Experimental Noise in Data Acquisition and Evaluation

I. Noise Amplitude Distribution Function on High Resolution FT-NMR Instruments

II. Signal-to-Noise Measurements in NMR

Stanislav Sykora and Jürgen Vogt

Bruker Report, Issue **2**, 15-18 (1979)

Abstract:

I. The noise reaching the analog-digital converter of a HR-NMR instrument has been analyzed for any deviations from normal distribution. Several statistical normality tests had been employed, such as skewness, kurtosis, second and fourth moment and the correlation between them. It has been found that any such deviation, if present at all, is below the threshold of detection at 95% confidence level. Tentative reasons for this fact are listed. The fact that the noise is indeed normal implies that standard data averaging techniques are indeed the optimal ones and it has no sense to look for better ones.

II. The simple way of estimating the signal-to-noise ratio (S/N) in HR-NMR spectra consists of estimating the height of a standard peak (p) and the peak-to-peak amplitude of a segment of the spectrum containing only noise (n) and setting S/N = 2.5*(p/n). This Note investigates to which extent such a procedure can be considered objective and reliable. The maximum and the minimum of a segment of noise with n data points is a well-defined random variable whose statistical properties can be easily analyzed. In particular, its median and quantiles for any confidence level are given by explicit formulae. From these it follows that the standard method is depends too much of the size of the noise segment and should be therefore replaced by a more sophisticated method.

Keywords: Noise, Experimental data, NMR

Experimental Noise in Data Acquisition and Evaluation

I. Noise Amplitude Distribution Function on High Resolution FT-NMR Instruments.

by

S. Sýkora, Bruker Spectrospin Italiana Srl,
Via Miglioretti 2, 20161 Milano, Italy.

J. Vogt, Spectrospin AG, Industriestrasse 26,
8117 Zürich-Fällanden, Switzerland.

This being the first of a series of articles regarding various aspects of experimental noise in data acquisition and evaluation systems, we would like to start with a short explanation of the aim and the scope of the whole series. The project started with a plan to write a comprehensive review of data evaluation problems in presence of noise. Soon it turned out, however, that only the foreword to such a review requires a dozen or so pages. We have therefore decided to postpone the synthetic theoretical approach to a later stage and to start with a series of short studies on well-delimited practical problems.

In this paper we investigate the amplitude distribution function of the noise reaching the A/D converter of a high-resolution NMR spectrometer.

The noise in the receiver channel of an NMR spectrometer arises from a number of sources: noise from the input transistor of the preamplifier, thermal noise from the equivalent shunt resistance of the tuned probehead[1], noise arising from the sample, combined noise from all the receiver components (mixers, gates, IF and AF amplifiers, relay contacts, filters, cables, etc.), pick-up noise from other parts of the spectrometer (switching transients of logic circuitry, relays, thyristors, etc.), and pick-up noise from the surroundings (atmospheric discharges, power line transients, etc.). Even though the noise from some of these sources has been investigated theoretically, the combined effect can really be estimated only empirically.

Considered as a random function of time, any noise is a very complex phenomenon which can sometimes tell us as much about its source as any old-fashioned signal. The limited frequency response of any electronic circuitry implies that any voltage or current noise is always a continuous function of time. Even so, its full description requires an infinite set of parameters[2]. Contrary to intuitive statements which occasionally appear in the literature, no such thing as perfect randomness exists for the simple reason that it cannot be defined.

There are certain simple properties, however, which experimental noise often satisfies to a good degree of approximation. Thus a noise is said to be **stationary** if its properties are independent of any shift in the origin of time. In this study we will assume that this is indeed the case (possible exceptions: some piezoelectric and ferroelectric powders can produce a terrible transient noise decaying for quite a long time after each pulse). A noise is called **white** if the properties of its Fourier transform (a random function by itself) are independent of any shift of frequency origin and/or phase. Obviously, a white noise would spread over **all** frequencies and thus carry infinite mean energy, which is impossible except as an approximation. The autocorrelation function of a white noise is necessarily an infinitely sharp delta-function. In this study, though, we will not investigate the noise autocorrelation properties; we will simply assume that – chosing the final stage filters wider than the sweep width – the correlation between two successive points of the digitized noise is too small to be of any importance. Finally, a noise is called **normal** if its sampled values are distributed according to the normal (Gaussian) distribution function[3]. It is this aspect of the noise which we have tested.

