TRANSACTIONS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 296, Number 2, August 1986

HYPOTHESIS TESTING IN INTEGRAL GEOMETRY

PETER WAKSMAN

ABSTRACT. Probability distributions are defined relative to a fixed plane domain and are calculated explicitly when the domain is a union of coordinate rectangles. The theory of approximating step functions by the resulting special functions gives an interpretation of the problem of guessing a domain given a random sample of observations.

Introduction. The purpose of this paper is to define, following Pohl [2],[1] certain probability distributions associated to plane domains, to calculate the distributions explicitly in the case of domains which are the union of coordinate rectangles, and using these, to consider naive strategies for guessing the shape of a domain given random measurements.

The idea of hypothesis testing is to be contrasted with the idea of direct measurement; rather than identifying an object directly from its measurements, one compares its measurements to those of known objects, thus testing the hypothesis that the unknown is one of the known objects. The best guess is the known object whose measurements are the "closest" to those of the unknown object. This is illustrated in the following example. Suppose we have two nonstandard dice: one with twos on five faces and a one on the sixth face, the other with twos on three faces and ones on three faces; then if someone reported to us the results of tossing one of these dice it would not take long to guess which one it is. The reasoning might be that with each die is associated a known pure probability distribution on the set $\{1, 2\}$, that averaging the results of the tosses gives an empirical distribution which is "close" to one of the pure distributions, and we guess accordingly. This is not to say that we can determine an unknown die in this way—for example we cannot distinguish between two dice with different arrangements of the same number of ones and twos.

One can define (a family of) pure and empirical distributions with a plane domain playing the role of the die and with information collected along a random line playing the role of tossing the die once. Here I calculate these quantities for rectangles and unions of coordinate rectangles and consider the problem of *guessing*

Received by the editors December 17, 1984.
1980 *Mathematics Subject Classification*. Primary 60D05, 62F03; Secondary 53C65.
Key words and phrases. Guessing shape, random lines, statistics.

[1] This paper was given, in a shortened form, at the A.M.S. Summer Research Conference in Integral Geometry at Bowdoin College in July 1984.

0002-9947/86 $1.00 + $.25 per page

shape from amongst this restricted class of domains; this is not the same as trying to determine an unknown domain, although the guessing procedure could be applied to measurements of domains which are not made of rectangles. Even with the restriction to domains made of rectangles, there may be different domains which are indistinguishable by the measurements considered here, and so the result of the guessing procedure may not be unique. Uniqueness is not considered here.

Restricting to domains which are the union of coodinate rectangles is not a severe restriction, and these domains have the advantage of being easy to parametrize when varying continuously, or to discretize. Also comparison of empirical distributions to pure ones would not be possible without the explicit calculation for these domains.

I consider several methods for comparing distributions: the method of maximum likelihood, the methods of mean and median, and the "L^p strategies"; these are meaningful in the context of many observations, but the author cannot resist an unjustified application in the case of just one or two observations. Thus I ask about the "best" square guess given a single random chord length. A more sophisticated analysis [**1**] is possible (with the calculations of this paper) only in the case of rectangles and convex domains; however the unsophisticated methods remain valid in the nonconvex case assuming enough observations.

There are several reasons for a statistical approach to shape recognition. For example we need not remember and store the coordinates of the line giving the random observation; this could be useful for recognizing moving objects, or simply in interpreting partial mesurements. Further, the quantities considered here, involving the intersection of lines with domains, are only one example of the different kinds of quantities which can, and in some cases must, be measured randomly. It is hoped that the approach of hypothesis testing based on explicit calculations for a restricted class of objects (the union of coordinate rectangles here) will have other applications to the study of the measurement of shape.

1. Preliminaries. Let S be the closure of a bounded open domain in $\mathbf{R}^2$ with piecewise smooth boundary C; there is a natural correspondence between oriented lines in the plane meeting the interior of S (denoted $\dot{S}$) and ordered pairs (x, y) of distinct points of C: (x, y) corresponds to the line joining x and y oriented from x to y. Thus the rotation/translation invariant measure dl on the space of oriented lines pulls back to an invariant measure on $\mathscr{C} = C \times C \setminus \{(x, x) \mid x \in C\}$ [**4**]. Let $\mathscr{L}(S)$ be the oriented lines meeting $\dot{S}$. Then from [**4**] we have

PROPOSITION 1.1. *On* $\mathscr{C}$

$$dl = |\sin\theta|\, d\theta\, dx$$

where θ *is the angle between the tangent to* C *at* x *and the line joining* x *and* y, *and* dx *is the arc-length one-form at* $x \in C$.

We can use $\mathscr{C}$ to parametrize $\mathscr{L}(S)$ locally and in some cases (e.g. when S is convex) calculate the total measure of $\mathscr{L}(S)$; dividing by this factor we can study the probability distributions of functions (random variables) on $\mathscr{L}(S)$.

2. Convex domains. Throughout this section assume that S is convex. A well-known formula of Crofton [**4**] describes the total measure of $\mathscr{L}(S)$:

PROPOSITION 2.1 (CROFTON'S FORMULA). *If L is the length of C the total measure of $\mathscr{L}(S)$ is $2L$.*

The proof is a simple application of 1.1.

DEFINITION 2.2. Given a measure space $(X, d\mu)$ with $\mu(X) < \infty$ and a measurable $f\colon X \to R$.

