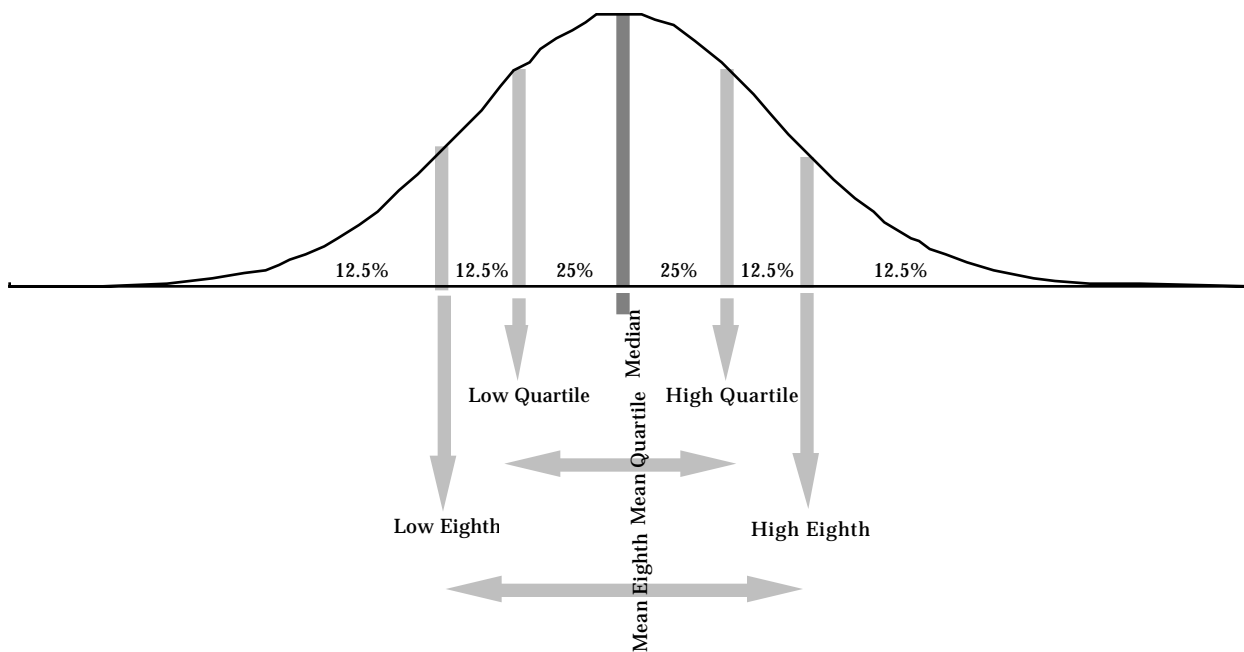


Is it Symmetrical?

Whether or not a variable has a symmetrical distribution is exceedingly important both for descriptive analysis and for more advanced statistical methods. In a simple case there is no need for “high tech” to judge symmetry. Looking back at the people per physician data I feel perfectly competent, on the authority of my eyeball, to look at the picture and assert that the distribution, using people, is not symmetrical. And I feel perfectly competent to look at the second distribution, using logs, is more symmetrical than the first. But for less blatant cases of asymmetry I need a procedure. How should I decide whether data are or are not symmetrical?

The trick is to return to the picture of symmetry and put some numbers on what the eyeball “sees” and identifies as symmetry.



