

Mean-Based Budgeting

My grade school textbooks, in Chicago, told me that the model of democracy could be found in the New England town meeting. That was where neighbors came together, argued their visions of the good the true and the beautiful in fair debate, and wrote the laws they would live by. Displacing heaven to New England saved my grade school teachers from potentially interesting but much less theoretical discussions of the experience we lived with in Chicago.

My adult experience in New Hampshire brings me fact to face with the New England town meeting. And I can say this much for it: If the stuff that teachers feed nine year olds had turned out to be true, then civic life in New England would have been rather dull, which it is not.

The key item on the agenda each year is the budget. There it is: A town of 3,979 people; last year's appropriations, \$3,301,133; proposed budget \$3,077,903 (not including the school budget). "What is your pleasure on the budget?"

So what do *you* think? Is \$3,077,903 a good budget? That's a little hard to answer, isn't it. Well the budget comes to the voters

in a booklet disclosing some of the detail. And long before it is brought to the voters, "What is your pleasure on the budget?", it goes through a budget committee, and through Selectmen, and through public hearings.

Would detail help you evaluate the budget?
O.K.:

Welfare: Direct Assistance

Actual appropriations prior year:
\$44,000

Actual expenditures prior year:
\$22,738

Selectmen's Recommended Budget:
\$29,000

Budget Committee Recommendation:
\$31,500

That doesn't really help much. And I don't recall that the wispy picture of democracy drawn in my Chicago text book offered much guidance for creating budgets, spending real money on welfare, roads, salaries, and so forth.

There is no formula that will translate the grand democratic vision into a budget—ultimately it is a matter of vision and

interest, what is needed, who says so, what is right, what is legitimate, who is responsible, what can we afford. More than numbers, a budget debate includes demands from department heads, from affected citizens, from community groups, businesses, and other governments — all of which represent themselves with varying degrees of effectiveness and persuasion.

But the discussion can be given context (followed by argument about which context is relevant). Most easily: What did we spend last year? The context does not establish an ethical calculus that will translate beliefs into numbers. But it helps: Just as the shape of a statistical distribution indicates what is average (without claiming that it is right), the size of last year's budget indicates where the discussion will begin. Just as the extremes of a statistical distribution indicate what is atypical, unusual numbers in a budget will focus the debate.

The advantage of last year's numbers as a context for this year's debate is that the numbers are available. The trouble with last year's numbers as a context is that they are insensitive to deviations that grow by degree: Small deviations accumulate into large excursions, imperceptibly, year by year: One year a department merits an increase 2 or 3 percent in excess of other departments. Next year the department needs a new piece of equipment. The next year the equipment requires maintenance. After that there should be a capital reserve. Next year there

is an unrepeatable emergency. After that there is a building on the market at a price which, in the long run, saves money. Each change is reasonable and not excessively different from the year before. And then, year by year, insignificant deviations accumulate into significant distortions — which are undetectable in the comparison between one year and the last. It is like adding a few grains of sugar to a cup of coffee, a few grains at a time. The change is undetectable. But eventually the coffee will surely become too sweet.

A better context can protect against these local excursions into madness by asking not, "What did we do last year?", but asking "What does everyone else do?", where "everyone else" means the voters in other towns in the state. Other towns in the state operate with the same state laws, (governing, for example, whether or not the "town" budget includes the school budget), with the same weather attacking the roads, and with similar economies. In this context the mean provides a base line. And the shape of the distribution and the extremes provide a reality check free of local personalities, free of local credit or blame: "Three quarters of the towns in this state are spending half as much per person on welfare as we are. Why is that?" It gives focus to the discussion.

Until recently mean based budgeting was a good plan that was impossible to implement: You knew the next town. You could look at

their books. But budgets were on paper. Budgets followed different accounting categories. It was impractical to develop the empirical base that would support these local decisions. It was impractical to begin with the simple statement

“On the average, towns spend x-dollars per capita on welfare [using the median]. And fifty percent of the towns voted appropriations between a-dollars per capita and b-dollars per capita [using the two quartiles].”

followed up by the simple question

“Our town spends 20% more (or 20% less) than the ‘normal’ range? Why?”

Now, it is practical. For better or for worse, towns have accommodated themselves to computers, states have archived these computerized accounts and imposed common accounting categories. Archives are open to public access.

The statistics are straight forward: If the object is to place the total budget in context, then “regress” total budget as a function of total population. Fit a line to the data, compute the residuals, and that’s it: X-burg is ___ above or ___ below what you would expect for a town of its size. If the objective is to place the welfare budget in context, or the road budget, or the budget for the town office in context: regress the objective on population size, and compare X-burg to the statistical norms for a town of its size.

The statistics are straightforward. But first, there is a question. Is this comparison valid? Is there any reason to believe that budgets either are (or should be) proportional to population? Is there evidence that this statistical criterion matches the problem to which it is being applied?

Folklore and common sense are richly contradictory on this question:

For example, everyone knows that large cities have problems of crime not encountered by small towns. Therefore the per capita cost of public safety will be (or should be) higher in large cities. Everyone knows that you can not compare the budgets of small towns to the budgets or larger cities, even on a per capita basis: Cities cost more.

Common sense, with its rich supply of contradictory advice also tells us that it is easier (more efficient) to large groups than small ones. A town like mine can, on the average, expect about .05 fires at any hour of the day or night, but it still has to have one full fire crew at the ready when the call comes in. A large city can match its capacity much more closely to the demand, lowering the cost per capita. On a per capita basis small towns cost more.

So common sense, as usual, is eloquently useless, able to support any proposition, or deny it, or both confirm it and deny it at the same time. Perhaps the facts can do better: Is it valid to compare costs from town to town, on a

per capita basis. Empirically, is there a relation between cost and population? What is that relation? Does it provide a usable base for comparing budgets, town to town?

The Relation Between Population and Total Budget

Beginning with the variable that sums up the rest, beginning with the “bottom line”, here are the population and budget numbers for all towns of the state of New Hampshire in 1994.

