Where do you stand? Notions of the statistical 'centre'

KEYWORDS:
Teaching;
Centrality;
Minimisation;
Mean;
Median;
Minimax.

James A Hanley and Abby Lippman
McGill University, Montreal, Canada.
e-mail: jimh@epid.lan.mcgill.ca
abbyl@epid.lan.mcgill.ca

Summary
We present a simple problem that can help highlight the decision-theoretic properties of various centrality measures. We also illustrate how the 'centrality' parameters of the classical distributions in statistics would have to be modified if we were constrained to fewer defining principles.

◆ INTRODUCTION ◆

IN MANY of our daily decisions, we implicitly use various centrality parameters. Formal courses in statistics seldom explain these parameters in the context of everyday decisions. Instead, the various centrality measures are presented as descriptive statistics and their relative merits are discussed in terms of their degree of computational ease and resistance to extreme data values. The decision theoretic properties of the various centrality parameters - and indeed the very origins of the words average, mean and median used as parameters - are seldom mentioned.

We present here a simple thought experiment, which can be supplemented with various surveys. It illustrates how these measures are, in fact, used while at the same time providing a way to teach about them. We then discuss the responses of some of our students and explore how the experimental scenarios could be changed to enrich their teaching value.

◆ THE ELEVATOR PROBLEM: ◆ A SIMPLE THOUGHT EXPERIMENT

You are delivering a heavy article to a high-rise building which is served by 3 elevators. None of them has an indicator light to show its current location, and there is but one button for all three. Once you have pushed the button, any one of the 3 elevators may respond - with equal probability. You must take whichever one arrives next. As shown by the positions of the 3 e's on figure 1, these elevators are unequally spaced along one wall. You will need to carry the heavy article to the elevator from wherever along the wall you stand and wait. You can take as long as is required to reach the elevator. Thus the distance to the elevator is your only concern.

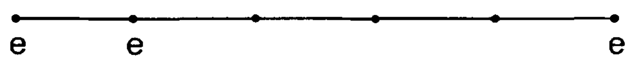

Fig 1.

1. (Without doing any formal calculations) where along the wall would you stand and wait? Mark the position with an X.

2. Why would you choose this position?

3. Would you choose differently if you were faced with this situation each day as part of your job?

   If so, indicate the spot with an O.

   Either way, explain your answer to 3.

THE RESPONSES AND WHAT ◆ THEY CAN TEACH US ABOUT ◆ DIFFERENT CENTRAL VALUES

In figure 1, units are marked off in 10s so the two rightmost elevators are 10 and 50 units from the one at the left. When we gave this problem to a convenience sample of students and staff, we found that their responses clustered in three main groups. These are shown schematically in figure 2.

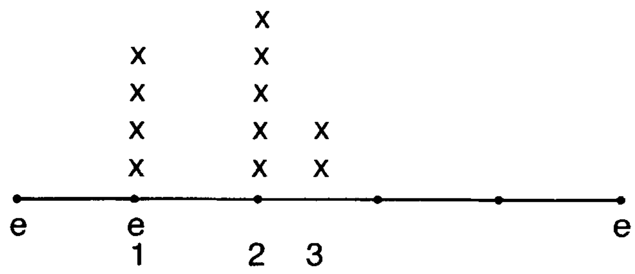

Fig 2.

1 - those who would stand at the second of the 3 elevators, i.e. 1/5th of the way from the left extreme;

2 - (the largest group) who would stand somewhat to the right of this "middle" elevator, but still left of centre, i.e. about 2/5ths of the way from the left extreme;

3 - a (smaller) group who would stand halfway between the two extremes.

To use this problem and the responses heuristically, students who chose to stand at "1" (directly at the left-of-centre "middle" elevator) might first be asked the reasons for this choice. Students in group "2", likely to be the majority, can be asked what they are trying to accomplish by choosing their position. The classroom discussion can then be steered to consider explicitly the different possible (informal statistical) criteria that have guided the decisions. What should emerge is how decisions have something to do with optimisation or minimisation: the distance one has to walk, or something about it, is being - in some way - minimised. Finally, to link these "intuitive" criteria to more formal centrality principles, students can be asked to carry out calculations to see which option actually gives better performance on average.

The average distances travelled by the various groups are shown in table 1. Those in group 2 probably thought that, by standing slightly to the right of the second elevator, they were "playing it safe". They will be surprised by the results. Many of them may be especially surprised to see that their strategy does not fit with their intention. To underscore the basis for this - to them - counterintuitive situation, they might calculate the averages again, but this time focusing on the squared distance rather than distance itself (also shown in table 1).

These results may reassure members of group 2; they may now say, "of course, we were just being a bit more cautious in protecting ourselves against large distances". But these results are also of heuristic use in illustrating that group 2 members stood at the MEAN elevator position (1/3)(0 + 10 + 50) = 20, or 2/5ths of the way from the left, and that a defining property of the mean is that it is the value about which the average SQUARED deviation is minimised.

