

objective ...determined by and emphasizing the features and characteristics of the object, or thing dealt with, rather than the thoughts, feelings, etc., or the artist, writer, or speaker.

subjective ...of or resulting from the feelings or temperament of the subject, or person thinking, rather than the attributes of the object thought of ...

Websters New World Dictionary, College Edition 1956.

Transforming the Complex into the Simple: Well-Behaved Variables

Data analysts attempt to reach objective conclusions. But the path to objective results is full of seeming contradictions one of which is that the path itself is not objective, only the conclusion. If there is one way of assigning a unit of measure to the unit of analysis, then there is *always* a second, and a third and infinitely many way of assigning units of measure. And the analyst must choose among them. The choice will have consequences. It will affect the path of the research. Yet the choice must be made before the result is known. So — among the many ways to proceed with one set of data — which one is right?

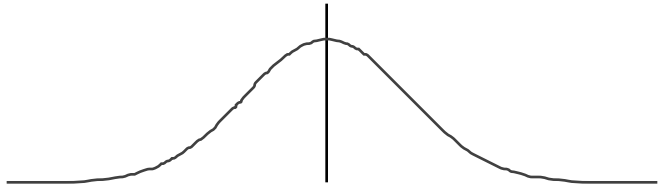
The answer depends on the concept of a *well-behaved variable*. Eventually I will provide reasons why this concept “should” be as useful as it is. Eventually, I will philosophize with respect to its meaning. But make no mistake: The “proof of the pudding” is that this thing — the concept of a well-behaved variable — works. Logical argument as to why this concept should work may or may not be convincing. My explanation of the reason why this concept works may or may not even be correct. No matter. It works.

Five properties identify “well-behaved” variables. A well-behaved variable is:

1. Symmetrical
 2. Homeoscedastic
 3. Linear
 4. Additive
- and
5. It makes sense.

1. Symmetrical

The distribution of a well-behaved variable is symmetrical around the center of the distribution: The upper quartile is as far above the median as the lower quartile is below the median. Often a well behaved variable is both symmetrical and bell-shaped, suggesting an idealized form known by several names as a “bell shaped curved”, or “normal distribution” or “Gaussian distribution”.

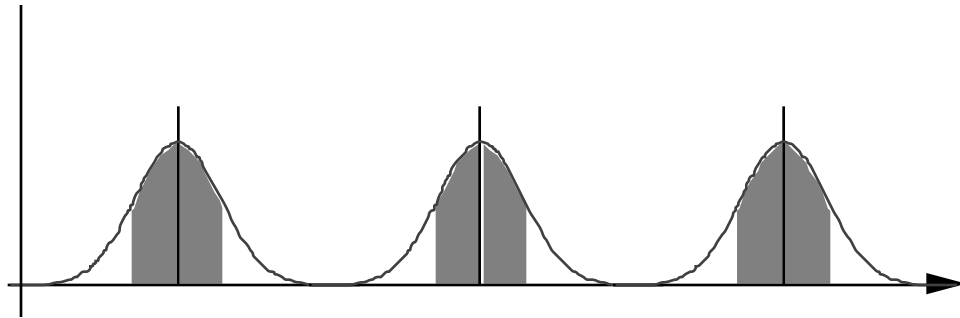


2. Homeoscedastic

The variation of a well-behaved variable is constant from case to case. For example, if individual wealth is a well-behaved variable, then the variation of wealth in the United States in 1960 and the variation of wealth in the United States in 1990, must be (more or less) constant from 1960 to 1990 — even though the average income will have increased considerably during those thirty years. *If* individual wealth

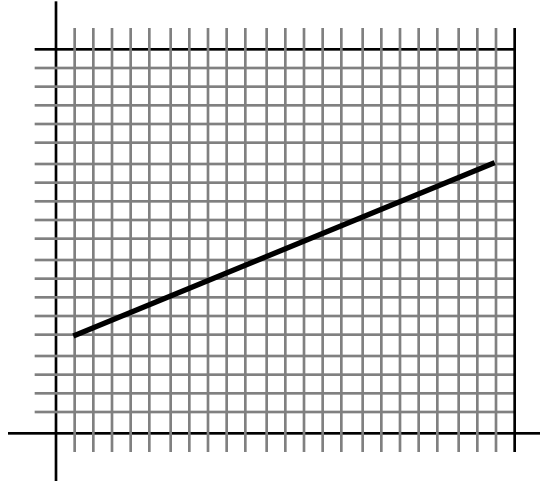
is well-behaved, then although the average income will have changed between 1960 and 1990, the variation will not.

Similarly, if individual wealth is a well-behaved variable, then the variation in wealth among those who have completed high school and the variation in wealth among those who have completed college must be (more or less) the same. The average income of the college graduate will exceed the average income of the high school graduate, but the variation of income within each educational group will be the same.