Note that the properties of stationarity, whiteness, and normality are totally unrelated among themselves; none of them does a-priori imply or exclude any other. Moreover, even a noise function satisfying all these properties cannot be considered "perfectly random" since it is not unique. Many such functions can be devised, differing widely among themselves in other aspects.

On NMR spectrometers of the WH- and WP-series, the signal passes through a number of narrow-band stages (probehead, preamp, IF-amplifier) followed by the output filters at the AF level. This represents a chain of filters of different types. It is well known[4] that, given any random function with finite standard deviation, the application of **any** kind of filter always brings its amplitude distribution function closer to the normal distribution. Thus, even if the noise **at the source** were non-normal, we would still expect it to be nearly normal at the A/D converter input. Given the extreme similarity between all WH- and WP-type instruments, it is sufficient to carry out the normality tests at only one of them. We have chosen the WP-80 spectrometer (H1 channel, $CHCl_3/CDCl_3$ test sample).

A total of six single-scan tests have been run at widely different settings of sweep width (dwell time) and filter width. For each of the sweep widths chosen, two filter widths were adopted. One was the standard value set automatically by the FT-NMR program, while the other was about three times larger. In each test, 16K (i. e. 16384) data points were taken. After correcting these data for any DC offset, the second (m_2), third (m_3), and fourth (m_4) moments were calculated. These were used to calculate[5] the skewness

$$\gamma_1 = m_3/(m_2)^{3/2} \qquad (1)$$

and the curtosis

$$\gamma_2 = m_4/(m_2)^2 - 3 \qquad (2)$$

of the sample. These are both very sensitive indicators of any deviation from normality as far as the symmetry (γ_1) and the center/wings balance (γ_2) of the distribution are concerned. For normal distribution $\gamma_1 = \gamma_2 = 0$. The experimental data are reported in Table 1.

The question is whether the deviations from zero are statistically significant or not. In order to answer, we must estimate the expected deviations under the assumption of normal parent population. Let us use the brackets [. . .] to denote

Table 1. Summary of the noise normality tests

No.	Sweep width (Hz)	Filter width (Hz)	γ_1	γ_2	$\chi^2/\nu; \nu=25$
1	60	100	−0.007	+0.065	0.51
2	60	300	−0.005	−0.053	1.55
3	3000	3800	+0.015	−0.048	1.76
4	3000	15000	−0.026	−0.056	1.55
5	31250	39100	+0.019	+0.021	2.04
6	31250	125000	−0.001	+0.009	0.76
	***	Expected values:	0.000	0.000	1.00
		Probable errors $\varepsilon(\gamma)$:	±0.019	±0.038	–
	95 %	Confidence intervals:	±0.037	±0.074	1.51

*** The probable errors of γ_1 and γ_2 were calculated for samples of 16384 points.

mean values. If r is a random variable (in our case the noise amplitude) then the moments (provided they exist) are defined as

$$\mu_k = [r^k]. \qquad (3)$$

The variance of r^k, denoted as Δ_k^2, is given by

$$\Delta_k^2 = [(r^k - [r^k])^2] = [r^{2k}] - [r^k]^2$$
$$= \mu_{2k} - \mu_k^2. \qquad (4)$$

Now, if a sample of N numbers $\{r_i, i = 1, 2, \ldots, N\}$ is taken, Δ_k^2 exists, and N is a large number, then – by the central limit theorem[3] – the k-**th** *moment of the sample*

$$M_k = (1/N) \sum_{i=1}^{N} r_i^k \qquad (5)$$

has an approximately normal distribution with center at μ_k and standard deviation σ_k given by

$$\sigma_k^2 = \Delta_k^2/N = (\mu_{2k} - \mu_k^2)/N. \qquad (6)$$

For the normal distribution with standard deviation σ and center at zero, μ_k is zero for odd k and

$$\mu_{2n} = (1/\sigma\sqrt{2\pi}) \int_{-\infty}^{+\infty} r^k \exp\left(-\tfrac{1}{2} r^2/\sigma^2\right) dr$$
$$= \sigma^{2n}(2n)!/(2^n\, n!). \qquad (7)$$