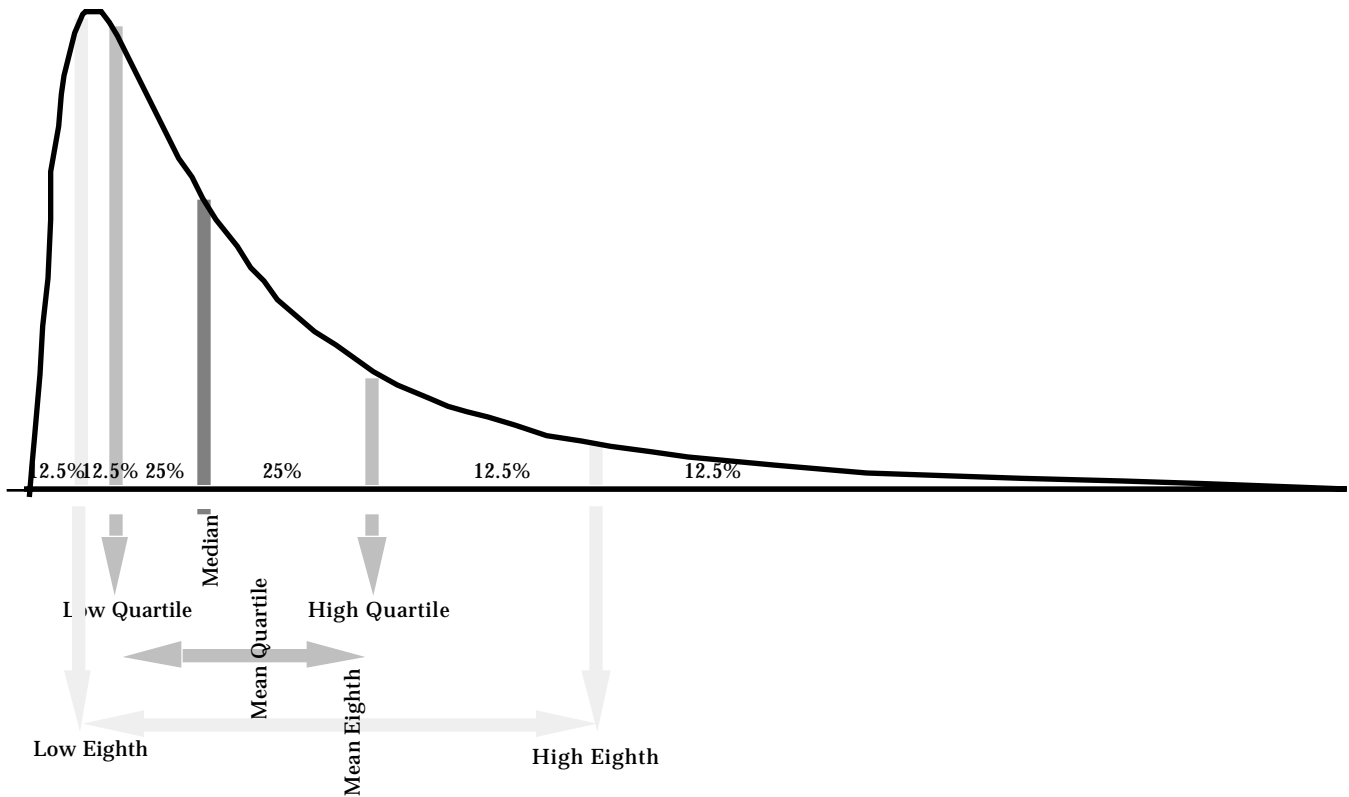
If the distribution is symmetrical, then the quartiles will be located, symmetrically, at equal distances from the median. And, therefore, if the distribution is symmetrical then the point exactly half way between the two quartiles will be equal to the median.

If symmetry, then mid-quartile = median

That's easy enough to test: You simply compute the mean quartile and compare it to the median. But generally, two numbers computed from data are rarely equal, they do not match precisely and out to infinite numbers of decimal digits. So we need a test that is a little more clever. For that purpose, following Tukey's Exploratory Data Analysis, compute two more numbers, the two "eighths" and compute the "mid-eighth". Defining terms: As the two quartiles mark the two outer quarters of the distribution, the two eighths mark the two outer eighths of the distribution. And the mid eighth is the point midway between the two eighths. And again, if the distribution is symmetrical then the mid eighth will be equal to the median.

If symmetry, then mid-eighth = median

Now I can get a practical test of symmetry, referring to the asymmetrical distribution in Figure 2: In practice, if there is a trend among the three numbers, from the median to the mid-quartile to the mid eighth, then there is evidence of asymmetry. If the mid-eighth is greater than the mid quartile and the mid quartile is greater than the median, then the distribution is asymmetrical with a tail to the right. If the mid-eighth is less than the mid quartile and the mid quartile is less than the median, then the distribution is asymmetrical with a tail to the left. And if there is no trend, then the distribution is symmetrical. Or — to be very precise (using a double negative): If there is no trend, then there is no evidence of asymmetry.



If you want greater certainty, then you continue the investigation: Adding the mid-sixteenth, the mid-thirty-second ... as much as your data will allow.

Defining the "eighths"

To be sure that there is no ambiguity let me specify the step by step computation for the eighths: We find them by mimicking the procedures that have already been used to define the median and the quartiles. Recall that for the fifty-fifty split,

n = number of values in the data

$$m = \text{location of median} = (n+1)/2$$

And, to repeat, if the result is a whole number then the number of values that are greater than or equal to the median is m . If the result is a whole number, then the m th value, in rank order, is the median. If the result is a fraction, then m lies between two values whose mean is the median.

For the quartiles, splitting off twenty-five percent at each end, we compute m which is the integer part of m (lopping off the fraction if there is one) and use it to compute the locations of the quartiles

$$m = \text{number of values greater than or equal to the median}$$

$$q = \text{location of quartiles} = (m+1)/2$$

Mimicking the logic for the median: if the result, q , is a whole number then the two q -th values, in order from each end of the distribution, are the quartiles. If the result is a fraction then the m -th value at each end lies between two values whose mean is the quartile

are found by counting in q values from each end of the data

identifies the location, then the number of values that are greater than or equal to the median is the integer part of m , m .

And now for the eighths, splitting off twelve and one-half percent at each end, we compute q which is the integer part of q (lopping off the fraction if there is one) and use it to compute the locations of the eighths.

$$q = \text{number of values greater than or equal to the quartile}$$

$$e = \text{location of the eighths} = (q+1)/2$$

If the result, e , is a whole number then the two e -th values, in order from each end of the distribution, are the eighths. If the result is a

fraction then the e -th value at each end lies between two values whose mean is the eighth.

