Determining the Statistical Significance of the Difference Between Arbitrary Curves: A Spreadsheet Approach

Kalina Hristova[1] and William C. Wimley[2*]

[1]Department of Materials Science and Engineering, Whiting School of Engineering, Johns Hopkins University, Baltimore MD, 21218

[1]Johns Hopkins Institute for NanoBiotechnology. Baltimore, MD 21218

[2]Department of Biochemistry and Molecular Biology, Tulane University School of Medicine, New Orleans, LA 70112

*To whom correspondence should be addressed

ABSTRACT

We present a simple, spreadsheet-based method to determine the statistical significance of the difference between any two arbitrary curves. This modified Chi-squared approach includes a critical correction for the deviation from normality in measurements with small sample size, which are typical in biomedical sciences. Statistical significance is determined without regard to the functionality of the curves, or the signs of the differences. Numerical simulations are used to validate the procedure. Example experimental data are used to demonstrate its application, and an Excel spreadsheet is provided for performing the calculation.

INTRODUCTION

Biological phenomena are often characterized by the measurement of curves portraying a dependent experimental output as a function of an independent parameter such as concentration, dilution factor, or time. Some biological phenomena are well described by explicit, physically meaningful functions, and thus can be reduced by curve fitting to a set of physically meaningful fit parameters, with associated uncertainties, that can be compared statistically. But many biological phenomena obey complex, unknown functions that cannot be described with mathematically simple functions or with functions that have physically meaningful parameters. Thus, determining the statistical significance of the difference between arbitrary curves is a commonly faced problem, and is one that does not have a widely known and easily implemented solution.

The problem of comparing arbitrary curves is one of testing the null hypothesis that the two curves, such as the simulated data in **Fig. 1**, were sampled from the same parent population, at each value of X, but without any assumptions about how the curve values change with X. Ultimately comparing arbitrary curves requires a comparison of the total sum of squared differences between the curves, to the distribution of differences expected from random sampling when the null hypothesis is true. This goal is not accomplished by two-factor ANOVA or by any other commonly used method in biostatistics.

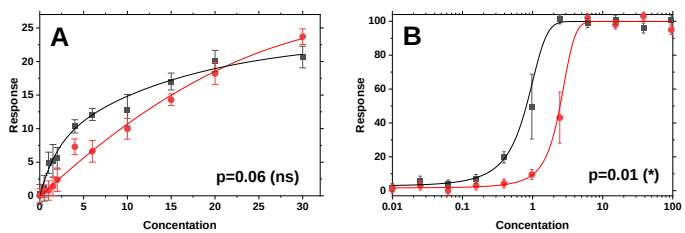

Figure 1. Simulated pairs of curves. Each point is sampled randomly from a Gaussian distribution with N=4. **A.** Curves plotted on a linear X-axis. In this example, uncertainties are constant (independent of Y) and are shown as standard errors calculated individually from randomly sampled data for each point. **B.** Curves naturally plotted on a log axis. In this case uncertainties scale with the slope of the curve. Using the method described here, any pair of arbitrary curves with known uncertainties can be compared to produce a p-value for the null hypothesis that both curves were sampled from the same parent population at each value of X.

Here we present a simple spreadsheet-based method of determining the statistical significance of the overall difference between any pair of arbitrary curves. Our approach is based on the commonly used Chi-squared test for goodness of fit[1] but with an essential correction for the deviation from normality that arises with small to moderate sample sizes, which are nearly universal in biomedical research. Similar to one-factor ANOVA, this method enables a user to calculate an overall p-value for the null hypothesis that, at every value of X, the measurements for two curves were sampled from the same parent population and it does so with a false positive rate (at least one Tyle I error over all comparisons) equal to α, the chosen cutoff for statistical significance, typically 0.05. The method we describe also provides a post-hoc test for calculation

of p-values for the points ateach value of X, with a correction for multiple comparisons, so the user can determine which part of the curve contributes most to the measured overall statistical significance.

