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 M. Gentili 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, 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 circuits, 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. 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 a 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 - choosing 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. 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 alone 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 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, CHCl3/CDC13 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 (m2), third (m3), and fourth (m4) moments were calculated. These were used to calculate the skewness

\[ \gamma_1 = m_3/m_2^{3/2} \]

and the curtosis

\[ \gamma_2 = m_4/m_2^{2} - 3 \]

of the sample. These are both very sensitive indicators of any deviation from normality as far as the symmetry (\(\gamma_1\)) and the center/wings balance (\(\gamma_2\)) 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 \([\ldots]\) to denote
Table 1. Summary of the noise normality tests

<table>
<thead>
<tr>
<th>No.</th>
<th>Sweep width (Hz)</th>
<th>Filter width (Hz)</th>
<th>( \gamma_1 )</th>
<th>( \gamma_2 )</th>
<th>( x^2/\nu; \nu = 25 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>60</td>
<td>100</td>
<td>-0.007</td>
<td>+0.065</td>
<td>0.51</td>
</tr>
<tr>
<td>2</td>
<td>60</td>
<td>300</td>
<td>-0.005</td>
<td>-0.053</td>
<td>1.55</td>
</tr>
<tr>
<td>3</td>
<td>3000</td>
<td>3800</td>
<td>+0.015</td>
<td>-0.048</td>
<td>1.76</td>
</tr>
<tr>
<td>4</td>
<td>3000</td>
<td>15000</td>
<td>-0.026</td>
<td>-0.056</td>
<td>1.55</td>
</tr>
<tr>
<td>5</td>
<td>31250</td>
<td>39100</td>
<td>+0.019</td>
<td>+0.021</td>
<td>2.04</td>
</tr>
<tr>
<td>6</td>
<td>31250</td>
<td>125000</td>
<td>-0.001</td>
<td>+0.009</td>
<td>0.76</td>
</tr>
</tbody>
</table>

**Expected values:**

<table>
<thead>
<tr>
<th>Probable errors ( \varepsilon ) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>95 % Confidence intervals:</td>
</tr>
<tr>
<td>±0.019</td>
</tr>
<tr>
<td>±0.038</td>
</tr>
</tbody>
</table>

The probable errors of \( \gamma_1 \) and \( \gamma_2 \) 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

\[ \mu_k = [r^k]. \]  

(3)

The variance of \( r^k \), denoted as \( \Delta^k \), is given by

\[ \Delta^k = [r^k - [r^k]^2]^2 = [r^{2k} - [r^k]^2] \]

(4)

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

\[ \gamma_1 = \text{moment} \]

(5)

Now, if a sample of \( N \) numbers \( \{r_i, i = 1, 2, \ldots, N\} \) is taken, \( \Delta^2 \) exists, and \( N \) is a large number, then by the central limit theorem \( \gamma \) is the \( k \)-th moment of the sample

\[ M_k = (1/N) \sum_{i=1}^{N} r_i^k \]

(5)

has an approximately normal distribution with center at \( \mu_k \) and standard deviation \( \sigma_k \) given by

\[ \sigma_k^2 = \Delta_k^2/N = (\mu_k - \mu_k^2)/N. \]

For the normal distribution with standard deviation \( \sigma \) and center at zero, \( \mu_k \) is zero for odd \( k \) and

\[ \mu_{2n} = (1/\sqrt{2\pi} \cdot 2^n \cdot \Gamma(n+1/2) \exp(-1/2 \cdot s^2) \cdot dr = s^{2n}(2n)/((2\pi n)) . \]

(7)

Hence

\[ r = 0, \mu_2 = \sigma^2, \mu_3 = 0, \mu_4 = 3 \sigma^4, \quad \text{(8a)} \]

and

\[ \sigma_k = \mu_k^{1/2}, \quad \text{(8b)} \]

where

\[ \beta_1 = \sqrt{2/N}, \beta_2 = \sqrt{2\pi/N}, \quad \beta_4 = \sqrt{96/N}. \]

In order to estimate the expected errors in \( \gamma_1 \) and \( \gamma_2 \), we must take into account the fact that by eliminating the DC offset we have

\[ m_0 = 0 \]

and curvatures (center/wings balance) are concerned.

