MAA
MATHEMATICAL ASSOCIATION OF AMERICA

GLOBAL MATH PROJECT
Uplifting Mathematics for All

CURRICULUM INSPIRATIONS: www.maa.org/ci

Innovative Online Courses: www.gdaymath.com
Tanton Tidbits: www.jamestanton.com

WHOA! COOL MATH!

CURIOUS MATHEMATICS FOR FUN AND JOY

May 2017

THIS MONTH'S PUZZLER

Is there a non-constant function F with the property that F of the average value of any given set of data values is sure to equal the average value of F of those data values (that is, for all a, b, c, … we have

$$F\left(\frac{a+b}{2}\right)=\frac{F(a)+F(b)}{2}$$

and

$$F\left(\frac{a+b+c}{3}\right)=\frac{F(a)+F(b)+F(c)}{3}$$

and so on)?

IDEA VERSUS REAL DATA

If I measure the height of a particular tree three times in a row then, in an ideal world, I should get the exact same measurement three times in a row: x feet, x feet, and x feet, where x is the true height of the tree. But in reality, our abilities to measure are not exact and I will likely obtain three slightly different measurements: a feet, b feet, and c feet. But we do expect those three values to each be close to some common value of x feet.

Phrasing this slightly differently, this mean we expect the set of numbers (a,b,c) to be close to an ideal set (x,x,x) for some (unknown) value x.

Can we work out the "best" value for x suggested by the data, one that gives the closest ideal point to the point of data values? This is a geometry problem.

USING THE DISTANCE FORMULA: EUCLIDEAN DISTANCE

The Euclidean geometry we study in school uses a distance formula based on Pythagoras's Theorem. Here the distance between (a,b,c) and (x,x,x) is

$$\sqrt{(x-a)^2+(x-b)^2+(x-c)^2}\ .$$

What value of x gives the smallest value for this distance?

Working with square roots is awkward. But we can avoid dealing with them by noting that the quantity above is minimal when the quantity under the square root sign is minimal. Thus it suffices to find the value of x that minimizes

$$(x-a)^2+(x-b)^2+(x-c)^2.$$

This is equivalent to the expression

$$3x^2-2(a+b+c)x+(a^2+b^2+c^2)$$

and this quadratic has minimal value at

$$x=\frac{a+b+c}{3}.$$

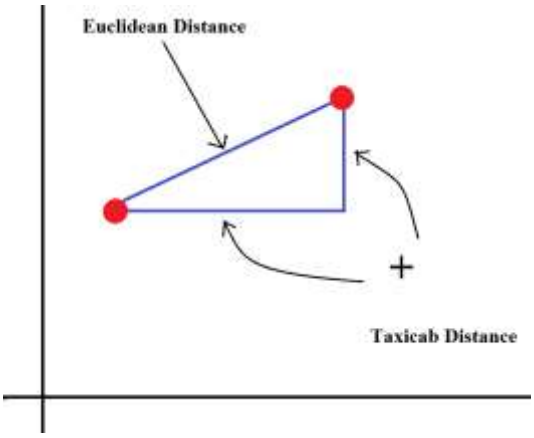
[Either use $x=-b/2a$ or use Tanton's "find two symmetrical points on a symmetrical curve" approach.]

Challenge: Show that for n data values $(a_1, a_2, ..., a_n)$ the value x that gives the closest ideal point $(x,x,...,x)$ to the point of data values, via Euclidean distance, is $x=\dfrac{a_1+a_2+\cdots+a_n}{n}$.

This shows that if we favor the standard distance formula in our lives, then the best ideal data point for a set of data values is given by the *arithmetic mean* of that data.

USING THE TAXI CAB METRIC

Another way to measure distances between points is with the taxicab metric. In a plane, this is the east/west distance between points plus the north/south distance between them. (That is, it is the length of the path a taxicab driver would follow to travel from one point to the other following a city's perpendicular streets and avenues.)



In three dimensions one adds the vertical distance between the points too.

In this context, the distance between the point of data values (a,b,c) and an ideal point (x,x,x) is

$$|x-a|+|x-b|+|x-c|.$$

Let's assume that the data is arranged in order, $a \le b \le c$. Then, for $x \le a$, this distance is given by

$a - x + b - x + c - x = a + b + c - 3x$.
(This expression represents a line of slope -3.)

For $a \le x \le b$, the distance is given by

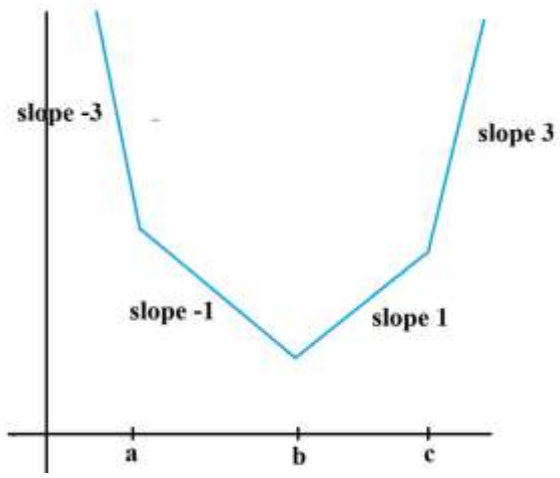
$x - a + b - x + c - x = (-a + b + c) - x$.
(This expression represents a line of slope -1.)

