

One's bed is "unreal." The Idea of the bed, existing eternally in some distant empyrean, is the true reality. Any bright Athenian could have made the obvious objection to this stratospheric nonsense.

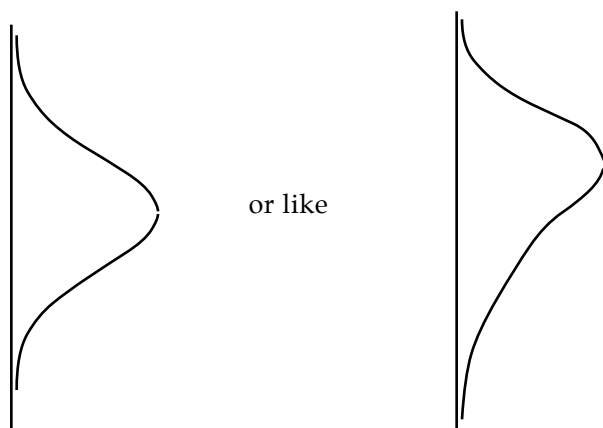
I. F. Stone, *The Trial of Socrates*, p. 73.

deviate: ... to turn aside (from a course direction, standard, doctrine, etc.) ...

Webster's New World Dictionary of the American Language.

Things Vary

The moral of the story that follows is simple. The moral is "Things vary." There is a pleasant feeling of certainty to wrapping up your data with one definitive number: "The average income is \$50,000." "The average number of people per physician is 512." That's it, a neat clean description of reality. But, reality is not that neat. Numbers vary. Data vary. It is almost guaranteed that the data are a lot messier than anything that could be reported by a single number. Look at the stem and leaf for a set of data and you will see a shape: Generally, you will see something with a central "hump" that we can think of, roughly, as the center of the distribution. Schematically, which is to say smoothing over the roughness of real data, the hump may be in the middle of the range or closer to one end, like



The average, either the median or the mean, puts a number on the location of the center: “The median income of this population is \$50,000.” “The mean income of this population is \$60,000.” But we can do better by thinking of the “center” as a range of data, a hump that may be tightly wrapped around the average, or loosely spread. Generically, this tightness of looseness around the center is called “variation” and it makes a great difference to your analysis of data.

Measure your height, measure your weight, measure these things again and again, and the answers will vary. This is not because you used a bad yardstick or a cheap scale. It is because things vary.

I don’t understand just why variation, not constants, are the reality of our experience, but the idea that there is such a thing as *the answer* is just that: an idea. By contrast, what we know for sure, the evidence before our eyes, our experience, is variation.

For example, consider the weight of the standard 10 gram weight that resides at the United States Bureau of Standards. What does it weigh? That should be simple enough, a question

shorn of the usual subtleties plaguing measurement in the social and physical sciences. Or is it?

The US 10 gram weight is not “the” 10 gram weight. The U.S. standard is a copy of the International standard 10 gram weight in France. It is an imperfect copy, a little bit lighter than the original. But what does it weigh?

Table 1 records observations made by the U.S. Bureau of Standards itself (presented by David Freedman in *Statistics*¹). Imagine the resources of the Bureau of Standards — fully capable of commanding whatever resources it takes to do the job — which in this case is to measure the weight of this little piece of metal.² And the answer? It varies. The figure shows one hundred observations (and nine different answers).

How much does it weigh? The measurements exhibit the distribution of values shown in Figure 1. It is *about* 0.4 milligrams light. *Most* of the measurements lie in a range between 9.99950 and 9.99960 grams, a range of uncertainty equivalent in weight to the weight of about half a centimeter (about three-sixteenths of an inch) of human hair.

1 *Statistics*, Second Edition, by Freedman, Pisani, Purves, and Adhikari, Norton, 1991, page 93.

2 The US 10 gram weight is not “the” 10 gram weight. The U.S. standard is a copy of the International standard 10 gram weight in France. It is an imperfect copy, a little bit lighter than the original. But what does it weigh?

