



---

# Well-Behaved Variables

## The Unit of Measure

The Stem and Leaf procedure and its cousin the histogram are powerful devices for making sense out of data: You look at the picture, you construct mental hypotheses by examining first the shape and then the positions of the different units — soy bean cereal is high protein, rice cereal is low protein. Much of the numerical technology invented for the purpose of data analysis provides ways to summarize what you see in these pictures, ways to test the hypotheses you construct, and ways of adding precision to what you have inferred from these pictures. Yet the picture is primary and that is why such a technique as simple as the Stem and Leaf, with little technological challenge, is nevertheless an extremely powerful device in the hands of a data analyst.

However, even here there is a challenge. There are choices to be made. You, the analyst, can not just take the data as given, build a few pictures, report a few numbers, and be done with it — claiming you have analyzed the data. Would that it were that easy.

There are too many pictures that you *could* draw, too many numbers that you *could* compute. So the analyst has to acquire skills that go beyond the routine pictures and mechanical computation of numbers. Which pictures? Which numbers? Why these? And what are their implications. For example, suppose I had set up the breakfast cereal data in terms of grams of protein *per calorie* as the unit of measure instead of setting it up in terms of protein alone.. Using protein *per calorie*, the idea would be to look for high protein qualified by the number of calories that have to be consumed to obtain the measure of



---

protein. Using the ratio of protein to calories would affect the analysis by building-in some compensation for sugar-filled foods as well as air-filled foods that provide neither protein nor anything else in each air-filled serving.

If the idea is to protein to calories, the idea allows at least two implementations. There is a choice: I could compare protein to calories by computing grams of *protein per calorie*. Or, I could compare one to the other by computing *calories per gram of protein*.

On the face of it, it should make no difference: Whether it is protein per calorie or calories per protein it is the same information — or is it?



1	<u>Rice Krisp Kell; Rice Cerl Gerb; Rice puffed Quak; Rice Flak</u>
2	<u>KixGM; PostToast; BranRaisPost; BranRaisKel; KrumbKell; WheatShred; WheatiesGM; Wheat Chex; MuffetsQuak; BranFlk40%Post; BranFlk40%Kell</u>
3	<u>BranKel; CheerioGM; BarleyGerb; WheatFlkQuak; WheatPuffQuak; Mixed Gerb</u>
4	<u>Oatmeal</u>
5	<u>Special K Kell</u>
6	<u>Hi Pro GM</u>
7	
8	
9	
10	<u>HighProGerb</u>

Figure 1a  
 Grams of Protein per 100 Calories  
 Stems: 1 gram per 100 calories, 2 grams per  
 100 calories, ... 10 grams per 100 calories.

10	<u>HighProGerb; HiProGM; SpecialK Kell</u>
20	<u>Oatmeal; MixedGerb; WheatPuffQuak; WheatFlkQuak; BarleyGerb</u>
30	<u>Cheerio; BranKel; BranFlk40%Kell; BranFlk40%Post; MuffetsQuak; WheatChex; WheatiesGM; WheatShred; KrumbKell</u>
40	<u>BranRaisKel; BranRaisPost; PostToast; KixGM</u>
50	<u>RiceFlak</u>
60	<u>RicepuffedQuak; RiceCerlGerb; RiceKrispKell</u>

