

Description Using Logs

Now, using the concept of a well-behaved variable, and using the strategy of re-expression, it's back to the basics of description. But this time we have more tools that can be used for the job, and thus more alternatives that require more thinking. The example will be the size of nations, by population, asking for a brief description of the sizes of nations and how their populations have changed between 1975 and 1990.

Country	Pop in 1,000's			
	1975	1990		
			Cape Verde	292 375
			Central African Republic	1,790 2,877
			Chad	3,947 5,017
			Chile	10,253 13,083
Afghanistan	19,280	15,564	China / People's Republic of	
Albania	2,482	3,273	China / Mainland	838,803 1,133,683
Algeria	16,792	25,337	Colombia	25,890 33,076
Andorra		52	Comoros	306 460
Angola	6,394	8,449		
Antigua and Barbuda		64	Congo	1,345 2,242
Argentina	25,384	32,291	Costa Rica	1,994 3,033
Armenia		3,357	Croatia	4,686
Aruba		64	Cuba	9,481 10,620
Australia	13,809	17,037	Cyprus	673 702
			Czechoslovakia	14,793 15,683
Austria	7,538	7,644	Denmark	5,026 5,131
Bahamas	200	249	Djibouti	337
Bahrain	260	520	Dominica	85
Bangladesh		113,930	Dominican Republic	5,118 7,241
Bangladesh	73,746			
Barbados	245	254	Ecuador	7,090 10,507
Belgium	9,846	9,909	Egypt	37,543 53,212
Belize		220	El Salvador	4,108 5,310
Benin	3,074	4,674	Equatorial Guinea	313 369
Bhutan	1,173	1,566	Estonia	1,584
			Ethiopia	28,134 51,407
Bolivia	5,410	6,989	Fiji	577 738
Bosnia Herzegovina		4,517	Finland	4,652 4,977
Botswana	691	1,224	France	52,913 56,358
Brazil	109,730	152,505	Gabon	521 1,068
Brunei		372		
Bulgaria	8,793	8,934	Gambia	509 848
Burkina		9,078	Georgia	5,479
Burkina			Germany	79,123
Burma	31,240	41,277	Germany East	17,127
Burundi	3,765	5,646	Germany West	61,682 63,232
			Ghana	9,873 15,130
Byelarus		10,257	Greece	8,930 10,028
Cameroon	6,433	11,092	Grenada	100 84
Canada	22,801	26,538		

Guatemala	6,129	9,038	Netherlands	13,599	14,936
Guinea	4,416	7,269	New Zealand	3,031	3,296
			Nicaragua	2,318	3,602
Guinea-Bissau	525	999	Niger	4,600	7,879
Guyana	791	753	Nigeria	63,049	118,819
Haiti	4,552	6,142	Norway	4,007	4,253
Honduras	3,037	4,804	Oman	770	1,481
Hong Kong	4,225		Pakistan	70,560	114,649
Hungary	10,534	10,569	Panama	1,678	2,425
Iceland	216	257	Papua New Guinea	2,716	3,823
India	613,217	852,667			
Indonesia	136,044	190,136	Paraguay	2,647	4,660
Iran	32,923	57,003	Peru	15,326	21,906
			Philippines	44,437	64,404
Iraq	11,067	18,782	Poland	33,841	37,777
Ireland	3,131	3,500	Portugal	8,762	10,354
Israel	3,417	4,436	Puerto Rico	2,902	
Italy	55,023	57,664	Qatar	90	491
Ivory Coast			Romania	21,178	23,273
Cote d'Ivoire	4,885	12,478	Russia		148,254
Jamaica	2,029	2,469	Rwanda	4,233	7,609
Japan	111,120	123,567			
Jordan	2,688	3,273	Saint Kits and Nevis		40
Kampuchia / Cambodi	8,110	6,991	Santa Lucia		150
			Saint Vincent and the Grenadines	80	113
Kenya	13,251	24,342	San Marino		23
Kiribati		70	Sao Tome and Principe		125
Korea North	15,852	21,412	Saudi Arabia	8,966	17,116
Korea South	34,663	42,792	Senegal	4,418	7,714
Kuwait	1,085	2,124	Serbia		9,883
Kyrgystan		4,394	Seychelles	60	68
Laos	3,303	4,024	Sierra Leone	2,983	4,166
Latvia		2,695			
Lebanon	2,869	3,339	Singapore	2,248	2,721
Lesotho	1,148	1,755	Somalia	3,170	6,654
			South Africa	24,663	39,539
Liberia	1,708	2,640	Soviet Union frmr		
Libya	2,255	4,223	Spain	35,433	39,269
Liechtenstein		28	Sri Lanka	13,986	17,198
Lithuania		3,726	Sudan	18,268	26,245
Luxembourg	342	384	Suriname	422	397
Madagascar	8,020	11,801	Swaziland	469	837
Malawi	4,909	9,197	Sweden	8,291	8,526
Malaysia	12,093	17,556			
Maldives	120	218	Switzerland	6,535	6,742
Mali	5,697	8,142	Syria	7,259	12,483
			Tajikistan		5,342
Malta	329	353	Taiwan / Republic of China	16,453	20,435
Mauritania	1,283	1,935	Tanzania	15,388	25,971
Mauritius	899	1,072	Thailand	42,093	56,002
Mexico	59,204	88,010	Togo	2,248	3,674
Moldova		4,393	Trinidad and Tobago	1,009	1,271
Mongolia	1,446	2,187	Tunisia	5,747	8,104
Morocco	17,504	25,630	Turkey	39,882	57,285
Mozambique	9,223	14,539			
Namibia		1,453	Turkmenistan		3,658
Nepal	12,572	19,146			

