



Geometry, Statistics, Probability: Variations on a Common Theme

Author(s): Peter Bryant

Source: *The American Statistician*, Vol. 38, No. 1 (Feb., 1984), pp. 38-48

Published by: [American Statistical Association](#)

Stable URL: <http://www.jstor.org/stable/2683558>

Accessed: 10/11/2010 17:34

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/page/info/about/policies/terms.jsp>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/action/showPublisher?publisherCode=astata>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.



American Statistical Association is collaborating with JSTOR to digitize, preserve and extend access to *The American Statistician*.

<http://www.jstor.org>

In this section, *The American Statistician* publishes articles and notes of interest to teachers of the first mathematical statistics course and of applied statistics courses. To be suitable for this section, articles

and notes should be useful to a substantial number of teachers of such a course or should have the potential for fundamentally affecting the way in which the course is taught.

Geometry, Statistics, Probability: Variations on a Common Theme

PETER BRYANT*

This article draws together some common geometrical ideas and their statistical and probabilistic analogs and outlines them for teaching elementary statistical ideas to students inside and outside the mathematical sciences. The main benefit from this approach is an appreciation of the surprising power of a small number of underlying principles. The approach emphasizes the equivalence of the notions, expressed in different "languages," rather than any one expression by itself.

KEY WORDS: Projections; Elementary statistics; Geometry.

1. INTRODUCTION

For the past five years or so I have presented elementary statistical ideas in a manner that emphasizes the equivalent expression of some common ideas in the different languages of geometry and statistics. The presentations varied from two-hour lectures to week-long seminars to full-semester courses. All enjoyed some success and followed, more or less, the approach described in this article.

Margolis (1979) points out that geometry seems to be the natural way to emphasize the unity of the fundamental ideas. The projection gives the best fit, and the angle measures how good that fit is. Students from outside the mathematical sciences should understand this

point: There are not really that many *formulas*; there are, however, many *variations* on a common theme. If they can learn that theme and understand that the variations are *only* variations, they will be better able to apply the fundamental principles, and thus will, perhaps, understand some of the fascination mathematicians feel for the geometric expression of statistical ideas.

In reviewing an earlier version of this article, a perceptive referee noted that getting students to understand this underlying unity

... is, of course, *the* problem in teaching any mathematics. The ability to transfer common ideas from one context (language) to another is evidently not naturally present in most students. Thus we have, for example, biological, psychological, business, nursing statistics. So although ... it is important to show students the common thread which holds the subject together, I would not be surprised to learn that many, if not most, teachers and students find it easier to teach and learn the separate ideas, blissful in their ignorance of the common thread.

Indeed, as Herr (1980) points out, the geometric approach has fallen out of favor after heavy use in the early days of mathematical statistics.

I feel that perhaps one reason for this lack of unity is that the relevant material has not been published in the appropriate elementary-level literature. Thus, although the expression of similar ideas in terms of geometry, analytic geometry, and statistics is exploited systematically in Dempster (1968), for example, it is done at a high mathematical level, and nothing seems to be available at the level of *The Teacher's Corner*.

Sections 2 through 5 contain the relevant ideas of geometry, analytic geometry, statistics, and probability. Section 6 contains comments on their presentation and some suggested references. None of the material is new.

2. GEOMETRY

In this section, we show how to express ordinary geometric ideas—lines, planes, length, distance, angles, and projections—in terms of vectors and inner prod-

*Peter Bryant is Associate Professor of Management Science and Information Systems, University of Colorado at Denver, 1055 Wazee St., Denver, CO 80204. The author thanks Lucinda Bryant, Henry Ludlam, Paul Jedamus, an anonymous referee, and (particularly) the associate editor for helpful comments on earlier drafts of this article. He also thanks the students at the department of EPO Biology, University of Colorado at Boulder, and at CENARGEN, Brasilia, Brazil, for their participation in the development of the material and Estella Breitler for help with the manuscript. Some of the work was done while the author was at the International Business Machines Corporation.

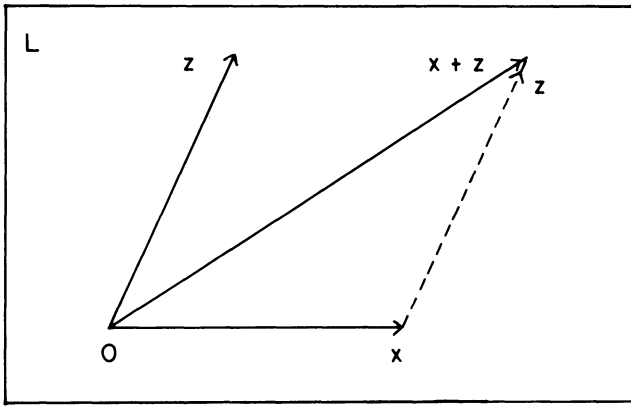


Figure 1. Vectors and vector addition.

ucts. We do this to introduce two important ideas: (a) We can express familiar (in this case geometric) ideas in another language, the language of vector spaces. (b) Whenever something has the properties of a vector, we can interpret it geometrically, however we may have thought of it at first.

Vectors

Using vectors and some simple operations on them, we can express the ideas of line segments, lines, planes, and plane figures such as triangles. In Figure 1, consider the plane L and a particular point in L , called the *origin*, 0 (zero). We use letters such as x and z to denote other points in L . We also use letters like x and z to denote the directed line segments from 0 to the points x and z , respectively. We drew arrows on the lines of Figure 1 to show this. In practice, such ambiguity of notation is not so troublesome as one might at first imagine.

We call directed line segments *vectors*. Thus the vector x is the directed line segment from 0 to the point x . Vectors have *length* and *direction*. The vector from 0 to x is not the same as the vector from x to 0 , for although they have the same lengths, they go in opposite directions. In general, we call two vectors *equal* if they have the same length and direction, even if they start at different points. By this rule, for example, the vector shown by a dashed line in Figure 1 is equal to the vector z , because it has the same direction and length as z , even though it starts at point x , not at 0 . We mentally move all vectors to begin at the origin 0 , retaining their original directions and lengths. ("Equivalent" might seem more appropriate here than "equal," but the latter is convenient later.)

We speak of adding the vectors x and z , obtaining a third vector $x + z$. Geometrically, we mean by this that we start at 0 and proceed in a direction and for a length given by vector x . We then proceed in a direction and for a length given by vector z , finally arriving at a point we call $x + z$. Adding x and z means placing the beginning of z on the end of x . The sum is the vector from 0 to the end of z . In this way, x , z , and $x + z$ form the triangle in Figure 1. We will see in Section 3 how this

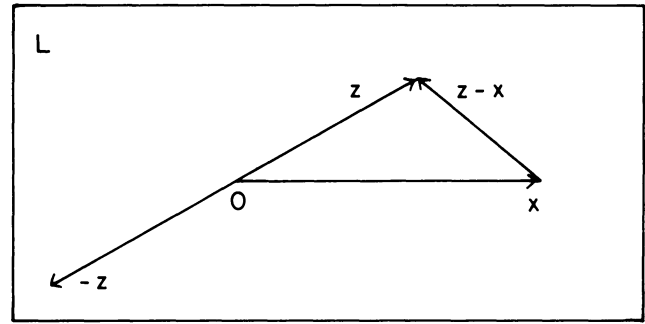


Figure 2. Vector subtraction.

corresponds to our ordinary idea of addition of numbers. For now, it is just a definition.

