Diversity, Variability, and the Probable Error: Building Intuition

Robert W. Jernigan
Department of Mathematics and Statistics, American University, Washington, DC 20016

Abstract

Students new to statistical concepts often struggle with the notion of measuring variability. Our standard measures, variance and standard deviation, can seem unduly cumbersome to the uninitiated. Full understanding of the why and how of these measures often needs more mathematical background than many students possess. We illustrate how the concepts of diversity and quantitative variability are confusing. So much so that, in our textbooks even our most basic measures of variability, must be defined, computed, and accepted before intuition about them can be developed. This seems backwards. As an introductory measure of variability, I propose reviving an old measure called the probable error (one half of the inter-quartile range). Examples are given to motivate this measure, building on students' more intuitive ideas about quantitative variability. Relationships with outlier bounds and 68-95-99.7-type rules of thumb are also illustrated and discussed. Some historical perspective is given including why probable error, at best only an introductory measure of variability, is very much misnamed.

Key Words: Inter-quartile Range, Standard Deviation, Empirical Rule

1. Introduction

Students have trouble with the concept of variability. Not only the details of how to compute it, but more fundamentally what it is, how to recognize it, and how to measure it. Some confuse variability with sample size. The reasoning seems to be that a bigger sample must, of course, have a more varied collection of measurements. This misunderstanding is often reinforced when we display simple frequency histograms for two samples of different sizes. As instructors know well, a larger sample will have histogram bins with larger frequency counts, resulting in taller histogram bars. The resulting histogram for the larger sample will actually be larger than the histogram for a smaller sample. To some students this indicates greater variability. Of course, this histogram size issue can be avoided by plotting relative frequency histograms instead.

But beyond this misconception students confuse quantitative variability with qualitative diversity, as illustrated in Figure 1. Two hypothetical fleets of rental cars are shown. The students first see the top fleet of six cars, all of different colors and the bottom fleet of six cars of only two colors. Initially, gas mileage figures are not shown. The students are then asked, “Which fleet is more variable?” This is not a trick question. Without mileage figures the only thing differentiating the fleets is color. The top fleet with six colors is more variable than the bottom fleet with only two colors.

