Sampling Issues in the Design of Experiments for the Undergraduate Engineering Laboratory

B. Terry Beck, David A. Pacey
Mechanical and Nuclear Engineering Department
Kansas State University
Manhattan, Kansas

Abstract
An extremely important aspect of the proper design of an experiment is specification of the sample size, sample rate, and duration of test. When sampling real signal data from the wide variety of transducers currently available, the presence of noise, generated from many sources, usually makes it necessary to sample the associated input signal numerous times in order to determine accurate statistical information; typically mean and standard deviation. From these statistics, and the associated sample size, it is possible to arrive at a reasonable estimate of the confidence interval for the sampled signal mean. Furthermore, in some applications, for example in acquiring statistical information about turbulence intensity, the focus may also be on the standard deviation itself.

This paper provides a simple intuitive, but quantitative, procedure that students can follow in the undergraduate engineering laboratory to ascertain reasonable estimates of the required sampling information. The procedure has the advantage of not requiring complex background in sampling theory. In particular, it does not require detailed theoretical understanding of auto-correlation or cross-correlation statistical concepts that the typical undergraduate student has not yet acquired, which makes it readily applicable to the introductory undergraduate engineering laboratory. The technique is illustrated using simulations of time-series data generated by LabVIEW, as well as from real sampled signals. The paper also addresses the confidence interval for the standard deviation, which is frequently not given specific attention in the undergraduate laboratory experience.

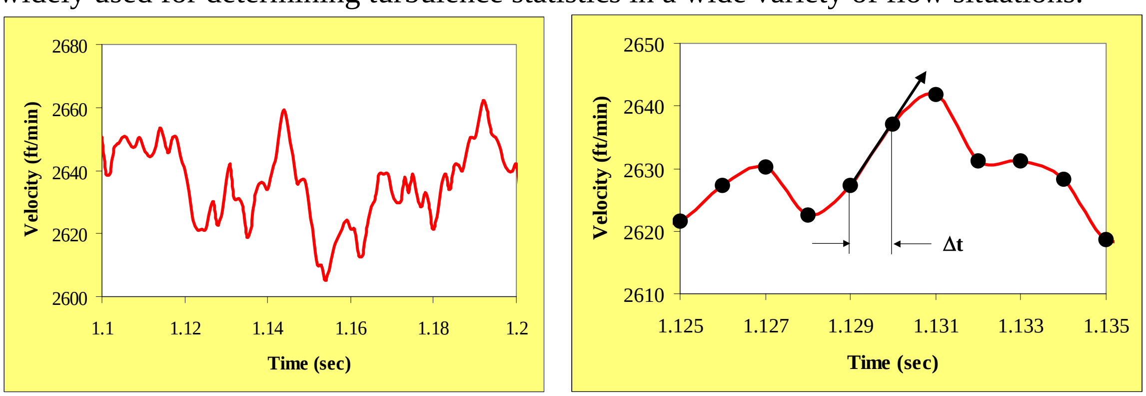
Introduction
One of the most important instruction issues encountered in the introductory engineering laboratory is that of introducing students to modern instrumentation and measurement techniques. The wide variety of electronic sensors currently used in industry, and the large number of industrial research, product development and testing applications that engineering students may encounter as they start their careers, requires that the students have a good working knowledge of such techniques. For the most part these applications will involve some form of automated (computerized) data collection and data reduction. Hence, it is a necessity that students gain laboratory experience with a wide variety of computerized data acquisition principles, which include Analog to Digital (A/D) conversion, and statistical sampling procedures.

The need to employ statistical sampling is not new, but is even more important in modern measurement applications that may involve real-time process assessment and evaluation of numerous time-varying signals of various types for quality control purposes.

Computerized data acquisition principles have long been an important component of the introductory Measurements and Instrumentation (ME 535) Laboratory class in the Department of Mechanical and Nuclear Engineering at Kansas State University. One of the difficulties encountered when dealing with modern sampling procedures in particular, is that our students have limited background in statistics and the practical application of statistical principles. Even with a background in prerequisite statistics course work, without supplementary instruction, the students still have rather limited practical understanding of how to apply these basic principles to laboratory measurements that involve real time-varying signals. Courses involving the detailed statistical treatment of time-dependent random signals are not part of the MNE curriculum since they generally have prerequisite requirements beyond the reach of our typical undergraduate students. In addition, while available course textbooks (e.g., [1], [2]) usually provide a good discussion of the statistical treatment of random errors, they do not generally address the practical issue of how to actually perform independent sampling of time-series data.