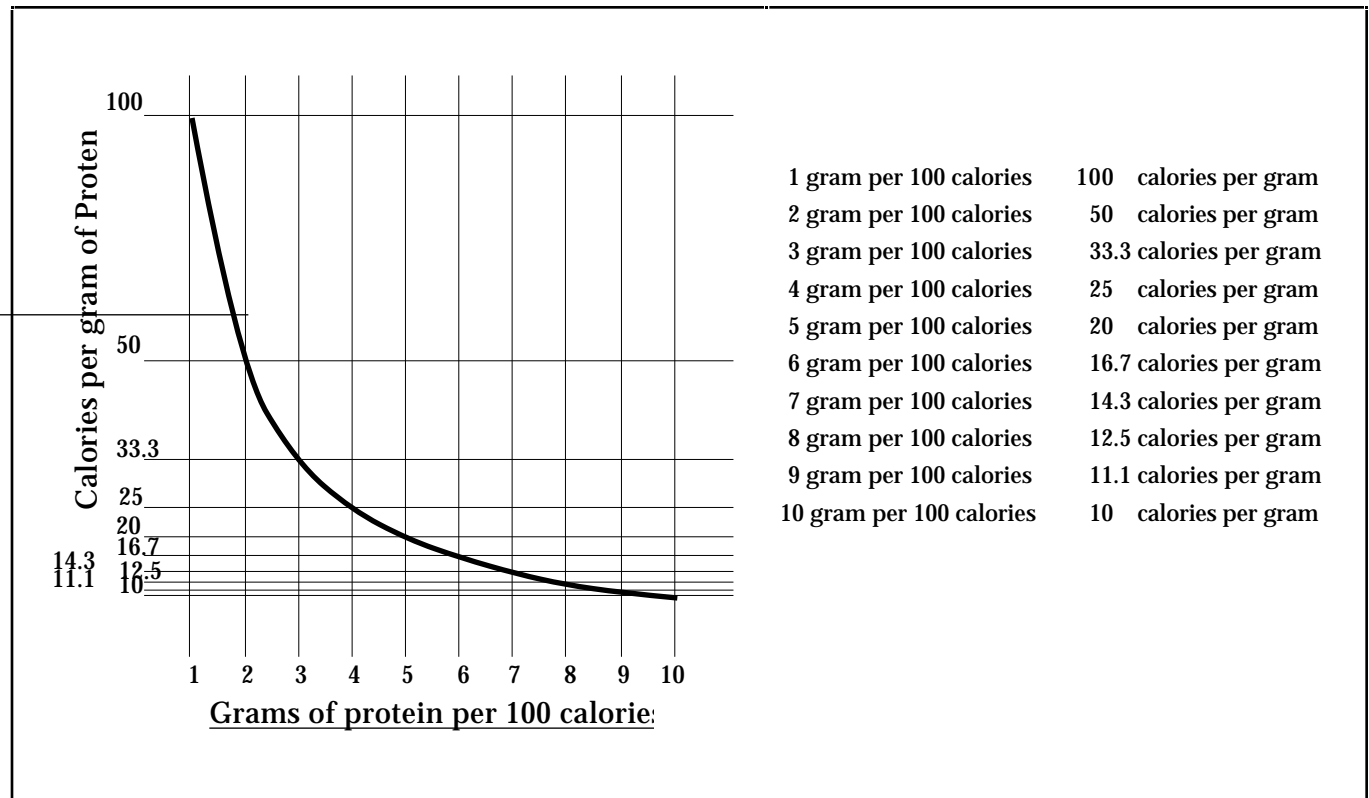
Figure 1b  
 Calories per gram of Protein  
 Stems: 10 Calories per gram, 20 calories per gram,  
 ... 60 calories per gram.

The two figures display the two Stem and Leaf Procedures using, in the first case, grams of protein per 100 calories and, in the second case, calories per gram of protein. The first S & L, Figure 1a is similar to what we saw earlier with four cases standing out on the high side, one of which, Gerber's High Protein is extreme.

The second S & L is a different shape. There is a dip at 50 followed by a bump at 60, and together these two stems identify four extreme cases. But these are not the same cases — these are all rice cereals. The Gerber's High Protein has migrated to the other end of the distribution. And it no longer stands out, demanding attention. In this picture of the data it is just one leaf on a stem with two other leaves.

This is, for me, “distressing”. When I analyzed the breakfast cereal data my intuition and the course of my investigation leaned heavily on what I saw. Here I “see” two different pictures. I have to assume that in the long run whichever picture I use I will end up with the same understanding of the underlying nature of these data — there is one reality behind these data and I had better find it. But it is also clear that the intuitions and the course of my investigation will start off in different directions depending on which picture I use at the beginning of my research.

What's going on? The two different procedures alter the unit of measure. In the first case, one unit, two units, three units, 1, 2, 3, .... counts grams of protein associated with 100 calories. In the second case, one unit, two units, three units counts calories associated with a gram of protein. This has changed the units and, most important, it has changed the intervals between the values.