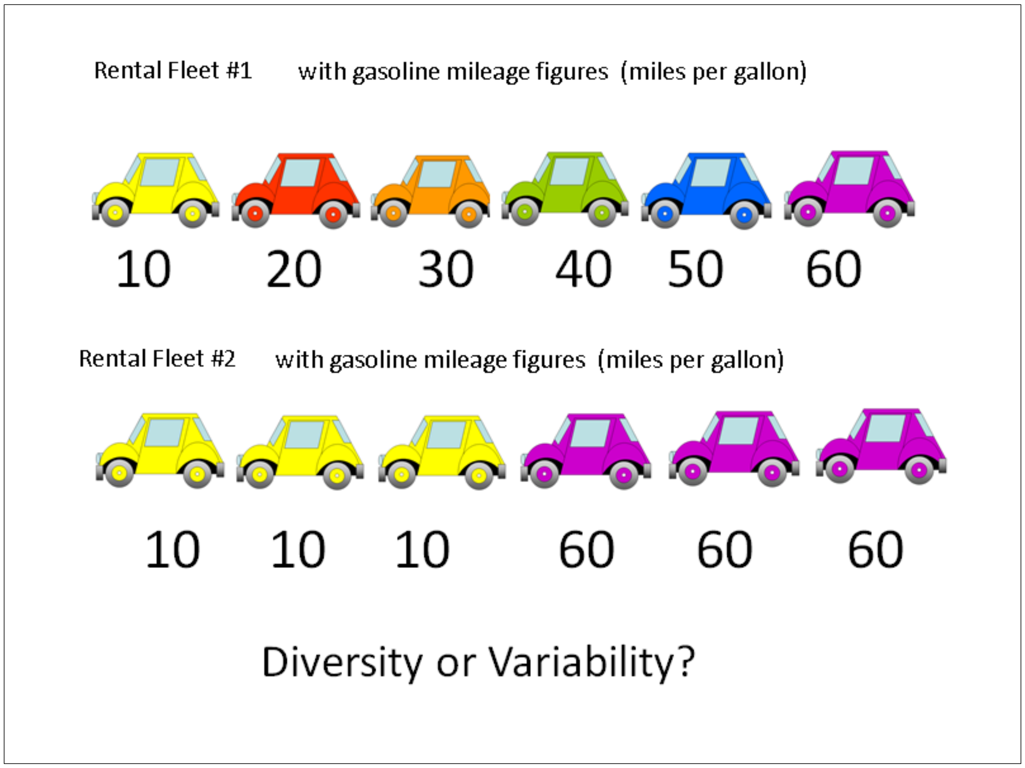
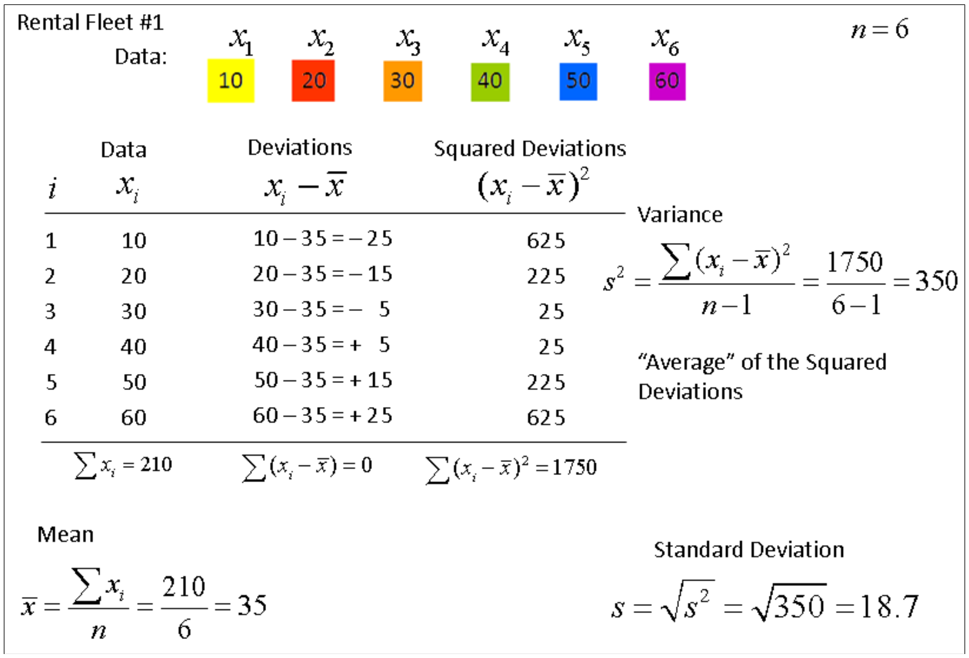

Rental Fleet #1 with gasoline mileage figures (miles per gallon)
10 20 30 40 50 60
Rental Fleet #2 with gasoline mileage figures (miles per gallon)
10 10 10 60 60 60
Diversity or Variability?

Figure 1. Two hypothetical fleets of rental cars with gas mileage figures listed below. See text for more details.

Now the mileage figures are shown and the question is asked again. The top fleet has six different mileage figures, just like its six different colors. The bottom fleet has only two different mileage figures, again just like its two different colors. By the same reasoning as with the car colors, many are likely to say again that the top fleet (with more mileage figure values) is more variable. But when students reason this way they are considering qualitative diversity not quantitative variability. Qualitatively they have seen six different colors or six different mileage numbers in the first fleet and only two different colors and numbers in the second. The first fleet must therefore be more variable. It is more qualitatively diverse. To measure this, imagine selecting two cars at random with replacement from each fleet. The probability that two cars are the same color can be used as a measure of diversity. For the top fleet this probability is 5/6 for the bottom fleet it is 1/2. So the top fleet has greater qualitative diversity.