The Sampling Problem

It is typically assumed that the samples of measured variables used in statistical analysis (for random uncertainty estimates) are all independent; however, this may not be the case when dealing with real time-series data. When dealing with real time series data, there is a definite need to address the possible dependence of sampled data, in order to arrive at good estimates of the statistical characteristics of the signals. Consider the real time-series signal, X(t), sampled at a frequency of 1 kHz from a real velocity signal obtained using a hot-film high frequency probe as shown in Figure 1. Such probes continue to be widely used for determining turbulence statistics in a wide variety of flow situations.



(a) Actual Velocity Signal **(b) Close-up of Sampling**

Figure 1: Time Series of Actual Velocity Signal

The basic measurement problem is to determine an appropriate sampling time interval, Δt (or sampling frequency, $f_S = 1/\Delta t$), along with the total duration of the test, T. The total number of samples is then $N = T/\Delta t$ if N is large. Furthermore, if the samples are all independent, the mean may be estimated from $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X(t)$, and standard deviation may be estimated from $S_X = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X(t)-\bar{X})^2}$. However, when considering real signals sampled at high frequency, the signal value at some time t is highly related (or correlated) to its value at a small time $t + \Delta t$ later, as suggested from Figure 1 (b). In fact, in an extreme case where the sampling frequency is sufficiently large, effectively all of the samples together could actually constitute but a single instantaneous measurement. If the samples are somewhat correlated, the mean may be estimated from the same relationships given above for $\bar{X}$ and S_X, except that the actual sample size is replaced by the "effective" sample size, $N_{eff} = T/T_u$ where T_u is a measure of the time between uncorrelated samples. There are different ways available to estimate T_u ; in particular, if measurements of N, the actual sample size are compared to N_{eff} , from the relationships above $T_u/\Delta t = N/N_{eff}$. The main focus of this paper is to develop a tool which will demonstrate how to "experimentally" determine T_u , and also verify the behavior of this procedure with sampled signal data obtained from a LabVIEW simulation. The goal is to provide the students with an intuitive method of determining the appropriate sampling interval T_u (and the associated sampling frequency, $f_S = 1/T_u$) that can be readily used in an introductory instrumentation and measurements laboratory. The same approach can also be applied to real signals, to the extent that meaningful values of the required sample statistics can be obtained experimentally in the undergraduate laboratory.

Random Signal Generation

The first step in the development of the tool described above is to generate simulated random signals with known amplitude and frequency characteristics to represent "real" signals. LabVIEW software is particularly suited to this task. It is an "icon" based commercially available software package that is widely used for data acquisition both in industry and in research. It contains numerous VI (or virtual instrument) programs which are assembled in a graphical interface to perform all required operations associated with signal measurement and processing of results. In addition, LabVIEW also contains numerous built-in VI's for generation of random numbers with known statistical characteristics, and it also contains a variety of different random signal generation VI's with known amplitude and frequency content. In particular, it contains a source of Gaussian distributed "white noise" waveform. According to the documentation contained within LabVIEW Version 7, which is currently being used in our Department, the noise generation is based on a triple-seeded Very-Long-Cycle pseudorandom random number generation algorithm that produces approximately 2^{90} random numbers before the pattern repeats. The result is a pseudorandom Gaussian noise waveform pattern containing some 2,147,483,647 (2^{31} - 1) elements with a "white noise" uniform frequency spectrum[1], which provides an extremely large population for sampling purposes. Furthermore, the students that take our introductory Instruments and Measurements Laboratory utilize this software extensively throughout the course for data acquisition, after they receive basic instruction on its use.

Figure 2 shows the basic block diagram of the subVI used for generation of a modified signal called "band-limited white noise," which has uniform frequency spectrum over a limited (and known) frequency bandwidth. The band-limited white noise is generated by passing the white noise through a simple 1st order Butterworth low-pass filtering process.

[1] White noise is an idealistic noise time-series signal that has uniform frequency content (so-called spectral intensity) over a very large frequency bandwidth.