The change in the unit of measure changes the intervals. That is why Gerber's High Pro has one third of the S&L to itself, in one picture, while Gerber's shares its stem with two other cereals in the other picture of the same data.

In one picture I may say that Gerber's High Pro has approximately double the protein per calorie of its nearest competitor. In the other picture I say that Gerber's High Pro has approximately the same number of calories per gram of protein as General Mills Hi Pro and Kellog's Special K. Both statements are correct. But they steer your intuition in different directions.

This is the proverbial “tip of the iceberg”. Consider, for example, the data comparing the number of physicians to the number of doctors in a collection of countries. (1975 data).

In physicians per 1,000 people, the median among these 137 countries is 0.385 physicians per thousand people, the number for Mauritania. There is a small number of countries with relatively large numbers of physicians per person among which the USSR and the United Arab Emirates are so extreme as to warrant consideration as special cases. The inner fences establish a “normal” range of variation from -1.5 physicians per 1,000 to 2.8 physicians per thousand.

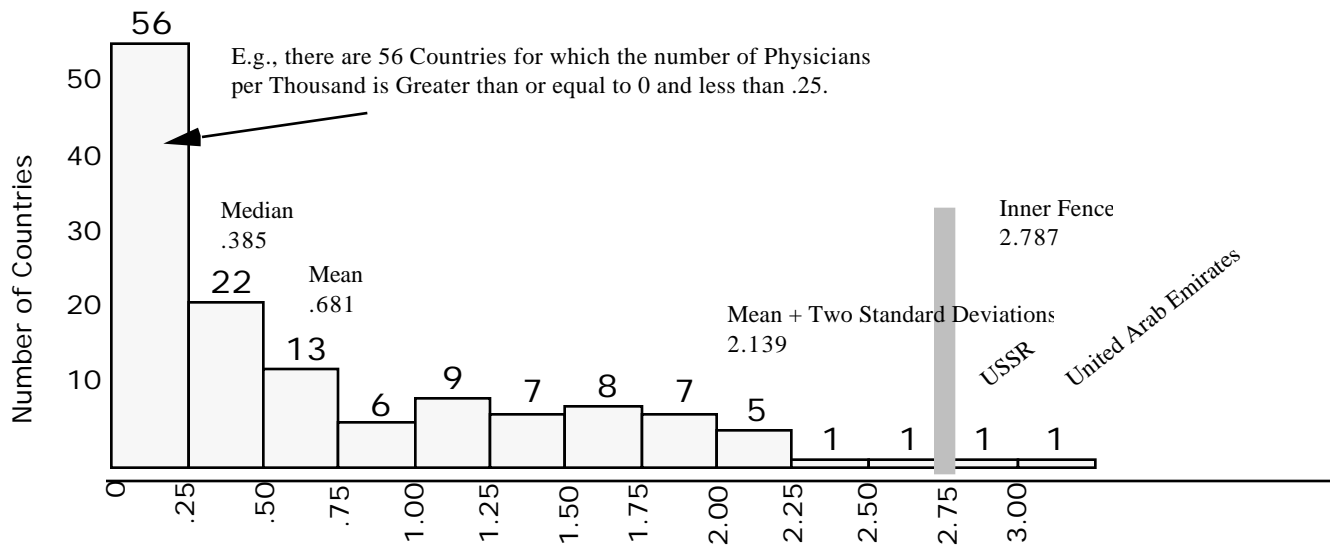


Figure 2

Distribution of 137 Countries with Respect to Physicians per Thousand Population, 1975 Data.

(In physicians per thousand, the median is .385 and the lower and upper quartiles are .077 and 1.161. The mean is .681 and the standard deviation is .729.

---

That's what the numbers say: They establish a reasonable range from -1.5 to 2.8 physicians per thousand, directing us that nothing within this range is so unusual as to warrant attention as an exception. I have my trouble with any method that directs me not to worry about a negative number of physicians per thousand, that's silly. But ignoring that, they direct my attention to the physician-intensive end for two unusual cases.

