

UC Santa Barbara

UC Santa Barbara Previously Published Works

Title

A review of dimensional assessment for statistics and probability

Permalink

<https://escholarship.org/uc/item/236782qs>

Journal

Journal of Statistical Theory and Practice, 10(4)

ISSN

1559-8608 1559-8616

Author

Hughes, Gary B

Publication Date

2016-07-25

DOI

10.1080/15598608.2016.1214852

Peer reviewed

A review of dimensional assessment for statistics and probability

Gary B. Hughes

Statistics Department, California Polytechnic State University, San Luis Obispo, California, USA

ABSTRACT

Numeric values that are meant to describe physical attributes are given a precise context by association with measurement units. Units define the quantification of the attribute's dimension within a specific measurement system. Science and engineering disciplines have long relied on measurement systems and dimensional reasoning. This article explores the (long-held) notion that dimensional awareness can also stimulate unique insight into statistics and probability concepts. Linear algebra developed for dimensioned physical quantities provides a rigorous framework for analyzing applied equations. Units associated with descriptive statistics can be determined by dimensional assessment of basic definitions. Dimensioned linear algebra can be used to trace units through more advanced statistical constructs. Any dimensional context precludes a non-dimensionless probability unit, and requires probability density to have a dimension that is reciprocal to that of the random variable. The physical meaning of distribution parameters becomes clear in a dimensional context, and several pervasive inconsistencies in graphics and basic equations are clarified. The ideas presented in this article are not new. Dimensional awareness also contributes to a broader system-level foundation for utilizing statistics and probability concepts as tools in applied scenarios.

ARTICLE HISTORY

Received 7 June 2015

Accepted 16 July 2016

KEYWORDS

Dimensional homogeneity; dimensioned linear algebra; unit cancelation; dimensions of statistics; dimensions of probability density

AMS SUBJECT CLASSIFICATION

00A73

1. Introduction

An oft-cited example of the importance of considering dimension in science and engineering is NASA's Mars Climate Orbiter (MCO), which was designed to be a Martian weather satellite. It was launched in December 1998; mission controllers lost contact with MCO on September 23, 1999, during the Mars orbit insertion maneuver. NASA convened an interdisciplinary Mishap Investigation Board (MIB) tasked with determining the spacecraft's fate. Conclusions of the MCO MIB are described in the official report (Stevenson et al. 1999, 6):

The MCO MIB has determined that the root cause for the loss of the MCO spacecraft was the failure to use metric units in the coding of a ground software file, "Small Forces," used in trajectory models. Specifically, thruster performance data in English units instead of metric units was used in the software application code titled SM_FORCES (small forces).

Thrust values were calculated in English units, but were expected by the receiving module to be in metric units, so the amount of thrust generated did not accomplish the

envisioned trajectory modifications. The spacecraft likely passed too close to the planet's surface, where insurmountable interaction with the Martian atmosphere likely occurred, disintegrating the spacecraft.

Conclusions from the MCO MIB are persistently trumpeted in engineering classrooms to stress the importance of attention to measurement units. Thrust calculations were performed correctly, but a fault occurred when numeric values were passed from one subsystem to another, and each subsystem interpreted the values in a different way. The error might have been avoided with heightened system-level thinking, which advocates that individual subsystems are better understood through the perspective of interactions with each other, rather than in isolation. Numeric values alone, that is, without consideration for appropriate measurement units, are insufficient to describe physical entities or phenomena, echoing a long-standing doctrine of the broader discipline of classical dimensional analysis (CDA) (Huntley 1967, chap. 1; Hart 1995, chap. 1.1).

Scientists and engineers have long utilized dimensional concepts, such as CDA and the “algebra of quantities” (i.e., “unit cancelation”). But with the exception of a small number of statistical studies incorporating CDA, dimensional concepts are conspicuously scarce in the statistics and probability literature. This article explores the notion that fundamental statistics and probability concepts can be viewed within the context of dimensional reasoning. Incorporating dimensional concepts in statistics and probability is not a new idea, having been explicitly discussed in previous works such as Hand (1996) and Finney (1977), and embraced by Davis (2013). Direct gains of dimensional assessment are realized by clarification of several pervasive inconsistencies in graphics and basic equations, and by revealing common misconceptions. Viewing fundamental concepts within a dimensional framework also yields enhanced understanding of values, variables, and parameters that pervade statistics and probability concepts. Collateral benefits include improved synergy with applied disciplines.

2. Salient dimensional background from physical sciences

2.1. Dimensioned physical quantities

Applied statistics is a science of data. Data are acquired by many processes, including measurements or assessments of physical entities or phenomena. For numeric variables, the act of measuring produces dimensioned physical quantities. In an applied sense, the number 3 has a less distinct meaning than the quantity 3 m, which conveys a precise and universally known quantity of length. Dimension refers to the category or type of a physical quantity, such as length or mass. Dimensions are quantified within a system of measurement, the most common being the metric system, officially *Le Système International d'Unités* (the International System of Units), abbreviated SI (*Bureau International des Poids et Mesures* 2006, hereafter SI brochure). SI is based upon seven fundamental dimensions (also called primary or base dimensions) and associated measurement units: length (m), mass (kg), time (s), electric current (A), thermodynamic temperature (K), amount of substance (mol), and luminous intensity (cd). The fundamental dimensions form a basis that can be used to express any other physical dimension.

With a suitable measurement system, CDA is used to develop dynamical laws by analyzing the fundamental dimensions of each system component, and without requiring

knowledge of underlying physical principles. Early formalism for CDA was provided by Bridgman (1922), a thorough history is provided by Macagno (1971), and a comprehensive modern account can be found in Szirtes (2007). Only sparsely have statisticians employed CDA. Rutherford (1975) reduced the number of factors in designed experiments using CDA. Finney (1977) performed consistency checks for statistical algebra. Vignaux and Scott (1999) analyzed the development of regression models, and concluded that CDA can facilitate discovery of an optimal regression equation.