3. Linear

If two well-behaved variables are related at all, then the relation between two well-behaved variables is likely to be linear — This becomes important later when we look for relations and correlations between variables.



4. Additive

If a variable is well-behaved then effects that serve to increase or decrease its value will decrease it or increase it additively. Perhaps the most commonly used examples of non-additive variables are related to health where it is often suggested that risk factors (e.g., smoking, lack of exercise, overweight, poor diet, heredity as risk factors associated with cardiovascular disease) are spoken of as multiplicative in their consequences for disease. This too becomes important later.

5. Makes Sense

If the *logarithm* of personal income is well-behaved, then the logarithm will have an interpretation and it will make sense.

If the cube root of the weight of organisms is well-behaved, while the weight itself is not, then the cube root of weight is a correct unit of measure. The cube root will have an interpretation and there will be good reason why the cube root makes sense.

Memorize this list: *Symmetrical, homoscedastic, linear, additive, and correct.* Much of the magic an experienced data analysis can perform, much of our ability to go beyond common sense, to find order in data, and then to make sense of it, depends on the use of variables that are *symmetrical, homoscedastic, linear, additive, and correct*. That is to say, much of the power that a data analyst and exercise depends on well-behaved variables.

Exercises

1. Compute means and standard deviations from useful subsets of the data for breakfast cereals. Does protein content appear to be homoscedastic?

2. Ditto -- the ten gram weight

3. Tukey Viscosity. Tukey, *Exploratory Data Analysis*, page 25, quoting

In 1963, McGlanery and Harban gave the values in panel A, showing how well they could measure the viscosity of liquids with a device called a capillary rheometer. Make appropriate stem-and-leaf displays for each of the three samples; comment on the appearance of each.

Run (for) each sample	Viscosity in 100,000's of poises		
	Sample I	Sample II	Sample III
1	.384	.661	3.54
2	.376	.671	3.66
3	.376	.688	3.42
4	.371	.644	4.10
5	.385	.668	4.09
6	.377	.648	3.77
7	.365	.706	4.17
8	.384	.715	3.91
9	.365	.647	4.61
10	.384	.682	3.87
11	.378	.692	
12		.729	

(Original source: R. M. McGlanery and A. A. Harban 1963, "Two instruments for measuring the low-shear viscosity of polymer melts," *Materials Research and Standards* 3: 1003-1007. Table 2 on page 1004.)

4. Get the income data referred to in the text.

Transformation

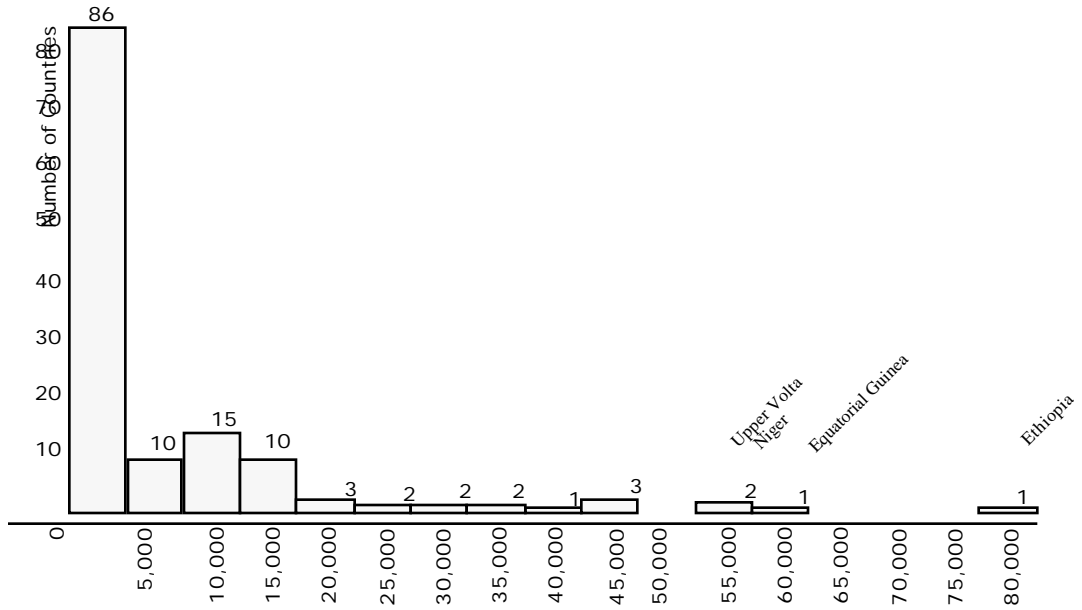
Whenever possible data analysts do not fit their tools to the data, we fit the data to our tools. Data analysts *could* develop tools for variables that are not well behaved, for variables whose distribution is not symmetrical, for relations that are not linear, and so forth, but we do not. Instead we transform the data to make it well behaved.

