INTRODUCTION TO ROBUST STATISTICS AND DATA FILTERING

Gernot M. R. Winkler
U.S. Naval Observatory
3450 Massachusetts Avenue NW
Washington DC 20392-5420

Automatic and/or remote time and frequency measurements produce data that are sometimes contaminated by errors which can arise from single bit failures and can be arbitrarily large. The detection of such errors and the justification for their rejection will be discussed and simple methods given how to accomplish this automatically and objectively. Such methods, and an understanding of the principles involved, are indispensable for the correct evaluation of frequency standards and the operation of timed systems. The treatment offered can be considered as part of Robust Statistics but will emphasize the practical aspects. A related but substantially different application arises in the operation of electronic navigation receivers, where the prompt detection of system integrity failures cannot rely on postprocessing of the observations. This problem will also be discussed briefly in an introductory way.

CONTENTS

INTRODUCTION

Many applications of precise time and time interval (PTTI), especially the use of time measurements in navigation or positioning, or the use of remotely collected time or frequency measurements, often confront us with data which are questionable. This happens, of course, in the practical use of any data, but in our field a decision of what to do is usually more critical and pressing. An overview of robust statistics with practical examples will be presented in these notes with the aim of alerting the practitioner to the benefits of using even the simplest analytical tools of this specialty.

The assumption of a Gaussian (normal) distribution of errors is the standard starting point of almost all critical data processing done in practice. However, with the advent of electronic data transmission and automatic data collection, this assumption cannot be made tacitly. Arbitrarily large errors can and do occur and the analyst is ill advised to use blindly the statistical tools which he may have learned at the beginning of his career. Even more important are real-time applications where people depend on measurements immediately. We must alert users to the need to build judgment into their routines that will prevent the blind automatic acceptance of dangerous outliers or system failures.

As an example, we can take the Global Positioning System (GPS) in its present state. We have observed from time to time spurious timing errors of very large magnitude, up to hundreds of microseconds. With the introduction of the Block II satellites, this has become very rare. Nevertheless, we must have some way to recognize and to reject such erroneous measurements. Fortunately, there

are techniques, some of them very simple, yet not widely known, which can be used with contaminated data. The more sophisticated techniques are not so clear to the occasional user and for this reason have been neglected by most practitioners in the past. However, as Tukey has said, "It is not so important what you do as that you do something about this problem."

What we can do easily is to familiarize ourselves with at least a simple but safe form of robust data analysis, and discuss the principles involved. I hope that these examples will give encouragement for going deeper into critical data analysis.

Box (of Box and Jenkins fame, [1]) introduced the concept of robustness as a name for a class of statistics that are insensitive to large errors ("outliers"). The idea, however, had been considered much earlier, by Gauss and Laplace, [2]. In principle, we can approach the outlier problem on the basis of the internal consistency of the dataset. If we have a set of measurements, then we want to find an estimate for the "real" value that is to be measured, and we also need a measure for the magnitude of the scatter as an estimate for the uncertainty of our result. The (sample) mean M and the (sample) standard deviation s are the accepted estimates for the center and the scatter of the distribution at hand. They are estimates because they are based only on the sample at hand, which does not exhaust the total of all possible measurements that is meant when we talk about the "population". However, if our data have been contaminated with extraneous large errors, then M and s will be affected and can be substantially different from the uncontaminated measures which are needed as a characterization of the process.

But why have M and s been accepted as standard measures in the first place? The reason is that, for a normal distribution of errors, it can be shown that the sample mean M is the most efficient estimate for the"center" or "true" value and s, the sample standard deviation, is the most efficient estimate for the variance or scatter (sigma). But this is true only for a normal distribution. In the real world, we find that often a normal process with its characteristic measures of M and s gets contaminated by blunders, extraneous noise, transmission errors, or other interference with orderly data collection. In fact, the frequency with which such "outliers" appear in measurements is so high that we have no justification for our standard methods of statistical analysis (based on the assumption of normal distributions) unless we find and reject these extraneous data at the beginning. With this model in mind, we have to find systematic, objective ways to identify and then filter the outliers.

First, we will deal with sets of measurements of a fixed quantity and later with quantities that vary with time (time series). In the first case we deal with the classical problem of multiple observations of a single quantity. In the second case we observe a process, but a process that is only partially deterministic and, in addition, we must expect also a few "blunders" (additive outliers). It is important to keep these two cases conceptually separate, even though we can note that the second case can be reduced to the first if we consider the distribution of the residuals from a mathematical model which has been fitted to the data by some criterion such as least squares. The mathematical model represents the deterministic part and the residuals the stochastic. Of course, we have the additional problem of finding a realistic model, i.e., a model that produces residuals without systematic trends. The application of robust statistics for the determination of M and s will reappear as filtering of the data for the purpose of getting to the uncontaminated model for the description of the time variability.

One can distinguish robust measures that include the set as a whole from robust methods that attempt to find the uncontaminated sample which can then be treated classically. We could distinguish these two cases also as inclusive vs. exclusive methods. This presentation will concentrate on the second approach, better described as an iterative method, as opposed to the first, which amounts to a block procedure. For a discussion of block procedures I have to refer you to the literature, especially [3], [4], [5], and [6]. Among these methods, one distinguishes M-estimates derived from maximum likelihood arguments; L-estimates, which are a linear combination of order statistics (our rejection criterion to be discussed later would belong here); R-estimates based on rank tests such as the Spearman rank-order correlation; and finally techniques derived from control and filtering theory, such as the Kalman filter, which can be made somewhat robust if the necessary precautions are taken. To summarize, we have to accomplish two tasks: We have to find robust measures to characterize the uncontaminated distribution of our dataset; and we have to reject those points which do not belong to this distribution. The first part is robust statistics; the second part is of concern to the practitioner.

Back to Contents

IS A REJECTION OF OUTLIERS JUSTIFIED?

What is an outlier? Outliers appear to be inconsistent with the remainder of the data. They are "surprisingly" far off the consensus.

But this remains a subjective judgment until we analyze the data population as a whole. Then we may be able to say whether we have a contaminant (a genuine outlier) or an extreme value. Extreme but legitimate values may be important; they carry some information which should not be ignored. The simplest, but at this stage, inadvisable method would be to reject a priori a certain small number, say 5%, of all points. Of course, we want to reject points which are farthest from the mean. But this cannot be done without judgment unless we want to accept the risk of throwing away perfectly good data. In addition, cutting the tail of a distribution will introduce a noticeable bias in M and s, particularly if the tails are cut in reference to the contaminated mean. Therefore, this process is not objective and it cannot in this form be automated. What we need is an adaptive filter which is capable of making simple judgments about outliers. In other words, the problem is much more involved than it may appear to be at first sight. It has been a subject of intense interest to astronomers and geodesists ever since the days of Gauss. An extensive modern discussion of outliers can be found in [7].

The rejection of dubious points is a good example for the general decision problem in statistics. In making decisions on the basis of uncertain data, we must take into consideration two types of errors: type I and type II. Committing an error of type I in our case consists in wrongly rejecting a genuine member of the distribution; we make an error of type II if we retain a real outlier. The preference for one or the other of these errors (or better, willingness to accept them)reflects a basic strategy of decision. These preferences can be precisely quantified. A "conservative" decision-maker prevents the loss of a dubious value by preferring a small risk I. The "radical", or "liberal", in reverse, prefers a small risk II and is willing to risk the loss of valid data because he dreads contamination. (One could say that he is a conservative, too; but, perhaps simplistically, he aims at conserving truth and not data points).

