CAMBRIDGE

Cambridge University Press
978-1-107-02976-7 — Quantitative Methods of Data Analysis for the Physical Sciences and Engineering
Douglas G. Martinson
Excerpt
More Information

Part I

Fundamentals

Analysis of data requires the use of a broad range of techniques spanning a wide range of complexities. Common to most of these techniques is a fundamental foundation consisting of a nomenclature (not always consistent from one author to the next) as well as a set of mathematical and statistical tools. The purpose of Part I is to define that nomenclature and those basic tools.

For this Part, the *order* in which the observations occur is *not* important. This is in contrast to **sequential data**, which are generically referred to as **time series** (whether they vary with time or not), the subject of Part II. For the latter, the order in which the observations occur *is* important.

Part I is dominated by techniques of classical statistics, such as regression analysis, though some newer techniques such as nonparametric and resampling (bootstrap) statistics represent valuable additions to that traditional arsenal. While the tools of statistics are extensive and span a broad range of approaches, the concepts of *Expectation* and *Maximum Likelihood Estimators* are particularly useful in data analysis and will be stressed throughout. Because resampling statistics offers a significant increase in our processing capabilities, it too will be presented for general analyses.

© in this web service Cambridge University Press www.cambridge.org

Cambridge University Press
978-1-107-02976-7 — Quantitative Methods of Data Analysis for the Physical Sciences and Engineering
Douglas G. Martinson
Excerpt
More Information

1 The Nature of Data and Analysis

1.1 Analysis

The Random House College Dictionary defines **analysis** as “the separation of any material or abstract entity into its constituent elements.” For our analysis to be meaningful, it is implicit that the data being analyzed contain some “signal” representing the phenomenon of interest (or some aspect of it). Satisfying this, we might attempt to separate the signal from the noise present in the data. Then we can characterize the signal in terms of its robust features and, in the case of complex phenomena, separate the signal even farther into constituents, each of which may afford additional insights regarding the character, behavior or makeup of multiple processes contributing to our single phenomenon.

Mathematically, we often desire to rewrite a data set, y_i, as

$$y_i = a_1\varphi_{1i} + a_2\varphi_{2i} + \ldots \quad + a_n\varphi_{ni}, \tag{1.1}$$

where the constituents of the data are now described by functions or vectors, φ_{ki}. These constituents (or some subset) with the appropriate weights a_k can then be recombined to **synthesize** (reconstruct) the original data y_i (or just the signal portion of it). Typically, the fewer constituent terms you can use to describe the greatest amount of the data is better (presuming that the signal is contained in a few, hopefully understandable, constituents).

Equation (1.1) represents a simple linear foundation from which a large number of the techniques and analysis tools developed in this text will build on.

Caution: Most analysis techniques will produce something satisfying that technique’s definition of signal, even when performed on pure noise, so be aware that the analysis result may actually be nothing more than a statistical construct. A proper interpretation of your analysis is possible when multiple pieces of evidence support or refute a hypothesized answer to the question being addressed.

1.2 Data Nomenclature

Data (plural) represent measurements of quantities or of variables (a single data point is a **datum**). The variables being measured are classified as discrete or continuous.

Discrete variables are those having discontinuous or individually distinct possible outcomes. Examples include

© in this web service Cambridge University Press

CAMBRIDGE

Cambridge University Press
978-1-107-02976-7 — Quantitative Methods of Data Analysis for the Physical Sciences and Engineering
Douglas G. Martinson
Excerpt
More Information

1) flipping of a coin or rolling of dice
2) counts of individual items or groups of items
3) categorization or classification of measurements

Continuous variables are those having an uninterrupted range of possible outcomes (i.e., with no breaks). Examples include

1) concentrations of a quantity
2) percentage of an item (such data, forced to a constant sum, sometime require special care and attention)
3) magnitude, such as length, mass, etc.

The data (i.e., the measurements of the variables) are also classified according to how they are recorded:

Analog data are those which have been recorded "continuously," such as by a strip recorder (though, technically, even this is not purely continuous, given a non-instantaneous response time of the recorder).