That's what the numbers say, or maybe it isn't.

In people per physician, the median among these 137 countries is 2,600 people, the number for Mauritania. The inner fences establish a “normal” range of variation from 861 physicians per person to 13,000 physicians per person. The only unusual cases are at the population-intensive end. On the physician-intensive end seven countries exceeding the inner fence and four more countries exceeding the outer fence. Together, the USSR and the United Arab Emirates show the smallest number of people per physician, but neither of these ratios is sufficiently different from adjacent values in the distribution to warrant attention as being different in kind from other countries.

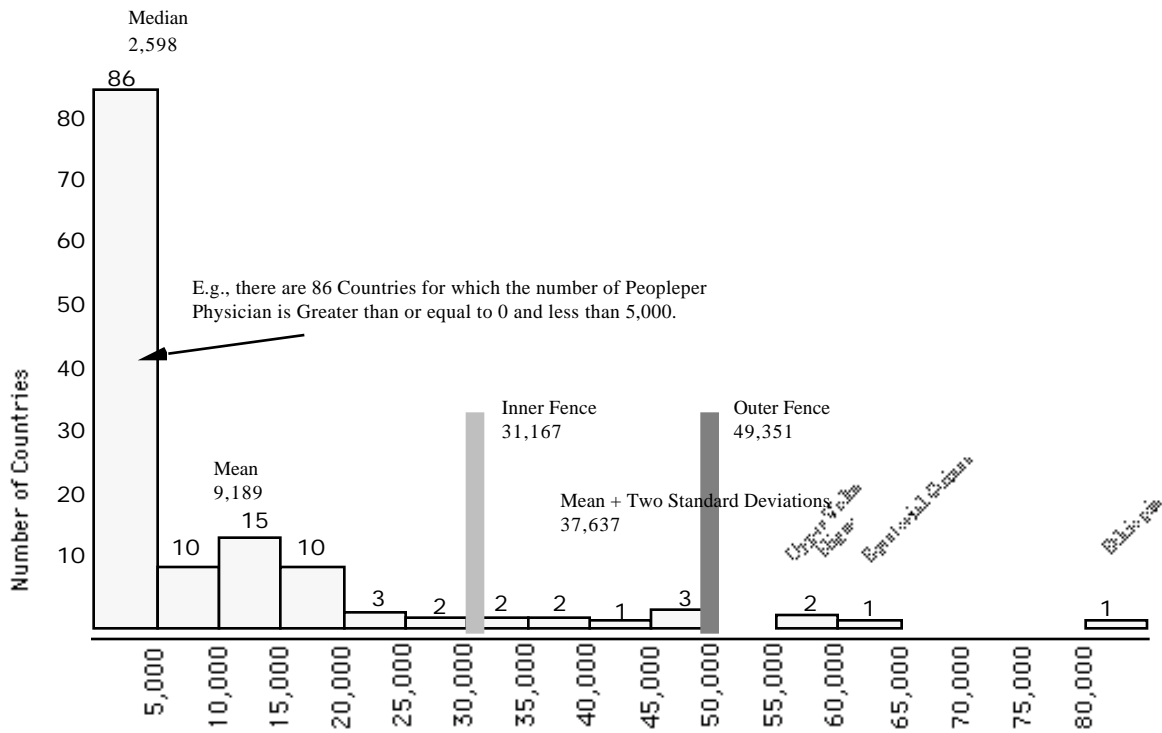


Figure 3

Distribution of 137 Countries with Respect to People per Physician, 1975 Data.

In physicians per person, the median is 2,598 and the lower and upper quartiles are 861 and 12,984. The mean is 9,189 and the standard deviation is 14,224.

This is the kind of now-you-see-it / now-you-don't stuff that gives statistics a bad name — it appears that if I don't like one picture of the data, then I am free to create another. Do I want to use school enrollment data to show that certain schools have excessively large classes — never mind the facts? Then I would use enrollment data organized to count *students per class*. In that form the data will tend to show a “tail” at the high end of the distribution, suggesting that something is wrong or out of line among the largest entries. Do I want to use school enrollment data to show that certain schools have extremely *favorable* faculty student ratios? Then I would organize the same enrollment data to count faculty per student. In that form the data will tend to produce a tail at the opposite end of the distribution where there are large faculty to student ratios.