More recently, Davis (2011) sought to show that “dimensional analysis should take its place at the heart of experimental design in engineering applications.” Albrecht et al. (2013) described advantages associated with CDA relative to standard approaches for experimental design, including (a) a priori dimension reduction and (b) scalability of results. Lin and Shen (2013) expressed some concerns about the general applicability of techniques described by Albrecht et al.. Shen et al. (2013) described several benefits of CDA for experimental design, echoing the advantages of dimension reduction and scalability that were presented by Albrecht et al., and further noting that CDA produces “analytical insights into the relations among variables that it generates.”

2.2. Assessing dimensions of applied equations

Previously cited works indicate that CDA can make worthwhile contributions to experimental design, and perhaps to other statistical subfields. But exploring the use of CDA in statistics is not among the objectives of this article. Rather, the aim here is to explore how accounting for data dimensions by analyzing measurement units can enhance understanding of fundamental statistics and probability concepts. One aspect of CDA plays a central role: *Dimensional homogeneity* is the requirement that an equation have equivalent physical dimensions on both sides. Dimensional homogeneity is ensured whenever measurement units on each side of the equation are “algebraically reducible” to the same form.

Algebraic methods that can deal with value–unit pairs have a long history. Axiomatic approaches for algebra with dimensioned (scalar) quantities have been developed and refined, for example, in works by Drobot (1953), Brand (1957), Thun (1960), Parkinson (1964), Kurth (1965), Quade (1967), Page (1967), Whitney (1968a; 1968b), Luce (1971), Szekeres (1978), Carlson (1979), and Luce et al. (1990). Methods are also available for dealing with equations that govern physical systems, including operations with vectors and matrices whose elements are dimensioned physical quantities. General approaches for working with dimensioned vectors and matrices could be described as dimensioned linear algebra (DLA). A comprehensive axiomatic development and extensive DLA were described by Hart (1995), including vector and matrix operations where elements are dimensioned values. Palanthandalam-Madapusi et al. (2007) revisited some of Hart’s results, including a self-contained presentation of the “dimensional structure of a matrix under which standard operations involving the inverse, powers, exponential, and eigenvalues are valid.” Palanthandalam-Madapusi et al. also proposed the concept of different types of dimensionless quantities, which could be defined in terms of the constituent physical dimensions that comprise a ratio.

Dimensional homogeneity and DLA provide a rigorous framework for dimensional assessment of applied equations. Definitions of summary statistics are applied equations,

and so can be assessed dimensionally. Subsequent sections of this article describe dimensional assessment of statistics and probability concepts and identify specific areas where dimensional assessment provides enhanced insight.

Axiomatic developments for operations with dimensioned quantities rely on a dimensional basis. In the typical sense of a basis for a vector space, dimensional basis elements are “independent” and the basis “spans” the desired space of physical dimensions. Notation varies somewhat in the DLA literature; in Hart’s approach, a dimensioned physical quantity is defined as an ordered pair consisting of a numeric value, x_v , and a vector of numeric powers $[e_1, e_2, \dots, e_n]$ corresponding to the elements of a dimensional basis $[B_1, B_2, \dots, B_n]$. A dimensioned physical scalar $x = (x_v, [e_1, e_2, \dots, e_n])$ is then written as a product:

$$x = x_v \cdot \prod_{i=1}^n [B_i]^{e_i} \quad (1)$$

Any physical quantity in the span of the basis can be represented in terms of a magnitude and a vector of basis-element powers. The magnitude of a dimensioned physical quantity depends on which measurement system is used, as well as the chosen basis within the measurement system.

The seven SI base units provide a basis that spans the space of physical dimensions (although more parsimonious bases could be defined; see discussions in Hart ([1995, 37, footnote 21] and Szirtes [2007, 37]). A natural SI dimensional basis is formed by the vector

$$B_{SI} = [1 \text{ m}, 1 \text{ kg}, 1 \text{ s}, 1 \text{ A}, 1 \text{ K}, 1 \text{ mol}, 1 \text{ cd}]$$

Using the natural SI basis B_{SI} , and referring to the construction of derived units as described by Thompson and Taylor (2008, back cover), the physical quantity 5 J (5 joules) is denoted by the ordered pair and product:

$$\begin{aligned} 5 \text{ J} &= 5 \text{ kg m}^2/\text{s}^2 \\ &= (5, [2, 1, -2, 0, 0, 0, 0]) \\ &= 5 \cdot (1 \text{ m})^2 \cdot (1 \text{ kg})^1 \cdot (1 \text{ s})^{-2} \cdot (1 \text{ A})^0 \cdot (1 \text{ K})^0 \cdot (1 \text{ mol})^0 \cdot (1 \text{ cd})^0 \end{aligned}$$

2.3. Algebra of quantities

Essential elements of DLA can be distilled to form a system of definitions and rules that are sufficient for analyzing applied equations; for examples, Fröhlich (2010) developed an algebra for a specific application in economics. In an appropriate algebra, operations should be defined in such a way as to replicate the essential mathematics of dimensioned quantities that describe physical systems. The sum $x_1 + x_2$ is only defined when both physical quantities have the same dimensions; otherwise, addition is not defined. The algebra must be closed under multiplication. For dimensioned scalars, exponents and arguments of transcendental functions must be dimensionless.

The system of definitions and rules presented in this article expresses the familiar approach of “unit cancelation” for dimensioned physical scalars, in a formulation suitable for statistics and probability. These definitions and rules emphasize the notion of dimensional assessment; this is not to suggest that the expanded notation be adopted whenever

dimensioned physical quantities are written. Insights gained from dimensional assessment with the expanded notation will also be characteristics of standard notational constructions without including dimensional symbols.

Definition 2.3.1. *A quantified dimensioned physical quantity x is the product of a numerical value x_v and a quantified dimension $||x||_{M_B}$ within the measurement system M using the dimensional basis B , called “measurement units”:*

$$x = x_v \cdot ||x||_{M_B} \quad \text{examples : } 3 \text{ m, } 5 \text{ J, } 2 \mu\text{F}$$

Standard notation uses $[x]$ to indicate the dimension of x . In this article, the standard notation is extended to $||x||_{M_B}$ which indicates that the dimension has been quantified within a suitable measurement system M , such as SI, with a specified basis B , such as B_{SI} .