Working with the 100 observations of the 10 gram weight, shown in rank order in Table 1, $n = 100$. So

$$n = 100$$

$$m = (n+1)/2 = (100+1)/2 = 50.5$$

The median is the mean of the 50-th and 51-st values, median = $(9.999596+9.999596)/2 = 9.999596$

Then m is the integer part of m :

$$m = 50$$

$$q = (m+1)/2 = (50+1)/2 = 25.5$$

The high quartile is the mean of the 25th and 26th values in rank order from the high end, $Q_+ = (9.999599+9.999599)/2 = 9.999599$. And the low quartile is the mean of the 25th and 26th values in rank order from the low end, $Q_- = (9.999593+9.999593)/2 = 9.999593$.

Then q is the integer part of q :

$$q = 25$$

$$e = (q+1)/2 = (25+1)/2 = 13$$

The high eighth is the 13th value in rank order from the high end, $E_+ = 9.999601$. And the low eighth is the 13 value in rank from the low end, $Q_- = 9.999590$.

Rank High to Low	Rank Low to High	Item	Weight in Grams	Rank High to Low	Item	Weight in Grams
1	100	94	9.99962551	50	89	9.999596
2	99	63	9.99960852	49	100	9.999596
3	98	85	9.99960753	48	19	9.999595
4	97	26	9.99960354	47	40	9.999595
5	96	11	9.99960255	46	41	9.999595
6	95	97	9.99960256	45	54	9.999595
7	94	4	9.99960157	44	62	9.999595
8	93	16	9.99960158	43	3	9.999594
9	92	22	9.99960159	42	6	9.999594
10	91	23	9.99960160	41	37	9.999594
11	90	25	9.99960161	40	38	9.999594
12	89	29	9.99960162	39	46	9.999594
13	88	43	9.99960163	38	52	9.999594
14	87	2	9.99960064	37	65	9.999594
15	86	17	9.99960065	36	72	9.999594
16	85	32	9.99960066	35	80	9.999594
17	84	74	9.99960067	34	82	9.999594
18	83	7	9.99959968	33	96	9.999594
19	82	9	9.99959969	32	98	9.999594
20	81	15	9.99959970	31	13	9.999593
21	80	18	9.99959971	30	27	9.999593
22	79	28	9.99959972	29	35	9.999593
23	78	30	9.99959973	28	45	9.999593
24	77	34	9.99959974	27	53	9.999593
25	76	59	9.99959975	26	64	9.999593
26	75	77	9.99959976	25	70	9.999593
27	74	83	9.99959977	24	92	9.999593
28	73	90	9.99959978	23	21	9.999592
29	72	91	9.99959979	22	68	9.999592
30	71	5	9.99959880	21	75	9.999592
31	70	14	9.99959881	20	79	9.999592
32	69	20	9.99959882	19	81	9.999592
33	68	24	9.99959883	18	1	9.999591
34	67	39	9.99959884	17	42	9.999591
35	66	44	9.99959885	16	48	9.999591
36	65	50	9.99959886	15	73	9.999591
37	64	60	9.99959887	14	95	9.999591
38	63	8	9.99959788	13	33	9.999590
39	62	10	9.99959789	12	56	9.999590
40	61	12	9.99959790	11	57	9.999590
41	60	31	9.99959791	10	58	9.999590
42	59	67	9.99959792	9	55	9.999589
43	58	99	9.99959793	8	71	9.999588
44	57	49	9.99959694	7	84	9.999588
45	56	51	9.99959695	6	93	9.999588
46	55	61	9.99959696	5	47	9.999587
47	54	66	9.99959697	4	88	9.999585
48	53	69	9.99959698	3	87	9.999582
49	52	76	9.99959699	2	36	9.999577
50	51	78	9.99959600	1	86	9.999563

Now back to the point, which is to estimate whether or not these data are symmetrical. What we would like is equality: with the median having exactly the same value as the mean quartile and the mean eighth but with real data that is unlikely. What we settle for is a comparison of the median, the mean quartile, and the mean eighth that shows no trend.

For the ten gram weight, what is the evidence:

The median is 9.999596 grams

The mean quartile is $(9.999593 + 9.999599)/2 = 9.999596$ grams

The mean eighth is $(9.999590 + 9.999601)/2 = 9.9995955$ grams

Reasoning negatively: The numbers do not show clear evidence of asymmetry, so I do not have convincing reason to reject the hypothesis that the measurement errors are described by the hypothesis.

Homework:

1. Pick some easily measured number such as your own pulse (counting for a full 60 seconds to gain precision), or your own blood pressure, or the weight of a coin or the diameter of a coin if you have the equipment. Get at least ten estimates. What is the shape of the distribution for your ten or more estimates?

2. There is a certain ambiguity about the numbers for the ten gram weight: The mean quartile is indistinguishable from the median; the mean eighth is a bit less than the mean quartile. Having more data here, 100 observations, pursue this a fit further: Compute the mean sixteenth and the mean thirty-second. Interpret the whole set of mean value numbers

3. Return to the data for People per Physician, using the logarithm as the unit of measure. Is it symmetrical? Push to the mid sixteenth or further. Is it symmetrical?

Stretching and Shrinking: The Construction of an Interval Scale

One way to understand the concept of a well behaved variable is by the use of another concept employed by data analysts and mathematical modelers. Roughly defined, numerical interval scale must have a correct relation to comparisons among the objects the scale is supposed to represent: If you have measured an object with numbers 1,2,3, then the substance of the differences among the objects must correspond to the differences among the numbers that represent them.