METHODS

Chi-squared comparison of two means. The chi-squared (χ^2) probability distribution was first described in 1900 by Karl Pearson to compare categorical data[2], but it has also found many applications in continuous data, especially in regression and curve fitting[1]. In goodness of fit tests, χ^2 takes the form

$$\chi_i^2 = \left(\frac{O_i - E_i}{\sigma_i} \right)^2 \tag{1}$$

where O_i is the observed value of point i, E_i is the expected value of point i based on the fit curve, and σ_i is the standard deviation of O_i in the Y-dimension at point i. When comparing a pair of measured means, O_i^j, where j=dataset a or b, each determined from N^j independent measurements, χ_i^2 takes the form

$$\chi_i^2 = \left(\frac{O_i^a - O_i^b}{SE_i^{diff}} \right)^2 \tag{2}$$

In this equation, O_i^a and O_i^a are the two means and SE_i^{diff} is the standard error of the difference, defined by

$$SE_i^{diff} = \sqrt{\left(SE_i^a \right)^2 + \left(SE_i^b \right)^2} \tag{3}$$

where se^j is the standard error of each measurement, calculated from standard deviation (σ)

$$SE_i^j = \sigma_i^j \Big/ \sqrt{N_i^j} \quad . \tag{4}$$

The utility of the χ^2 formalism is that the individual χ_i^2 values for pairs of points can be summed over any number of pairs to determine an overall statistic for the difference between two arbitrary curves, and this can be done without increasing the probability of false positive results from the multiple comparisons. This sum of χ^2 enables calculation of a probability, the p-value, for the null hypothesis that the two curves were sampled from the same parent population, i.e. that the differences between the curves arose only due to random sampling. Neither the signs of the differences nor the absolute values of the points influence the

final results of this analysis. They are dependent only on the magnitude of the difference between points, relative to the standard error of the difference.

Here, we will show how to use this approach to generate p-values for the pairwise comparison of any two curves and will validate the approach by simulation.

Using the t-distribution to correct for small N. The χ^2 formalism just described assumes that the parent probability distributions in Y around each point follow a Gaussian/normal distribution shape. However, as Gossett (aka "Student") first showed in 1908[3], this is only true when the mean of each point is determined with large N. In biomedical sciences, this condition is rarely satisfied and undersampling is typial. For N $\leq$ 20 the deviations from normality are not trivial, and for N$\leq$10, they are large and cannot be ignored. Gossett thus defined the t-distributions as the appropriate probability distributions to use for small N values. To correct for the effect of small N, while still using the χ^2 formalism, we will determine a corrected χ^2 for each pair of points using t-distributions, instead of directly calculating it using equation 2.

In this approach, the t-distribution is used to calculate the corrected probability from the overlap between each pair of points, assuming the null hypothesis is true and the samples are normally distributed at large N. We note that the t-test probability does not assume that the SD of the two points are the same. Thus, the t-value and probability are equivalent to those from a Welch's corrected, two-sample t-test, which does not assume that SDs are equal. This probability is then used to calculate an equivalent, thus corrected, value for χ_i^2 .

To use the t-distribution to calculate probability, we first calculate t for each pair of points

$$t_i = \frac{\overline{O_i^a - O_i^b}}{SE_i^{diff}} \qquad (6)$$

where the terms are defined above. With t and dF^t, the probability for each pair of means is determined

$$t_i \ and \ dF_i^t \xrightarrow{t-distribution} p_i \qquad (7)$$

which in turn is used to calculate a corrected Chi squared with an inverse Chi squared calculator using dF=1, for the pair of points.

$$p_i \ and \ dF = 1 \xrightarrow{\chi^2-distribution} \chi_i^2 \qquad (8)$$

Calculating the overall statistical significance. To calculate the overall p-value of the difference between curves, over all pairs of points taken together, the corrected χ^2 values are summed

$$\chi^2_{sum} = \sum_{i=1}^{M} \chi_i^2 \tag{9}$$

and then the overall p-value for the comparison of two curves is determined using a χ^2 calculator using dF=M, where M is the number of pairs of points.

$$\chi^2_{sum} \ and \ dF^{curves}(M) \xrightarrow{\chi^2-distribution} p_{overall} \tag{10}$$

The null hypothesis and p-value for comparison of curves. The null hypothesis for comparison of two arbitrary curves is that each pair of measurements O_i^a and O_i^b collected at the same value of X, were sampled from the same parent population at that point. There is no assumption about how the parent populations differ from one another at different X values, thus the differences between any pair of curves with arbitrary shape can be assessed.