We can also subtract vectors. The vector from x to z is the difference vector $z - x$, for it is what must be added to x to obtain z according to our new definition of addition. See Figure 2. By this definition, then, $z = z - 0$ and $-z = 0 - z$. Note that $-z$ has the same length as z , but goes in the opposite direction—it is what we would have to add to z to obtain 0 (since we do not care about the starting point).

We also speak of multiplying vectors by numbers (called *scalars* to distinguish them from vectors). See Figure 3. By the vector $3z$ we mean a vector in the same direction as z , but three times as long. (Equivalently, it is the vector $z + z + z$.) Geometrically, this is an extension of the line segment from 0 to z . In general, the vector cz is c times as long as z . If $c > 0$, cz has the same direction as z ; if $c < 0$, cz has the opposite direction. Thus $(-\frac{1}{2})x$ is as shown in Figure 3.

A vector x determines a line called $L(x)$. This line is the indefinite extension in both directions of the line segment from 0 to x . In vector terms, $L(x)$ consists of all scalar multiples of x . Every point on $L(x)$ is the endpoint of some multiple of the vector x . $L(x)$ and $L(z)$ are indicated by dashed lines in Figure 3.

We can reach any point in the plane L of Figure 3 by starting at zero, proceeding in direction x for some distance and from there proceeding in direction z for some (other) distance; that is, any point in L can be expressed

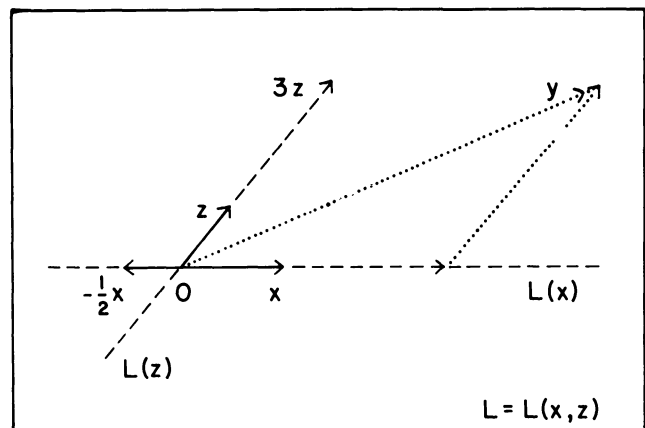


Figure 3. Scalar multiplication and subspaces.

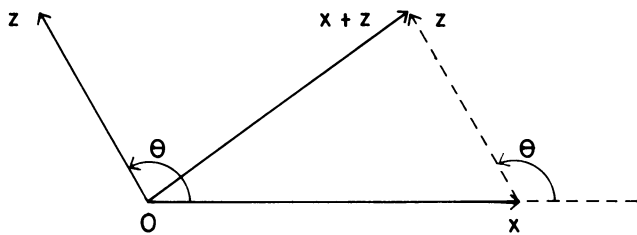


Figure 4. The law of cosines.

as $ax + bz$ for some scalars (numbers) a and b , using the preceding definitions of addition and scalar multiplication. As long as x and z are not parallel to each other, there is one and only one such *linear combination* of x and z for every point in L . In this sense, x and z together determine the plane in the same way in which x and z individually determine $L(x)$ and $L(z)$, respectively. We thus write $L = L(x, z) =$ all linear combinations of x and z . $L(x)$, $L(z)$ and $L(x, z)$ are examples of *linear spaces*—spaces by analogy with our ordinary ideas of space (as we will see), and linear because they are composed of linear combinations. Since $L(x)$ is contained within $L(x, z)$, we call it a *subspace*. Note that linear subspaces always contain 0.

Line segments correspond to vectors; the sides of plane figures are associated with vector addition; and planes and lines through the origin are analogous to linear combinations of vectors.

Lengths, Distance, Angles, and Inner Products

By the *length* of a vector x , $|x|$, we mean the length of the line segment from 0 to x . The *distance* from x to z is the length of the line segment from x to z ; that is, it is the length of $z - x$, $|z - x|$. In Figure 4, consider the triangle whose vertices are 0, x , and $x + z$. The law of cosines says that

$$|x + z|^2 = |x|^2 + |z|^2 + 2|x||z|\cos(\theta), \quad (2.1)$$

where θ is the angle between the vectors x and z . When $\theta = 90^\circ$ (x and z are at right angles), the triangle is a right triangle, $\cos(\theta) = 0$, and (2.1) reduces to the familiar law of Pythagoras:

$$|x + z|^2 = |x|^2 + |z|^2. \quad (2.2)$$

The difference between (2.1) and (2.2) is the quantity $2|x||z|\cos(\theta)$, which is a measure of how much the Pythagorean relation (2.2) fails to hold—the extent to which x and z are *not* at right angles. This quantity is a fundamental description of the relationship between the two vectors x and z , and we shall study it further.

Given two vectors x and z , the quantity

$$\langle x, z \rangle = |x||z|\cos(\theta), \quad (2.3)$$

where θ is the angle between x and z , is called the *inner product* of x and z . Using this notation, (2.1) becomes

$$|x + z|^2 = |x|^2 + |z|^2 + 2\langle x, z \rangle, \quad (2.4)$$

which corresponds to the ordinary algebraic equation $(a + b)^2 = a^2 + b^2 + 2ab$.

The fundamental nature of the inner product becomes clearer when we realize that once we know the inner product of two vectors, we also know the lengths, distances, and angles related to the two vectors:

$$\text{Length: } |x|^2 = \langle x, x \rangle$$

$$|z|^2 = \langle z, z \rangle$$

$$\begin{aligned} \text{Distance: } |z - x|^2 &= |x|^2 + |z|^2 - 2\langle x, z \rangle \\ &= \langle x, x \rangle + \langle z, z \rangle - 2\langle x, z \rangle \end{aligned}$$

$$\begin{aligned} \text{Angle: } \cos(\theta) &= \langle x, z \rangle / (|x||z|) \\ &= \langle x, z \rangle / \sqrt{\langle x, x \rangle \langle z, z \rangle} \end{aligned}$$

Once we know the inner product, then, we know the geometry. Inner products satisfy

$$\langle x, z \rangle = \langle z, x \rangle$$

$$\langle x, x \rangle \geq 0 \text{ and } \langle x, x \rangle = 0 \text{ iff } x = 0$$

$$\langle cx + dy, z \rangle = c\langle x, z \rangle + d\langle y, z \rangle. \quad (2.5)$$

In particular, two vectors x and z are orthogonal if and only if

$$\langle x, z \rangle = 0. \quad (2.6)$$

Vector Spaces

We saw previously that plane geometry can be expressed in terms of vectors, vector addition, scalar multiplication, and inner products. We can do the same for three-dimensional geometry. Instead of a point x in a plane, we have a point x in space. The directed line segment from the origin to x is a vector x . The inner product of any two vectors x and z is still defined by (2.3), for although the vectors are in space, any two of them lie in some plane, and the angle between them is well defined.