TOWNNAME	Popula- tion	Log Popu- lation (Base 10)	1994 Final Appropri- ation (Excludin g School Budget)	Log 1994 Appropri- ation (Base 10)					
HARTS LOCATION	36	1.56	16,000	4.20	SOUTH HAMPTON	740	2.87	324,468	5.51
ELLSWORTH	74	1.87	30,808	4.49	GOSHEN	742	2.87	334,423	5.52
WINDSOR	107	2.03	58,917	4.77	GILSUM	745	2.87	266,667	5.43
WATERVILLE VALLEY	151	2.18	1,740,814	6.24	MONROE	746	2.87	456,612	5.66
EASTON	223	2.35	90,963	4.96	ACWORTH	776	2.89	387,067	5.59
CLARKSVILLE	232	2.37	145,950	5.16	BATH	784	2.89	450,261	5.65
ORANGE	237	2.37	143,221	5.16	SPRINGFIELD	788	2.90	777,247	5.89
ROXBURY	248	2.39	76,421	4.88	BRIDGEWATER	796	2.90	488,662	5.69
CHATHAM	268	2.43	84,028	4.92	FRANCONIA	811	2.91	742,157	5.87
ERROL	292	2.47	211,349	5.33	HILL	814	2.91	421,541	5.62
SHARON	299	2.48	138,500	5.14	WARREN	820	2.91	329,435	5.52
GROTON	318	2.50	241,847	5.38	DALTON	827	2.92	594,631	5.77
DUMMER	327	2.51	184,018	5.26	NEW CASTLE	840	2.92	1,067,544	6.03
BENTON	330	2.52	56,121	4.75	RICHMOND	877	2.94	327,495	5.52
LANDAFF	350	2.54	181,648	5.26	DANBURY	881	2.94	413,890	5.62
EATON	362	2.56	299,695	5.48	NEWFIELDS	888	2.95	741,162	5.87
RANDOLPH	371	2.57	221,855	5.35	PITTSBURG	901	2.95	678,073	5.83
HEBRON	386	2.59	267,954	5.43	GRAFTON	923	2.97	549,815	5.74
LYMAN	388	2.59	345,987	5.54	STRATFORD	927	2.97	1,443,059	6.16
DORCHESTER	392	2.59	178,009	5.25	FREEDOM	935	2.97	1,067,095	6.03
SHELBURNE	437	2.64	302,111	5.48	WILMOT	935	2.97	513,349	5.71
SUGAR HILL	464	2.67	520,871	5.72	EFFINGHAM	941	2.97	664,421	5.82
BROOKFIELD	518	2.71	276,403	5.44	LEMPSTER	947	2.98	504,417	5.70
STARK	518	2.71	274,850	5.44	JEFFERSON	965	2.98	375,173	5.57
CARROLL	528	2.72	569,912	5.76	HARRISVILLE	981	2.99	546,486	5.74
NELSON	535	2.73	302,008	5.48	NEWINGTON	990	3.00	2,865,784	6.46
ALBANY	536	2.73	447,992	5.65	CENTER HARBOR	996	3.00	828,850	5.92
LANGDON	580	2.76	286,900	5.46	ORFORD	1,008	3.00	731,668	5.86
STODDARD	622	2.79	395,507	5.60	STEWARTSTOW N	1,048	3.02	429,166	5.63
PIERMONT	624	2.80	287,212	5.46	SALISBURY	1,061	3.03	515,522	5.71
CROYDON	627	2.80	292,970	5.47	SANDWICH	1,066	3.03	1,451,472	6.16
WASHINGTON	628	2.80	865,975	5.94	WOODSTOCK	1,167	3.07	1,564,914	6.19
WENTWORTH	630	2.80	529,955	5.72	MIDDLETON	1,183	3.07	521,861	5.72
MARLOW	650	2.81	345,725	5.54	ALEXANDRIA	1,190	3.08	717,376	5.86
COLUMBIA	661	2.82	202,933	5.31	TEMPLE	1,194	3.08	606,798	5.78
SURRY	667	2.82	233,351	5.37	MASON	1,212	3.08	675,578	5.83
JACKSON	678	2.83	976,718	5.99	FRANCESTOWN	1,217	3.09	960,260	5.98
SULLIVAN	706	2.85	256,805	5.41	LINCOLN	1,229	3.09	2,960,520	6.47
					BENNINGTON	1,236	3.09	864,319	5.94
					GRANTHAM	1,247	3.10	872,350	5.94
					LYNDEBOROUGH	1,294	3.11	694,035	5.84
					MILAN	1,295	3.11	435,206	5.64
					UNITY	1,341	3.13	575,507	5.76
					NEWBURY	1,347	3.13	1,216,672	6.09
					EAST KINGSTON	1,352	3.13	641,087	5.81

MADBURY	1,404	3.15	670,873	5.83	SUNAPEE	2,559	3.41	3,753,511	56.57
BRADFORD	1,405	3.15	1,095,186	6.04	FREMONT	2,576	3.41	768,185	5.89
WEBSTER	1,405	3.15	696,631	5.84	BRENTWOOD	2,590	3.41	771,805	5.89
RUMNEY	1,446	3.16	533,389	5.73	GILMANTON	2,609	3.42	1,493,586	6.17
SUTTON	1,457	3.16	1,030,573	6.01	ROLLINSFORD	2,645	3.42	835,310	5.92
DUBLIN	1,474	3.17	938,545	5.97	CHESTER	2,691	3.43	1,013,028	6.01
LYME	1,496	3.17	1,209,899	6.08	GREENLAND	2,768	3.44	1,204,935	6.08
HAMPTON FALLS	1,503	3.18	912,400	5.96	NOTTINGHAM	2,939	3.47	1,091,841	6.04
THORNTON	1,505	3.18	1,371,051	6.14	MOULTON-BOROUGH	2,956	3.47	3,506,437	6.54
GREENFIELD	1,519	3.18	672,559	5.83	STRAFFORD	2,965	3.47	893,047	5.95
WESTMORELAND	1,596	3.20	410,500	5.61	CANAAN	3,045	3.48	1,839,983	6.26
HANCOCK	1,604	3.21	1,076,495	6.03	WAKEFIELD	3,057	3.49	1,805,051	6.26
NEW HAMPTON	1,606	3.21	1,264,561	6.10	BARNSTEAD	3,100	3.49	1,608,588	6.21
KENSINGTON	1,631	3.21	581,154	5.76	CHESTERFIELD	3,112	3.49	2,286,910	6.36
CORNISH	1,659	3.22	895,466	5.95	WILTON	3,122	3.49	2,033,603	6.31
LISBON	1,664	3.22	1,221,587	6.09	DEERFIELD	3,124	3.49	1,937,144	6.29
CANTERBURY	1,687	3.23	615,537	5.79	NORTHWOOD	3,124	3.49	1,418,169	6.15
HOLDERNESS	1,694	3.23	1,498,571	6.18	GORHAM	3,173	3.50	3,827,227	6.58
MADISON	1,704	3.23	1,198,574	6.08	NEW LONDON	3,180	3.50	3,462,182	6.54
DEERING	1,707	3.23	862,174	5.94	WALPOLE	3,210	3.51	2,425,109	6.38
ALLESTEAD	1,721	3.24	765,745	5.88	NEW BOSTON	3,214	3.51	1,932,824	6.29
DUNBARTON	1,759	3.25	664,876	5.82	TILTON	3,240	3.51	1,719,738	6.24
MONT VERNON	1,812	3.26	863,697	5.94	ALTON	3,286	3.52	2,958,199	6.47
TUFTONBORO	1,842	3.27	983,087	5.99	OSSIPEE	3,309	3.52	4,037,042	6.61
ANDOVER	1,883	3.27	562,335	5.75	NEWTON	3,473	3.54	1,414,819	6.15
WHITEFIELD	1,909	3.28	2,011,131	6.30	LANCASTER	3,522	3.55	5,302,740	6.72
ASHLAND	1,915	3.28	5,991,549	6.78	CANDIA	3,557	3.55	1,215,657	6.08
MARLBOROUGH	1,927	3.28	1,484,826	6.17	BOSCAWEN	3,586	3.55	1,440,818	6.16
CHICHESTER	1,942	3.29	696,019	5.84	EPSOM	3,591	3.56	1,129,369	6.05
NEW DURHAM	1,974	3.30	1,709,885	6.23	NORTH HAMPTON	3,637	3.56	2,621,011	6.42
FITZWILLIAM	2,011	3.30	1,105,147	6.04	MILTON	3,691	3.57	1,640,989	6.22
BETHLEHEM	2,033	3.31	1,695,314	6.23	PITTSFIELD	3,701	3.57	2,221,767	6.35
PLAINFIELD	2,056	3.31	1,252,835	6.10	LEE	3,729	3.57	1,623,753	6.21
TROY	2,097	3.32	1,174,027	6.07	HINSDALE	3,936	3.60	1,797,216	6.25
SANBORNTON	2,136	3.33	1,485,044	6.17	ENFIELD	3,979	3.60	3,301,133	6.52
TAMWORTH	2,165	3.34	1,159,218	6.06	NEW IPSWICH	4,014	3.60	1,431,500	6.16
GREENVILLE	2,231	3.35	1,303,897	6.12	WINCHESTER	4,038	3.61	2,301,339	6.36
WARNER	2,250	3.35	1,513,819	6.18	SANDOWN	4,060	3.61	1,504,893	6.18
BARTLETT	2,290	3.36	1,143,921	6.06	AUBURN	4,085	3.61	1,373,330	6.14
ANTRIM	2,360	3.37	1,688,818	6.23	LOUDON	4,114	3.61	1,752,196	6.24
CAMPTON	2,377	3.38	1,040,835	6.02	HENNIKER	4,151	3.62	2,616,758	6.42
BROOKLINE	2,410	3.38	1,113,194	6.05	HAVERHILL	4,164	3.62	1,252,612	6.10
COLEBROOK	2,444	3.39	1,357,725	6.13	NORTHFIELD	4,263	3.63	1,809,819	6.26
NORTH-UMBERLAND	2,492	3.40	1,371,920	6.14	HILLSBOROUGH	4,498	3.65	6,107,675	6.79
DANVILLE	2,534	3.40	858,789	5.93	RYE	4,612	3.66	4,159,532	6.62
BRISTOL	2,537	3.40	2,498,282	6.40					