Item	Weight in Grams	Difference: Weight in Grams Minus 10 Grams	Item	Weight in Grams	Difference: Weight in Grams Minus 10 Grams
1	9.999591	-0.000409	51	9.999596	-0.000404
2	9.999600	-0.000400	52	9.999594	-0.000406
3	9.999594	-0.000406	53	9.999593	-0.000407
4	9.999601	-0.000399	54	9.999595	-0.000405
5	9.999598	-0.000402	55	9.999589	-0.000411
6	9.999594	-0.000406	56	9.999590	-0.000410
7	9.999599	-0.000401	57	9.999590	-0.000410
8	9.999597	-0.000403	58	9.999590	-0.000410
9	9.999599	-0.000401	59	9.999599	-0.000401
10	9.999597	-0.000403	60	9.999598	-0.000402
11	9.999602	-0.000398	61	9.999596	-0.000404
12	9.999597	-0.000403	62	9.999595	-0.000405
13	9.999593	-0.000407	63	9.999608	-0.000392
14	9.999598	-0.000402	64	9.999593	-0.000407
15	9.999599	-0.000401	65	9.999594	-0.000406
16	9.999601	-0.000399	66	9.999596	-0.000404
17	9.999600	-0.000400	67	9.999597	-0.000403
18	9.999599	-0.000401	68	9.999592	-0.000408
19	9.999595	-0.000405	69	9.999596	-0.000404
20	9.999598	-0.000402	70	9.999593	-0.000407
21	9.999592	-0.000408	71	9.999588	-0.000412
22	9.999601	-0.000399	72	9.999594	-0.000406
23	9.999601	-0.000399	73	9.999591	-0.000409
24	9.999598	-0.000402	74	9.999600	-0.000400
25	9.999601	-0.000399	75	9.999592	-0.000408
26	9.999603	-0.000397	76	9.999596	-0.000404
27	9.999593	-0.000407	77	9.999599	-0.000401
28	9.999599	-0.000401	78	9.999596	-0.000404
29	9.999601	-0.000399	79	9.999592	-0.000408
30	9.999599	-0.000401	80	9.999594	-0.000406
31	9.999597	-0.000403	81	9.999592	-0.000408
32	9.999600	-0.000400	82	9.999594	-0.000406
33	9.999590	-0.000410	83	9.999599	-0.000401
34	9.999599	-0.000401	84	9.999588	-0.000412
35	9.999593	-0.000407	85	9.999607	-0.000393
36	9.999577	-0.000423	86	9.999563	-0.000437
37	9.999594	-0.000406	87	9.999582	-0.000418
38	9.999594	-0.000406	88	9.999585	-0.000415
39	9.999598	-0.000402	89	9.999596	-0.000404
40	9.999595	-0.000405	90	9.999599	-0.000401
41	9.999595	-0.000405	91	9.999599	-0.000401
42	9.999591	-0.000409	92	9.999593	-0.000407
43	9.999601	-0.000399	93	9.999588	-0.000412
44	9.999598	-0.000402	94	9.999625	-0.000375
45	9.999593	-0.000407	95	9.999591	-0.000409
46	9.999594	-0.000406	96	9.999594	-0.000406
47	9.999587	-0.000413	97	9.999602	-0.000398
48	9.999591	-0.000409	98	9.999594	-0.000406
49	9.999596	-0.000404	99	9.999597	-0.000403
50	9.999598	-0.000402	100	9.999596	-0.000404

Figure 1 (Facing)

One hundred measurements of the weight of the U.S. Ten Gram Weight. (Statistics, Second Edition, by Freedman, Pisani, Purves, and Adhikari, 1991, Norton, page 93.)

Weight Interval in Grams	Distribution
9.999625	
9.999620	
9.999615	
9.999610	
9.999605	
9.999600	
9.999595	
9.999590	
9.999585	
9.999580	
9.999575	
9.999570	
>=9.999565 to <9.999570	
>=9.999560 to <9.999565	