Hence

$$\mu^1 = 0, \mu_2 = \sigma^2, \mu_3 = 0, \mu_4 = 3\sigma^4, \quad (8\,a)$$

and

$$\sigma_k = \beta_k \sigma^k, \qquad (8\,b)$$

where

$$\beta_1 = \sqrt{1/N},\ \beta_2 = \sqrt{2/N},$$
$$\beta_3 = \sqrt{15/N},\ \beta_4 = \sqrt{96/N}. \qquad (8\,c)$$

In order to estimate the expected errors in γ_1 and γ_2, we must take into account the fact that by eliminating the DC offset we have **forced** m_1 to be zero. This implies an error in the offset parameter equal to $-M_1$ which propagates to the higher moments according to the well-known formulae[3]

$$m_2 = M_2 - M_1^2, \qquad (9\,a)$$
$$m_3 = M_3 - 3M_1M_2 + 2M_1^3, \qquad (9\,b)$$
$$m_4 = M_4 - 4M_1M_3 + 6M_1^2M_2 - 3M_1^4. \qquad (9\,c)$$

Substituting Eqs. 9 into Eq. 1, expanding with respect to deviations δM of the M's from their means, and truncating the expansions at terms decreasing as $N^{-1/2}$ for large N, we obtain

$$\gamma_1 = m_3/(m_2)^{3/2}$$
$$\doteq (\delta M_3)/\sigma^3 - 3(\delta M_1)/\sigma. \qquad (10)$$

γ_1 is therefore a difference of two random variables with standard deviations β_3 and $3\beta_1$. The calculation of the standard deviation of γ_1 is further complicated by the fact that, as could be expected, δM_1 and δM_3 are not independent. Their correlation coefficient $\rho_{1,3}$ is given by

$$\rho_{1,3} = [(r-[r])(r^3-[r^3])]/\Delta_1\Delta_3$$
$$= [r^4]/\Delta_1\Delta_3 = 3/\sqrt{15.} \qquad (11)$$

The expected error in γ_1 is therefore

$$\varepsilon(\gamma_1) \doteq [\beta_3^2 - 2\rho_{1,3}\beta_3(3\beta_1) + (3\beta_1)^2]^{1/2}$$
$$= \sqrt{6/N.} \qquad (12)$$

An analogous calculation must be carried out for γ_2. The expansion of γ_2 in terms of δM's is

$$\gamma_2 \doteq (\delta M_4)/\sigma^4 - 6(\delta M_2)/\sigma^2. \qquad (13)$$

The correlation coefficient $\rho_{2,4}$ between δM_4 and δM_2 is easily calculated as

$$\rho_{2,4} = [(r^2-[r^2])(r^4-[r^4])]/\Delta_2\Delta_4 =$$
$$= ([r^6]-[r^2][r^4])/\Delta_2\Delta_4 = 3/\sqrt{12.} \qquad (14)$$

The expected error in γ_2 is therefore

$$\varepsilon(\gamma_2) \doteq [\beta_4^2 - 2\rho_{2,4}\beta_4(6\beta_2) + (6\beta_2)^2]^{1/2}$$
$$= \sqrt{24/N.} \qquad (15)$$

Comparing the numerical values of the expected errors (Table 1) with the experimental data we see that there is no statistically significant deviation from normal distribution as far as the skewness (symmetry) and curtosis (center/wings balance) are concerned.

Since skewness and curtosis are just two global parameters, certain types of deviations from normality might still have passed unnoticed.

We have therefore decided to complement each test with a still more powerful check for normality. This consisted in constructing a 49 channel histogram of the experimental sample, with each channel having the width of 0.25 σ, σ being the standard deviation of the sample. The channels were numbered from −24 to +24. The occurence frequencies in the lowest channels from −13 to −24 were summed together and the sum denoted o_l. Analogously, o_u was defined as sum of the occurence frequencies in the highest channels from +13 to +24. The occurence frequencies in the remaining channels were denoted as o_k, where k = −12, −11, ..., +11, +12. The expected occurence frequencies e_k were calculated for the same channels by expanding the normal probability density function in each interval into a Taylor series and integrating over the interval. The resulting formula is

$$e_k = N\,(0.0994758 + 0.0000162\,k^2)$$
$$\exp(-k^2/32). \qquad (16)$$

The expected frequency $e = e_l = e_u$ was simply read from the tables of normal distribution function: e = 0.000893 N. The expected and observed occurence frequencies were then used to calculate the χ^2 value using the Yates's formula[6]

$$\chi^2 = \sum_{k=1,-12}^{+12,u} (|o_k - e_k| - \tfrac{1}{2})^2/e_k. \qquad (17)$$

The number of degrees of freedom is in this case $\nu = 25$ since two parameters (sum of all occurence frequencies and centre of the distribution) are fixed. The reduced χ^2/ν

values are reported in the last column of Table 1.

