Acta Crystallographica Section A
Foundations of Crystallography

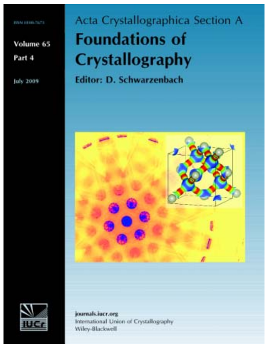
ISSN 0108-7673

Editor: **D. Schwarzenbach**

Probability plots based on Student's *t*-distribution

Rob W. W. Hooft, Leo H. Straver and Anthony L. Spek

Acta Cryst. (2009). A**65**, 319–321



Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 16 February 2009
Accepted 17 March 2009

Probability plots based on Student's *t*-distribution

Rob W. W. Hooft,[a]* Leo H. Straver[a] and Anthony L. Spek[b]

[a]Bruker AXS, PO Box 811, 2600 AV Delft, The Netherlands, and [b]Bijvoet Center for Biomolecular Research, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands. Correspondence e-mail: rob@hooft.net

The validity of the normal distribution as an error model is commonly tested with a (half) normal probability plot. Real data often contain outliers. The use of *t*-distributions in a probability plot to model such data more realistically is described. It is shown how a suitable value of the parameter ν of the *t*-distribution can be determined from the data. The results suggest that even data that seem to be modeled well using a normal distribution can be better modeled using a *t*-distribution.

1. Introduction

Abrahams & Keve (1971) introduced 'probability plots' as a tool in crystallography to verify how the distribution of errors in any set of observed values visually compares with a general presumed error distribution. This is done by creating a scatter diagram of observed *versus* theoretically expected deviations. In Abrahams & Keve (1971), and in many subsequent papers using this technique, the distribution that is used for the comparison is the Gaussian or 'normal' distribution, and the resulting probability plot is called a 'normal probability plot'.

Unfortunately the uncertainties in many day-to-day observations do not follow a normal distribution. Distributions encountered in real experiments often have a much larger incidence of highly deviating observations in the tails than predicted by the normal distribution. The incidence of deviations of at least 10σ following a normal distribution is extremely low (8×10^{-24}). These kinds of deviations, however, are encountered in practice and will result in a normal probability plot that shows an inverted S curve (Fig. 1).

The likelihood of outliers from a normal distribution is not only very small, but also counter-intuitive. The incidence of deviations of at least 11σ is approximately 40 000 times less likely than deviations of at least 10σ. This is completely contrary to experience: in practice it is observed that once a measurement deviates wildly from expected values, it does not make much difference by how much.

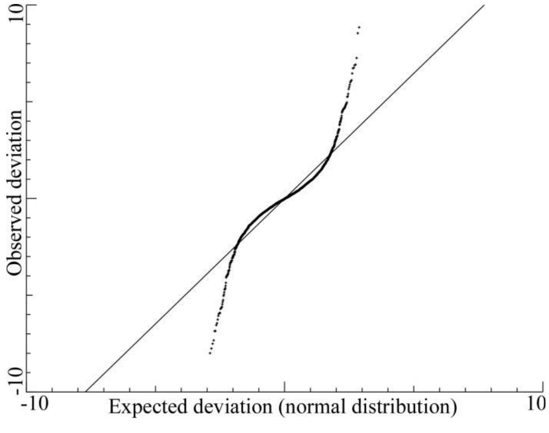


Figure 1
Normal probability plot of Bijvoet differences of a small-molecule crystal-structure data set obtained using a point detector, showing curves due to non-normal behavior of the errors. The diagonal straight line represents a least-squares fit; its slope is larger than 1.0.

It follows that to describe real-world experiments a distribution should be found that is more permissive of outliers. A good candidate is Student's *t*-distribution (Student, 1908). Originally, the *t*-distribution was derived to describe statistical experiments where the population variance must be estimated from a limited set of observations. Over the years, the *t*-distribution has found much wider applications than Student's original intention, most notably in robust statistical modeling of data (Lange *et al.*, 1989). Based on this, we propose to use it in a probability plot as well.