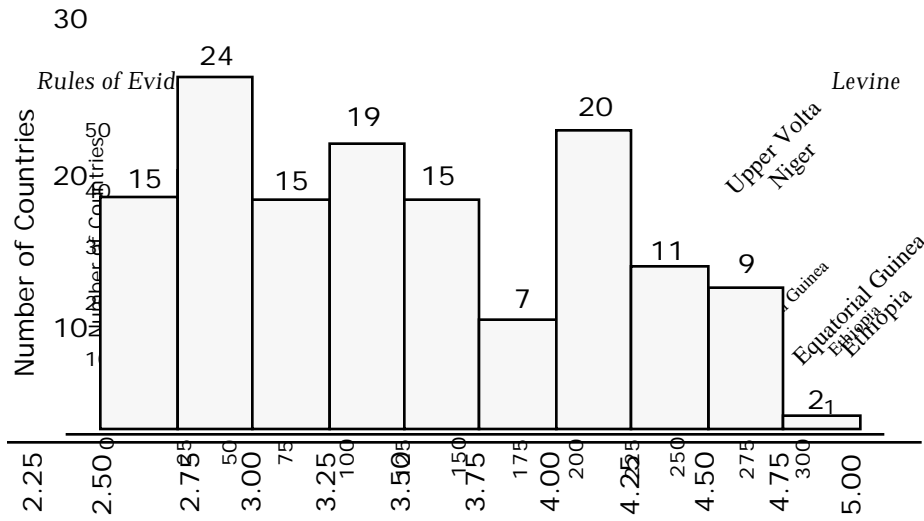
Mathematicians do something similar and it gives them great power: When you have a relatively “difficult” problem like a multiplication problem, a transformation changes a multiplication problem to an addition problem (applied to the logarithms). When you have a difficult problem like the analysis of the sound wave of a musical instrument, a transformation changes combinations of sine waves and cosine waves into linear combinations (of their Fourier transforms).

It is a general strategy for handling complicated problems. Instead of tackling them head on, the genius of the mathematics is to figure how to transform the problem into something simple. The rest is (relatively) easy. Data analysts uses the same strategy — transforming the unit of measure in order to create a well-behaved variable. After that the rest of the analysis is easier.

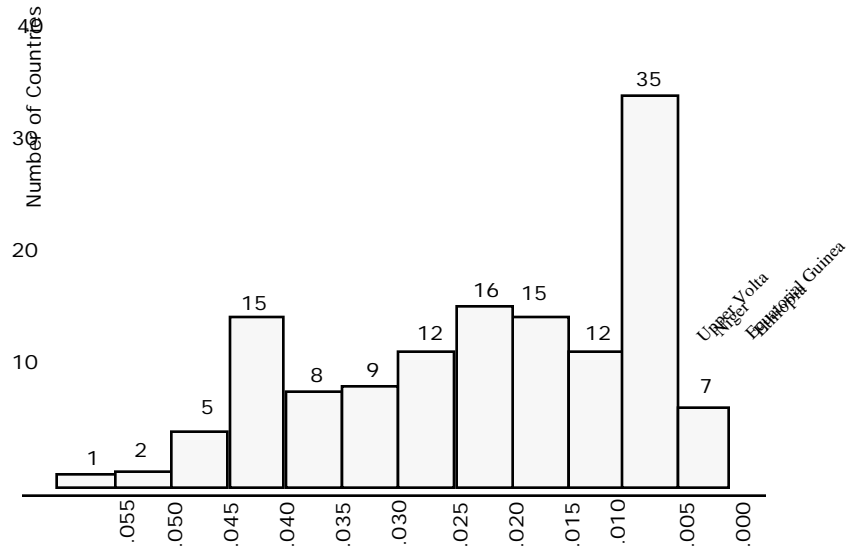
For example, here is a preview of coming attractions: Figure _ shows five transformations of people per physician, transforms ranging from the identity transformation (identical to the variable as given) to the inverse transformation. Between the two extremes you see three intermediate results corresponding to a square root transformation, a logarithmic transformation, and an inverse square root transformation, five in all. (Remember — we don’t have to make sense out of all of these things, only the one that is well behaved. I don’t have to make sense out of the square root or the inverse square root of people per physician — unless it is well-behaved.)

Look at the shapes of these five pictures of the data. In sequence the histograms show a systematic change in the behavior of the distributions. The first, the identity transformation is assymetrical with four physician poor countries, Ethiopia, Equatorial Guinea, Niger, and Upper Volta in extreme positions, out on the tail, away from the main body of the data (without a corresponding tail at the other end of the distribution).