Thus by addition, scalar multiplication, and inner products of our vectors in space, we obtain the usual notions of three-dimensional Euclidean geometry. Indeed, the very notion of *dimension* is related to these ideas. A line such as $L(x)$ is determined by a single vector x —it is *one-dimensional*. A plane $L(x, y)$ is determined by two collinear vectors, x and y . Our ordinary idea of space is three-dimensional: Any point can be represented as a linear combination of three non-coplanar vectors, and we could call it $L(x, y, z)$. $L(x, y)$ would be a two-dimensional *subspace* of $L(x, y, z)$ —a plane through the origin. $L(y, z)$ would be another, different subspace. In general, the minimum number of vectors required to determine a subspace is called the *dimension* of the subspace.

A (finite-dimensional) vector space is an abstraction of this approach to geometry. We say we have such a vector space whenever (a) we have a collection of elements, called vectors, that can be added and multiplied by scalars as above; and (b) the inner product of any two vectors is defined and satisfies (2.5). In this approach we do not describe what vectors are; we describe how they behave. Anything that behaves this way is a vector.

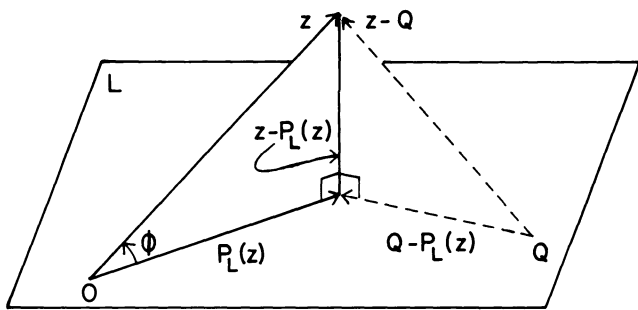


Figure 5. Projections.

Two important complementary ideas emerge from this approach: (a) Though the idea is inspired by common geometrical ideas, the elements of a vector space need not appear geometric on the surface. (b) Once an inner product is defined, we may interpret the elements of a vector space geometrically, whatever their “true” nature. For example, consider random variables x and y . Their sum and difference are also random variables, and so is $cx + dy$ for any numbers c and d . That is, random variables can be added and multiplied by scalars, which makes them vectors. If an appropriate inner product is defined, we can think of random variables geometrically; then linear combinations of random variables are like planes through the origin, and so forth. We will return to this example in Section 5.

Projections

In Figure 5, we have the *orthogonal projection* $P_L(z)$ of a point z onto a plane L containing 0. Alternatively, we could say the *vector* $P_L(z)$ is the orthogonal projection of the vector z onto L . Two properties define the orthogonal projection:

Property A: $P_L(z)$ is in the plane L ; and

Property B: $z - P_L(z)$ is orthogonal to every vector in L ; that is, $\langle x, z - P_L(z) \rangle = 0$ for every x in L .

The following simple theorems give some important properties of projections. The proofs illustrate well the remarkable power of inner products. They also make good exercises for advanced students. (Use (2.5), (2.6), Properties A and B, and prove them in the order listed.)

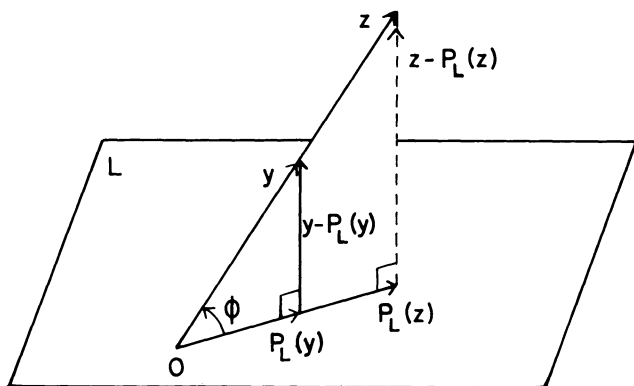


Figure 6. Measuring distance to a plane.

Theorem 1 (T1). z is in L if and only if $P_L(z) = z$. This says that if z is already in the plane L , projecting it does not change anything; conversely, any vector unchanged by projecting it into L must have been in L to begin with.

Theorem 2 (T2). z is orthogonal to L if and only if $P_L(z) = 0$. This says that if z is orthogonal to L , projecting it into L yields nothing; conversely, anything that yields nothing when projected into L must have been orthogonal to L to begin with.

Theorem 3 (T3). $P_L[P_L(z)] = P_L(z)$. Projecting the projection does not change anything, for $P_L(z)$ is in L , and thus (T1) applies.

Theorem 4 (T4). $|z|^2 = |P_L(z)|^2 + |z - P_L(z)|^2$. This is the law of Pythagoras.

Theorem 5 (T5). For each vector z , the projection is unique. That is, if $P'_L(z)$ and $P''_L(z)$ are two projections of the same point z onto the same subspace L , then $P'_L(z) = P''_L(z)$.

Theorem 6 (T6). Of all the vectors in L , $P_L(z)$ is the closest to z .

Theorem 7 (T7). Projections are linear: $P_L(ay + z) = aP_L(y) + P_L(z)$.

The proof of (T6) is instructive (see Figure 5). Consider any other point Q in L . From Property B, $z - P_L(z)$ is orthogonal to the vector $Q - P_L(z)$. The law of Pythagoras then implies that

$$|z - Q|^2 = |z - P_L(z)|^2 + |Q - P_L(z)|^2 \geq |z - P_L(z)|^2.$$

The inner product form of this argument is

$$\begin{aligned} |z - Q|^2 &= |z - P_L(z) + P_L(z) - Q|^2 \\ &= \langle z - P_L(z) + P_L(z) - Q, z - P_L(z) + P_L(z) - Q \rangle \\ &= \langle z - P_L(z), z - P_L(z) \rangle \\ &\quad + \langle P_L(z) - Q, P_L(z) - Q \rangle \\ &\quad + 2\langle z - P_L(z), P_L(z) - Q \rangle \\ &= |z - P_L(z)|^2 + |P_L(z) - Q|^2 \\ &\quad \text{(using Property B)} \\ &\geq |z - P_L(z)|^2. \end{aligned}$$

Such parallel expression of ideas in the “languages” of geometry and inner products is a recurrent theme in the discussion that follows.

Pedagogically, this proof offers the additional advantage of solving a minimization problem by appealing to familiar geometric notions rather than to calculus. Formally, of course, we would have to verify the existence of a vector with Properties A and B, and so forth.

Consider Figure 6. How far away from the plane L is the vector y ? The vector z ? One natural measure of these distances comes from the squared distances to the

closest points in the plane, the numbers $|y - P_L(y)|^2$ and $|z - P_L(z)|^2$. These measure the absolute squared distances from the endpoints of the vectors to the plane. According to these measures, z is further away from L than y is. Another measure is the angle ϕ between z and L . The quantity $\cos^2(\phi) = |P_L(z)|^2/|z|^2$ measures, on a scale of 0 (when z is orthogonal to L) to 1 (when z is in L), the extent to which z is close to L , that is, the extent to which z can be represented as a linear combination of vectors in L . Both measures have their uses. The expression in $|z - P_L(z)|$ is naturally interpreted as distance, whereas the angle has the advantage that if the vector z is extended or shrunk, the measure of closeness is unchanged. The vectors y and z are equally close to L by this measure, and this can be useful in statistical applications. In summary:

Criterion 1. The projection $P_L(z)$ is the closest vector in L to z .