This is a hidden assumption in virtually any numerical procedure applied to data. Consider the mean for example. The mean is so transparent an object that it might seem strange to say that the use of the mean requires certain usually unstated assumptions. That's why I choose it. Recall what a mean is: The mean of a set of numbers is a center that is close to all of the numbers. It is close to them in the sense that it minimizes the squared deviations between the center and the numbers for which it is the center.

There is the key: the deviations. The deviations are a set of intervals: For the first number in the set of data, the deviation is $x_1 - \bar{x}$. That is an interval. For the second number in the set of data, the deviation is $x_2 - \bar{x}$. So when I use the mean, I am assuming that the meanings of these intervals are appropriately represented by the numbers.

When you use the fences to mark out the limits of reasonable variation, you add a number to the high quartile and you subtract a number from the low quartile — which assumes that being so many units above the quartile has a meaning directly comparable to being so many units below the quartile. When you use the standard deviation to mark out limits, again there is an assumption of symmetry, that it is as normal to be one standard deviation above the mean as it is to be one standard deviation below the mean.

Very often these symmetries are not realized as you saw in blatant terms where the boundary for the number of physicians two standard deviations below the mean number of physicians (or below the lower fence) was a negative count — negative physicians — which is ridiculous. That is to say, the moral of the story is that the arithmetic of most data analysis requires interval scales. Without an interval scale even so low tech a computation as the mean is not a valid operation on the numbers. And sometimes the result is not only wrong but obviously wrong as, for example, when it puts the data analyst in the embarrassing position of using numbers that refer to negative people or perhaps negative age or negative income.

In data — as they are presented to the analyst — meaningful numbers are far from guaranteed: For me, counting money as money in hand, the differences between ten dollars in my wallet and twenty dollars and the between ten thousand dollars in my wallet and ten thousand are not the same. From ten to twenty is doubling. From ten thousand to ten thousand and ten the difference is lost in the small change.

But, I have to admit that this statement about unequal intervals is not guaranteed. It depends on context: To an accountant ten dollars is ten dollars. Ten dollars has the same effect on the total (the bottom line) whether it is contributed by an account with little more than ten dollars or one with a great deal more. In this context ten contributes ten to the total wherever it comes from.

If I am measuring traces of a chemical compound, the difference between no trace of the element and one molecule may be extremely important while the difference between one hundred grams of the compound and one hundred and one may have relatively little effect on the conclusions or direction of my research.

For mathematics the differences between numbers may be established by mathematical definition. For the scientist using math

to process of assignment of numbers requires some care and depends on context. The use of transformations speaks to the problem of changing the intervals of the scale. The mathematics of these transformations stretches some parts of a scale relative to others, with the consequence that the change of unit can change the behavior of the variable. For example, comparing dollars as the unit of measure to the logarithm of the number of dollars as the unit of measures, note how the logarithm stretches the equal dollar scale at the left in Figure 1. Using the dollar as the unit of measure, the four different incomes, \$25,000, \$50,000, \$75,000, and \$100,000 are separated by three equal intervals, in dollars.

Re-expressed in logs at the right, the intervals change, stretching the distance between $\log(25,000)$ and $\log(50,000)$ as compared to the distance between $\log(50,000)$ and $\log(100,000)$.

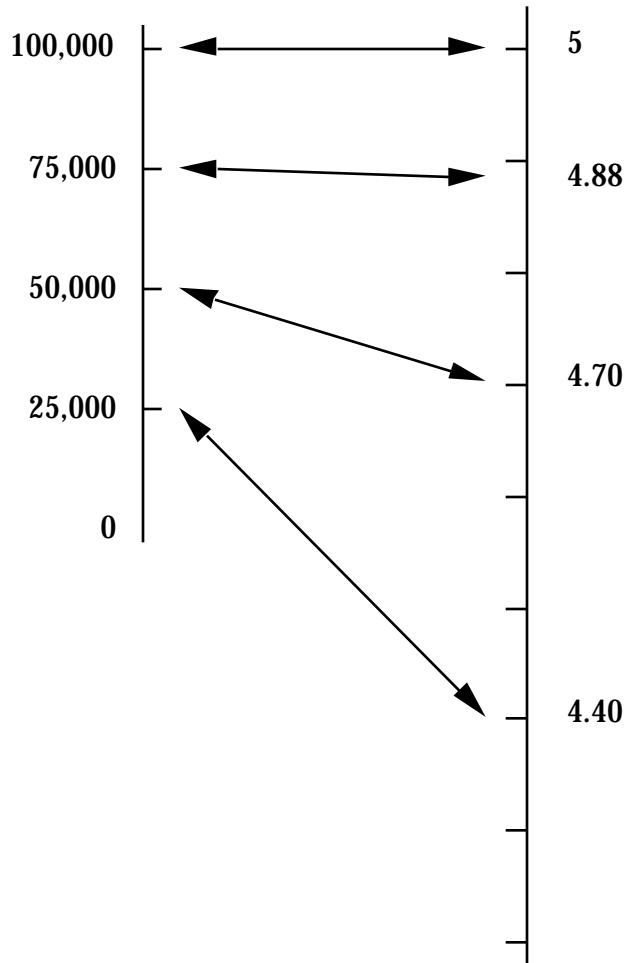


Figure __
Re-Expression of Dollar Values as Logarithmic Values, Using
Logarithms Base 10.