The first fleet has more varied types (colors or numbers), but this ignores the size of numbers. This reasoning considers the numbers as types not indications of quantity. This reasoning does not examine quantitative variability. The most common measure of quantitative variability is the standard deviation. At this point, the common approach, illustrated in Figure 2, would be to list the sample, compute its mean, and compute the deviations from the mean, to measure how far each car’s mileage is from this mean mileage.



Rental Fleet #1

Data:

x_1	x_2	x_3	x_4	x_5	x_6
10	20	30	40	50	60

$n = 6$

i	Data x_i	Deviations $x_i - \overline{x}$	Squared Deviations $(x_i - \overline{x})^2$
1	10	$10-35=-25$	625
2	20	$20-35=-15$	225
3	30	$30-35=-\ \ 5$	25
4	40	$40-35=+\ \ 5$	25
5	50	$50-35=+15$	225
6	60	$60-35=+25$	625
	$\sum x_i = 210$	$\sum(x_i - \overline{x}) = 0$	$\sum(x_i - \overline{x})^2 = 1750$

Variance

$$s^2 = \frac{\sum(x_i - \overline{x})^2}{n-1} = \frac{1750}{6-1} = 350$$

“Average” of the Squared Deviations

Mean

$$\overline{x} = \frac{\sum x_i}{n} = \frac{210}{6} = 35$$

Standard Deviation

$$s = \sqrt{s^2} = \sqrt{350} = 18.7$$

Figure 2: Typical classroom illustration of the computation of variance and standard deviation for the top fleet of rental cars shown earlier.

These deviations, of course, add to zero (not a very useful measure of variability). So we square them, to get rid of their algebraic signs. Then we “average” these squared deviations and define this to be the variance: our measure of quantitative variability. Of course, our “average” is defined in an odd way, dividing by (n-1), rather than the usual and better understood averaging method of dividing by (n). The top fleet has a variance of 350. Similar computations show that the bottom fleet has a variance of 750. The top fleet has less quantitative variability.

All this is typically done before the student is offered much understanding or intuition about what a measure of quantitative variability is or what it should be measuring. After these findings one can, of course, alert the students to the fact that the top fleet has several cars closer to the mean mileage than the bottom fleet. But his observation comes after the definition, computation, and comparison of variabilities.

This all seems backward. Concepts and intuition should come before formal computations.

2. Building Intuition

We can build greater intuition about quantitative variability by considering practical estimation problems noting both under- and over-estimation, as illustrated in the carton in Figure 3. Arrows are shown that fall short or over shoot a target. The median locations of the shortfall or overshoot provide estimates of these types of errors in an archery contest. Averaging the typical extent of these errors can serve as an easily understandable measure of quantitative variability.

