-many relations

Another fundamental viewpoint in informetrics different than that given above is the 'many-to-many' relationship, as was pointed out by S. Robertson (oral communication). It contrasts with the above techniques in that it studies (e.g. in the 2-dimensional version) many sources versus many items. We have here the study of the relationship between a set S_1 and a set I_1 such that S_1 is a subset of the source set S and I_1 is a subset of the item set I, where the 'device' function is now

$$f : S_1 \rightarrow I_1 \qquad (\text{or } \leftarrow) \ .$$

An example is offered by :

I_1 = a set of index words ,
S_1 = the set of papers ,
f = the function 'stating' that the sources in S_1 have the index words in I_1.

IV.8.8. Time-dependent studies and problems

1. Even in two-dimensional informetrics, one can require a time-dependent theory. This problem has been raised by B.C. Brookes (oral communication). So far, only one-dimensional time-dependent studies have been carried out. The most important topics in these studies are growth and obsolescence. The reader can note that the success-breeds-success arguments (see Simon (1955), Price (1976)) and other dynamic models (such as Schubert and Glänzel (1984b) and Rao (1980)) are time-dependent arguments on two-dimensional IPP's. These explanations are indeed 'dynamic' but their results (the distributions) are time-independent (for example, Price derives a time-independent cumulative advantage distribution, while Schubert and Glänzel derive the time-independent Waring distribution).

Thus, we indicate here the problem of the time evolution of the dual system

$$(S(t), I(t), V_t) ,$$

the time-dependent version of the IPP's defined in Chapter IV.3 and of the informetric laws that apply to them. More specifically, make a model of the growth of such an IPP. Preliminary calculations of Egghe have indicated that an exponential growth of the sources in conjunction with an exponential growth of the items is not likely to occur. The exponential function, however, seems to be the basic one used to describe evolution in time for both obsolescence (a^t, $0 < a < 1$, see III. 6.3) and growth (a^t, $a > 1$); see, for example, Price (1963). However, already in Price (1963), one can read the remark that in real life growth cannot be exponential forever. Consequently, a deflection point, changing the curve from convex to concave, must occur somewhere in the growth curve. Moreover, a horizontal asymptote is reached when t goes to ∞. This aspect is also called the 'principle of zero growth' (where t is high). See Fig.IV.8.3.

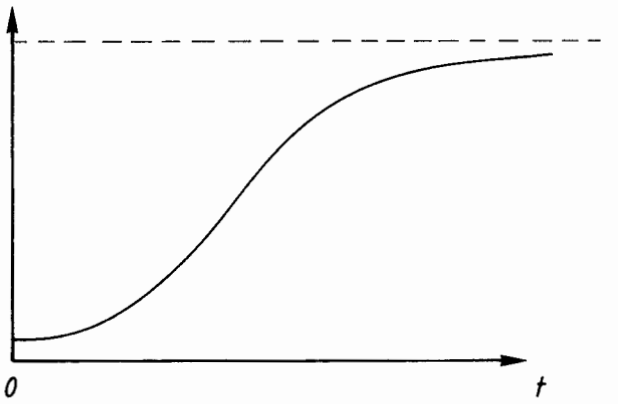

Fig.IV.8.3 Growth curve in 'real life' : the logistic curve

Such curves are called 'logistic curves' (see Price (1963) or Rescher (1978) for examples). Incidentally, Zipf (1949, e.g. p.529) also discusses the logistic growth of several social phenomena. The mathematical difference between exponential growth and logistic growth is given as follows :

Exponential growth :

The increase is proportional to population size P(t) :

$$\frac{dP(t)}{dt} = \lambda P(t) .$$

λ is called the 'Malthusian parameter' of the population. The solution is

$$P(t) = P(0)\, e^{\lambda t} , \qquad [IV.8.5]$$

which is indeed exponential. The law of exponential growth is also called 'Adam's law' (see Rescher (1978)).