Note that the re-expression using logs stretches intervals among small values relative to intervals among the large values.

This “stretching” changes everything: It changes the shape of the distribution, it changes the variation, it changes the relation between one variable and another, and it changes the meaning of the variable. And, in particular, it is capable of transforming a poorly-behaved variable into a well-behaved variable. Here for example is the histogram of the wealth of nations for 19__, first in dollars, and then in log dollars.

Figure: Histograms of gross national products, in dollars and in log dollars.

Exercise

Describe the distribution of gross national products of states of the Western Hemisphere, without logarithms, and with logarithms, in 19__ and 19__ **Get the data**

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.

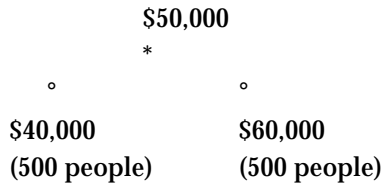
Then, second, using the nation as the unit of analysis and percent of population (first year) as the unit of measure, apply one variable technique, shape of the distribution, measures, and examples, to obtain a brief report of change.

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

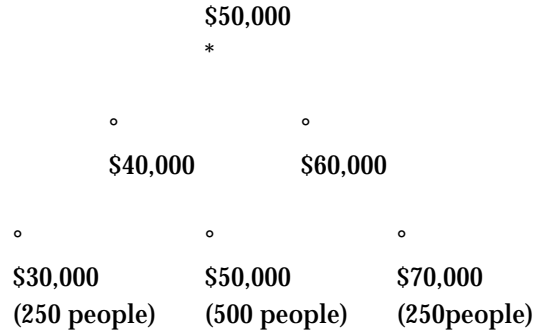
Transformations

I have tried to convince you by logical argument that things, things out there in the real world, “should” have symmetrical bell-shaped distributions whereas, on the other hand, truth is they do not — not even close. Why? Well, to give you an explanation that tries to salvage both the argument and the reality, consider two hypothetical models of personal income.

Let me imagine a group of 1,000 people, all of whom have an income of \$50,000, and watch what happens to them over time. Life can be good and life can be bad: At the end of a year, half of them get a \$10,000 increase, half get a \$10,000 decrease, half get a \$10,000 increase. Now I’ve got 500 people with \$40,000 incomes, 500 people with \$60,000 incomes.



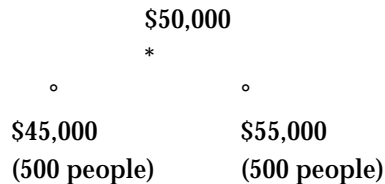
Life goes on and again, half get a \$10,000 increase and half get a \$10,000 decrease. That gives me 250 people with \$30,000, 250 people who dropped to \$40,000 and then bounced back to \$50,000, 250 more people who rose to \$60,000 and then went down to \$50,000, and 250 people at \$70,000.



Let life run on run again, again suppose half go up \$10,000 and half go down \$10,000

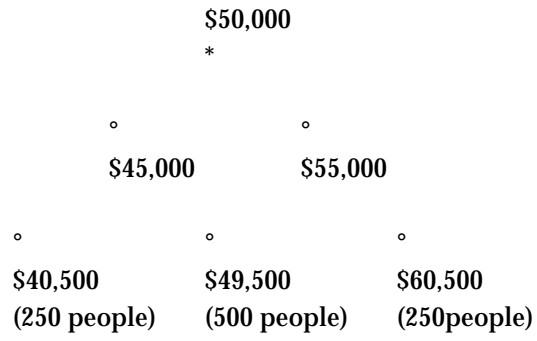
The process seems perfectly ordinary: A few people will got to the top. Some will get to the bottom. The result of their performance, their income distribution, will be the symmetrical result of a symmetrical process.

That's one look at a hypothetical income process. Here's another. This time let me start with a group of 1,000 people, all of whom have an income of \$50,000, and watch what happens to them over time and then, at the end of a year, half of them get a \$10,000 increase, half get a 10% decrease, half get a 10% increase. Now I've got 500 people with \$40,000 incomes, 500 people with \$55,000 incomes.