Percival [6] defines a filtering algorithm as "robust if deviant behavior by a small percentage of the frequency standards in a time scale does not unduly influence the resulting time scale." This definition is characteristic for the approach using block procedures because it allows some, albeit "not undue," degradation, in contrast to our stated aims that seek to eliminate such influences altogether. The justification for our aims is the, possibly naive, principle that no contamination whatever should be allowed. The conservative, and we confuse the issue a bit by a priori identifying the block procedure adherents with tolerance for a high type II risk, will reply that we can only know with some finite confidence that a value may be an outlier. Therefore, we should keep it in the set but with a correspondingly lower weight. This is a perfectly sound principle and most of Robust Statistics is based on it. Nevertheless, we must realize that very often our policy will be influenced by the circumstances rather than by a priori principles. A paucity of data will most likely induce a conservative attitude, in contrast to the case where an abundance of data allows a more generous, liberal, and in my opinion also safer, rejection method.

We adopt the following definitions: The risk I is the probability α that a genuine value is rejected. Conversely, the risk II is the probability ß that a real outlier is retained. Of course, we can only evaluate these probabilities to the extent that we know the distributions for the genuine values and for the real outliers. Usually, we will know very little about the outliers (unless we have some a priori ideas about their statistics) and, therefore, ß is almost irrelevant for us. This is in contrast to the general case in statistics where we decide between two alternate hypotheses. All we can say here is that we must expect almost any value as an outlier. For this reason we can only make ß small by allowing α to grow. α , however, is known to the degree of our ability to estimate the uncontaminated distribution. Alpha can also be called the probability of false alarm, while ß is the probability of a missed detection. Sometimes, the probability of correctly identifying the outlier is called Γ and is referred to as the power of the test.

$$\Gamma = 1 - ß$$

Standard choices for α can be recommended (following [8] p. 119): If a type I error is serious, adopt α =0.1% or at most 1%. If, on the other hand, a type II error is serious, then a larger α , possibly 5%, should be adopted. The following example will make this clear: A 4.5% value for α would imply a rejection limit of 2 σ of the undisturbed normal distribution (see scale E of Fig. 1). It seems clear that the first case, when a type I error is serious, will be better handled by the adoption of a conservative method ([3], [4],[5], [6]). As an additional aid in deciding, one should also look at Fig. 2.1 of [7], which gives more detail on the different kinds of outliers that could be expected, such as additive vs. innovative outliers. In general, I would recommend that the less we know about the outliers, the more we should be inclined to the simple rejection methods advocated here. That would, indeed, lead to the adoption of an α =5% as a rule of thumb, to be on the safe side. Many experienced practitioners of statistics do this in a wide variety of applications ([8]). This is, however, not what we rejected at the beginning of this section as not being objective. There we argued against a blind, a priori rejection. But now we have decided that we must estimate the undisturbed distribution, and then, we also realize the importance of diagnostics in our procedures to guard us against cases which do not agree with our assumptions about the

basic distribution. In addition, as we shall see, the data rejection that we will propose will not necessarily eliminate all values beyond the adopted limit. We only reject, starting with the largest deviation (because that is the most likely outlier) until we reach the number that we can expect on the basis of our estimate of the main distribution.

To summarize the present attitude regarding outliers, we can say that today, as practitioners in contrast to the theoretical statisticians, we are more in favor of a rejection of suspicious data as opposed to the earlier attitude of keeping them at a lower weight: we collect more data and we can be more generous about them. A main reason for that is the experience that today's automatic methods not only produce vast amounts of data, but they also can produce arbitrarily large errors that clearly come from processes that are different from the one under investigation. However, even before our modern data collection methods presented us with the problem of "cleaning up large but noisy data files," volumes have been written about the justification of rejecting any data, however suspicious. An example is the following table, which contains a series of measurements of the vertical semidiameter of the planet Venus made in 1846. The measurements have been discussed by several authorities who classified -1.40 as a discordant value [9]:

-1.40	-0.44	-0.30	-0.24	-0.22	-0.13	-0.05	+0.06
+0.10	+0.18	+0.20	+0.39	+0.48	+0.63	+1.01	

Table 1. Semidiameter of Venus

The opinion is still widespread that there is no theoretical basis for the rejection of any authentic, bona fide observations. This is considered controversial. However, the rejection of data is only controversial if we fail to define exactly what we intend to do. If we intend to summarize the particular observational process with which we are concerned, then indeed we must not reject any data. We must not reject anything because even the wildest "outlier" is part of the process and conveys important information. It is part of the distribution of errors that are a consequence of the details of the process chosen. If, on the other hand, the process is only one of all conceivable ways of obtaining information on what is behind the process, the "real" value if there is one, then of course, we must consider wild outliers as not belonging to the distribution of errors that we expect in our process of interest: in the estimation of that"real" value in which we are interested. Clearly, the second case is of wider importance and is usually tacitly assumed. In what follows, we assume this latter case and, therefore, we will discuss methods for an objective identification of outliers. The first case, however, is not unimportant. In the discussion of observational techniques, or of data collection methods, the occurrence of widely discordant results is a very important fact and must not be disregarded. In the Venus observation above, the theoreticians should reject the outlier, but the observers should not; they must use it in their critique of the observations. We will discuss practical cases of frequency stability analysis from this point of view.

We shall test our methods and see how they will deal with sets such as this Venus case in an unambiguous way. Our problem is, again, the design of procedures that will make rejection decisions automatically, and these decisions cannot, we repeat, be based on subjective grounds. To proceed, we must first use realistic levels of significance and, furthermore, we must adopt a criterion for randomness. The prototype for the concept of randomness is the coin-tossing experiment where we can most clearly see randomness in action. If we say that, with significance of 1%, it is beyond chance that a fair coin falls on the same side in 7 consecutive tossings, then this is how we arrive at the statement: each time the coin shows one side (out of two possible) that happens with probability 2. For this to happen seven times in a row, the probability is $(2)^7$ because the probabilities of independent events multiply. But this means that pure chance of this happening is only $(2)^7$, which is less than 0.01. For other significance levels, we can easily construct the following table:

n	2^{-n}	Level
1	0.50000	
2	0.25000	

3	0.12500	
4	0.62500	<10%
5	0.03125	< 5%
6	0.01562	
7	0.00781	< 1%
8	0.00391	
9	0.00195	
10	0.00098	< 0.1%
14	0.00006	< 0.01%

Table 2. Levels of Significance for Coin Tossings

A necessary assumption for all this is, of course, that the outcomes of our coin tossings are really random, i.e., that the results are independent of each other and are not in any way systematically influenced. Only then can we multiply the individual probabilities for one outcome (assumed to be 2 in our case). As data analysts, we will probably not be dealing with a distribution such as the above unless we work in Las Vegas or have to deal with reliability problems where the binomial distribution is fundamental. Otherwise we will have to deal predominantly with two different ideal situations:

Case one: Our data are normally distributed with the exception of a few extraneous outliers, which we identify on the basis of the probability that they cannot belong to our distribution at an accepted level of significance, say 5%. The problem in this case is that we have to find the dispersion of the uncontaminated distribution so that we can set the rejection criterion properly.

Case two: Our data, even without disturbance, are not normally distributed, but follow a principally different distribution, such as a log-normal distribution or (another important practical case) a Poisson distribution.

In this paper, we will not consider case two further, because the principles of robust data analysis can be discussed more clearly on the basis of the first case. Appendix 1 does, however, give a brief discussion of the Poisson distribution.