The χ^2 test did reveal significant deviations from normality in several cases (see references [5, 6, 7] for tables of the χ^2 distribution). In one case (test No. 5), the deviation was significant even at the 99.5% level, and three other cases superated the 95% significance level.

A more detailed analysis of the problem has shown, however, that these deviations are due simply to the digitization round-off errors. The standard deviations of the samples varied in fact between 60 and 90 steps of the A/D converter. This gives about 15 steps for each histogram channel, so that the round-off error may reach in the worst case (1/30) **th** of the actual occurence frequency. A simple calculation shows that this may increase the χ^2/ν values for our samples by as much as 0.7 – enough to account for the observed anomalies.

We may therefore conclude by stating that even a very detailed and careful analysis did not reveal any significant deviation of the noise amplitude distribution from the normal distribution function, except for the obvious interference of the digitization round-off process.

So far we were concerned only with the time-domain noise in a single scan. Since the sum of any number of random variables with normal distribution is itself a normally distributed random variable, our result implies that the noise remains normal even after an accumulation of any number of scans. It is in fact true that even if the single scan noise were not normal, the accumulation process would still lead to a normal noise in the limit of large number of scans (a straightforward consequence of the central limit theorem [3]).

By a slight modification of the central limit theorem, it is also possible to show [4] that for large number of data points the Fourier transform of **any** noise approaches a normal random function. Whenever the t-domain noise is normal, the resulting ω-domain noise is always exactly normal.

By demonstrating the normality of the single-scan noise, we have therefore proved also the normality of the t-domain noise in accumulated data, as well as the normality of any ω-domain noise. Even though the experimental part of this study regards only the WP-80 instrument, the theoretical arguments invoked above guarantee that the results are likely to apply to any modern high-resolution FT-NMR instrument.

References

1. Abragam A., The Principles of Nuclear Magnetism, Chapter IIIB, Clarendon Press, Oxford 1961.
2. Wang M. C., Uhlenbeck G. E., Rev. Mod. Phys. **17**, 323 (1945); reprinted in "Noise and Stochastic Processes", Ed. Wax N., Dover, New York 1954.
3. Feller W., An Introduction to Probability Theory and Its Applications, 2 volumes, J. Wiley, New York 1968. For the central limit theorem, see Chapter VIII.
4. Bracewell R., The Fourier Transform and Its Applications, Chapter 16, McGraw-Hill, New York 1965.
5. Abramowitz M., Stegun I. A., Editors, Handbook of Mathematical Functions, Section 26, Dover, New York 1972.
6. Alder H. L., Roessler E. B., Introduction to Probability and Statistics, W. H. Freeman & Co., San Francisco 1968.
7. Bevington P. R., Data Reduction and Error Analysis for the Physical Sciences, McGraw-Hill, New York 1969.

Experimental Noise in Data Acquisition and Evaluation

II. Signal-to-Noise Ratio Measurements in NMR.

by

S. Sýkora, Bruker Spectrospin Italiana Srl,
Via Miglioretti 2, 20161 Milano, Italy.
J. Vogt, Spectrospin AG, Industriestrasse 26,
8117 Zürich-Fällanden, Switzerland.

The signal-to-noise ratio S_N is best defined as

$$S_N = kS/\sigma, \qquad (1)$$

where k is a conventional constant, S is any quantity proportional to signal intensity (usually the peak height), and σ is the standard deviation (root mean square) of the noise.

In early days of NMR, when computer evaluation of the data was a rarity, Varian Associates suggested using the empirical formula

$$S_N = 2.5\ S/N_{pp}, \qquad (2)$$

where N_{pp} is the peak-to-peak amplitude of a chosen segment of the noise. This formula has found wide acceptance and it is in general use even today.