Criterion 2. Either $|z - P_L(z)|$ or the angle ϕ (or some function of it, such as its cosine) may be used to measure how close $P_L(z)$ is to z .

We will return to these principles in Section 4. Note that both the projection and the angle depend on the inner product.

3. ANALYTIC GEOMETRY

In Section 2, vectors and inner products were introduced and used to describe geometric ideas. It is only when we express the ideas in terms of coordinates, though, that we can actually *compute* anything. We review coordinate geometry briefly in this section and use it to lead to the ideas of n -dimensional geometries.

Coordinates

Consider an ostensibly new vector space L , defined as follows:

1. The vectors (elements) of L are all the triples of numbers of the form $z = (z_1, z_2, z_3)$, $x = (x_1, x_2, x_3)$, and so on.
2. Addition of any two vectors is defined component by component: $(z + x) = (z_1 + x_1, z_2 + x_2, z_3 + x_3)$.
3. Scalar multiplication is defined similarly: $cz = (cz_1, cz_2, cz_3)$.
4. The inner product of any two vectors is defined by

$$\langle z, x \rangle = z_1 x_1 + z_2 x_2 + z_3 x_3. \quad (3.1)$$

It is easy to verify that the definitions of a finite-dimensional vector space are satisfied for L . The length of a vector z is then given by

$$|z|^2 = z_1^2 + z_2^2 + z_3^2. \quad (3.2)$$

Consider the vectors $i_1 = (1, 0, 0)$, $i_2 = (0, 1, 0)$, and $i_3 = (0, 0, 1)$. Using (3.1) and (3.2), we see that $|i_1|^2 = |i_2|^2 = |i_3|^2 = 1$ and

$$\langle i_1, i_2 \rangle = \langle i_1, i_3 \rangle = \langle i_2, i_3 \rangle = 0. \quad (3.3)$$

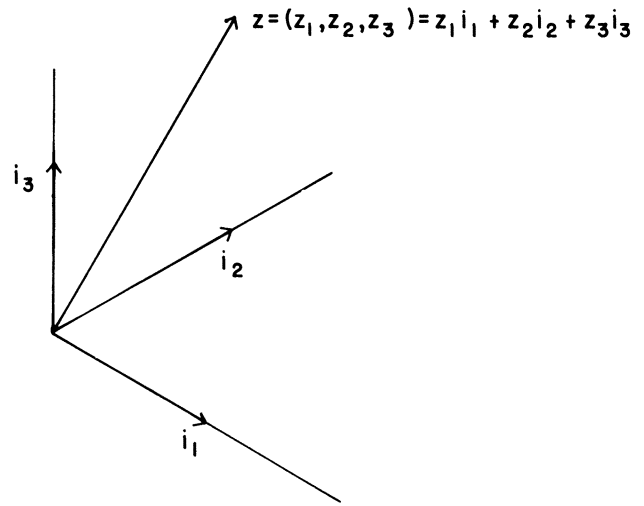


Figure 7. Coordinates and axes.

That is, the vectors i_1 , i_2 , and i_3 are *unit* vectors (of length 1) that are mutually orthogonal. Further, any vector $z = (z_1, z_2, z_3)$ can be written in the form

$$z = z_1 i_1 + z_2 i_2 + z_3 i_3 \quad (3.4)$$

as a linear combination of i_1 , i_2 , and i_3 . The vector space L is thus $L = L(i_1, i_2, i_3)$, the three-dimensional space determined by i_1 , i_2 , and i_3 . Figure 7 shows this relationship geometrically. The *axes* i_1 , i_2 , and i_3 are shown as mutually orthogonal. The coefficients z in (3.4) are called the *coordinates* of z relative to these axes.

A triple of numbers (z_1, z_2, z_3) may be interpreted geometrically, then, if one thinks of it as a vector in space determined by (3.4), relative to some imagined set of mutually orthogonal axes i_1 , i_2 , and i_3 and inner product (3.1). It can be verified that the angles, cosines, and so on derived from this inner product agree with our ordinary geometric notions. In particular, the vectors x and z are orthogonal when (3.1) vanishes. A standard result in analytic geometry is that

$$\cos^2(\phi) = \frac{(x_1 z_1 + x_2 z_2 + x_3 z_3)^2}{[(x_1^2 + x_2^2 + x_3^2)(z_1^2 + z_2^2 + z_3^2)]}, \quad (3.5)$$

which agrees with what (3.1) suggests. Equation (3.2) is a three-dimensional law of Pythagoras. We thus have a correspondence between geometric ideas and triples of numbers, achieved by interpreting the triples as a vector space with inner product given by (3.1).

From Section 2, we know that if we have vectors and an inner product, we have essentially determined a geometry. From this point of view, there is no reason to limit ourselves to three dimensions. We could think of any collection of n numbers z_1, z_2, \dots, z_n as a vector $z = (z_1, z_2, \dots, z_n)$, or as the point whose coordinates are the z 's relative to some imagined set of n orthogonal axes. In this scheme, the vectors are the n -tuples of the form $z = (z_1, z_2, \dots, z_n)$, and addition and scalar multiplication are defined in terms of the individual coordinates. Our axes are in the directions of $i_1 = (1, 0, \dots, 0)$, $i_2 = (0, 1, \dots, 0)$, and $i_n = (0, 0, \dots, 1)$. Our n -dimen-

sional space is $L(i_1, i_2, \dots, i_n)$, the set of all linear combinations of i_1, \dots, i_n . The inner product of x and z is

$$\langle x, z \rangle = x_1 z_1 + x_2 z_2 + \dots + x_n z_n. \quad (3.6)$$

It is easy to verify that with these definitions our n -tuples behave like vectors, and (2.5) is satisfied. In an abstract way, then, n -tuples of numbers are like vectors in an n -dimensional space. They have lengths $|z|^2 = \langle z, z \rangle$, and so forth. In three dimensions ($n = 3$), it is exactly our ordinary geometry. When $n > 3$, it is an extended (imaginary?) analog of our ordinary ideas. For example, $L(x, z)$, the set of all linear combinations of x and z , is now a two-dimensional plane through the origin of an n -dimensional space, and so on. Although the definitions like (3.6) are given in coordinate form, we could by this means think about n tuples of numbers geometrically, if it were useful.

Other Axes

We can compute inner products from (3.6) if we know the coordinates. This formula is for one particular set of axes. If we were to use a different set of orthogonal axes, each geometric point z would have a new set of coordinates to be used in computing inner products (and from them, lengths, angles, etc.). As long as the axes are orthogonal, the formula will be of the general form (3.6). If the axes are not orthogonal, a more complicated formula is required.

On the other hand, fundamentally geometric notions like length, angles, and so forth determined from inner products are the same for any coordinate system. For

example, the angle between two vectors does not depend on which axes you use. In this sense, the axes are a computational crutch. We can use any set we like, and we will get the same answers no matter which ones we use, at least for the geometric aspects of things.

