Supporting Information:

Statistical Tools for Analyzing Measurements of Charge Transport

William F. Reus,[1] Christian A. Nijhuis,[2] Jabulani R. Barber,[1] Martin M. Thuo,[1] Simon Tricard,[1] George M. Whitesides[1]*

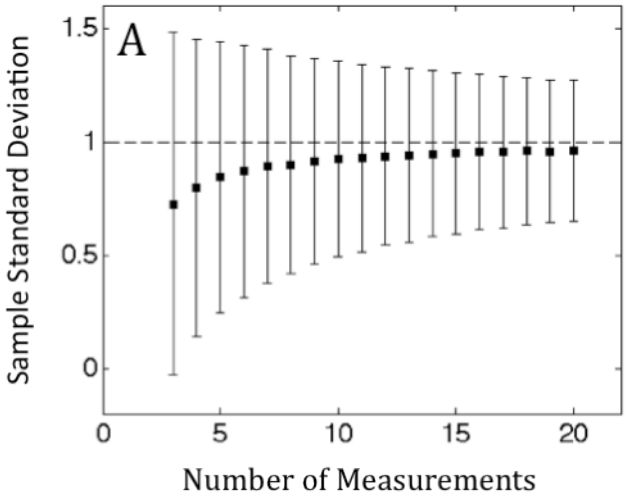
[1] Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, MA, 02138

[2] Department of Chemistry, National University of Singapore, 3 Science Drive, Singapore 117543

*Author to whom correspondence should be addressed: gwhitesides@gmwgroup.harvard.edu

How many data are sufficient for accurate and precise statistical analysis? There is no hard and fast answer to this question, but Figure S1 gives an impression of how the number of data affects the statistical analysis of that data. Figure S1 shows the results of a simulation: i) starting with $N = 3$, N data were sampled from a (computer generated) normally distributed population with mean zero and standard deviation unity, ii) the arithmetic mean (μ_A) and standard deviation (σ_A) of the sample were calculated, iii) steps i and ii were repeated in 1000 trials, after which the average value of σ_A for $N = 3$ was plotted in Figure S1A and the fraction of trials in which the interval ($\mu_A - \sigma_A$, $\mu_A + \sigma_A$) contained zero (the true population mean) was plotted in Figure S1B, and iv) steps i – iii were repeated for $N = 4 – 20$. The implication of Figure S1A is that small samples tend to *underestimate* the true standard deviation of the population, while the implication of

Figure S1: The effect of sample size on the accuracy of a measurement taken from a normally distributed population with a mean of 0 and a standard deviation of 1. (A) The sample (measured) standard deviation as a function of the sample size. Each point is the average of 10^4 measured standard deviations, and the error bars represent one standard deviation from that average. The dashed line indicates the actual standard deviation of the population being measured. (B) The probability that the actual mean of the population lies within one sample standard deviation of the sample mean. This probability is plotted as a function of sample size.


A
Sample Standard Deviation
1.5
1
0.5
0
0
5
10
15
20
Number of Measurements

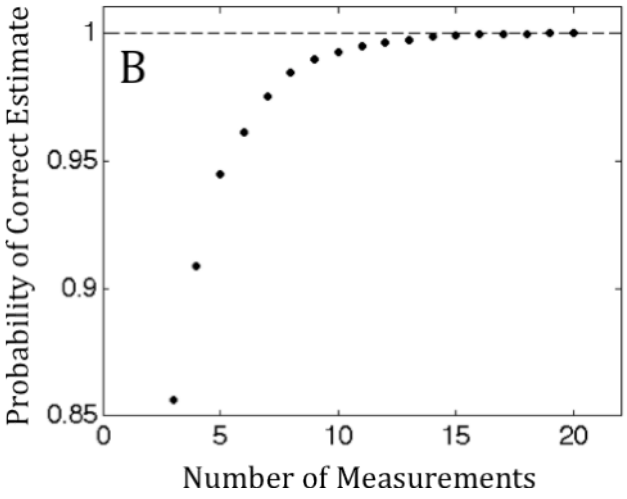

B
Probability of Correct Estimate
1
0.95
0.9
0.85
0
5
10
15
20
Number of Measurements

Figure S1B is that, for small sample sizes, the sample mean, μ_A, is frequently far from the actual population mean. For samples with fewer than 10 data, these effects can be significant. The value of collecting more than a few measurements is, therefore, apparent.