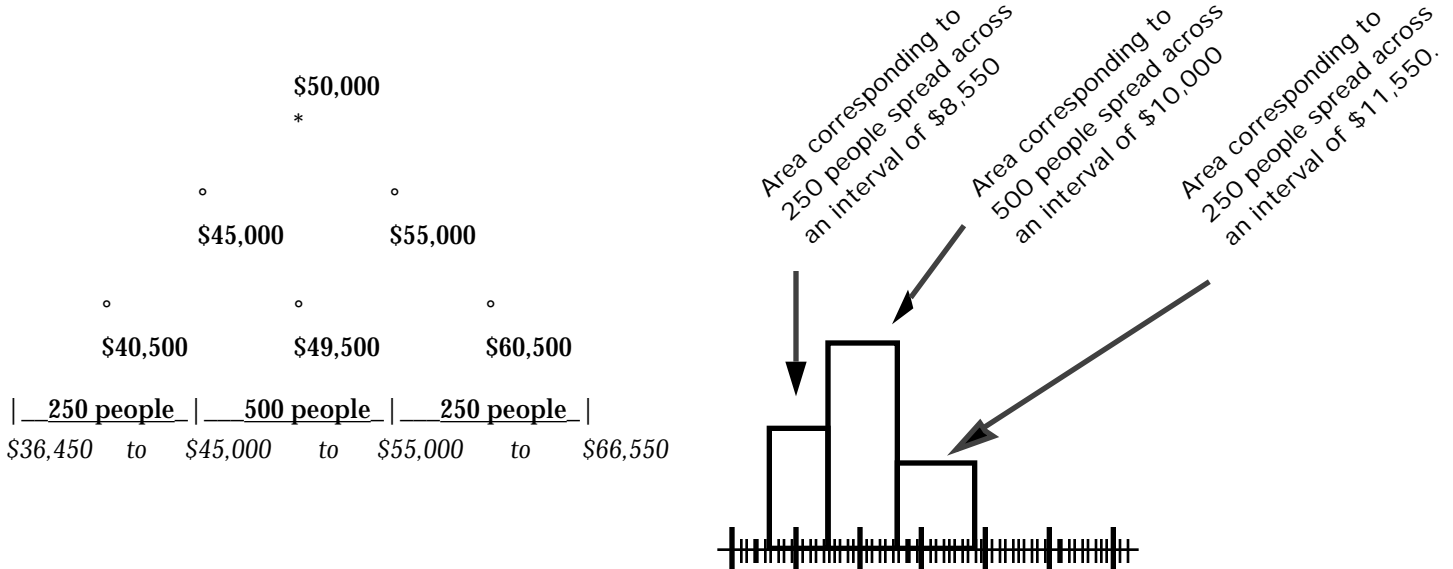
Life goes on and again, half get a 10% increase and half get a 10% decrease. That gives me 250 people with \$44,500, 250 people who dropped to \$45,000 and then bounced back to \$49,500, 250 more people

who rose to \$55,000 and then went down to \$49,500, and 250 people at \$60,500.



Again, let life continue for these people, again suppose half go up 10% and half go down 10%. This second process also seems perfectly ordinary: A few people will get to the top,. Some will go to the bottom. If anything this is probably more realistic — these people had income changes that were proportional to the income they already had, some percent up or some percent down. And the second process too has a feel of symmetry about it. But look at the result: These things aren't equally spaced: The gap between the 250 people at the left and the 500 people in the center is \$9,000. But the gap between the 500 people at the center and the one at the right is \$11,000.

As a result, if we collected these hypothetical data and organized them into a histogram, the histogram would be assymetrical, skewed to the right.



This histogram is only a little bit “off” of symmetry, but it would get worse if I followed it out to allow more and more “bounces” to affect this population, some up and some down. So how do I reconcile this with the privileged place of bell-shaped symmetrical distributions?

The answer is to transform the data. And the reason that that answer is right is because the process itself is not equally spaced in dollars, The process is being performed in percentages. And when you transform the data to a unit of measure that is consonant with the unit in terms of which the process itself is behaving, the result is symmetry.

Data analysts will go one step further, transforming the data using logs rather than percentages. The reason for this is that percentages don't add up: On an interval scale you want an interval of 1 added to an interval of 1 to add up to an interval of 2, one plus one (should be) equal to 2. But for percentages a 1% increase followed by a second 1% increase does not add up to a 2% increase, not quite. (They combine to a 2.01% increase.) Percentages do not add up. So if you try to draw percentages as an interval scale you get into trouble, more trouble with larger percentages. Percentages are good summary measures because people accept their intuitive meaning. But they get you into trouble if you try to use them in an analysis, even so simple an analysis as a histogram or a stem and leaf.

Logarithms, as compared to percentages “add up”. So we use them where common sense would have us use percentages — because we know that the idea is right but that percentages do not quite do the job.

So for this problem the symmetry of the problem makes itself visible in the picture of the data — using *logarithms*. My people start at log \$50,000. Those whose money increases go up from log 50,000 to log 50,000 plus log (1.1): That corresponds to multiplying the \$50,000 by 1.1 (increasing it by 10%), except that, using logs, I simply add the logarithm of 1.1.

Those whose money decrease below \$50,000 go down from log 50,000 to log 50,000 minus log (1.1): Transformed using logs that is

$$\begin{array}{ccc}
 & \log(\$50,000) & \\
 & * & \\
 & \circ & \circ \\
 & \log(\$50,000)-\log(1.1) & \log(\$50,000)+\log(1.1) \\
 & \circ & \circ & \circ \\
 \log(\$50,000)-2\log(1.1) & \log(\$50,000) & \log(\$50,000)+2\log(1.1) \\
 250 & 500 & 250 \\
 \text{people} & \text{people} & \text{people}
 \end{array}$$

And now, both the symmetry of the values (in logs) and the symmetry of the counts (in people) are restored.