Definition 2.3.2. *Dimensionless physical quantities have dimension $[1]$, and associated measurement units of $||1|| = 1$.*

Values that are formed by ratios of two physical quantities of the same dimension are dimensionless. Radian measure is defined as the ratio of arc length s and radius r , which both have dimensions of length, hence radian measure is dimensionless. The scheme of Palanthandalam-Madapusi et al. classifies radian measure as a “lengthian” dimensionless type, which they denote $[1]_m$.

Definition 2.3.3. *Counts are dimensionless.*

From the SI brochure (section 2.2.3):

Another class of dimensionless quantities are numbers that represent a count, such as a number of molecules, degeneracy (number of energy levels), and partition function in statistical thermodynamics (number of thermally accessible states). All of these counting quantities are also described as being dimensionless, or of dimension one, and are taken to have the SI unit one.

Rule 1. Addition and subtraction are defined only for physical quantities with the same measurement units.

Adding dimensioned quantities with the same measurement units produces a sum with the same units, for example, $3 \text{ cm} + 2 \text{ cm} = 5 \text{ cm}$. Adding quantities with different dimensions is incommensurable: It is not meaningful to add $5 \text{ m} + 3 \text{ kg}$. It is possible to interpret a sum with consistent dimension, but inconsistent units: $3 \text{ m} + 2 \text{ cm} = 3.02 \text{ m}$. In such cases, the numerical operation of addition can only be performed when all values are quantified using the same measurement system basis. This rule illustrates the intent behind the (unwieldy) notation $||x||_{M_B}$, which indicates that units are a combination of a measurement system, in addition to a dimensional basis within the measurement system. The notation is not suggested here for general use, but rather is offered to concisely convey the (common sense) concept of quantified dimensioned physical quantities in a

notational form. The rule applies to dimensionless types, since it is also unclear what the result would be of adding 2 radians to Mach 2.

Rule 2. Multiplication and division are defined for physical quantities with any measurement units, except when dividing by the quantity zero.

From the SI brochure (section 5.3.1):

Symbols for units are treated as mathematical entities. In expressing the value of a quantity as the product of a numerical value and a unit, both the numerical value and the unit may be treated by the ordinary rules of algebra. This procedure is described as the use of quantity calculus, or the algebra of quantities.

The normal rules of algebraic multiplication or division can be used with unit symbols, since the dimensional component of the product in Eq. (1) can be treated as an algebraic variable.

Rule 3. Scalar exponents and scalar arguments of transcendental functions must be dimensionless quantities.

Fröhlich (2010) reasoned that Rule 3 follows from Rule 1 by citing the Taylor expansion for an exponential:

$$e^x = \sum_{i=1}^{\infty} \frac{x^i}{i!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

If the scalar quantity x were to possess a dimension, the sum would be undefined because every term would have different measurement units, namely, $|[x]|_{M_B}^i$, violating Rule 1. Similar arguments can be made for power, logarithmic, and trigonometric functions with scalar arguments.

Rule 4. Expressions on both sides of a valid equation must be reducible to the same measurement units.

Dimensional homogeneity can be verified using “algebra of quantities” with equations containing dimensioned physical quantities.

3. Dimensional reasoning for statistics

3.1. Descriptive statistics

Dimensional homogeneity (Rule 4) dictates that the notion of equality must include both numerical and dimensional character. Measurement units associated with descriptive statistics can be determined by dimensional assessment, for example,

$$\bar{x}_n = \frac{1}{n|[1]|} \cdot \sum_{i=1}^n (x_{v_i} \cdot |[x]|_{M_B}) \sim \frac{|[x]|_{M_B}}{1} \sim |[x]|_{M_B}$$

The sum is defined under Rule 1 only when all elements of the sum are quantified to the same measurement units. By Definition 2.3.3, multiplying by $1/n$ does not contribute to

the eventual units of the sample mean. Hart uses the notation $x \sim y$ to indicate that the variables x and y have the same dimensional form; that is, for dimensioned scalars, $x \sim y$ if, and only if, $[x] = [y]$. The notation is extended here to quantified dimensions: $x \sim y$ whenever $||x||_{M_B} = ||y||_{M_B}$.

The coefficient of variation is permissible only for ratio-scale measurements. The statistic is used to describe the dispersion of values in a way that does not depend on the unit of measurement. When two variables have the same units, covariance has the same units as variance, providing a familiar gauge for interpretation of the covariance value. In general, however, covariance can be determined for variables that have different units, rendering the scale of covariance less familiar and possibly more difficult to interpret, since in the general case covariance may not be expressed in a familiar unit of measurement. Correlation scales the numerical value of covariance to a more readily interpretable (nondimensional) range, even when the two variables have different units. The coefficient of determination is dimensionless, consistent with the interpretation that it represents the proportion of variability in response that is explained by a model of the independent variable. Order measures, including quantiles, are simply data values (or in some cases, linear combinations of data values), and thus have the same dimension as the data values themselves.

3.2. Histograms

Common graphical displays for univariate data include various types of histograms. For frequency and relative frequency histograms, the statistic for each bin has units of $||1|| = 1$, since only counts are involved. Bin boundaries $b_1, b_2, \dots, b_k$ are locations along the x -axis of the histogram, and therefore possess the same units as the variable being summarized. The statistic in each bin of a frequency density histogram therefore has units $1/||x||_{M_B}$. The area of an individual bin is

$$\begin{aligned} \text{area}_i &= (b_{v_{i+1}} \cdot ||x||_{M_B} - b_{v_i} \cdot ||x||_{M_B}) \cdot \left(\text{density}_i \cdot \frac{1}{||x||_{M_B}} \right) \\ &\sim ||x||_{M_B} \cdot \frac{1}{||x||_{M_B}} \sim 1 \end{aligned}$$

Total area under a frequency density histogram is calculated by summing the areas for all individual bins, evaluating to a dimensionless sum to coincide with the meaning of the area, which is a proportion. The area of each bin is only dimensionless when the bin density value has units of $1/||x||_{M_B}$.

Histograms are graphical data summaries. Among the goals of graphical excellence are clarity, precision, and efficiency in the presentation of statistical information (Cleveland 1985; Tufte 2001; Levine et al. 2001). Precision should require that histogram axis labels display the measurement unit that applies to the numeric values in the graph. For frequency and relative frequency histograms, the vertical axis represents dimensionless quantities, so no unit label is needed. However, since frequency density has units $1/||x||_{M_B}$, axes of such histograms should be labeled with appropriate measurement unit descriptions. For example, if data values represent length in meters, the y -axis of a frequency density histogram should be labeled with, for example, “frequency density (m^{-1}).”

