Analysing data - by means and by no means

by

Nick Ridler[1]

National Physical Laboratory
% DRA, St Andrews Road, Malvern, Worcs WR14 3PS, UK

Abstract

This paper reviews some different methods of analysing measurements made using RF and microwave network analysers. These measurements are peculiar, in a statistical sense, in that each measurement value has two-dimensions. (Current standards on the use of statistics to analyse data concentrate on situations where the measurements are scalar, or one-dimensional, quantities.) The methods discussed in this paper are divided into two categories based on; (i) mean-based statistics, and (ii) order statistics. The relative merits of the categories are discussed.

1 INTRODUCTION

The use of statistics to analyse data is a vital part of most measurement processes. One of the key areas where statistics can be applied is in the interpretation of measurement results. For example, where something has been measured a number of times, statistics can be used to produce an average value and an indication of the scatter, or dispersion, of values around the average value. The average value can be used as our best estimate of the real value, and the scatter gives some information about the reliability of that estimate.

Average values and measures of dispersion are key concepts in the evaluation of random errors during measurement uncertainty estimates and in the treatment of results supplied by participants during measurement comparison exercises.

This paper looks at how we can apply these concepts to vector network analyser (VNA) measurements where the values are two-dimensional. In section 2 of the paper, mean-based statistics are reviewed and extended to allow for the two-dimensional case, and in section 3, order-based statistics are used to show how a resilience to the presence of poor measurements in a sample can be achieved without biasing the estimated average and losing confidence in its value.

2 MEAN-BASED STATISTICS

2.1 AVERAGE VALUES

The statistic most often used to provide an average value from a series of observations is the arithmetic mean, often just called the mean. It is simply calculated by summing the values of all the observations, then dividing this sum by the total number of observations.

---

[1] Mr Ridler is employed by Assessment Services Limited, and is working under contract at the National Physical Laboratory.

I.e., for a number, n, of observations, x_i, the mean, $\bar{x}$, is given by;

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

This works well for well-behaved scalar observations (*i.e.*, a series of single numbers). But if the observations are vector quantities (like those generated by VNAs), then each observation consists of *two* values, *e.g.*, the magnitude and phase of a reflection/transmission coefficient, or its real and imaginary components.

This raises the question; "How do we apply the above equation to a series of vector observations?"

2.1.1 Real and Imaginary versus Magnitude and Phase

There are two obvious approaches:

Approach 1 - calculate the mean of the real components and the mean of the imaginary components,

Approach 2 - calculate the mean of the magnitude values and the mean of the phase values.

Ide [1] has investigated the above two approaches and demonstrated that, in general, the two approaches yield different mean values. He also showed that different mean values can be produced depending on whether the phase values are expressed on either a ±180° scale or a 0° to 360° scale. Ide's article concluded with the recommendation that the statistical analysis of two-dimensional vector measurements should be performed on the real and imaginary components of the complex number (approach 1) and *not* the equivalent magnitude and phase components (approach 2). If, as is usually the case, final values are required in terms of magnitude and phase then these are arrived at by transforming the mean real and imaginary components to their respective magnitude and phase.

For a series of complex numbers, x_i, such that $x_i = Re(x_i) + j\ Im(x_i)$, (where $j = \sqrt{-1}$), Ide's recommendation (approach 1), expressed mathematically, is as follows;

Calculate (i) the mean of the real components, $\overline{Re(x)}$;

$$\overline{Re(x)} = \frac{1}{n} \sum_{i=1}^{n} Re(x_i)$$

(ii) the mean of the imaginary components, $\overline{Im(x)}$;

$$\overline{Im(x)} = \frac{1}{n} \sum_{i=1}^{n} Im(x_i)$$

The complex mean, $\bar{x}$, is then given by;

$$\bar{x} = \overline{Re(x)} + j\ \overline{Im(x)}$$

The **magnitude of the mean,** $|\bar{x}|$, is given by;

$$|\bar{x}| = \sqrt{\overline{Re(x)}^2 + \overline{Im(x)}^2}$$

and the **phase of the mean,** $\phi(\bar{x})$, is given by;

$$\phi(\bar{x}) = \tan^{-1}\left(\frac{\overline{Im(x)}}{\overline{Re(x)}}\right)$$

Approach 2, expressed mathematically, is as follows;

Calculate (i) the **mean of the magnitude** values;

$$|\bar{x}| = \frac{1}{n}\sum_{i=1}^{n} |x_i|$$

(ii) the **mean of the phase** values;

$$\phi(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i)$$

These two values are used as mean magnitude and phase values.