CHARLESTOWN	4,630	3.67	2,602,431	6.42
ALLENSTOWN	4,649	3.67	2,352,666	6.37
HOPKINTON	4,806	3.68	2,856,304	6.46
WOLFEBORO	4,807	3.68	11,206,835	7.05
MEREDITH	4,837	3.68	6,951,966	6.84
RINDGE	4,941	3.69	1,841,645	6.27
STRATHAM	4,955	3.70	2,125,127	6.33
EPPING	5,162	3.71	2,423,933	6.38
ATKINSON	5,188	3.71	2,125,331	6.33
PETERBOROUGH	5,239	3.72	4,995,961	6.70
JAFFREY	5,361	3.73	6,983,140	6.84
BOW	5,500	3.74	4,110,238	6.61
LITCHFIELD	5,516	3.74	1,725,133	6.24
KINGSTON	5,591	3.75	2,254,780	6.35
HOLLIS	5,705	3.76	3,674,216	6.57
FARMINGTON	5,739	3.76	2,610,103	6.42
BELMONT	5,796	3.76	3,352,536	6.53
PLYMOUTH	5,811	3.76	4,317,324	6.64
LITTLETON	5,827	3.77	4,210,515	6.62
GILFORD	5,867	3.77	5,824,931	6.77
NEWPORT	6,110	3.79	5,242,011	6.72
BARRINGTON	6,164	3.79	2,023,182	6.31
WEARE	6,193	3.79	2,865,631	6.46
SWANZEY	6,236	3.79	2,276,072	6.36
SEABROOK	6,503	3.81	12,510,753	7.10
PEMBROKE	6,561	3.82	6,894,823	6.84
HAMPSTEAD	6,732	3.83	2,466,294	6.39
NEWMARKET	7,157	3.85	4,467,355	6.65
PLAISTOW	7,316	3.86	3,451,783	6.54
CONWAY	7,940	3.90	5,988,938	6.78
FRANKLIN	8,304	3.92	6,792,378	6.83
RAYMOND	8,713	3.94	3,804,497	6.58
HOOKSETT	8,767	3.94	7,225,433	6.86
WINDHAM	9,000	3.95	5,286,475	6.72
AMHERST	9,068	3.96	4,672,970	6.67
HANOVER	9,212	3.96	8,928,686	6.95
PELHAM	9,408	3.97	5,087,464	6.71
SOMERSWORTH	11,249	4.05	7,619,638	6.88
MILFORD	11,795	4.07	8,529,696	6.93
DURHAM	11,818	4.07	7,989,037	6.90
BERLIN	11,824	4.07	9,251,705	6.97
LEBANON	12,183	4.09	21,029,363	7.32
HAMPTON	12,273	4.09	13,434,218	7.13
EXETER	12,481	4.10	12,223,190	7.09
BEDFORD	12,563	4.10	9,122,876	6.96
CLAREMONT	13,902	4.14	10,445,650	7.02
GOFFSTOWN	14,621	4.16	8,220,285	6.91
LACONIA	15,743	4.20	11,776,111	7.07
HUDSON	19,530	4.29	13,193,546	7.12
LONDONDERRY	19,781	4.30	13,812,112	7.14
MERRIMACK	22,156	4.35	15,351,214	7.19
KEENE	22,430	4.35	22,915,928	7.36
DOVER	25,042	4.40	20,808,160	7.32
SALEM	25,746	4.41	29,189,769	7.47
PORTSMOUTH	25,925	4.41	20,979,565	7.32
ROCHESTER	26,630	4.43	19,804,052	7.30
DERRY	29,603	4.47	19,299,092	7.29
CONCORD	36,006	4.56	37,787,640	7.58
NASHUA	79,662	4.90	45,170,090	7.65
MANCHESTER	99,567	5.00	87,173,729	7.94
Low Quartile	937	2.97		
2nd Quartile (median)	2,117	3.33		
High Quartile	4,644	3.67		
Median (checking)	2,097			
Mean	4,740			
Sum/234 (Checking)	4,740			

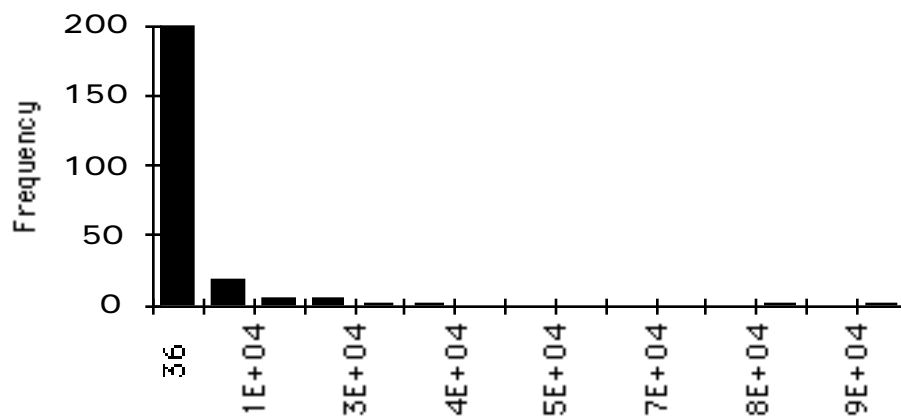
And where does analysis of the relation between population and budget (or the relation between any two variables) begin? With well behaved variables — with an examination of the units of measure,

translating if necessary to the forms of well behaved variables.