3.3. Scatter plots and regression lines

A set of ordered pairs will in general have different units for the independent and dependent variables. The regression line equation must have the same units on both sides (Rule 4). Requiring the sum to be performed between values with the same units (Rule 1) dictates that the regression line parameters must possess units:

$$y_v \cdot |[y]|_{M_B} = \beta_{0_v} \cdot |[y]|_{M_B} + \beta_{1_v} \cdot \frac{|[y]|_{M_B}}{|[x]|_{M_B}} \cdot x_v \cdot |[x]|_{M_B}$$

The required units are clearly consistent with the interpretation of the parameters as the slope and intercept of the regression line.

Enforcing dimensional homogeneity in the regression line equation dictates the units of parameters. The same result is obtained by tracking units through the process of estimating regression parameters from a sample, typically accomplished by solving a linear system:

$$\begin{aligned} \begin{bmatrix} \sum_i (x_{v_i} \cdot |[x]|_{M_B})^2 & \sum_i (x_{v_i} \cdot |[x]|_{M_B}) \\ \sum_i (x_{v_i} \cdot |[x]|_{M_B}) & n \cdot |[1]| \end{bmatrix} \cdot \begin{bmatrix} \beta_{1_v} \cdot \frac{|[y]|_{M_B}}{|[x]|_{M_B}} \\ \beta_{0_v} \cdot |[y]|_{M_B} \end{bmatrix} \\ = \begin{bmatrix} \sum_i (x_{v_i} \cdot |[x]|_{M_B}) \cdot (y_{v_i} \cdot |[y]|_{M_B}) \\ \sum_i (y_{v_i} \cdot |[y]|_{M_B}) \end{bmatrix} \end{aligned}$$

The linear system is dimensionally homogeneous whenever the parameter estimates have units that are consistent with the slope and intercept of the regression line.

When the dimensional component of the leading matrix A can be written as the outer product of a constant vector and the first row of A , the numeric and dimensional parts of a matrix can be treated separately in an equation (Hart 1995, 78–83). For the regression line system, the dimensional component of the leading matrix can be written as such an outer product:

$$[A] \sim a \cdot b^{\sim} \quad \text{where : } a = \begin{bmatrix} 1 \\ \frac{1}{|[x]|_{M_B}} \end{bmatrix}, b^{\sim} = \begin{bmatrix} |[x]|_{M_B}^2 & |[x]|_{M_B} \end{bmatrix}, \text{ and } [A^{\sim}]_{ij} = \frac{1}{A_{ji}}$$

It is possible to execute Gaussian elimination for the dimensional component in terms of the typical LDU factorization, and also to determine the dimensional structure of A^{-1} :

$$[A] \sim (a \cdot a^{\sim}) \cdot (a \cdot b^{\sim}) \cdot (b \cdot b^{\sim}), \text{ and } [A^{-1}] \sim b \cdot a^{\sim}$$

For the regression line matrix equation,

$$[A^{-1}] \sim b \cdot a^{\sim} \sim \begin{bmatrix} \frac{1}{|[x]|_{M_B}^2} \\ \frac{1}{|[x]|_{M_B}} \end{bmatrix} \cdot \begin{bmatrix} |[1]| & |[x]|_{M_B} \end{bmatrix} \sim \begin{bmatrix} \frac{1}{|[x]|_{M_B}^2} & \frac{1}{|[x]|_{M_B}} \\ \frac{1}{|[x]|_{M_B}} & |[1]| \end{bmatrix}$$

Multiplying the dimensional matrix $[A^{-1}]$ by the dimensional right-hand side vector gives the expected parameter dimensions:

$$\begin{bmatrix} \frac{1}{|[x]|_{M_B}^2} & \frac{1}{|[x]|_{M_B}} \\ \frac{1}{|[x]|_{M_B}} & |[1]| \end{bmatrix} \cdot \begin{bmatrix} |[x]|_{M_B} \cdot |[y]|_{M_B} \\ |[y]|_{M_B} \end{bmatrix} \sim \begin{bmatrix} \frac{|[y]|_{M_B}}{|[x]|_{M_B}} \\ |[y]|_{M_B} \end{bmatrix}$$

Hart (1995, sections 6.4 and 6.5) describes algorithms for obtaining the dimensions of the solution vector, along with the numerical values. The algorithms are general, and so are expected to be suitable for any linear system that arises in the suite of statistical procedures.

4. Dimensional reasoning for probability

4.1. Probability units

Hart (1995, 168) argues that probability must be a dimensionless quantity, that is, that the idea of a non-dimensionless unit of probability is inconsistent with any rigorous formulation of DLA:

Note that one might wish to define a nondimensionless unit of probability, P , as in Ellis [1965, 150], changing the normalization axiom to

$$\sum_i p_i = 1P$$

but this causes problems. One such problem is that $E[x] \sim [P]x \not\sim x$ unless a corrective constant $1/P$ term is arbitrarily inserted in the definitions of the expectation. In addition, P would need a ‘polarity axiom’ to allow one to meaningfully state that probabilities and probability densities are always non-negative. Thirdly, this would identify a distinguished dimensioned quantity, ‘ $1 P$ ’, in violation of the basis-independence principle.

In addition to the points made by Hart, a non-dimensionless unit of probability would create inconsistencies, for example, with some probability mass functions (PMFs) such as binomial:

$$P(X = k|n, p) = \begin{cases} \binom{n}{k} \cdot p^k \cdot (1-p)^{(n-k)} & 0 \leq k \leq n \\ 0 & \text{otherwise} \end{cases}$$

The standard interpretation is that p represents the probability of success in a Bernoulli trial, and so would carry the putative probability unit. The resulting binomial probability would evaluate to possess the putative probability unit raised to the power n . The inconsistency is insurmountable under dimensional homogeneity (Rule 4). The same issue would occur in several discrete PMFs, including negative binomial and geometric distributions, but not in others, including Bernoulli and hypergeometric distributions.