But this kind of cheating is not data analysis, this is bad data analysis or, perhaps, clever propaganda using numbers and data to create a pretense that its conclusions are objective. This kind of manipulation may be done intentionally, in order to create a picture which is as favorable as possible with respect to the interests of the analyst or the client of the analyst. It may be done unintentionally — because the data were presented in one form and the computations were made on the data as given, without thought for the consequences or the alternatives. Whatever the reason, in the trade we have standards that go a long way toward preventing this kind of lying with statistics, and a long way toward detecting it when it is committed by others. In a phrase, the solution is the “well behaved variable.”

090 Phy/Cap people/Phy

	A	B	C	D	E	F	G	H
1	Country	Doctors	Total Population '75		Doctors/Person	Doctors Per 1,000 People		People per Doctor
2	UAR United Arab Emirates	681	220,000		0.003095455	3.095		323
3	USSR	733,744	255,038,000		0.002876999	2.877		348
4	Israel	9,144	3,417,000		0.002676032	2.676		374
5	Czechoslovakia	35,385	14,793,000		0.00239201	2.392		418
6	Bulgaria	18,773	8,793,000		0.002134994	2.135		468
7	Austria	15,702	7,538,000		0.002083046	2.083		480
8	Italy	114,228	55,023,000		0.002076005	2.076		482
9	Greece	18,423	8,930,000		0.002063046	2.063		485
10	Hungary	21,131	10,534,000		0.002005981	2.006		499
11	Germany West	122,069	61,682,000		0.001979005	1.979		505
12	Denmark	9,896	5,026,000		0.001968961	1.969		508
13	Argentina	48,687	25,384,000		0.001918019	1.918		521
14	Belgium	18,510	9,846,000		0.001879951	1.880		532
15	Germany East	31,308	17,127,000		0.001827991	1.828		547
16	Mongolia	2,604	1,446,000		0.00180083	1.801		555
17	Switzerland	11,469	6,535,000		0.001755011	1.755		570
18	Iceland	372	216,000		0.001722222	1.722		581
19	Poland	58,240	33,841,000		0.001720989	1.721		581
20	Norway	6,884	4,007,000		0.001717994	1.718		582
21	Canada	39,104	22,801,000		0.001715012	1.715		583
22	Sweden	14,045	8,291,000		0.001694006	1.694		590
23	USA	348,484	213,925,000		0.001629001	1.629		614
24	Netherlands	21,826	13,599,000		0.001604971	1.605		623
25	SPAN	54,992	35,433,000		0.001552	1.552		644
26	France	77,888	52,913,000		0.001472001	1.472		679
27	Finland	6,699	4,652,000		0.001440026	1.440		694
28	New Zealand	4,110	3,031,000		0.001355988	1.356		737
29	Romania	28,548	21,178,000		0.001348003	1.348		742
30	United Kingdom	75,612	56,427,000		0.001339997	1.340		746
31	Yugoslavia	27,143	21,322,000		0.001273004	1.273		786
32	Portugal	11,101	8,762,000		0.001266948	1.267		789
33	Ireland	3,773	3,131,000		0.001205046	1.205		830
34	Japan	133,344	111,120,000		0.0012	1.200		833
35	Puerto Rico	3,479	2,902,000		0.001198828	1.199		834
36	Malta	382	329,000		0.001161094	1.161		861
37	Libya	2,586	2,255,000		0.001146785	1.147		872
38	Luxembourg	368	342,000		0.001076023	1.076		929
39	Venezuala	13,105	12,213,000		0.001073037	1.073		932
40	Qatar	96	90,000		0.001066667	1.067		938
41	Kuwait	1,089	1,085,000		0.001003687	1.004		996
42	Cuba	8,201	9,481,000		0.000864993	0.865		1,156
43	Paraguay	2,229	2,647,000		0.000842085	0.842		1,188
44	Panama	1,404	1,678,000		0.00083671	0.837		1,195
45	Cyprus	547	673,000		0.000812779	0.813		1,230