Logistic growth :

A Malthusian population makes no allowance for the effects of crowding or the limitation of resources (cf. also the philosophical note on Zipf's law). So, instead of a constant Malthusian parameter, one can let it depend on the population itself. The following model can be traced all the way back to Verhulst (1845) :

$$\frac{dP(t)}{dt} = \lambda\left(1 - \frac{P(t)}{K}\right) P(t) .$$

This is the logistic equation which has a solution (method of separating the variables) :

$$P(t) = \frac{K}{1 + \left(\frac{K}{P(0)} - 1\right) e^{-\lambda t}} , \qquad [IV.8.6]$$

the curve of which is shown in Fig.IV.8.3.
For further information on population growth we refer the reader to Webb (1985) (mathematical work) or Rescher (1978) (philosophical work).

Another problem is the time evolution of $\rho(A)$ or, if you wish, of y_m (cf. [IV.5.1]), the production of the most productive source. It might also be interesting to conduct a study of the success-breeds-success principle (cf. Section IV.2.1) in connecting time dependent IPP's.

2. Once time dependent IPP's have been studied, one might look into the problem of applying 'stopping times' to the IPP's. Stopping times (see, for example, Egghe (1984)) are measurable functions, stating which objects in the population 'stop living' at which times.Such developments might be very important for all kinds of IPP's in practice, such as bibliographies over a long time period, authors papers bibliographies, patent bibliographies and, in general, all IPP's (since sources have finite lives).

3. The reader interested in time-dependent model studies (mainly one-dimensional) is referred to Part III (obsolescence), to Part II (library circulations) and to Kot (1987), Diamond Jr. (1987), Diamond Jr. (1984),

Kochen and Blaivas (1981), Brookes (1970), Rao (1980), Rouse (1979), Hubert (1976), Ware (1973), Kochen (1969) and Gama de Queiroz and Lancaster (1981), concerning growth problems.

IV.8.9. Bradford

1. In fitting Bradford's law (see Section IV.6.1), we chose the number p of Bradford groups more or less freely in $\mathbb{N}$. This idea is mainly based on the fact that if an IPP satisfies a certain informetric law (such as Leimkuhler's law), it satisfies Bradford's law (group-free as well as with p groups, for any $p \in \mathbb{N}$, naturally within the finite limits of the IPP, when we are dealing with a practical bibliography).

Of course, in practice perfect Bradfordian situations never arise. The question is then whether one choice of $p \in \mathbb{N}$ is better than another one. But what does a 'better Bradfordian division' really mean? Brooks (1989a) presented an interesting method for making this distinction. Since the number of items in every Bradford group must be equal (in the case of perfect Bradfordian IPP's), we can define a first Bradfordian division better than another one if Pratt's measure (cf. Subsection IV.7.1.3.2), calculated over the groups of items of the first one, is smaller than the one calculated over the groups of items of the second one. In this connection Brooks defines the term 'perfect Bradford multiplier' if this multiplier is linked to a Bradford division for which C, calculated over the groups of items, is zero.

Brooks then checks some Bradford fitting methods, namely the one described above in Section IV.6.1 and the one of Goffmann and Warren (1969). They determine the maximum possible number of Bradford groups, such that each group has the minimum possible number of sources. This is determined by requiring that in the second to last Bradford group, there must be at least one source with two items. If not then the second to last group will then contain only sources with one item. Since this is then obviously also the case with the last group, the fact that all groups have the same number of items would imply here that the last two Bradford groups would contain the same number of sources. Consequently, this would result in a Bradford factor of $k = 1$, contradicting the requirement that k must be > 1. In such a situation in which there is a 'minimum number of sources per group', the sources in the first Bradford group are called the 'minimal core (or nucleus) sources'. In Egghe (1989d), based on results in Egghe (1986a), it is shown that this limiting situation occurs precisely when $\log k = 0.5$, i.e. when $k = 1.6487$.

Brooks (1989b) makes use of this minimal core Bradford division to determine a 'clustering index' in the IPP, used to measure the degree of

clustering in IPP's, i.e. a comparison between the number of sources with more than one item and the number of sources with one item. The future will show the exact place of this promising clustering index in informetrics.

2. In some publications (see, for example, Yablonsky (1980), Goffmann and Warren (1969)), the Bradford factor k is interpreted as an average μ. This is not true for at least one reason : k is p-dependent. In Egghe (1989d) the exact function $\frac{k}{\mu}$ has been studied and it has been shown that neither k nor K, the group-free Bradford factor, can be interpreted as averages.