HOW TO LIE WITH STATISTICS

Darrell Huff

Copyright 1954

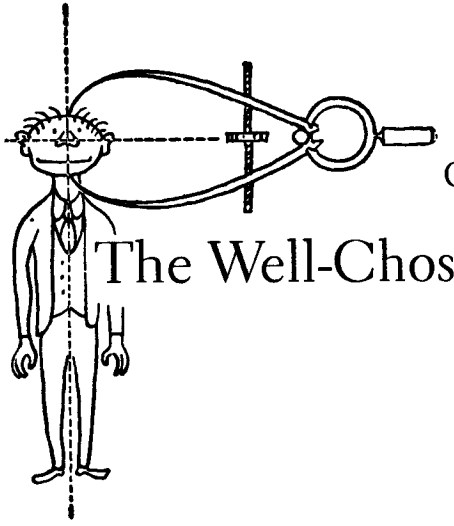
Over Half a Million Copies Sold—
An Honest-to-Goodness Bestseller

You have pretty fair evidence to go on if you suspect that polls in general are biased in one specific direction, the direction of the *Literary Digest* error. This bias is toward the person with more money, more education, more information and alertness, better appearance, more conventional behavior, and more settled habits than the average of the population he is chosen to represent.

You can easily see what produces this. Let us say that you are an interviewer assigned to a street corner, with one interview to get. You spot two men who seem to fit the category you must complete: over forty, and urban. One is in clean clothes, neat. The other is dirty and he looks surly. With a job to get done, you approach the more likely-looking fellow, and your colleagues all over the country are making similar decisions.

Some of the strongest feeling against public-opinion polls is found in liberal or left-wing circles, where it is rather commonly believed that polls are generally rigged. Behind this view is the fact that poll results so often fail to square with the opinions and desires of those whose thinking is not in the conservative direction. Polls, they point out, seem to elect Republicans even when voters shortly thereafter do otherwise.

Actually, as we have seen, it is not necessary that a poll be rigged—that is, that the results be deliberately twisted in order to create a false impression. The tendency of the sample to be biased in this consistent direction can rig it automatically.



CHAPTER 2

The Well-Chosen Average

You, I trust, are not a snob, and I certainly am not in the real-estate business. But let's say that you are and I am and that you are looking for property to buy along a road that is not far from the California valley in which I live.

Having sized you up, I take pains to tell you that the average income in this neighborhood is some $15,000 a year. Maybe that clinches your interest in living here; anyway, you buy and that handsome figure sticks in your mind. More than likely, since we have agreed that for the purposes of the moment you are a bit of a snob, you toss it in casually when telling your friends about where you live.

A year or so later we meet again. As a member of some taxpayers' committee I am circulating a petition to keep

the tax rate down or assessments down or bus fare down. My plea is that we cannot afford the increase: After all, the average income in this neighborhood is only $3,500 a year. Perhaps you go along with me and my committee in this—you're not only a snob, you're stingy too—but you can't help being surprised to hear about that measly $3,500. Am I lying now, or was I lying last year?

You can't pin it on me either time. That is the essential beauty of doing your lying with statistics. Both those figures are legitimate averages, legally arrived at. Both represent the same data, the same people, the same incomes. All the same it is obvious that at least one of them must be so misleading as to rival an out-and-out lie.

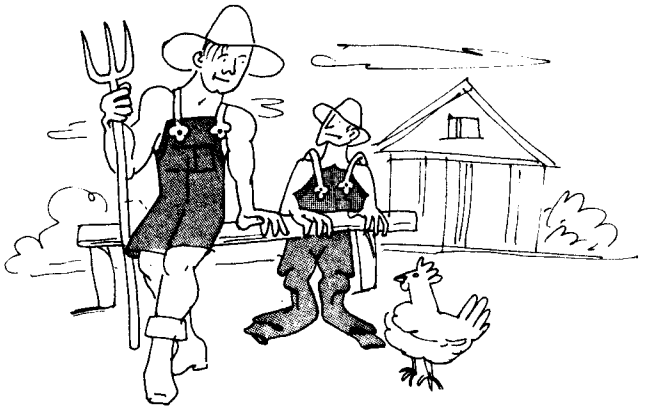
My trick was to use a different kind of average each time, the word "average" having a very loose meaning. It is a trick commonly used, sometimes in innocence but often in guilt, by fellows wishing to influence public opinion or sell advertising space. When you are told that something is an average you still don't know very much about it unless you can find out which of the common kinds of average it is—mean, median, or mode.

