Appendix 1: Tables and Figures

A dot over a digit in the tables signifies recurring.

Table 0

no. requests	sig. %	S>	F_A>	for 95% power P>	$p_A - p_B$ =5% N>	% of pool required as assessment sample: RECALL rel = 5 % pool	rel =10 % pool	rel =25 % pool	rel =50 % pool	rel =75 % pool	PRECISION retr = 25 % pool	retr = 50 % pool	retr =100 % pool	retr =250 % pool	retr =1000 % pool
300	5	2.0	167	.605	15	*	*	60	30	20	60	30	15	6.0	1.5
500	5	2.0	272	.581	9	*	90	36	18	12	36	18	9	3.6	0.9
700	5	2.0	376	.569	7	*	70	28	14	9.3	28	14	7	2.8	0.7
900	5	2.0	480	.561	6	*	60	24	12	8	24	12	6	2.4	0.6

Figures all approximate

sig % = significance level of test
S = critical region of test
F_A = number of measurements favouring strategy A
P = probability that measurement on strategy A will exceed that on B over the request set
N = evaluation sample size for hypothesised difference between strategies A and B, i.e. the number of documents of known relevance status required in the output of A, and of B

<u>Table 0</u> : The Pool method: precentage of pool requiring assessment for specified evaluation samples; approximate figures taken from the design study report (Sparck Jones and Bates (1977))

no. requests	sig, %	S>	F_A>	for 95% power P>	$p_A - p_B$ =5% N>	RECALL rel = 5 % pool	rel =10 % pool	rel =25 % pool	rel =50 % pool	rel =75 % pool	PRECISION retr = 25 % pool	retr = 50 % pool	retr =100 % pool	retr =250 % pool	retr =1000 % pool
300	5	1.96	167	0.604	15	*	*	60	30	20	60	30	15	6	1.50
	1	2.58	172	0.620	19	*	*	76	38	$25.\dot{3}\dot{3}$	76	38	19	7.60	1.90
400	5	1.96	220	0.591	11	*	*	44	22	$14.\dot{6}\dot{6}$	44	22	11	4.40	1.10
	1	2.58	226	0.606	15	*	*	60	30	20	60	30	15	6	1.50
500	5	1.96	272	0.581	9	*	90	36	18	12	36	18	9	3.60	0.90
	1	2.58	279	0.595	12	*	*	48	24	16	48	24	12	4.80	1.20
600	5	1.96	324	0.574	8	*	80	32	16	$10.\dot{6}\dot{6}$	32	16	8	3.20	0.80
	1	2.58	332	0.587	10	*	*	40	20	$13.\dot{3}\dot{3}$	40	20	10	4	1
700	5	1.96	376	0.568	7	*	70	28	14	$9.\dot{3}\dot{3}$	28	14	7	2.80	0.70
	1	2.58	384	0.580	9	*	90	36	18	12	36	18	9	3.60	0.90
800	5	1.96	428	0.564	6	*	60	24	12	8	24	12	6	2.40	0.60
	1	2.58	436	0.574	8	*	80	32	16	$10.\dot{6}\dot{6}$	32	16	8	3.20	0.80
900	5	1.96	479	0.560	5	100	50	20	10	$6.\dot{6}\dot{6}$	20	10	5	2	0.50
	1	2.58	489	0.571	7	*	70	28	14	$9.\dot{3}\dot{3}$	28	14	7	2.80	0.70
1000	5	1.96	531	0.557	5	100	50	20	10	$6.\dot{6}\dot{6}$	20	10	5	2	0.50
	1	2.58	541	0.567	6	*	60	24	12	8	24	12	6	2.40	0.60

Table 1 : The Pool method: percentage of pool requiring assessment for specified evaluation samples; accurate comprehensive figures

sig % = significance level of test
S = critical region of test
F_A = number of measurements favouring strategy A
P = probability that measurement on strategy A will exceed that on B over the request set
N = evaluation sample size for hypothesised difference between strategies A and B, i.e. the number of documents of known relevance status required in the output of A and of B

Table 2 (a)

Exhaustive judgement

RECALL

Number of relevant documents per request	Number of requests required for 5% significance	Number of requests required for 1% significance
5	830	1135
10	424	581
25	172	235
50	89	121
75	60	82
100	46	63

PRECISION

Number of retrieved documents per request	Number of requests required for 5% significance	Number of requests required for 1% significance
5	830	1135
10	424	581
25	172	235
50	89	121
75	60	82
100	46	63
250	19	26
500	12	17
1000	8	12

Table 2 : The Pool approach: ideal evaluation conditions
(a) number of requests required for specified numbers of documents of known relevance status, assuming exhaustive assessments

Table 2 (b)

<u>Exhaustive judgement</u>

Number of requests	Significance	RECALL Number of relevant documents	PRECISION Number of retrieved documents
10	5	445	445
	1	785	785
50	5	85	85
	1	117	117
100	5	40	40
	1	57	57
200	5	21	21
	1	28	28
300	5	15	15
	1	19	19
400	5	11	11
	1	15	15
500	5	9	9
	1	12	12
600	5	8	8
	1	10	10
700	5	7	7
	1	9	9
800	5	6	6
	1	8	8
900	5	5	5
	1	7	7
1000	5	5	5
	1	6	6

Table 2 : The Pool approach: ideal evaluation conditions
(b) number of documents required for specified numbers of requests

Table 3

Recall illustrated

100 documents in the pool

Number of relevant documents required to be assessed

Number of relevant documents in pool		5	6	7	8	9	10	11	12	13	14
	25	32	36	40	44	49	53	57	61	64	68
	50	15	18	20	22	24	27	29	31	33	35
	75	9	11	12	14	15	17	18	20	21	23