(1) The *cumulative distribution* (c.d.) of f is the function

$$\omega_f(t) = \int_{f \leqslant t} d\mu.$$

(2) The *normalized cumulative distribution* (n.c.d.) of f is $\omega_f(t)/\mu(X)$.

(3) The *probability distribution* (p.d.) of f is

$$\frac{1}{\mu(X)} \frac{d}{dt} \omega_f(t).$$

(this will be defined a.e. $[t]$ in the cases considered here).

Let $t\colon \mathscr{L}(S) \to \mathbf{R}$ assign to each line $l \in \mathscr{L}(S)$ the length of the chord $l \cap \dot{S}$ (note that t never take the value zero). I wish to calculate the above distributions for $f = t$ on $\mathscr{L}(S)$ when S is an $m \times n$ rectangle (by an abuse of notation t is regarded either as a function or as a value of the function).

DEFINITION 2.3. If $0 \leqslant m \leqslant n$ let

$$B_{mn}(t) = t1_{[0,m]} + m1_{[m,n]} + (m + n - t)1_{[n,d]} + n\left(\frac{\sqrt{t^2 - m^2}}{t}\right)1_{[m,d]} + m\left(\frac{\sqrt{t^2 - n^2}}{t}\right)1_{[n,d]},$$

where $d = \sqrt{m^2 + n^2}$ and

$$1_{[a,b]} = 1_{[a,b]}(t) = \begin{cases} 1 & \text{if } a \leqslant t \leqslant b, \\ 0 & \text{otherwise.} \end{cases}$$

For $t \geqslant d$, $B_{mn}(t)$ is defined to be constant at $B_{mn}(d)$.

THEOREM 2.4. *The (c.d.) of t on $\mathscr{L}(S)$ is $4B_{mn}(t)$; its (n.c.d.) is $B_{mn}/(m + n)$ and its (p.d.) is $(d/dt)B_{mn}/(m + n)$.*

PROOF. One calculates directly using 1.1 and letting x vary along two consecutive sides of the rectangle.

The formula for $(d/dt)B_{mn}$ is simpler; for reference define $\beta_{mn} = (d/dt)B_{mn}$—i.e.,

$$\beta_{mn}(t) = 1_{[0,m]} - 1_{[n,d]} + n\left(\frac{\sqrt{t^2 - m^2}}{t}\right)'1_{[m,d]} + m\left(\frac{\sqrt{t^2 - n^2}}{t}\right)'1_{[n,d]}. \tag{2.4.1}$$

REMARKS 2.4.2. (a) $\beta_{mn} = \beta_{nm}$ and so also $B_{mn} = B_{nm}$ (since $\lim_{t \to 0} b_{mn}(t) = 0$).

(b) $\beta_{0,k} = \beta_{k,0} = 0$ and so also for $B_{0,k}$ and $B_{k,0}$.

(c) $\beta_{mn} \geqslant 0$ so B_{mn} is increasing.

(d) $\int_0^\infty \beta_{mn} = B_{mn}(d) = m + n$.

The (p.d.) $\beta_{mn}/(m + n)$ is a curious looking distribution; it is not even continuous. Figure 1 shows the graphs of $4B_{mn}$ and $4\beta_{mn}$ for $(m, n) = (1, 1), (1, 2)$ and $(2, 2)$.

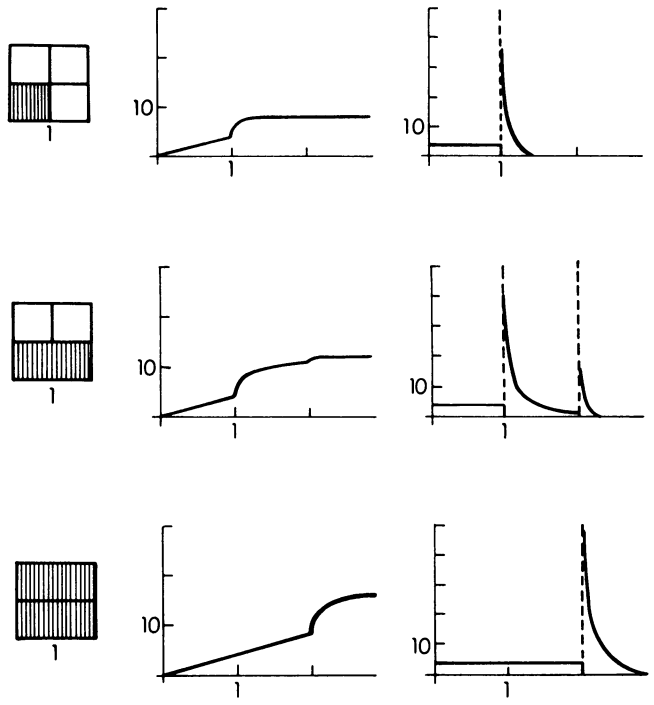

FIGURE 1. Showing an $m \times n$ rectangle S, $B_S = 4B_{mn}$ and $\beta_S = (d/dt)B_S$.

Thus if S is an $m \times n$ rectangle the probability a random line meeting S is a chord of length $\leqslant t$ is $B_{mn}(t)/(m+n)$.