To summarise, approach 1 calculates the magnitude and phase of the mean value, whereas approach 2 calculates the mean of the magnitude and phase values. The reason that the two approaches produce different mean values is that the functions that transform real and imaginary components into magnitude and phase components, and vice versa, are non-linear.

2.1.2 Analysing phase data

Although the foregoing discussion recommended, in general, that a statistical analysis of complex data should be performed on the real and imaginary components on the data, this might not always be possible (*e.g.*, if only the phase component of the complex vector is measured) or ethical (*e.g.*, if participants in a measurement comparison exercise are asked to supply results in terms of magnitude and phase, should the data analyst then transform these parameters before analysing the data?) In these cases, the numeric phase values are analysed.

Special considerations are needed when analysing phase data because phase is measured on a scale which is periodic in nature. This means that a single observation can be represented with more than one value. For example, the values +45°, +405° (=45°+360°), -315° (45°-360°), +765° (45°+2×360°), etc, all represent the same data point.

For phase measurements from an ANA it is conventional to give the phase, ϕ, in the interval -180° to +180°. However when performing calculations on phase data it can be convenient to use other intervals, such as -360° to 0° or 0° to +360°, although the results of the calculation will usually be quoted within the original -180° to +180° interval.

Example: What is the mean of +179° and -177°?

(i) Using the interval $-180° < \phi \leq +180°$:

$$mean = \frac{+179° - 177°}{2} = +1°.$$

(ii) Using the interval $0° \leq \phi < +360°$:

$$mean = \frac{+179° + 183°}{2} = +181°.$$

Re-expressing this result in terms of the conventional interval gives:

$$mean = -179°$$

This second result seems a more sensible interpretation of the mean in this case.

For the above reasons, the range over which the values for the phase data varies needs to be taken into consideration when analysing phase data. Often, re-expressing a value (or values) will yield a more sensible result, as seen above.

2.2 MEASURES OF DISPERSION

It is often helpful, when reporting an estimated value, to provide some indication of the reliability of the estimate so that those using the estimate can assess its usefulness. In the case of a series of repeated determinations of the same observation, the dispersion of values gives an indication of the reliability of the estimated value.

The most common dispersion statistic, used in conjunction with the mean, is the estimated standard deviation, *s(x)*. It is calculated as follows;

$$s(x) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

As with the mean, this is useful for scalar observations but requires further consideration when applied to the vector quantities produced by VNAs. The problem is caused by the two-component nature of vector data, and the subsequent two-dimensional scatter of measured values.

Having chosen a mean value calculated from the mean of the real components and the mean of the imaginary components, we require an estimator for the scatter about this value.

2.2.1 Standard deviation circles

This problem was addressed recently by Ridler and Medley [2] during a measurement uncertainty investigation for vector reflectometers. They proposed the following modification to the classical standard deviation expression;

$$s(x) = \frac{1}{n-1} \sum_{i=1}^{n} |x_i - \bar{x}|^2$$

where $|x_i - \bar{x}|$ is the *distance* (a scalar quantity) between each observation and the two-dimensional mean. The resulting value for $s(x)$ represents the radius of a circle, centred on the mean, the size of which indicates the measure of dispersion about the mean.

This expression is interesting because it produces a scalar quantity (the circle's radius) as a measure of scatter in the vector data. It is also very simple to convert this circular measure of dispersion into a measure of the more commonly required dispersion in the magnitude and phase components of the results. Figure 1 shows a diagram of the situation, where Γ_m is the calculated mean vector with magnitude M and phase ϕ. The radius of the circle represents simply the variation in the magnitude, U_M, and, by constructing a right-angled triangle, the variation in the phase, U_ϕ, is given by;

$$U_\phi = \sin^{-1}\left(\frac{U_M}{M}\right)$$

This works well when the circular dispersion indicator is away from the origin in the complex plane. However, as a circle of this nature moves towards the origin, the width of the cone characterising the phase variation increases. When the circle is close enough to include the origin it is no longer possible to construct the cone and so all values of phase become admissible. Under these circumstances (*e.g.*, measuring the reflection from a well-matched load), the phase is essentially undefinable.