Since skewness and curvatures 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 \( \sigma \), \( \sigma \) being the standard deviation of the sample. The channels were numbered from -24 to +24. The occurrence frequencies in the lowest channels from -13 to -24 were summed together and the sum denoted \( o_k \). Analogously, \( o_k \) was defined as sum of the occurrence frequencies in the highest channels from +13 to +24. The occurrence frequencies in the remaining channels were denoted as \( o_k \), where \( k = -12, -11, \ldots, 11 + 12 \). The expected occurrence 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

\[ e_k = N \cdot (0.00475 + 0.0000162 k^2) \exp(-k^2/32). \]

We then used the table of normal distribution function: \( e = 0.000893 N \). The expected and observed occurrence frequencies were then used to calculate the \( \chi^2 \) value using the Yates's formula

\[ \chi^2 = \sum_{k=-12}^{12} (|o_k - e_k| - 1/2)^2/e_k. \]

The number of degrees of freedom is in this case \( \nu = 25 \) with the two parameters (sums of all occurrence frequencies and centre of the distribution) are fixed. The reduced \( \chi^2 \) value
The $x^2$ test did reveal significant deviations from normality in several cases (see references 5, 6, 7) for tables of the $x^2$ distribution. In one case (test No. 5), the deviation was significant even at the 99.5% level, and three other cases surpassed 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 (1/300)th of the actual occurrence frequency. A simple calculation shows that this may increase the $x^2/n$ 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 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).

By a slight modification of the central limit theorem, it is also possible to show 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 $o$-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 o-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

Experimental Noise in Data Acquisition and Evaluation
II. Signal-to-Noise Ratio Measurements in NMR.
by S. Sjökvist, 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

$$ S/N = kS/\sigma $$

where $k$ is a conventional constant, $S$ is any quantity proportional to signal intensity (usually the peak height), and $\sigma$ 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

$$ S/N = 2.5 S/N_p $$

where $N_p$ 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 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

$$ N_p = \max (r_i - \min (r_i)) $$

This is a typical range statistic. The distribution function of $N_p$ is found to be

$$ D_n(x) = \int \left| F(x+y) - F(y) \right|^n dy $$

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

$$ F(x) = \frac{1}{\sqrt{2\pi} \sigma} \exp \left( -\frac{x^2}{2\sigma^2} \right) dy $$

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

$$ M = \max (r_i) \quad \text{and} \quad m = \min (r_i) $$

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. 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

$$Z_n(x) = F(x)^n.$$  \hfill (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, $0 \leq a \leq 1$, is that value of $x$ for which the probability that $M < x$ is equal to $a$).

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

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

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

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

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

$$F(x) = \frac{1}{2}(1 + \exp(-x^2/2)).$$  \hfill (10)

and

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

The function $P(x)$ has been plotted in Fig. 1, using the approximation to the error function reported in reference. 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 easier to determine. The $a$-quantile $M_a(n)$ is the solution of the equation

$$Z_a(x) = a.$$  \hfill (12)

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

$$Q_a(x) = \ln(a)/\ln(F(x)) = n.$$  \hfill (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 $M^*(1000) = 3.1$ and the median of $M$ is $M_{0.5}(1000) = 3.2$. $M$ will be smaller than $M_{0.01}(1000) = 2.65$ in only 1% of cases and larger than $M_{0.99}(1000) = 4.3$; 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 $N_{pp}(1000) = 6.2$. This, substituted into Eq. 2 and compared with Eq. 1, would imply $k = 2.5/6.2 = 0.4$. 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 $k$ for a chosen section of the noise can be made by the computer. The probable error of such an estimate is $\sqrt{\frac{v^2}{n}}$ 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.

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.

---

2. Freeman H., Introduction to Statistical Inference, Chapter 33, Addison-Wesley, Reading (Mass.) 1963.