Tuvalu		9	Venezuala	12,213	19,698
UAR United Arab Emirates	220	2,254	Vietnam	43,451	66,171
Uganda	11,353	18,016	Vietnam North	23,800	
Ukraine		51,711	Vietnam South	19,650	
United Kingdom	56,427	57,366	Western Somoa	160	186
Upper Volta	6,032		Yemen		9,746
Uruguay	3,108	3,102	Yemen (Aden)	1,660	
USA	213,925	250,410	Yemen (Sana)	6,668	
USSR	255,038		Yugoslavia	21,322	
Uzbekistan		20,569	Zaire	24,450	36,613
Venuatu		165	Zambia	5,004	8,154
			Zimbabwe	6,272	10,394

Table 1
Countries of the World: 1975 and 1990 Population

Source, 1975: *World Handbook of Political and Social Indicators, Volume I*, Taylor and Jodice. Original source Labour Force Estimates and Projections, 1950-2000, ILO, Geneva, 1977, and Demographic Yearbook, 1977.. Source, 1990: *Statistical Abstract of the United States*, 1991 Table 1,359, compiled by the U.S. Census Bureau from various original sources.

World Population: The Work

So now, from the beginning: Who, What, Where, ...: The data are from the United Nations and in turn from national sources. How good are the data? Well, the text in the secondary sources I am using, from *World Handbook of Political and Social Indicators* and the *Statistical Abstract of the United States* warns me that standards differ from country to country. For example, some do and some do not count aborigines, nomadic peoples, displaced persons, or refugees. The separate counts are based on varying methods including attempts at complete counts, including samples, including registration censuses based on voting or tax registers. So, the data are a mixed lot. But, the data are also the best I can get — the United Nations sources have attempted to adjust for these inconsistencies. And were I reject these data on population, notwithstanding their blemishes, I would be acting as if there

were no data on populations — because I would have rejected the best. So, I'll accept the data, with caution.

Now for a first look at the data, stem and leaf. The national populations range from a low of “9”, which is nine thousand, to a high of 1,133,683, which is one billion. If I attempt to break this range into approximately ten equal stems, dividing the range into intervals of 100,000 each (one hundred million each), I will get a mess — I can see that coming by just looking at the counts, without completing the stem and leaf:

Stems		Leaves	
0- 100,000			170 countries
100,000 - 199,999			7 countries
200,000 - 299,999			1 country

800,000 - 899,999			1 country

1,000,000 -			1 country.

No point in completing this stem and leaf, I already know what it looks like: Most of the countries are “piled up” at the low end of the scale. There are a few very large countries forming a “tail” at the high end of the distribution.

I could persist with the stem and leaf, changing scales, omitting very large nations, and doing it again. But I'm in a hurry, I'm always in a hurry, so I'll compromise by simply putting the countries in order by size. The printed page, ranked by size, gives me much of what I need from the stem and leaf and, since this is an extremely “skewed” distribution, (very asymmetrical, with a few very large values) it avoids the work of the stem and leaf — which (I already know) is unlikely to pay off with a good-looking stem and leaf.

	Country	Pop in 1,000's			Description	Using Logs
		1975	1990			
			54		Liberia	1,708
			55		Latvia	2,640
			56		Singapore	2,695
			57		Central African	2,721
1	Tuvalu	9	57		Costa Rica	2,877
2	San Marino	23	58		Uruguay	1,994
3	Liechtenstein	28	59		Albania	3,108
4	Saint Kits and Nev	40	60		Jordan	2,482
5	Andorra	52	61		New Zealand	2,688
6	Antigua and Barbud	64	62		Lebanon	3,031
7	Aruba	64	63		Armenia	2,869
8	Seychelles	60	64		Ireland	3,357
9	Kiribati	70	65		Nicaragua	3,131
10	Grenada	100	70		Turkmenistan	2,318
11	Dominica		84		Togo	2,248
12	Saint Vincent and	80	85		Lithuania	3,658
13	Sao Tome and Princi		113		Papua New Guinea	2,716
14	Santa Lucia		125		Laos	3,303
15	Venuatu		150		Sierra Leone	2,983
16	Western Somoa	160	165		Libya	2,255
17	Maldives	120	186		Norway	4,007
18	Belize		218		Moldova	4,393
19	Bahamas	200	220		Kyrgystan	3,417
20	Barabados	245	249		Israel	2,647
21	Iceland	216	254		Bosnia Herzogovin	3,074
22	Djibouti		257		Paraguay	2,647
23	Malta	329	337		Benin	3,074
24	Equatorial Guine	313	353		Croatia	4,686
25	Brunei		369		Honduras	3,037
26	Cape Verde	292	372		Finland	4,652
27	Luxembourg	342	375		Chad	3,947
28	Suriname	422	384		Denmark	5,026
29	Comoros	306	397		El Salvador	4,108
30	Qatar	90	460		Tajikistan	5,342
31	Bahrain	260	491		Georgia	5,479
32	Cyprus	673	520		Burundi	3,765
33	Fiji	577	702		Haiti	4,552
34	Guyana	791	738		Somalia	3,170
35	Swaziland	469	753		Switzerland	6,535
36	Gambia	509	837		Bolivia	5,410
37	Guinea-Bissau	525	848		Kampuchia / C	8,110
38	Gabon	521	999		Dominican Rep	5,118
39	Mauritius	899	1,068		Guinea	4,416
40	Botswana	691	1,072		Rwanda	4,233
41	Trinidad and T	1,009	1,224		Austria	7,538
42	Namibia		1,271		Senegal	4,418
43	Oman	770	1,453		Niger	4,600
44	Bhutan	1,173	1,481	100	Tunisia	5,747
45	Estonia		1,566	101	Mali	5,697
46	Lesotho	1,148	1,584	102	Zambia	5,004
47	Mauritania	1,283	1,755	103	Angola	6,394
48	Kuwait	1,085	1,935	104	Sweden	8,291
49	Mongolia	1,446	2,124	105	Bulgaria	8,793
50	Congo	1,345	2,187	106	Guatemala	6,129
51	UAR United Arab	220	2,242	107	Burkina	9,078
52	Panama	1,678	2,254	108	Malawi	4,909
53	Jamaica	2,029	2,425	109	Yemen	9,746
			2,469	110		