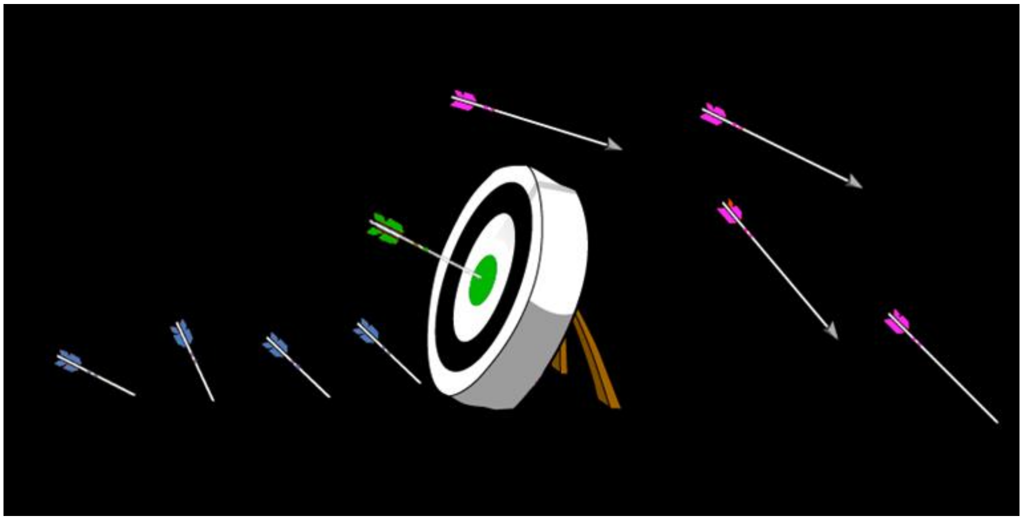


Figure 3. The cartoon shows the results of an archery contest. One arrow (in green) has hit the target. Four arrows (in blue) have fallen short, under-estimating the aim or force needed to hit the target. Four arrows (in pink) have overshot the target, over-estimating the aim or force needed hit the target. The median horizontal location of the blue arrows serves as a typical value of our shortfall. Likewise, the median of the final position of the pink arrows serves as a typical value of our overshot.

To be more specific, and a bit historical, I examined the summary percentiles on an experiment performed by Galton (1907), described more recently in Surowiecki (2004). Having attended an agricultural fair, Galton collected 787 contest entries submitted to guess the weight of an ox. He reports the percentiles (5 to 95 by 5) of the submitted guesses. As an illustrative data set I have used these percentiles to produce a sample of 787 simulated guesses having percentiles matching those of Galton.

Students know well that the median is often a good choice for estimating or guessing a typical value from a general set of data, (although the median may not always be the best choice). The median (*M*) of the guesses was 1207 lbs. very close to the weight of the ox, 1198 lbs. Galton uses this to extol the possible virtues of democratic decisions, and Surowiecki uses this close result as an example of the wisdom of crowds. But what of the variability in these guesses?

Being the median, half of the guesses are under-estimates and half are over-estimates. Considering only those guesses that fall below the median, students are asked “What is a typical value for this under-estimate?” Knowing the median is a good choice they compute the median of just the under-estimates. This, of course, is the first quartile $(Q1)$ of the entire set of 787 guesses. Likewise, considering only those guesses that fall above the median, a typical value is taken to be the median of the over-estimates, yielding the third quartile $(Q3)$ of the entire set of guesses. Figure 4 shows these results for the Galton example. In this context, *Q1* and *Q3* are useful summary statistics of the shortfall and overshot. They provide greater motivation than their common use as merely defining the range of the middle 50% of the data!