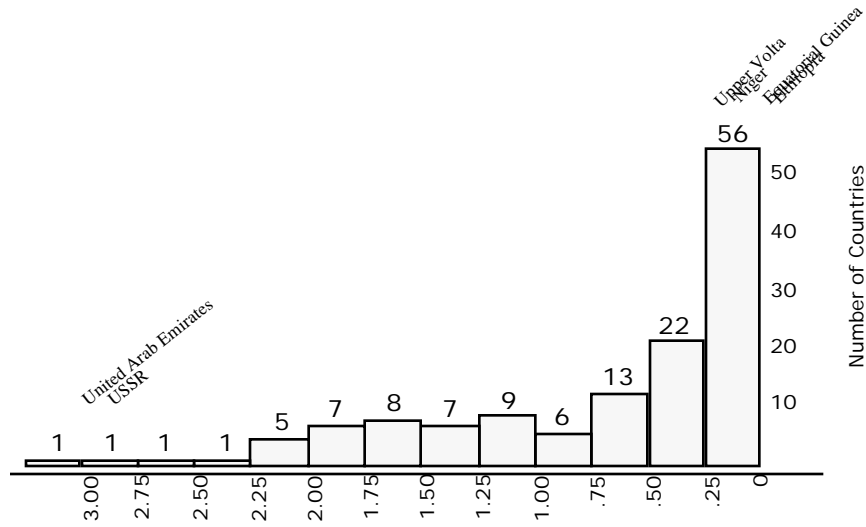




Histogram for Log (base 10) of Persons per Physician



Histogram for Inverse Square Root of Persons per Physician



Histogram for Inverse of Persons per Physician , i.e., for Physicians per Person
(Recorded as Phsycians per 1,000 Persons)

The next transformation, using the 1/2 power instead of the first power is less extreme. The behavior has grown a tail on the left while the tail on the right is less extreme.

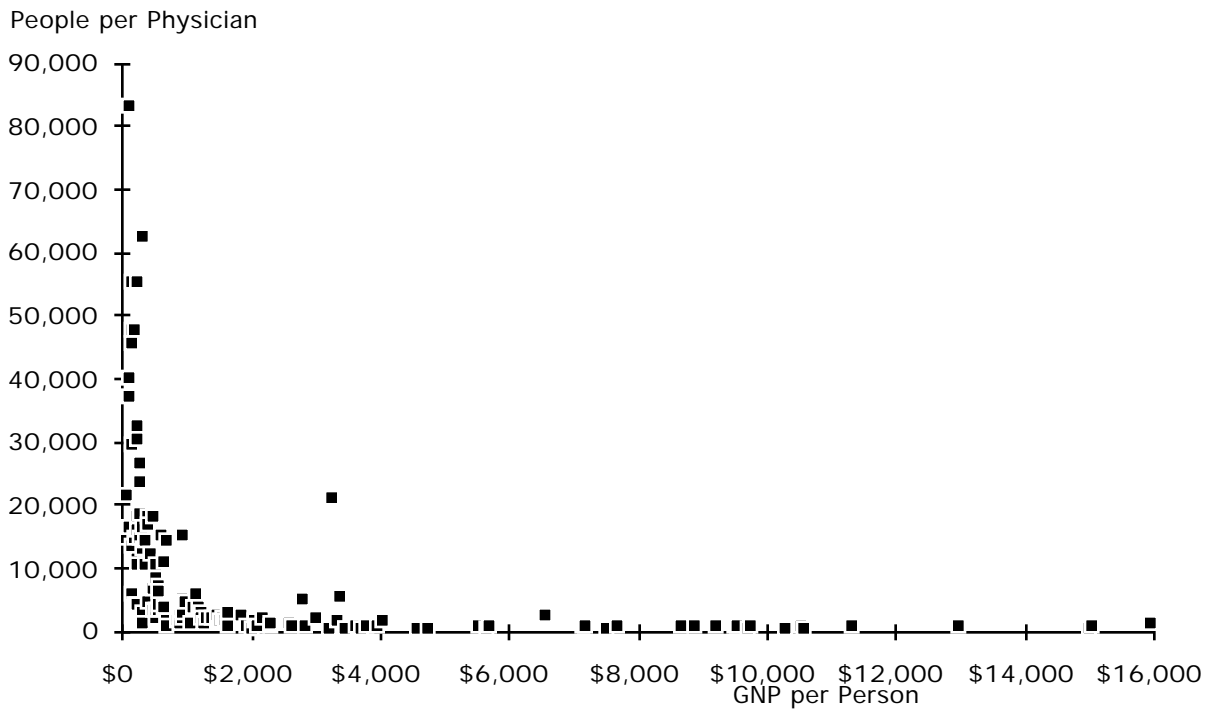
The next transformation using the logarithm instead of the .5 power is (relatively) symmetrical.

The next transformation, using the -.5 power of the variable, shows a long tail on the left and a short tail on the right. For example, the two countries that were previously extreme are now packed in close to the the center.