Determination of SE values. An accurate measurement of uncertainty based on multiple independent experiments is essential for calculating accurate p-values using this approach. When comparing two curves, ideally one should use uncertainties that are individually determined for each point. If uncertainties are expected to be equal across the range of X, averaged SE values can be used as long as they are experimentally determined using independent measurements. In simulations like those described below, we found that there is no systematic difference between individually measured and averaged uncertainties, as long as they are not systematically different across the curve.

Degrees of freedom for t-distribution. Typically, the individual means being compared will both be measured with the same N points, such that $dF^t = 2N-2$. When the two means are measured with different number of points, N_{larger} and $N_{smaller}$, the smaller N dominates the deviation from normality. In this case, the best estimate for dF^t is $dF^t = 2N_{smaller} - 2$.

Spreadsheet approach We show next how the overall p-value for the comparison of two arbitrary curves can easily be determined using a Microsoft Excel spreadsheet. A spreadsheet with these functions coded as described below is included in **Supplemental Information**. These functions are easily recreated in MatLab, Python, and other programming environments.

To calculate mean and SE from raw data in EXCEL

Enter $Y_i^{j=a}$ raw measurements for curve *a* in rows, one row per X-value

O_i^a - Calculate mean of point i in curve j=*a* with [=AVERAGE($Y_1^a : Y_N^a$)] where N is N_i^a

SD_i^a - Calculate SD of point i in curve j=a with [=STD.S($Y_1^a : Y_N^a$)] where N is N_i^a

N_i^a - Calculate number of measurements for each point i in curve j=a with [=COUNT($Y_1^a : Y_{all}^a$)]

SE_i^a - Calculate SE of point i in curve j=a with [=STD.S($Y_1^a : Y_N^a$)/SQRT(N)] where N is N_i^a

Repeat above steps for curve b to get O_i^b , SD_i^b , N_i^b and SE_i^b

Paste these values into the worksheet for Chi-squared calculation.

To calculate corrected χ^2 values for each pair of points in Excel

(Column designations match the supplementary Spreadsheet included)

Col A: Enter X_i , the X-values for the two curves, a and b

Col B: Enter O_i^a (Mean of Y for point i, Curve a)

Col C: Enter SD_i^a (SD of point i, Curve a) (Optional. SD is not used directly in the calculations)

Col D: Enter N_i^a (Number of measurements for point i, Curve a)

Col E: Enter SE_i^a (SE of point i, Curve a)

Col F: Enter O_i^b (Curve b)

Col G: Enter SD_i^b (Curve b)

Col H: Enter N_i^b (Curve b)

Col I: Enter SE_i^b (Curve b)

[characters within brackets refer to excel formulas with the same column designations as the supplmemtary spreadsheet]

Col K: Calculate absolute value of difference between the means of the ith pair of points with [=ABS(Fi-Bi)]

Col L: Calculate standard error of the difference between ith pair of points, with [=SQRT(Ei^2+Ii^2)]

Col M: Calculate t-value for ith pair of points with [=Ki/Li]

Col N: Determine N for dF calculation. If the number of data points is unequal, choose the smaller of the two N values, with [=IF(Hi<Di,Hi,Di)]

Col O: Calculate dF for t-calculation with [=2*Ni-2]

Col P: Calculate the probability from t and dF with [=T.DIST.2T(Mi,Oi)]

Col Q: Calculate corrected χ_i^2 using the probability in Col P and dF=1 with [=CHISQ.INV.RT(Pi,1)]