Well-Behaved Population

First population: Because my spread sheet program is good at putting things in rank order , with names attached, and because it is poor at stem and leaf (and because I value my own time) I will forego the Stem and Leaf and rely on the rank order for equivalent detail. And because my spread sheet program is good at bad histograms and more cumbersome for good ones, I will settle for a less than friendly histogram — it will suffice to give me an overview of the shape of the distribution. Using the population data, as given (with the person as the unit of measure), here, is the histogram. It is not well behaved; on the contrary, it is extremely skewed with a tail extending in the direction of the larger values.

Histogram

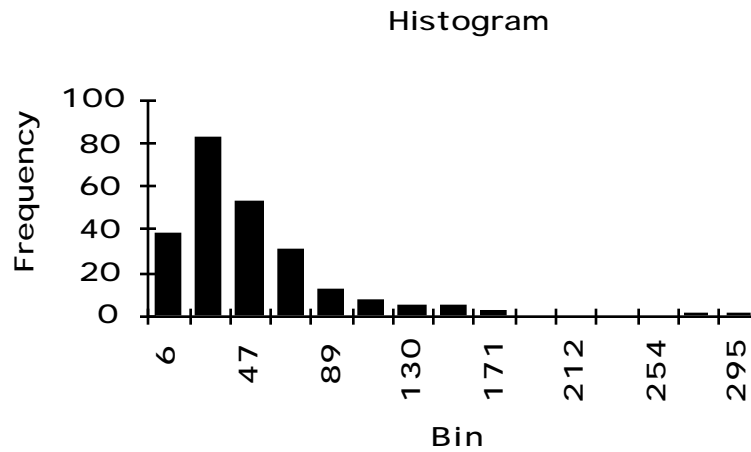


Putting numbers (on what is already obvious from the picture), the mid values give numerical expression to the skew in this picture: The mid-quartile is larger than the median. The mid-eighth is larger than the mid quartile, and so forth.

Count	Population	Population	Mid Value	Examples		
n=234	117.5	2,116.5	2,116.5	2,117	Median	Troy; Sanbornton
59	935	4,649	2,792	Mid Quartile	Freedom	Allenstown
30	624	8,304	4,464	Mid Eighth	Piermont	Franklin
15.5	356.0	13,232.5	6,794	Mid Sixteenth	Landaff; Easton	Bedford; Claremont
8	248	25,042	12,645	Mid Thirty-Second	Roxbury	Dover
4.5	187	28,103	14,145		Waterville Valley; East	Rochester; Derry
2.5	91	57,834	28,962		Ellsworth; Windsor	Concord; Nashua
1	36	99,567	49,802		Harts Location	Manchester

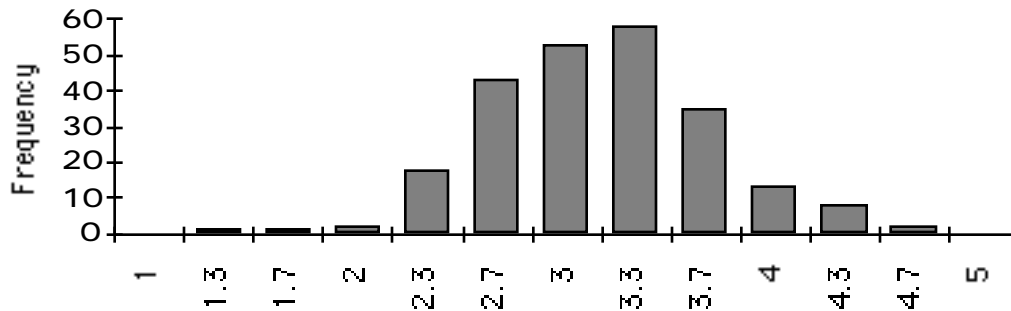
Manchester at 99 thousand people is, by itself, equal to the cumulative populations of the 107 smallest towns. That makes it large, but not necessarily different.

Pursuing the well-behaved form, changing the unit of measure from the person to the square roots improves the symmetry, but not enough.



Pursuing the well behaved form by changing to the log-arithm, it is clear (to the eye) that logs are close

Histogram of Town Populations (Using Logarithms, Base 10)



Putting numbers on the image, the mid value numbers support the visual appearance. For comparison, the table below shows 3 transformations, including the logs as well as two power transformations, one a little weaker than the logarithm, one a little stronger. Mid values based on the weaker transformation, the .1 power, still show a slightly increasing trend of values,

still indicating a tail in the direction of the larger population values. Mid values based on the logarithmic transformation wander — as they will when data are symmetrical. Mid values based on the stronger transformation, the $-.1$ power, also wander, like the mid values for the logarithm. So the negative $.1$ power is also close. That makes it user's choice: I'll use the logarithm as the well behaved unit of measure for these populations.

power				
0.1	2.15	2.15	2.15	Median
	1.98	2.33	2.15	Mid Quartile
	1.90	2.47	2.18	Mid Eighth
	1.86	2.58	2.22	Mid Sixteenth
	1.74	2.75	2.24	Mid Thirty-Second
	1.68	2.78	2.23	
	1.57	2.97	2.27	
	1.43	3.16	2.30	

log				
	3.33	3.33	3.33	Median
	2.97	3.67	3.32	Mid Quartile
	2.80	3.92	3.36	Mid Eighth
	2.79	4.12	3.46	Mid Sixteenth
	2.39	4.40	3.40	Mid Thirty-Second
	2.26	4.45	3.36	
	1.95	4.73	3.34	
	1.56	5.00	3.28	

power				
-0.1	0.46	0.46	0.46	Median
	0.50	0.43	0.47	Mid Quartile
	0.53	0.41	0.47	Mid Eighth
	0.54	0.39	0.46	Mid Sixteenth
	0.58	0.36	0.47	Mid Thirty-Second

	0.59	0.36	0.48
	0.64	0.34	0.49
	0.70	0.32	0.51

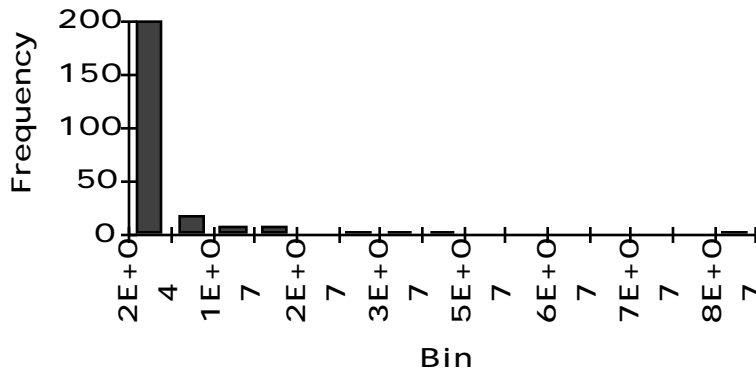
Using logs, and computing the fences, Nashua and Manchester exceed the inner fences on the high end. Nothing is so large that it exceeds the outer fences. On the low end, as on the upper end, two towns are below the inner fences. None are below the outer fences: Using logarithms for the unit of measure, the variable is well-behaved, with a note of caution at each end.