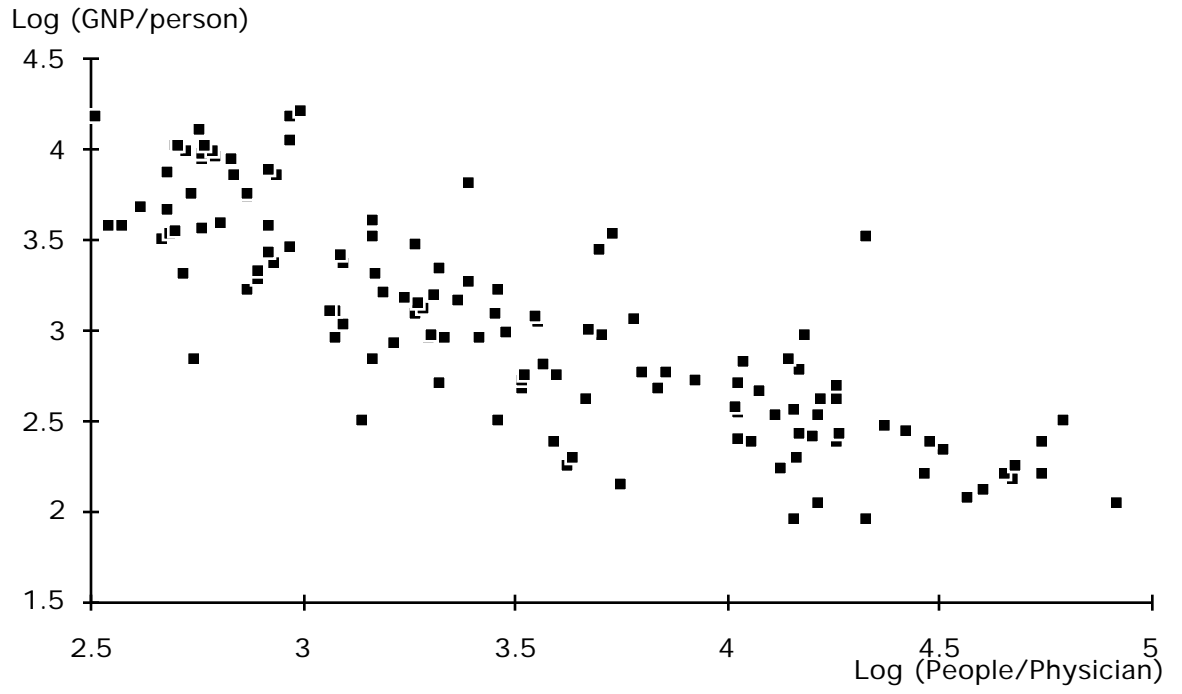
And finally the inverse transformation shows extreme behavior. It is decidedly asymmetrical, but here the tail is tacked to the opposite end as compared to the behavior of the original unit of measure.

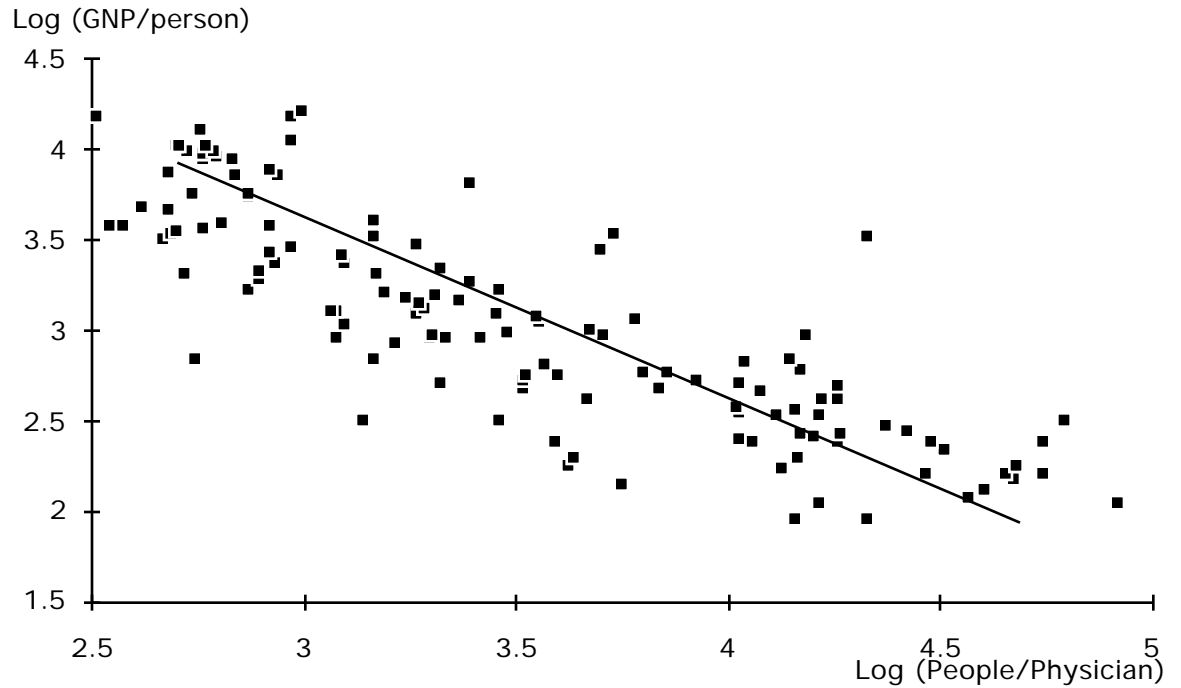
That is a demonstration of the power of transformation to change the picture. From these the data analyst chooses how the data will be and how the data should be transformed.