To calculate the overall p-value in Excel using χ^2

Col T, Cell 2: Calculate sum of χ_i^2 with [=SUM(Q3:Qm)] where m is the row containing the last pair of points

Col T, Cell 3: Determine dF for χ_i^2 = M, the number of pairs of points, with [=COUNT(Q3:Qm)]

Col T, Cell 4: Calculate overall p-value for the comparison of the two curves using the sum of χ^2 and dF with [=1-CHISQ.DIST(T2,T3,TRUE)]

RESULTS and DISCUSSION

Validation by simulation To validate this entire approach, including the correction for deviation from normality, and the choices for degrees of freedom, we performed numerical simulations of 10,000 pairs of curves created by randomly sampling points from a Gaussian parent population to generate pairs of curves from a simulated experiment that satisfies the null hypothesis of having all points sampled from the same parent population. For each simulation, we determined the overall p-value using the modified χ^2 approach as above. We tested datasets with 3 and 10 measurements per point (N) and 3 or 10 points per curve (M). Because the simulated data are random, and satisfy the null hypothesis, by definition we can validate the entire χ^2 approach by showing that the frequency distributions of observed p-values and observed χ^2 values obtained from the simulations match the expected frequencies based on the χ^2 probability distribution. For this comparison, the observed frequencies of simulation p-values can be compared to their predicted frequencies. More importantly, the frequency of simulation outliers that give false positive results (i.e. Type I errors) can be compared to the critical right tail of the asymmetric $\chi 2$ distribution.

In the simulations shown in **Fig 2A**, we sampled points with N=3 from a Gaussian parent distribution. Simulations with 3 measurements per point are a worst case for the necessary correction for deviation from normality. We created curves with 3 or 10 pairs of points each. We compared the curves using the modified Chi Squared approach, and accumulated 10,000 overall Chi squared values and 10,000 overall p-values. We binned the p-values into 6 bins; two 0.025 wide probability bins, each, on the left and right side of the distribution and the remaining 0.90 probability bin in the middle of the distribution. To test for outliers, we also measured the probability of p≤0.001. The results in **Fig. 2A&C** show that the Type 1 error frequencies are exactly correct in all cases; for bin sizes of 0.025 and 0.05 and for both left and right tails, which are asymmetric, and for the outliers at p>0.001. In **Fig. 2B&D** we compare the observed χ^2 frequency distributions with the expected probability distribution. In this case again, the observed and expected χ^2 distribution are indistinguishable. We note that <u>any errors</u> in implementation of the algorithm, such as incorrectly determined dF values, lead to deviations from the expected values in these simulations.