The representation of the scatter of two-dimensional vectors by a circle is a simplified treatment of a general problem of treating variation in two variables measured simultaneously. It has the advantage of being easy to calculate, interpret, and translate to the more often used magnitude and phase components. These advantages make it useful for providing customers with easy to understand information about the scatter of complex data.

2.2.2 The bivariate approach

The disadvantage with the standard deviation circles outlined above is that it tends to mask effects caused by variations in the individual components of the complex indication and also makes no allowance for any correlation which may be present between the two components.

These considerations lead towards a more general approach to the problem of a set of measurements drawn from a population represented by a joint probability density function of the several (more than one) variables under consideration. This approach introduces methods based on multivariate analysis, the details of which can be found in the literature [3,4]. Although multivariate analysis is an established tool in the field of statistics, it is rarely encountered in metrology - most metrologists inhabit a scalar world of measurements! We will outline briefly here this approach to the two-component situation, termed bivariate analysis. (Our components are the real and imaginary components of the VNA's vector indication.)

Firstly, a bivariate approach examines the variation in each component separately. This produces a dispersion indicator which can be characterised by an *ellipse*, instead of the previously determined circle. The shape of the ellipse is determined by the variation in the individual components. Where one variable varies considerably more than the other, the ellipse will be long and thin. When the variation in the two variables is the same, the ellipse degenerates to a circle.

Secondly, the degree of correlation between the two variables is examined. When there is no correlation between the two variables, the axes used to plot the variables coincide with the axes of the ellipse – see figure 2. When there is correlation between the two variables, the correlation coefficient, ρ, causes the axes of the ellipse to rotate with respect to the axes used to plot the

variables. Values of ρ vary between -1 and +1. When ρ = +1 or -1, the ellipse degenerates to a straight line. Figure 3 shows the ellipses produced for values of ρ = 0.0. 0.5, 0.7, 0.9 and 0.98.

The correlation coefficient between two variables, p and q, can be calculated from the estimated standard deviation (from before) for each variable, $s(p)$ and $s(q)$, and the estimated covariance between the two variables, $s(p,q)$. The covariance is estimated using;

$$s(p,q) = \frac{1}{(n-1)} \sum_{i=1}^{n} (p_i - \bar{p})(q_i - \bar{q})$$

The correlation coefficient, $r(p,q)$ estimates the degree of correlation between the two variables, using;

$$r(p,q) = \frac{s(p,q)}{s(p)\, s(q)}$$

In summary, the variation in the two individual components, characterised by their respective standard deviations, determines the length and width of the dispersion ellipse, and the degree of correlation, characterised by the correlation coefficient, affects the angle of rotation of the ellipse with respect to the parameter axes.

Although the bivariate approach is a more rigorous treatment of the problem of two-dimensional measurements, it has the possible disadvantage of being relatively difficult to calculate and translation to other parameter representations (such as magnitude and phase) is likely to produce asymmetric bounds which a customer may have more difficulty in using.

2.3 PROBLEMS WITH MEAN-BASED STATISTICS

The assumption underlying all the foregoing treatment is that the data under investigation has been drawn from a Gaussian, *i.e.* normal, distribution. However, there is little evidence to indicate that data produced by actual measurement systems conforms to this assumption of normality! Some of the main causes for this departure from normality are; (i) the presence of outlying observations in the data set, far-removed from the majority of values, (ii) a finite range of admissible values (non-infinite tails) for the parameter, (iii) deformations (skewness, peakedness etc) in the main body of the distribution, and (iv) numeric truncation of values produced by digital instruments.

This brings into question the validity of the use of the above statistics which, by nature, only perform optimally in the Gaussian environment. The main justification for the use of Gaussian statistics seems to be; "well, it's what we've always used" - which is a less than convincing argument!