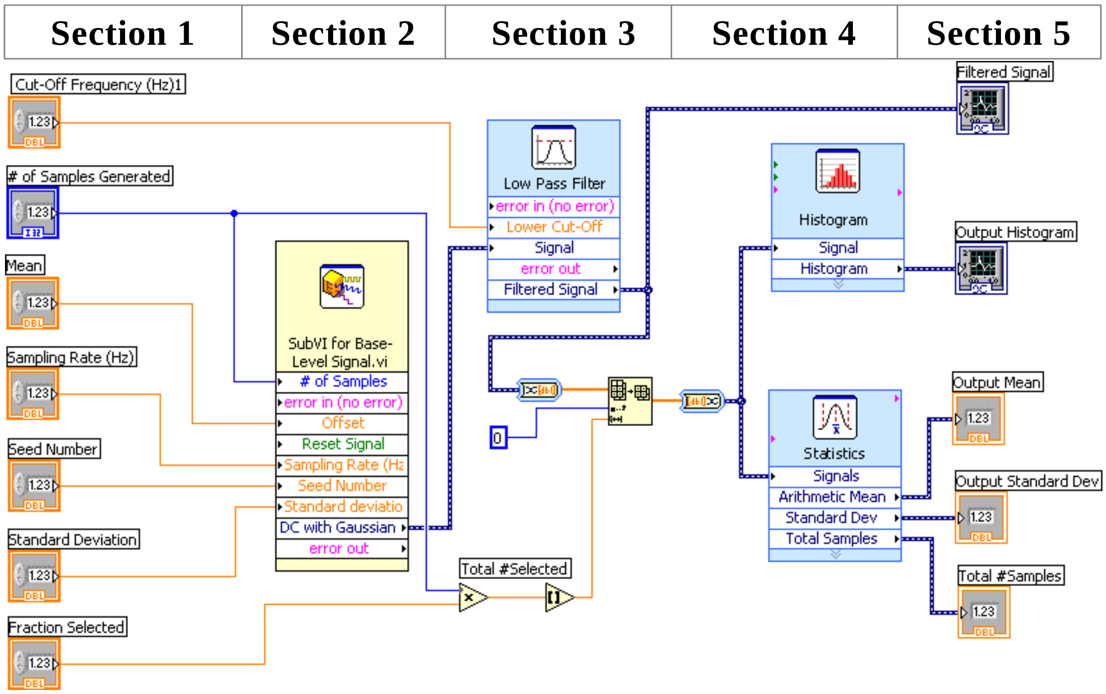


Figure 2: LabVIEW SubVI for Generating Band-Limited Gaussian White Noise

There are five main sections in the block diagram shown in Figure 2, arranged in a left-to-right calculation sequence. Section 1 of the input block specifies the main statistical characteristics of the generated noise, which includes the mean and standard deviation of the pure Gaussian white noise, the cut-off frequency (or bandwidth in Hz) of the filtered or band-limited Gaussian noise, the number of samples generated, and the specified sample rate in Hz. A "seed" number is used to initiate pseudorandom number generation process, and is used to either initiate the same random time-series signal, or a random sequence of signals. Section 2 is the LabVIEW SubVI that generates the pure Gaussian white noise signal. Section 3 passes this noise through a 1st order Butterworth Low-Pass filter, resulting in the desired band-limited Gaussian white noise. Section 4 performs statistical analysis on the time-series data, which includes generation of a histogram as well as determining mean and standard deviation of the sampled signal. Lastly, section 5 outputs the results in both numerical and graphical format. This band-limited white noise signal was used for all of the simulations presented in this paper.

The upper level VI, shown in Figure 3, uses the SubVI described above to generate multiple sets of independent sampled signals. A large number of independent sampled signals can be generated, which allows for determination of not only statistics of the individual time-series signals, but also statistics of the means of the signals of particular interest to the present paper. This SubVI is arranged in four sections in a left-to-right sequence as shown. Section 1 provides the necessary inputs that specify the number of sampling experiments (i.e., the number of independent time-series segments) to be performed, and the statistics of the noise that feed into the lower level SubVI described earlier. Section 2 is a "FOR-Loop" calculation sequence in which the time-series generation and statistical characteristics of each independent time-series experiment is repeated the specified number of times. Section 3 then performs statistics on the statistics of the independent "experiments," and Section 4 outputs these characteristics in both graphical and numerical format. The output statistics include the mean and standard deviation of the means, and a histogram of the means.