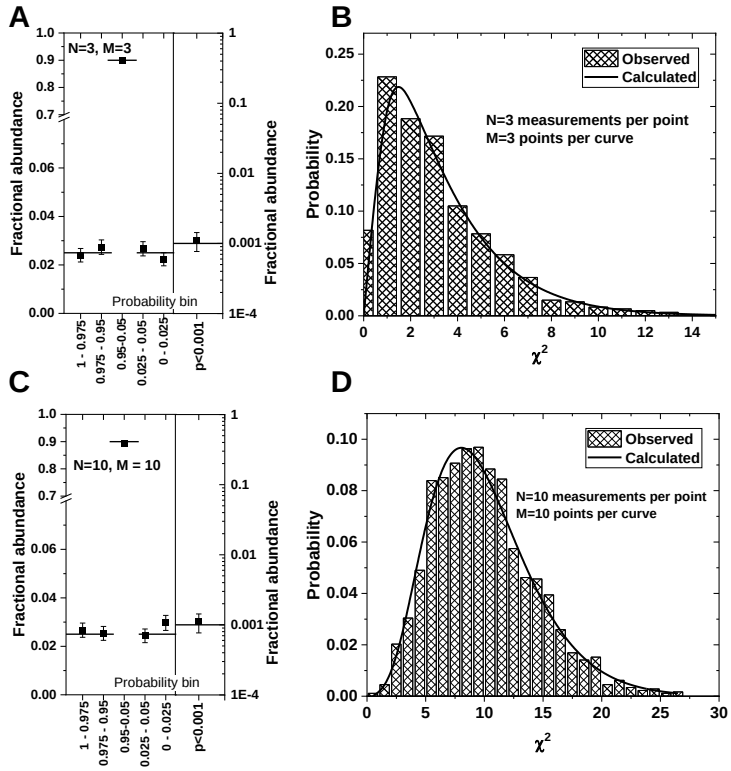

Figure 2 Validation of the modified Chi Squared approach using numerical simulations. Pairs of curves were simulated by sampling M pairs of points where each point is comprised of N random samples from a Gaussian parent population. **A,B:** Simulations with N=3 and M=3. **C,D:** Simulations with N=10 and M=10. **A,C:** Expected and measured probability of calculated overall p-values, which are shown ± SD. Horizonal bars indicate the expected values for each probability bin. **B,D:** Expected and measured distribution of overall χ^2.

Post hoc analysis. The probabilities between pairs of points calculated using the t-test above (Col P, above) enable the *post hoc* determinization of which pairs of points contribute the most to the overall difference between the curves. To state a standard statistical significance for each pair of points, one has to correct for multiple comparisons. For simplicity and stringency, we use the Bonferroni correction which enables

calculation of a corrected cutoff for statistical significance. Formally, the Bonferroni correction can be described as

$$1-\alpha_{overall} = (1-\alpha_{corr})^M \quad or \quad 1-\alpha_{corr} = \sqrt[M]{(1-\alpha_{overall})} \qquad 11)$$

Where $\alpha_{overall}$ is the desired total type 1 error rate over all M comparisons, typically 0.05, and α_{corr} is the corrected cutoff for significance of each of M comparisons needed to achieve the desired $\alpha_{overall}$. However, the Bonferroni-corrected α can be closely approximated by

$$\alpha_{corr} = \alpha_{overall} / M \qquad 12)$$

where the corrected cutoff for significance is equal to the uncorrected cutoff (0.05) divided by M, the number of pairs of points being compared. The power of the χ^2 approach is that a highly significant difference between curves can be identified even when none of the differences between the individual points rise to the level of significance.

We note that the Bonferroni correction is considered inappropriately stringent for curves containing many comparisons. In this case, the Benjamini-Hochberg[4] or similar methods for performing multiple comparisons using a false discovery rate approach may be more appropriate.

Multiple pairwise comparisons of curves. If more than one pairwise comparison between curves is needed, such as multiple curves versus a control or multiple curves versus each other, the χ^2 approach remains applicable, but