The $15,000 figure I used when I wanted a big one is a mean, the arithmetic average of the incomes of all the families in the neighborhood. You get it by adding up all the incomes and dividing by the number there are. The smaller figure is a median, and so it tells you that half the families in question have more than $3,500 a year and half have less. I might also have used the mode, which is the most frequently met-with figure in a series. If in

this neighborhood there are more families with incomes of $5,000 a year than with any other amount, $5,000 a year is the modal income.

In this case, as usually is true with income figures, an unqualified "average" is virtually meaningless. One factor that adds to the confusion is that with some kinds of information all the averages fall so close together that, for casual purposes, it may not be vital to distinguish among them.

If you read that the average height of the men of some group of people is only five feet, you get a fairly good idea of the stature of these people. You don't have to ask whether that average is a mean, median, or mode; it would come out about the same. (Of course, if you are in the business of manufacturing overalls for these men you would want



more information than can be found in any average. This has to do with ranges and deviations, and we'll tackle that one in the next chapter.)

The different averages come out close together when you deal with data, such as those having to do with many human characteristics, that have the grace to fall close to what is called the normal distribution. If you draw a curve to represent it you get something shaped like a bell, and mean, median, and mode fall at the same point.

Consequently one kind of average is as good as another for describing the heights of men, but for describing their pocketbooks it is not. If you should list the annual incomes of all the families in a given city you might find that they ranged from not much to perhaps $50,000 or so, and you might find a few very large ones. More than ninety-five percent of the incomes would be under $10,000, putting them way over toward the left-hand side of the curve. Instead of being symmetrical, like a bell, it would be skewed. Its shape would be a little like that of a child's slide, the ladder rising sharply to a peak, the working part sloping gradually down. The mean would be quite a distance from the median. You can see what this would do to the validity of any comparison made between the "average" (mean) of one year and the "average" (median) of another.

In the neighborhood where I sold you some property the two averages are particularly far apart because the distribution is markedly skewed. It happens that most of your neighbors are small farmers or wage earners employed in

a near-by village or elderly retired people on pensions. But three of the inhabitants are millionaire week-enders and these three boost the total income, and therefore the

arithmetic average, enormously. They boost it to a figure that practically everybody in the neighborhood has a good deal less than. You have in reality the case that sounds like a joke or a figure of speech: Nearly everybody is below average.

That's why when you read an announcement by a corporation executive or a business proprietor that the average pay of the people who work in his establishment is so much, the figure may mean something and it may not. If the average is a median, you can learn something significant from it: Half the employees make more than that; half make less. But if it is a mean (and believe me it may be that if its nature is unspecified) you may be getting nothing more revealing than the average of one $45,000 income—the proprietor's—and the salaries of a crew of underpaid workers. "Average annual pay of $5,700" may conceal both the $2,000 salaries and the owner's profits taken in the form of a whopping salary.