For $b \le x \le c$, the distance is given by

$x - a + x - b + c - x = (-a - b + c) + x$.
(This expression represents a line of slope $+1$.)

For $c \le x$, the distance is given by

$x - a + x - b + x - c = (-a - b - c) + 3x$.
(This expression represents a line of slope $+3$.)



It is clear that the minimal taxicab distance occurs if we choose $x = b$.

Challenge: a) Show that for an odd number of data values $(a_1, a_2, ..., a_n)$ arranged in increasing order, the value x that gives the closest ideal point $(x, x, ..., x)$ to the data (via the taxicab metric) is x equal to the middle value in the list.
b) Show that for an even number of data values, the taxicab distance has the same minimal value for all x between the middle two values in the data list. (You might as well take x to be the average of those two middle values.

If we favor taxicab distance formula in our lives, then the best ideal data point for a set of data is given by the *median* of that data.

GETTING WEIRD
Assume each of the data values $(a_1, a_2, ..., a_n)$ are positive numbers and we seek the closest ideal point $(x, x, ..., x)$ with x a positive number. Show that if we use the function

$$\sqrt{(\log(x) - \log(a_1))^2 + (\log(x) - \log(a_2))^2 + \cdots + (\log(x) - \log(a_n))^2}$$

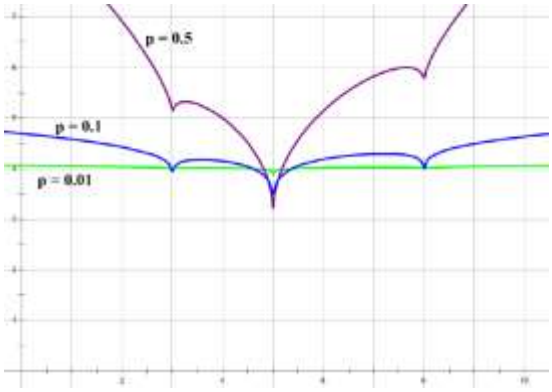
to measure distance, then the "best" value for x is $\sqrt[n]{a_1 \cdot a_2 \cdot \cdots \cdot a_n}$, the *geometric mean* of the data values.

MEAN, MEDIAN, and …

Choose a real number $p > 0$ and look at a measure of distance between a point of data values $(a_1, a_2, ..., a_n)$ and an ideal data point $(x, x, ..., x)$ given by

$$d(x) = |x - a_1|^p + |x - a_2|^p + \cdots + |x - a_n|^p$$

Here are the graphs of this function for the data set $(3,5,5,8)$ and $p = 0.5$, 0.1, and 0.01.



In each case, the minimum seems to occur at $x = 5$, the mode of the function. (One can see a slight dip in the green curve here too.) And as $p \to 0$ it seems that these functions approach, more or less, the function with constant output 4, the count of data values.

We can explain that.

For any non-zero value r we know that $r^0 = 1$. So, for any value x different from a_1, a_2, …, a_n, we have that

$$d(x) = |x - a_1|^p + |x - a_2|^p + \cdots + |x - a_n|^p$$

approaches $1 + 1 + \cdots + 1 = n$, the number of data points, as $p \to 0$.

If k of the data values a_i are the same, then, for x equal to that common data value,

$$d(x) = |x - a_1|^p + |x - a_2|^p + \cdots + |x - a_n|^p$$

is a sum of $n - k$ nonzero terms. As $p \to 0$, the sum approaches a sum of $n - k$ 1s, and so has value $n - k$.

Thus, the metric

$$d(x) = \lim_{p \to 0^+} |x - a_1|^p + \cdots + |x - a_n|^p$$

is given by

$$d(x) = \begin{cases} n & \text{if } x \text{ is different from a data value} \\ n - k & \text{if } x \text{ equals a data value with frequency } k \end{cases}$$

and so has minimum value the *mode* of the data set. (This metric is known as the Hamming Code distance: the distance between two points is simply the count of coordinates with differing entries.)

***

How fun that measures of central tendency in data can be viewed as natural ideals in different geometries. In fact, all the measures we discussed are based on minimizing the function

$$d(x) = |x - a_1|^p + |x - a_2|^p + \cdots + |x - a_n|^p$$

for different values of p. For $p = 2$, the minimum occurs at the arithmetic mean of the data, for $p = 1$,the median of the data, and for " $p = 0$," the mode of the data.

RESEARCH CORNER

1. OTHER AVERAGES?
Is there a measure of distances between points that yield the midrange of a data set as the ideal common data value? How about other classic averages of data, the quadratic mean, the harmonic mean, and the like?

2. FOR $0 < p < 1$.
Can you prove that
$d(x) = |x - a_1|^p + |x - a_2|^p + \cdots + |x - a_n|^p$
also has minimal value at the mode of the data for all $0 < p < 1$?

Care to analyze this formula for $p > 1$.

3. OPENING PUZZLER.
Certainly functions of the form
$F(x) = mx + k$ have the property described in the opening puzzler. Must all examples of continuous functions with this property have this form? Are there interesting discontinuous functions that work?

tanton.math@gmail.com