3 ORDER-BASED STATISTICS

Another family of statistics which can be used to analyse measurement data is based on *order statistics* – applying analytical methods to the data set after first arranging the data in ascending numerical order of size. The ordered data set can be analysed in a variety of ways to give indications (as before) of an "average" value and a measure of dispersion. We will discuss here the use of some order statistics primarily to show that there are useful alternative ways of analysing data which can provide superior estimates to their Gaussian counterparts.

3.1 AVERAGE VALUES

3.1.1 The median

Perhaps the easiest value to find, and understand, from a set of order statistics is the median. Put simply, the median is the middle value (in order of size) of the results. If the number of results is an even number, a unique middle value does not exist, so the median is the midpoint of the middle pair of values.

The median was used recently to provide an average during the analysis of data obtained from a measurement comparison exercise [5]. On that occasion, the median was chosen due to its resilience to values within the data set which vary dramatically from the majority of values. (Such values are often called "outliers".) Part of the data analysis required deriving an average value for the data set which was used to assess the performance of each participant's result. The presence of outliers in some of the data sets meant that the median provided a more suitable average than the more conventional arithmetic mean.

To illustrate how the median can exhibit superior performance over the mean we will look at both statistics for one of the data sets obtained during the comparison exercise. Measurements of S_{11} of a 3 dB attenuator at 1 GHz, made by seven participants, were as follows; +118.2°, +120.6°, +115.2°, +108.9°, +112.9°, +116.3° and +114.0°. The mean value is ≈ **115.2°**. However, if we include a measurement made by another of the participants (who measured the phase to be -58.3°) and re-calculate the mean, we get ≈ **93.5°**. Including this extra measurement has caused the mean to change dramatically. This indicates that the mean is struggling to estimate a reliable average value.

To find the median for the first data set, we arrange the values in ascending order of size;

+108.9° +112.9° +114.0° +115.2° +116.3° +118.2° +120.6°

The middle value, or median, is **115.2°**.

For our eight results, arranged in order of size, we get;

-58.3° +108.9° +112.9° +114.0° +115.2° +116.3° +118.2° +120.6°

The middle two values are +114.0° and +115.2°, so the median is now (114.0+115.2)/2 = **114.6°**.

The median has remained relatively unaffected by the inclusion of the outlying observation. We can conclude, therefore, that the median is more likely to provide a reliable average value when a relatively small series of results, containing one or two contaminated values, is being analysed.

3.1.2 More refined average estimators

There are a great number of statistical tools which can be used to analyse measurement data. In recent times, emphasis has been given to developing statistical tools which are not adversely affected by some of the quirks exhibited by real measurement data – non-normality, skewness, outlier contamination etc. This has given rise to a branch of data analysis called "robust statistics". A review of robust statistics is beyond the scope of this present paper although there are several excellent textbooks on the subject (see, for example, Andrews et al [6] and Huber [7]). It is anticipated that, in future, robust techniques will play an increasingly important role in the field of measurement data analysis as these techniques get written into standards [8] and applied in actual measurement situations [9].

A considerable number of robust procedures cannot be expressed as closed-form expressions and therefore are usually evaluated iteratively. In most cases, a good starting point for the iterative calculation is the sample median – subsequent iterations usually take into account, and weight, the data with respect to successive adjustments in the estimated average value.

3.2 MEASURES OF DISPERSION

Measures of dispersion, used in conjunction with the median, can also be derived from order statistics. We will examine two of the simplest and most often used – the inter-quartile range (IQR) and the median absolute deviation (MAD).

3.2.1 Inter-quartile range

The IQR is the difference between the values for the upper and lower quartiles. Having ordered the data, the lower quartile divides the data so that 25% of the observations are below its value and the upper quartile has 25% above its value. However, ordered data cannot always be split exactly into four equal parts. This calls for more precise definitions for the quartiles, which are given in terms of a calculated 'depth', *i.e.*, how far to go through the data to find the required value. The depth of the quartiles is derived from the depth for the median. For n observations, the depth for the median is;

$$\textit{Depth of median} = \frac{n + 1}{2}$$

and the depth for the quartiles is;

$$\textit{Depth of quartiles} = \frac{\textit{Depth of median} + 1}{2}$$

The depth for the lower quartile is found by entering the data from the low end (*i.e.*, starting with the lowest value) and the depth for the upper quartile is found by entering from the high end (starting with the highest value).