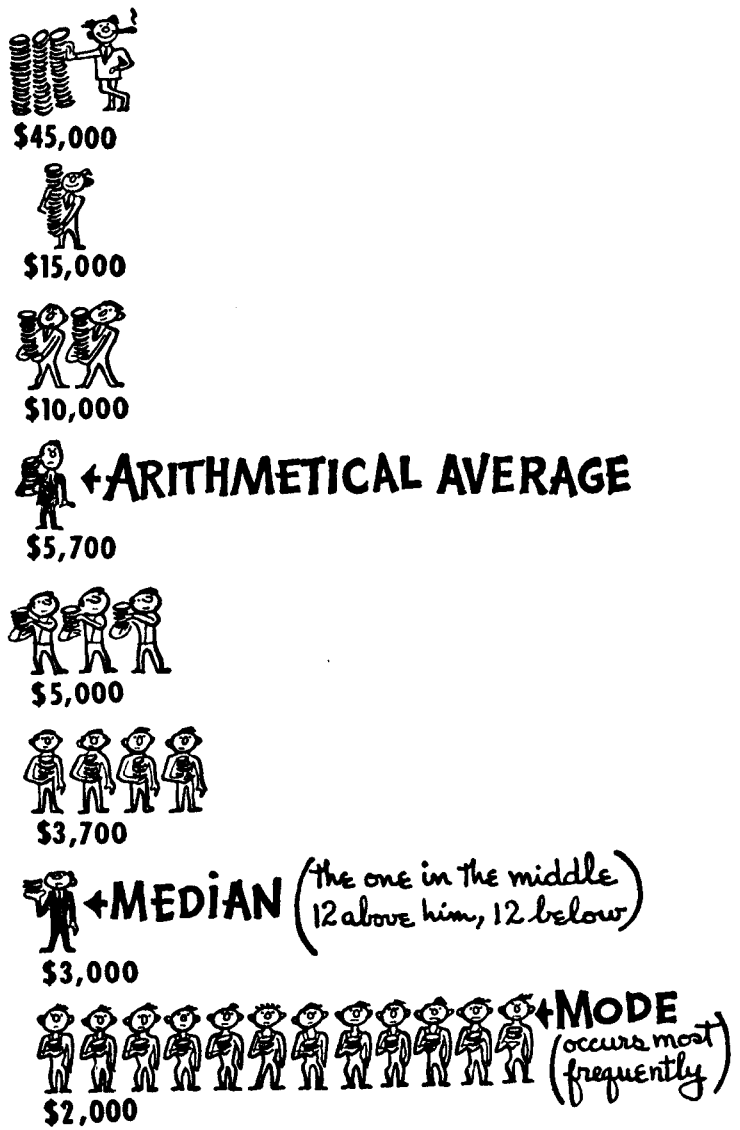
Let's take a longer look at that one. The facing page shows how many people get how much. The boss might like to express the situation as "average wage $5,700"—using that deceptive mean. The mode, however, is more revealing: most common rate of pay in this business is $2,000 a year. As usual, the median tells more about the situation than any other single figure does; half the people get more than $3,000 and half get less.

How neatly this can be worked into a whipsaw device in which the worse the story, the better it looks is illustrated in some company statements. Let's try our hand at one in a small way.

You are one of the three partners who own a small manufacturing business. It is now the end of a very good year. You have paid out $198,000 to the ninety employees who do the work of making and shipping the chairs or whatever it is that you manufacture. You and your partners have paid yourselves $11,000 each in salaries. You find there are profits for the year of $45,000 to be divided equally among you. How are you going to describe this? To make it easy to understand, you put it in the form of averages. Since all the employees are doing about the same kind of work for similar pay, it won't make much difference whether you use a mean or a median. This is what you come out with:

Average wage of employees	$ 2,200
Average salary and profit of owners	26,000

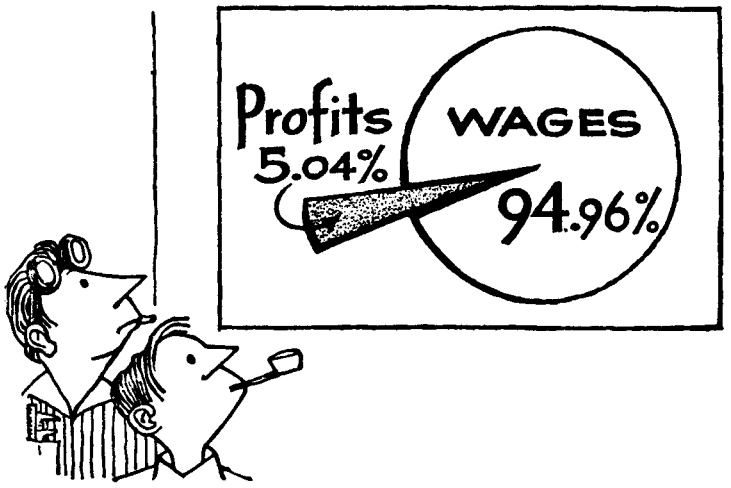
That looks terrible, doesn't it? Let's try it another way.