It is sometimes convenient to consider a "viewing screen" V which is a domain containing S. A convention for lines meeting $\dot{V}$ which do not meet $\dot{S}$ is that they have chords of length zero, that is these lines are counted when considering the measure of lines whose chord lengths are $\leqslant t$. Thus, for example, if V is a $k \times l$ rectangle containing the $m \times n$ rectangle S the probability that a random line meeting $\dot{V}$ meets $\dot{S}$ in a chord of length $\leqslant t$ is

$$(h + 4B_{mn}(t))/4(k+l) \tag{2.5.1}$$

where $h = 4(k+l) - 4(m+n)$ is the measure of lines not meeting $\dot{S}$.

The corresponding (p.d.) is

$$(h\delta_0 + 4\beta_{mn}(t))/4(k+l) \tag{2.5.2}$$

where δ_0 is the Dirac point mass at zero. Since

$$B_{mn}/(m+n) \neq (h + 4B_{mn})/4(k+l),$$

probability results will differ depending on whether S is contained in a "viewing screen".

3. The glance problem for convex domains. To "glance" at a domain S means to consider how a random line meets $\dot{S}$; if S is convex the information contained in this glance is the length of $l \cap \dot{S}$ (this will be formalized later). The central problem

considered in this paper is that of *guessing a domain S given one or several "glances"*. By analogy with the dice problem of the introduction a "glance" corresponds to tossing a die, and the restricted problem I consider is to compare the known distributions $B_{mn}/(m+n)$ with the observed distribution of glance lengths.

I take it that a single "glance" meeting a convex domain S in a length a corresponds to the "cumulative distribution"

$$H_a(t) = \begin{cases} 1 & \text{if } t \geqslant a, \\ 0 & \text{otherwise.} \end{cases}$$

Also given glances of lengths $a_1 < a_2 < \cdots < a_k$ (always assume distinct glance lengths) the corresponding distribution is

$$H_{a_1 a_2 \cdots a_k}(t) = \frac{1}{k} \sum_{i=1}^{k} H_{a_i}(t)$$

(differentiating these, the probability distributions are weighted sums of Dirac point masses).

Methods of maximum likelihood. Based on a hypothesis one can calculate the pure probability of an observed event—called the *likelihood* of the event under the given hypothesis. If that probability is small, yet the event occurs, the indication is that the hypothesis is bad. Conversely, the principle of maximum likelihood [3] says that the best hypothesis is that which maximizes the likelihood of what is observed.

There are different versions of what event is being observed given a *single* glance of length a. One version is that we are observing a length in the interval $(a, a+\Delta x)$. The likelihood of this event under the hypothesis of an $m \times n$ rectangle is

$$B_{mn}(a+\Delta x)/(m+n) - B_{mn}(a)/(m+n).$$

Dividing by Δx and doing the unjustified interchanging of the limiting as $\Delta x \to 0$ with the maximization, we seek to maximize $\beta_{mn}(a)/(m+n)$. If $m = n$ one sees that this maximum *cannot be attained* because β_{mm} has a pole. Set $\hat{m}$ = the value of $t > m$ where $\beta_{mm}(t)/2m = \frac{1}{2}$, any m for which $m < a < \hat{m}$ yields a larger value than $\frac{1}{2}$ and we should take m as close as possible to a but still strictly less than a. On the other hand since

$$\frac{\partial}{\partial m} \frac{\beta_{mm}(a)}{2m} = \begin{cases} > 0 & \text{if } m < a < \sqrt{2m}, \\ < 0 & \text{if } 0 < a < m, \end{cases}$$

it seems reasonable to take $m = a$.

It is more difficult to get an analogous answer given *independent* glances of lengths $a_1 < a_2 < \cdots < a_k$, where we seek to maximize

$$\prod_{i=1}^{k} \frac{\beta_{mn}(a_i)}{m+n}. \tag{3.1}$$

It would be nice to clarify this method. Note that it views observations as functionals on the hypotheses involving evaluation of the distribution at the observed points a_i, and that the functional should be maximized.

Another version of the event being observed, given a glance of length a is that *for each fixed t* we are observing the truth or falsity (success or failure) of the statement:

"the chord length is $\leqslant t$". Given glances of lengths $a_1 < a_2 < \cdots < a_k$ we have for $a_i < t < a_{i+1}$ that we are observing i out of k successes. Since $B_{mn}(t)/(m+n)$ is the probability of one success, for each fixed t, the binomial trials formula tells us the probability of i out of k successes is

$$\binom{k}{i}\left(\frac{B_{mn}(t)}{m+n}\right)^i\left(1-\frac{B_{mn}(t)}{m+n}\right)^{k-i}.$$

Let $G = kH_{a_1 a_2 \cdots a_k}(t)$; for each t we are observing $G(t)$ successes, thus as a function of t, the probability to be maximized is

$$\binom{k}{G(t)}\left(\frac{B_{mn}(t)}{m+n}\right)^{G(t)}\left(1-\frac{B_{mn}(t)}{m+n}\right)^{k-G(t)}. \tag{3.2}$$

The unclear part of this method is as to how to simultaneously maximize (3.2) for all t. We can integrate in $(0, d)$ where $d = \sqrt{m^2+n^2}$ and use dt or some convenient measure. The interested reader could consider (3.2) when $k \to \infty$.