111	Serbia		9,883	153	SPAN	35,433	39,269
112	Belgium	9,846	9,909	154	South Africa	24,663	39,539
113	Greece	8,930	10,028	155	Burma	31,240	41,277
114	Byelarus		10,257	156	Korea South	34,663	42,792
115	Portugal	8,762	10,354	157	Ethiopia	28,134	51,407
116	Zimbabwe	6,272	10,394	158	Ukraine		51,711
117	Ecuador	7,090	10,507	159	Egypt	37,543	53,212
118	Hungary	10,534	10,569	160	Thailand	42,093	56,002
119	Cuba	9,481	10,620	161	France	52,913	56,358
120	Cameroon	6,433	11,092	162	Iran	32,923	57,003
121	Madagascar	8,020	11,801	163	Turkey	39,882	57,285
122	Ivory Coast / C4	8,885	12,478	164	United Kingdom	56,427	57,366
123	Syria	7,259	12,483	165	Italy	55,023	57,664
124	Chile	10,253	13,083	166	Germany West	61,682	63,232
125	Mozambique	9,223	14,539	167	Philippines	44,437	64,404
126	Netherlands	13,599	14,936	168	Vietnam	43,451	66,171
127	Ghana	9,873	15,130	169	Germany		79,123
128	Afghanistan	19,280	15,564	170	Mexico	59,204	88,010
129	Czechoslovakia	14,793	15,683	171	Bangladesh		113,930
130	Australia	13,809	17,037	172	Pakistan	70,560	114,649
131	Saudi Arabia	8,966	17,116	173	Nigeria	63,049	118,819
132	Sri Lanka	13,986	17,198	174	Japan	111,120	123,567
133	Malaysia	12,093	17,556	175	Russia		148,254
134	Uganda	11,353	18,016	176	Brazil	109,730	152,505
135	Iraq	11,067	18,782	177	Indonesia	136,044	190,136
136	Nepal	12,572	19,146	178	USA	213,925	250,410
137	Venezuela	12,213	19,698	179	India	613,217	852,667
138	Taiwan / Rep	16,453	20,435	180	China / Peop	838,803	1,133,683
139	Uzbekistan		20,569	181	Bangladesh	73,746	
140	Korea North	15,852	21,412	182	Burkina		
141	Peru	15,326	21,906	183	Germany East	17,127	
142	Romania	21,178	23,273	184	Hong Kong	4,225	
143	Kenya	13,251	24,342	185	Puerto Rico	2,902	
144	Algeria	16,792	25,337	186	Soviet Union frmr		
145	Morocco	17,504	25,630	187	Upper Volta	6,032	
146	Tanzania	15,388	25,971	188	USSR	255,038	
147	Sudan	18,268	26,245	189	Vietnam North	23,800	
148	Canada	22,801	26,538	190	Vietnam South	19,650	
149	Argentina	25,384	32,291	191	Yemen (Aden)	1,660	
150	Colombia	25,890	33,076	192	Yemen (Sana)	6,668	
151	Zaire	24,450	36,613	193	Yugoslavia	21,322	
152	Poland	33,841	37,777				

Ah, I notice immediately from the rank order on 1990 population, that I don't have a 1990 population for all of these countries — the list of nations varies from year to year. Checking the names, I see that the change in nations is a result of fusion and fission. Do I attempt to compensate for this, changing the units of the 1975 list to correspond to the units of the 1990 list, standardizing my list? No, at some point that may be called for, but to decide on the "correct" standardization I

would have to have a clear purpose in mind (with respect to which I could decide what was “right”.) Having no particularly subtle purpose in mind, beyond description, changing the list now would simply change the data — to no apparent end. So, I’ll continue, acknowledging that it may be difficult to compare the 1975 data to the 1990.

Now, I’m going to stop for a moment: Why? Because I was about to yield to a mindless reflex: I was about to compute averages and measures of variation on everything in sight. That kind of mindless reflex is to be treated with great caution. So, thinking about these data, before I commit to a lot of computation, what do I already know and what do I suspect: Invoking the list of 4 properties for well-behaved variables, I know that population fails criterion #1 and therefore I suspect that population will fail on criterion #4. I know that the distribution is not symmetrical, criterion #1. And therefore, I suspect that the unit of measure is the wrong unit, criterion #4.