The Normal Distribution, the Log-normal Distribution, and Other Statistical distributions. The statistical distribution of a random variable X is specified by a function (the probability density function, or pdf) that gives the probability, $p(x)$ of observing a particular value (x) of X, as a function of that value. Probably the most common statistical distribution is the normal distribution, an example of which is shown in Figure 3A in the main text, and whose pdf is given in eq. **2**, also in the main text. For a true normal distribution, the Gaussian mean, median, and arithmetic mean are all equal, as are the Gaussian standard deviation, the adjusted median absolute deviation, and the arithmetic standard deviation. The former values indicate the center of the peak, while the latter values give an impression of the width of the peak.

The log-normal distribution is related to the normal distribution. If the variable X is normally distributed, then the variable $Y = e^X$ is an exponential function of a normally distributed variable. The transformed variable $\log Y = X$ is, therefore normally distributed. Thus, Y is said to be log-normally distributed. In the same way, if d (in the Simmons model; see main text, eq. **1**) is normally distributed, then since J depends exponentially on d, J is log-normally distributed.

Histograms. A histogram is a method for plotting a sample of values in order to observe the statistical distribution of that sample and, hopefully, to infer something about the distribution of the underlying population. When analyzing any relatively large dataset, the histogram is one of the most fundamental and useful tools available. Histograms are useful for visualizing the shape of a particular distribution, and for comparing the shape of the experimental (measured) distribution against a statistical model.

To construct the histograms in Figures 2, 3, and 5 of the main text, we i) partitioned the range covered by the distribution into bins, ii) assigned each value in the distribution to the bin that included that value, and iii) plotted the number of values in each bin against the location of the center of the bin.

While it is not necessary to plot a histogram of a sample in order to calculate the median or the arithmetic mean of the sample, plotting a histogram is still useful, especially with large samples, for performing a visual check of the shape of the distribution. For determining the Gaussian mean, on the other hand, constructing a histogram is necessary, because the fitting algorithm uses the histogram as the object to which it fits a Gaussian function. Choosing the number (or width) of bins in the histogram can have an effect on the Gaussian mean, but unless the number of bins is very small (< 10) or very large (approaching or exceeding the number of data in the sample), then this effect is probably negligible. There is no standard rule for choosing the number of bins in a histogram, but some have suggested that, over the range of data in the sample, the number of bins should be approximately the square root of the number of data in the sample. We have chosen, for consistency, to plot all of our histograms with 10 bins per unit of $\log|J|$ (i.e. ~ 40 bins covering the bulk of the data for most samples).

Details of Fitting Gaussian Functions to Histograms of log|J|. We used the curve-fitting tool in MATLAB 7.10.0.499 (R2010a). This tool is used by calling the function "fit()" from the command line, and selecting 'gauss1' (i.e. one Gaussian peak) as the model to fit. The options we used are the standard options in MATLAB for the 'gauss1' model: no excluded data, no weights, no bounds for the mean, a lower bound of zero for the standard deviation, no "robust" fitting options selected, and a trust-region-reflective algorithm (this algorithm was recommended, but the Levinberg-Marquardt algorithm gave similar results). The first parameter determined by the fitting algorithm specifies the area under the curve (we ignored this parameter, because it is specific to a sample, and not a general characteristic of the population), the second parameter is the Gaussian mean (μ_G), and the third parameter must be divided by the square root of two, in order to find the Gaussian standard deviation (σ_G). The coefficients of determination, R^2, for the fitted Gaussian functions for n = 9 – 18 are given in Table S1.

Calculating Quantiles. For a sample, in which N values of x are sorted in increasing order, the q quantile is the ith value of x, x_i, where $i = q(N+1)$, if i is an integer. If i is not an integer, then if j = integer(i) (i.e. if j is i rounded down), then the q quantile is $x_j + q(x_{j+1} - x_j)$. In other words, the q quantile is linearly interpolated from the sorted values of x.

Values of Confidence Intervals for μ_G, m, and μ_A. In the main text, we explained how to calculate confidence intervals for estimates of the locations of samples of log|J|. We

Table S1. Coefficients of Determination for Gaussian Fits to Histograms of log|*J*| for $S(CH_2)_{n-1}CH_3$

n	R^2
9	0.6986
10	0.6438
11	0.7191
12	0.7222
13	0.6973
14	0.7887
15	0.7544
16	0.8177
17	0.7798
18	0.7845

plotted the 99.9% confidence intervals on the Gaussian mean (μ_G), median (m), and arithmetic mean (μ_A) in Figure 7, and here, we give the numerical values of those confidence intervals in Table S2.