Still other objections to a non-dimensionless unit of probability are evident. Consider the expression for conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

If each probability were to carry the putative unit, then the equation for conditional probability would not be dimensionally homogeneous. The only consistent remedy for all of the issues raised is that probability is a dimensionless quantity.

Palanthandalam-Madapusi et al. defined different types of dimensionless quantities, in terms of the constituent dimensions that comprise a ratio. It is not meaningful to add a radian measure to a probability, for example, even though both quantities are dimensionless. The classification scheme of Palanthandalam-Madapusi et al. should also include a ratio of dimensionless quantities, including a “probabilistian” category that would imbue concepts like expected value with a probabilistic character that is not evident from expressions under notational paradigms that lack a dimensional component.

4.2. Discrete random variables

4.2.1. Binomial distribution

Discrete PMFs are defined as equations, and each term can be viewed with associated units. For example, the binomial PMF for $0 \leq k \leq n$ is

$$P(X = k|n, p) = \binom{n \cdot |[1]|}{k} \cdot |[1]| \cdot (p \cdot |[1]|)^{k \cdot |[1]|} \cdot (1 \cdot |[1]| - p \cdot |[1]|)^{(n \cdot |[1]| - k \cdot |[1]|)}$$

The values X and k represent a number of successes in n Bernoulli trials; all are dimensionless counts. The nCk coefficient is a quotient of products of dimensionless quantities. All exponents are dimensionless, and the Binomial PMF equation is dimensionally consistent.

4.2.2. Poisson distribution

The PMF for a Poisson random variable X with parameter $\lambda > 0$ is often written as:

$$P(X = k|\lambda) = \begin{cases} \frac{\lambda^k \cdot e^{-\lambda}}{k!} & 0 \leq k \\ 0 & \text{otherwise} \end{cases}$$

The distribution parameter, λ , is typically described as representing the average number of discrete events that occur in a window of fixed size within some continuous space. This definition, however, can lead to dimensional confusion. For example, Devore (2008, 121), gives the Poisson PMF as just shown, and then states, “The value of λ is frequently a rate per unit time or per unit area.” In general, if λ specifies a rate, then λ is not dimensionless, and cannot appear alone in an exponent. A clarification is possible that preserves the interpretation that λ represents a rate, and that also renders the Poisson PMF dimensionally consistent, for example, as presented in Yates and Goodman (2005). Define λ to be the average rate of discrete events per continuous window of unit size. Choose a fixed window of interest on the continuous space to be a particular size, say W . Then the average number of discrete events within W as it is moved to every position in the continuous space is $\lambda \cdot W$. The dimensions of λ and W are reciprocal, and their product is dimensionless whenever they are expressed in the same measurement units. The PMF for a Poisson random variable X is dimensionally consistent for $0 \leq k$ by writing:

$$P(X = k|\lambda, W) = \frac{(\lambda \cdot W)^k \cdot e^{-\lambda \cdot W}}{k!}$$

The expected value of a Poisson random variable X is $\lambda \cdot W$, a dimensionless quantity, consistent with the expected value representing a count.

4.3. Continuous random variables

4.3.1. Dimensions of probability density

For continuous random variables, probabilities are found by integrating a continuous probability density function (PDF) over some interval. In general, the definite integral of an integrable function $f(x)$ over the interval $[a, b]$ can be regarded as the limit of a Riemann sum of f over a partition Δ of $[a, b]$ as the norm of the partition approaches zero:

$$\int_a^b f(x) dx = \lim_{\|\Delta\| \rightarrow 0} \sum_{i=1}^n [f(c_i) \cdot (x_i - x_{i-1})]$$

where:

$$\Delta = x_i : i = 0, 1, \dots, n \mid a = x_0 < x_1 < \dots < x_n = b, x_{i-1} \leq c_i \leq x_i$$

Units of x are tracked through the mathematical operations that comprise f to determine units of elements in the range, denoted $||f(x)||$. For any partition Δ of $[a, b]$, every term in the Riemann sum has the same dimensions:

$$f(x_i \cdot |[x]|_{M_B}) \cdot (x_i \cdot |[x]|_{M_B} - x_{i-1} \cdot |[x]|_{M_B}) \sim [f(x)] \cdot |[x]|_{M_B}$$

In the limit, the norm of the partition does not reach zero. Thus, even in the limit, every subinterval $(x_i - x_{i-1})$ in the Riemann sum will always have units $|[x]|_{M_B}$, no matter how close adjacent partition boundaries become. All terms in the Riemann sum have dimensions $[f(x)] \cdot |[x]|_{M_B}$, so the Riemann sum is defined per Rule 1. Units are treated as a constant multiplier, and the limit of the sum will have the same units as the sum. That is, the definite integral of a function $f(x)$ is generally a dimensioned quantity:

$$\int_a^b f(x) dx \sim [f(x)] \cdot |[x]|_{M_B}$$

The outcome is a standard result of algebra for dimensioned quantities, for example as stated in Hart (1995, 71).

A function $f(X)$ is a PDF for the (real-valued) continuous random variable X if the following equation holds true for any interval $[a, b]$ of the real line:

$$P(a \leq X \leq b) = \int_a^b f(X) dX$$

For non-dimensionless random variables, the integral has units $||f(X)|| \cdot |[X]|_{M_B}$. Probability is dimensionless, and dimensional homogeneity in the PDF equation requires

$$[1] = [f(X)] \cdot |[X]|_{M_B} \Rightarrow [f(X)] = \frac{1}{|[X]|_{M_B}}$$

Probability density has units $1/[X]|_{M_B}$, the reciprocal of whatever units are ascribed to the random variable X . Hart (1995, 167) points out that these are the units one would expect of a density. Several additional scenarios illustrate this fundamental result.

- (1) Frequency density was shown to have units $1/[x]_{M_B}$. Units of probability density agree with the interpretation that a continuous PDF is analogous to an “idealized” frequency density histogram.
- (2) The expected value of a continuous random variable X with PDF $f(X)$ is defined as

$$E(X) = \int_{-\infty}^{+\infty} (X \cdot f(X)) dX \sim \left([X]_{M_B} \cdot \frac{1}{[X]_{M_B}} \right) \cdot [X]_{M_B} \sim [X]_{M_B}$$

The expected value of X is a location along the X -axis of the PDF, and thus should be interpreted to possess the same units as X . Hart (1995, 168) points out the necessity of ascribing the same units to $E(X)$ that X possesses: The two values are added together in numerous statistical definitions, such as variance:

$$\text{Var}(X) = E([X - E(X)]^2)$$

- (3) A PDF $f(X)$ can also be recovered by differentiating the associated cumulative distribution function (CDF) $F(x)$, a result that follows directly from the second fundamental theorem of Calculus:

$$\text{if } F(x) = \int_{-\infty}^x f(X) dX \quad \text{then} \quad \frac{d}{dx}[F(x)] = f(x)$$

The CDF evaluates to a probability, and so must be dimensionless.