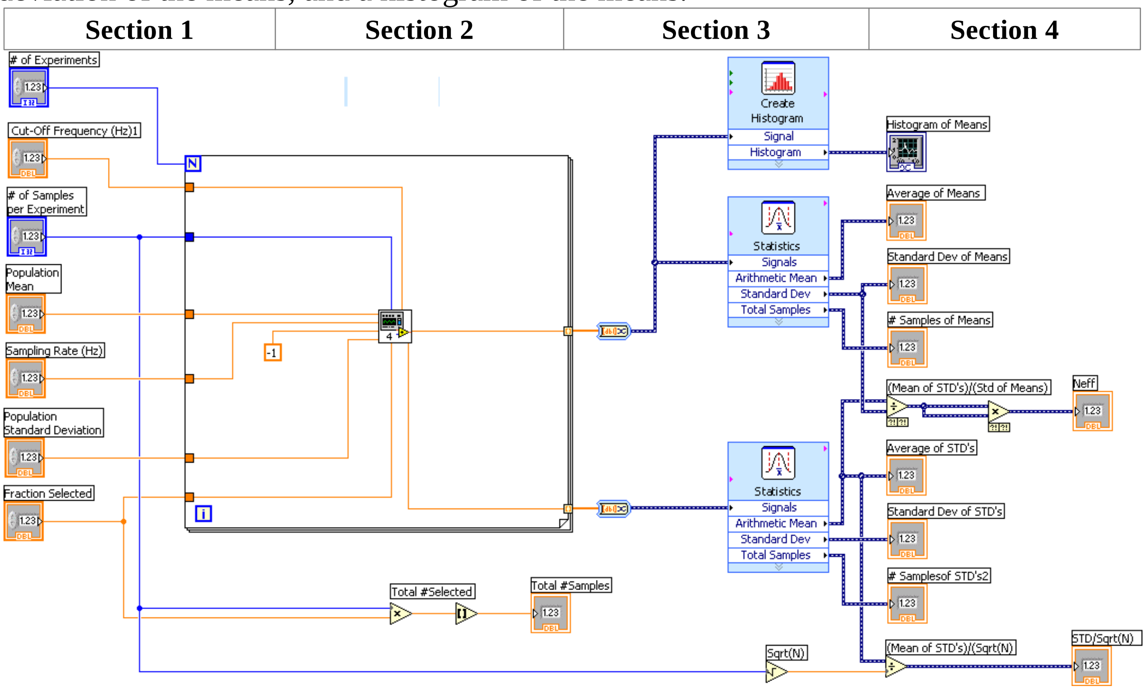


Figure 3: Block Diagram for Generation of Multiple Signal Sample Statistics

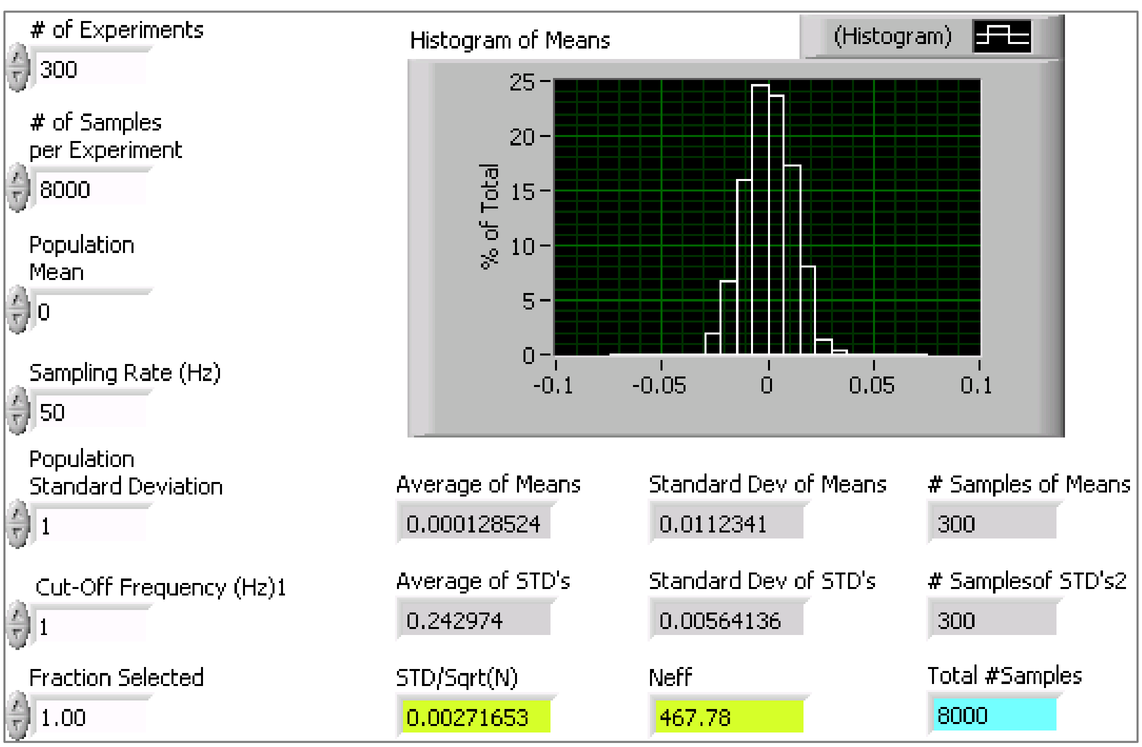
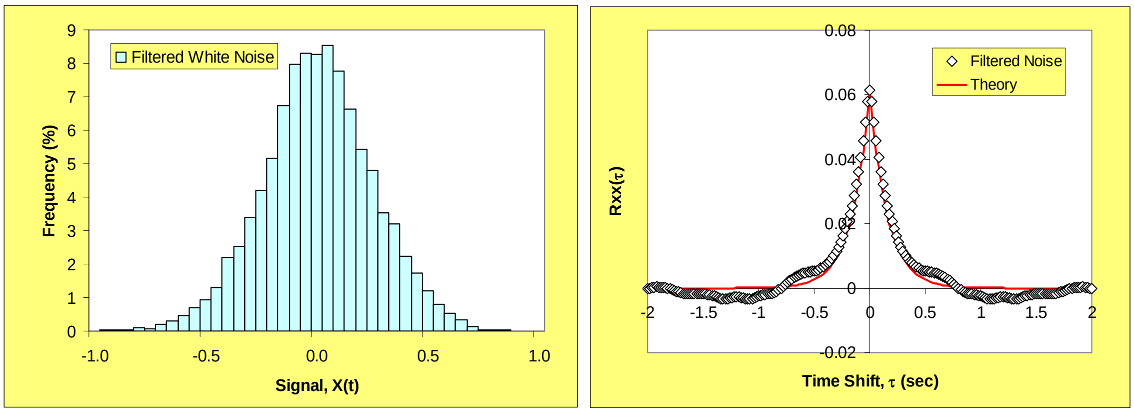


Figure 4: Front Panel of Multiple Signal Sampling VI

Hundreds of independent samples of the mean and standard deviation of the individual time-series signals can easily be acquired, sufficient to yield well-defined histograms of the signal means and standard deviations.

Verification of Generated Time Series Statistics