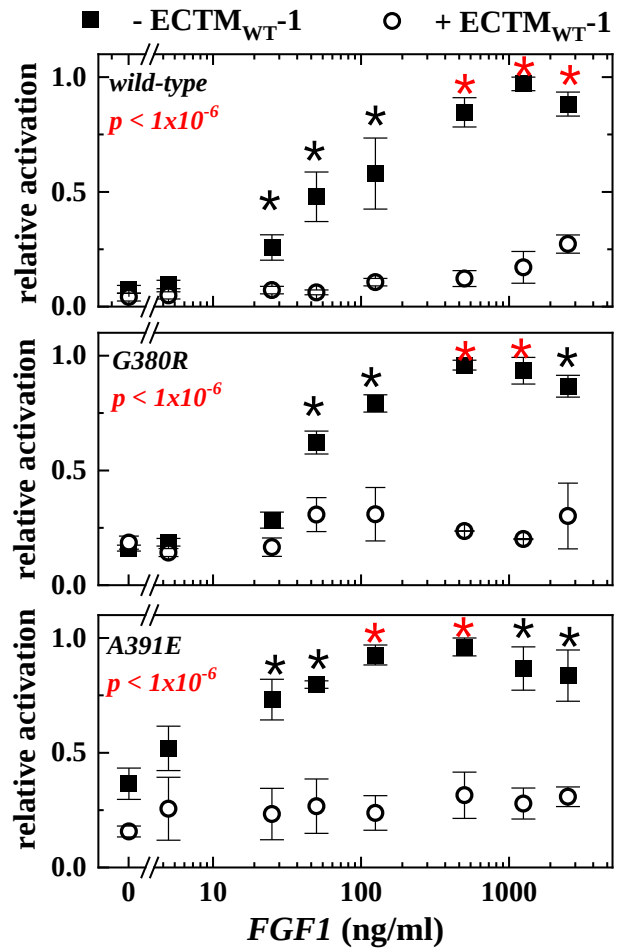

Figure 3. Phosphorylation of the kinase domain of the receptor FGFR2. The activation levels for Wildtype and two mutants were measured in the presence and absence of an inhibitory wildtype construct, ECTMwt, comprised of extracellular and transmembrane domains, but without a functional kinase domain. For each point, three independent experiments were done. Values shown are means and error bars SE. A black asterisk indicates a pair of points with p<0.05. A red asterisk indicates a pair of points that are significant after the Bonferroni correction for multiple comparisons, p<0.006.

a correction must be made for multiple comparisons between curves. Here again, the Bonferroni correction to α, the cutoff for statistical significance, is appropriate and statistically robust,

$$\alpha_{corr} = \alpha_{overall} \Big/ Number\ of\ comparisons \qquad 13)$$

Example data analysis

Next, we show the application of the Chi squared approach using a previously published data set.

Heterodimerization of FGFR3 mutants. In previously published work, we characterized the heterodimerization of the receptor tyrosine kinase FGFR3 with two pathogenic mutants that cause achondroplasia in humans, A391E and G380R. Experimentally, we measured autophosphorylation of the receptors as functions of ligand concentration, both in the presence and absence of enzymatically inactive truncated FGFR3, called $ECTM_{WT}$-1. If heterodimerization between WT and mutant occurs, the presence of $ECTM_{WT}$-1 will inhibit the mutant receptor phosphorylation.