The *t*-distribution is modulated by a parameter ν ($\nu > 0$, not restricted to integer values). This parameter describes the number of degrees of freedom in the statistical sample. For $\nu = \infty$, the *t*-distribution is equal to the normal distribution. For lower values like ν = 10, the central part of the distribution hardly differs from the normal distribution, but the tails become very different (Fig. 2). At ν = 10, a deviation of at least 10σ has a likelihood of 8×10^{-7}, and deviations of at least 11σ are only 2.4 times less likely than that. Overall, lower values of ν will result in distributions that are more permissive of outliers. Experiments with different fixed values of ν to model real data have been reported in the literature (*e.g.* Yuh & Hogg, 1988).

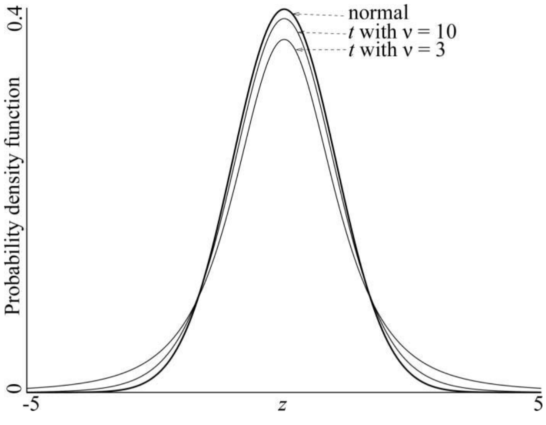


Figure 2
Probability density function of a normal distribution and *t*-distributions with two different values of ν.

doi:10.1107/S0108767309009908

In this paper we will detail how a probability plot can be based on Student's *t*-distribution. We suggest calling this a '*t*-probability plot' or tPP. We will not derive the basis of the *t*-distribution nor validate our use of this distribution for our goals.

2. Method

Because there is more complex mathematics involved when using the Student's *t*-distribution than when using the normal distribution, we will explain each of the steps that are needed to use this distribution in a probability plot. We will first describe the *t*-distributions and compare them with the normal distribution. Along the way we will derive what is needed to use any distribution as the basis for a probability plot. After that, we will show how the ν parameter can be used.

2.1. Calculating probability functions

The most natural way to look at a distribution function is to describe its probability density function (PDF). The PDF for the normal distribution is

$$\mathrm{pdf}_n(z) = [1/(2\pi)^{1/2}]\exp(-z^2/2). \tag{1}$$

Herein z is the variable of the so-called 'standard normal' distribution with mean value of 0 and a standard deviation of 1, and it can be obtained from any normal deviate x using the transformation

$$z = (x - \mu)/\sigma. \tag{2}$$

Herein μ is the 'correct' or 'expected' value for x, which can be approximated as $\langle x \rangle$ in the case of a homogeneous population. Similarly, σ is the expected standard uncertainty and can be approximated by the square root of the population variance $[\langle s^2 \rangle^{1/2}]$ in the case of a homogeneous population. The PDF for the *t*-distribution is

$$\mathrm{pdf}_t(z|\nu) = \frac{\Gamma[(\nu+1)/2]}{(\nu\pi)^{1/2}\,\Gamma(\nu/2)}\left(1+\frac{z^2}{\nu}\right)^{-(\nu+1)/2}. \tag{3}$$

In this formula Γ constitutes the gamma function, a mathematical extension of the factorial to real numbers.[1]