Back to Contents

STATISTICAL DISTRIBUTIONS

The normal (Gaussian) distribution (Fig. 1) will be assumed from now on. It is so frequent because it is generated by many small disturbances which act additively but randomly. These disturbances prevent us from measuring always the same value. This is essentially also the basis for the "central limit theorem" which states that the normal distribution results from the action of many small independent errors of any individual distribution. An entirely different distribution of the data, the log-normal distribution, results if many disturbances act multiplicatively, i.e., if they influence their successors. This is usually the case in biological or sociological problems. The rich get richer because their money helps in earning more money. Income is nearly log-normally distributed (it is called a Pareto distribution if it has a sharp lower cut-off and a very long upper tail as we see it in the income case). Yet another case, the Poisson distribution, is used in problems that arise in the counting of rare and mutually independent events (isolated events in a continuum). Examples for approximate Poisson distributions would be misprints per page, disintegration of radium atoms per second, telephone calls arriving at an exchange, plane arrivals at a large airport, etc. (see Appendix 1). In most of these cases, however, it cannot be assumed apriori that these events are indeed independent of each other. Hence the need for testing the distribution. The binomial distribution, finally, is of foremost interest in reliability questions. Both of these, the Poisson and the binomial distributions, tend to the normal distribution as their limiting case with large numbers.

In position fixing, target shooting, and in diffraction phenomena, we deal with two-dimensional error distributions. Many small, random disturbances affect the "impact" point. If the errors in both dimensions have normal distributions with equal sigmas, and if

they are uncorrelated, then we have a Rayleigh distribution. The often used "c.e.p." (circular error probable, circle of equal probability, or circular equal probability) is the 50% circle of the radial errors.

At any rate, it is essential for the identification of outliers that we know which distribution we should assume as the main distribution of the data. By distribution we mean, of course, the idealized histogram of the data known as the probability density function (pdf). The pdf gives the probability of finding data at a given value. Since probability is in practice the number of considered cases over the total number in the population, we must express the pdf as the number of data over the small interval considered, divided by the total number, i.e., a density.

Parenthetically, we can note that the popular idea of statistics is superficial and distorted, which leads to many wrong conclusions. Only the unsophisticated will compress the information which is contained in a set of related values by forming the arithmetic mean and stop. What we obtain this way is only a sample mean. Without further work, it is entirely unjustified to assign to this value any meaning other than that it is the average of the data at hand. However, in practice, there is a tacit belief that we have somehow obtained the "real value". But, of course, who knows whether there is any single "real value"? In contrast to this naive position, a main principle of statistical work is to find the distribution of the data; only then can the "expected" value be found by (Lebesgue) integration of this distribution. In other words, the concept of a distribution is fundamental.

The expected value is defined through the distribution and the sample mean is just an estimate, and often a poor one, for the expected value E{z} of the random variable z. The expected value can only be found if we sum (or integrate) over all values z multiplied with their probability. Because the probability of each value is given by the pdf of the population and not by the histogram of the possibly contaminated sample, our methods require some assumption regarding the undisturbed distribution. Nevertheless, we must look at our sample distribution very carefully as an indispensable diagnostic. We will approach the problems by making the initial assumption of an underlying normal distribution of the errors, but this assumption must be checked and cannot be blindly maintained on faith. See [10] for more detail on testing for a normal distribution.

Incidentally, we have just come across a point that is of fundamental importance throughout science and philosophy. What we learn from experiments depends on the question we ask and on the assumptions we make when we analyze our data. These assumptions are, in turn, the result of some basic structures and functions in our brain, and all prior experience in the form of our preconceived notions, and cannot be presumed to be facts of nature which exist independently of us. Perceived reality is different to different people and to different times and, simply stated, our view of the world is to a large extent what we think it is! Seen from this angle, science is the struggle to identify and reduce these subjective parts as much as possible, but they cannot be eliminated completely. Statistics is exactly in the same situation. We attempt to be as objective as possible, but this cannot be done without any assumptions whatever. It is utterly necessary to be as clear and honest as possible about these initial assumptions. For us this means that without any assumptions, there would be no outliers!

For practical purposes, we will concentrate on the main principles and, in the following, we assume that we have indeed a normal distribution of the undisturbed data. Therefore, the outliers are assumed to come from a process which is different from the one that produces the "regular" (small, random) errors. We will also suggest a few simple diagnostic measures in addition to the use of the histogram program. Incidentally, if we were to find a log-normal distribution, then we could transform it into a normal distribution by changing the measurements (transforming the abscissa) to their logarithms.

The diagnostics that are included with our routines will allow some monitoring of the degree of normalcy in the data set. For that purpose, we must discuss the various measures of dispersion, estimates of the "center", and the rough shape of the distribution. To summarize, we must know the undisturbed distribution in order to be able to state the probability for the occurrence of an unusually large discordant value in the data. If that probability is, say, less than 0.001, then we can state with a significance of 0.1% (0.001) that the datum in question is not a legitimate member of the data set if we have less than 1000 measurements. And, in order to improve objectivity, we should better state or decide the critical significance level to be adopted in advance of our data inspection.

In summary, we say that from a set of measurements alone, we cannot know exactly which values, if any, are outliers. But it is also unknowable what the exact "true value" is. Only from the internal consistency of the data, from the distribution, can we estimate a "best" value with some stated uncertainty; and we can identify values as outliers with a stated probability. Precise statements are always uncertain, while certain statements are necessarily imprecise (as stated in [8]).

Back to Contents

THE NORMAL DISTRIBUTION

This distribution is the ideal bell-shaped histogram. It is symmetric about its mean (the population mean is usually denoted as μ) and has as its (population) variance the mean-square deviation. The square root of the variance is the standard deviation (sigma). Mean and standard deviation (or variance) fully characterize the distribution. The ideal histogram (the pdf) represents the probability density as a function of the value. One usually standardizes by drawing the abscissa in terms of sigma, with the Mean as the center. The total area under the curve is one. The integral of the pdf, i.e., the area up to the point in question, is the cumulative distribution function (cdf). The value of cdf(x) is the probability to find a value that is equal to or smaller than x. For large negative deviations, the cdf starts at zero and increases steadily up to one in an S shaped curve. The slope is, of course, the density of our data and we see that this is largest around the mean and is very flat at large values. In practice, the cdf is even more revealing than the pdf because in dealing with actual sample sets and not with the theoretical population, the histogram appears uneven or discontinuous in contrast to the cdf. Of course, this is because the cdf is the integral of the pdf.

The sigma is a measure for the scatter in the same units as the values themselves. The variance σ^2, on the other hand, is more useful in theory. If we have two disturbances acting together but independently, the variances can be added to give the variance of the result. This cannot be done with the sigmas.

The formula for the ideal histogram, the probability density of a normal distribution, is as follows:

$$\text{pdf}\,(x \mid \mu, \sigma) = (1/\sigma \cdot \sqrt{(2\pi)})\, e^{(-½)[(x-\mu)/\sigma]^2}$$

The inflection points of the curve are at $\pm 1\sigma$. The value of 0.674 σ is known as the probable error (p.e.) because 50% of all values are within ± 1 p.e. If we know or can estimate the p.e., e.g., by counting and ordering all errors and finding the half point, then we can estimate the sigma: $\sigma = (\text{p.e.}) \cdot 1.483$ which we will use later. We should remember the main probabilities listed in Table 3 for the normal distribution:

Within ± This Range	Are This % of All Values	Remarks
1 p.e.	50.000	
1 σ	68.270	
1.5σ	86.640	
2σ	95.450	4.55 > 2 σ
2.5σ	98.760	1.24 > 2.5 σ
2.75σ	99.400	
3σ	99.730	0.27 > 3 σ
4 σ	99.994	0.0063 > 4 σ

Table 3

A rejection of outliers beyond 4σ corresponds with a significance level of 0.01%. We allow only one error in ten thousand to exceed 4σ because only one could be expected larger than 4σ in our sample. One such value would then still have to be considered as a legitimate member of our data set. In other words, we are justified to reject this value at the previously stated level of significance only if the number of our data points is less than 10000. Therefore, if we adopt, say, a level of significance of 0.27%, then we would reject outliers, starting with the largest on either side and eliminating them one by one until we have 27 per 10000 left which exceed 3σ (or correspondingly fewer for a sample smaller than 10000). This, then, is an objective method with a specified level of significance which we can use in the rejection task, always remembering, however, that we assume a normal distribution.