	0.697		Quartile Spread
	1.045		Step Size
In Population	In Logs	In Logs	In Population
84	1.9264	712	51,544
8	0.8815	757	571,483
			inner fences
			Outer Fences

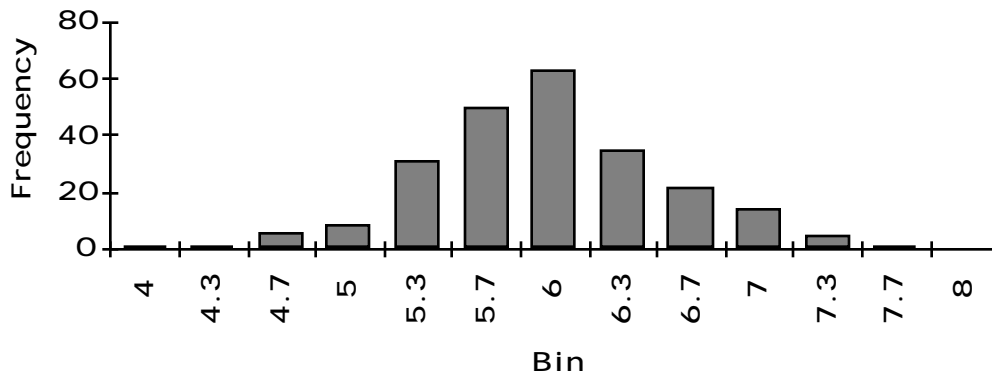
Well-Behaved Appropriations

That is one variable done, one more to go. For Appropriations: Using dollars as the unit of measure, the distribution is, like its mate, sharply skewed, with a few large values at the high end. Using logarithms as the unit of measure the behavior changes, close to symmetry. Attempting to verify this with the mid values, the result, using logs, is disconcerting. The numbers do not support what the eyeball has suspected. the distribution is not really symmetrical: It shows a consistent trend of mid values, 6.10, 6.11, 6.18, etc., indicating that even using logs as the unit of measure, the distribution is skewed with a tail extending in the direction of the higher appropriations.

Histogram



Histogram



Count					
n=234					

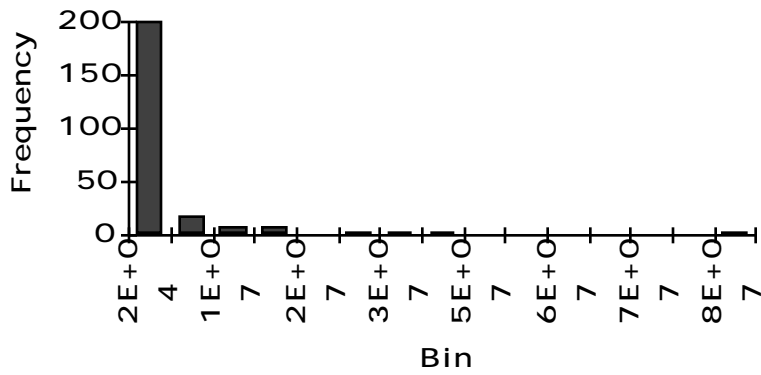
117.5	6.10	6.10	6.10	Median	Haverhill; Plainfield	
59	5.76	6.46	6.11	Mid Quartile	Unity	Newington
30	5.51	6.84	6.18	Mid Eighth	South Hampton	Pembroke
15.5	5.34	7.09	6.22	Mid Sixteenth	Errol; Randolph	Exeter; Seabrook
8	5.14	7.32	6.23	Mid Thirty-Second	Baron	Dover
4.5	4.82	7.41	6.12		Windsor; Roxbury	Kennebunk; Salem
2.5	4.62	7.80	6.21		Ellsworth; Benton	Nashua; Concord
1	4.20	7.94	6.07		Harts Location	Manchester

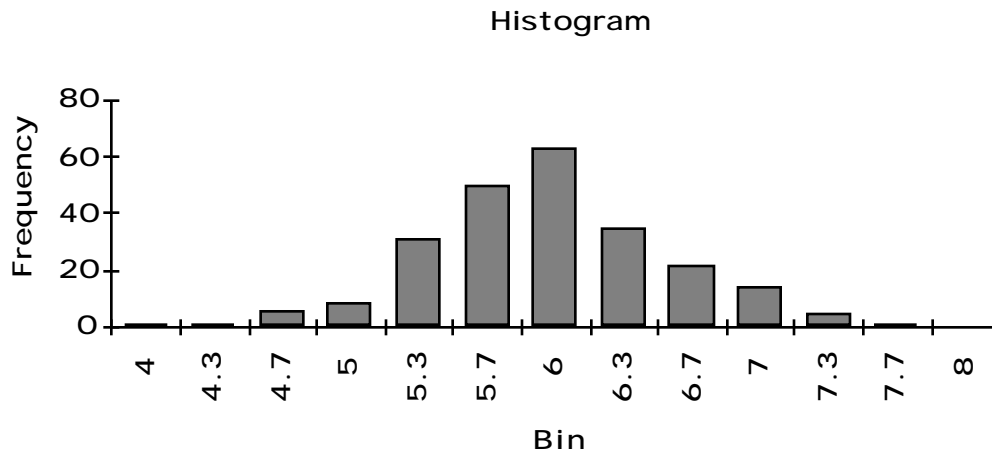
Can I accept that as symmetry? Is it close enough? I don't know. The only way to answer the question is to try a stronger transformation of the unit of measure and see how strong a transformation it takes to eliminate this trend among the mid values.

What it takes to break the trend is the $-.11$ power. So my choice is between the logarithm and the negative $-.11$ power. I'll take the log as close enough. (Had it required the -1 power to break the trend, or even the $-.5$ power, I would have worried. Fortunately I don't have to figure out what I would have done.) (Note that the $-.11$ power reverses the rank order of the numbers, high becomes low and low becomes high. So the "high values" at the mid thirty-second and later correspond to low values on the original scale of the variable. If this were a tail, it would be a tail toward the small values appropriations, implying that the transformation had been too strong.)

power					
	-0.11	0.213	0.213	0.213	Median
		0.232	0.195	0.214	Mid Quartile
		0.248	0.177	0.212	Mid Eighth
		0.259	0.166	0.212	Mid Sixteenth
		0.272	0.157	0.214	Mid Thirty-Second
		0.289	0.153	0.221	
		0.311	0.145	0.228	
		0.345	0.144	0.244	