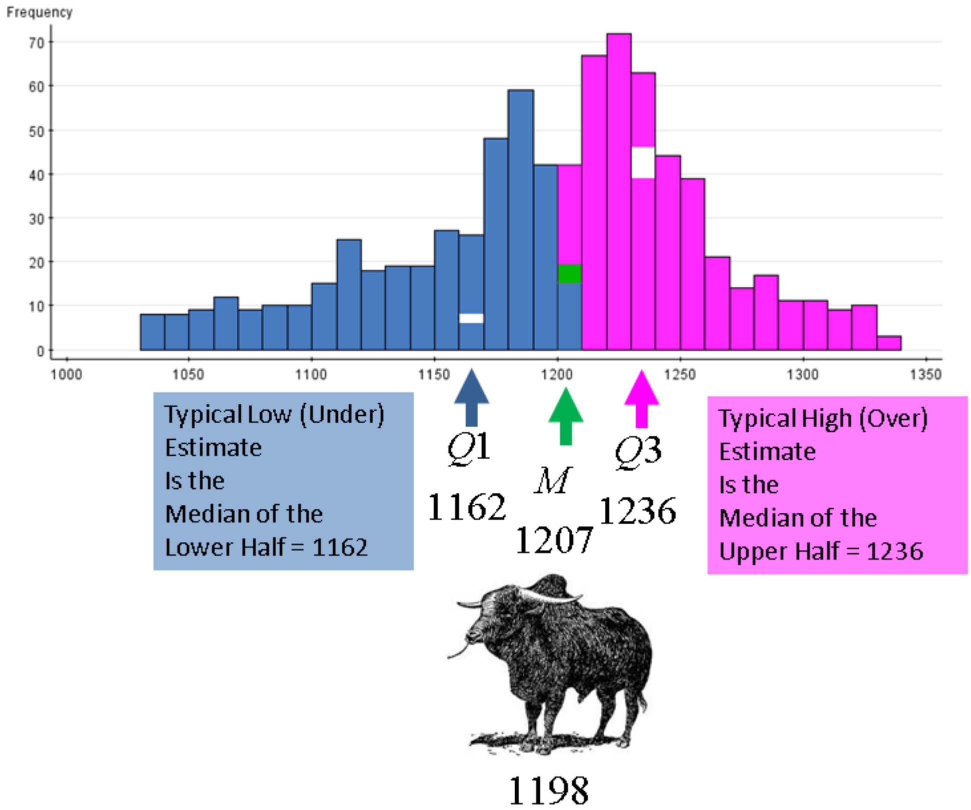


Figure 4. In guessing the weight of the ox, a typical low (under-estimate) is the median of the lower half of the guesses, 1162 (*Q1* for the entire sample). A typical high (over-estimate) is the median of the upper half of the guesses, 1236 (*Q3* for the entire sample).

A typical low-side error in our guessing is then the difference: $M - Q1$. This measures a typical amount by which we might fall short of the correct answer. For the ox data $M - Q1 = 45$. Likewise a typical high-side error in our guessing is the difference: $Q3 - M$. This measures a typical amount by which we might overshoot the correct answer. For the ox data $Q3 - M = 29$. Adding these two errors we get the Inter-quartile Range: $IQR = (M - Q1) + (Q3 - M) = Q3 - Q1 = 74$. Averaging the two errors yields $PE = IQR/2 = 37$, what Galton called the probable error. So if we fall short or over shoot, a typical size for either type of deviation is 37 lbs. It is easily understood and useful measure of variability that can be used in much the same way that the less intuitive and more involved standard deviation.

Galton(1889, p.57) did not like the name "probable error", since for a standard normal distribution this probable error would have a value of $PE = 0.6745 \approx \tfrac{2}{3} Std.Dev.$ where as a literal reading one would have to admit that the most probable normal error is zero. Galton preferred the term probable deviation, which seemed to pave the way for Pearson's standard deviation.

2. Using the Probable Error

The connection between *PE* and the *IQR* allows for a reexamination of our usual outlier bounds or fences. The upper outlier fence is $Q3+1.5\cdot IQR$. The factor of 1.5 seems arbitrary to many. It is rooted in the behavior of outliers for the normal distribution, chosen so that less than 1% of a normal distribution would be labeled as outliers. Using the *PE* the upper outlier fence can be changed to: $Q3+1.5(2PE)=Q3+3PE$ and for the lower outlier fence: $Q1-3PE$.

For symmetric distributions we can expect $Q3=M+PE$ and $Q1=M-PE$ so that the outlier fences take on a more satisfying form defined around the center of the data rather than around upper or lower extremes. The upper outlier fence is: $Q3+3PE=M+4PE$ and the lower outlier fence is: $Q1-3PE=M-4PE$.

For the normal distribution we can also define 68-95-99.7%-type rules, sometimes called Empirical Rules, based on the *PE*. Figures 5 and 6 show such rules for some integer and half-integer values of the *PE*.