Discrete (or digital) data are those that have been recorded at discrete intervals. All data, when represented on digital computers, are discrete.

In either case, the data must be discretized before they are analyzed in any computational manner that we will consider.

Regardless of how the data are recorded, a sequential series of data are classified as time series.

A **sequential data series** consists of measurements of a quantity that vary as a function of time *or space*, and *the order in which the measurements occur is important*. In this case, the variable being measured is typically referred to as the **dependent variable**, while the time or space variable is the **independent variable**.

Sequential series are commonly referred to by other names such as **time series, traces, records, data series, spatial series**, etc., though "time series" is the most common. The independent variable need not be restricted to time or space. In many instances, it is desirable to measure a quantity as it varies with some other variable whose order of occurrence is important. Regardless, as long as the order of occurrence of the measurements is important, the name "time series" still is commonly applied.

Time series can represent measurements of either discrete or continuous variables and are recorded in either analog or discrete fashion. However, since time and space vary continuously themselves, the discrete or continuous variables being measured often vary continuously as a function of the independent variables. Therefore, discretization of time series data may involve both the dependent and independent variables.

Some authors distinguish between digital and discrete time series as follows. **Discrete series** are sequential series that are discrete in the independent variable but continuous in the dependent variable. **Digital series**, on the other hand, are series that are discrete in both the dependent and independent variables. I will make no such distinction here.

Multidimensional data are those in which the *dependent variable* varies as a function of two or more independent variables simultaneously. For example, weather (measured by a quantity such as temperature) or seismic activity, both of which vary in space and time.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-02976-7 — Quantitative Methods of Data Analysis for the Physical Sciences and Engineering
Douglas G. Martinson
Excerpt
More Information

Multivariate data are those in which there are *multiple dependent variables* varying as a function of a single independent variable. Or, if they are not time series or sequential data, then they simply represent a data set that includes two or more dependent variables. An additional, slightly more restrictive criterion is added to this term as used in probability and statistical applications (Chapter 2).

Real versus Complex Data. "Real" data are what we deal with in the real world, but treating them mathematically as complex numbers (described in more detail in later chapters) affords us the ability to conveniently consider rotation (or phase) of a quantity, as well as offering several additional mathematical conveniences. Therefore, real data will sometimes be organized as if they are complex quantities. Contrary to the name, complex quantities are often considerably easier to deal with than real ones.

Data are further classified according to statistical considerations (e.g., samples, realizations, etc.). These are presented in the next chapter (Probability Theory). Since the analyses presented in this text involve computer manipulation, all data are considered to be discrete.

1.3 Representing Discrete Data and Functions as Vectors

It is helpful to use the most convenient and standard form to represent data. This involves the concept of vectors and matrices. A more detailed summary of the matrix techniques utilized in this text is presented in Appendix 1. Here, only the concept of how one stores discrete data and mathematical functions in vectors is presented.

Typically, data are stored in tables (matrices). For example, if one measures the temperature at noontime on each of *m* days, at *n* different locations, the data are stored in a table as shown in Table 1.1:

Table 1.1

	Site 1	Site 2	Site 3	⋯	Site *n*
Day 1	20.1	23.2	24.8		23.6
Day 2	23.2	23.0	23.6		19.8
Day 3	24.8	23.6	24.2		20.5
⋮					
Day *m*	23.6	19.8	21.9		19.4

Alternatively, you can store temperatures in columns and different locations in rows (*preferred*) as shown in Table 1.2:

Table 1.2

	Day 1	Day 2	Day 3	⋯	Day *m*
Site 1	20.1	23.2	24.8		23.6
Site 2	23.2	23.0	23.6		19.8
Site 3	24.8	23.6	24.2		21.9
⋮					
Site *n*	23.6	19.8	20.5		19.4

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-02976-7 — Quantitative Methods of Data Analysis for the Physical Sciences and Engineering
Douglas G. Martinson
Excerpt
More Information

Which of the two storage schemes you use is a matter of personal taste, but you must pay close attention to the storage form when performing the mathematical manipulations so that the appropriate values are being manipulated as required. For consistency between matrix and vector operations, the form of Table 1.2 is the form used throughout this text.