1990 population surely fails the first property of a well behaved variable — it is anything but symmetrical. And, ordinarily that would be enough to stop me, but for pedagogical purposes, let me show you the kind of trouble I would get in to if I yielded to mindless reflex.

[Check how Excel defines median and how it uses missing data: I got a different value for the median, using the median built in function, than I got by picking the median out of the rank order.]

	1975	1990	
Sum	4,021,291	5,347,251	Added because I didn’t trust the difference between the means and wanted, therefore, to look at the difference between the sums
Mean	25,944	29,707	Verify directly
Standard Deviation (Excel function stdevp)	87,477	111,616	Verify directly
Median	5,722	6,398	Verify
Low Quartile	1,994	1,699.5	Verify
High Quartile	17,127	17,786	Verify
Quartile Spread	15,133	16,119.5	

[Looking that over, first look at the means: increasing from 26 million to 30 million, about 16 percent. That seems a little odd, 16 percent increase during fifteen years is approximately one percent a year. That seems low — my memory tells me that the growth rate for world population is about 2 percent per year. Shouldn't the average also be growing by about 2 percent?

Well, still thinking, maybe and maybe not: This is not *world population*, it is average size of nations, which is different. So, maybe. Let me re-assure myself by checking world population, adding up the populations: Ah, 4 billion in 1975 up to 5.3 billion, up about 33%. So, yes, the total seems in line with what I expect, approximately 2 percent per year. That warns me to be careful about the unit: these are nations (actually states). The means may be showing the trace of the breakup of the Soviet Union into smaller countries. O.K. I'm ready to continue.]

First, look at those standard deviations: In both cases they exceed the means, substantially — the standard deviations are more than three times greater than their respective mean. That's strange: If those are the numbers, then those are the numbers, barring numerical error. But still, think of what those numbers are supposed to mean: They are supposed to represent — and put a number on — what you saw in the picture. The “standard deviation” is supposed to describe standard or typical average variation around the mean but these numbers are much too large for that purpose: For 1975, 150 of the 155 countries are *less than* one standard deviation away from the mean.

Putting it another way, remember that we will use a standard deviation to describe the middle range of data. But it just doesn't do the job with this picture: By the numbers, the middle range of data would be between - 61,533 (minus 62 million) and + 113,421 (plus 113 million) — calculating the mean minus one standard deviation and the mean plus one standard deviation. And, intuitively, that's just silly as a description of “typical” population: What is a negative population?

These numbers fail to do their job, which is to represent what the facts in the picture — I can do better than that without even

looking at the data (because I know full well that there are *no* countries with negative populations). And so, in these standard deviations I see both the second criterion of well-behaved variables being violated — the standard deviations change sharply between 1975 and 1990 — and I see the fourth criterion beginning to get shaky — the standard deviation does not describe the picture of the data and its values are difficult to interpret.

The medians and the quartiles, of course, give me interpretable numbers — they have to because they always refer to particular cases. That is one reason why medians and quartiles are used, often, in the process of research (although means and standard deviations are more often what is shown in a final report.) But I also note, somewhat uncomfortably, that there is a substantial difference between the “average” I get from one method versus the average I get from the other, between the mean and the median — two very different reports about the middle of the distribution: The median says the average is 6 million people while the mean says the average is 26 million people (for 1975). The median says the average is 6 million people while the mean says the average is 30 million people (for 1990). Now, its not that I can't cope with these numbers and their oddities. Such peculiarities are typical of badly-behaved variables and I can cope with them if I must. But I don't need to.

Now, let's look at the logarithms of these numbers, changing the unit of measure to the logarithm of population, using logarithms base 10.

Nation	Population in 1,000's		logarithms Base 10	
	1975	1990	1975	1990
1 Tuvalu		9		0.954
2 San Marino		23		1.362
3 Liechtenstein		28		1.447
4 Saint Kits and Nevis		40		1.602
5 Andorra		52		1.716
6 Antigua and Barbuda		64		1.806
7 Aruba		64		1.806
8 Seychelles	60	68	1.778	1.833
9 Kiribati		70		1.845

Rules of Evidence		Description Using Logs			
10	Grenada	100	84	2.000	1.924
11	Dominica		85		1.929
12	Saint Vincent and the Gren	80	113	1.903	2.053
13	Sao Tome and Principe		125		2.097
14	Santa Lucia		150		2.176
15	Venuatu		165		2.217
16	Western Somoa	160	186	2.204	2.270
17	Maldives	120	218	2.079	2.338
18	Belize		220		2.342
19	Bahamas	200	249	2.301	2.396
20	Barabados	245	254	2.389	2.405
21	Iceland	216	257	2.334	2.410
22	Djibouti		337		2.528
23	Malta	329	353	2.517	2.548
24	Equatorial Guinea	313	369	2.496	2.567
25	Brunei		372		2.571
26	Cape Verde	292	375	2.465	2.574
27	Luxembourg	342	384	2.534	2.584
28	Suriname	422	397	2.625	2.599
29	Comoros	306	460	2.486	2.663
30	Qatar	90	491	1.954	2.691
31	Bahrain	260	520	2.415	2.716
32	Cyprus	673	702	2.828	2.846
33	Fiji	577	738	2.761	2.868
34	Guyana	791	753	2.898	2.877
35	Swaziland	469	837	2.671	2.923
36	Gambia	509	848	2.707	2.928
37	Guinea-Bissau	525	999	2.720	3.000
38	Gabon	521	1,068	2.717	3.029
39	Mauritius	899	1,072	2.954	3.030
40	Botswana	691	1,224	2.839	3.088
41	Trinidad and Tobago	1,009	1,271	3.004	3.104
42	Namibia		1,453		3.162
43	Oman	770	1,481	2.886	3.171
44	Bhutan	1,173	1,566	3.069	3.195
45	Estonia		1,584		3.200
46	Lesotho	1,148	1,755	3.060	3.244
47	Mauritania	1,283	1,935	3.108	3.287
48	Kuwait	1,085	2,124	3.035	3.327
49	Mongolia	1,446	2,187	3.160	3.340
50	Congo	1,345	2,242	3.129	3.351
51	UAR United Arab Emirates	220	2,254	2.342	3.353
52	Panama	1,678	2,425	3.225	3.385
53	Jamaica	2,029	2,469	3.307	3.393
54	Liberia	1,708	2,640	3.232	3.422
55	ILatvia		2,695		3.431
56	Singapore	2,248	2,721	3.352	3.435
57	Central African Republic	1,790	2,877	3.253	3.459
58	Costa Rica	1,994	3,033	3.300	3.482
59	Uruguay	3,108	3,102	3.492	3.492
60	Albania	2,482	3,273	3.395	3.515
61	Jordan	2,688	3,273	3.429	3.515

