Thoughts on measures of central tendency

Frederick Gabriana

Introduction

Students of high school statistics learn three basic measures of central tendency: the mean, median, and mode. All of them have their strengths and weaknesses. The mean has a disadvantage of being easily affected by extreme values. The median, on the other hand, while being a true measure of the center, does not give sufficient information about the data because it is unaffected by the actual values of the points. The mode tends to be too unstable.

To work around these weaknesses, analysts would sometimes discard the highest and lowest values and take the mean of the remaining, or would discard the first and fourth quartiles and get the mean of the middle 50%, or some other variation of this. The disadvantage of this approach is that it involves some arbitrariness in the selection of how far from the center we consider significant enough to include. These approaches build on the idea that values that are farther from the center should be considered less significant than those closer to the center, but do not provide a quantitative way to determine the actual significance of each data point.

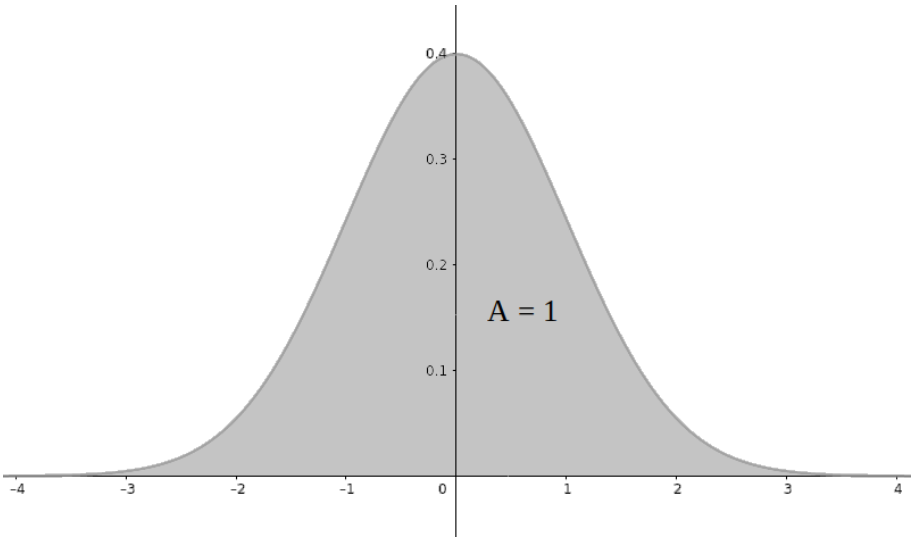
The mean is just an average where all values have equal significance, while the median is one where all values but the middle one (or two) have zero significance. In my mind we should not be limited to these two extremes, but instead there must be a middle ground, a way to mathematically express the significance of each data point.

Central Tendency

Thus I decided to think out of the box, and think of an approach to measure central tendency in a manner that quantifies the significance of each data point according to how far away they are from the center. The true center being the median, my thought was to get a weighted average based on how far away a point is from the median. The question would now be how to assign appropriate weights for each point.

I considered several ways to do this, some more mathematical (a function of how far away they are from the mean or median), some based on physics (an inverse square relationship like how bodies that are farther away from the center of a force field have less effect on the system), but decided to work on an approach that is purely statistical.

While thinking about how the means of random subsets of a population would still form a normal distribution even though the population is non-normal (Central Limit Theorem), it randomly occured to me that a normal curve can actually be used, even for non-normal distributions, to quantify how far an element is from the center. This is because bell curve peaks at the center and vanishes at both ends, and we usually assign the area under the entire curve to be 1, thus can be appropriate to assign weights.



The area under the curve is 1.

The problem with this approach is that the distribution assumes an uncountably infinite number of data points. Thus the need for a way to apply this idea to discrete values.

Binomial Distribution

The coefficients of a power of a binomial can be expressed in terms of Pascal's Triangle, as seen in the following diagram we have learned in school. Each row represents a sequence of coefficients for a given power *n* of a binomial, where rows and columns are counted from 0.