However, the problem remains that we do not yet have the undisturbed statistics available for making the above decision. We have to search for robust estimates for the center of our distribution, μ , and of the scatter, σ , but the simple sample values can't be used because of their high efficiency: they are too much disturbed by outliers. We need estimates that are not sensitive to the presence of outliers. With robust estimates we could recognize these outliers, remove them from the data set by using the procedure just outlined, and only then use the standard statistical methods to evaluate the undisturbed set. But before we forget, we have assumed that the undisturbed distribution is normal! If tests should show that this assumption is not justified, then we cannot use this method unless we can transform the distribution to a normal one or modify the routines to account for the particular distribution on hand. But how normal must our distribution be so that we can apply our methods? This is why we need good diagnostic measures.

Let us discuss our standard estimates first. The most efficient estimate for the μ of the normal distribution is the sample mean (Mean or M):

$$M = \Sigma\, x/n$$

By most efficient, we mean that, for the given sample size, we obtain estimates with the smallest variance from sample to sample because this measure extracts a maximum of information from the sample. The most efficient measure of the sigma is the sample standard deviation σ . It is the root mean square deviation from the mean. This can also be written as:

$$\sigma = \{[\Sigma\, x^2 - (\Sigma\, x)^2)/n]/(n-1)\}^{1/2}$$

This unusual looking formula is simply a different form for the rms deviation from the sample mean with a denominator of (n-1) instead of n. The reason for (n-1) in the denominator is that using n would produce too small a value because the deviations are only available against the sample mean and not against the unknown μ, which is slightly different. The sample mean is the value which minimizes the sample deviations, which makes the s appear too small; this is compensated for by using (n-1). We express this by saying that using n in the denominator would bias the estimate of the population variance (and its standard deviation as well). On the other hand, the computation with n produces a "maximum likelihood estimator." As one can see, it is extremely important to keep a clear distinction between the ideal measures of the population (μ and σ), and their estimates computed from the sample (M and s). Our formula appears in a form which allows the computation of the two sums concurrently with the arriving data without the need for waiting until we have computed the mean. The benefit is that we do not need to keep the data in memory, but only the two sums. This formula is well known but, as has been pointed out by Dr. James Barnes of Austron, Inc., the use of this formula can lead to errors because we may be subtracting two large numbers from each other. A way around this is adding the squares of the differences x(i) - x(1). We leave it as an exercise for the reader to figure out how to correct this after M has been computed so that the equivalent to the above formula is obtained. Of course, it is even simpler to follow exactly the definition and compute the deviations after the mean has been obtained. By root-mean-squaring of the sum divided by n-1, we then obtain the standard deviation without these potential accuracy problems.

Other simple measures of diagnostic interest are the median (med), the mode, the median absolute deviation (meddev), the skewness, and the kurtosis. The great and so often ignored value of the median as an estimator of the center of a symmetric population lies in its insensitivity to the magnitude of the outliers. While the mean must reflect the presence of an outlier xout with an error of xout/n, the median is hardly affected; in fact, it is totally insensitive to the magnitude of that outlier because it extracts only position and no value from the data. The same thing is true for the median (absolute) deviation (meddev) vs. the standard deviation. We see that efficiency (use of all information) and sensitivity to outliers go hand in hand. You cannot have one without the other. Part of the reason for the relative neglect of the median as means to characterize the distribution is that, naturally, its variance is greater than the variance of the standard estimators mean and s. But nothing is free. If we need insensitivity, then we have to accept a slightly less efficient measure. A contributing factor for the neglect by the professional theoreticians may be that the medians are not "analytic" and for this reason do not lend themselves easily to mathematical analysis in closed form. Nevertheless, it is a widely recognized fact that the median is a valuable estimator, particularly for distributions with broad tails [11].

What follows belongs to what is known as order statistics. We order the data by magnitude. The central data point is then the Median:

med = x[(n+1)/2] for n odd, and

med = {[x(n/2) + x[(n/2)+1]}/2 for n even.

This assumes that the errors are listed in an array that starts with position one.

For symmetrical distributions (such as the normal distribution), the mean and the median (i.e. the population parameters, but not the sample measures, which will always show some variance) are identical. Another important concept, albeit of less practical value, is the mode. It is the maximum point in the distribution (i.e. in the pdf, approximated by the maximum in the histogram). Sometimes there are two maxima; in that case, we speak of a bimodal distribution. For a symmetrical unimodal distribution, the mode agrees with the mean.

The shape of the actual distribution will often deviate from the theoretical normal distribution. If it is longer tailed to the larger values, then we speak of a positive skewness, or to smaller values, of a negative skewness. If the distribution is peaked more than the normal distribution, then we speak of a leptokurtic, or in the opposite case, of a platykurtic, distribution. The shape measures are defined as follows:

$$skew = (1/n)\Sigma\,[(x-Mean)/s]^3$$

$$kurt = (1/n)\Sigma\,[(x-Mean)/s]^4 - 3$$

A strongly leptokurtic distribution (kurt1.0) indicates the presence of data in the tails of an otherwise well concentrated distribution. Aplatykurtic characteristic (negative kurtosis) can be an indication for a time variability in the data. If the measurements drift with time, then the result is a flat distribution. Values for the kurtosis of less than -1 are a danger sign and suggest a closer look at the data: we may have to subtract a trend before we investigate the residuals. Similarly, values in excess of 1.0 are also suspicious, in this case there may be outliers. Values between -0.5 and 0.5 are probably harmless and a result of the considerable variance in the measure of kurtosis itself, which is very sensitive to outliers. All real world data have somewhat more points in the tails than expected from the normal distribution and skewness and kurtosis are important diagnostics to find out quickly whether and where problems are. An excellent visual aid is, of course, the histogram of the measurements. A simpler and almost more useful trick is the display of the ordered (by absolute size) measurements which have been numbered to provide for an abscissa. The inverted S shape gives an immediate idea about the distribution of the errors. This plot is, of course, nothing else than the cumulative distribution function (cdf), plotted with the axes inverted.

There are various alternatives to estimate the dispersion in the data. The simplest is the range R:

R = x(max) - x(min)

or, for an ordered set,

R = x(n) - x(1).

One also uses the "relative range" = R/s also called normalized range. The range has several serious disadvantages: it depends very much on the sample size n, has a substantial variance, and is totally affected by outliers. Table 4 gives estimates for the "n" dependency of the range [8] as expressed in units of the standard deviation (which is the normalized range).

n	Range	n	Range
2	1.13	11	3.17
3	1.69	12	3.26
4	2.06	13	3.34

5	2.33	. . . . .	. . . . .
6	2.53	50	4.50
7	2.70	100	5.03
8	2.85	1000	6.49
9	2.97	5000	6.67
10	3.08	10000	≈ 7

Table 4. Estimates for the Range (n)

In fact, the range could be used as a good outlier detector in combination with the variance, if that were not so inconvenient because of the dependence on n and its lack of efficiency (its considerable variance) to boot. However, the variance itself is also quite sensitive to outliers! A much better detector is indeed a combination of the standard deviation as an outlier sensitive measure with a measure which is insensitive. Such a robust measure for the dispersion is the median of the absolute deviation (meddev); indeed, it is the least sensitive dispersion estimate of them all.

meddev = med{|x - Median|}

The meddev should be an excellent estimator for the probable error as defined above, but one finds little about this in the literature. Simulation tests with pseudorandom data sets indicate that this estimator for the dispersion of the distribution is not only extremely robust (because we use only position), but its efficiency is just sufficient to serve for the first phase of our routines, the detection process, at least for reasonably large sample sizes. Of course, for very small samples (n<16) it becomes impossible to find a center of the distribution with any confidence. Without having seen much analytical work regarding this estimator, it is, nevertheless, in my opinion certain that its variance goes with $n^{-½}$ (because we are only counting). Therefore, we should improve our relative precision with the square root of n. Another question is whether one should not use the sample mean, instead of the median, as the reference point for the deviations. The median is less sensitive to the outliers. This method of outlier detection and elimination is quite different from the excess reduction method outlined above, and it is simpler and often sufficient. We can call our first method the (significant) excess method, and the use of the sigma/meddev ratio criterion just the ratio criterion method.