Number of relevant documents required to be assesed

Number of relevant documents in pool		15	16	17	18	19	20	21	22	23	24	25
	25	72	75	79	82	85	88	91	94	96	99	100
	50	38	40	42	44	46	48	-	-	-	-	58
	75	24	25	27	28	30	31	-	-	-	-	38

Number of relevant documents required to be assessed

Number of relevant documents in pool		30	35	40	45	50	55	60
	25	*	*	*	*	*	*	*
	50	67	76	85	94	100	*	*
	75	45	51	58	64	71	77	83

500 documents in the pool

Number of relevant documents required to be assessed

Number of relevant documents in pool		5	6	7	8	9	10	11	12	13	14
	25	163	186	208	230	250	270	290	309	328	346
	50	85	97	109	121	133	144	155	166	177	188
	75	57	65	74	82	90	97	105	113	120	128
	100	43	49	55	61	67	73	79	85	91	96
	250	16	18	21	23	25	28	30	33	35	37

Number of relevant documents required to be assessed

Number of relevant documents in pool		15	16	17	18	19	20
	25	364	381	398	414	429	444
	50	199	210	220	230	241	251
	75	135	143	150	157	164	171
	100	102	108	113	119	124	130
	250	39	42	44	46	48	51

Table 3 contd

(500 documents in the pool)

Number of relevant documents required to be assessed

Number of relevant documents in pool		25	30	35	40	45	50	55	60
	25	500	*	*	*	*	*	*	*
	50	300	347	392	434	472	500	*	*
	75	206	240	273	306	338	369	400	429
	100	156	183	209	234	259	284	308	332
	250	61	72	83	94	104	115	125	136

1000 documents in the pool

Number of relevant documents required to be assesed

Number of relevant documents in pool		5	6	7	8	9	10	11	12	13	14
	25	328	374	418	461	502	542	582	620	657	694
	50	172	197	221	245	268	290	313	335	357	379
	75	116	133	150	166	182	197	213	228	243	258
	100	87	100	113	125	137	149	161	173	184	195
	250	34	39	44	49	54	59	64	69	73	78
	500	16	18	21	23	26	28	-	-	-	-

Number of relevant documents required to be assessed

Number of relevant documents in pool		15	16	17	18	19	20	21	22	23	24	25
	25	729	763	797	829	860	889	917	943	966	986	998
	50	400	422	443	464	484	504	-	-	-	-	602
	75	273	288	303	317	331	346	-	-	-	-	415
	100	207	218	230	241	252	263	-	-	-	-	316
	250	83	87	92	96	101	105	-	-	-	-	128
	500	-	-	-	-	-	-	-	-	-	-	-

Number of relevant documents required to be assessed

Number of relevant documents in pool		30	35	40	45	50	55	60
	25	*	*	*	*	*	*	*
	50	696	786	870	946	1000	*	*
	75	483	550	616	680	742	802	859
	100	369	421	472	522	572	621	669
	250	149	172	193	215	236	257	278

- = not computed, * = not necessary

Table 3 : The Pool method, modified: numbers of documents to be assessed for specified pool sizes, illustrated for recall

Table 4

(a) Assuming $1 - \frac{n_A}{N} \sim 1$.

α = upper bound to probability × 100	d = lower bound to divergence	$n_A \geq$ number of relevant or retrieved in pool
10%	0.05	272
	0.04	425
	0.03	756
	0.02	1701
	0.01	6806
5%	0.05	384
	0.04	600
	0.03	1067
	0.02	2401
	0.01	9604
1%	0.05	665
	0.04	1040
	0.03	1849
	0.02	4160
	0.01	16641

(b) 'Exact' form ($\frac{n_A}{N} > \frac{1}{10}$, say)

α = 10% d = 0.05

N	n≥
500	176
1000	214
1500	230
2000	239
2500	245

N = number of documents in the pool

n≥ = lower bound for n

α = 10% d = 0.04

N	n≥
500	229
1000	298
1500	331
2000	350
2500	363
3000	372
3500	379
4000	384

Table 4 contd

α = 10%   d = 0.03

N	n≥
1000	430
2000	548
3000	603
4000	635
5000	656
6000	671
7000	682

α = 10%   d = 0.02

N	n≥
2000	919
5000	1269
10000	1453
15000	1527
20000	1567

α = 10%   d = 0.01

N	n≥
10000	4049
20000	5078
30000	5547
40000	5816
50000	5990
60000	6112
70000	6202

α = 5%   d = 0.05

N	n≥
500	217
1000	277
1500	305
2000	322
2500	332
3000	340
3500	346

α = 5%   d = 0.04

N	n≥
1000	375
1500	428
2000	461
2500	484
3000	500
3500	512
4000	521
4500	529
5000	535
5500	541
6000	545

Table 4 contd

α = 5% d = 0.03

N	n≥
2000	696
3000	787
4000	842
5000	879
6000	906
7000	925
8000	941
9000	954
10000	964

α = 5% d = 0.02

N	n≥
3000	1333
5000	1622
10000	1936
15000	2069
20000	2143
25000	2190

α = 5% d = 0.01

N	n≥
10000	4899
20000	6488
30000	7275
40000	7744
50000	8056
60000	8278
70000	8445
80000	8574
90000	8678

α = 1% d = 0.05

N	n≥
1000	399
2000	499
3000	544
4000	570
5000	587
6000	598
7000	607

Table 4 contd

α = 1% d = 0.04

N	n≥
2000	684
4000	825
6000	886
8000	920
10000	942
12000	957

α = 1% d = 0.03

N	n≥
2000	961
5000	1350
10000	1560
15000	1646
20000	1692

α = 1% d = 0.02

N	n≥
5000	2270
10000	2938
20000	3443
30000	3653
40000	3768

α = 1% d = 0.01

N	n≥
25000	9990
50000	12485
75000	13619
100000	14266
125000	14685
150000	14979
175000	15196

Table 4 : The Pool method: requirements for accuracy of estimators

Table 5

Suppose that only 90% of the relevant documents are contained in the pool.