Other Inner Products

The inner product (3.6) is the Euclidean inner product because it corresponds to ordinary Euclidean geometry when $n = 3$. It is by far the most common inner product used for n -tuples of numbers and the only one we need in this article, but it is by no means the only one that might be defined. Any function of the x 's and z 's that satisfies (2.5) is a satisfactory inner product. When interpreted according to our conventions (e.g., $\langle z, z \rangle$ is the squared length of z), it will generate its own geometry, which may or may not agree with our ordinary notions.

We will see another inner product in Section 5, though not for n -tuples of numbers. We may summarize the results of Sections 2 and 3 as shown in the first two columns of Table 1: Each notion of geometry has a corresponding expression in coordinate form, and we can think of them in whatever way seems convenient. Instructors may want students to construct Table 1 or its equivalent, with successive columns filled in as the various topics are covered.

4. STATISTICS

Consider the following common situations of descriptive statistics:

Table 1. Vector Space Ideas Expressed in Different Ways

Vector Space Idea	Geometric Expression		Statistical Expression	Probabilistic Expression
	Coordinate-Free Form	Analytic Form		
Vector z	Directed line segment from 0 to z	$z = (z_1, \dots, z_n)$	Data z_1, \dots, z_n	Random variable z
Vector $z - y$	Directed line segment from y to z	$z - y = (z_1 - y_1, \dots, z_n - y_n)$	Differences $z_1 - y_1, \dots, z_n - y_n$	Random variable $z - y$
Inner product of x and z , $\langle x, z \rangle$	Measure of Nonorthogonality	$x_1 z_1 + \dots + x_n z_n$	Sum of cross products	$E(xz)$
$ z ^2 = \langle z, z \rangle$	Squared distance, 0 to z	$z_1^2 + \dots + z_n^2$	Sum of squares of z	$E(z^2)$
$ z - y ^2$	Squared distance, y to z	$(z_1 - y_1)^2 + \dots + (z_n - y_n)^2$	Sum of squared differences	$E(z - y)^2$
$\langle x, z \rangle = 0$	Orthogonality: x, z perpendicular			
All vectors of the form $ax + by + cz + \dots$	Plane $L(x, y, z, \dots)$ through 0 determined by vectors x, y, z, \dots		All models of the form $ax + by + cz + \dots$	All square-integrable functions of x, y, z, \dots
A vector P in $L(x, y, \dots)$ such that $\langle x - P, Q \rangle = 0$ for every Q in L	Projection $P_L(z)$ of z onto $L(x, y, \dots)$		The best-fitting model of the form $ax + by + cz + \dots$	$E(z x, y, \dots)$
"45-degree" vector J	$J = (1, 1, \dots, 1)$		Constant $(1, 1, \dots, 1)$	Constant 1
$ x - P_J(x) ^2$	$\sum (x_i - \bar{x})^2$		$(n - 1) \times$ sample variance of x	Variance of x
$\frac{\langle x - P_J(x), z - P_J(z) \rangle}{ x - P_J(x) \cdot z - P_J(z) }$	Cosine of angle between $x - P_J(x)$ and $z - P_J(z)$		Sample correlation of x and z	Correlation of x and z
$\langle x - P_J(x), z - P_J(z) \rangle = 0$	$x - P_J(x)$ and $z - P_J(z)$ are perpendicular		x and z are uncorrelated	x and z are uncorrelated

(Q1) I wish to summarize my data (z_1, z_2, \dots, z_n) by a single number a . What is the best number a to use? How good a summary is it?

(Q2) I have measurements (y_1, y_2, \dots, y_n) and (z_1, z_2, \dots, z_n) on two characteristics y and z of my n experimental units. How strongly related (if at all) are these characteristics?

(Q3) I suspect that certain characteristics z , x , and y of my n experimental units are related (at least approximately) by a linear function $z = a + bx + cy$. What linear function best describes the relationship? How good a description is it?

In the classical approach to these situations, we use the *least squares* criterion to define best fit. In this section, we show how that approach is related to the geometric ideas of Sections 2 and 3. Once that relationship is clear, the least squares approach offers a convenient and unified approach to various situations that are often handled by individual formulas. Formal principles of inference also justify the least squares approach, but for many pedagogical purposes, its unity, simplicity, and intuitive geometric appeal are justification enough.

It will be convenient to think of the data in (Q1)–(Q3) as a matrix containing n rows (one for each observation or experimental unit) and one column for each characteristic or variable, as follows:

1	z_1	x_1	y_1	...
1	z_2	x_2	y_2	...
·	·	·	·	·
·	·	·	·	·
1	z_n	x_n	y_n	...

and to interpret the n -tuples that make up the columns of the matrix as vectors. The constant vector $J = (1, 1, \dots, 1)$ plays a special role in what follows. We will interpret the n -tuples $z = (z_1, z_2, \dots, z_n)$ geometrically, of course.

According to the least squares principle, the best single-number summary in (Q1) is the number a that minimizes

$$\sum (z_i - a)^2. \quad (4.1)$$

Using the preceding definitions and the Euclidean inner product, (4.1) can be written as

$$|z - aJ|^2. \quad (4.2)$$

Geometrically, then (see Figure 8), the least squares approach to (Q1) says that we should consider as possible models all multiples aJ of J and choose that one closest to z . "Least squares" means "shortest distance," using the Euclidean inner product.

The geometry of Section 2 tells us that the closest point is the projection $P_J(z)$ of z onto J . Using the inner product characterization of projection, we can then derive the value for a :

$$\text{By property A, } P_J(z) = aJ \text{ for some } a. \quad (4.3)$$

$$\text{By property B, } \langle z - P_J(z), J \rangle = 0. \quad (4.4)$$

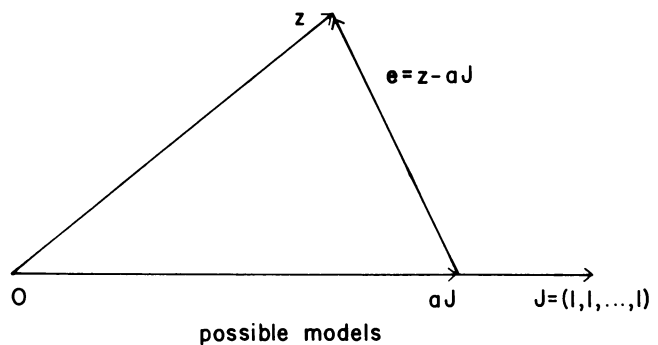


Figure 8. A simple statistical model.

Substituting (4.3) into (4.4), we obtain $\langle z - aJ, J \rangle = 0$, or

$$\langle z, J \rangle = a \langle J, J \rangle. \quad (4.5)$$

Using the Euclidean inner product, (4.5) becomes $\sum z_i = an$, or

$$a = \bar{z} = (1/n) \sum z_i. \quad (4.6)$$

The first question of (Q1) is thus answered: The sample mean \bar{z} is the best single number summary of the data. Geometry suggests the solution; analytic geometry allows us to compute it. The solution portrayed in Figure 9 indicates how we have decomposed z into two orthogonal components: $\bar{z}J$, the component along J , and $e = z - \bar{z}J$, the remainder or error. If all the z 's were identical, z would lie on J exactly, and we would have $|e|^2 = 0$. Thus we can think of $\bar{z}J$ as the part of z that can be satisfactorily explained by a constant. To the extent that $|e|^2 > 0$, this explanation is imperfect because the data vary. We interpret e as the *variations* in z . Thus $z = \text{constant component} + \text{variations} = \bar{z}J + (z - \bar{z}J)$.