1

1 1

1 2 1

1 3 3 1

1 4 6 4 1

1 5 10 10 5 1

1 6 15 20 15 6 1

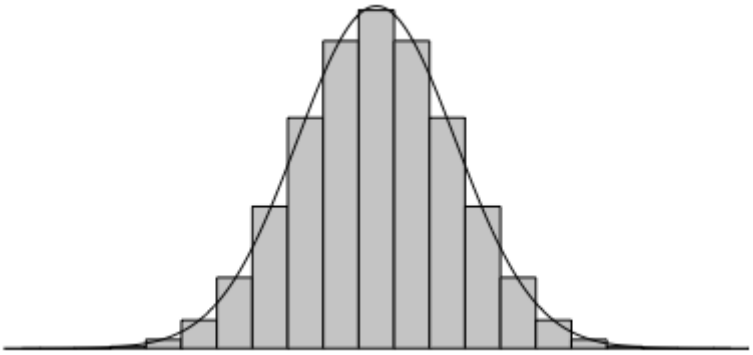
Pascal's Triangle

These numbers form a binomial distribution, and can be written in terms of combinations, e.g., the nth row will have the following entries:

$$\binom{n}{0}, \binom{n}{1}, \binom{n}{2}, \ldots, \binom{n}{n-1}, \binom{n}{n}$$

where each term $\binom{n}{k} = {}_nC_k = \frac{n!}{r!(n-k)!}$ is the number of combinations of *n* taken *k* at a time.

If we examine each row of the triangle, we see that the number at the center is the largest, and tapers off to 1 at both ends. The beauty of this sequence is that when n is very large, the shape of its graph looks that of the bell curve, as illustrated below:



The sequence of binomial coefficients form a bell curve.

If we get the sum of each row, we see that the entries of the nth row just add up to 2^n. Thus we can convert it to a sequence whose sum is 1 by dividing each term by 2^n, as follows,

$$\frac{\binom{n}{0}}{2^n}, \frac{\binom{n}{1}}{2^n}, \frac{\binom{n}{2}}{2^n}, \ldots, \frac{\binom{n}{n-1}}{2^n}, \frac{\binom{n}{n}}{2^n}$$

Thus the area under the curve becomes 1. We can therefore write this as a function defined by

$$f(k,n) = \frac{{}_nC_k}{2^n}$$

where n and k are non-negative integers and $k < n$. This function, called the *probability mass function* $f(k, n, p)$ for $p = 0.5$, is a discrete function that approaches a normal distribution as $n \longrightarrow \infty$.

Applying Weights

I saw this as a way to apply a set of weights to discrete data points so that values farther away from the center become less significant.

We first arrange the set of values in increasing order, as we would when getting the median, but instead of finding only the central value, we assign the centermost terms the largest weights, waning as it goes farther away from the median in the list order.

Let us take for example a set of seven numbers N_1, N_2, N_3, N_4, N_5, N_6, N_7 arranged in increasing order such that N_1 is the smallest and N_7 is the largest. We get the 7th row in Pascal's Triangle ($n = 6$, since we start counting at 0). We see that the values are 1, 6, 15, 20, 15, 6, 1. We then apply these as weights, as follows:

$$1N_1 + 6N_2 + 15N_3 + 20N_4 + 15N_5 + 6N_6 + 1N_7$$

The sum of the coefficients is 1 + 6 + 15 + 20 + 15 + 6 + 1 = 64, which we can also see from the formula $2^n = 2^6 = 64$.

When getting an average in the usual sense, we divide the sum by 7 which is the number of terms. But since this is a *weighted* average, we divide it by 64 which is the sum of the weights:

$$\frac{1N_1+6N_2+15N_3+20N_4+15N_5+6N_6+1N_7}{64}$$

This gives us a weighted average where the median N_4 has *20 times* more significance than the endpoints N_1 and N_7, but where each endpoint still contributes one-64th of the value. Unlike the median, every term contributes some weight. But unlike the mean, extreme values are less likely to influence our measure. The more data points there are, the smaller the significance of the endpoints.

Take note that the assigned weights only depend on the order which they appear on the sequence, and is not affected at all by their actual values. This is on purpose, in order to keep the idea more in line with the median. By doing this, we are able to calculate a central value that is more representative of the data than either the mean or the median.

The formula

I started with an example to show how this works. But now I attempt to write down a general formula that can be used to calculate it in a spreadsheet or a computer algorithm.

Let N_1, N_2, N_3, ..., N_{n-1}, N_n be a set of n data points, arranged in increasing order such that N_1 is the smallest and N_n is the largest. The formula for my modified measure of central tendency can therefore be written as:

$$\frac{\sum_{k=0}^{n}\binom{n}{k}N_k}{2^n}$$

I do not have a name for it yet, nor a symbol for it. I intentionally did not think of one because I am unaware if such a thing already exists, or what it is called. But I decided to write this down in order to help those for whom this idea might be useful.

Whatever the case, this formula represents a measure of central tendency that, in my opinion, could be more representative of a central value for a set of data points than either the mean, median, or mode, and should be useful in many situations.