To calculate a probability, the integral over the PDF for the appropriate interval must be computed. The integral with lower bound $-\infty$ is called the cumulative distribution function (CDF). The integral of the PDF for any interval can be computed as the difference between two values of the CDF. The CDF for the normal distribution can be expressed by means of the *error function*:

$$\mathrm{erf}(x) = (2/\pi^{1/2})\int_0^x \exp(-t^2)\,\mathrm{d}t. \tag{4}$$

The error function and the complementary error function ('erfc') are often used and are included in many standard mathematical libraries. The CDF for the normal distribution is given as

$$\mathrm{cdf}_n(z) = (1/2)[1 + \mathrm{erf}(z/2^{1/2})], \tag{5}$$

but is more conveniently calculated (especially for $z \ll 0$) as

$$\mathrm{cdf}_n(z) = (1/2)[\mathrm{erfc}(-z/2^{1/2})]. \tag{6}$$

The CDF for the *t*-distribution is

$$\mathrm{cdf}_t(z|\nu) = \frac{1}{2} + x\Gamma\left(\frac{\nu+1}{2}\right)\frac{{}_2F_1[1/2, (\nu+1)/2; 3/2; -z^2/\nu]}{(\pi\nu)^{1/2}\,\Gamma(\nu/2)}. \tag{7}$$

In this equation, ${}_2F_1$ is the hypergeometric function (*e.g.* Abramowitz & Stegun, 1972; *Wikipedia*, 2009)

To be able to use a probability distribution in a probability plot, it is necessary to calculate the value of z that corresponds to a known value of the CDF. The function required to perform this calculation is called the inverse CDF (cdf^{-1}). The inverse CDF of the normal distribution cannot be written in closed functional form, but is readily available as an approximated function with sufficient accuracy in libraries for many programming languages. Unfortunately, the situation with the *t*-distribution is not so easy, especially since there are infinitely many *t*-distributions for different values of ν. The only practical approach is to implement the inverse CDF of the *t*-distribution as an iterated (binary) search using the CDF.

The inverse CDF is defined for values between 0 and 1. The values $p_1 \ldots p_N$ for the horizontal (expected) axis of the probability plot with N data points are calculated as

$$p_i = \mathrm{cdf}^{-1}(x_i) \tag{8}$$

with

$$x_i = (i - 1/2)/N. \tag{9}$$

2.2. Choice of the number of degrees of freedom

Having described the functions involved in equations (7) and (8), we can now make a probability plot based on a *t*-distribution. What is still missing is a method for estimating the value for ν. For the original purpose of the *t*-distribution, ν is the number of degrees of freedom of the data set; most often two less than the number of data points. When the *t*-distribution is used for robust statistical modeling, as in this paper, the best choice of ν is not obvious. Different practical ranges have been suggested in the literature. Yuh & Hogg (1988) suggested using ν = 11 for lightly tailed distributions and ν = 3 for heavily tailed distributions. They also suggested how to decide whether a distribution has a light or heavy tail.

In the case of a probability plot the situation is easier. We can make different probability plots corresponding to different values of ν. The best probability plot corresponds most closely to a straight line.

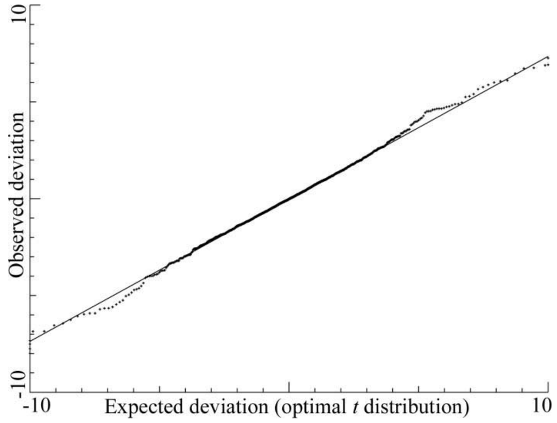


Figure 3
t-probability plot for the same data as shown in Fig. 1. A *t*-distribution with ν = 2.2 was found to be optimal to model this data set. The least-squares line shows that the slope is much smaller than 1.0, showing overestimation of the standard uncertainty. Some points with expected deviations of larger than 10σ have been left out of the plot to make the axes identical.

[1] Fortran code implementing all functions described in this section can be obtained from Netlib, http://www.netlib.org/random/dcdflib.f.tar.gz. Equivalent python code is available from the IUCr electronic archives (Reference: ZM5057). Services for accessing the code are described at the back of the journal.