In the case of a single glance ($k = 1$) of length a we seek, by this method, to maximize

$$\int_0^a 1 - \frac{B_{mn}}{m+n}\,dt + \int_a^d \frac{B_{mn}}{m+n}\,dt$$

which is equivalent to *minimizing* the L^1 norm $\|B_{mn}/(m+n) - H_a\|_1$. For $m = n$ the minimum occurs when

$$0 = \pi/4 - 1 - 2\sec^{-1}(a/m) + \tfrac{1}{2}(a/m)^2 \quad \text{where } 1 < a/m < \sqrt{2}\,.$$

Other methods of comparing distributions. A crude method is to compare means and medians of (p.d.)'s.

PROPOSITION 3.3. (a) *The mean of* $\beta_{mn}/(m+n)$ *is*

$$\int_0^\infty t\frac{\beta_{mn}}{m+n}\,dt = \frac{\pi}{2}\frac{mn}{m+n}.$$

(b) *The median of* $\beta_{mn}/(m+n)$ *is the solution to* $B_{mn}(x)/(m+n) = \frac{1}{2}$ *which occurs when* $x = 2mn/\sqrt{3n^2+2mn-m^2}$; *which is equal to m when* $m = n$.

The proof of (a) is in (5.4) below; (b) follows by a direct calculation.

Thus by these methods one takes the mean or median of the $\{a_i\}$ and chooses a square (or rectangle) whose distribution has, at least, the same mean or median.

L^p methods. Given an $m \times n$ rectangle being observed, the $H_{a_1 \cdots a_k}$ converge pointwise to $B_{mn}/(m+n)$ by the law of large numbers, i.e. by definition of "random line". The convergence is dominated by 1 and so for any $0 < p < \infty$ the convergence is also in $L^p(0, d)$. For $p = \infty$ the convergence is guaranteed by the Cantelli Lemma [**1**]. Thus to *minimize*, over all m and n, the quantity

$$\left\| B_{mn}/(m+n) - H_{a_1 a_2 \cdots a_k} \right\|_p \qquad (0 < p \leqslant \infty)$$

is a reasonable procedure for guessing rectangular shape, given large k.

L^∞ strategy. To approximate increasing step functions by increasing continuous functions, in L^∞, we use the following:

LEMMA 3.4. *If H is an increasing step function with steps of uniform height $\varepsilon > 0$, then for any increasing continuous function f*

$$\|f - H\|_\infty \geq \frac{\varepsilon}{2}$$

with equality iff $f(x) = (H(x^+) + H(x^-))/2$ for all discontinuities x of H. (So the L^∞ strategy can be viewed as an interpolation problem.)

COROLLARY 3.5. *Given a single glance of length a*

$$\|B_{mm}/2m - H_a\|_\infty \geq \tfrac{1}{2} \quad \text{with equality iff} \quad B_{mm}(a)/2m = \tfrac{1}{2}.$$

Since $(B_{mm}/2m)^{-1}(\frac{1}{2}) = m$, this method says the best guess is with $m = a$; this is the same answer as for the first version of maximum likelihood and for the method of the median.

Given glances of lengths $a < b$ and the problem of finding the best rectangular guess, using the L^∞ strategy, there is a $0 < \delta < b$ such that the minimum of $\frac{1}{4}$ is not achieved by

$$\|B_{mn}/(m+n) - H_{a,b}\|_\infty \quad \text{if } 0 < a < \delta.$$

Thus the interpolation problem is, at best, solved approximately.

It is worth comparing L^p distances to H_a for two domains contained in the 1×2 rectangle: S_1 = the 1×1 square and S_1 = the 1×2 rectangle itself. By (2.5) the distributions are $(1 + B_{11})/3$ for S_1 and $B_{12}/3$ for S_2.

A glance of length zero cannot occur for the rectangle and a glance of length greater than $\sqrt{2}$ cannot occur for the square. So it is desired of the L^p strategy that we have

$$\left\| \frac{1 + B_{11}}{3} - H_a \right\|_p \lesseqgtr \left\| \frac{B_{12}}{3} - H_a \right\|_p \quad \begin{matrix} \text{if } a = 0, \\ \text{if } a > \sqrt{2}. \end{matrix}$$

This is true for all p, but the a for which the inequalities reverse depends on p.

In conclusion, we have the two methods of maximum likelihood as well as the methods of mean or median, and also the L^p methods. They are all justified for a large number of observations but are ad hoc when applied to few observations. Some of these methods agree for a single observation. In principle, given the explicit form of the B_{mn}'s, a sophisticated analysis is possible [**1**] even for a small number of observations.

4. Nonconvex domains. There are several ways that the function t on $\mathscr{L}(S)$ can be generalized to nonconvex domains S. For example, the length of the single component of $l \cap \dot{S}$ can be replaced by the sum of the lengths of the components of $l \cap \dot{S}$. This function from $\mathscr{L}(S)$ to R is called the "Radon tranform" of S. It is difficult to

compute examples of the distribution of this random variable although in some cases it is easy to measure (for example CAT scanning). I consider a different generalization.

Let (a_1, b_1), $(a_2, b_2), \ldots, (a_q, b_q)$ be disjoint intervals on an oriented line with $a_i < b_i < a_{i+1} < b_{i+1}$.