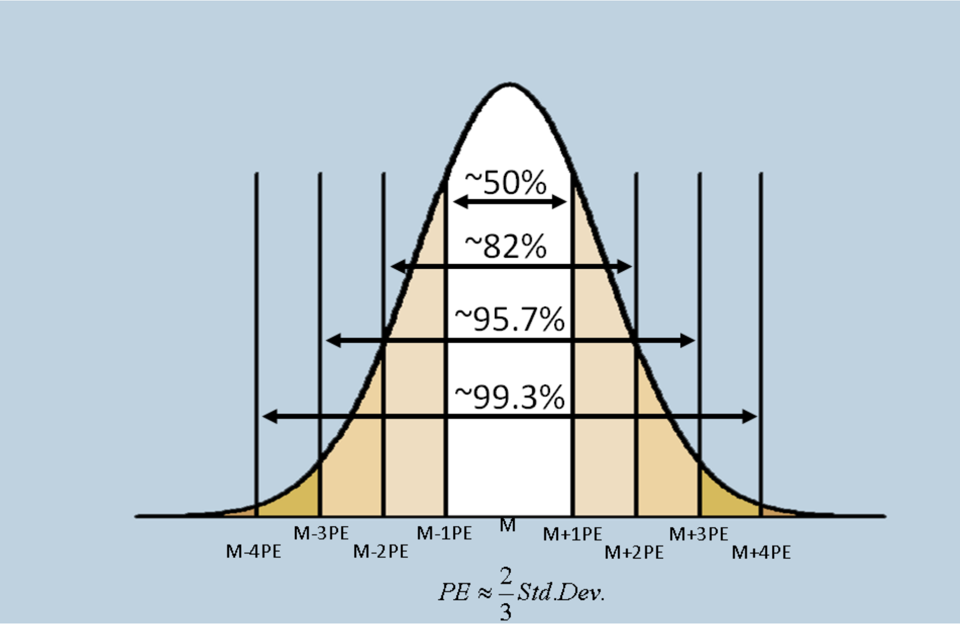


Figure 5: The percentages of a normal distribution that can be expected to fall within one, two, three, and four probable errors (*PE*) of the median (*M*).

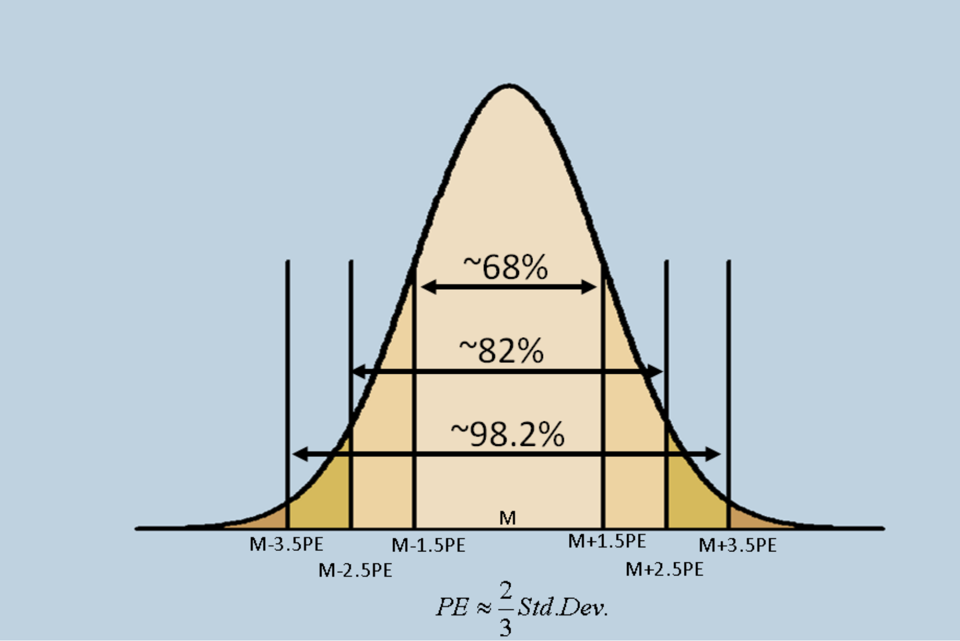


Figure 6: The percentages of a normal distribution that can be expected to fall within 1.5, 2.5 and 3.5 probable errors (*PE*) of the median (*M*).