Histogram





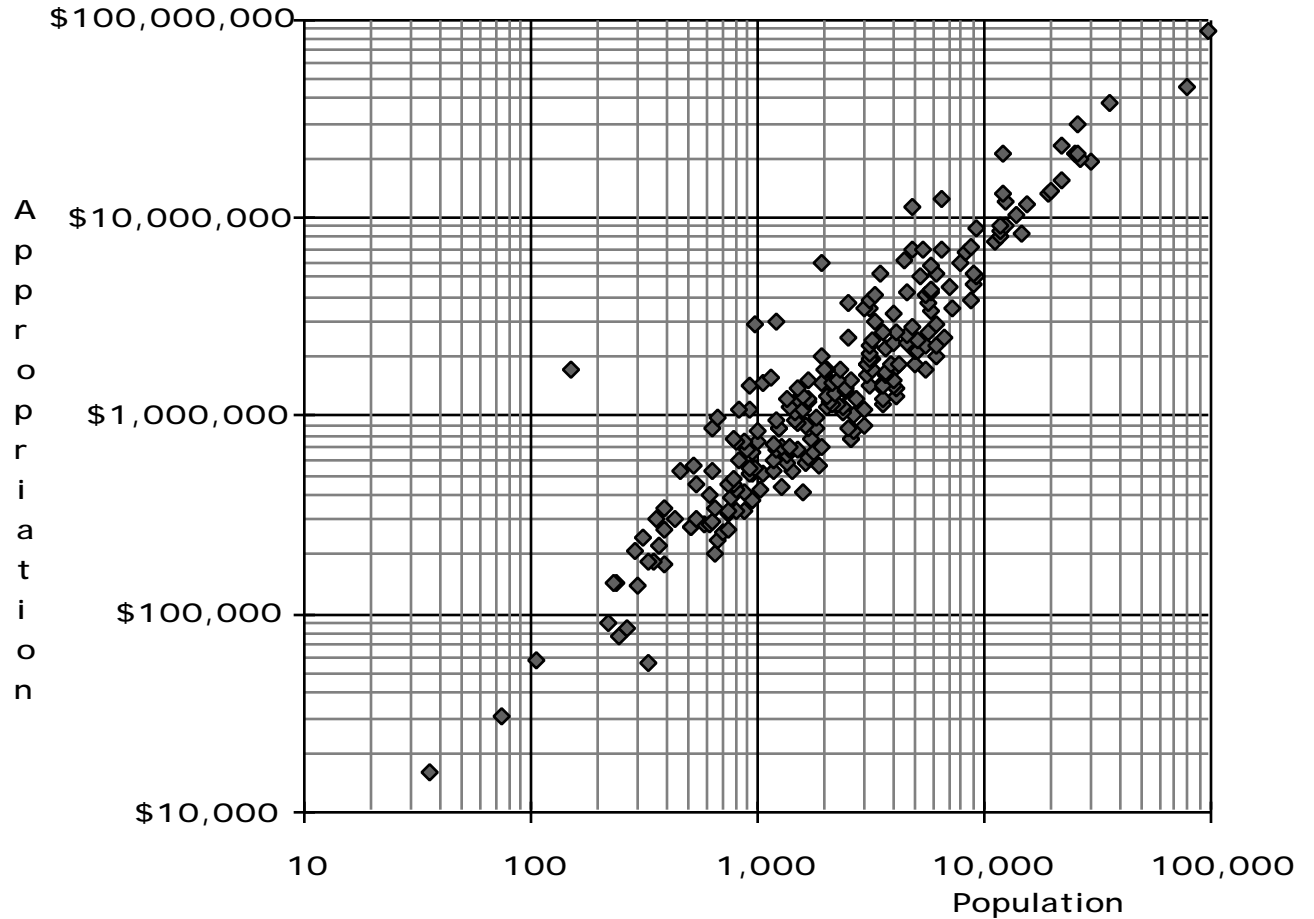
The Relation Between Population and Total Appropriation

Is there a relation between population and appropriation? That was the question I had to answer. If there is a relation? And if the relation is a relation of strict proportionality, double the population and you will double the appropriation (on the average) — then I have support for the procedures of mean-based budgeting: Strict proportionality, empirically, establishes reason to use appropriations per capita as a standard for the budget — regardless of the size of the town.

Searching each separate variable for its well-behaved form, tells me to look for the relation between these two variables by examining their log log form, using the log form of each variable.

Ordinarily, I would actually do the log log graph on “wallpaper”, making it about three feet wide. It takes a graph of this size to provide proper labels that literally spell out the name of each town on the graph. Then I would feast my eyes on the result, locating Waterville Valley, a ski town in the mountains, locating Manchester, an old industrial town, locating Concord, the state capitol, Hanover, a college town, ... getting a “feel for the data”.

But, limiting myself to a publishable piece of paper, I can at least inspect the shape of the graph, using dots.

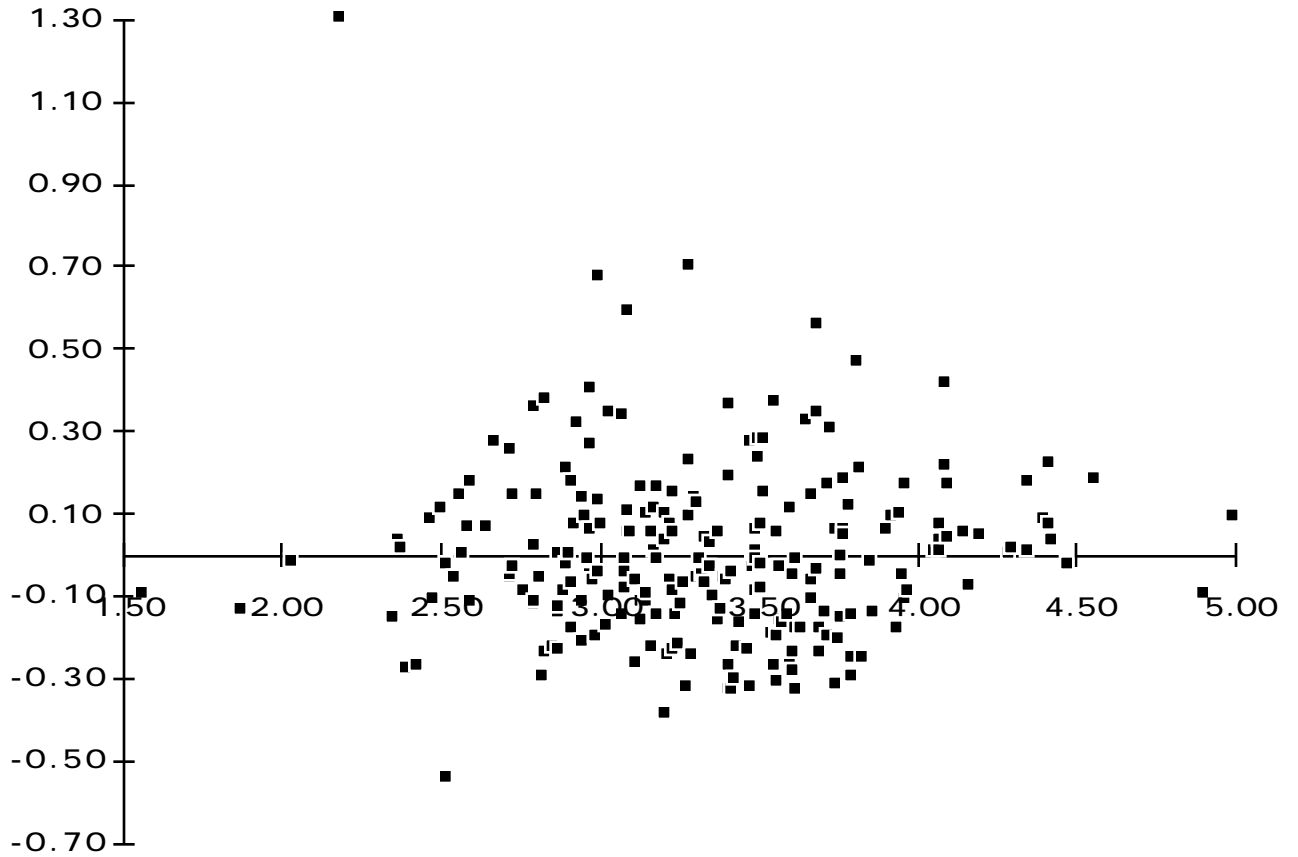


My first reaction to the graph is relief: So far, without getting serious (that is, without looking at the residuals) it appears that the relation is a line, not a curve. And the slope is tantalizingly close to 1, at a height showing that a population of 1,000 will have, on the average an appropriation of about \$600,000, about \$600 dollars per person. There is hope: This

could be a linear relation, in logs, and a strictly proportional relation in dollars and people.