These methods can be combined easily. A combination is useful because a single outlier that is far off is immediately recognized by the ratio method while the excess method requires several steps to find out how far out a single outlier is. On the other hand, the excess method is useful in other cases where one wants to have a hard, clear criterion on when to stop rejecting data. The ratio method, of course, also gives this criterion, but counting numbers is more robust than looking at a slowly changing ratio. Therefore a combination is a good idea in principle. This combination will be even better if we include the kurtosis as an indicator that there are data in the wings which may not belong there.

Back to Contents

TIME SERIES, MODELING, AND FILTERING OF DATA

Until this point, we have considered a collection of measurements without reference to their possible time dependency. If such a dependency exists, then the computation of statistics, such as the mean or the standard deviation, becomes meaningless. Such estimates make sense only for such portions of data as can be assumed to have been drawn from the same (constant) distribution. However, by modeling the time dependency, it is possible here, too, to separate data that we assume reflect the process of interest from extraneous disturbances that appear to contaminate our measurements. The random part of the "signal", some of it being the contamination (the residuals), is now called noise instead of deviation as in the static case (the extreme residuals still qualify as outliers, if we can show that they come from a distribution different from the rest). It should be noted, however, that in time series analysis it is usually even more important to recognize and evaluate the internal correlations of the random residuals than to do a detailed analysis of their magnitude distribution. That analysis is only required for outlier rejection. The systematic part, the part which we remove by fitting a model to the data, is also called trend, and its successful treatment, the de-trending, is by far the most important part of our work.

An important problem must be considered first, even though it is only related, but does not strictly belong, to the subject of filtering or rejection of outliers. This is the fact that, for practical purposes, we look at our data with a view towards predicting future developments. In this case, it is important to realize that, a priori, we can state that the most recent data are the most meaningful for prediction, because old data lose their significance fast due to the ever present changes in all processes. This means that we may want to filter our data with a priori weights assigned that increase towards the end of the data. But this also means that we are giving greater weight to an outlier if it should occur at the end of our data. This is another aspect of the statement that a causal filter is not as effective as a filter which can look both ways, forward and backward in time. Whether there is a genuine change or whether it is just an outlier, this question can only be decided in retrospect when we have data coming after the outlier. In robust statistics the distinction is often made between additive outliers (points outside the trend) and innovative outliers which start a different trend, or an offset in the trend. In the second case, a new trend has to be started, whereas in the first case, a simple rejection of individual points will take care of the problem.

The alternative of unequal weighting can, of course, also be applied in the static case if, for some reason, we have a priori reasons to have particular confidence in some points (or suspicion in others). Such an inclusion of a priori information is more effective but also more dangerous (again, there is nothing free!) compared with the unweighted case. It is for this reason that many statisticians, as a matter of principle, reject the inclusion of a priori information in the derivation of statistical information, because such inclusion cannot be done entirely objectively. Nevertheless, for that reason alone we should not refrain from using our best judgment if it is based on our total experience (and not just on beliefs and prejudices).

The principle of filtering is again the same as the one in the static case, except now we apply a least-squares fit to create a mathematical model such as a linear fit, look at the residuals by comparing the rms value with the median absolute deviation, and, if necessary, reject the point that is farthest from the model. Then we repeat the fit until the accepted criterion is reached. One criterion is the ratio of the rms deviation to the meddev of 1.5, which is slightly larger than what a purely random Gaussian (normal) distribution of residuals produces. If we lower that limit, then we will, of course, delete more data points and the routine will search for, and eventually settle on, the densest part of the distribution.

It is instructive to experiment with artificial data generated by a random-number generator. In this case, we know the outcome and can see the results of introducing one or several outliers. For this reason, a program has been included in the diskette for the generation of data with a uniform and a normal distribution. One can also experiment with data from actual data sets, such as those from the USNO automated data service (ADS). (I don't want to imply, however, that our data are so noisy as to always require robust methods!). As a further aid for such exercises, we provide several auxiliary programs which assist in editing data or adding line numbers to them, a simple way to make a time series out of an otherwise amorphous data set, or to facilitate the generation of the cdf.

The filtering by linear fits is the simplest way, but quite effective, for the suppression of noise. An even simpler smoothing would be a running average, which is totally independent of any models. In this case, however, it will be harder to develop criteria for the rejection of outliers. Yet, this is the direction in which we find the most sophisticated modeling: the Moving Average, the Autoregression, the mixture of the two (ARMA), and the integrated version (ARIMA). The authoritative reference for this is the text by Box & Jenkins [1]. Several commercial statistics packages provide programs based on these models.

It is not our goal, however, to compete with these software sources, but rather to discuss the principles and alert to possible pitfalls in the application of robust methods. It is also not at all established that the most sophisticated methods are also the best in the sense of safety. It is for these reasons that a simple, well understood, and yet extremely robust "smoothing" procedure is so robust because it uses "windowed medians," with the window width adjustable according to the number of data points available (a number which depends on the sampling rate). The windowed medians have been tested on a variety of data and the program can handle very noisy data indeed. This is particularly true if the data have been sampled at a high rate, which is just another way of saying that, in the presence of noise, you must have as much data as possible. Another advantage of this method is that it is independent of any model assumptions, i.e., it is a strictly non-parametric procedure that follows the general trends, whatever they are. For each point, the filter uses the set of the neighboring 2n+1 points, with n taken as the one-sided width of this 'window' of data points. If we now take the median of this set of points and assign this value to the point in question, then the claimed robust filtering takes place. This method is excellent for the purpose of providing an estimate for the center values but it has some drawbacks in other regards. Most importantly, it does not produce a "continuous" result and the values can only come from the set of given data points.

You find in the literature of robust data analysis that almost every author seems to favor his own special method. It is very hard to obtain estimates of comparable merits of these methods. I am, however, convinced that the principles which we have used remain valid in time series analysis: Use a robust method to identify and reject outliers and then use a method based on the assumption of a normal distribution of the remaining random part of the data. Or, in other words, fit a model after outliers have been rejected. This will require some iteration of experimental fits, with outlier rejection alternating with state-of-the-art model fitting.

Back to Contents

DANGERS AND PITFALLS IN DATA MODELS

A particularly notorious danger in model fitting is the over-confidence that one can develop, particularly if we apply "canned" programs without really understanding the basic problems. One is tempted to give too much credence to results which may be completely wrong and utterly misleading. This can happen if noisy data are fitted with higher order mathematical models, particularly polynomials, in the belief that greater complication will necessarily produce superior results. The danger inherent in this must be discussed from several angles. First, the normal equations for the solution of a least-squares fit are notoriously ill conditioned (the determinants are close to singularity) unless a data transformation is made before the analysis to bring the coordinate origins into the range of the data given. If that is not done, then even small rounding errors (forget the disaster due to real outliers!) will produce unacceptable errors that make the fit worse than useless.