Several tests were conducted to verify that the time-series signals had the desired amplitude and frequency characteristics. For the tests presented in this paper, the inputted pure Gaussian white noise had a specified mean of zero and a standard deviation of 1. The bandwidth of the filtered noise was chosen to be 1 Hz, and the sample rate was 50 Hz. This provided an "over sampled" signal to give emphasis to the difference between the actual number of samples and the "effective" sample size.



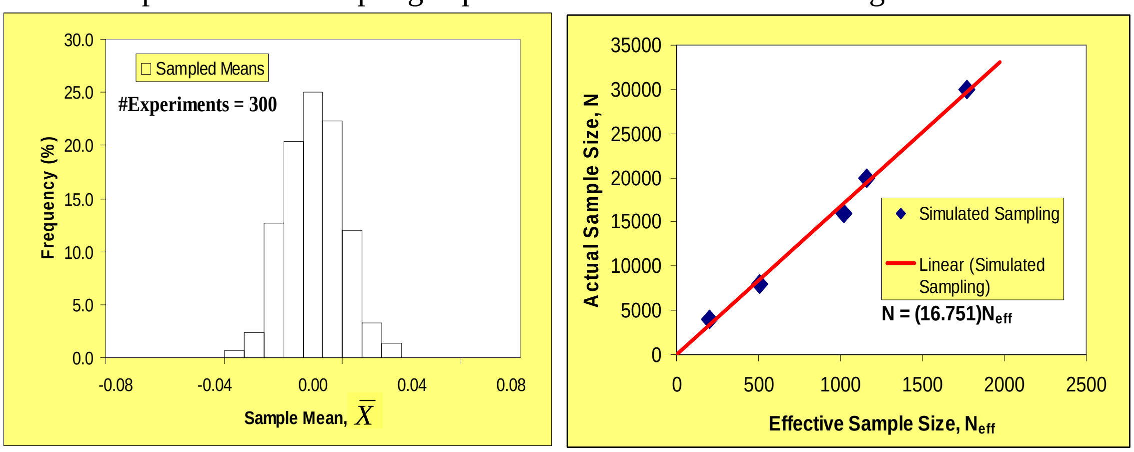
(a) Histogram of Amplitude **(b) Autocorrelation**