Now I've got to get serious, using least squares regression to estimate a line at the center of the cloud of data and then looking at the residuals. So, allowing my spread sheet program to execute a least squares regression of log appropriations on log population, using common logs (base 10), I get an estimate of a line at the center of this cloud of data, a line with intercept 2.689 and slope of 1.032.

That is a little worrisome: I note that the anti log of 2.689, base 10, is 489, \$489 per person — a little low compared to my eyeball estimate of \$600. But then, checking, \$600 per person would have corresponded to an intercept of 2.778. The difference between 2.689 and 2.778 may be too fine for eyeball discrimination. So \$4889 may be acceptably close to my original estimate. Looking at the slope, the slope at 1.032 is also nice — passably close to a very simple number, passably close to the magic number 1 — at which I can establish strict proportionality. Using this intercept and this slope, as estimated by the computer, and plotting residuals (represented in logs), the residuals are



The plot of the residuals makes the extremes in appropriations per capita stand out. I see that exceptional point at the top of the graph and I have to look: That “dot” is Waterville Valley, site of a large ski development in the White Mountains.

		In logs	In Ratios
--	--	---------	-----------

Low Quartile	2.97	-0.14	0.72
Mean	3.33	0.00	1
High Quartile	3.67	0.10	1.25
		In logs	In Ratios
	Quartile Spread	0.24	1.73
	Step Size	0.36	2.28
	lower inner fence	-0.50	0.32
	Lower Outer Fence	-0.86	0.14

The mean residual is 0, as it must be in least squares regression. In logs, using the quartiles, fifty percent of the towns lie in a range between -.14 and +.10. In ratios those two numbers translate into a ratio that is less than one, .72, and a ratio that is greater than one, 1.25. In percentages, that means that fifty percent of the town budgets lie in a range from 28% below the average to 25% above the average of appropriations per capita for the entire state. Looking for outliers, one of these residuals is below the inner fence on the low end. But, as is obvious on the graph, there are several outliers exceeding the fences at the high end of appropriations per capita — five exceed the inner fence of appropriations per capita, one of the five exceeds the outer fence: That is Waterville Valley again, with 151 in its official population, and budget of \$1.7 million dollars — that is \$11,500 per capita, definitely an outlier.

Leaving this one point out, and re-estimating, the revised estimates are now intercept = 2.604 (\$402 per capita), slope = 1.056.

Interpreting the slope itself, in logs the relation is

$$\log(\text{Appropriation}) = 2.604 + 1.056(\log \text{population})$$

Taking anti-logs on both sides of the equation and restoring the original units, that is

$$\text{Appropriation} = (\$402 \text{ per person}) \times (\text{population})^{1.056}$$

Thinking about the Exponent: The Relation Between Population and Budget

Thinking: How would I *like* this relation to turn out? That's clear. I want the exponent to be one. That would establish that the relation between budget and population is independent of the size of the city. That would give me empirical support for the standard by which I can begin to sort the budgets of the state of New Hampshire and, in particular, the budget of my town.

Carl Sagan suggests that the essence of scientific method is skepticism. I am surely a skeptic — I don't even trust myself. That's why it is important to be up front about what I would like to find and, therefore, to be particularly careful and suspicious when, lo and behold I find, at the end of my analysis that I have "discovered", exactly what I was looking for at the beginning. Maybe, but I have to be careful.

So what I have so far is a proportionality between Appropriations and the 1.056th power of population, not the first power. Can I just lop off that .056, declare the power to be 1 (close enough). If I were to do that I would argue something sophisticated like

"On grounds of parsimony, I will simplify that 1.056 to 1, which establishes that the appropriations are directly proportional to the population."

But how do I know that .056 is small? Compared to what? Like analyzing budgets, data analysis is

often a “game” of establishing contexts. How do I know that .056 is small? Can I leave it out?

I will figure that out by trusting the mathematics and getting a feel for it. What 1.056 says, in contrast to 1.000, is that larger cities spend more, per capita (and on the average), than small towns. How much more? Following the math, I will figure out the numbers with and without that .056. Using strict proportionality (using 1.000), suppose that a town of one thousand people could be expected to appropriate \$400,000. Using strict proportionality, by comparison, a town of 10,000 people would be expected to appropriate ten times more, \$4,000,000, and a town of 100,000 (Manchester) would be expected to appropriate one hundred times more \$40,000,000.

Log Population	Population	Prediction in logs	Prediction in dollars	Ratio (to first row)	Ratio of the appropriation to the appropriation of a town of 1,000 people.
3	1,000	log appropriation = intercept + log population	\$appropriation = $10^{\text{intercept}}$ population		1
4	10,000	log appropriation = intercept + log population	\$appropriation = $10^{\text{intercept}}$ population	$\frac{10^{\text{intercept}} 10,000}{10^{\text{intercept}} 1,000} = 10$	10 to 1

5	100,000	log appropriation = intercept + log population	\$appropriation = $10^{\text{intercept}}$ population	$\frac{10^{\text{intercept}} 100,000}{10^{\text{intercept}} 1,000} = 100$	100 to 1
---	---------	--	---	---	----------

Now, by contrast how big is that 1.056? “.056” looks small but, actually the increase is not obviously so small that it can be ignored: Compared to a town of 1,000 people, a town that is 10 times larger would have an appropriation that is larger by the ratio of $(10^{1.056}) / (1^{1.056})$. That is 11.38 to 1. And it means that the relative budget is 14% larger than would be expected under strict proportionality. And Manchester, the extreme at with approximately 100,000 people would have an expected budget that is 29% larger than would be expected under strict proportionality.

Log Population	Population	Prediction in logs	Prediction in dollars	Ratio (to first row)	Ratio of the appropriation to the appropriation of a town of 1,000 people.
3	1,000	log appropriation = intercept + 1.056 log population	\$appropriation = $10^{\text{intercept}}$ population ^{1.056}		1

4	10,000	log appropriation = intercept + 1.056 log population	\$appropriation = $10^{\text{intercept}} \text{population}^{1.056}$	$\frac{10^{\text{intercept}} 10,000^{1.056}}{10^{\text{intercept}} 1,000^{1.056}} = 10^{1.056}$	11.38 to 1 $10^{1.056}$
5	100,000	log appropriation = intercept + 1.056 log population	\$appropriation = $10^{\text{intercept}} \text{population}^{1.056}$	$\frac{10^{\text{intercept}} 100,000^{1.056}}{10^{\text{intercept}} 1,000^{1.056}} = 100^{1.056}$	129.4 to 1 $100^{1.056}$

So, I can't casually throw it away: With or without that .056 tacked on to the 1 (in the exponent), I would not or would find a 29% larger budget (to be typical or large). It is not a large amount. It affects only one large city and (at this magnitude) it only affects the comparison between the largest city and the smaller towns, not the comparison to the average. But it is worth attention.