RECALL

no. requests	sig. %	N>	% of pool required as assessment sample rel = 5 %	rel =10 %	rel =25 %	rel =50 %	rel =75 %
300	5	15	*	*	$66.6\dot{6}$	$33.3\dot{3}$	$22.2\dot{2}$
	1	19	*	*	$84.4\dot{4}$	$42.2\dot{2}$	$28.1\dot{5}$
400	5	11	*	*	$48.8\dot{8}$	$24.4\dot{4}$	16.30
	1	15	*	*	$66.6\dot{6}$	$33.3\dot{3}$	$22.2\dot{2}$
500	5	9	*	100	40.00	20.00	$13.3\dot{3}$
	1	12	*	*	$53.3\dot{3}$	$26.6\dot{6}$	$17.7\dot{7}$
600	5	8	*	$88.8\dot{8}$	$35.5\dot{5}$	$17.7\dot{7}$	11.85
	1	10	*	*	$44.4\dot{4}$	$22.2\dot{2}$	14.81
700	5	7	*	$77.7\dot{7}$	$31.1\dot{1}$	$15.5\dot{5}$	10.37
	1	9	*	100	40.00	20.00	$13.3\dot{3}$
800	5	6	*	$66.6\dot{6}$	$26.6\dot{6}$	$13.3\dot{3}$	$8.8\dot{8}$
	1	8	*	$88.8\dot{8}$	$35.5\dot{5}$	$17.7\dot{7}$	11.85
900	5	5	*	$55.5\dot{5}$	$22.2\dot{2}$	$11.1\dot{1}$	7.41
	1	7	*	$77.7\dot{7}$	$31.1\dot{1}$	$15.5\dot{5}$	10.37
1000	5	5	*	$55.5\dot{5}$	$22.2\dot{2}$	$11.1\dot{1}$	7.41
	1	6	*	$66.6\dot{6}$	$26.6\dot{6}$	$13.3\dot{3}$	$8.8\dot{8}$

Suppose that only 90% of the output of future strategy searches is in the pool.

PRECISION

			retr = 25	retr = 50	retr =100	retr =250	retr =1000
300	5	15	$66.6\dot{6}$	$33.3\dot{3}$	$16.6\dot{6}$	$6.6\dot{6}$	$1.6\dot{6}$
	1	19	$84.4\dot{4}$	$42.2\dot{2}$	$21.1\dot{1}$	$8.4\dot{4}$	$2.1\dot{1}$
400	5	11	$48.8\dot{8}$	$24.4\dot{4}$	$22.2\dot{2}$	$4.8\dot{8}$	$1.2\dot{2}$
	1	15	$66.6\dot{6}$	$33.3\dot{3}$	$16.6\dot{6}$	$6.6\dot{6}$	$1.6\dot{6}$
500	5	9	40.00	20.00	10.00	4.00	1.00
	1	12	$53.3\dot{3}$	$26.6\dot{6}$	$13.3\dot{3}$	$5.3\dot{3}$	$1.3\dot{3}$
600	5	8	$35.5\dot{5}$	$17.7\dot{7}$	$8.8\dot{8}$	$3.5\dot{5}$	$0.8\dot{8}$
	1	10	$44.4\dot{4}$	$22.2\dot{2}$	$11.1\dot{1}$	$4.4\dot{4}$	$1.1\dot{1}$
700	5	7	$31.1\dot{1}$	$15.5\dot{5}$	$7.7\dot{7}$	$3.1\dot{1}$	$0.7\dot{7}$
	1	9	40.00	20.00	10.00	4.00	1.00
800	5	6	$26.6\dot{6}$	$13.3\dot{3}$	$6.6\dot{6}$	$2.6\dot{6}$	$0.6\dot{6}$
	1	8	$35.5\dot{5}$	$17.7\dot{7}$	$8.8\dot{8}$	$3.5\dot{5}$	$0.8\dot{8}$
900	5	5	$22.2\dot{2}$	$11.1\dot{1}$	$5.5\dot{5}$	$2.2\dot{2}$	$0.5\dot{5}$
	1	7	$31.1\dot{1}$	$15.5\dot{5}$	$7.7\dot{7}$	$3.1\dot{1}$	$0.7\dot{7}$
1000	5	5	$22.2\dot{2}$	$11.1\dot{1}$	$5.5\dot{5}$	$2.2\dot{2}$	$0.5\dot{5}$
	1	6	$31.1\dot{1}$	$15.5\dot{5}$	$6.6\dot{6}$	$2.6\dot{6}$	$0.6\dot{6}$

Table 5 : The Pool method, weakened assumptions: percentage of pool requiring assessment for specified evaluation samples, pool not exhaustive

Table 6

no. requests	sig.	S>	T_w>	For 95% power P>	N
300	5	1.96	5894.698	0.619	19
	1	2.58	7759.347	0.639	25
400	5	1.96	9069.825	0.604	14
	1	2.58	11938.850	0.621	19
500	5	1.96	12670.721	0.593	11
	1	2.58	16678.807	0.609	15
600	5	1.96	16651.938	0.585	10
	1	2.58	21919.389	0.599	12
700	5	1.96	20980.099	0.579	8
	1	2.58	27616.661	0.592	11
800	5	1.96	25629.336	0.574	7
	1	2.58	33736.575	0.586	9
900	5	1.96	30578.836	0.570	6
	1	2.58	40251.733	0.581	8
1000	5	1.96	35811.377	0.566	5
	1	2.58	47139.466	0.577	7

T_w = measure of difference between strategies A and B; other columns as Table 1.

