WHAT IS THE DIFFERENCE BETWEEN A MEAN AND A MEDIAN?

Much of the reporting on housing, employment, and family income statistics is often reported in the press in terms of mean and median without the authors making a clear distinction between the two. This fact can often lead to misinterpretation by the public and allows for a form of deception discussed in the excellent book by Darrell Huff on "How to Lie with Statistics" by groups wishing to further their own agenda. Let us here clarify things a bit. First of all by the mean of a group of numbers we are generally talking about their arithmetic mean which is just the sum of all the input values divided by the total number of items in the data base. Thus, given the set of numbers $[x_1,x_2,x_3,\ldots x_N]$, we have as the arithmetic mean-

$$\overline{x} = \frac{[x_1 + x_2 + x_3 + ... + x_N]}{N} = \frac{\sum_{n=1}^{N} x_n}{N}$$

Thus the data set [2,2,4,6,7,9] has an arithmetic mean of (2+2+4+6+7+9)/6=5. We can also take this concept directly over to continuous functions y(x) by the use of calculus. Thus the mean of the function y(x) in the interval a<x<b is-

$$\overline{y} = \frac{\int_{x=a}^{b} y(x)dx}{(b-a)}$$

The function $y(x)=x(x^2-3x+2)$ has a mean value of $\overline{y}=1/4$ in the range 0<x<1. This type of mean also occurs in mechanics when discussing the center of gravity which is defined as that point of a 2D lamina about which the sum of the moments equals zero. Taking the case of a uniform density equilateral triangle of side length 1. One finds the center of gravity along a bilateral symmetry axis at-

$$\overline{y} = \frac{4}{\sqrt{3}} \int_{0}^{\sqrt{3}/2} y(1 - \frac{2}{\sqrt{3}}y)dy = \frac{1}{2\sqrt{3}}$$

This point lies one third of the way up from the triangle base along any one of the three lines of bilateral symmetry. It also happens to be the point which centers the largest inscribed circle possible for such a triangle.

A modification of the arithmetic means is the geometric mean G. It is defined as the Nth root of the product of N numbers in a data base. That is-

$$G = [x_1 \cdot x_2 \cdot x_3 \cdot \ldots \cdot x_N]^{1/N} = [\prod_{n=1}^{N} x_n]^{1/N}$$

If we take the logarithm of this expression we find the identity-

$$\ln(G) = \frac{1}{N}\sum_{n=1}^{N} \ln(x_n$$

That is, the logarithm of the geometric mean equals the mean of the logs of x_n. To demonstrate the geometric mean consider the data set [2, 5, 8]. Its arithmetic mean is 15/3=5. The geometric mean has the different value $G=(80)^{1/3}$ =4.30886.. . Note ln(G)=[0.69314+1.60943+2.07944]/3=1.4606. The G finds some application in both finances and in mathematics. In the latter case it plays a role in the AGM method of Gauss for evaluating integrals of the type-

$$I(a_0, b_0) = \int_{x=0}^{\infty} \frac{dx}{\sqrt{[(a_0\, b_0) + x^2]\left[\frac{(a_0 + b_0)^2}{4} + x^2\right]}}$$

You will notice here that the constant terms appearing in the radical are just the squares of the geometric and arithmetic means. By carrying out the iterations –

$$a_{n+1} = \frac{(a_n + b_n)}{2} \quad and \quad b_{n+1} = \sqrt{a_n b_n}$$

enough times the geometric and arithmetic means become equal to the same value M. Thus one can solve the integral in closed form as-

$$I(a_o, b_o) = \int_{x=0}^{\infty} \frac{dx}{(M^2 + x^2)} = \frac{\lim}{x \to \infty}\{\frac{1}{M}\arctan(\frac{x}{M})\} = \frac{\pi}{2M}$$

Taking the case of a_0b_0=1 and $(a_0+b_0)/2$=2 we find $a_4=b_4$=1.4567910≈M. Thus we have the seven place accurate result-

$$\int_{x=0}^{\infty} \frac{dx}{\sqrt{(1 + x^2)(4 + x^2)}} = \frac{\pi}{2(1.456791_0)} = 1.078257_8$$

Let us next look at the median $\mu_{1/2}$ of a set of numbers $[x_1,x_2,x_3,\ldots,x_N]$. It is defined as that value of a data set for which half of the samples lie above in value and half below. To demonstrate, take the seven item set [2,7,8,3,1,9,4] and arrange things in ascending order [1,2,3,4,7,8,9]. Next remove the three values on the left and three

values on the right nearest the bracket ends. One is left with the middle value of 4 which represents the median of the set. Note that the arithmetic mean in this case is 34/7=4.85714.. and so differs slightly from $\mu_{1/2}$. When the number of items in a list is even then the median will be the average of the middle two numbers in the rearranged array. When dealing with a large number of input data points it becomes of advantage to treat the problem as one of continuous input such as for example in discussing the age distribution of a selected population or median family income. Consider the continuous function $y(x)=(x^4+x^2)/2$ over the range $0<x<1$. Its mean value is-

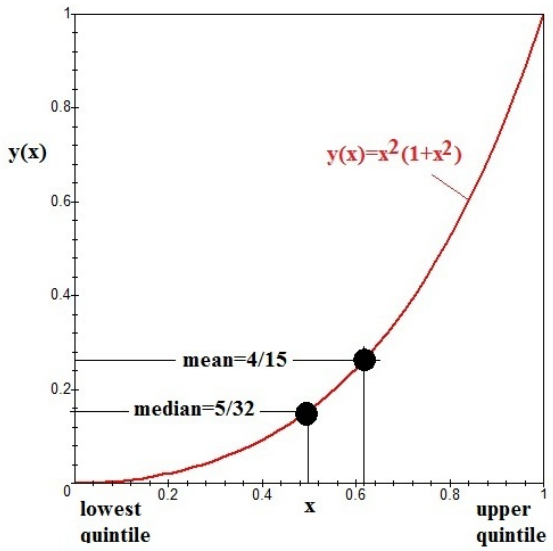
$$\bar{y} = 0.5\int_{x=0}^{x=1} x^2(1+x^2)\,dx = \frac{4}{15} = 0.2666..$$

However the median occurs at x=0.5 and equals-

$$\mu_{1/2} = 5/32 = 0.15625$$

This clearly shows a difference between the mean and median value of the function y(x) in $0<x<1$. This curve is somewhat reminiscent of the well known Pareto Curve used by economists when discussing the distribution of wealth in a specified population group. I have marked up the graph of the above function y(x) to simulate a Pareto Curve. Here is the graph-

PARETO CURVE SHOWING WEALTH DISTRIBUTION VERSUS POPULATION RANK. MEAN AND MEDIAN VALUES ARE INDICATED



In it we see that most wealth in a country is held by only a small fraction of the population. This is reflected by having the median lie well below the mean. The original

statement of the Pareto Principle is that about 80% of a countries wealth is controlled by about 20% of the population. When this ratio gets out of whack , as it has in this country where the upper 1% of the population controls almost 40% of all wealth , it has always in the past led to political instability including economic collapse and/or revolution. The present median family income in the US is about $50,000 per annum. Notice that a straight line distribution y(x)=x for the Pareto Curve would have the mean and median family income be the same.

U.H.Kurzweg
March 17, 2012