By this time, group 1 may realise that they had positioned themselves at the MEDIAN, and that a defining property of the median is that it is the value about which the average ABSOLUTE deviation is minimised.

Discussion can now focus on the minority group 3 and the basis for their choice. They should quickly realise they were being very conservative. In fact, they took a position that minimised the MAXIMUM distance they would have to walk, thereby putting into practice the "minimax" principle. For completeness, we list the maximum distances in table 2, along with the calculations and rankings under the other two previously examined principles (that imply using the mean and median respectively).

Group/Location	Distances			Average	Squared Distances			Average
	L	M	R		L	M	R	
1	10	0	40	50/3	100	0	1600	1700/3
2	20	10	30	60/3	400	100	900	1400/3
3	25	15	25	65/3	625	225	625	1475/3

L: Leftmost M: "Middle" R: Rightmost

Table 1. (Average) distances and squared distances travelled by three groups (1,2,3) to three elevators (L,M,R)

Group/Location		**Principle**		
		Maximum Distance	Average Absolute Distance	Average Squared Distance
1	MEDIAN	40	**50/3**	1700/3
2	MEAN	30	60/3	**1400/3**
3	MINIMAX	**25**	65/3	1475/3

Table 2. Rankings of locations in relation to 3 principles (distance in **bold** is the minimum of the 3 distances in column)

◆ DISCUSSION ◆

The elevator problem has several "right" solutions. Each of these has heuristic value in uncovering some principles of centrality on which daily decisions may be made. The problem acquires further use if the simple question we presented is expanded. For instance, it could be supplemented with specific directives as to what should be optimised when a position is selected. This would allow students to assess the concordance between the way they say they behave and the principles they say they follow.

As an example of a situation where one's choice of centre can have more serious consequences, one might wish to replace the three elevators in the diagram by three communities, and ask about the placement of a vehicle that must respond to calls from these communities. One might wish to minimise the average distance or average fuel consumption, in which case the median becomes the preferred choice. If the vehicle were an ambulance and one wished to minimise the largest response time, one would choose the midpoint. But it would be more difficult to imagine an objective that implied the mean. How could one justify minimising the average *squared* distance rather than, say, the average cubed distance or the distance to the power of 1.5?

The analyses of these "where-do-you-stand" problems tend to diminish the attractiveness of the mean as a centrality parameter, and reveal the surprising performance of the median. The question might even be raised as to whether the mean is in any sense a "natural" middle. If centrality parameters had to be re-invented, would the mean be, as it is now, a privileged candidate? There is some doubt. With modern computers used for data analysis, the mean no longer need be favoured over the median just because calculating the latter requires more than a single pass through the data. Complaints that the sampling distribution of the median statistic is not easily estimated are also moot. From this perspective then, the mean becomes a parameter primarily for accountants and other "bottom-line" people who worry about collectives but not about individuals. For example, while an accountant would immediately recognise the impact of raising the average salary of professional football players by 5%, the effect of an increase presented as a percentage of the median salary would probably be hard to grasp. Even the referents of the two parameters differ: the mean refers simply to $1/n$th of the total salaries, whereas the median has a more human, even individual, referent: it is the salary of the "middlemost" *player*.

A further heuristic spin-off from a "where-do-you-stand" problem might be the opportunity to become lexical and explain the derivation of the word average from the Latin word *havaria*. *Havaria* was the premium each Roman ship owner paid towards a central insurance pool to compensate owners of trading ships that were lost. This, too, emphasises the focus on collectives.

Finally, and in the same spirit of de-throning the mean, we raise the following question. Suppose one "banned" all use of the mean, and likewise all centrality parameters which could - to a first order of recursion - be defined in terms of a mean. In these circumstances, even the median, which can be derived as the value which minimises the *mean* absolute deviation, would itself have to be banned as a centrality statistic. However, what if we were allowed to define a new centrality parameter, about which the *median* absolute deviation were minimised? Where would this new "middle" be?

This principle is not as whimsical as it might at first appear. It has direct relevance to the distribution of emergency response times, which is sometimes summarised using the mean but more often by a quantile, such as the 50th or other higher percentile. The "centres" that minimise the 50th percentile of the absolute deviations for some specific skewed distributions are given in table 3.

Distribution †	**Centrality Parameter**			
	Mode	**New***	**Median**	**Mean**
Beta[2,8]	0.125	0.151	0.180	0.2
Exponential[1]	0.000	0.346	0.693	1.0
FRatio[3,10]	0.278	0.489	0.845	1.25
Gamma[2,4]	4.000	5.049	6.713	8.00
HalfNormal[1]	0.000	0.443	0.845	1.00
LogNormal[1,1]	1.000	1.630	2.718	4.482

† Mathematica notation

* Value about which the *median* absolute deviation is minimised

Table 3. Centrality parameters - three old and one new - for selected skewed distributions