The second danger is a blind belief in a meaningfulness of higher polynomial coefficients that are obtained by using a "canned" routine. The pitfalls come from the fact that even in the case of a data transformation as discussed above, the matrices involved in such solutions are close to singularity where rounding errors can lead to completely erroneous results if there is not a sufficient basis for the computation of these higher order coefficients. In simple terms, one should not fit a ninth order polynomial to a noise contaminated straight line even if we have thousands of points! The right way to avoid these pitfalls in a quasi automatic fitting of higher order polynomials is the solution of such a problem by the use of the singular value decomposition (SVD). This is discussed exhaustively in [12].

Yet another danger is the use of routines without some control by visually inspecting the results in a plot or graph. For example, linear fits may completely overlook nonlinear trends. Visual inspection is also of great importance if any of the automatic ARMA or ARIMA models are being used, because the solution of the model coefficients depends critically on the estimates of the autocorrelation (again, particularly with the higher coefficients of the more complicated models). These estimates can be seriously distorted by outliers in the sample. The same thing is true for Kalman filtering, which in its inner workings is completely opaque to the user; the more so, the greater the number of parameters involved. Visual inspection is virtually the only really safe way to understand data. Therefore, the use of some plotting routines is highly recommended.

Lastly, an often neglected principle: make as few assumptions as you can without doing injustice to the data. This is basically "Occam's razor," which is now more often referred to as the principle of "parsimony" (especially by Box and Jenkins, [1]). This means simply, that a linear fit is better than a higher power fit, unless there is a very clear basis for the inclusion of higher powers. The SVD method is ideally suited for this, because the lack of meaning of higher power coefficients is reflected by the smallness of the corresponding singular values. The same thing is true for the use of orthogonal polynomials, which had been in favor before the SVD was shown to accomplish the same thing with less effort. Similarly, the use of weights must also be counted as an inclusion of additional hypotheses in our routines and must be looked upon with a priori suspicion.

The deep reason for parsimony is that the inclusion of higher terms, or of more details, in any model or theory more or less inevitably will bring purely random and accidental noise power into the systematic part of the model, making it misleading for extrapolations. This is the reason why the celebrated Ernst Mach was wrong when he postulated that science selects the simplest laws because of an "economy of thought." It is not economy but facts of information theory that yield the ideas that are important in our context. As Jeffreys explains it, the simplest laws have the greatest probability of making correct predictions in the largest number of cases (see [9] and also, for empirical evidence, [13]). Another interesting and persuasive example for the need to keep parsimony in mind is given by Scheid [14]. His example is particularly instructive because it corroborates the importance of parsimony in the basic assumptions in a simple and transparent case of finding the power of a noise contaminated low order polynomial.

We have several choices in building models. The interpolating functions are only useful if we have great confidence in the integrity of the data and their freedom from random noise. In this case, a cubic spline is superior over polynomial fits because the spline connects the given points under a more "physical" constraint (a minimum total curvature) than polynomials, which are purely mathematical fictions.

All approximating models, Fourier series, or general orthogonal function fits, often produce stability problems (the matrices involved in the least-squares solutions are close to singularity), and the use of the Singular Value Decomposition (see [12]) is generally advantageous, not only in fitting polynomials. The best known orthogonal functions, Chebyshev polynomials, spread out the residual errors more smoothly. The Chebyshev polynomials are minimax functions in the sense that they minimize the largest deviations. But that makes them least robust! These functions are excellent for interpolation of tabulated precise functions. They are useless for the analysis of real data.

In general, if we know nothing about the underlying physics of the process, we must avoid by all means the complicated global models. I prefer piecewise fitting with linear or quadratic trends. A failure to look for breaks in the data and instead, fitting the data blindly with high order polynomials is a frequent but cardinal mistake of data analysts. Natural processes have a tendency to continue for a while until a break occurs. From then on, a new process must be fitted. That is the reason for the preference of piecewise fits. It is also the reason why splines are more useful than higher order polynomials. In the latter case, any remaining outlier will affect the whole data set. In contrast, in a local model, only the region in which the outlier occurs, is affected. An entirely different situation exists when the environment is the determining factor, such as in seasonal effects. In this case, the coefficients will not remain the same over time, but the phase will more likely be preserved. Again, the criterion for goodness of fit is the absence of trends in the residuals. If the residuals are random with a white spectrum, all the systematic information has been extracted into the model.

Our concept of identifying breaks in the trends, and the use of piecewise fits with simple trends, is just another way of treating the"innovative" outliers mentioned in the Introduction. In contrast to the "additive" outliers which do not change the trend but add short term noise to the "signal", the innovative outlier signals the beginning of a different process. In the case of GPS carrier tracking, e.g., a cycle "slip" will be an innovative outlier. And again, the decision about the difference can only be made in hindsight, after more data have been considered.

The statistical models, such as the ARIMA models, have found great favor during the last two decades. The problem remains, however, to first subtract the trends. ARIMA methods as such cannot, therefore, be included under robust methods without these qualifications.

In my own, admittedly biased opinion, the best way to handle large amounts of critical data would be the following: Obtain as high a sampling rate as possible and then, establish a raw but robust model by filtering with the "median window". On the basis of this tentative reference, look for, and reject outliers that are beyond an adopted limit. In many cases, e.g., where the trend is approximately linear, the rejection can be done without that tentative reference and pre-filtering, by directly fitting a linear model with a program. However, in the general case it is easier and better to filter with a non-parameteric program because, this method makes no assumptions whatever about the trends. A direct model fit must be done, however, if the rate of change of the trends is high compared with the data density.

Either way, we meet the same problem of rejection limit setting that we have discussed in the case of the normal distribution of a set of measurements, except now we consider the residuals from the tentative model as the set of values that are somehow distributed with the possible admixture of genuine outliers. Again, a visual inspection of the filter process is invaluable to prevent excessive filtering. This concludes the part of data analysis which can be called robust. From now on we deal with interpolation or prediction on the basis of the filtered remaining data and standard methods can be used.

If, after filtering, we have confidence in the signal-to-noise ratio of the remaining data, we might use a cubic spline as the final interpolator and model. In the case that the residuals are dominated by noise, however, a statistical model should be used after the trend has been modeled separately by least-squares fit of an appropriate model such as Fourier series or polynomials. The production of this trend model is usually by far the more important and critical issue compared with the structure of the residuals, which is more of diagnostic interest. References [15] and [16] can be recommended for further study at this point.

Back to Contents

ESTIMATES FOR THE MAGNITUDE OF NOISE IN TIME SERIES

The noise in a time series can be easily estimated if we have produced a good model fit. The desired measure is simply the rms residual, which we obtain as a by-product of the fit. It is possible, however, to obtain excellent estimates of the noise even without fitting by using the rms successive difference of the data. Within the PTTI community, this is better known as the Allan variance, because David Allan introduced it (with an additional factor of 2) for the measurement of frequency instability of oscillators [17]. The use of the successive differences can, however, be a possible source of confusion. We always form the first differences of the quantity in question. In the case of frequency measurements, we form the "meansquare" of the second differences of phase. And this is really the first difference of the quantity which we use for the mean-square successive difference analysis of noise content, and we have no exception to the rule.

The principle of this noise characterization is simple: Instead of looking for the sample standard deviation, which for a time-variable measure may not converge to a finite value if we add more and more data(if the measurements move around), we form the mean-square successive difference (and divide by 2 for the Allan variance) as a measure of the noise present, the $\sigma_y(\tau)$. By comparing this value with the sample variance, we obtain additionally an estimate for the kind of noise present. For this purpose, we form the ratio B_1 of $\sigma(\tau)$ over $\sigma_y(\tau)$. A value of unity of B_1 is indicative of a random "white" (no frequency- dependent) noise (which allows meaningful averaging over longer periods). The more internal correlation is present in the data, the more the sample statistics are going to wander around. This is part of what is meant by non-stationarity of the data, and it will be reflected in an increase of the sample variance relative to the Allan variance, i.e., an increase of B_1 beyond 1.