Table 6 : Wilcoxon test

Table 7

95% power assumed

sig.	n	p_b+p_c i.e. Π_R	b+c *	b>	$\frac{p_b}{p_b+p_c}$ i.e. $\Delta \geq$
5	50	0.25	6	5	0.935
1				-	-
5			12.5	10	0.923
1				11	0.966
5			19	14	0.866
1				15	0.902
5	50	0.50	18	13	0.858
1				14	0.897
5			25	17	0.810
1				19	0.871
5			32	22	0.804
1				23	0.853
5	50	0.75	31	21	0.797
1				23	0.848
5			37,5	25	0.778
1				27	0.823
5			45	29	0.751
1				31	0.789
5	100	0.25	17	13	0.890
1				14	0.928
5			25	17	0.810
1				19	0.871
5			34	23	0.792
1				25	0.839
5	100	0.50	40	26	0.761
1				28	0.804
5			50	32	0.743
1				34	0.778
5			61	38	0.718
1				41	0.762
5	100	0.75	67	42	0.718
1				44	0.745
5			75	46	0.701
1				49	0.738
5			84	51	0.691
1				54	0.724

Table 7 contd

sig.	n	p_b+p_c	b+c	b>	$\frac{p_b}{p_b+p_c} \geq$
5	500	0.25	106	63	0.670
1				66	0.697
5			125	73	0.655
1				77	0.685
5			144	84	0.649
1				87	0.669
5	500	0.50	228	129	0.619
1				133	0.636
5			250	140	0.611
1				145	0.631
5			272	152	0.608
1				157	0.626
5	500	0.75	356	196	0.594
1				202	0.610
5			375	206	0.592
1				212	0.607
5			394	216	0.589
1				223	0.607
5	1000	0.25	223	126	0.619
1				131	0.641
5			250	140	0.611
1				145	0.631
5			277	155	0.608
1				160	0.626
5	1000	0.50	469	256	0.584
1				262	0.597
5			500	272	0.581
1				279	0.594
5			531	288	0.578
1				295	0.591
5	1000	0.75	724	388	0.567
1				397	0.579
5			750	402	0.566
1				410	0.577
5			776	415	0.565
1				424	0.576

Table 7 contd

sig.	n	p_b+p_c	b+c	b>	$\frac{p_b}{p_b+p_c} \geq$
5	2000	0.25	462	252	0.584
1				259	0.599
5			500	272	0.581
1				279	0.594
5			538	292	0.578
1				299	0.591
5	2000	0.50	938	499	0.559
1				509	0.570
5			1000	531	0.557
1				541	0.567
5			1062	563	0.556
1				573	0.565
5	2000	0.75	1462	768	0.547
1				780	0.555
5			1500	788	0.547
1				800	0.555
5			1538	807	0.546
1				820	0.554
5	3000	0.25	704	378	0.568
1				386	0.579
5			750	402	0.566
1				410	0.577
5			796	426	0.565
1				434	0.575
5	3000	0.50	1446	760	0.548
1				772	0.556
5			1500	788	0.547
1				800	0.555
5			1554	816	0.546
1				828	0.554
5	3000	0.75	2205	1149	0.539
1				1163	0.545
5			2250	1171	0.538
1				1186	0.545
5			2295	1194	0.538
1				1209	0.544

Table 7 contd

sig.	n	p_b+p_c	b+c	b>	$\frac{p_b}{p_b+p_c} \geq$
5	4000	0.25	946	503	0.559
1				513	0.569
5			1000	531	0.557
1				541	0.567
5			1054	559	0.556
1				569	0.565
5	4000	0.50	1956	1021	0.541
1				1035	0.548
5			2000	1044	0.541
1				1058	0.548
5			2044	1066	0.540
1				1080	0.547
5	4000	0.75	2946	1526	0.534
1				1543	0.539
5			3000	1554	0.533
1				1571	0.539
5			3054	1581	0.533
1				1598	0.539
5	5000	0.25	1190	629	0.553
1				640	0.562
5			1250	660	0.552
1				671	0.560
5			1310	690	0.550
1				701	0.558
5	5000	0.50	2431	1264	0.537
1				1279	0.543
5			2500	1299	0.536
1				1315	0.543
5			2569	1334	0.536
1				1350	0.542
5	5000	0.75	3690	1905	0.530
1				1923	0.535
5			3750	1935	0.530
1				1954	0.535
5			3810	1965	0.530
1				1985	0.535

Table 7 contd

sig.	n	p_b+p_c	b+c	b>	$\frac{p_b}{p_b+p_c} \geq$
5	10000	0.25	2415	1256	0.537
1				1271	0.543
5			2500	1299	0.536
1				1315	0.543
5			2585	1342	0.536
1				1358	0.542
5	10000	0.50	4902	2520	0.526
1				2541	0.531
5			5000	2569	0.526
1				2591	0.530
5			5098	2619	0.526
1				2641	0.530
5	10000	0.75	7415	3792	0.521
1				3819	0.525
5			7500	3835	0.521
1				3862	0.525
5			7585	3878	0.521
1				3905	0.525

* The three values given for b+c for each value of n represent the lower bound, expected value, and upper bound, for 95% confidence.