- (4) Differentiating the CDF produces the slope of a tangent line on the CDF graph. The slope of any line on a CDF graph has units of the y -axis (dimensionless probability) divided by units of the x -axis, giving units of the slope to be $1/[x]_{M_B}$.

4.3.2. Uniform and normal distributions

The dimensionality of probability density is also evident when using familiar continuous PDFs to calculate probabilities. For example, a uniform random variable X over an interval $[a, b]$ on the real line, with measurement units $[X]_{M_B}$, has PDF

$$f(X) = \begin{cases} \frac{1}{(b \cdot [X]_{M_B} - a \cdot [X]_{M_B})} & a \leq X \leq B \\ 0 & \text{otherwise} \end{cases}$$

The height of the uniform PDF must have units that are the inverse of whatever units belong to the X -axis in order to make the area under the uniform PDF dimensionless.

The normal (Gaussian) PDF with mean μ and standard deviation σ is

$$f(X) = \frac{1}{\sigma \cdot [X]_{M_B} \cdot \sqrt{2\pi}} \cdot e^{-\frac{1}{2} \left(\frac{X \cdot [X]_{M_B} - \mu \cdot [X]_{M_B}}{\sigma \cdot [X]_{M_B}} \right)^2}$$

The units of μ and σ are the same units ascribed to the random variable X , rendering the exponent dimensionless. The σ in the denominator of the normalizing coefficient carries the same units as X , and thus the units of the normal PDF are the reciprocal of the units of X .

4.3.3. Exponential, gamma and χ^2 distributions

For a Poisson process in time that has an average rate, λ events per unit time, the interarrival time of events, X , can be modeled with an exponential distribution:

$$f(X) = \begin{cases} \lambda \cdot \frac{1}{|[X]|_{M_B}} \cdot e^{-\lambda \cdot \frac{1}{|[X]|_{M_B}} \cdot X \cdot |[X]|_{M_B}} & X \geq 0 \\ 0 & X < 0 \end{cases}$$

In the exponent, λ and X have reciprocal units, rendering the exponent dimensionless. The units of $E(X) = 1/\lambda$ are the same as X . An exponential distribution is shown in [Figure 1](#), with $\lambda = 2.0 \text{ s}^{-1}$, that is, a Poisson process with an average of two events per second. The PDF crosses the y -axis at λ , which has units that are consistent with probability density. $E(X)$ is a location along the X -axis, and has the same units as X . Height of the CDF gives the probability that $X \leq x$, a dimensionless quantity. The slope of a tangent line to the CDF has units $1/|[X]|_{M_B}$, consistent with a probability density. At a given value of X , the slope of a tangent line to the CDF is equal to the height of the PDF, and both quantities have the same units, namely, $1/|[X]|_{M_B}$.

The gamma distribution with shape parameter $k = 1$ is an exponential distribution, where the gamma scale parameter θ is the inverse of the exponential rate parameter, $\theta = 1/\lambda$. Units of the scale parameter θ are $|[X]|_{M_B}$. The shape parameter is dimensionless, and the gamma PDF is dimensionally consistent.

The χ^2 distribution with ν degrees of freedom is an accurate theoretical model for the distribution of a sum of squares of ν independent standard normal random variables. Any normally distributed random variable, X , with mean μ and standard deviation σ can be converted to a dimensionless standard normal variate by the transformation:

$$Z = \frac{X - \mu}{\sigma}$$

The χ^2 statistic can be transformed to a sum of dimensionless squares. In this context, the X -axis of the χ^2 PDF represents dimensionless values of the χ^2 statistic. The χ^2 distribution with ν degrees of freedom is also a special case of the gamma distribution with

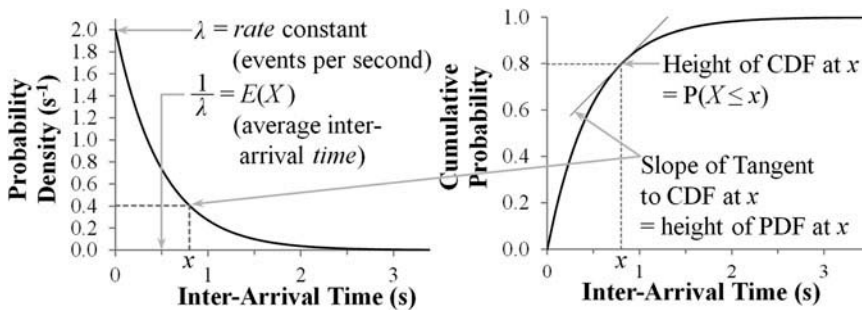


Figure 1. PDF (left) and CDF (right) for an exponentially distributed random variable X with rate parameter $\lambda = 2.0 \text{ s}^{-1}$. Units of the rate parameter λ and $E(X)$ are associated with the dimensions of each axis. The slope of a tangent line on the CDF graph at any location x is equal to the height of the PDF at x . The slope of a tangent line on the CDF has units $1/|[X]|_{M_B}$. It follows that probability density also has units $1/|[X]|_{M_B}$.

dimensionless shape parameter $k = \nu/2$ and scale parameter $\theta = 2$. Careful examination of the values in the χ^2 PDF using interpretations of the constants based on the gamma distribution confirms that the χ^2 PDF has units $1/|[X]|_{M_B}$:

$$f(X) = \frac{1 \cdot |[1]|}{\Gamma\left(\frac{\nu \cdot |[1]|}{2 \cdot |[1]|}\right) \cdot (2 \cdot |[X]|_{M_B})^{\frac{\nu \cdot |[1]|}{2 \cdot |[1]|}}} \cdot (X \cdot |[X]|_{M_B})^{\frac{\nu \cdot |[1]|}{2 \cdot |[1]|} - 1} \cdot |[1]| \cdot e^{\frac{-X \cdot |[X]|_{M_B}}{2 \cdot |[X]|_{M_B}}}$$