The IQR is then simply the difference between these values. We will use the data used previously to demonstrate finding the median to illustrate finding the IQR. This will also demonstrate the resilience of the IQR to outlying observations.

Firstly, we will use the seven results which, when ordered, become;

+108.9° +112.9° +114.0° +115.2° +116.3° +118.2° +120.6°

The depth for the median is (7 + 1)/2 = 4, so the depth for the quartiles is (4 + 1)/2 = 2.5, *i.e.*, the quartiles are midway between the second and third values (when entering the data from either end).

$$\text{Lower quartile} = (112.9 + 114.0)/2 = 113.45°$$

$$\text{Upper quartile} = (116.3 + 118.2)/2 = 117.25°$$

$$\text{IQR} = 117.25 - 113.45 = \mathbf{3.8}°$$

Now, repeating the calculations using all eight results, we have;

-58.3° +108.9° +112.9° +114.0° +115.2° +116.3° +118.2° +120.6°

The depth for the median is (8 + 1)/2 = 4.5, and the depth for the quartiles is (4.5 + 1)/2 = 2.75, *i.e.*, the quartiles are three-quarters of the way between the second and third values (when entering the data from either direction).

For the lower quartile, the second and third values are +108.9° and +112.9°, and three-quarters of the way between these two values (in an ascending direction) is +111.9°.

For the upper quartile, the second and third values from the top are +118.2° and +116.3°, and three-quarters of the way between these two values (in a descending direction) is +116.775°.

The IQR is therefore 116.775 - 111.9 = 4.875° ≈ **4.9°**

The change in the IQR when the outlying observation (-58.3°) is included in the results being analysed is relatively small – increasing from 3.8° to 4.9°. (Compare this with standard deviation calculations for both samples – for the seven results, $s(x) \approx 3.8°$, and for the eight results, $s(x) \approx 61.4°$.) This indicates that the IQR is less vulnerable to outlier contamination than the standard deviation and therefore makes a useful dispersion indicator when a relatively small number of observations, containing a contaminated value, are being analysed.

3.2.2 Median absolute deviation

The MAD is defined as the median of the absolute difference of each value from the median. This can be expressed mathematically as;

$$MAD = median\{|x_i - \hat{x}|\}$$

where $\hat{x}$ is the median. Again we will use the same set of results used previously to demonstrate finding the MAD. For the seven results the median was 115.2°. Therefore the values for the $|x_i - \hat{x}|$ are;

6.3°	2.3°	1.2°	0.0°	1.1°	3.0°	5.4°

Ordering this data gives;

0.0°	1.1°	1.2°	2.3°	3.0°	5.4°	6.3°

The median of these ordered absolute deviation values is the middle value, *i.e.*, MAD = **2.3°**.

Now, repeating the process for the eight results (including the outlying observation in the set) with median = 114.6°, the $|x_i - \hat{x}|$ values are;

172.9°	5.7°	1.7°	0.6°	0.6°	1.7°	3.6°	6.0°

Ordering this data gives;

0.6°	0.6°	1.7°	1.7°	3.6°	5.7°	6.0°	172.9°

The median of these ordered absolute deviation values is (1.7 + 3.6)/2 = 2.65°, *i.e.*, MAD ≈ **2.7°**

The change in the MAD when the outlying observation is included (like the change in the IQR) is relatively small – increasing from 2.3° to 2.7°. This indicates that the MAD, like the IQR, exhibits a high degree of resilience to outlier contamination. In fact the MAD is more resilient than the IQR since it uses fewer of the values of the data set to establish its value.

There are some useful equivalence relations between MAD, IQR and the standard deviation $s(x)$ for large sample sizes drawn from a normal distribution which can be used to compare intervals produced by each statistic. These are;

$$\text{IQR} \approx 2 \times \text{MAD}$$

$$\text{IQR} \approx 1.349 \times s(x)$$

$$s(x) \approx 1.483 \times \text{MAD}$$

These only apply if the distribution is normal.