n = contingency table total
Π_R = probability that a document is retrieved by one strategy only
b+c = number of documents retrieved by one strategy only
b = number of documents retrieved by strategy A alone
$\frac{p_b}{\Pi_R}$ = probability that a document is retrieved by strategy A given that it is retrieved by one strategy

Table 7 : The Squares method: number of documents to be retrieved by strategy A alone to support statement that A is better than B

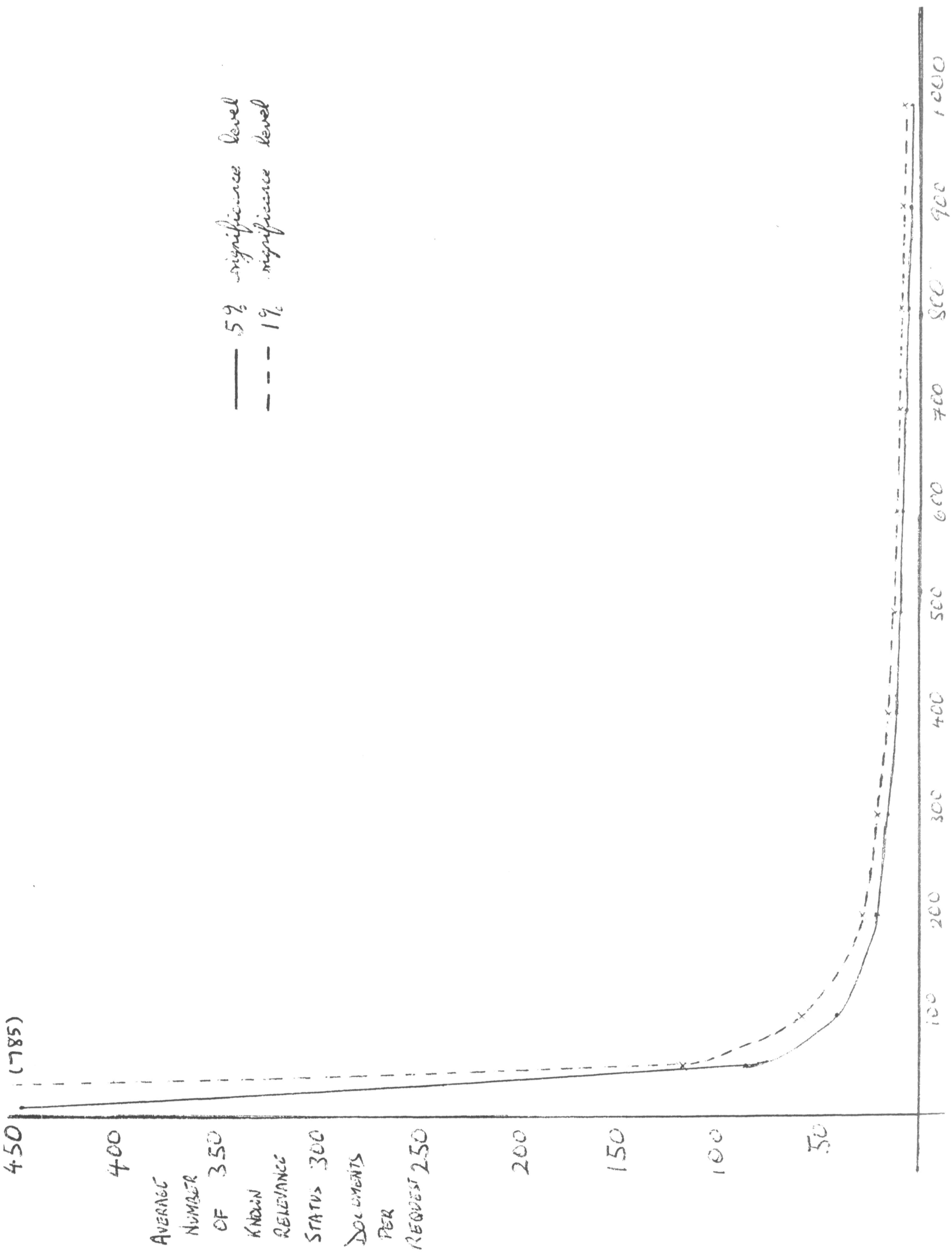

Figure 1: The Pool method: illustration of Table 2

Figure 2

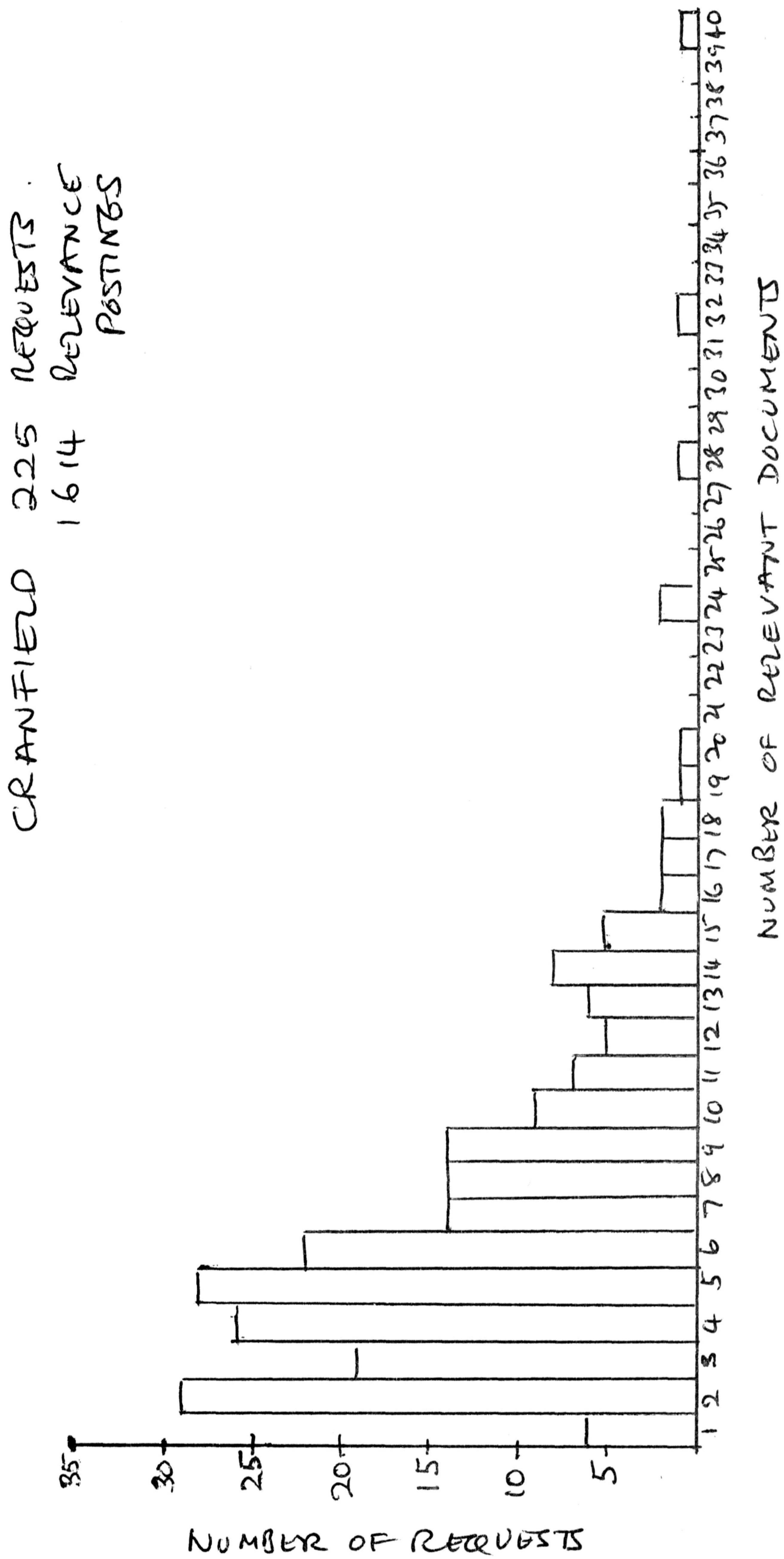
CRANFIELD 225 REQUESTS
1614 RELEVANCE POSTINGS
NUMBER OF RELEVANT DOCUMENTS
NUMBER OF REQUESTS
5 10 15 20 25 30 35
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40