DEFINITION 4.1.

$$\tau(a_i, b_j) = \tau(b_j, a_i) = +1, \qquad \tau(a_i, a_j) = \tau(b_i, b_j) = \begin{cases} 1 & \text{if } i \neq j, \\ 0 & \text{if } i = j. \end{cases}$$

DEFINITION 4.2.1. Given an oriented line l meeting $\dot{S}$ in intervals (a_i, b_i) where $a_1 < b_1 < a_2 < \cdots < a_q < b_q$ the *t-crossing number* of l (relative to S) is

$$H_{l \cap S}(t) = \sum_{\substack{d(x,y) \leqslant t \\ x \leqslant y \text{ on } l}} \tau(x, y)$$

where $x, y \in \{a_1, b_1, \ldots, a_q, b_q\}$ and $d(x, y)$ is the Euclidean distance between x and y. Thus given a line with intervals of lengths $\alpha_1, \alpha_2, \ldots, \alpha_q$ separated by spaces $\beta_1, \beta_2, \ldots, \beta_{q-1}$ we have that

$$\begin{aligned} H_{l \cap S} &= H_1 + H_2 + \cdots + H_{2q-1} \quad \text{where} \\ H_1 &= H_{\alpha_1} - H_{\alpha_1+\beta_1} + H_{\alpha_1+\beta_1+\alpha_2} - \cdots + H_{\alpha_1+\beta_1+\cdots+\alpha_q} \\ H_2 &= H_{\beta_1} - H_{\beta_1+\alpha_2} + H_{\beta_1+\alpha_2+\beta_2} - \cdots - H_{\beta_1+\alpha_2+\beta_3+\cdots+\alpha_q} \\ &\vdots \\ H_{2q-1} &= H_{\alpha_q}. \end{aligned} \tag{4.2.1.1}$$

DEFINITION 4.2. The *t-crossing number* for a domain S is the function

$$B_S(t) \equiv \int_{\mathscr{L}(S)} H_{l \cap S}(d)\, dl.$$

REMARKS 4.3. (1) If S is convex, B_S agrees with the previous (*non*normalized) (c.d.) of t on $\mathscr{L}(S)$ since

$$\int_{\substack{t(l) \leqslant t \\ l \in (S)}} dl = \int_{\mathscr{L}(S)} H_{t(l)}(t)\, dl = \int_{\mathscr{L}(S)} H_{l \cap S}(t)\, dl.$$

When we normalize, for nonconvex S, the function is not necessarily increasing and so cannot be interpreted as a probability; however for each fixed t *it is the expected value of the t-crossing number* for lines, so it can still be measured statistically. (Figure 2 shows nonnormalized B_S and its derivative for some nonconvex domains.)

(2) Let $X_t(l) = H_{l \cap S}(t)$. X_t is a one-parameter family of integer valued random variables on $\mathscr{L}(S)$ (that is, a *stochastic process*) after normalizing B_S we have the expectation of the random variable as a function of t. $X_t(l)$ is the t-crossing number for a line; for each nonnegative integer i let

$$\omega_t(i) = \frac{\text{measure of lines with } X_t(l) = i}{\text{measure of lines considered}}.$$

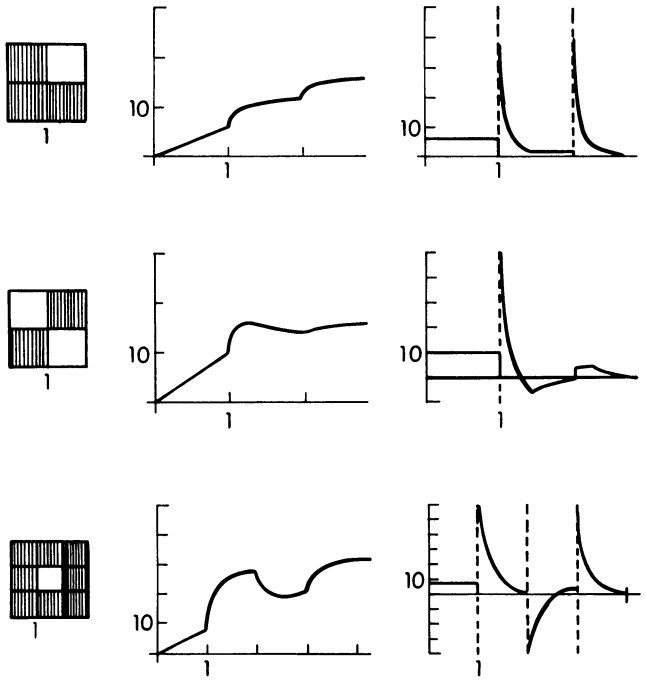

FIGURE 2. Showing examples of nonconvex S, $B_S = \Sigma(\pm)B_{mn}$ and $\beta_S = (d/dt)B_S$. For the top figure $B_S = 4B_{11} + 2B_{22}$, for the middle figure $B_S = 16B_{11} - 8B_{12} + 2B_{22}$, for the bottom figure $B_S = -4B_{11} + 16B_{12} - 8B_{22} + 4B_{33}$.