Rules of Evidence

Description Using Logs

62	New Zealand	3,031	3,296	3.482	3.518
63	Lebanon	2,869	3,339	3.458	3.524
64	Armenia		3,357		3.526
65	Ireland	3,131	3,500	3.496	3.544
66	Nicaragua	2,318	3,602	3.365	3.557
67	Turkmenistan		3,658		3.563
68	Togo	2,248	3,674	3.352	3.565
69	Lithuania		3,726		3.571
70	Papua New Guinea	2,716	3,823	3.434	3.582
71	Laos	3,303	4,024	3.519	3.605
72	Sierra Leone	2,983	4,166	3.475	3.620
73	Libya	2,255	4,223	3.353	3.626
74	Norway	4,007	4,253	3.603	3.629
75	Moldova		4,393		3.643
76	Kyrgystan		4,394		3.643
77	Israel	3,417	4,436	3.534	3.647
78	Bosnia Herzegovina		4,517		3.655
79	Paraguay	2,647	4,660	3.423	3.668
80	Benin	3,074	4,674	3.488	3.670
81	Croatia		4,686		3.671
82	Honduras	3,037	4,804	3.482	3.682
83	Finland	4,652	4,977	3.668	3.697
84	Chad	3,947	5,017	3.596	3.700
85	Denmark	5,026	5,131	3.701	3.710
86	El Salvador	4,108	5,310	3.614	3.725
87	Tajikistan		5,342		3.728
88	Georgia		5,479		3.739
89	Burundi	3,765	5,646	3.576	3.752
90	Haiti	4,552	6,142	3.658	3.788
91	Somalia	3,170	6,654	3.501	3.823
92	Switzerland	6,535	6,742	3.815	3.829
93	Bolivia	5,410	6,989	3.733	3.844
94	Kampuchia / Cambodia	8,110	6,991	3.909	3.845
95	Dominican Republic	5,118	7,241	3.709	3.860
96	Guinea	4,416	7,269	3.645	3.861
97	Rwanda	4,233	7,609	3.627	3.881
98	Austria	7,538	7,644	3.877	3.883
99	Senegal	4,418	7,714	3.645	3.887
100	Niger	4,600	7,879	3.663	3.896
101	Tunisia	5,747	8,104	3.759	3.909
102	Mali	5,697	8,142	3.756	3.911
103	Zambia	5,004	8,154	3.699	3.911
104	Angola	6,394	8,449	3.806	3.927
105	Sweden	8,291	8,526	3.919	3.931
106	Bulgaria	8,793	8,934	3.944	3.951
107	Guatemala	6,129	9,038	3.787	3.956
108	Burkina		9,078		3.958
109	Malawi	4,909	9,197	3.691	3.964
110	Yemen		9,746		3.989
111	Serbia		9,883		3.995
112	Belgium	9,846	9,909	3.993	3.996
113	Greece	8,930	10,028	3.951	4.001
114	Byelarus		10,257		4.011