Figure 5: Band-Limited Gaussian White Noise

Figure 5 shows some basic statistical characteristics of the generated band-limited Gaussian white noise signal. In figure 5(a), a typical histogram of the noise amplitude distribution is given, and it is seen to have the expected Gaussian form with a standard deviation of approximately S_X = 0.2480 and a mean very close to zero. Figure 5(b) shows a plot of the autocorrelation, $R_{XX}(\tau)$, of the noise signal as a function of the time shift parameter, τ. The autocorrelation provides a measure of the relatedness of adjacent signal values in the sampled signal. If the random signal was completely uncorrelated with adjacent signal values of itself, $R_{XX}(\tau)$ would drop immediately to 0 after time shift $\tau = 0$. Real random signals drop off more gradually. For a stationary random process representing band-limited Gaussian white noise, the autocorrelation function is of exponential form and is given by [3] $R_{XX}(\tau) = \sigma_X^2 \exp(-2\pi f_C |\tau|)$, for $\tau < \infty$, where $R_{XX}(0)$ corresponds to the variance of X. This theoretical result is seen to compare very well with the sampling results shown in Figure 5(b), where it is clear that the signal is for all intents and purposes uncorrelated with itself after a time shift of just over 0.5 seconds.

Examples of Time Series Sampling

Now that individual samplings of the time series have been shown to behave with expected statistical characteristics, the multiple sampling characteristics can be investigated using the main LabVIEW VI shown in Figures 4 and 5 above. The purpose here is to investigate the effective sample size [4], and to illustrate a simple method of estimating the effective number of samples for a given sampling scheme. The relationship between actual sample size, N, and effective sample size, N_{eff}, can be estimated from the mean statistics of multiple samples. First, if all samplings of a random signal of duration T at the sample frequency, f_S, were independent, then the standard deviation of the mean for multiple samples would be $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{N}}$. However, if the samples are not all independent then the true standard deviation of the mean can be related to the effective number of samples as $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{N_{eff}}}$. Hence, N can be related directly to N_{eff} according to the ratio $N_{eff} = \left(\frac{\sigma_X}{\sigma_{\bar{X}}}\right)^2$. Thus, the actual and effective sample sizes can be related from simple estimates of σ_X and $\sigma_{\bar{X}}$. Figure 6(a) shows a histogram for mean signal, $\bar{X}$, resulting from 300 simulated independent sampling experiments of duration $T = N/f_S = 160\ sec$, where $N = 8000$ samples and $f_S = 50\ Hz$. This corresponds to the sampling experiment illustrated earlier in Figure 4.



(a) Histogram of Means **(b) Effective Sample Size Measurement**

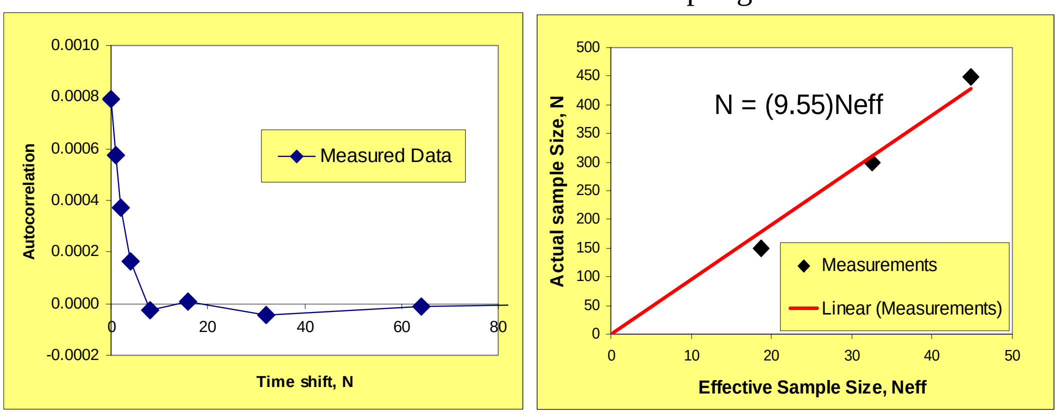
Figure 6: Multiple Sampling Experimental Results