For another preview, consider physicians per person, the same variable, compared to GNP per capita for the same countries. Figure _ shows the graph of the relation between these two variables. Neither variable is well behaved. Neither is symmetrical (first criterion). And their relation is not well behaved — not linear (third criterion). I would not care to go forward with an analysis of the relation displayed in this graph — too complicated: The picture suggests a chevron shaped distribution with two distinct wings.



Now, transforming both variables, here is the new picture of the relation (using the same data):





This picture gives me a place to begin. Both figures “describe” exactly the same reality, but in the second figure provides my intuition and my sense with ample cues with which to interpret the relation. The second picture is approximately linear. Moreover, the slope is close to being negative one — not some relatively complicated number like 2 or 3, but negative 1. I can begin to make sense of that. (It tells me that to a first approximation the number of physicians in the country is proportional to the wealth of the country.)

The most common choices among transformations are organized according to the power of the transformation, where power refers to the exponent of the transformation. Here we have considered five, where the “0” power is considered the log. The five transformations had a progressive effect. An increase of power decrease the appearances of tails on the right and increases the appearances of a tail on the left. A

decrease in power increases the appearances of tails on the left and decreases the appearances of a tail on the right.

Power	Transformation	May Indicate
...	...	
1	x^1	Identity transformation (no change)
1/2	$x^{1/2}$	Square Root
0	$\log(x)$	Logarithm
-1/2	$x^{-1/2}$	Inverse square root
-1	x^{-1}	Inverse
...	...	

Common Transformations and Their Indication

For $p=1$ the picture shows a highly assymetrical distribution with the United Arab Emirates and USSR out on the tail away from the main body of the data. By contrast, for $p=-1$ the behavior is highly assymetrical in the opposite direction, with Ethiopia, Equatorial Guinea, Niger, and Upper Volta solidly close to the main body of the data — the tail is at the other end. In between there is a transition from one extreme shape to the other. And we narrow the choice among alternative units of measure by choosing the one that is well-behaved.

Interpreting The Data

For people per physician, the well-behaved choice requires a compromise between the first criterion, symmetry, and the fifth criterion, sense. By itself, the first criterion would lead to something like the negative one tenth power, $p = -.1$. But the fifth criterion leaves me in trouble attempting to interpret the $-.1$ power whereas the logarithm, which is close, is easy to interpret. So I will use the logarithm of physicians per capita as my unit of measure. The sense of the logarithm is that adding and subtracting to the logarithm corresponds to multiplying or dividing the original. Well behaved logarithms imply that two or more values of the original variable should be compared using ratios or percentages.

To actually write it up I have to speak to two audiences. One is me. For me I have to keep it simple — well behaved variables, well-behaved relations between variables when we get to relations. The other audience is a “general public” that will be none to pleased by a statement like “the median number of people per physician is approximately 3.4 in logs base 10”.

So, in order:

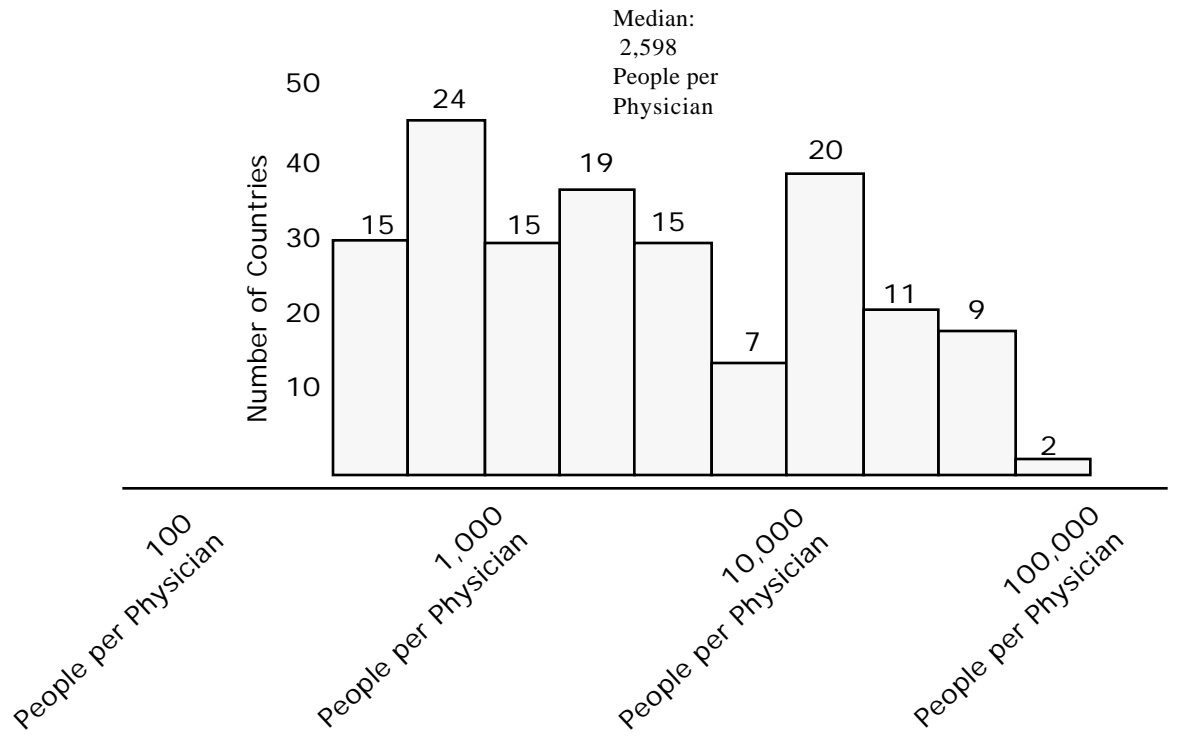
1. Transform the data to a well behaved variable.
2. Analyze the transformed data.
3. Translate the analysis into units of measure that are “friendly” to a non-technical consumer of the data.

For example, using physicians per persons, using logarithms and the rank order statistics here is a brief description of the facts.

<p>Using the logarithm of the number of people per physician, in 1970 the typical country showed a median of 3.41. For example, Mauritania, Saudi Arabia, and Iraq were all close to the median value. The shape of the distribution, Figure __, shows a large range of values but no clear evidence of polarization into two distinct groups, as for example, rich and poor, with nothing in between. Nevertheless the middle fifty per cent of the distribution shows a large range from 2.94 to 4.11. While the full range extends from 2.51 to 4.92, even at the extremes, none are so low or so high as to suggest a sharp differentiation making some of the countries radically different from the rest.</p>	<p>Transform the data</p> <p>Central value</p> <p>Examples</p> <p>Implicit recognition of the shape.</p> <p>Range of typical values</p> <p>Full Range</p>
---	---

And now translating

<p>In 1970 the typical country showed a median of approximately 2,600 people per physician. For example, Mauritania, Saudi Arabia, and Iraq were all close to the median value. The shape of the distribution, Figure __, shows a large range of values but no clear evidence of polarization into two distinct groups, as for example, rich and poor, with nothing in between. Nevertheless the middle fifty per cent of the distribution includes more than ten fold contrasts, from fewer than 100 people per physician to more than 1,000 people per physician.</p> <p>The full range extends from an extreme of only 300 people per physician (United Arab Emirates) to an extreme of 80,000 people per physician (Ethiopia), a 300 fold contrast from the most physician intensive to the least physician intensive society.</p> <p>Even at the extremes, none are so low or so high as to suggest a sharp differentiation making some of the countries radically different from the rest.</p>	<p>Invert the transform and translate.</p> <p>The median and quartiles are easy to translate, because the median country is the median country regardless of the unit of measure.</p> <p>The Figure should be translated by relabelling the x axis in physicians per person even while the shape is computed using log physicians per person.</p> <p>Use words suggesting multiplication (ten fold) because “plus and minus” in terms of logs (original analysis), corresponds to multiplication in terms of people per physician.</p>
--	--