Figure 2 contd

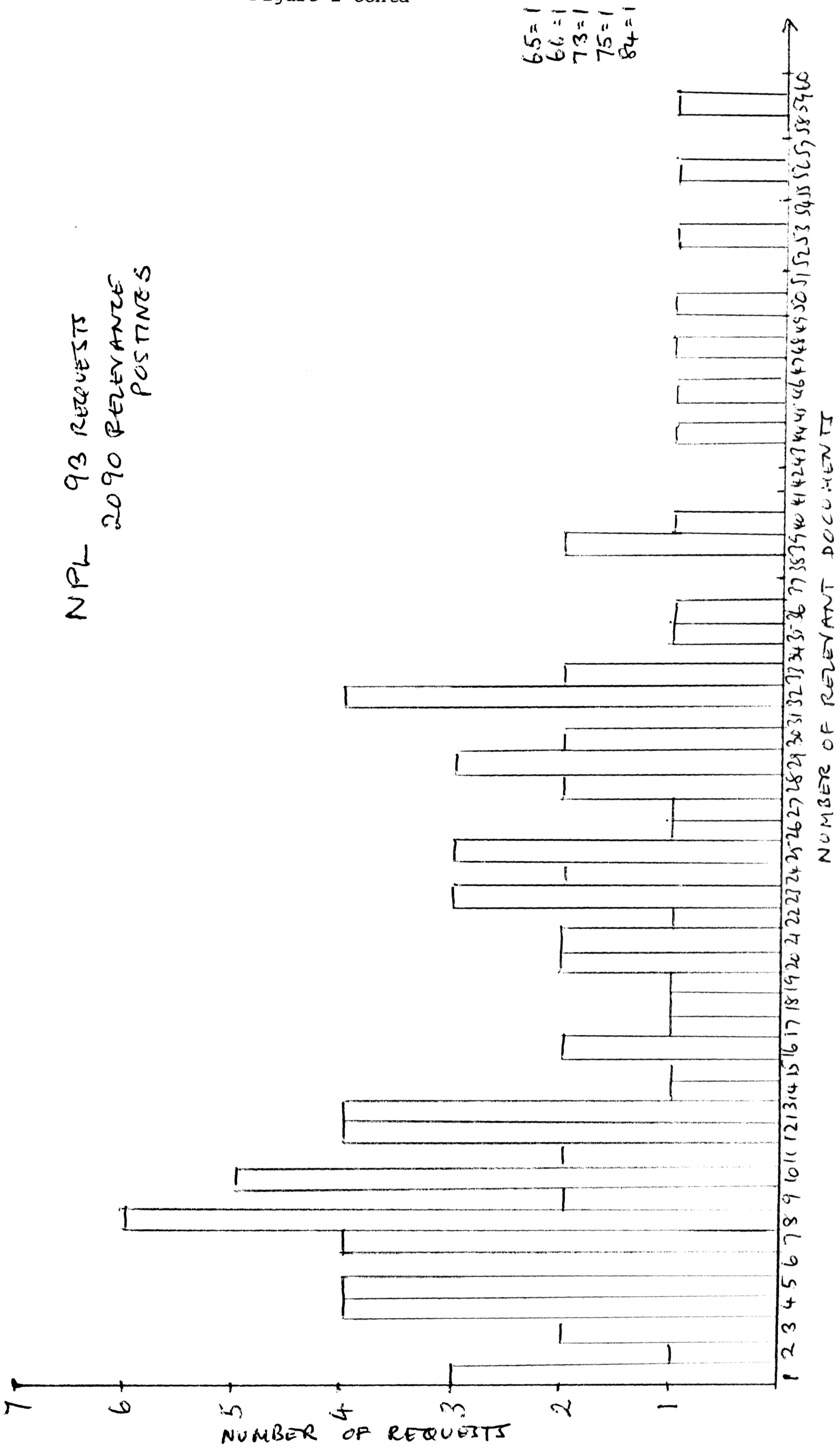

Figure 2 contd

UKCIS 182 REQUESTS
10715 RELEVANCE POSTINGS

no. rel.	no. req.
1	6
2	6
3	7
4	7
5	3
6	2
7	4
8	7
9	3
10	3
11	3
12	2
13	3
14	6
15	4
16	3
18	2
19	1
20	2
21	2
22	4
23	2
24	3
25	4
26	3
27	2
28	2
29	1
30	3
31	2
34	1
35	3
36	2
37	1
40	2
42	2
43	2
44	1
45	1
46	1
47	2
48	1
49	1
52	3
53	2
54	1
56	2
57	2
58	1

no. rel.	no. req.
60	3
61	1
64	1
67	1
74	1
75	1
76	1
82	1
85	1
89	1
93	2
94	1
98	1
99	1
100	1
103	1
105	1
108	1
109	1
112	1
113	1
123	1
125	1
127	1
130	1
136	1
140	1
149	2
150	1
151	1
154	1
159	1
177	1
180	1
183	1
198	1
231	1
233	1
266	1
319	1
408	1
413	1
510	1
511	1
554	1

Figure 2: Illustrations of test collection relevant document distributions

Appendix 2: Computing

For convenience we repeat the statement of the problem which the program is trying to solve.

Suppose that (considering recall) one has K relevant documents altogether in the pool and that at least n of these are required to be assessed. Suppose also that there are N documents in total in the pool. Then if a simple random sample of size S is taken for assessment, the probability that n of the relevant documents have been assessed is

$$\frac{\sum_{y=n}^{\min(S,K)} \binom{K}{n}\binom{N-K}{S-n}}{\binom{N}{n}} \qquad \text{(hypergeometric distribution).}$$

Unfortunately the hypergeometric distribution is quite messy to calculate, so quite often it is approximated to by the binomial distribution, this approximation holding whenever $\frac{SK}{N} > 25$.

Similarly it can be approximated to by the normal distribution, whenever the normal distribution is a suitable approximation to the binomial (that is, $\frac{S}{N} < 0.1$).

Unfortunately, neither condition is expected to hold in general, and so the approximations were not considered to be feasible.

Now the magnitude of the numbers with which we are dealing is very large (N takes values from 100 - 10,000) and the IBM computer (a 370/165) cannot cope with N!: so we have to approximate N!. To do this Stirling's theorem is used, which says that

$$N! \sim \sqrt{(2\pi N)}\left(\frac{N}{e}\right)^N \quad .$$

Taking logs one finds that

$$\log N! \sim 0.39909 + (n + \tfrac{1}{2})\log_{10}N - 0.4342945N \quad . \qquad (1)$$

This expression is only accurate to $O(\frac{1}{N})$ and so was only used for $N > 30$. Once this has been done the antilog is taken and the probability of obtaining at least n relevant documents is calculated. The value of S is then varied until the largest value of S is obtained which has probability less than 0.95, and this value (incremented by 1) is the value tabulated in Table 3.

The fact that (1) is only accurate to $O(\frac{1}{N})$ is the reason for the error mentioned in chapter B2; but as was said then, the probability of achieving the required number of relevant documents should still be greater than 0.93.

Note that the order the data was read into the computer was such that N and K would remain unchanged while X ran through all its possibilities; then K would change , and eventually N would change. (In the program n was denoted by X.)

If N and K are the same and X has increased by 5, say, then one would only expect S to increase by a small amount.

The 'IF' statements at the beginning of the program (concerning A and B) are making use of this fact in order to speed up the search for the required sample size S. This improvement in technique is very important since to calculate the hypergeometric distribution is very costly in resources.

Another way of speeding up the program would be to make use of the recursive property of the distribution. However the way the value of S jumps about makes this very awkward (although it could be used at the stage when S is always being decreased by 1).

A lot of time was spent during the project on debugging and improving the program. Since it was proving to be so expensive it was decided to try and fit the normal distribution anyway.

This was tried with data for which S had already been calculated by the hypergeometric distribution. Since there is some degree of inaccuracy already in the calculations (see earlier) it was decided that the normal approximation would be acceptable if the value of S it resulted in was within 5 of the previous value of S.