In this paper we will show that despite its popularity, Formula (2) is inherently contradictory. We will eventually proceed to suggest a different approach to the S_N evaluation which, apart from being objective, precise, and free of contradictions, should suit well all feasible purposes.

Consider a digitized noise with normal amplitude distribution (we have shown [1] in Part I of this series that the noise in high-resolution FT-NMR is indeed normal to a very high degree of precision). Suppose n mutually uncorrelated points $\{r_i,\ i = 1, 2, \ldots, n\}$ are taken into account. We would like to know the distribution function for the quantity

$$N_{pp} = \mathrm{Max}_i\,(r_i) - \mathrm{min}_i\,(r_i). \qquad (3)$$

This is a typical range statistic. The distribution function of N_{pp} is found[2,3] to be

$$D_n(x) = n\int [F(x+y) - F(y)]^{n-1}\, d\{F(y)\}, \qquad (4)$$

where the Stieltjes integral is employed and F(x) is the distribution function of the parent population from which the r's are drawn. In our case F(x) is the normal distribution function

$$F(x) = (1/\sigma\sqrt{2\pi})\int_{-\infty}^{x} \exp\left(-\tfrac{1}{2}y^2/\sigma^2\right) dy. \qquad (5)$$

Unfortunately, Eq. 4 is rather cumbersome and unsuitable for explicit calculations. The complexity can be traced primarily to the correlation between the maximum and minimum values of r_i for small n. Since it is obvious that this correlation must vanish for large N, we can restrict the analysis to this limit and investigate the distribution functions of the quantities

$$M = \mathrm{Max}_i\,(r_i) \text{ and } m = \mathrm{min}_i\,(r_i). \qquad (6)$$

For symmetric parent distributions, the distribution function of m is identical to that of M with changed sign of the argument, so that it is sufficient to analyse only the latter.

The distribution function $Z_n(x)$ of M can be easily deduced[2,3]. By definition, $Z_n(x)$ is the probability that M is smaller than x, which is obviously equal to the product of the separate probabilities that each r_i in the sample is smaller than x. Hence

$$Z_n(x) = [F(x)]^n. \qquad (7)$$

Even though this formula is not suitable for determination of the mean value of M, it allows an easy explicit determination of its most probable value and of any of its quantiles (the a-quantile, $0 \leqq a \leqq 1$, is that value of x for which the probability that $M < x$ is equal to a).

Considering that $dZ_n(x)/dx$ is the probability density function of M, the most probable value $M^*(n)$ of M is given by the solution of the equation

$$d^2Z_n(x)/dx^2 = 0. \qquad (8)$$

From Eqs. 7 and 8 it follows that $M^*(n)$ is the root of the equation

$$P(x) = 1-[F(x)F''(x)]/[F'(x)]^2 = n. \qquad (9)$$

Scaling the parent distribution in such a way that $\sigma = 1$, Eq. 5 becomes

$$F(x) = \tfrac{1}{2}[1+\mathrm{erf}(x/\sqrt{2})] \qquad (10)$$

and

$$P(x) = 1+\sqrt{2\pi}F(x)\,x\exp(x^2/2). \qquad (11)$$

The function P(x) has been plotted in Fig. 1, using the approximation to the error function reported in reference[4]. According to Eq. 9, the most probable value $M^*(n)$ of M for n data points can be read out from Fig. 1 as that value of x at which $P(x) = n$.

The quantiles are even easier to determine. The a-quantile $M_a(n)$ is the solution of the equation

$$Z_n(x) = a. \qquad (12)$$

It follows from Eqs. 7 and 12 that $M_a(n)$ is the root of the equation

$$Q_a(x) = \ln(a)/\ln F(x) = n. \qquad (13)$$

The functions $Q_a(x)$ are plotted in Fig. 1 for a=0.01, 0.05, 0.5 (median), 0.95, and 0.99. The readout of the $M_a(n)$ value from the graph of $Q_a(x)$ is identical to the readout of $M^*(n)$ from the graph of P(x).

As an example, consider $n = 1000$ (a likely practical value). From Fig. 1 we determine that the most probable value of M is $M^*(1000) = 3.1\,\sigma$ and the median of M is $M_{0.5}(1000) = 3.2\sigma$. M will be smaller than $M_{0.01}(1000) = 2.6\,\hat{\sigma}$ in only 1 % of cases and larger than $M_{0.99}(1000) = 4.3\sigma$ also in only 1 % of cases.