So, back to the question: How do I reconcile the argument with the facts, the argument that says data should be symmetrical with the fact that data usually are not symmetrical? I reconcile the two by asserting that the data usually *are* symmetrical. But to see the symmetry you have to express the data in units compatible with the process.

If the process is multiplying people incomes or dividing them, then represent the process in logarithms: In logarithms, equal intervals in terms of the logs will correctly represent equal multipliers in terms of the process. And, more interesting: *If* a process looks symmetrical when it is examined in terms of logs, then I infer that the process was symmetrical with respect to multiples.

(Tukey, Chapter 3.) Homework: Look at the distribution of gross national products per capita, by nation. You have the data. And you have the methods for checking for symmetry. So, I ask you, are these data symmetrical in terms of dollars? Are these data symmetrical in terms of log dollars?

And, going further, do the numbers, Tukey style: Using dollars, does the Tukey analysis suggest that some of these nations are not just wealthier than others but different in kind (i.e., beyond the fences)?

Using log dollars, does the Tukey analysis suggests that some of these nations are not just wealthier than others but different in kind (i.e., beyond the fences)? Using different scales — calibrating the Galton board that sorted these nations, but calibrating it in the two different scales, you get two different answers to the last question. Show the two answers. Discuss the discrepancy. And then, practice looking at the world the way I look at it: Argue why someone should take the second interpretation (based on logs) as the correct interpretation. Convince a skeptic.

Thinking About Intervals Using the Tools of Elementary Calculus

One way to understand the transformations is to state a simple question and then use the calculus to derive the answer — which is a transformation.

Here's the question: I have a variable, x , which changes from case to case. I imagine some cause, c , though I do not assume that I actually know what this cause might be. And I want to look at changes in x related to changes in c .

If I want simple changes in x , there is no problem. I just look at

$$\frac{x(c') - x(c)}{c' - c}$$

And you should recognize from definitions used in elementary calculus, if I look for the limiting form of the relation between x and c as c' approach c , then this thing becomes the simple derivative for x as a function of c .

$$\frac{dx(c)}{dc} = \lim_{c' \rightarrow c} \frac{x(c') - x(c)}{c' - c}$$

Thus the derivative, of the calculus, is a device for expressing simple comparisons.

Now suppose I want to qualify the changes in x by referring them to some other value. For example, suppose I wish to qualify changes in x by comparing them to the size of x itself. Can I find a new variable y such that simple changes in y act like these qualified changes in x ?

I can state that question as an equation: Is there a y such that simple changes in y correspond to qualified changes in x ?

$$\frac{dy(c)}{dc} = \frac{\frac{dx(c)}{dc}}{x(c)}$$

Fortunately, the equation has a solution. So the answer is “Yes”. The answer uses one of the first differential equations in introductory calculus: Simplifying the equation, it says.

$$dy = \frac{dx}{x}$$

And this differential equation has the solution

$$y = \ln(x)$$

So the answer is, “Yes, use the logarithm of x instead of x itself.”

For the data analyst this has two tactical applications. First, if you want a variable that acts like another variable — but weighted according to the size of the values that are changing, then switch from the original variable to the logarithm of the original variable. (Exercise to the reader: It does not matter which base you use for your logarithms, as long as you are consistent. Prove it.)

Second, the same logic works in reverse: In reverse, suppose I know empirically that the logarithm of a variable is well behaved. I have to ask why: What does it mean when the logarithm of a variable is well-behaved? I answer this question by reverse engineering problem: Knowing that the logarithm is well behaved, what does this tell me about the original variable whose logarithm is well behaved? It tells me that I should be looking at weighted changes, weighted in proportion to size, not simple change.

Generalizing to Other Transformations

Square Root

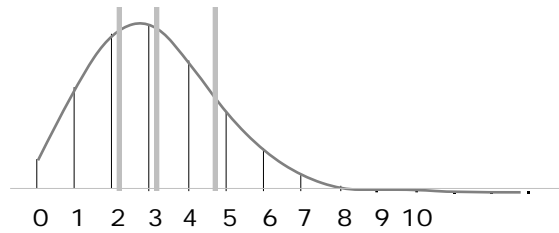
Empirically, counts of objects, tend to have a predictable behavior. Suppose that we are counting the number of people who have incomes between \$50,000 and \$100,000. Let me suppose that in the general population the number of people in this income category is unknown — some percent of the total. And let me suppose that the data available provides a sample of 1,500 people from the general population. In that sample the number of people with incomes between \$50,000 and \$100,000 is probably not *exactly* 10%. It is usually a little bit high or a little bit low.

Suppose that another sample of 1,500 becomes available. Again the number of people with incomes between \$50,000 and \$100,000 will probably not be exactly 10%. It is usually a little bit high or low.

And suppose that yet another sample becomes available. Eventually, with more and more samples, the count will trace a distribution. There will be an average count and there will be a standard deviation for the counts.

So what is the true percentage of the population within this income category? We still don't know. But we can use the mean of the counts computed in these separate samples to estimate the percentage of the general population within this income category?

Both experience and statistical theory tell us certain things about the distribution of counts. Experience tells us that it is likely to have a long tail. And statistical theory tell us that the shape is likely to follow what is called a Poisson distribution. Schematically, it will look something like this.