Rules of Evidence

Description Using Logs

115 Portugal	8,762	10,354	3.943	4.015
116 Zimbabwe	6,272	10,394	3.797	4.017
117 Ecuador	7,090	10,507	3.851	4.021
118 Hungary	10,534	10,569	4.023	4.024
119 Cuba	9,481	10,620	3.977	4.026
120 Cameroon	6,433	11,092	3.808	4.045
121 Madagascar	8,020	11,801	3.904	4.072
122 Ivory Coast / Cote d'Ivoire	4,885	12,478	3.689	4.096
123 Syria	7,259	12,483	3.861	4.096
124 Chile	10,253	13,083	4.011	4.117
125 Mozambique	9,223	14,539	3.965	4.163
126 Netherlands	13,599	14,936	4.134	4.174
127 Ghana	9,873	15,130	3.994	4.180
128 Afghanistan	19,280	15,564	4.285	4.192
129 Czechoslovakia	14,793	15,683	4.170	4.195
130 Australia	13,809	17,037	4.140	4.231
131 Saudi Arabia	8,966	17,116	3.953	4.233
132 Sri Lanka	13,986	17,198	4.146	4.235
133 Malaysia	12,093	17,556	4.083	4.244
134 Uganda	11,353	18,016	4.055	4.256
135 Iraq	11,067	18,782	4.044	4.274
136 Nepal	12,572	19,146	4.099	4.282
137 Venezuela	12,213	19,698	4.087	4.294
138 Taiwan / Republic of China	16,453	20,435	4.216	4.310
139 Uzbekistan		20,569		4.313
140 Korea North	15,852	21,412	4.200	4.331
141 Peru	15,326	21,906	4.185	4.341
142 Romania	21,178	23,273	4.326	4.367
143 Kenya	13,251	24,342	4.122	4.386
144 Algeria	16,792	25,337	4.225	4.404
145 Morocco	17,504	25,630	4.243	4.409
146 Tanzania	15,388	25,971	4.187	4.414
147 Sudan	18,268	26,245	4.262	4.419
148 Canada	22,801	26,538	4.358	4.424
149 Argentina	25,384	32,291	4.405	4.509
150 Colombia	25,890	33,076	4.413	4.520
151 Zaire	24,450	36,613	4.388	4.564
152 Poland	33,841	37,777	4.529	4.577
153 SPAN	35,433	39,269	4.549	4.594
154 South Africa	24,663	39,539	4.392	4.597
155 Burma	31,240	41,277	4.495	4.616
156 Korea South	34,663	42,792	4.540	4.631
157 Ethiopia	28,134	51,407	4.449	4.711
158 Ukraine		51,711		4.714
159 Egypt	37,543	53,212	4.575	4.726
160 Thailand	42,093	56,002	4.624	4.748
161 France	52,913	56,358	4.724	4.751
162 Iran	32,923	57,003	4.517	4.756
163 Turkey	39,882	57,285	4.601	4.758
164 United Kingdom	56,427	57,366	4.751	4.759
165 Italy	55,023	57,664	4.741	4.761
166 Germany West	61,682	63,232	4.790	4.801
167 Philippines	44,437	64,404	4.648	4.809

Rules of Evidence	Description Using Logs			
168 Vietnam	43,451	66,171	4.638	4.821
169 Germany		79,123		4.898
170 Mexico	59,204	88,010	4.772	4.945
171 Bangladesh		113,930		5.057
172 Pakistan	70,560	114,649	4.849	5.059
173 Nigeria	63,049	118,819	4.800	5.075
174 Japan	111,120	123,567	5.046	5.092
175 Russia		148,254		5.171
176 Brazil	109,730	152,505	5.040	5.183
177 Indonesia	136,044	190,136	5.134	5.279
178 USA	213,925	250,410	5.330	5.399
179 India	613,217	852,667	5.788	5.931
180 China / People's Repu	838,803	1,133,683	5.924	6.054
181 Yemen (Aden)	1,660		3.220	
182 Puerto Rico	2,902		3.463	
183 Hong Kong	4,225		3.626	
184 Upper Volta	6,032		3.780	
185 Yemen (Sana)	6,668		3.824	
186 Germany East	17,127		4.234	
187 Vietnam South	19,650		4.293	
188 Yugoslavia	21,322		4.329	
189 Vietnam North	23,800		4.377	
190 Bangladesh	73,746		4.868	
191 USSR	255,038		5.407	
192 Burkina				
193 Soviet Union frmr				

Beginning again, the first thing you “see” using logs is that the numbers are unfamiliar. That’s not good, it leads to error, so I’ve introduced a few bench marks by using logs base 10. Using logs base 10, a “2” in logs, corresponds to 100 without logs. So, St. Vincent with approximately 100 (approximately one hundred thousand people) will have a log, base 10, of approximately 2. Using logs base 10, a “3” in logs corresponds to 1,000 without logs. So Guinea-Bissau? with approximately 1,000 (approximately one million people) will have a log, base 10, of approximately 3. Belgium with approximately 10,000 (approximately ten million people) will have a log, base 10, of approximately 4. And, while you will become accustomed to these numbers, there is no harm done by keeping the original values in the table, for backup.

Now for the picture, the shape of the stem and leaf. Preparing to select boundaries for the stems, I check the range, finding a range between .954 and 6.05. Using convenient boundaries to get about ten

stems, and just checking the counts I would get for these stems, I get Figure _. I will not actually construct the stem and leaf because with the rank ordering in Figure _, including the names, and the shape shown in Figure _, I have what I need.

0.5 -	.999		1 country
1.0 -	1.499		2 countries
1.5 -	1.999		8countries
2.0 -	2.499		10 countries
2.5 -	2.999		15 countries
3.0 -	3.499		23 countries
3.5 -	3.999		53 countries
4.0 -	4.499		36 countries
4.5 -	4.999		22 countries
5.0 -	5.499		8 countries
5.5 -	5.999		1 country
6.0 -	+		1 country

That's good — decidedly closer to symmetry than the original and suggesting that log population may be a well-behaved (or relatively well-behaved) variable. Let's find the numbers that are supposed to describe the picture.

	1975	1990	
Mean	3.724	3.688	Verify
Standard Deviation (Excel function stdevp)	.787	.911	Verify
Median	3.759	3.8055	Verify
Low Quartile	3.3035	3.222	Verify
High Quartile	4.2295	4.278	Verify
Quartile Spread	.926	1.056	

Looking at the averages:

This is certainly different. Let's take it apart. The mean is approximately 3.7 to 3.8 and so is the median. Whichever average I choose, I get approximately the same result. That's another sign of symmetry (Symmetry will place the both the median and the mean in the middle, near the peak. So this is another suggestion that the variable is well-behaved.) And keeping myself grounded in the units for which I have intuition, 3.7 is the log, base 10, of a number slightly less than 10,000 (with 4 zeroes), implying approximately 10 million people as the central value for the distribution of populations — approximately the size of Denmark, the Dominican Republic, or Guatemala and probably a bit of a jolt to the intuition of an American living in a nation of 250 million).