In summary, we have to compute:

a. the sample variance and its square root σ.

b. the mean-square successive difference, divided by 2 (for making it commensurate with σ , following Allan's procedure), and extracting the square root. For the sampling time τ this is now the two-sample (Allan) deviation $\sigma_y(\tau)$.

c. Find the ratio $B_1=(\sigma/\sigma_y(\tau))^2$ as a simple test for the whiteness of the data (B_1=1?).

An even better test is available by computing the autocorrelation function of the residuals from a model fit. If the residuals are white, then the autocorrelation function must drop to near zero immediately at lags0. One has to make sure, however, that the data values are listed without gaps. Confidence intervals for the rms difference (without the factor 2 in the denominator because the Allan measure is not used outside the PTTI community) and for B_1 can be found in the literature ([8], para. 4.7, pp. 373-375).

Back to Contents

DEALING WITH REAL-TIME NAVIGATIONAL DATA

All of the above cases assumed that we can afford the luxury of fitting and filtering data post facto. In this case the critical decision of what is random and what is systematic, or what trends must be subtracted before the filtering, can be handled reasonably well. However, a main reason why this is relatively easy is the possibility of using non-causal filters. In other words, if we see where the trend went after a questionable point, then we can decide to reject or not to reject, or as the case may be, to break the model and start a new trend. In the case of real-time data, the decision between the case of an additive outlier or a break (new trend) cannot be made immediately but only after some delay. This is discussed extensively by Salzmann[20] who uses Kalman filtering (and gives many references). In contrast to our discussion (he uses the concept of "Reliability" of the information obtained and not robustness of the procedure), his treatment is concerned with block procedures that require iteration if "slips" (innovative outliers) are detected. He must, therefore, accept a small delay in the detection process. He also mentions briefly "robust" Kalman filters and gives references. However, he agrees that this can only deal with additive 'model errors'. But it is possible to avoid the delay and obtain almost

instantaneous detection of integrity failure. To this end we must search for additional information that can help us in making the rejection decision. This is the typical case in electronic navigation (GPS) where a decision about acceptance or rejection of data must really be made on the spot.

Essentially, there are only two things we can do: We can build criteria on the basis of rate of change, which is very unreliable for the reason that we must differentiate a noisy signal; or, much better and effective, we must have redundant information and/or obtain information from elsewhere. In many cases, this is easier than it appears, but it does extract some extra costs for this additional information. In the case of aviation applications and the GPS, the problem has been extensively investigated under the acronym RAIM, receiver autonomous integrity monitoring. Additional information is available from a variety of sources: if we monitor more satellites than the minimum four; if we use (inexpensive) inertial information as acceleration reference; if we add altimeter information to the navigation processor; and finally and very effectively, if we use a sufficiently stable clock as a short term (<600s) system time reference. The more information we provide, the simpler and more reliable the detection of outliers (or the integrity failure) becomes. Only with the help of such additional information can an effective and reliable rejection criterion be used. At that point, after eventual rejection of some data, the redundant information will provide us with the additional and very substantial benefit that we can improve performance by appropriate integration. This can lead to a great improvement of accuracy, not only because of noise reduction, but even more so because with the minimum number of satellites, the notorious geometric dilution of precision (GDOP) comes into play each time when position lines come close together. This situation can be completely avoided with more than the minimum information as the basis for the navigation solution. By avoiding the occurrence of the large uncertainties during periods of poor geometry altogether, the average position accuracy is then hardly affected by GDOP. This makes the RAIM a very cost effective solution.

Back to Contents

APPENDIX 1: The Poisson Distribution

The probability that exactly x events take place during the unit interval is $P(x \mid \lambda) = [(\lambda^x)e^{-\lambda}]/x!$ where λ is the average rate of occurrence per unit interval (>0) and x is a positive integer. Therefore, the distribution is fully characterized by λ, which is also equal to the mean and the variance of this distribution: $\mu = \lambda$; $\sigma^2 = \lambda$. Example: On the average we count 10 cars passing on the highway per minute. What is the probability that we once count 5 cars in a minute? $P(5 \text{ per m}) = [(10^5)\cdot 4.54\cdot 10^{-5}]/120 = 0.038$ We can expect to count 5 cars in a minute in about 4% of the cases. The Poisson distribution is not symmetric for small λ (it has a positive skewness), but becomes increasingly symmetric as λ increases. The maximum is at the largest integer x that is less than or equal to λ . That means that, for $\lambda < 1$, the most frequent occurrence is zero, as can be seen on the table below:

λ	0.1	0.2	1.0	2.0
P(x=0)	0.905	0.819	0.368	0.135
P(x=1)	0.090	0.164	0.368	0.271
P(x=2)	0.005	0.017	0.264	0.594

For a large λ , the Poisson distribution approaches the normal distribution, which is its limiting case. If, e.g., we count on the average 10000 items per unit of time, then the actual distribution of our counts will look very much like a normal distribution around the mean 10000 with a σ of %10000 =100. This is a good example for the counting of cesium atoms arriving at the detector in an atomic clock, and we can see why the frequency variations decrease with increasing times. The counting variance is equal to the counts and the standard variation goes with the square root. Therefore, the relative fluctuation of the total counts goes with 1/%n. That is the origin of the -2 slope in the logarithmic sigma-tau plots. However, all this is true only as long as our assumption of statistical independence of the individual events remains true. And indeed, for very long times it becomes increasingly hard to build clocks which can preserve this independence, because of environmental sensitivities and internal relaxation phenomena.

Back to Contents

ACKNOWLEDGEMENT

I am grateful for discussions with Dr. James Barnes of Austron, Inc., who advised me on several important points in a variety of subjects, including statistics and frequency measurements.

Back to Contents

REFERENCES, NOTES, AND LITERATURE

[1] George E. Box and Gwilym M. Jenkins (1976) "Time Series Analysis, Forecasting and Control," Holden Day, San Francisco. 77-79534. The standard work on ARMA and ARIMA modeling.

[2] S. M. Stigler (1973) "Simon Newcomb, Percy Daniell, and the History of Robust Estimation 1885-1920," J. Am. Statistical Assn., Vol. 68 pp. 872-879.

[3] P. J. Huber (1981) "Robust Statistics," John Wiley & Sons, New York.

[4] R. L. Lauer and G. N. Wilkinson (eds.) (1979) "Robustness in Statistics," Academic Press. These proceedings of a workshop on robustness contain useful contributions for a deeper study of the subject, particularly robust smoothing and robust estimation for time series autoregressions.

[5] Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel (1986) "Robust Statistics: The Approach Based on Influence Functions," Wiley-Interscience, New York. This new book of which I know from Huber's review in the American Scientist (July-August 1987, p.436) promises to be an important source of information concerning robust statistics. Hampel's approach (see also [19]), the influence functions, is intuitive, more realistic, and less "academic" than most other theoretical treatments of the subject, possibly with the exception of Huber's game theoretic approach (see [3]) which optimizes the worst-case performance of a statistic. At any rate, for all who want to go beyond the necessarily practical and even a little simplistic approach taken here in these notes, the study of a work such as this will be advisable.

[6] D. B. Percival (1982) "Use of Robust Statistical Techniques in Time Scale Formation," Report under Contract N70092-82-M-0579 (Naval Observatory). This was presented at the Second Symposium on Atomic Time Scale Algorithms, June 1982, Boulder, Colorado.