090 Phy/Cap people/Phy

	A	B	C	D	E	F	G	H
46	Bahamas	161	200,000		0.000805	0.805		1,242
47	Lebanon	2,301	2,869,000		0.000802022	0.802		1,247
48	Togo	1,623	2,248,000		0.000721975	0.722		1,385
49	Peru	10,514	15,326,000		0.000686024	0.686		1,458
50	Hong Kong	2,881	4,225,000		0.000681893	0.682		1,467
51	Bahrain	177	260,000		0.000680769	0.681		1,469
52	Barabados	166	245,000		0.000677551	0.678		1,476
53	Costa Rica	1,292	1,994,000		0.000647944	0.648		1,543
54	Nicaragua	1,400	2,318,000		0.000603969	0.604		1,656
55	Brazil	62,656	109,730,000		0.000571002	0.571		1,751
56	Trinidad and Tobago	550	1,009,000		0.000545094	0.545		1,835
57	Turkey	21,696	39,882,000		0.000544005	0.544		1,838
58	Mexico	31,556	59,204,000		0.000533005	0.533		1,876
59	Korea South	17,851	34,663,000		0.000514987	0.515		1,942
60	Colombia	12,997	25,890,000		0.000502008	0.502		1,992
61	Ecuador	3,517	7,090,000		0.000496051	0.496		2,016
62	South Africa	12,060	24,663,000		0.000488992	0.489		2,045
63	Suriname	202	422,000		0.000478673	0.479		2,089
64	Bolivia	2,581	5,410,000		0.000477079	0.477		2,096
65	Dominican Republic	2,375	5,118,000		0.000464048	0.464		2,155
66	Vietnam South	9,000	19,650,000		0.000458015	0.458		2,183
67	Chile	4,419	10,253,000		0.000430996	0.431		2,320
68	Iraq	4,504	11,067,000		0.000406976	0.407		2,457
69	Saudi Arabia	3,613	8,966,000		0.000402967	0.403		2,482
70	Mauritius	346	899,000		0.000384872	0.385		2,598
71	Seychelles	21	60,000		0.00035	0.350		2,857
72	Iran	11,358	32,923,000		0.000344987	0.345		2,899
73	Western Somoa	55	160,000		0.00034375	0.344		2,909
74	Syria	2,403	7,259,000		0.000331037	0.331		3,021
75	Philippines	13,464	44,437,000		0.000302991	0.303		3,300
76	Honduras	920	3,037,000		0.000302931	0.303		3,301
77	Guyana	237	791,000		0.000299621	0.300		3,338
78	Jamaica	570	2,029,000		0.000280927	0.281		3,560
79	Jordan	745	2,688,000		0.000277158	0.277		3,608
80	El Salvador	1,117	4,108,000		0.000271908	0.272		3,678
81	Pakistan	17,922	70,560,000		0.000253997	0.254		3,937
82	Grenada	25	100,000		0.00025	0.250		4,000
83	India	145,946	613,217,000		0.000238001	0.238		4,202
84	Sri Lanka	3,245	13,986,000		0.000232018	0.232		4,310
85	Egypt	8,034	37,543,000		0.000213995	0.214		4,673
86	Tunisia	1,213	5,747,000		0.000211067	0.211		4,738
87	Oman	153	770,000		0.000198701	0.199		5,033
88	Guatemala	1,207	6,129,000		0.000196933	0.197		5,078
89	Gabon	96	521,000		0.000184261	0.184		5,427
90	Burma	5,561	31,240,000		0.000178009	0.178		5,618
91	Malaysia	2,007	12,093,000		0.000165964	0.166		6,025