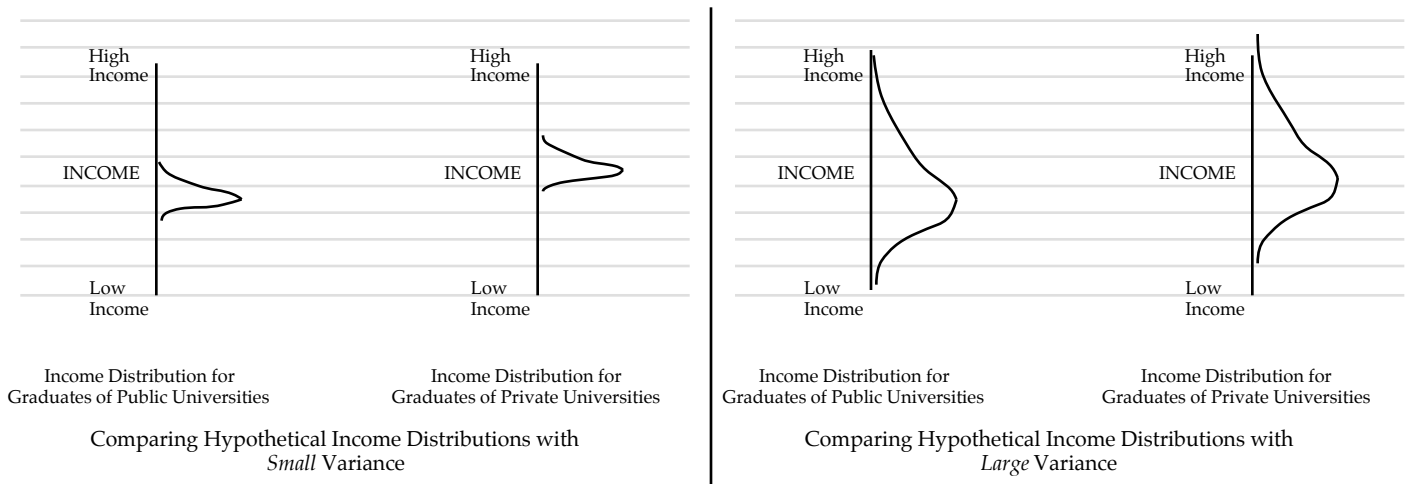
Table 1
Histogram of One Hundred Measurements of the 10 Gram Weight.

So here is our task: To summarize the facts displayed by a Stem and Leaf, or by a histogram, we need a number to represent the center of the variation, and we need a number (or numbers) to represent the variation.

Implicitly, you are already paying attention to variation: For example, if you had found that the variation of protein content among breakfast cereals was small, with protein content lying between a low of 2.3 grams of protein and a high of 2.4 grams of protein, tightly wrapped around the average (which it is not) then you would have concluded (from the *lack* of variation) that there was no need to pursue the data — it doesn't matter: If the variance were this small then the content of your cereal would not depend on your choice of cereal and the protein content of your breakfast cereal would depend more on the size of the bowl than on the choice of the cereal. By contrast, in fact, you found that the variation of protein content was large, with protein content ranging from a low of 0.8 grams of protein to a high of 10.2 grams of protein — which led us forward in search of an explanation.

If I were comparing the personal incomes of two groups of people, then again I would have to pay attention to the variation. For example suppose, without real data, that the average income of college graduates from private universities exceeded the average income of graduates from public universities. Even assuming that that is a fact, my evaluation of that fact would depend on the variation: If the variation of incomes in each of the two groups is small, corresponding to the two drawings in Figure 1a, then the difference between public schools and private schools is worthy of examination: Because the variation is small, most of the people in the second group have higher incomes than most of the people in the first group, which would force you to conclude that the difference between the two groups is to be taken seriously. By contrast, if the variation of incomes in each of the two groups is large, corresponding to the two drawings in Figure 1b, then you would

conclude that the difference between public schools and private schools is real, but small compared to the overall variation of income. In each case, Figure 1a and Figure 1b, the contrast between the two averages is the same, but your evaluation of these averages depends on the variation.



To get this information out of the picture and put it into numbers, we think of the center as a range of data tightly or loosely spread around the average. What remains to be said is exactly what range of data we will report as the “center”.

The Median and the Mean as Centers

I would like you to pretend, for the moment, that you had never heard of a mean or a median and were facing, as if for the first time, the problem of coming up with a number to that represents the center of a collection of numbers. How do you do it? Statisticians have created not so much an answer to this question as they have created a strategy capable of producing answers — sometimes different answers for different occasions.