3.2.3 Other measures

In recent years a range of robust dispersion estimators have been developed to accompany the robust average estimators. A detailed examination of these robust scale estimators is beyond the scope of this paper – they will be discussed only briefly.

Most of the work in the field of robust interval estimation has been aimed at producing a variance-like interval (or standard deviation) which can be suitably scaled using a Student's t type statistic (often with an adjusted number of degrees of freedom) to provide a level of confidence for the interval. (See, for example, Gross [10].) The fruits of this labour show great promise for metrological applications since the metrologist's "end product" is a stated result accompanied by a level of confidence in the result.

A significant number of the robust dispersion estimators use as their basic value the MAD and apply refinements and higher-order expressions derived from the classical standard deviation to achieve their confidence interval.

4 CONCLUSIONS

This paper has reviewed some of the statistical techniques used in the analysis of measurement data, specifically, that produced by RF and microwave vector network analysers. Gaussian statistical techniques have been re-evaluated in the context of a vector environment and less well-known techniques (median, IQR, MAD) based on order statistics have also been discussed. The proven resilience to outlier contamination can be very useful in the case of measurement blunders, which inevitably happen in real measurement processes, since "blunders in recording or analysing data can introduce a significant unknown error in the result of a measurement. . . . Measures of uncertainty are not intended to account for such mistakes" [11].

Relatively new statistical techniques exhibiting robustness to departures from normality in the measurement data have also been discussed briefly. It is anticipated that these newer techniques will make a significant impact in the field of metrology in years to come, as more and more reliable information will be expected from measurement data produced by increasingly more sophisticated and automated test equipments.

5 ACKNOWLEDGEMENTS

The author would like to thank his colleagues John Medley, Jezz Ide and Tim Hodgetts for useful discussions concerning several of the points discussed in this paper.

6 REFERENCES

[1] IDE, J.P. "A comedy of errors!" or "That's not what I mean by 'mean'", in RIDLER, N.M. (Ed), ANAMET – ANA*lyse* Notes numbers 1 to 10, *National Physical Laboratory Report DES 140*, to be published, October 1995.

[2] RIDLER, N.M. and MEDLEY, J.C. An uncertainty budget for VHF and UHF reflectometers, *National Physical Laboratory Report DES 120*, May 1992.

[3] WILKS, S.S. *Mathematical Statistics*. Princeton University Press, Princeton, New Jersey, 1943.

[4] KENDALL, M.G. *The advanced theory of statistics, I and II.* Griffin and Co, Ltd, London, 1946.

[5] RIDLER, N.M. and MEDLEY, J.C. A comparison of complex scattering coefficient measurements in 50 ohm coaxial line to 26.5 GHz, *National Physical Laboratory Report DES 138*, June 1995.

[6] ANDREWS, D.F., BICKEL, P.J., HAMPEL, F.R., HUBER, P.J., ROGERS, W.H. and TUKEY, J.W. *Robust estimates of location: Survey and advances*, Princeton University Press, Princeton, New Jersey, 1972.

[7] HUBER, P.J. *Robust statistical procedures*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pennsylvania, 1977.

[8] ISO/CD 5725-5, *Accuracy (trueness and precision) of measurement methods and results – Part 5: Alternative methods for the determination of the precision of a standard measurement method.* (Draft for Public Comment – Available from BSi Standards.)

[9] RIDLER, N.M. and MEDLEY, J.C. Recent improvements to the traceability of APC-7 impedance measurements at lower RF in the UK", to be published, seventh British Electromagnetic Measurements Conference (BEMC'95) Digest, November 1995.

[10] GROSS, A.M. Confidence interval robustness with long-tailed symmetric distributions. *J. Amer. Statist. Assoc.*, 1976, **71**, (354), 409-416

[11] ISO. *Guide to the expression of uncertainty in measurement*, section 3.4.7, 1st Ed., 1993. (BSI equivalent publication: PD 6461: Part 3: 1995.)