According to what was said above about the lack of correlation between M and m for large values of N, the most probable peak-to-peak value (range) $N^*_{pp}(n)$ is approximately $2M^*(n)$. In the above example $N^*_{pp}(1000) \doteq 6.2$. This, substituted into Eq. 2 and compared with Eq. 1, would imply $k = 2.5/6.2 = 0.40$. The problem is, however, that $N^*_{pp}(n)$ varies with the number of data points n. It is equal to 1.41 for $n = 2$, 4.8 for $n = 100$, 6.2 for $n = 1000$, 7.4 for $n = 10\,000$, and 8.6 for $n = 100\,000$. Moreover, the expected deviations in the N_{pp} values are considerable; for $n = 1000$, e. g., any value between 5.7 and 7.5 is quite likely. This, of course, brings to mind the everlasting discussions upon whether noise peaks can be ignored and, if so, which and how many.

In our opinion, the only solution of this problem consists in abandoning Formula (2) and returning to Formula (1). The estimation of $\hat{\sigma}$ for a chosen section of the noise can be made by the computer. The probable error of such an estimate is $\hat{\sigma}/\sqrt{N}$ which is at least one order of magnitude more precise than in the case of peak-to-peak value. The only obstacle is that the user should have sufficient manual and visual control over the whole procedure.

Our proposal consists in a routine incorporated into the standard FT-NMR software which would i) enable the user to specify a section of the spectrum free of any coherent signal, ii) calculate the standard deviation $\hat{\sigma}$ for this section and iii) substitute the data within the section by a step function offset with respect to the mean by $-C\hat{\sigma}$ in the first half and by $+C\hat{\sigma}$ in the second half of the interval. S_N would then be measured as 2.5 times the ratio of the signal S to the height of this step.

By chosing $C = 2.5$ (i. e., $k = 0.5$ in Eq. 1), we would ensure an approximate correspondence with the S_N values determined in the traditional way. Moreover, this would lead to placing the two horizontal lines of the step function at approximately the same position as in the peak-to-peak method (i. e., with just a few peaks of the noise exceeding the lines), thus enabling a visual check of the correctness of the calculation.

We invite a strong critical feedback from our readers. The proposal, if adopted, would eventually influence a procedure carried out daily in every NMR laboratory. It should therefore be discussed thoroughly before attempting to "push it" into practice.

1. Sýkora S., Vogt J., Bruker Report **2/79**
2. Freeman H., Introduction to Statistical Inference, Chapter 22, Addison-Wesley, Reading (Mass.) 1963.
3. Rao C. R., Linear Statistical Inference and Its Applications, Page 214, Problems 10 to 10.3, J. Wiley, New York 1973.
4. Abramowitz M., Stegun I. A., Editors, Handbook of Mathematical Functions, Eq. 7. 1. 25, Dover, New York 1972.

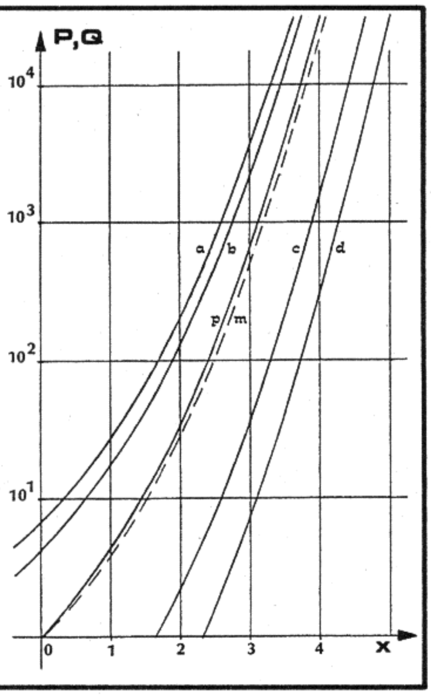

Figure 1

Graphs of the functions P (x) and Q_a (x). Curve p represents P (x), curve m represents the median $Q_{0.5}(x)$, and curves a, b, c, and d represent Q_a (x) for a=0.01 0.05, 0.95 and 0.99, respectively.