This is a probability distribution on the nonnegative integers; the natural confusion between convex and nonconvex domains is resolved when we observe that for convex domains, ω_t is supported on $\{0,1\}$ and that such distributions are equivalent to their mean (the expectation of their random variable) by the formula

$$\omega_t(i) = (B_S/2L)^i(1 - B_S/2L)^{1-i} \qquad (i = 0,1).$$

For nonconvex domains there is no such simple formula; nor will ω_t be easy to calculate for nonconvex domains, even polygonal ones where it is a complicated relation between all types of tuples of sides rather than of pairs of sides. However the nonnormalized expectation B_S *is* easy to calculate as we will see. To do maximum likelihood analysis or anything more sophisticated will not be possible using only this expectation; so we will study the method of the mean, making no use of the empirically calculated ω_t except to find its mean, and this amounts to pointwise approximating the observed average t-crossing number by the hypothetical expected value of the t-crossing number: the normalized B_S. In any case, X_t may be of more general interest as a common framework for studying different geometric quantities.

To calculate $B_S(t)$ *for S a union of coordinate rectangles* we take a signed sum of B_{mn}'s over all ordered pairs of corners. We need to distinguish between black corners (three of the quadrants around the corner are white) and white corners (three

surrounding quadrants are black); the case of two black corners with the corner point in common can be treated equally as two white corners.

THEOREM 4.4.

$$B_S(t) = \sum_{\substack{\text{ordered pairs} \\ \text{of corners}}} (\pm) B_{mn}(t)$$

where for an ordered pair of corners, m and n are the horizontal and vertical separation between the corners, and the $(\pm)$ *sign is determined by*

$$(\pm) = \begin{cases} + & \text{if the corners are in the same quadrant } (\mathrm{mod}\, 2), \\ - & \text{if the corners are in different quadrants } (\mathrm{mod}\, 2) \end{cases}$$

if both corners are black, or both are white; the reverse if one is black and the other is white.

PROOF. A function $f_{S_1,S_2}(t)$ depending on an ordered pair of domains S_1 and S_2 is *additive* if whenever $S_1 \cap S_2$ has empty interior

$$f_{S_1 \cup S_2, S_1 \cup S_2} = \sum_{i,j} f_{S_i,S_j}.$$

The theorem is proved by interpreting left- and right-hand sides of (4.4) as of the form $f_{S,S}$ for an additive function f and then observing that additive functions that agree when $S_1 = S_2 =$ rectangle, must also agree when S is a union of coordinate rectangles.

Step 1. Agreement on rectangles: If S is an $m \times n$ rectangle, B_S as defined by (4.2.2) is the (c.d.) of t on $\mathscr{L}(S)$ which is $4B_{mn}$ by (2.4). This is also the right-hand side of (4.4) since there are four ordered pairs of corners separated by (m, n), each of which contributes $+B_{mn}$ to the sum, the other ordered pairs contribute $B_{0,k} = B_{k,0} = 0$.

Step 2. We define $B_{S_1,S_2}(t)$ as in (4.2.2) replacing $H_{l \cap S}$ by

$$H_{l \cap (S_1,S_2)}(t) = \sum_{\substack{d(x,y) \le t \\ x \leqslant y \\ x \in S_1,\, y \in S_2}} \tau(x, y)$$

and note that $B_{S,S} = B_S =$ left-hand side of (4.4). Also define

$$D_{S_1,S_2}(t) = \sum_{\substack{\text{ordered pairs of corners} \\ \text{1st from } S_1,\ \text{2nd from } S_2}} (\pm) B_{mn}$$

and note that $D_{S,S}$ is the right-hand side of (4.4).

Claim. B_{S_1,S_2} and D_{S_1,S_2} are additive.

D_{S_1,S_2} is additive: this is true in general for functions defined by a sum over ordered pairs of corners with this $(\pm)$ sign convention: because the pairs of corners of a union are the union of the pairs of corners, except for those corners (taken care of by the quadrant convention) which disappear in the union.

B_{S_1,S_2} is additive: Fix a line l, to show

$$(*) \qquad H_{l \cap (S_1 \cup S_2)} = H_{l \cap S_1} + H_{l \cap S_2} + H_{l \cap (S_1,S_2)} + H_{l \cap (S_2,S_1)}$$

whenever $S_1 \cap S_2$ has empty interior. Let P be the set of boundary points of $l \cap \dot{S}_1$,

and let Q be the set of boundary points of $l \cap \dot{S}_2$, and let R be the subset of $P \cap Q$ not present in the boundary of $l \cap (\widehat{S_1 \cup S_2})$. For the purposes of defining the sign $\tau(x, y)$, R is viewed in two ways with opposite orientations: as a subset of P or of Q denoted by R_P and R_Q. Now we can expand the left-hand side of (*) as

$$\sum_{\substack{x \in P \setminus R \\ y \in P \setminus R}} \tau(x, y) + \sum_{\substack{x \in Q \setminus R \\ y \in Q \setminus R}} \tau(x, y) + \sum_{\substack{x \in P \setminus R \\ y \in Q \setminus R}} \tau(x, y) + \sum_{\substack{x \in Q \setminus R \\ y \in P \setminus R}} \tau(x, y)$$