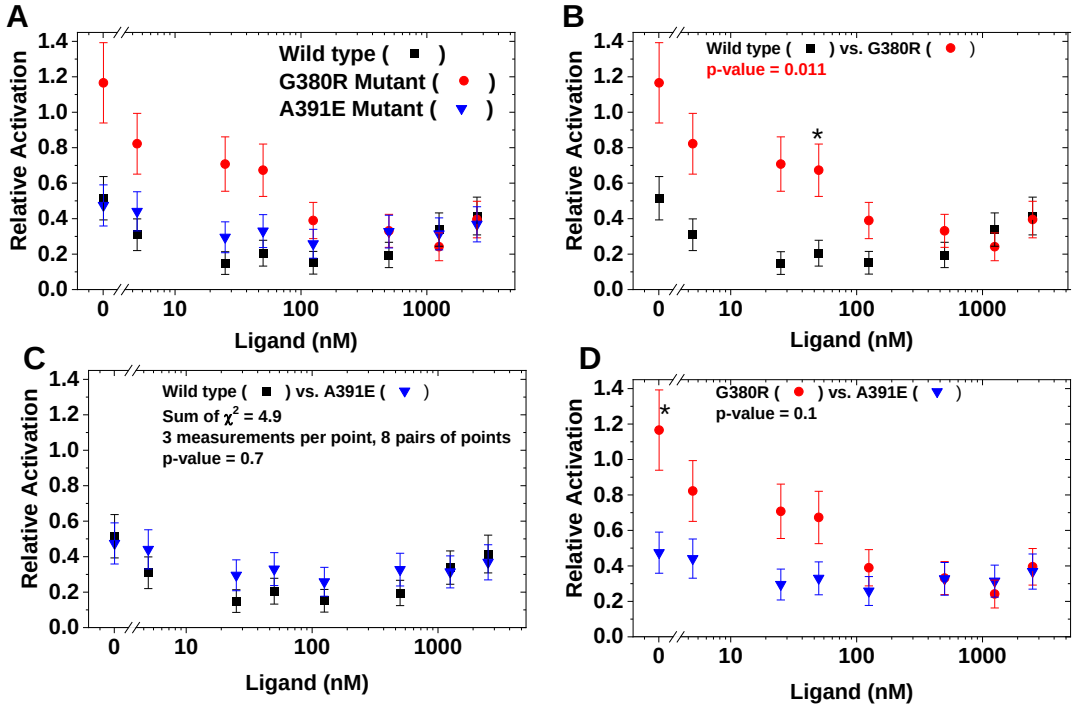

Figure 4. Relative activation curves calculated by taking the ratio of -ECTMwt/+ECTMwt from the previously published experiments, shown in **Fig. 3**. A value of 1 indicates no inhibition by ECTMwt and a value of 0 indicates complete inhibition. For each point, three independent ratios were averaged. Values shown are means and error bars are SE.

In **Fig. 3**, we show the phosphorylation versus ligand concentration for WT, A391E and G380R, with and without the inhibitory construct. Each point is comprised of three independent measurements of phosphorylation, and is shown as mean ± SEM.

We applied the modified χ^2 algorithm, described above, to the three pairs of curves. All pairwise comparisons are highly significant and give overall p-values of less than $1x10^{-6}$. The significance of each individual pair of points is calculated using a 2-sample t-test as described. Differences, with $p<0.05$ are shown with a black asterisk and those with $p<\alpha_{corr}$, the Bonferroni corrected cutoff of 0.0065 for the eight comparisons, are shown with a red asterisk.

Relative activation curves.

Relative activation of Wild type and G380R mutant receptors								
	Wild Type FGFR3				**G380R Mutant FGFR3**			
Ligand (nM)	**Mean**	**SD**	**N**	**SE**	**Mean**	**SD**	**N**	**SE**
0	**0.515**	**0.212**	**3**	**0.122**	**1.166**	**0.392**	**3**	**0.227**
5	**0.309**	**0.155**	**3**	**0.089**	**0.822**	**0.297**	**3**	**0.172**
25	**0.149**	**0.111**	**3**	**0.064**	**0.708**	**0.265**	**3**	**0.153**
50	**0.205**	**0.126**	**3**	**0.073**	**0.673**	**0.256**	**3**	**0.148**
125	**0.151**	**0.111**	**3**	**0.064**	**0.390**	**0.177**	**3**	**0.102**
500	**0.195**	**0.123**	**3**	**0.071**	**0.331**	**0.161**	**3**	**0.093**
1250	**0.338**	**0.163**	**3**	**0.094**	**0.241**	**0.136**	**3**	**0.079**
2500	**0.415**	**0.184**	**3**	**0.106**	**0.395**	**0.179**	**3**	**0.103**

Table 1. Example data from the analysis shown in **Fig 4**. Mean relative activation data (described in[5]) are given here for the pair of curves shown in **Fig. 4B**. At each ligand concentration, we measured relative activation in 3 independent experiments. The data in the table are the mean, SD, N and SE for each combination of mutant and ligand concentration.

When we compared the degree to which each sequence, WT, A391E and G380R, is inhibited by the enzymatically inactive construct, the curves are superficially similar as shown in **Fig. 4A.** The Chi-squared approach we describe here was initially developed to address whether these curves are significantly different[5], however the correction for the deviation from normality was not applied at that time. Here, we reanalyze the three comparisons with the fully developed Chi-squared approach.

All points in **Fig. 4** are the result of 3 independently measured ratios of activation with and without the inactive construct. Means and standard errors are shown. There are 8 pairs of points in each comparison. The three activation curves are shown together in **Fig. 4A** and the pairwise comparisons and modified Chi squared test results are shown in **Fig. 4B-D**. Because we are making three pairwise comparisons, we apply the Bonferroni correction to obtain an adjusted cutoff of $\alpha=0.05/3 = 0.017$, for statistical significance of the overall p-value. Based on this approach, we conclude that there is statistically significant, difference

between the inhibition of Wildtype versus the G380R mutant, that is only slightly smaller than the cutoff for significance, and that there is no significant difference the other comparisons of curves.