4.4. Dimensionally dubious PDFs

The χ^2 PDF is dimensionally consistent when each of the constants is interpreted within the context of a χ^2 statistic. However, some functions that are offered as exercises in statistics and probability texts require contrivances to attain dimensional consistency. It is technically possible to force any function to have units $1/|[X]|_{M_B}$, by assigning necessary units to constant values, and including “constants” with value 1 and appropriate units at required locations. However, such contrivances typically have no reasonable interpretation within the context of the random variable. For example, Navidi (2011, 115, problem 22) offers the following function as a density model for the error in the length, X , of a part, in millimeters:

$$f(X) = \begin{cases} \frac{e^{-X}}{1-e^{-1}} & 0 < X < 1 \\ 0 & \text{otherwise} \end{cases}$$

If X has units of millimeters, then its appearance in the exponent requires a dimensional contrivance, that is, multiplication by a “phantom” dimensioned value 1 mm^{-1} . Furthermore, both terms in the denominator must have units $|[X]|_{M_B}$ to render the “PDF” dimensionally consistent. Two serious issues are apparent. First, there is no discussion in the text regarding the well-known rule that exponents must be dimensionless, and the example explicitly violates the rule. Second, it is unlikely that the required dimensionalization of constants and inclusion of phantom dimensioned values have any reasonable scientific interpretation in the context of the random variable. It should be noted that both dimensionality in exponents of statistical formulas and dimension of probability density are discussed explicitly in Finney (1977).

Another example illustrates a third serious issue. Navidi (2011, 115, problem 18) offers the following function as a lifetime density model:

$$f(X) = \begin{cases} \frac{18}{(X+3)^4} & X \geq 0 \\ 0 & X < 0 \end{cases}$$

This density function could be made dimensionally consistent by assigning appropriate units to the constants. Once again, however, the dimensionalized constants likely do not have reasonable scientific interpretations in the context of the random variable. This example also illustrates another serious issue: There does not appear to be a theoretical justification for using the given function as a lifetime density model. Some guidance about the importance of theoretical pertinence for density models is offered in the National Institute of Standards and Technology (NIST) Engineering Statistics Handbook (NIST/SEMATECH *e-Handbook of Statistical Methods*: Croarkin and Tobias 2012, section 8.2.1):

Models like the lognormal and the Weibull are so flexible that it is not uncommon for both to fit a small set of failure data equally well. Yet, especially when projecting via acceleration models to a use condition far removed from the test data, these two models may predict failure rates that differ by orders of magnitude. That is why it is more than an academic exercise to try to find a theoretical justification for using a particular distribution.

5. Conclusions

The object of statistical analyses is “data” that are, in general, dimensioned physical quantities. Data scientists should be, at minimum, cognizant of results required by dimensional reasoning. The notion of equality for applied equations must include both numerical and dimensional character. Dimensioned linear algebra (DLA) provides a framework for assessing dimensional consistency in statistics and probability equations. Quantified dimensions, that is, measurement units, that are associated with statistics can be determined; for example, the sample mean has associated units:

$$\bar{x}_n \sim |[x]|_{M_B}$$

The sample mean is undefined whenever a sample data set contains elements with different dimensions, or whenever the values have not all been quantified to the same measurement units. Dimensional reasoning serves the goal of accuracy in graphical displays. Counts are dimensionless, so the vertical axes of frequency and relative frequency histograms do not require unit labels. Frequency density histogram bin values have units $1/|[x]|_{M_B}$, and the vertical axis must be labeled appropriately.

Probabilities must be dimensionless. Probability density has units $1/|[X]|_{M_B}$, a result that follows from the requirement that probability is a dimensionless quantity coupled with the second fundamental theorem of calculus. Requirements for a PDF should incorporate dimensional reasoning:

- (1) $F(X) \geq 0, -\infty < X < \infty$.
- (2) $\int_{-\infty}^{+\infty} F(X) dx = 1$.
- (3) $|[F(X)]| = \frac{1}{|[X]|_{M_B}}$.
- (4) $P(a \leq X \leq b) = \int_a^b F(X) dX, \quad \forall \quad a \leq b \in (-\infty, +\infty)$.

Functions that do not have units $1/|[X]|_{M_B}$ cannot be used as probability models for non-dimensionless random variables. Inclusion of phantom dimensioned constants that infix a function with the proper dimensionality is ill advised, since it is doubtful that such constants have any reasonable scientific interpretation in the context of the random variable. It is imprudent to suggest that an arbitrarily fashioned function can serve as a PDF; distributions should always be derived from theoretical models related to the random variable. Tenets of graphical excellence also require inclusion of dimensional labels on PDF graphs. Nearly all of the ideas presented here are completely absent from modern, widely used statistics and probability textbooks. Many such textbooks contain outright dimensional errors and inconsistencies. Furthermore, consideration for dimensional ideas can lead to enhanced understanding of some statistics and probability

concepts; for example, the calculus relationship between continuous PDFs and CDFs is clearly illustrated in Figure 1.

System-level thinking advocates that individual subsystems are better understood through the perspective of interactions with each other, rather than in isolation. Data dimensions and measurement units provide a common language for subsystem interaction. Dimensional reasoning contributes to a foundation for system-level thinking by providing a broader perspective of data as dimensioned physical quantities, and by promoting awareness that numerical values alone are insufficient to describe physical attributes or phenomena. Multidisciplinary teams are today's status quo, especially for the design and construction of complex systems. Team members should be proficient in system-level thinking. Applied statisticians are, more than ever, essential members of multidisciplinary teams, and preparation in dimensional reasoning contributes to a common ground for cross-disciplinary efforts.