Initially, however, we will deal predominantly with single-column vectors of data. In the above examples, this is equivalent to having the temperature measured each noontime at one site only (e.g., the first column of Table 1.1, or the first row in Table 1.2), or the noontime temperature of one particular day at n different sites (e.g., the first column of Table 1.2).

Data storage in organized rows and columns is precisely the method used in a matrix. Indeed, each column of numbers represents a column vector. For simplicity, we will assume all vectors are column vectors and thus will drop the descriptor "column" (row vectors are indicated as the *transpose* of a column vector).

In addition to storing discrete observations or data values in table (matrix) form, the storage of mathematical functions, conveniently expressed as formulas when dealing with continuous data, must be presented at discrete values to represent them in vectors, hence the indexing of the constituent terms, φ_{kj} in equation (1.1), where the k represents *which* constituent term (function or vector), and j, the jth discrete value of the term.

1.4 Data Limits

1.4.1 Domain

Domain represents the spread or extent of the *independent variable* over which the quantity being measured varies. It is usually given as the maximum value of the independent variable minus the minimum value. Because no phenomenon is observed over all time or over all space, data have a limited domain (though the domain may be complete relative to the phenomenon of interest, e.g., the finite Earth's surface).

1.4.2 Range

Range represents the spread or extent over which the *dependent variable* (i.e., the quantity being measured, possibly as a function of time or space) can take on values. You will typically present range as the maximum value of the dependent variable minus the minimum value. No measuring technique can record (or transmit) values that are arbitrarily large or small. The lower limit on very small quantities is often set by the noise level of the measuring instrument.

Dynamic Range (**DR**) is the actual range over which dependent variables are measured or reproduced. Often this is less than the true range of the variable. You present dynamic range on a logarithmic scale in decibels (dB):[1]

[1] You can use some other form to express this if you are uncomfortable with decibels; just make it clear what your form is.

CAMBRIDGE
Cambridge University Press
978-1-107-02976-7 — Quantitative Methods of Data Analysis for the Physical Sciences and Engineering
Douglas G. Martinson
Excerpt
More Information

$$\mathrm{DR} = 10 * \log_{10}\left(\frac{\mathit{largest\ power}}{\mathit{smallest\ (nonzero)\ power}}\right). \tag{1.2}$$

or

$$\mathrm{DR} = 20 * \log_{10}\left(\frac{|\mathit{largest\ value}|}{|\mathit{smallest\ (nonzero)\ value}|}\right) \tag{1.3}$$

The first formula (1.2) is used if the data represent a measure of power (a squared quantity such as variance or square of the signal amplitude). Otherwise the second formula (1.3) is used. Use the smallest nonzero value for measurement devices that report zero values.

Since power is a quantity squared, the first formula (1.2) is related to the second (1.3) by

$$\mathrm{DR} = 10 * \log_{10}\left[\left(\frac{|\mathit{largest\ value}|}{|\mathit{smallest\ (nonzero)\ value}|}\right)^2\right]. \tag{1.4}$$

Therefore, the two formulas yield the same answer, given the appropriate input. This is especially useful for instruments that return measurements proportional to variance or power.

An increment of 10 in DR equates to a factor of 10 in R_p (the power ratio).

$$\mathrm{DR} = 10 * \log_{10}(\mathrm{R}_p), \tag{1.5}$$

so

$$\mathrm{R}_p = 10^{(\mathrm{DR}/10)}. \tag{1.6}$$

E.g., DR = 30 so R_p = 1,000; DR = 40, so R_p = 10,000.

Finally, because of the limited range and domain of data, any data set, say $y(x)$, are constrained by

$$X_S < x < X_L \tag{1.7}$$

and

$$|y(x)| < M, \tag{1.8}$$

where X_S, X_L and M are finite constants. *Such functions are manageable and can always be integrated.* This seemingly esoteric fact proves extremely useful in the practical analysis of data.