χ^2 based comparison of Wild type versus G380R mutant curves							
Ligand (nM)	**Difference**	**SE Diff**	**t-value**	**N/pt**	**dF**	**P-value**	**Chi Sqr**
0	**0.651**	**0.257**	**2.527**	**3**	**4**	**0.065**	**3.409**
5	**0.513**	**0.194**	**2.652**	**3**	**4**	**0.057**	**3.627**
25	**0.558**	**0.166**	**3.363**	**3**	**4**	**0.028**	**4.814**
50	**0.468**	**0.165**	**2.841**	**3**	**4**	**0.047**	**3.952**
125	**0.238**	**0.121**	**1.972**	**3**	**4**	**0.12**	**2.419**
500	**0.136**	**0.117**	**1.161**	**3**	**4**	**0.31**	**1.029**
1250	**0.097**	**0.123**	**0.791**	**3**	**4**	**0.47**	**0.514**
2500	**0.020**	**0.148**	**0.135**	**3**	**4**	**0.90**	**0.016**
					Sum of χ^2		**19.8**
					dF		**8**
					p-value		**0.011**

Table 2. Example calculations of the Chi-squared test applied to the two curves shown in Fig. 3B and in Table 1. The values in the table are the intermediate calculations in the Chi-squared test and the final result. In this case the sum of Chi-squared is 19.8, which at 8 degrees of freedom is equivalent to a p-value of 0.011.

Example data

In **Table 1** we provide the relative activation data that is plotted in **Fig. 4B** for Wild type FGFR3 and the G380R mutant. In **Table 2**, we provide the intermediate Chi-squared calculations outlined above. These validated example data and calculation results are provided for users to validate any spreadsheet or coded application of the modified Chi-squared method.

Restrictions and assumptions

A limitation of this modified Chi-squared approach for comparing arbitrary curves is that the two curves must have data measured at the same values of X to enable pairwise comparisons. Data from one curve without corresponding data at the same X in the other must be ignored, or an appropriate extrapolation of data and uncertainty must be carried out using nearby points. With large data sets containing many points, binning data is an effective way to ensure that curves can be compared at the same X-values. Bins do not have to be equally spaced, or equally populated, but binning can only be done over ranges of X for which the Y value does not change substantially.

A second significant restriction is that the data, at each value of X, must approach a Gaussian distribution for large N, an assumption that is usually satisfied in biological data sets. Dataset normality can be tested with a D-Agostino-Pearson or other normality test. Non-Gaussian data can be log-transformed, which frequently gives a normal distribution, when the linear data are not Gaussian. Use of log transformed data provides a rigorous solution to the problem of non-normal data.

Conclusions

We present here a simple modified Chi-squared based method to determine the statistical significance of the difference between two arbitrary curves. This simple and rigorous approach can be performed using a spreadsheet program and is applicable to any pair of X-Y datasets in which the X-values are matched in pairs and in which the standard errors of the individual points have been measured in independent experiments. A functional Excel spreadsheet is provided as Supplemental Information.

1. Bevington, P. R., *Data reduction and error analysis for the physical sciences*. McGraw-Hill Book Company: New York, 1969.
2. Pearson, K., On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine and Journal of Science* **1900,** *50* (302), 157-175.
3. Student, The probable error of a mean. *Biometrica* **1908,** *6*, 1-25.
4. Hochberg, Y.; Benjamini, Y., More powerful procedures for multiple significance testing. *Stat Med* **1990,** *9* (7), 811-8.
5. He, L.; Shobnam, N.; Wimley, W. C.; Hristova, K., FGFR3 heterodimerization in achondroplasia, the most common form of human dwarfism. *J Biol Chem* **2011,** *286* (15), 13272-81.