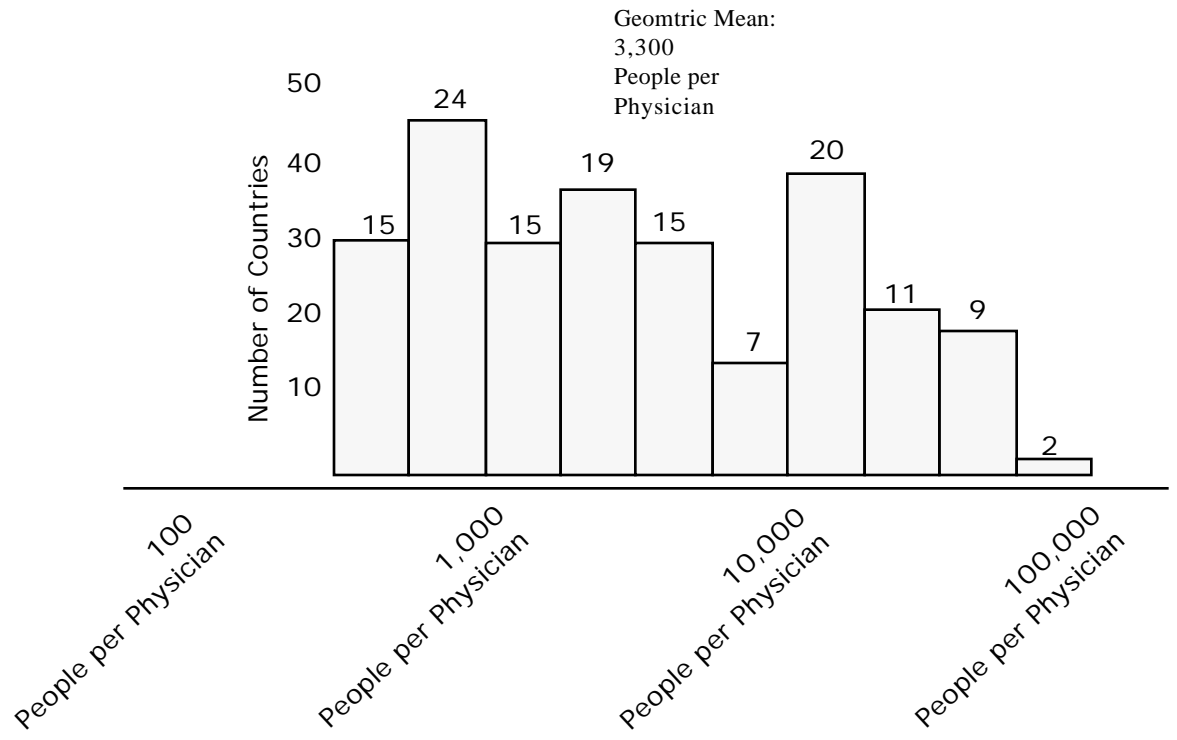


And in least square statistics:

<p>Using the logarithm of the number of people per physician, in 1970 the typical country showed a mean of 3.52. For example, the Phillipines, Syria, and Honduras were all close to the median value. The shape of the distribution, Figure __, shows a large range of values but no clear evidence of polarization into two distinct groups, as for example, rich and poor, with nothing in between. Nevertheless the variation is large with a standard deviation of .64.</p> <p>The full range extends from 2.51 to 4.92, with physician-poor Equatorial Guinea and Ethiopia standing approximately two standard deviations away from the mean at one end of the distribution at the extremes, none are so low or so high as to suggest a sharp differentiation making some of the countries radically different from the rest.</p>	<p>Central Value</p> <p>Interpretation of the shape</p> <p>Reporting and interpreting the standard deviation</p> <p>Range</p> <p>Marking the extremes using two standard deviations</p>
---	---

And now translating

<p>In 1970 the data for 138 countries showed a geometric mean of 3,300 people per physician. For example, the Phillipines, Syria, and Honduras were all close to the mean value. The shape of the distribution, Figure __, shows a large range of values but no clear evidence of polarization into two distinct groups, as for example, rich and poor, with nothing in between. Nevertheless the variation is large with a standard deviation around the central value that corresponds to a factor of 4.4.</p> <p>The full range extends from 300 people per physician to 80,000 people per physician, with physician-poor Equatorial Guinea and Ethiopia at extreme values showing more than twenty times the mean value of people per physician.</p>	<p>In order: first transform the data, using logs. Then compute the mean of the logarithms. Then compute the anti-log of the mean. The result is called the geomtric mean</p> <p>The Figure should be translated by relabelling the x axis in physicians per person even while the shape is computed using log physicians per person.</p> <p>Use words suggesting multiplication (ten fold) because “plus and minus” in terms of logs (original analysis), corresponds to multiplication in terms of people per physician.</p> <p>Dodging on my use of plus or minus two standard deviations. The problem is that there is no term in general use for the anti-log of the standard deviation of the log. You would expect it to be called the “geometric standard deviation, but it just does not get named. So, I use it to make an intepretive statement.</p>
--	---



Exercise: Consider the data for nations. Using population as the unit of measure, write a brief report summarizing the report, including what is large (and very large). Then, by contrast, use the logarithm of population as the unit of measure and write another brief report. Compare the two? Is China is certainly the largest, by population. But how large? Is it an outlier — so large as to be unrelated to the rest? Or is it merely the largest and not otherwise remarkable?

Exercise: Consider the population data for nations, two different years, and compute the change in population:

First, using the nation as the unit of analysis and millions of people as the unit of measure, apply one variable technique, shape of the distribution, measures, and examples, to obtain a brief report of change.

Then, second, using the nation as the unit of analysis and percent of population (first year) as the unit of measure, apply one variable technique, shape of the distribution, measures, and examples, to obtain a brief report of change.

Exercise: As above for GNP (or immigration, or imports v/s imports as a percentage of GNP).