The following table shows the results obtained:

N	K	X	$\frac{S}{N}$	$\frac{SK}{N}$	S hyper	S normal
100	25	5	0.320	8.00	32	34
100	25	20	0.880	22.00	88	89
100	50	10	0.270	13.50	27	28
100	50	30	0.670	33.50	67	68
100	75	5	0.090	6.75	9	10
100	75	20	0.310	23.25	31	32
500	25	5	0.320	8.15	163	177
500	25	20	0.880	22.20	444	449
500	50	10	0.280	14.40	144	151
500	50	30	0.690	34.70	347	351
500	100	5	0.086	8.60	43	47
500	100	50	0.568	56.80	284	287
500	250	10	0.056	14.00	28	29
500	250	30	0.144	36.00	72	74
1000	25	5	0.328	8.20	328	357
1000	25	20	0.889	22.20	889	899
1000	50	10	0.290	14.50	290	305
1000	50	30	0.696	34.80	696	704
1000	100	5	0.870	8.70	87	96
1000	100	50	0.572	57.20	572	578
1000	250	10	0.590	14.75	59	62
1000	250	30	0.149	37.25	149	153

There are quite a few occasions when the difference is greater than 5 (indeed it can be as large as 28). Also there seems to be no pattern as to when the fit is good and when it is not. Many times when $\frac{S}{N} > 0.1$ and $\frac{SK}{N} < 25$ there is only a difference of 1 between the two estimates (see for example the first row). So no pattern could be found and the idea of the normal approximation was abandoned.

This program could be used for any particular values of N, K and X (X is the required number of assessed relevant documents) by filing them in I6 format and putting a delimiter of 0 at the end.

A listing of the program follows.

```
FORTXCLG (PROGRAM=%H? FT05F001=%H#) SECC=57
       IMPLICIT REAL*8 (A-H), REAL*8 (O-Z)
       INTEGER N,K,X,S,T,V,A,B,G
       A=0
       B=0 .
25     READ (5,8) N,K,X
       IF (N.EQ.0) GOTO 20
       IF (N.NE.A) GOTO 401
       IF (K.NE.B) GOTO 401
       S=G
402    S=S+5
       P=PROB(N,K,X,S)
       IF (P.GE.0.95D0) GOTO 6
       GOTO 402
401    V=N/2
       IF (V.GT.X) GOTO 550
       S=X
       GOTO 2
550    S=V
2      P=PROB(N,K,X,S)
       IF (DABS(P-0.95D0).LT.1.0E-6) GOTO 6
       IF (P.GE.0.95D0) GOTO 1
       T=S
       S=S+(N-S)/4
       IF (S.NE.T) GOTO 2
       S=S+1
       GOTO 2
1      S=S-1
       P=PROB(N,K,X,S)
       IF (DABS(P-0.95D0).LT.1.0E-6) GOTO 6
       IF (P.LE.0.95D0) GOTO 6
       GOTO 1
6      WRITE (6,9) N,K,X,S,P
       G=S
       A=N
       B=K
       GOTO 25
20     STOP
8      FORMAT (3I6)
9      FORMAT (19H # OF DOCS IN POOL=,I6,13H # RETRIEVED=,I5,
      $18H # TO BE ASSESSED=,I3,13H SAMPLE SIZE=,I6,4H P= ,F8.6)
       END
       DOUBLE PRECISION FUNCTION PROB(N,K,I,J)
       IMPLICIT REAL*8 (A-H), REAL*8 (O-Z)
       INTEGER F
       PROB=0.0D0
       W=DFLOAT(N)
       Y=DFLOAT(K)
       Z=DFLOAT(J)
       IF (J.GT.K) GOTO 250
       M=J
       GOTO 251
250    M=K
251    D=Z-W+Y
       F=DINT(D)
       IF (F.GT.I) GOTO 150
       F=I
```

```
150     DO 7 L=F,M
        C=DFLOAT(L)
        R=BINLOG(Y,C)+BINLOG(W-Y,Z-C)-BINLOG(W,Z)
        H=DABS(R)
        IF (H.GE.50.0D0) GOTO 7
        PROB=10.0D0**R+PROB
7       CONTINUE
300     RETURN
        END
        DOUBLE PRECISION FUNCTION FLOG(A)
        IMPLICIT REAL*8 (A-H), REAL*8 (O-Z)
        IF (A.LT.30.0D0) GOTO 75
        Q=0.4342945D0
        FLOG=0.39909D0+(A+0.5D0)*DLOG10(A)-Q*A
        GOTO 76
75      IF (A.GT.2.0D0) GOTO 77
        FLOG=DLOG10(A)
        GOTO 76
77      FLOG=FACTL(A)
76      RETURN
        END
        DOUBLE PRECISION FUNCTION BINLOG(A,B)
        IMPLICIT REAL*8 (A-H), REAL*8 (O-Z)
        IF (A.EQ.B) GOTO 31
        IF (B.NE.0.0D0) GOTO 33
31      BINLOG=0.0D0
        GOTO 32
33      BINLOG=FLOG(A)-FLOG(A-B)-FLOG(B)
32      RETURN
        END
        DOUBLE PRECISION FUNCTION FACTL(A)
        IMPLICIT REAL*8 (A-H)
        F=1.0D0
        L=DINT(A)
        DO 11 J=2,L
        F=F*DFLOAT(J)
11      CONTINUE
        FACTL=DLOG10(F)
        RETURN
        END
?
  1000    500     15
  1000    500     20
  1000    500     25
  1000    500     30
  1000    500     35
  1000    500     40
  1000    500     45
  1000    500     50
  1000    500     55
  1000    500     60
  2000     25     15
  2000     25     20
  2000     50      5
  2000     50     10
  2000     50     15
  2000     50     20
  2000     50     25
  2000     50     30
  2000     50     35
  2000     50     40
  2000     50     45
     0      0      0
#
```