Between 1975 and 1990, the mean has actually decreased a little, that's a little unsettling. Is this some weird result of using a weird unit of measure, the logs? No, but let's check: My intuition expects an increase of about two percent per annum, for world population, approximately 30% over the fifteen years between 1975 and 1990. (I got 30% for fifteen years by multiplying 2% per year, my personal expectation, times 15 years. That's a crude approximation because it ignores compound interest but then I pulled 2% out of my head anyway. For now, for the sake of a *quick* look and a quick think about the data, I can tolerate the crudeness of these approximations.)

But the mean population (in logs) hasn't increased at all and the median population (in logs) has increased by .0465. What is .0465? That is the difference between the logs so it corresponds to a ratio of populations, $10^{.0465} = 1.11$. So it says that the 1990 figure is 1.11 times the 1975 figure, an increase of 11%, not 30%. So both figures for the 1975 to 1990 change are too low — compared to what I am expecting. Again, is this just a peculiar punishment for using peculiar numbers, the logs? No, it can't be because the median country is the same country, regardless of the choice between population and log population. So there is something real here.

Unless — maybe my figure of 2% per annum for world population is wrong. I don't know how that number got into my head. And since my intuition is not matching the facts, I'd better check. And, I have the data: Adding up the 1975 populations, I get about 4 billion people. Adding up the 1990 populations, I get 5.3 billion. That's an increase of about 33% in *total* population — right on target for my intuition. So I still have to explain the difference between the increase for *total* population and the increase for *average* population. Ah — this begins to sound like a problem of units: For total population the unit is the world. For average population the unit is the nation. So, I've got a clue but I've still got something to worry about as I continue.

Looking at the variations:

What have I got for the variation? First, what am I expecting? I'm expecting or, to be more precise, I am hoping that the variation is a nice reasonable number — unlike the variation that including a negative range of populations (for population without logs). Taking the measures of variation one at a time (without comparing them), this part looks good. Using the standard deviations, the central range of the distribution is the range between the mean minus one standard deviation and the mean plus one standard deviation — for 1990 that's between 2.777 and 4.599. Unlike the measures of variation on population, which was comparatively poorly behaved, these numbers match the picture, including a range from about one and a half stems less than the stem with the largest number of leaves to about one and one half stems greater than the stem with the largest number of leaves. Bringing my intuition along, what is .911? It is the log of 8.15, ($10^{.911} = 8.15$). That tells me that the central range of the distribution lies between the center divided by 8.15 and the center multiplied by 8.15.

To make that a little easier, in plain English, I can introduce the term “geometric mean”: The geometric mean is the anti-log of the mean of the logs, i.e., the anti log of 3.688 in this case. So, I can build this information into a sentence by saying that “the central range of popula-

tions lies within a factor of 8 on either side of the geometric mean of the populations.”

The quartile spreads, used as alternative way to specify the central range of the distribution, pick out a narrower definition of the center, between 3.222 and 4.278. I would put that in words by using anti-logs and saying “the middle fifty percent of the distribution marks a range between 1.7 million and 19 million around a median of 6.4 million people.

What have I got for the comparison between variations? First, again, what am I expecting? I’m expecting, or hoping, that the variation is constant, more evidence of a well-behaved variable. Checking the facts, no. Both the standard deviations and the quartile spreads increase, 1975 to 1990. The standard deviation increases by .125. The quartile spread increases by .130. Again, how large are these logarithms? Checking, .125 is the log of 1.33, ($10^{.125} = 1.33$). .130 is the log of 1.35. Those ratios seem like large increases, thirty-three to thirty-five percent. I don’t like that — not well-behaved. It may be that these are the facts and there is nothing to do but report them. But that’s just deferring the thinking — somebody, probably me, is going to have to figure out what’s going on here. One strong cue I have is my knowledge that the computation of quartiles doesn’t even use the data beyond the quartiles, while the computation for the standard deviation, uses all of the data. That suggests I keep my eye on the small end or the large end of the distribution of nations. I’ll put this on the stack of questions I have to worry about as I continue.

Looking at the Limits:

Checking the upper and lower ends of the 1990 distribution: The difference between results based on the given unit of measure, population, as compared to the well-behaved unit of measure, log population, will differ most at the low end: Where the mean minus one standard deviation extended to negative populations, using people; the mean minus one standard deviation is still a credible number, using logs.

Doing the computations: Using two criteria, first using the criterion that builds on the standard deviation. Using straight population, no logs, the mean minus three standard deviations and the mean plus three standard deviations set the bounds of reasonable data at $-305,141$, that's negative 305 million, and $+364,555$, that's plus 365 million. Aside from the fact that there is no country with a negative population, whatever that means, there are three countries, the United States, India, and China, with unusually large populations. But, such measures don't make sense for a variable, population, that is not well-behaved.

Applying the same computations using logs as the unit of measure, three standard deviations sets the bounds of reasonable data at 1.855 and 5.521, corresponding to populations below 71.6 (below 72 thousand) and above 331,894.5 (above 339 million), marking one country, Seychelles, as atypically small and two countries, India and China, as atypically.