(where throughout the sum is defined over $d(x, y) \leqslant t$, $x \leqslant y$ on l). Also we can expand each term on the right-hand side of (*):

$$H_{l \cap S_1} = \sum_{\substack{P \setminus R \\ P \setminus R}} + \sum_{\substack{R_P \\ R_P}} + \sum_{\substack{P \setminus R \\ R_P}} + \sum_{\substack{R_P \\ P \setminus R}},$$

$$H_{l \cap S_2} = \sum_{\substack{Q \setminus R \\ Q \setminus R}} + \sum_{\substack{R_Q \\ R_Q}} + \sum_{\substack{Q \setminus R \\ R_Q}} + \sum_{\substack{R_Q \\ Q \setminus R}},$$

$$H_{l \cap (S_1, S_2)} = \sum_{\substack{P \setminus R \\ Q \setminus R}} + \sum_{\substack{R_P \\ Q \setminus R}} + \sum_{\substack{P \setminus R \\ R_Q}} + \sum_{\substack{R_P \\ R_Q}},$$

$$H_{l \cap (S_2, S_1)} = \sum_{\substack{Q \setminus R \\ P \setminus R}} + \sum_{\substack{R_Q \\ P \setminus R}} + \sum_{\substack{Q \setminus R \\ R_P}} + \sum_{\substack{R_Q \\ R_P}}.$$

Thus (*) is true because the first terms of these expansions of the four right-hand functions match the four terms in the expansion of the left-hand functions, all further terms can be paired so that R_P in one is replaced by R_Q in the other so that these terms cancel. This proves the claim and the theorem.

Thus we can explicitly calculate B_S for S a union of coordinate rectangles. The theorem also says that B_S is a finite linear combination of B_{mn}'s and so the statistical questions will amount to questions about approximating step functions by functions of this special type; that is, we should approximate step functions by special linear combinations of the B_{mn} if we wish to guess shape given a random sample of glances. Before going on to the statistical questions, it is worth considering some integral geometric formulas involving B_S.

5. Integral geometry. (1) If S is convex the "moments" of the Radon transform, also called *Blaschke chord power integrals*, can be defined by

$$I_j = \int_{\mathscr{L}(S)} t(l)^j \, dl.$$

These can be evaluated in terms of $B_S(t) = \int_{\mathscr{L}(S)} H_{t(l)} \, dl$, because the Laplace transform of $H_a(t)$ satisfies

$$\hat{H}_a(S) = e^{-aS} \quad \text{and} \quad (-d/ds)^j \hat{H}_a(0) = a^j.$$

Thus

$$I_j = (-d/ds)^j \hat{B}_S(0).$$

This no longer works when S is not convex; but it is worth asking if the I_j could be calculated in terms of the process X_t of (4.3.2). The question of determining S by these moments in the convex or nonconvex case goes back to Blaschke in the 1930's.

It can be shown that the I_j determine S when S is a sufficiently asymmetric convex polygon [**6**] and something can be said in the nonconvex case also [**5**]; all of which suggest that generically S is determined by B_S. Such questions, the "nonuniqueness" of the introduction, are difficult.

(2) We define a *glance* g to be an oriented line and a finite number of bounded disjoint intervals on the line (as in (4.2.1.1)) and H_g is defined in the same way as $H_{l\cap S}$ in (4.2.1.1). H_g is called the *glance function*. For such a glance g we let diam(g) = maximum t for which H_g has a discontinuity ($\alpha_1 + \beta_1 + \alpha_2 + \cdots + \alpha_q$ in (4.2.1.1))

$n(g)$ = the number of intervals (q in (4.2.1.1)),

$\sigma(g)$ = the total length of the intervals ($\alpha_1 + \alpha_2 + \cdots + \alpha_q$ in (4.2.1.1)).

These quantities are observables given g. Doing some algebra we see that

(a) $H_i \geqslant 0$ so $H_g \geqslant 0$ and also $B_S \geqslant 0$.

(b) $-(d/ds)\hat{H}_g(0) = \sigma(g)$ but in general $(-d/ds)^j\hat{H}_g(0) \neq \sigma^j(g)$.

(c) $\int_0^D H_g(t)\,dt = Dn(g) - \sigma(g)$ for all $D \geqslant$ diam(g).

Thus integrating $H_{l\cap\dot{S}}$ over all lines and using $D =$ diam$(\dot{S})$ we get

$$\int_0^D B_S\,dt = D\int_{\mathscr{L}(S)} n(l)\,dl - \int_{\mathscr{L}(S)} \sigma(l)\,dl$$

where $n(l)$ and $\sigma(l)$ are defined relative to the $l\cap\dot{S}$ glance. By (versions of) well-known formulas of integral geometry [**4**]

$$2L = \int_{\mathscr{L}(S)} n(l)\,dl \quad \text{and} \quad 2\pi A = \int_{\mathscr{L}(S)} \sigma(l)\,dl$$

where A is the area of S. Thus since

$$\int_0^D B_S\,dt = D\cdot 2L - 2\pi A,$$

and since $B_S \geqslant 0$, we get the inequality $D\cdot L \geqslant \pi A$.