Without having studied alternatives, we propose to choose a value of ν that maximizes the linear correlation coefficient of the probability plot. Maximizing the correlation coefficient focuses on the mass of data at the center of the distribution without ignoring deviations at the tails.

Abrahams & Keve (1971) remark that the positions of the extreme points in the probability plot are very sensitive to small changes in the measured values. These points can therefore disturb the determination of the value of ν at which the correlation coefficient is maximized. To avoid instability in the optimization, it would be possible to use a downweighting procedure that takes the uncertainty in each point into account in a quantitative way. In practice, however, for large data sets a very simple but seemingly arbitrary cutoff of five extreme data points at both ends gives sufficient stabilization.

3. Results and discussion

Fig. 3 gives the optimized *t*-probability plot for the same data set as represented in the normal probability plot of Fig. 1. The slope of the linear regression line is reduced from 1.33 for the normal probability plot to 0.76 for the optimized *t*-probability plot, showing that standard uncertainties have not been underestimated but are overestimated for the bulk of the data points. The correlation coefficient for the regression line increases from 0.92 to 0.998, showing a dramatic improvement of the error model.

Over the course of our studies we have analyzed many data sets. A few data sets required non-normal treatment of the standard uncertainties as became obvious from studying their normal probability plots. The description of the errors for these data sets could all be very significantly improved by use of a *t*-distribution as modeled in an optimized *t*-probability plot. Optimized values of ν for these data sets ranged between 2.3 and 5.6.

More surprisingly, we have found that the error model for almost all of the data sets that could be adequately described using a normal distribution could be significantly improved by using a *t*-distribution. Such data sets, identified by normal probability plots with correlations of their regression lines larger than 0.999, had significantly better correlation coefficients in a *t*-distribution plot with optimized values of ν ranging between approximately 12 and 30.

Our results show that we can always use an optimized *t*-probability plot where one would normally use a normal probability plot to model the standard uncertainties of a data set. The normal probability plot forms the limiting case at $\nu = \infty$ and does not need to be handled as a special case. In all but one of the data sets we have analyzed so far the optimization converged to $\nu < 100$, and significantly better fits were obtained than at $\nu = \infty$.

We have not studied how the optimization of ν can be performed in the case of smaller data sets. Our data sets generally contain many thousands of data points. We expect that the same procedure can be used for data sets as small as 100 points; with smaller data sets the difference between probability plots will become smaller and a simple binary decision about an appropriate value for ν as made by Yuh & Hogg (1988) may be more appropriate.

In the case where ν can be determined directly from the data, this may provide interesting information about the reliability of the experimental methods used to obtain or process the data. We have not studied this.

In cases when the reliability of the error model is of utmost importance the use of *t*-probability plots to model the standard uncertainties can improve the reliability of the calculations.

4. Conclusions

We have proposed a way of studying the standard uncertainties for large data sets that allows robust modeling of the data including any outliers. The method consists of an analysis of the errors by means of a probability plot using Student's *t*-distribution to provide expected deviations. We have shown that it is possible to determine the parameter ν of the *t*-distribution from the data themselves. We have seen that this procedure always improves the error modeling, even for data sets that, at first glance, would appear to behave in accordance with a normal distribution.

References

Abrahams, S. C. & Keve, E. T. (1971). *Acta Cryst.* A**27**, 157–165.

Abramowitz, M. & Stegun, I. A. (1972). Editors. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, 9th printing, ch. 15, pp. 555–566. New York: Dover.

Lange, K. L., Little, R. J. A. & Taylor, J. M. G. (1989). *J. Am. Stat. Assoc.* **84**, 881–896.

Student (1908). *Biometrika*, **6**, 1–25.

Wikipedia (2009). *Hypergeometric series*, http://en.wikipedia.org/wiki/Hypergeometric_series.

Yuh, L. & Hogg, R. B. (1988). *Biometrics*, **44**, 433–445.