This is predictable, but it is not “well-behaved” in the specific meaning of that phrase. (It is not symmetrical.)

Now suppose I want to compare two counts: Perhaps I have the count of people in this income category in one year and I want to compare it to the count of people in this income category in another year. Or, perhaps I have the count of people in this income category who are also college educated and I want to compare it to the count of people in this income category who have only a high school degree.

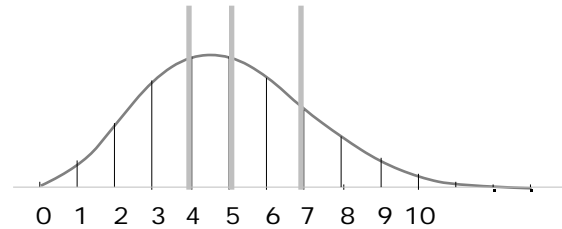
How do I compare the counts? The first cut at a comparison is simple: Subtract. That will tell you pretty quickly whether one count is greater than another and how much?

But how big a difference between two counts is a *big* difference? This is not so simple. Suppose that the difference is 2? In the sketch, I’ve assumed that the mean was three for the counts, and sketched-in three vertical lines for the median and the two quartiles. How big is a difference of “2”? If it is 2 above (if the count was 5), then this is a moderately big difference, slightly more than a quartile away. If it is 2 below (if the count was 1), then this is a big difference, much more than a quartile away.

So is “2” a big difference? It depends, 2 going up is less impressive than 2 going down. “2” at one part of the scale is not the same as “2” at another. That means for us, for those of us who have to interpret these numbers the intervals we are interested are not the intervals in which the data are being measured. That is one of the penalties for trying to

work with a variable that is not well-behaved, specifically the penalty for working with a variable that is not symmetrical.

It gets worse. Suppose we have a couple of samples, each of which gives us a number for the second count. Suppose that the mean for these counts for the second group is five. The distribution in this case would look approximately like this



Now, how big is a difference of “2”? The answer is different when this second distribution is used as a reference. So how big is “2”? Well it depends on whether you are going up or going down (asymmetry) and it depends on which distribution you are comparing it to because the variation is different in the two distributions (heteroscedasticity). That is another penalty we pay for failing to work with a well-behaved variable.

So, I want a transformation that is well-behaved. I also know, both empirically and from statistical theory that the standard deviation of a count (or a Poisson distribution) is equal to the square root of its mean. Let me look for a new unit of measure whose simple changes act like changes of counts qualified by comparison to their square roots.

$$dy = \frac{dx}{\sqrt{x}}$$

Solving the equation it tells me to use y as negative two times the square root of x and since the proportionality will not affect the behavior of the result I will use simply y equals the square root of x .

$$y = \sqrt{x}$$

So, with counts, try the square root transformation. If you want a variable that acts like another variable — but weighted according to the square root of the values that are changing, then switch from the original variable to the square root of the original variable. (Exercise to the reader: It does not matter whether you use $y = -2\sqrt{x}$ which is the solution to the equation or change the constant of proportionality to use $y = \sqrt{x}$, as long as you are consistent. Prove that if the transformation that is proportional to the square root gives you a unit of measure that is well behaved, then the simple square root itself will also be well behaved of these square root transformations is well-behaved.)

And in reverse, what does it mean when the square root of a variable is well-behaved? I answer this question by reverse engineering problem: Knowing that the square root is well behaved, I should be think that changes of the original variable had to be weighted in proportion to their square roots. So, the original variable is acting like a count.

Postscript on More General Transformations

The logic of this equation can lead to less commonly used transformations. The logic can lead to the inverse, but it is simpler to think of the inverse directly: The inverse of physicians per person is persons per physician. The inverse of time to completion (e.g., the time it takes a runner to complete a mile) is velocity: The inverse of 4 minutes per mile is 15 miles per hour.

The cases we have looked at have had a meaningful minimum at one end: zero people, zero doctors, zero counts, zero velocity. Another type of variable has a meaningful boundary at both ends. For example, what percent of a population is literate? The number is guaranteed to be bounded by 0 at one end and by 100 at the other. So you might wish to count a change from 1 percent literate to 2 percent literate to be a big change, doubling the literacy. By comparison changing the literacy from 50 percent literate to 51 percent literate is probably of little (relatively little) importance. By comparison again, changing the literacy rate from 98 percent to 99 percent is a difficult step, halving the number of illiterates.

By analogy, the equation for logs is comparing x to its lower bound.

$$dy = \frac{dx}{x - \text{lower bound}}$$

Where there are two bounds, the equation becomes

$$dy = \frac{dx}{(x - \text{lower bound})(\text{upper bound} - x)}$$

and the solution becomes

$$y = \log(x - \text{lower bound}) - \log(\text{upper bound} - x)$$

with percentages

$$y = \log(x) - \log(100 - x)$$

and with probabilities

$$y = \log(x) - \log(1 - x)$$

This is useful for data which have either mathematical limits, like percentages and probabilities or systemic limits where “no” production establishes a lower bound and the capacity of a system determines an upper bound.