The corresponding decomposition of the squared lengths is found from the law of Pythagoras:

$$|z|^2 = |\bar{z}J|^2 + |z - \bar{z}J|^2,$$

which reduces in this case to

$$\sum z_i^2 = n\bar{z}^2 + \sum (z_i - \bar{z})^2,$$

the usual decomposition of the sums of squares.

How good a summary is \bar{z} ? Various measures are used. From Criterion 2, either ϕ or $|e|^2$ could be used. $|e|^2$ is often called the *sum of squares error* (SSE). (Can

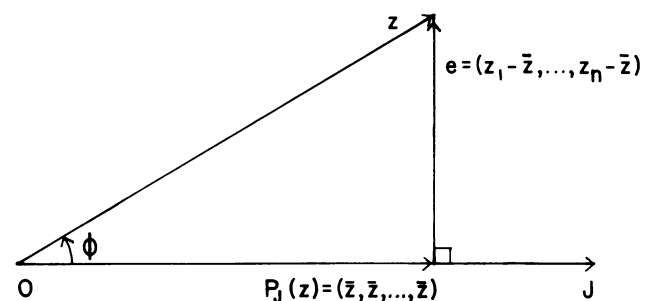


Figure 9. Derivation of the sample mean.

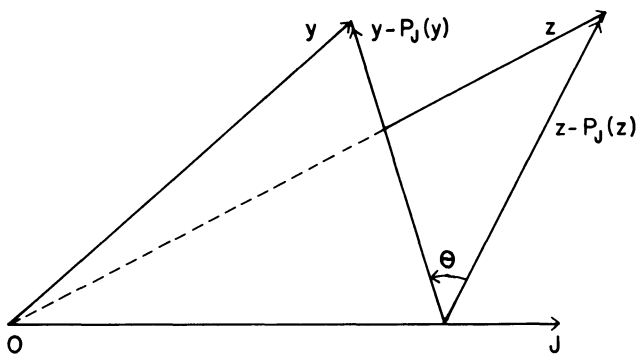


Figure 10. The simple correlation coefficient.

you find an interpretation of ϕ in terms of classical statistical quantities?) A variation of $|e|^2$ called the *mean square for error* (MSE) is often used (though not usually given this name in this situation). We expect $|e|^2$ to depend on n —the more observations, the more variation, at least in an absolute sense. To account for this, we may divide SSE by the dimension of the relevant subspace, obtaining an average error sum of squares “per dimension.” Since e is restricted to lie in the $(n - 1)$ -dimensional subspace orthogonal to J , we divide by $(n - 1)$, obtaining:

$$\text{MSE} = |e|^2 / (n - 1) = (n - 1)^{-1} \sum (z_i - \bar{z})^2,$$

the usual sample variance. Either SSE or MSE is a measure, then, of the extent to which the best single number summary \bar{z} fails to summarize the data. Incidentally, the dimension of the subspace within which a vector is restricted to lie is often called the *degrees of freedom* associated with that vector.

In situation (Q2) we must assess the strength of the relationship between two variables y and z . Using the same geometric interpretation of y and z as before, this means asking how close two vectors y and z are to each other (see Figure 10). By Criterion 2, we could measure this by $|y - z|^2$ or by some function of the angle between y and z . In practice, however, (Q2) is interpreted to mean “how strongly are *variations* in y related to *variations* in z ?” As in (Q1), variations in y and z are taken to be their components orthogonal to J : $y - \bar{y}J = y - P_J(y)$ and $z - \bar{z}J = z - P_J(z)$ in Figure 10. The usual measure of the relationship of variations in y and z is $\cos(\theta)$ in Figure 10, usually called the (simple) *correlation* of y and z . When $\cos^2(\theta) = 1$ ($\theta = 0^\circ$ or 180°) variations in y are exactly proportional to variations in z : $y - \bar{y}J$ and $z - \bar{z}J$ are collinear. When $\cos(\theta) = 0$ ($\theta = 90^\circ$) variations in y are unrelated to variations in z : $y - \bar{y}J$ and $z - \bar{z}J$ are orthogonal.

The inner product of $y - \bar{y}J$ and $z - \bar{z}J$, when divided by the appropriate degrees of freedom $(n - 1)$, is the sample *covariance* of y and z . In a more general setting, if we first project y and z onto the plane L determined by J, w, x, \dots , then the cosine of the angle between $y - P_L(y)$ and $z - P_L(z)$ is called the *partial correlation* of y and z , *removing the effects of* w, x, \dots . Most correlation analyses have natural geometric interpreta-

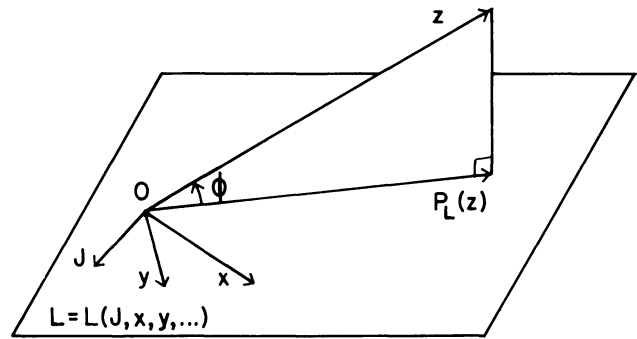


Figure 11. Regression.

tions—the only twist is that attention is usually restricted to the subspace orthogonal to J —the subspace of variation in the variables.

In (Q3) we have an example of the more general linear (or *regression*) model, $z = a + bx + cy + \dots$, and as before we interpret z, x , and y, \dots , as vectors. The least squares criterion requires that we choose a, b, c, \dots , to minimize $|z - aJ - bx - cy - \dots|^2$, and Criterion 1 gives the solution. If $L = L(J, x, y, \dots)$ is the plane determined by J, x, y, \dots , then the best fitting linear function $aJ + bx + cy + \dots$ is the projection $P_L(z)$ (see Figure 11). To measure the goodness of fit, Criterion 2 suggests we use either $\text{SSE} = |z - P_L(z)|^2$ or ϕ , the angle between z and $P_L(z)$, or something similar.

As in (Q1) and (Q2), we sometimes focus our attention on the subspace orthogonal to J , the subspace of variations. In particular, the angle θ between $z - P_J(z)$ and $P_L(z) - P_J(z)$ is a measure of how much better the fit is using J, x, y, \dots in the model than it is using J alone (see Figure 12). When $\theta = 0$ ($\cos^2(\theta) = 1$), z is in L , and we have a perfect fit. When $\theta = 90^\circ$ ($\cos^2(\theta) = 0$), the fit is no better than that obtained by using J alone. The quantity $\cos^2(\theta)$ is called the squared multiple correlation coefficient (R^2) and may be interpreted as is the simple correlation coefficient in (Q2). Note that from (T1)–(T7), $P_L(z) - P_J(z)$ is the projection of $z - P_J(z)$ onto L . Both $z - P_L(z)$ and $P_L(z) - P_J(z)$ are orthogonal to J . Thus $\cos^2(\theta)$ measures the extent to which variations in z ($z - P_J(z)$) may be expressed linearly by variations in x, y, \dots (measured by the nearest point, $P_L(z) - P_J(z)$, in that part of L orthogonal to J).