Take $30,000 of the profits and distribute it among the three partners as bonuses. And this time when you average up the wages, include yourself and your partners. And be sure to use a mean.

Average wage or salary	$ 2,806.45
Average profit of owners	5,000.00

Ah. That looks better. Not as good as you could make it look, but good enough. Less than six percent of the money available for wages and profits has gone into profits, and you can go further and show that too if you like. Anyway, you've got figures now that you can publish, post on a bulletin board, or use in bargaining.



This is pretty crude because the example is simplified, but it is nothing to what has been done in the name of accounting. Given a complex corporation with hierarchies of employees ranging all the way from beginning typist to president with a several-hundred-thousand-dollar bonus, all sorts of things can be covered up in this manner.

So when you see an average-pay figure, first ask: Average of what? Who's included? The United States Steel Corporation once said that its employees average weekly earnings went up 107 percent between 1940 and 1948. So they did—but some of the punch goes out of the magnificent increase when you note that the 1940 figure includes a much larger number of partially employed people. If you work half-time one year and full-time the next, your earnings will double, but that doesn't indicate anything at all about your wage rate.

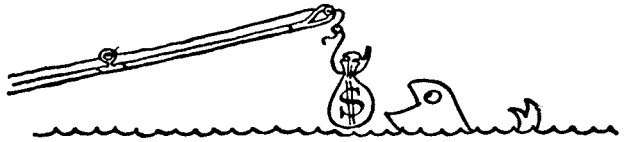
You may have read in the paper that the income of the average American family was $3,100 in 1949. You should not try to make too much out of that figure unless you also know what "family" has been used to mean, as well as what kind of average this is. (And who says so and how he knows and how accurate the figure is.)

This one happens to have come from the Bureau of the Census. If you have the Bureau's report you'll have no trouble finding the rest of the information you need right there: This is a median; "family" signifies "two or more persons related to each other and living together." (If persons living alone are included in the group the median slips to $2,700, which is quite different.) You will also learn if you read back into the tables that the figure is based on a sample of such size that there are nineteen chances out of

twenty that the estimate—$3,107 before it was rounded—is correct within a margin of $59 plus or minus.

That probability and that margin add up to a pretty good estimate. The Census people have both skill enough and money enough to bring their sampling studies down to a fair degree of precision. Presumably they have no particular axes to grind. Not all the figures you see are born under such happy circumstances, nor are all of them accompanied by any information at all to show how precise or unprecise they may be. We'll work that one over in the next chapter.

Meanwhile you may want to try your skepticism on some items from "A Letter from the Publisher" in *Time* magazine. Of new subscribers it said, "Their median age is 34 years and their average family income is $7,270 a year." An earlier survey of "old TIMErs" had found that their "median age was 41 years. . . . Average income was $9,535. . . ." The natural question is why, when median is given for ages both times, the kind of average for incomes is carefully unspecified. Could it be that the mean was



used instead because it is bigger, thus seeming to dangle a richer readership before advertisers?

You might also try a game of what-kind-of-average-are-you on the alleged prosperity of the 1924 Yales reported at the beginning of Chapter 1.

CHAPTER 3

The Little Figures That Are Not There

USERS report 23% fewer cavities with Doakes' tooth paste, the big type says. You could do with twenty-three percent fewer aches so you read on. These results, you find, come from a reassuringly "independent" laboratory, and the account is certified by a certified public accountant. What more do you want?

Yet if you are not outstandingly gullible or optimistic, you will recall from experience that one tooth paste is seldom much better than any other. Then how can the Doakes people report such results? Can they get away with telling lies, and in such big type at that? No, and they don't have to. There are easier ways and more effective ones.

The principal joker in this one is the inadequate sample—statistically inadequate, that is; for Doakes' purpose