1.4.3 Frequency

Most methods of data measurement cannot respond instantly to sudden change. The resulting data are thus said to be **band limited**. That is, they will not contain frequency information higher than that representing the fastest response of the recording

© in this web service Cambridge University Press www.cambridge.org

Cambridge University Press
978-1-107-02976-7 — Quantitative Methods of Data Analysis for the Physical Sciences and Engineering
Douglas G. Martinson
Excerpt
More Information

device, this is an invaluable constraint for some important analysis techniques, though it can be severely limiting regarding the study of certain high-frequency (rapidly varying) phenomena.

1.5 Data Errors

Data are never perfect. Errors can enter data through experimental design, measurement and collection techniques, assumptions concerning the nature and fidelity of the data, discretization and computational or analysis procedures. In analyzing and interpreting data it is important to attempt to estimate all of the potential errors (either quantitatively or qualitatively). That is, it is important to estimate the uncertainty contained in the measurements and their influence on your interpretation. Too often, errors related to one easily determined component are presented while others are completely ignored. You needn't be fanatic in your attempt to estimate the uncertainty – *scientific progress needn't be held hostage to unreasonable quantification* – rather, it is important to **make an honest assessment to the best of your ability to *estimate* the uncertainties associated with your measurements.** If you can't formally estimate an error, simply say so, then make your best educated guess or give a range for the error.

Box 1.1 Errors: Give Them Their Due

It is not uncommon to present, as the only error in the data, the scatter measured when making replicate measurements from a specific measurement sample (e.g., weighing a sample 100 times). While this is certainly a good estimate for the instrument error (sometimes called the "analytic error," "measurement precision" or "instrument precision"), it does not preclude the presence of a variety of other types of error that are likely present in the data. That is, how much scatter would occur if replicate samples were obtained (not just replicate measurements of the same sample)? Is there systematic bias in the instrument making the measurements? How representative is the sample of the process that you think you're sampling? (This is sometimes a dominant source of error that is completely overlooked.)

It is also important to provide an explanation as to how the estimate of the uncertainty was arrived at so the reader can make their own assessment of the techniques employed.

1.5.1 Instrument Error

Errors (uncertainties) in data are classified according to their source. Under the best circumstances, the quality of the data is predominantly controlled by the capabilities of the recording device. Measurement capabilities are classified according to:

1) **Precision** specifies how well a specific measurement of the same sample can be replicated. In statistical terms, the precision is a measure of the variance (or standard deviation) of the sample.

© in this web service Cambridge University Press

CAMBRIDGE
Cambridge University Press
978-1-107-02976-7 — Quantitative Methods of Data Analysis for the Physical Sciences and Engineering
Douglas G. Martinson
Excerpt
More Information

For example, if a substance was repeatedly weighed 100 times, giving a mean weight of 100 kg but with a scatter about this mean of 0.1 kg, then 0.1 kg would represent the precision of the measurement.

2) **Accuracy** specifies how well a specific measurement actually represents the true value of the measured quantity (often considered in terms of, say, a long-term instrument drift). In statistical terms, accuracy is often reported in terms of **bias**. For example, if a scale repeatedly returns a weight of ~100.0 kg for a substance, but its true weight is 105.3 kg – the mismatch between the measured value and true value reflects the bias of the measurement. So the scale is good to an accuracy of just over 5 kg, or the scale has a bias of ~5 kg.
3) **Resolution** specifies the size of a discrete measurement interval of the recording instrument used in the discretization process. In other words, it indicates how well the instrument (or digitized data) can *resolve* changes in the quantity being measured. For example, if a thermometer only registers changes in temperature of 0.01° C, then it cannot distinguish changes in temperature smaller than this resolution. One would achieve the same resolution if, in the process of digitizing higher-resolution data, all values were rounded off to the nearest 0.01° C.
4) **Response time** specifies how quickly an instrument can respond to a change in the quantity being measured. This will limit the bandwidth (range of frequencies) of a measured time series (discussed in more detail latter).