Using Confidence Intervals to Calculate p Values for Statistical Tests. The confidence interval on the value of β determined by the regression line on a plot of log|J| vs. n is given by eq. **S1**.

$$CI(\beta) = z_{\alpha/2}\frac{S}{\sqrt{S_{xx}}} = z_{\alpha/2}\frac{\sqrt{\frac{1}{N-2}\sum_{i=1}^{N}\varepsilon_i^{\,2}}}{\sqrt{N_{eff}\,\mathrm{var}(\mathrm{n})}} \qquad \textbf{(S1)}$$

N_{eff} is the effective sample size, defined by equations **6** and **7** in the main text. As explained in the main text, this type of confidence interval corresponds to the Z-test. The confidence interval corresponding to the t-test, which performs better than the Z-test, could be obtained by replacing $z_{\alpha/2}$ (the inverse of the standard normal cumulative distribution function, evaluated at $1 - \alpha/2$) with $t_{N-2,\alpha/2}$ (the inverse of the cumulative distribution function for a t-distribution, with $N - 2$ degrees of freedom, evaluated at $1 - \alpha/2$). Since $t_{N-2,\alpha/2} \approx z_{\alpha/2}$ for large N, the Z- and t-confidence intervals are approximately equivalent for our samples. In eq. **S1**, the numerator can be conceptually understood (*via* a rough analogy) as the standard deviation of the residuals (the differences between the measured values of log|J| and the fitted function). With this understanding, the confidence interval for β bears a formal resemblance to the confidence interval for the Gaussian mean, given by eq. **5** in the main text. The main conceptual difference is that the denominator also contains the variance of all values of n (the molecular length), to account for the fact that log|J| is supposed to vary as n varies.

Table S2. 99.9% Confidence Intervals for the Location Parameters Estimated by Methods 1 – 3.

n	μ_G	CI(99.9%)	*m*	CI(99.9%)	μ_A	CI(99.9%)
9	-1.784	(-1.840, -1.729)	-1.902	(-1.974, -1.730)	-1.990	(-2.074, -1.906)
10	-1.774	(-1.799, -1.749)	-1.910	(-2.015, -1.836)	-2.120	(-2.200, -2.040)
11	-3.234	(-3.321, -3.146)	-3.101	(-3.211, -3.022)	-3.040	(-3.127, -2.953)
12	-2.465	(-2.516, -2.415)	-2.525	(-2.569, -2.486)	-2.540	(-2.624, -2.456)
13	-3.864	(-3.937, -3.790)	-3.855	(-3.974, -3.704)	-3.730	(-3.812, -3.648)
14	-3.695	(-3.751, -3.640)	-3.762	(-3.813, -3.713)	-4.010	(-4.067, -3.953)
15	-4.928	(-5.010, -4.845)	-4.655	(-4.792, -4.559)	-4.450	(-4.572, -4.328)
16	-4.315	(-4.363, -4.267)	-4.324	(-4.388, -4.262)	-4.510	(-4.605, -4.415)
17	-5.815	(-5.841, -5.788)	-5.704	(-5.779, -5.613)	-5.460	(-5.572, -5.348)
18	-5.310	(-5.368, -5.252)	-5.255	(-5.306, -5.177)	-5.080	(-5.162, -4.998)

μ_G is the Gaussian mean (Method 1), *m* is the median (Method 2), and μ_A is the arithmetic mean (Method 3)

To perform a Z-test (which is approximately equivalent to a t-test, in our case) comparing β_{odd} and β_{even}, it is necessary to calculate the test statistic, Z, using eq. **S2**.

$$Z = \frac{\left|\beta_{odd} - \beta_{even}\right|}{\sqrt{CI\left(\beta_{odd}\right) + CI\left(\beta_{even}\right)}} \qquad \textbf{(S2)}$$

The probability, p, that the null hypothesis (that $\beta_{\text{odd}} = \beta_{\text{even}}$) is true, according to the Z-test, is then given by the standard normal cumulative distribution function, evaluated at Z. (For a t-test, one would evaluate the cumulative distribution function for the t-distribution, with $N - 2$ degrees of freedom, at Z).

The confidence interval for $\log|J_0|$ is given by eq. **S3**.

$$CI\left(\log\left|J_0\right|\right) = z_{\alpha/2} \frac{S\sqrt{\sum_{i=1}^{N} \mathrm{n}_i^{\,2}}}{\sqrt{N_{eff}\,\mathrm{var}(\mathrm{n})}} \qquad \textbf{(S3)}$$

This confidence interval also bears a formal resemblance to eq. **5** in the main text, but it includes an extra factor in the numerator. Because $\log|J_0|$ is the y-intercept of the fitted function, it must be determined by extrapolating from the domain of the data (n = 9 – 18) to the y-intercept (n = 0). The extra factor in the numerator of **S3** has the function of increasing the width of the confidence interval, to account for the uncertainty of this long extrapolation. The p value for the null hypothesis that $\log|J_{0,\text{odd}}| = \log|J_{0,\text{even}}|$ is calculated in the same manner as above, using eq. **S2**.