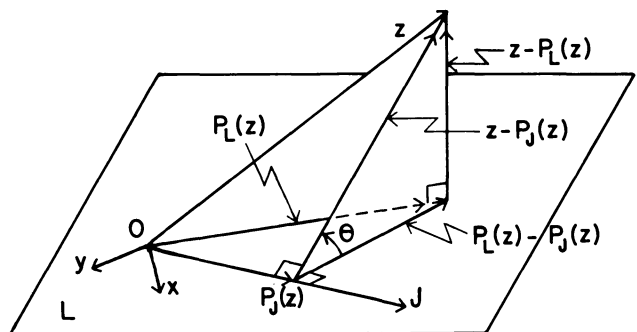


Figure 12. Regression and the multiple correlation coefficient.

All three of our situations (and, indeed, many other situations of classical statistics), can be handled according to our geometric rules, Criteria 1 and 2. The solutions are inspired by geometric principles: The best fit is the projection; they are assessed by considering an angle or a length; they may be best calculated by using the analytic equivalents.

5. PROBABILITY THEORY

In Section 2 we mentioned that we can think of random variables as vectors. We pick up this theme again here. To think about random variables geometrically, we must identify the vectors and the inner product.

The extension to this case (a potentially infinite dimensional vector space) is technically involved. We emphasize the geometric interpretation of the ideas and refer the reader to the references for technical detail.

Suppose we have some random variables z, x, y, \dots that are jointly distributed with distribution function $F(z, x, y, \dots)$. The *expected value* of a random variable is computed as $E(z) = \int z dF(z)$, where $F(\)$ denotes generically the distribution of the variables involved. Suppose our random variables have means $E(z) = \mu_z, E(x) = \mu_x$, and so forth. These random variables are our vectors. We add them, multiply them by constants (scalars), and so on in the conventional way. The quantity

$$\langle z, x \rangle = E(zx) = \int zx dF(z, x)$$

satisfies (2.5), as one can verify by doing the calculus involved; we take it here as our definition of the inner product of two vectors (random variables) z and x . With this definition, many ideas from probability theory correspond to ordinary ideas in geometry.

The constant 1 plays the same role here that J did in Section 4. For example, the projection of a random variable z on 1 must, by Properties A and B, satisfy $P_1(z) = a(1)$ for some a and $\langle z - P_1(z), 1 \rangle = 0$, which reduce to $E(z) = a$, so that $\mu_z(1) = \mu_z$ is the projection of z onto 1. The vector $z - \mu_z$ is the "centered" random variable. It plays the same role here as the variations $z - \bar{z}$ did in Section 4.

For any random variable $z, \langle z, z \rangle = E(z^2)$ is the squared length of z . The squared length of the corresponding centered random variable $z - \mu_z$ is given then by $|z - \mu_z|^2 = E(z - \mu_z)^2 = \sigma_z^2$, and this is called the *variance* of z . As in Section 3, the angle θ between the centered vectors $z - \mu_z$ and $x - \mu_x$ plays an important role (see Figure 13):

$$\cos(\theta) = \langle z - \mu_z, x - \mu_x \rangle / [|z - \mu_z| |x - \mu_x|]$$

is easily shown to be the *correlation* between z and x . When we say in probability theory that

$$\text{var}(z + x) = \text{var}(z) + \text{var}(x) + 2\text{cov}(z, x)$$

we are just repeating the law of cosines (2.1). When two random variables are uncorrelated, it is the analog of the corresponding two vectors' being orthogonal, and so on. The geometric interpretations exactly parallel

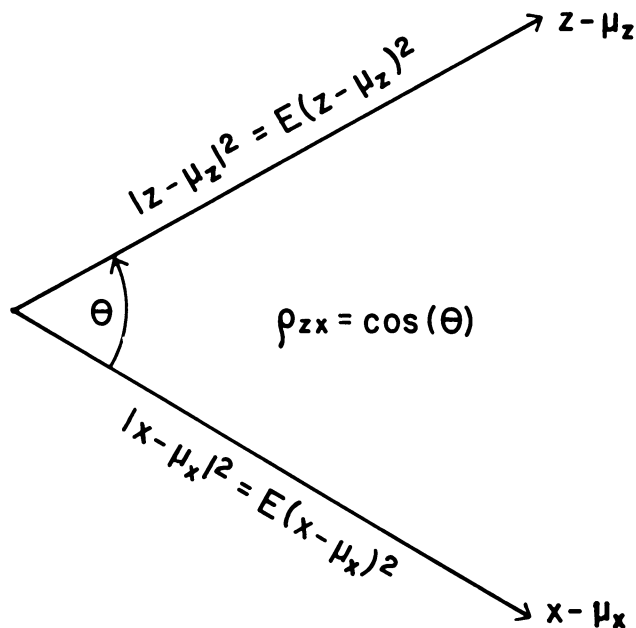


Figure 13. Random variables and correlation.

those for statistics in Section 4. For probability, we interpret a theoretical model. For statistics, we interpret the data themselves. The geometry is the same for both.

We considered above a specific case of projection—onto the constant 1. What is the general analog of projection in probability theory? One answer to this question is the conditional expectation. The expected value of z , given x, y, \dots , is $E(z | x, y, \dots)$ and is similar to the projection of z onto the plane determined by x, y, \dots , in the following way. By the plane $L(x, y, \dots)$, we mean roughly all square integrable functions of x, y, \dots . This space is closed under linear operations, and we may think of it as a linear subspace. Without dwelling on the technicalities here, note three properties of conditional expectation that can be routinely verified from their definitions:

$$E(z | x, y, \dots) \text{ is a function of } x, y, \dots; \quad (5.1)$$

$$E[E(z | x, y, \dots)] = E(z). \quad (5.2)$$

If g is "any" function of x, y, \dots , then

$$E(zg | x, y, \dots) = gE(z | x, y, \dots). \quad (5.3)$$

From (5.2) and (5.3) it follows that if g is "any" function of x, y, \dots , then

$$\langle g, z - E(z | x, y, \dots) \rangle = 0. \quad (5.4)$$

The point of all this is that in terms of Section 3, (5.1) says that $E(z | x, y, \dots)$ is in the plane L , while (5.4) says that $z - E(z | x, y, \dots)$ is orthogonal to any vector g in L . These are exactly the defining conditions for the projection $P_L(z)$. In this sense, then, the conditional expectation is the analog of projection (see Figure 14). Equality in this discussion means equality almost surely, and there are many other technical points to consider. Loève (1963) and Doob (1953) are appropri-

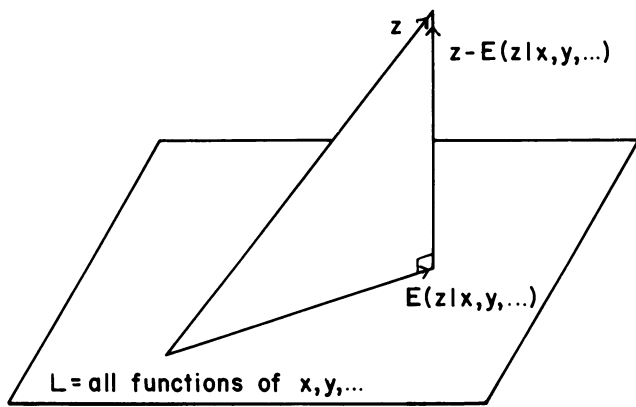


Figure 14. Conditional expectation.

ate references. Our interest here is in what the geometry suggests.