In general, accuracy *reflects the degree of systematic errors, while* precision *reflects the degree of random errors*. Statistics are well designed to treat the latter (precision), whereas they are not generally designed to address the former (accuracy). Accuracy must be estimated, using whatever means are practical and reasonable, by the person who understands the instrument.

1.5.2 Experimental/Observational Error

The experimental design, sampling program or observational methods may also lead to errors in precision and accuracy.

Precision. Consider estimating the precision of a specific brand of thermometer. If 100 of the thermometers were simultaneously used to measure the temperature of a well-mixed bath of water, the scatter about the mean temperature might typically be presented as the precision of the thermometers. However, this is really the precision of the estimated temperature and it reflects the precision of the experimental design. Any one particular thermometer may have significantly better or worse precision than that suggested by the scatter achieved between 100 different thermometers. Also, the "well-mixed" bath may actually contain temperature gradients to some extent, which will also influence the scatter observed in the measurements. Repeating the above calibration, only this time making 100 replicate measurements using a single thermometer, may include some scatter due to subtle changes in the water bath between replicate measurements. Thus, even that measured precision reflects some combination of instrument precision and experimental scatter.

© in this web service Cambridge University Press www.cambridge.org

Cambridge University Press
978-1-107-02976-7 — Quantitative Methods of Data Analysis for the Physical Sciences and Engineering
Douglas G. Martinson
Excerpt
More Information

In this respect, precision errors may well be attributable to a combination of both instrument and experimental error. This combination is responsible for the observed random scatter in replicate measurements, which is often referred to as **noise** in data. Noise can also represent any portion of the data that does not conform with preconceived ideas concerning the nature of the data – recall the expression that "one person's noise is another person's signal."

One of the goals in data analysis is to detect signal in noise or reduce the degree of noise contamination. Noise is sometimes classified according to its contribution relative to some more-stable (or non-fluctuating) component of the observations, referred to as the **signal**.

Signal-to-Noise Ratio (SN) is the common measure for comparing the ratio of signal to noise in a data set. As with dynamic range, you give this ratio on a logarithmic scale in decibels (dB):

$$\mathrm{SN} = 10 * \log_{10}\left(\frac{power\ of\ signal}{power\ of\ noise}\right) \tag{1.9}$$

or

$$\mathrm{SN} = 20 * \log_{10}\left(\frac{|amplitude\ of\ signal|}{|amplitude\ of\ noise|}\right). \tag{1.10}$$

Exactly how one determines the values to insert in the above formulas depends on the data and how they were collected. In some cases, it is appropriate to use the mean value of the data (or measured range, for time-series data) as the amplitude of signal and the (known) instrument error (or precision) as the amplitude of the noise. With time series, the noise may be alternatively estimated by computing the measured scatter in a series of replicate time-series measurements of the same quantity (under the same sampling conditions). The amplitude of signal might then be regarded as the observed range in the average of the replicate time series.

The signal-to-noise ratio is also given as the ratio of the variance (a power) of the signal to the variance of the noise, or it can be written in any other manner that essentially provides some ratio between the variance of the signal and that of the noise.

If one uses the mean of the data as the signal and the standard deviation as the noise, then it is often convenient to present this form of SN (actually, SN^{-1}) as a **coefficient of variation**:[2]

$$V = \frac{standard\ deviation\ of\ data}{mean\ of\ data}. \tag{1.11}$$

V can be presented as percentage or in dB, but its interpretation is conveniently intuitive. That is, in the vicinity of $V = 1$, it is a suggestion that the scatter (noise) in the data is comparable in size to the signal itself. At 200 percent (~3 dB), the scatter is twice the size of the signal. At 25 percent (~ −6 dB), the noise is only one-fourth of the signal amplitude.

[2] It is common to present many statistical quantities ("moments") as coefficients of the moment. For moment μ^k, a coefficient of the moment is given as μ^k/μ^{k-1} (as is the case for V here). This will make more sense after the discussion of moments in Chapter 2, "Probability Theory."

© in this web service Cambridge University Press www.cambridge.org