So I ask — what is the best measure for the center of a distribution of measurements. Specifically, very specifically, what do I mean by “best”?

Measuring The Center

Well, crudely, the best measure of the center should be close to all of the values in the data. And, of course, that leaves me with the problem of defining close: I'll say that the center, c , is close to the data if the difference between each data point and the center at c , " $x_i - c$ ", is consistently small. Actually, this is a bad definition, but I'll follow it out because I want you to see the process of inventing (or re-inventing) the average. So, I define the total deviation from the center c , $V(c)$, by summing up these differences between the data and the center and I write the total variation around each possible value of the center, c , as

$$V(c) = \sum_{i=1}^n (x_i - c) \quad \begin{array}{l} \text{First definition of deviation as a function of } c. \\ \text{Total Variation} \end{array}$$

There is one immediate objection to this measure of deviation: Using this measure a data set with many observations (large " n ") will always look worse than a data set with a small number of observations — because n is larger and more deviations get added together into this V .

So, I'll do better, this time correcting for different sizes of " n ".

$$V(c) = \frac{1}{n} \sum_{i=1}^n (x_i - c) \quad \begin{array}{l} \text{Second definition of deviation as a function of } c. \\ \text{n-adjusted Variation} \end{array}$$

Is this a good definition? I'll test it by example. Suppose I have three numbers $x_1=10$, $x_2=11$, and $x_3=12$. And suppose I choose 10 as the center. "10" is not the center of 10, 11, and 12, but let me suppose that it is and check out its effect on V . Here, if $c=10$, $V(c)=1$

Measuring The Center

		Center, c , Equal 10
i	x_i	Difference Between x_i and c
1	10	$10-10= 0$
2	11	$11-10= 1$
3	12	$12-10= 2$
		$V(c) = 1$

Can I find a better center, a center about which there is less variation? Certainly. Choosing 11 as the center the variation around this center, $V(c)$, is zero. Better. That looks promising. Is there any other "center" around which the variation would be even less?

		Center, c , Equal 11
i	x_i	Difference Between x_i and c
1	10	$10-11=-1$
2	11	$11-11= 0$
3	12	$12-11= 1$
		$E(c) = 0$

Unfortunately, yes, there is. Suppose I had tried $c = 12$. That is obviously a bad choice, 12 is not the center of these data, but what does the measure of variation around this center have to say? It says "-1" which may be ridiculous but it is certainly smaller than zero. So, this definition of "error", of the error that results from choosing c as the

Measuring The Center

center, produces ridiculous results — implying that 12 is more central than 11. So — out with the definition. I need a better one.

		Center, c, Equal 12
i	x_i	Difference Between x_i and c
1	10	$10 - 12 = -2$
2	11	$11 - 12 = -1$
3	12	$12 - 12 = 0$
		$V(c) = -1$

There are several ways of fixing up the definition, using alternative expressions of what it means for data to be “close” to their center. Suppose I fix up what I did above by saying “No, close is a matter of *distance* not *difference*. I should have used the *distance* between x_i and the center, not the difference.” That places “10” a distance of one unit away from 11 and it also places “12” a distance of one unit away from 11. Using this new working definition of variation

$$V(c) = \frac{1}{n} \sum_{i=1}^n |x_i - c|$$

Third definition of deviation as a function of c:
Absolute Deviation

Measuring The Center

		c = 10	c=11	c=12
i	x_i	Distance Between x_i and c	Distance Between x_i and c	Distance Between x_i and c
1	10	$ 10 - 10 = 0$	$ 10 - 11 = 1$	$ 10 - 12 = 2$
2	11	$ 11 - 10 = 1$	$ 11 - 11 = 0$	$ 11 - 12 = 1$
3	12	$ 12 - 10 = 2$	$ 12 - 11 = 1$	$ 12 - 12 = 0$
		$V(c) = 1$	$V(c) = 2/3$	$V(c) = 1$

Sure enough, among these three choices, $c=10$, $c=11$, and $c=12$, 11 is the center with respect to which the variation is smallest. Among these three choices, 11 is the best center.