One interpretation comes from the geometric idea that the projection gives the closest point in the plane to the vector z . If, for example, x, y, \dots represent our present knowledge and z represents a future value to be predicted, the analogy suggests that the projection is the function of our present knowledge (x, y, \dots), which comes closest to the future (z). Thus, to predict the future, use $E(z|x, y, \dots)$ —it gives the minimum mean squared error (see Chs. 5 and 9 of Breiman 1969).

Another theorem of probability theory is that

$$\text{var}(z) = \text{var}[E(z|x, y, \dots)] + E[\text{var}(z|x, y, \dots)]. \quad (5.5)$$

Note that since $E[z - E(z|x, y, \dots)] = 0$, we have

$$\text{var}[z - E(z|x, y, \dots)] = E[z - E(z|x, y, \dots)]^2.$$

Since the conditional variance is the variance relative to the conditional distribution, (5.2) further implies that this is also equal to $E[\text{var}(z|x, y, \dots)]$. Thus (5.5) can be written as

$$\text{var}(z) = \text{var}[E(z|x, y, \dots)] + \text{var}[z - E(z|x, y, \dots)],$$

that is, the law of Pythagoras.

6. COMMENTS

Order

I find the order given—geometry, analytic geometry, statistics, and probability—the most effective one, for it builds upon the ideas of a right angle and the law of Pythagoras, which most students remember, if only dimly. The probability can be omitted, although the expression of conditional expectation as a projection allows the statistical (“sample”) ideas to be tied more neatly to their probabilistic (model) equivalents.

Abstraction

Mathematicians like to talk about how things behave, rather than what they are. Nonmathematicians often

prefer the reverse. “Anything that behaves this way is a *vector* (because I said so)” is *not* a natural way of talking to the nonmathematician. Mathematicians ought to be more sensitive to this than they typically are. Repeating examples in different words, emphasizing the equivalence of the various means of expression rather than which expression came “first,” is important. Once students understand the ideas expressed in *some* language, they will be more willing to study the logical paths between them.

The choice of words is important. “Point” and “line” should be used only for the common geometric ideas they convey. “Vector” seems as good as any term for the abstract notion.

Extensions

Sections 2 through 5 form a bare-bones outline of a course. Instructors may add material to it in many ways, depending on time, interest, and student background. For example: (a) When transferred into matrix form, Properties A and B lead directly to the normal equations for linear models, without calculus. (b) Computer subroutines like those suggested by Beaton (1964) extend the material naturally. Each routine performs a function that is a natural computational atom with natural statistical and geometric interpretations as well. Such a combination is particularly effective when one is using interactive computer systems (see Schatzoff, Bryant, and Dempster 1975). (c) The more linear models we discuss (ANOVA, multiple regression, etc.), the more benefits we obtain from the investment in understanding the geometry. For example, to think of degrees of freedom as the dimension of an appropriate subspace will be a big intellectual leap for many, but once it is made, we no longer have to remember separate formulas for many individual cases. As many different models can be introduced as seem relevant: Criteria 1 and 2 apply to them all. Seber’s (1966,1977) books are good starting points. (d) The probability ideas can be considerably expanded in the case of normal distributions (see Dempster 1968, Ch. 14 and Loève 1963, Secs. 33–34). Note also that satisfactory courses can be put together using little or no probability theory, as the Dempster (1968) or Van de Geer (1971) books demonstrate. Such approaches focus attention on the basic statistical quantities and the methods of measuring them, without the mathematical overhead of probability theory. Robert Frost (1935) described writing free verse as “playing tennis with the net down,” and I suppose some would describe statistics without probability in the same way. Yet avoiding probability until the statistical ideas are well established may be pedagogically useful. There is no need for a net until the players understand that the object of the game is to hit the ball back and forth.

Further References

In addition to the references mentioned thus far, Kendall (1961) is a useful reference for formulas. Other

applications are described in Haberman (1978) (log-linear models) and Koopmans (1974) (spectral analysis of time series). Of course, almost any text on multivariate analysis makes at least a passing mention of the geometric interpretation of some quantities, but few exploit it systematically, and none, to my knowledge, do so at an elementary level.

[Received February 1982. Revised August 1983.]

REFERENCES

- BEATON, ALBERT E. (1964), "The Use of Special Matrix Operators in Statistical Calculus," *Research Bulletin*, RB-64-51, Educational Testing Service, Princeton, N.J.
- BREIMAN, LEO (1969), *Probability and Stochastic Processes With a View Toward Applications*, Boston: Houghton Mifflin.
- DEMPSTER, ARTHUR P. (1968), *Elements of Continuous Multivariate Analysis*, Reading, Mass.: Addison-Wesley.
- DOOB, J.L. (1953), *Stochastic Processes*, New York: John Wiley.
- FROST, ROBERT (1935), address at Milton Academy, Milton, Mass., 17 May 1935, in *The Oxford Dictionary of Quotations*, 3rd ed., Oxford: Oxford University Press, 219.
- HABERMAN, SHELBY J. (1978), *Analysis of Qualitative Data*, New York: Academic Press.
- HERR, DAVID G. (1980), "On the History of the Use of Geometry in the General Linear Model," *The American Statistician*, 34, 43-47.
- KENDALL, M.G. (1961), *A Course in the Geometry of n Dimensions*, London: Charles Griffin & Co.
- KOOPMANS, LAMBERT H. (1974), *The Spectral Analysis of Time Series*, New York: Academic Press.
- LOËVE, MICHEL (1963), *Probability Theory*, 3rd ed., Princeton, N.J.: D. Van Nostrand.
- MARGOLIS, MARVIN S. (1979), "Perpendicular Projections and Elementary Statistics," *The American Statistician*, 33, 131-135.
- SCHATZOFF, MARTIN, BRYANT, PETER, and DEMPSTER, ARTHUR P. (1975), "Interactive Statistical Computation With Large Data Structures," in *Perspectives in Biometrics*, ed. R. Elashoff, New York: Academic Press, 1-28.
- SEBER, G.A.F. (1966), *The Linear Hypothesis: A General Theory*, New York: Hafner Press.
- (1977), *Linear Regression Analysis*, New York: John Wiley.
- VAN DE GEER, JOHN P. (1971), *Introduction to Multivariate Analysis for the Social Sciences*, San Francisco: W.H. Freeman & Co.