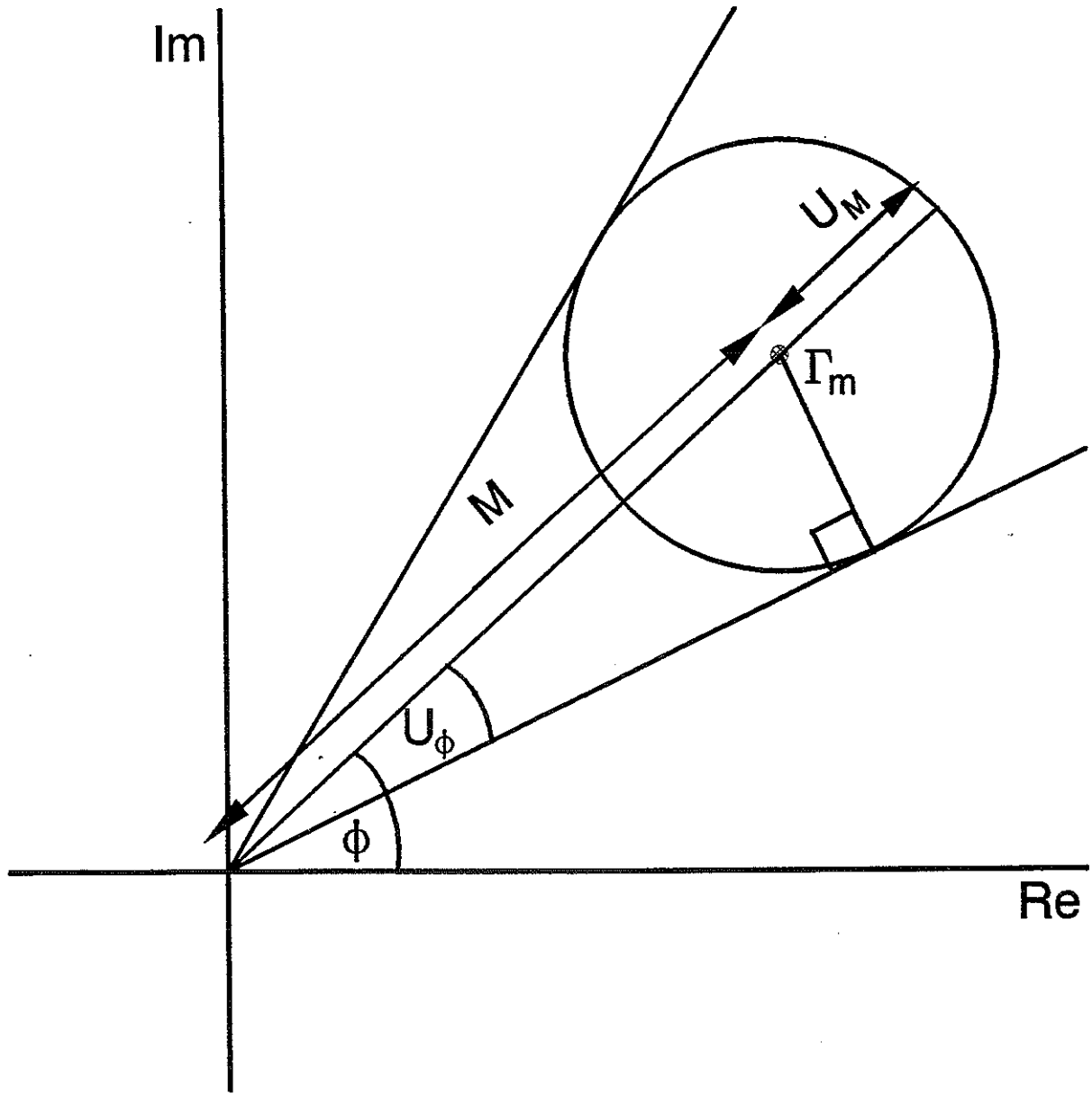

Figure 1: Diagram showing how a circular measure of dispersion in the complex plane can be converted to the equivalent dispersion in the magnitude and phase components.