Minimum Absolute Deviation

In the jargon of the trade, I have found the center “in the sense of minimum absolute deviation” (sometimes abbreviated MAD, *Minimum Absolute Deviation*). And the result is, I think, exquisite: I have just defined the median. What’s beautiful about it is that I haven’t said anything about rank ordering the data, or splitting it in half. I haven’t said anything about the median itself. I just defined a measure of “goodness of fit”, specifically, minimum absolute deviation, and I said “find the number that is close to the data, close in the sense of minimum absolute deviation.” That turns out to be the median. It takes a little bit of calculation to prove that, but it is true and it places the median in context: The median is the “best” measure of the center, “best” in the sense of MAD.

Least Squares

That is not the end of it: The same strategy is able to create other results when it is combined with another definition of “close”. The most widely used measure fixes up the difference (definition 1 and 2) by using squares to get rid of the negatives.

$$V(c) = \frac{1}{n} \sum_{i=1}^n (x_i - c)^2$$

Fourth definition of deviation as a function of c:
Squared Deviation

		c=10	c=11	c=12
i	x _i	Squared Distance Between x _i and c	Squared Distance Between x _i and c	Squared Distance Between x _i and c
1	10	(10 - 10) ² =0	(10 - 11) ² =1	(10 - 12) ² =4
2	11	(11 - 10) ² =1	(11 - 11) ² =0	(11 - 12) ² =1
3	12	(12 - 10) ² =4	(12 - 11) ² =1	(12 - 12) ² =0
		V(c)=2.5	V(c)= .667	V(c)=2.5

This too implies that 11 is the center for these hypothetical data. In the jargon of the trade I have found the center “in the sense of least squares”. And this logic leads to the mean. The mean is the central value of a distribution “in the sense of least squares”.

Homework:

In order of increasing difficulty:

1 I have asserted that *if* variation is defined in the sense of least squares *then* the best number for the center of the variation is the mean. Given the general statement that the variation around a center is minimized when that center is the mean, how small is that minimum variation. That is, what is the value of $V(c)$, V as a function of c , when c is equal to the mean?

$$V(c) = \frac{1}{n} \sum_{i=1}^n (x_i - c)^2$$

To see the answer (to see the reason for the question) more clearly, the better question is 'what is the *mean* value of the variation around c when c is the mean?' (Substitute \bar{x} into the equation for V and solve for $\sqrt{\frac{1}{n}V}$. To see the answer more clearly, when c is the best value (when c is equal to the mean) what is the square root of the mean value of V

2 Using calculus, prove that variance in the sense of least squares, V , is minimized when $c = \bar{x}$ (Differentiate $V(c)$ as a function of c . Set the derivative equal to zero. Solve for c .)

3 Using whatever you can improvise, *prove* that minimum absolute deviation is achieved when c equals the median.

Measuring The Variation

What you have seen is an example of the way statisticians have to be explicit: You can't just say what is the "best"? Not "what is the best way to represent these data?" That's not enough. You have to specify in what sense it is the best.

Once that is done, the rest is "easy": The best average for the data, best in the sense of minimum absolute deviation, is the median. The best average for the data, best in the sense of least squares, is the mean. You will see this strategy at work throughout statistics — when you need the "best" estimate of something.

And how large is the variation? We know that the data varies, but how much? The answer, or answers, to that question are already ordained since the definition of the center used a definition of variation.

So, in the sense of minimum absolute deviation, MAD:

What is the typical variation around the center? We represent it by average variation around the average or, more specifically, by median of deviations above and below the median. These are the quartiles. Again — I haven't defined this procedurally, telling you how to rank order your data and find quartiles (that will follow). I've given you a strategy whose logical consequence is the quartiles.

And, in the sense of least squares:

What is the typical variation around the center? We represent it by average variation around the average or, more specifically, by mean deviation around the mean. This is defined as the variance. Unfortunately, squared deviation is a little rough on the human intuition, squared grams of protein for example. So we also define the "standard deviation" as the

square root of the variance — putting the deviation in terms that the human intuition can handle.

Variance:

$$V(c) = \frac{1}{n} \sum_{i=1}^n (x_i - c)^2$$

Standard Deviation:

$$s_x(c) = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - c)^2}$$