[7] Vic Barnett and Toby Lewis (1984) "Outliers in Statistical Data,"Second Edition, John Wiley & Sons, New York, ISBN 0-471-90507-0

[8] Lothar Sachs (1984) "Applied Statistics," Second Edition, Springer, ISBN 0-387-90976-1, LCC # QA276.S213 84-10538. This is a handbook written for non-mathematicians by a practitioner with a medical and biometric background. It can give a sound basis for the whole field of applied statistics, of which robust statistics is just a specialty, albeit an important one.

[9] Harold Jeffreys (1939) "Theory of Probability," Oxford At The Clarendon Press. See p.4 and p.345, where J. discusses the reason for parsimony: "..variation must be taken as random until there is positive evidence to the contrary." This statement is a special case of Occam's famous principle as it applies to mathematical models. On p. 255, J.discusses Chauvenet's Venus data (and gives the reference). For J., Probability Theory is the theory of reasonable degrees of belief.

[10] David T. Mage (1982) "An Objective Graphical Method for Testing Normal Distributional Assumptions Using Probability Plots," The American Statistician Vol.3 6/2, pp. 116-120. This paper is a useful introduction into the testing for normalcy. It also discusses, and gives references to, the often used Kolmogorov-Smirnov test procedure.

[11] Edwin L. Crow (1964) "The Statistical Construction of a Single Standard from Several Available Standards," IEEE Transactions on Measurement and Control, pp. 180-185 (December).

[12] George E. Forsythe, Michael M. Malcolm, and Cleve B. Moler (1977)"Computer Methods for Mathematical Computations," Prentice-Hall, Englewood Cliffs, New Jersey, QA297.F568 76-30819, ISBN 0-13-165332-6. This is a very useful book, in many ways a forerunner of [18]. The programs given are in FORTRAN.

[13] Spyros Makridakis and Michele Hibon (1979) "Accuracy in Forecasting: An Empirical Investigation," J. R. Stat. Soc. Assn., Vol. 142, Part 2, pp. 97-145. The paper with discussion, in addition to its academic merits, is a good example for disagreements between experts on the meaning of a rather clear situation that is supported by a substantial amount of evidence. It is a study which gives empirical proof for what was derived above on purely theoretical insight. This is the need to keep the models as simple as possible. The study confirms that, not surprisingly. In a wide variety of applications, the simplest models have been superior to even the best Box-Jenkins predictions (which probably have been too complicated).

[14] F. Scheid (1968) "Numerical Analysis," Chapter 21, p. 250, Schaum's Outline Series, McGraw-Hill, New York. Scheid demonstrates that the finding of the degree of a noise contaminated polynomial is not a trivial affair. Again, it relates to the merits of making minimum assumptions, in this case using the lowest degree polynomial that leaves residuals without trends.

[15] Eric A. Aubanel and Keith B. Oldham (1985) "Fourier Smoothing without the Fast Fourier Transform," Byte, February 1985, pp. 207-218. This is a BASIC program which is quite simple and effective.

[16] J. Vondrak (1977) "Problem of Smoothing Observational Data II,"Bull. Astron. Inst. Czech. Vol. 28, No. 2, pp. 84-89. This paper references, and improves upon, an earlier contribution (1969, Vol. 20, pp. 349-355). The use of the method as a frequency filter is discussed and the use of ("natural") cubic spline interpolation is included to assure overall smoothness. V.'s method has become a quasi standard in the Earth orientation data processing. The method is not robust and, therefore, should only be used after pre- filtering. It is most important, however, to realize that this filtering must only look for"additive" outliers with genuine real changes left intact. This distinction can be made only in retrospect (with non-causal filters).

[17] CCIR: Recommendations and Reports of the CCIR, XVIth Plenary Assembly Geneva, 1986, Vol. VII, Standard Frequencies and Time Signals. Report 580-1, "Characterization of Frequency and Phase Noise," is the best and most complete current reference for the pertinent measurement practices in the PTTI community. It includes a very comprehensive list of references and will be updated regularly (i.e., revisions will appear as necessary at the next Plenary Assembly).

[18] William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling (1986) "Numerical Recipes," Cambridge University Press, ISBN 0 521 30811 9, LCC # QA297.N866 85-11397. This is a treasure house of ideas. It gives many useful FORTRAN routines for the practicing scientist. Recently, a special edition in 'C' has appeared which is even more useful.

[19] Frank R. Hampel (1974) "The Influence Curve and Its Role in Robust Estimation," J. Am. Statistical Assn., Vol. 69, #345, pp. 383-393. [20] Martin Salzmann (1993) "Least Squares Filtering and Testing for Geodetic Navigation Applications," Netherlands Geodetic Commission Publication on Geodesy, New Series Number 37, Delft, The Netherlands. ISBN 90 6132 245 6 (FAX 31-15-782348)

NOTE: Items [21] through [24] bring useful background information on spectral analysis and filtering. Items [25] through [27] cover numerical analysis in general, and item [28], although not primarily concerned with robust filtering or statistics, is nevertheless an excellent discussion of the statistics of measurement and it is going to be of great importance to all who report precision measurements in an internationally coordinated terminology and format.

[21] G.M. Jenkins and D.G. Watts (1968) "Spectral Analysis and Its Applications," Holden-Day, San Francisco.

[22] Norman Morrison (1969) "Introduction to Sequential Smoothing and Prediction," McGraw-Hill, 69-17187. This is an excellent comprehensive text on the estimation of deterministic functions in the presence of additive noise. It treats polynomial estimators very thoroughly and generalizes to a treatment of state estimation of deterministic processes governed by differential equations. However,

robustness is not considered as such.

[23] Arthur Gelb (editor) (1974) "Applied Optimal Estimation." The M.I.T. Press, 74-1604, ISBN 0-262-70008-5. The state space approach is used in many practical examples. The book includes an important chapter on sensitivity analysis for suboptimal filter design. Such designs use, often drastic, simplifications in the state space model in the interest of fast computability.

[24] A. H. Jazwinski (1970) "Stochastic Processes and Filtering Theory," Academic Press, New York. [25] Philip R. Bevington (1969) "Data Reduction and Error Analysis for the Physical Sciences," McGraw-Hill, LCC # 69-16942. The programs are in FORTRAN. There is no discussion on robustness in this book, but it is a fine collection of essentials for data analysis.

[26] Enders A. Robinson (1967) "Multichannel Time Series Analysis with Digital Computer Programs," Holden-Day, San Francisco, LCC # 67-28043. This is a valuable source for many FORTRAN programs and subroutines with particular emphasis on filtering, spectral analysis, and signal enhancement in geophysical applications. Unfortunately, thanks to the tireless efforts by some committees, many of these programs, written in"classical" FORTRAN, will have to be modified because FORTRAN is not what it used to be: simple. I agree completely with the comments by Press et al [18] who urge language experts to resist adding to their compilers complexities that are used but rarely, yet increase the size and cost.

[27] R. W. Hamming (1973) "Numerical Methods for Scientists and Engineers," Second Edition, McGraw-Hill, New York, ISBN 0-07-025887-2, QA297.H28 72-12643.

[28] ISO, IEC, OIML, & BIPM (1992) "Guide to the Expression of Uncertainty in Measurement" Draft Report by the ISO/TAG 4/WG 3, distributed by BIPM. It is hoped that this report, to be widely disseminated to concerned parties, will greatly assist in reaching an internationally accepted standard for the fundamental concepts that concern metrology. Although not directly related to the subject of these notes, the report should be extremely helpful to all who are in some way concerned with the reporting of standardized measurements.

NOTE: Manuscript closed by G. Winkler on 17 May 1993. Edited for HTML, references to programs eliminated and Appendix II removed by W. Riley on 8 October 1998.

Back to Contents