090 Phy/Cap people/Phy

	A	B	C	D	E	F	G	H
92	Congo	213	1,345,000		0.000158364	0.158		6,315
93	Sao Tome and Principe	12	80,000		0.00015	0.150		6,667
94	Zimbabwe	916	6,272,000		0.000146046	0.146		6,847
95	Swaziland	65	469,000		0.000138593	0.139		7,215
96	Thailand	5,009	42,093,000		0.000118998	0.119		8,403
97	Ghana	938	9,873,000		9.50066E-05	0.095		10,526
98	Kenya	1,246	13,251,000		9.40306E-05	0.094		10,635
99	Madagascar	754	8,020,000		9.4015E-05	0.094		10,637
100	Zambia	470	5,004,000		9.39249E-05	0.094		10,647
101	Botswana	63	691,000		9.11722E-05	0.091		10,968
102	Haiti	396	4,552,000		8.69947E-05	0.087		11,495
103	Liberia	142	1,708,000		8.31382E-05	0.083		12,028
104	Sudan	1,407	18,268,000		7.70199E-05	0.077		12,984
105	Maldives	9	120,000		0.000075	0.075		13,333
106	Morocco	1,243	17,504,000		7.10123E-05	0.071		14,082
107	Senegal	305	4,418,000		6.90358E-05	0.069		14,485
108	Bangladesh	5,088	73,746,000		6.89936E-05	0.069		14,494
109	Comoros	21	306,000		6.86275E-05	0.069		14,571
110	Mauritania	87	1,283,000		6.78098E-05	0.068		14,747
111	Nigeria	4,224	63,049,000		6.69955E-05	0.067		14,926
112	Ivory Coast / Cote d'Ivoire	322	4,885,000		6.59161E-05	0.066		15,171
113	Guinea	278	4,416,000		6.29529E-05	0.063		15,885
114	Indonesia	8,299	136,044,000		6.10023E-05	0.061		16,393
115	Somalia	193	3,170,000		6.08833E-05	0.061		16,425
116	Angola	384	6,394,000		6.00563E-05	0.060		16,651
117	Yemen (Sana)	367	6,668,000		5.5039E-05	0.055		18,169
118	Cameroon	354	6,433,000		5.50288E-05	0.055		18,172
119	Tanzania	846	15,388,000		5.49779E-05	0.055		18,189
120	Mozambique	507	9,223,000		5.49713E-05	0.055		18,191
121	Central African Republic	97	1,790,000		5.41899E-05	0.054		18,454
122	Singapore	106	2,248,000		4.7153E-05	0.047		21,208
123	Laos	155	3,303,000		4.6927E-05	0.047		21,310
124	Lesotho	49	1,148,000		4.26829E-05	0.043		23,429
125	Uganda	431	11,353,000		3.79635E-05	0.038		26,341
126	Afghanistan	656	19,280,000		3.40249E-05	0.034		29,390
127	Zaire	807	24,450,000		3.30061E-05	0.033		30,297
128	Benin	95	3,074,000		3.09044E-05	0.031		32,358
129	Nepal	339	12,572,000		2.69647E-05	0.027		37,086
130	Rwanda	106	4,233,000		2.50413E-05	0.025		39,934
131	Mali	142	5,697,000		2.49254E-05	0.025		40,120
132	Burundi	83	3,765,000		2.20452E-05	0.022		45,361
133	Chad	83	3,947,000		2.10286E-05	0.021		47,554
134	Malawi	103	4,909,000		2.09819E-05	0.021		47,660
135	Upper Volta	109	6,032,000		1.80703E-05	0.018		55,339
136	Niger	83	4,600,000		1.80435E-05	0.018		55,422

090 Phy/Cap people/Phy

	A	B	C	D	E	F	G	H
137	Equatorial Guinea	5	313,000		1.59744E-05	0.016		62,600
138	Ethiopia	338	28,134,000		1.20139E-05	0.012		83,237
139								
140					Median	0.385		2,598
141					Low Q	0.077		861
142					High Q	1.161		12,984
143								
144					Spread	1.084		12,122
145					Step Size	1.626		18,184
146								
147					Low inner fence	-1.549		-17,322
148					High inner fence	2.787		31,167
149								
150					Low Outer Fence	-3.175		-35,506
151					High Outer Fence	4.413		49,351
152								
153								
154					Mean	0.681		9,189
155					Standard Dev	0.729		14,224
156					Mean-2sd	-0.777		-19,260
157					Mean+2sd	2.139		37,637