Using the criteria based on quartiles, suggested by Tukey, the inner fences are at 2.2215 and 5.3895, corresponding to 166.5 (167 thousand) and 245,188 (two hundred and forty five million). And the outer fences are at .6375 and 6.9735, corresponding to 4.34 (four thousand) and 9,408,058 (9.4 billion). Using the inner fences, they classify 15 countries as very small and three as very large. Using the outer fences, no countries are "beyond the fence".

Summing it up and thinking:

So now, what have I got: The mean has become smaller while the median has become larger. This too may have something to do with using only the middle values, for the median, while using all the values for the mean. And now I also remember that there was a substantial difference between the list of nations for 1975 and the list for 1990. Looking at that list again, in rank order, I see one very big country has disappeared, replaced by ?? smaller ones. So I bet that the anomalies I need to deal with, one average increasing while the other

decreases, variations that get larger when, for I was looking for a constant, that these anomalies can depend on something that occurred among the extreme values, suggesting that this is the numerical indicator of the breakup of the Soviet Union.

I'll check. Removing the Soviet Union from the 1975 data produces

	1975	1990	
Mean	3.713	3.683	Verify ??
Standard Deviation (Excel function stdevp)	.777	.936	Verify
Median	3.7575	3.8445	Verify
Low Quartile	3.3000	3.1665	Verify
High Quartile	4.225	4.289	Verify
Quartile Spread	.925	1.056	

So much for that explanation, it doesn't work. — I've removed the USSR from the 1975 data and Estonia, Latvia, Armenia, Tukemenistan, Lithuania, Moldova, Kurqstan??. Tajikistan, Georgia, Buelarus, Uzbekistan, and the Ukraine from the 1990 data — but the effects I hoped to explain have persisted.

I'd better be more extreme: I'll use only those countries presenting data for both time periods: At 144 countries, this is a restricted subset, but it gives me a subset whose change, 1975 to 1990, depend more (though not exclusively) on population growth than political change (though not exclusively) — excluded other nations for the purpose of checking whether the curious changes of values are attributable to fission and fusion.

	1975	1990	
Mean	3.693	3.839	Verify
Standard Deviation (Excel function stdevp)	.790	.777	Verify

Rules of Evidence	Description Using Logs		
Median	3.721	3.892	Verify
Low Quartile	3.2425	3.4075	Verify
High Quartile	4.1935	4.336	Verify
Quartile Spread	.9510	.9285	

All right — somewhere between the full data set, which is different for the two dates, and the restricted data set, which is more similar for the two dates (though less complete for either date) some of the directions of change have reversed. With full data the mean decreases, with restricted data, the mean increases. With full data the standard deviations increase; with restricted data the standard deviations decrease. With the full data as well as the restricted data the medians increase. With full data the quartile spread increases; with restricted data the quartile spread decreases.

I'm now a little more at ease about the logarithm. It was supposed to give me a well-behaved variable. And it did gain symmetry and reasonable spreads. But it bothered me when my two measures of spread, the standard deviations and the quartile spreads of the logs increased. Now they decrease. That tells me I'm close: The size of the changes in the variation are within a range that can be influenced by changes in the set of countries. So the variations do not necessarily indicate that the measure is poorly-behaved. Phrasing that with a double negative: I have no clear evidence that log population is not homoscedastic. (Try these tests with the original numbers — in people as the unit the variable. The heteroscedasticity there **are/ is /should be [check]** much larger than might be explained by heteroscedastic with or without the adjustments in the set of nations.)

The Write Up

And now, wanting you to know how hard I've worked at this, but admitting, that you really don't care — except for the results: The write up:

The Size of Nations

What is the size of a country? United Nations data for 1990 show that the median population of countries is about 5 million, matching the population of Somalia and Haiti. Those of us who live in the United States naturally think of ourselves as “normal” and countries such as Somalia and Haiti as quite small, but in fact the middle fifty percent of national populations runs between 1.7 million, e.g., Estonia, and 19 million, Iraq. At the extremes, China and India stand out, including between themselves, 2 billion people, about half of the world’s population while, at the other end a few nations count populations of approximately 100,000 or less.

Curiously, even while the population of the world has increased by about 35%, approximately 2% per year, the population of the average nations shows no clear trend. Even the direction of change, up or down, depending on the precise measure that is used to compute the average. (The size of the median population has increased from 5.7 million to 6.4 million, a 12% increase.) In effect, as the total population has increased, the division of people into states has increased the fragmentation so that the 1990 average is close to the 1975 average and smaller than the 1975 average as a fraction of the total population of the world. This fragmentation appears to be the net result of changes in definitions and borders, including the breakup of the Soviet Union and Yugoslavia, the fusion of the Germanys, Vietnams, and Adens.

Who, What, Where ...

I’ve chosen the median because I can identify it with a country. That makes it easy to communicate. It also avoids the need to discuss the units because the median is the median (the middle is the middle) whether I use the logs or the original numbers.

I’m using bench marks to keep my reader oriented, using the size of the U.S. the names of well known countries, and rounded numbers.

These were convenient cutting points, using either the ratio of the size of a country to the size of the next smallest country, 3.4 for India as compared to China, 1.3 for China as compared to India, or easily remembered values, like 100,000.

I’ve chosen to make a virtue out of ambiguity with respect to the direction of change, acknowledging different results from different indicators.

I’m giving up on the distinction between nation and state — important professionally but not necessarily to my audience (depending on the audience)/