References

- Albrecht, M. C., C. J. Nachtsheim, T. A. Albrecht, and R. D. Cook. 2013. Experimental design for engineering dimensional analysis. *Technometrics* 55 (3):257–70.
- Brand, L. 1957. Dimensional analysis, scaling, and orders of magnitude. *Chemical Engineering Science* 52 (24):4681–98.
- Bridgman, P. W. 1922. *Dimensional analysis*. New Haven, CT: Yale University Press.
- Bureau International des Poids et Mesures, Sèvres, France. (2006). *Le Système International d'Unités (SI), The International System of Units (SI)*, 8th ed. [Commonly referred to as the BIPM SI Brochure.] http://www.bipm.org/en/si/si_brochure/general.html.
- Carlson, D. E. 1979. A mathematical theory of physical units, dimensions, and measures. *Archive for Rational Mechanics and Analysis* 70 (4):289–305.
- Cleveland, W. S. 1985. *The elements of graphing data*. Summit, NJ: Hobart Press.
- Croarkin, C., and P. Tobias. 2012. *NIST/SEMATECH e-Handbook of Statistical Methods*. <http://www.itl.nist.gov/div898/handbook/>.
- Davis, T. 2011. *Dimensional analysis in experimental design*. Talk presented at Isaac Newton Institute. <http://sms.cam.ac.uk/media/1191742>.
- Davis, T. 2013. Comment: Dimensional analysis in statistical engineering. *Technometrics* 55 (3):271–74.
- Devore, J. L. 2008. *Probability and statistics for engineering and the sciences*, 7th ed. Belmont, CA: Thomas Higher Education.
- Drobot, S. 1953. On the foundations of dimensional analysis. *Studia Mathematica* 14 (1):84–99.
- Ellis, B. 1966. *Basic concepts of measurement*. Oxford, UK: Cambridge University Press.
- Finney, D. J. 1977. Dimensions of statistics. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 26 (3):285–89.
- Fröhlich, N. 2010. *Dimensional analysis of price–value deviations*. http://www.boeckler.de/pdf/v_2010_10_29_froehlich.pdf.
- Hand, D. J. 1996. Statistics and the theory of measurement. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* 159 (3):445–92.
- Hart, G. W. 1995. *Multidimensional analysis: algebras and systems for science and engineering*. New York, NY: Springer-Verlag.
- Huntley, H. E. 1967. *Dimensional analysis*. New York, NY: Dover.
- Kurth, R. 1965. A note on dimensional analysis. *American Mathematical Monthly* 72 (9):965–69.
- Levine, D. M., P. P. Ramsey, and R. K. Smidt. 2001. *Applied statistics for engineers and scientists*. Upper Saddle River, NJ: Prentice Hall.
- Lin, D. K., and W. Shen. 2013. Comment: some statistical concerns on dimensional analysis. *Technometrics* 55 (3):281–85.

- Luce, R. D. 1971. Similar systems and dimensionally invariant laws. *Philosophy of Science* 38 (2):157–69.
- Luce, D., D. Krantz, P. Suppes, and A. Tversky, eds. 1990. *Foundations of measurement, Vol. III: Representation, axiomatization, and invariance*. New York, NY: Academic Press.
- Macagno, E. O. 1971. Historico-critical review of dimensional analysis. *Journal of the Franklin Institute* 292 (6):391–402.
- Navidi, W. 2011. *Statistics for engineers and scientists*, 3rd ed. New York, NY: McGraw-Hill.
- Page, C. H. 1967. The mathematical representation of physical entities. *IEEE Transactions on Education* 10 (2):70–74.
- Palanthandalam-Madapusi, H. J., D. S. Bernstein, and R. Venugopal. 2007. Dimensional analysis of matrices: State-space models and dimensionless units [Lecture notes]. *IEEE Control Systems Magazine* 27 (6):100–9.
- Parkinson, M. 1964. An axiomatic approach to dimensions in physics. *American Journal of Physics* 32:200–5.
- Quade, W. 1967. Mathematical appendix on the algebraic structure of dimensional analysis. In *Dimensional analysis for economists*, ed. F. J. De Jong. Amsterdam, The Netherlands: North-Holland. Appendix translated from the German article: Quade, W. 1961, Über die Algebraische Struktur des Gröenalküls der Physik. *Abhandlungen der Braunschweigischen Wissenschaftlichen Gesellschaft*, XII:24–65.
- Rutherford, J. R., 1975. Design of experiments and data analysis: a scientific approach. In *Applied Statistics*, ed. R. P. Gupta, 271–287. Amsterdam: North Holland.
- Shen, W., T. Davis, D. K. Lin, and C. J. Nachtsheim. 2013. Dimensional analysis and its application in statistics. *Journal of Quality Technology* 46 (3):185–204.
- Stevenson, A. G., L. S. LaPiana, D. R. Mulville, P. J. Rutledge, F. H. Bauer, D. Folta, G. A. Dukeman, R. Sackheim, and P. Norvig. 1999. Mars Climate Orbiter Mishap Investigation Board Phase I Report. ftp://ftp.hq.nasa.gov/pub/pao/reports/1999/MCO_report.pdf.
- Szekeres, P. 1978. The mathematical foundations of dimensional analysis and the question of fundamental units. *International Journal of Theoretical Physics* 17 (12):957–74.
- Szirtes, T. 2007. *Applied dimensional analysis and modeling*. New York, NY: McGraw-Hill.
- Thompson, A. and B. N. Taylor, eds. 2008. *Guide for the use of the International System of units (SI)*. National Institute of Standards and Technology (NIST) Special Publication 811. <http://www.nist.gov/pml/pubs/sp811>.
- Thun, R. E. 1960. On dimensional analysis. *IBM Journal of Research and Development* 4 (3):349–56.
- Tufte, E. R. 2001. *The visual display of quantitative information*, 2nd ed. Cheshire, CT: Graphics Press.
- Vignaux, G. A. and J. L. Scott 1999. Simplifying regression models using dimensional analysis. *Australian and New Zealand Statistics* 41 (1):31–42.
- Whitney, H. 1968a. The mathematics of physical quantities: Part I: Mathematical models for measurement. *American Mathematical Monthly* 75 (2):115–38.
- Whitney, H. 1968b. The mathematics of physical quantities: Part II: Quantity structures and dimensional analysis. *American Mathematical Monthly* 75 (2):227–56.
- Yates, R. D., and D. J. Goodman, 2005. *Probability and stochastic processes, A friendly introduction for electrical and computer engineers*, 2nd ed. Hoboken, NJ: John Wiley & Sons.