From the resulting 300 sample statistics, the effective sample size can be estimated as

$N_{eff} = \left(\frac{\sigma_X}{\sigma_{\bar{X}}}\right)^2 = \left(\frac{0.242974}{0.0112341}\right)^2 = 467.8$, which corresponds to the value indicated on the front panel diagram shown in Figure 4. An alternative way to calculate an estimate of N_{eff} would be to conduct several similar sets of 300 experiments involving the same sample rate, but with different total number of samples. Figure 6(b) shows the result of several samplings made in this manner. The slope of the approximately linear relationship illustrated in this Figure indicates that approximately 16.8 actual samples are necessary to produce a single effective sample.

The results of the simulated sampling experiments above can be verified by theoretically calculating the relationship between actual and effective sample size using the autocorrelation coefficient given in Figure 5(b). The effective time between uncorrelated (effectively independent) samples can be shown to be approximately twice the so-called integral time scale [4], T_X , where T_X is defined such that the area under the autocorrelation function curve equals $T_X \sigma_X^2$. In other words, $T_u \doteq 2T_X$, where integration of the exponential relationship for $R_{XX}(\tau)$ yields $T_X = \frac{1}{2\pi f_C} = \frac{1}{2\pi(1Hz)} = 0.1592\,\text{sec}$ for the current band limited noise signal. Thus, the theoretical effective sample size for the current simulation results is $N_{eff} = \frac{T}{T_u} = \frac{160\,\text{sec}}{2T_x} = \frac{160\,\text{sec}}{2(0.1592\,\text{sec})} = 502.5$, which is very close to the sampling result of 467.8 given above. Put another way, this results in a theoretical relationship for the slope of the line in Figure 6(b) of $\frac{N}{N_{eff}} = \frac{8000}{502.5} = 15.92$, which agrees well with the slope estimated from the sampled data.

Once the effective sample size has been estimated, it can be used to estimate other statistical characteristics, such as those associated with the standard deviation. In fluids, the standard deviation is related to turbulent intensity, which is an important measure of the amplitude of random fluctuations associated with commonly occurring turbulent flows. The standard deviation of the standard deviation represents a measure of the uncertainty in this physical quantity. An estimate of the standard deviation of the standard deviation (which is directly related to the uncertainty in the standard deviation) may be expressed from [5,6,7] as $\sigma_{\sigma_X} = \frac{\sigma_X}{\sqrt{2N_{eff}}}$, where N_{eff} can be estimated from the slope of the line in Figure 6. For the data shown in Figure 4, this relationship yields an estimate of $\sigma_{\sigma_X} = \frac{\sigma_X}{\sqrt{2N_{eff}}} = \frac{0.24297}{\sqrt{2\left(\frac{8000}{16.75}\right)}} = 0.00786$, which is reasonably close to the simulated value of 0.00564 determined from the 300 samplings of the standard deviation.



(a) Measured Autocorrelation **(b) Effective Sample Size Measurement**

Figure 7: Estimate of Effective sample Size from Real Measurements

This process of determining the effective sample size, described in the LabVIEW simulations above, can also be achieved with real data. Furthermore, the linear fit approach allows some smoothing of the results when dealing with a more limited number of experiments like would typically occur in practice using real, rather than simulated, experimental sampling results. Figures 7(a) and 7(b) show the results of utilizing this technique for sampling hot-wire velocity measurements of airflow. In this case the sample rate was only 1Hz and the test duration was 2.5 minutes. The sample statistics in this case were obtained from only five independent test runs, corresponding to five independent experiments as defined in the earlier simulations. Hence, it would be expected that the variation in the results would have a fairly large uncertainty. None-the-less, the basic principle still seems to give a reasonable, if somewhat nominally accurate, estimate of the effective sample size.