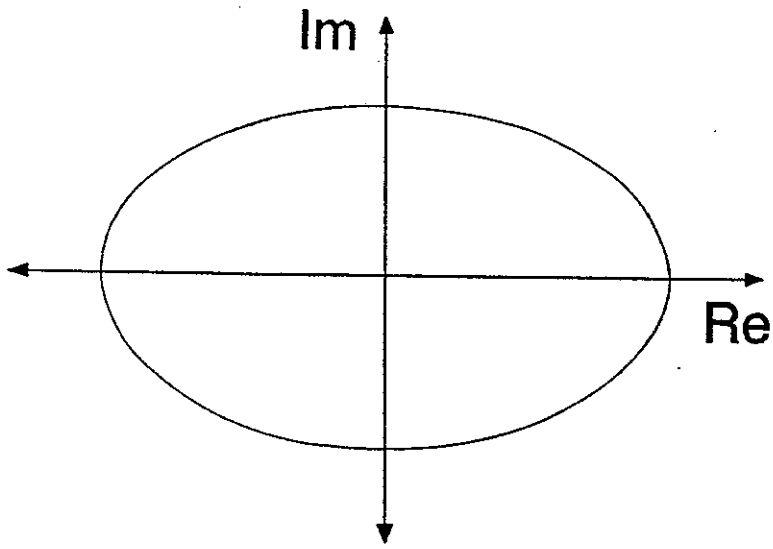

Figure 2: Elliptical measure of dispersion where the two components of the measurement are uncorrelated.

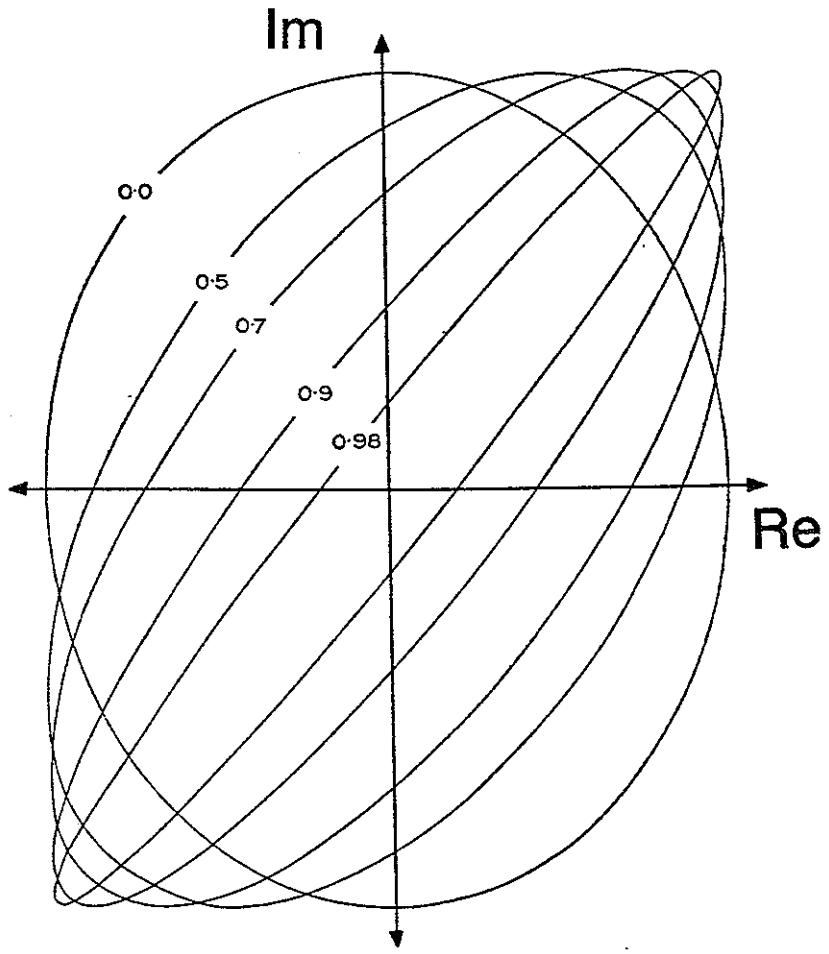

Figure 3: Elliptical measure of dispersion for correlation coefficient values of 0.0, 0.5, 0.7, 0.9 and 0.98.