Now I'm going to tackle it another way, asking "How much do I believe these numbers anyway?" These numbers are the facts, not a sample of the facts, so variability is not an issue. But I note that just removing one data point raised the slope from about 1.04 to about 1.05. I don't really believe that number out to as many digits as I can calculate. Is the contrast between 1 and a slope of 1.05, on the log log graph, within the "wobble" or uncertainty that is built in to my data? The slope and the intercept estimated by my computer minimize the squared deviations. How much larger would the squared deviations become if I were to impose a slope of 1.000? How sensitive is the squared error (by which least squares regression evaluates the result) to the contrast between the simple 1 and the observed slope of 1.056?

Starting with the best, starting with 1.056, the equation “explains” 89.07% of the variance:

By regression, $r = .9438$, $r^2 = .8907$, 89.07 percent of the variance is “explained”.

Intercept = 2.604,
slope = 1.056.

By contrast, imposing the exponent 1.000, the equation “explains” 88.77% of the variance:

Intercept = 2.304 (anti-log of 2.304 = \$637),
slope 1.000

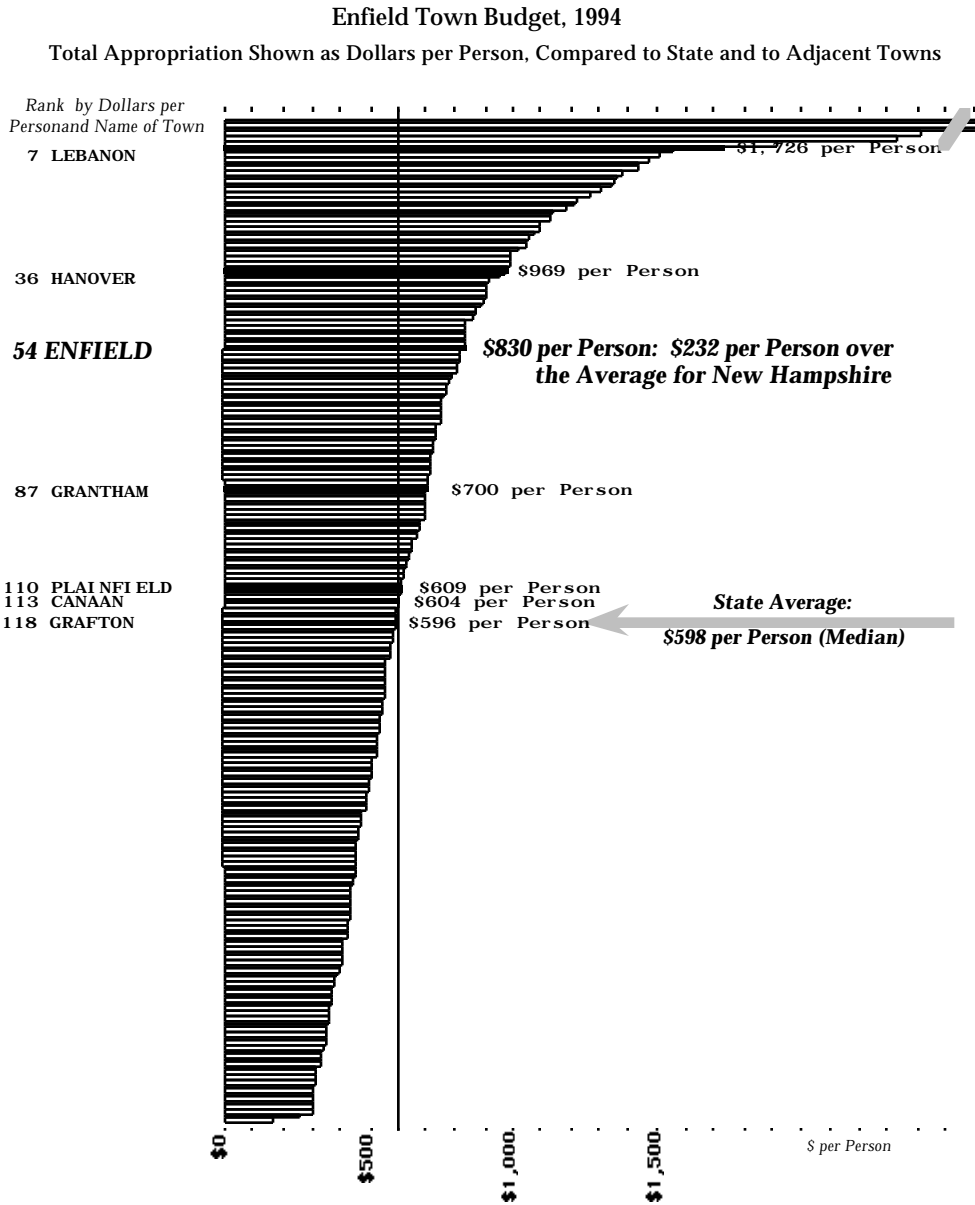
That almost convinces me that the difference can be ignored, 89.1% for the statistical best answer, 88.8% for the simple answer, 1. In fact, “1” may be the better answer. In the world of science I am free to and encouraged to look at details that may be a hint of some subtlety — but in my tentative conclusions I am obliged to choose what is simple unless there is compelling reason to do otherwise. I also note that imposing the slope 1 leads to a re-estimate of the constant so that it now corresponds to \$637. That is closer to what I saw in the graph and close to the median \$598 per person, which is less subject to error due to single cases like Waterville Valley or the other four high values of the residuals. In a sense by “helping” the statistics with their estimate of the slope, I may have been rewarded with a better estimate of the intercept.

I am not wholly happy, but I am willing to commit:

“On grounds of parsimony, I will simplify that 1.056 to 1, which establishes that the appropriations are directly proportional to the population — at a trivial loss of approximately 1% in the variance explained, and at a gain of considerable simplicity.

I will keep that 1.056 in mind and, in a professional publication I would be sure to alert the reader. But I would not use it unless subsequent research advanced the case for using the more complicated rather than the more precise answer in cases where both “explain” approximately the same amount of the variance.

And here is my first report:



The Enfield budget is \$3,301,133 for a 1990 population of 3,979. Reduced to the state average of \$598 per person, the Enfield budget would be to \$2,379,442. It would be \$921,691 below the present budget.
 Appropriation Data from State of New Hampshire Department of Revenue Administration
 Prepared by Joel H. Levine, RR1 Box 116, Enfield, New Hampshire 03748 8/1/95 b

The Ego of the Data Analyst — Postscript

I note, with pain, that this report use few words at all. My intent in this case was to build a chart, one page, few words, that would begin a discussion. I know, and in a report written to professionals you would know, that the comparisons implied by the chart are valid. I know and you know that it is valid to standardize budgets by computing the appropriation *per* person and to compare them in that form. And if someone had chosen to take up the question of validity, I would have been ready. It is not easy work or, to put it another way, you have to be up to a certain level of competence before it *is* easy work. But for the most part, few people will care. On the other hand, if someone does begin to ask you the right questions, you and the interrogator will, each of you, have found a worthwhile colleague.

Try it. Data grouped by the broad classifications of the state accounting categories are enclosed. Without ever setting foot in the Town of Enfield, without hours spent over the budget and discussions with the accountants, you can easily show that there was a \$400,000 purchase in need of explanation. Without any attention to the local press you will find that someone really should say a few words about a \$6,000,000 piece of goods purchased by the Town of Lebanon. These are simple statistical outliers brought into focus by the application of mean based budgeting.