Possible Classroom Implementation

One laboratory exercise where a problem related to non-independence of sampled data has occurred involves the use of a vane anemometer for the measurement of airflow. The vanes of the flowmeter make and break the beam of an infrared emitter-detector pair and the resulting pulse train becomes the input to a frequency-to-voltage converter. The voltage signal generally has a low-frequency noise component that students attempt to eliminate through averaging. The students are required to include an uncertainty analysis of their measurements and will develop a confidence interval for the true mean as a range about the measured mean. The range about the measured mean is the student-t value at the 95% confidence level times the standard deviation of the means, which is obtained by dividing the standard deviation of the sample by the square root of the number of measured values. Sometimes curious students will take multiple measurements at a given setting of flow to obtain an experimental set of sample means in an attempt to compare their variation with that predicted from a single set of data. These results are usually in very poor agreement because of the lack independence of the individual measured values in the sample resulting from sampling at too high of a rate. The procedure outlined in this paper will make it possible for the students to determine an appropriate sampling rate (and the associated sample size or test duration) needed to assure independence of individual measured values or to apply the concept of effective sample size so that predicted variations in sample means are correct.

Summary and Conclusions

This paper has presented a tool, based on the random waveform simulation capability of LabVIEW, for use in investigating time-series data sampling issues in the undergraduate engineering laboratory. The procedure involved utilization of Gaussian white noise inputted to a simple low-pass filter, resulting in well-defined band-limited Gaussian white noise. A simple VI block diagram for generating the time-series signals of known amplitude and frequency characteristics has been implemented and tested, along with a companion VI for generating large numbers of independent simulated sampling experiments. From these multiple sampling tests, a simple procedure was demonstrated for estimating the effective sample size associated with realistic simulated signals that, like real time-series data, exhibit correlation or relatedness between adjacent sampled data. It was also shown that realistic estimates of the sample statistics could be obtained from the effective sample size. The sample size estimation procedure was also demonstrated using real velocity measurements in an airflow situation. This approach represents a potentially very useful tool for enabling students to become intuitively familiar from an experimental point of view with a number of important laboratory sampling issues, without the need for prerequisite courses which are difficult and impractical for them to acquire in the usual undergraduate mechanical and nuclear engineering curriculum. Thus far the results appear to be very promising, and further testing of this approach, along with possible ways of implementing such an approach in the engineering laboratory classroom, are currently under investigation.

Bibliography

1. Beckwith, T. G., Marangoni, R. D. and J. H. Lienhard V, "Mechanical Measurements," 5th Edition, Addison-Wesley, 1993.
2. Figliola, R. S. and D. E. Beasley, "Theory and Design for Mechanical Measurements," John Wiley & Sons, 1991.
3. Bendat, J. S. and A. G. Piersol, "Engineering Applications of Correlation and Spectral Analysis," 2nd Edition, John Wiley & Sons, 1993.
4. Adrian, R. J. and Menon, R. K., "Data Acquisition, Processing and Analysis in Flow Measurements," TSI LDV Dada Analysis Course Text, 1989.
5. Lindgren, B. W., "Statistical Theory," 3rd Edition, MacMillan Publishing Company, 1976.
6. Shanmugan, K. S. and A. M. Breipohl, "Random Signals: Detection, Estimation and data Analysis," John Wiley & Sons, 1988.
7. Bendat, J. S. and A. G. Piersol, "Random Data: Analysis and Measurement Procedures," 2nd Edition, John Wiley & Sons, 1986.

Biography

B. TERRY BECK

Terry Beck is a Professor of Mechanical and Nuclear Engineering at Kansas State University and teaches courses in the fluid and thermal sciences. He conducts research in the development and application of optical measurement techniques, including laser velocimetry and laser-based diagnostic testing for industrial applications. Dr. Beck received his B.S. (1971), M.S. (1974), and Ph.D. (1978) degrees in mechanical engineering from Oakland University.

DAVID A. PACEY

David A. Pacey is a Professor of Mechanical and Nuclear Engineering at Kansas State University and teaches courses in measurement & instrumentation, machine design, and senior design projects. He is also Director of the Undergraduate Program. Dr. Pacey received his B.S. (1974) in agricultural engineering and his M.S. (1979) and Ph.D. (1989) degrees in mechanical engineering from Kansas State University.