Note some common results that agree with the usual Empirical Rule benchmarks: 68% within $1.5PE$ of the median (*M*), corresponds to within 1 Std.Dev. of the mean $[1.5PE \approx 1.5(\frac{2}{3}Std.Dev.) = 1\,Std.Dev.]$ and 95.7% within 3*PE* of the median (*M*), corresponds to being within 2 standard deviations of the mean: $3PE \approx 3(\frac{2}{3}Std.Dev.) = 2Std.Dev.$

For asymmetric distributions, like the left-skewed ox data, the shortfall error will be larger than the overshoot error and *PE*, the average of these errors, will misrepresent both. The *PE* will under-estimate the shortfall error, making $M - PE > Q1$. It will also over-estimate the overshoot error, making $M + PE > Q3$. Thus, for such asymmetric distributions, the interval from $M - PE$ to $M + PE$ may not contain the expected 50% of the distribution that the *Q1* to *Q3* interval defines, although for the ox data we do get almost exactly 50% even with the asymmetry.

The probable error is an easily understood, computed, and applied measure of quantitative variability.

3. More Theoretical Uses

Although what follows is not a development that beginning students need or could even follow, it might be useful for instructors to see that we can also develop statistical t-type tests using the *PE* (or *IQR*) as a measure of variability in ways similar to the usual t-test. Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with mean μ and variance σ^2. Let $Y_1 < Y_2 < \cdots < Y_n$ denote the order statistics from this sample. The *IQR* is given by: $IQR = Y_{[3n/4]+1} - Y_{[n/4]+1}$, where the square bracket notation $[x]$ denotes the greatest integer less than or equal to *x*. From well know properties of order statistics it follows that $IQR \xrightarrow{p} \Phi^{-1}(0.75) - \Phi^{-1}(0.25) \approx 1.34898\sigma$, where $\Phi^{-1}(p)$ represents the pth percentile of a standard normal distribution. From this it follows that $PE = IQR/2 \xrightarrow{p} \Phi^{-1}(0.75) = 0.674489\sigma$. We can define a location and scale invariant statistic similar to a t-statistic

$$g = \frac{0.674489\sqrt{n}\left(\bar{X} - \mu\right)}{PE} \quad .$$

By Slutsky's theorem it follows that $g \xrightarrow{d} N(0,1)$ and $n \to \infty$. For large samples, we could test hypotheses or build confidence intervals for μ using the statistic *g* and its asymptotic standard normal distribution. For smaller samples, numerical integration can be used to find the distribution of *g*. But because of the discrete and discontinuous nature of selected order statistics, the resulting behavior of tail probabilities of this statistic are jagged in a damped saw-toothed pattern as the sample size grows and converge to the expected normal distribution values. This erratic pattern results from the shifting position of *Q3* or *Q1* within intervals of order statistics. In general, *g* is not a statistic to be recommended except possibly for large samples.

4. Summary

We revive a measure of quantitative variability built on the grounding principle that the median is a good measure of location for a general set of data. With this principle if we fall short and under-estimate our location, *Q1* the median of all the shortfalls measures this underestimation. If we overshoot and over-estimate our location, *Q3*, the median of all the over shots measures this over-estimation. The difference *M-Q1* measures the error of under-estimation and *Q3-M* measures the error of over-estimation. The sum of these two errors is the Inter-quartile Range, *IQR*. The average of these two errors is the Probable Error, *PE = IQR/2*. The probable error provides a more intuitive and easily computed, and it offers a more understandable motivation for a quantitative measure of variability. As an introductory quantitative measure of variability, the probable error can be used to set up a student's intuition about variability and provide an easier transition to the more durable standard deviation which, quite rightly, is our standard way to measure quantitative variability.

References

Galton, F., (1889),
Natural Inheritance, MacMillan, London.

Galton, F., (1907), "Vox Populi", *Nature*, March 7, 1907, 450-451.

Surowiecki, James (2004). *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations* Anchor Books, New York.