(3) Whenever $f_{(m,n)}(t)$ is a function corresponding to an $m\times n$ rectangle and for S a union of coordinate rectangles we define

$$f_S = \sum(\pm)f_{(m,n)}$$

as in (4.4) we always have that

$$\sum(\pm)(m+n) = 2L \quad \text{and} \quad \sum(\pm)(m\cdot n) = 4A.$$

In the case of B_{mn}, these quantities are analytically as well as algebraically present:

$$\lim_{t\to\infty} B_{mn}(t) = \int_0^d \beta_{mn}\,dt = m+n$$

and

$$-\frac{d}{ds}\hat{B}_{mn}(0) = \int_{\mathscr{L}(S)} t(l)\,dl = 2\pi(m\cdot n) \qquad (t(l) = \sigma(l)).$$

Being linear, we can perform these same operations on the sum to compute A and L for S. That is

$$B_S(D) = \sum(\pm)B_{mn}(D) = \sum(\pm)(m+n) = 2L$$

and

$$-\frac{d}{ds}\hat{B}_S(0) = \sum(\pm) - \frac{d}{ds}\hat{B}_{mn}(0) = \sum(\pm)2\pi(mn) = 2\pi(4A) = 8\pi A.$$

(4) The mean of $\beta_S = (d/dt)B_S$ (*without normalization*) is

$$\int_0^\infty t\beta_S = tB_S]_0^D - \int_0^D dS = D \cdot 2L - (D2L - 2\pi A) = 2\pi A.$$

This gives a different method, not using the Laplace transform, for computing the area of S. For the (normalized) (p.d.) $\beta_{mn}/(m+n)$ we calculate the mean as follows: Let S be an $m \times n$ rectangle, so that $\beta_S = 4\beta_{mn}$; then

$$4\int_0^\infty t\beta_{mn} = \int_0^\infty t\beta_S = 2\pi A = 2\pi(mn).$$

Thus we have

$$\int_0^\infty t\frac{\beta_{mn}}{m+n}\,dt = \frac{\pi mn}{2(m+n)}.$$

(5) A final remark about B_S is that it has two nice general properties: (a) being of the form $f_{S.S}$ for an *additive* function f; and (b) being *homogeneous* with $B_{\lambda S}(\lambda t) = \lambda B_S(t)$ where λS is S magnified by a factor of λ. One wonders to what extent B_S could be characterized axiomatically through these properties. It is because they fail for X_t and ω_t that we cannot calculate the latter quantities easily.

6. The general glance problem. To get an expected value of the t-crossing number we must normalize B_S by dividing by the measure of all lines meeting $\dot{S}$. If S is connected (which I assume throughout this section) this quantity is twice the perimeter L^c of the convex hull of S, i.e. we study $B_S(t)/2L^c$.

If we are given glances $g_1, g_2, \ldots, g_k$ we define the *average glance function*

$$H_{g_1 g_2 \cdots g_k} = \frac{1}{k}\sum_{i=1}^{k} H_{g_i}.$$

This is not the only quantity which can be observed, but it is appropriate for comparing with the expected value of the t-crossing number. Since $B_S/2L^c$ and its derivative are not interpreted as probabilities, it makes no sense to use the method of maximum likelihood. For that, the $\omega_t(i)$ would be needed and the other information contained in the observations could be used. The method of the mean (3.3) suggests that for each t the best S is that for which

$$\left|B_S(t)/2L^c - H_{g_1 g_2 \cdots g_k}(t)\right| = 0.$$

We cannot expect to do this for all t. However, the pointwise convergence of $H_{g_1 \cdots g_k}$ to the $B_S/2L^c$ is still bounded, by a function of the maximum glance number $n(g)$, so the convergence is in L^p for $0 < p < \infty$; thus a practical solution, better for large k, is to minimize

$$\left\|B_S(t)/2L^c - H_{g_1 g_2 \cdots g_k}\right\|_p \qquad (0 < p < \infty).$$

We restrict the problem by assuming S is a union of coordinate rectangles.

A still more practical problem is to let $V_{N,K}$ be a square viewing screen with sides of length N, divided into K^2 smaller squares of side length N/K (K a positive integer). For S a connected union of these smaller squares let $h_S = 8N - 2L^c$, and

then the expected value of the t-crossing number for S given a random line meeting the interior of $V_{N,K}$ is $(h_S + B_S)/8N$. The discrete problem, given $H_{g_1 g_2 \cdots g_k}$ (where the glances must satisfy $\operatorname{diam}(g_i) \leqslant \sqrt{2}\,N$) is to minimize

$$\left\| (h_S + B_S)/8N - H_{g_1 g_2 \cdots g_k} \right\|_p$$

when S is a union of the smaller squares.

For small K a computer could do this by direct calculation, trying all possible S. For larger K a better strategy is desirable. For example the best choice in $V_{N,K}$ should guide the choice in $V_{N,K+1}$. At present not enough is known about the B_S to plan a good strategy.

For $p = \infty$ the convergence of $H_{g_1 \cdots g_k}$ to $(h_S + B_S)/8N$ is not guaranteed by the Cantelli Lemma since B_S is not necessarily increasing; nor is the L^∞ minimization directly related to an interpolation problem (as in (3.4)). It might still be of interest to test by computer the value of minimizing

$$\left| \frac{h_S + B_S(t)}{8N} - \frac{H_{g_1 \cdots g_k}(t^+) + H_{g_1 \cdots g_k}(t^-)}{2} \right|$$

over all discontinuities t of $H_{g_1 \cdots g_k}$.

7. Conclusions. The main result is that the expected value of the t-crossing number is a generalization of the distribution of chord lengths which is easy to compute ((2.3) and (4.4)) for domains which are (connected) unions of coordinate rectangles, or to measure statistically by glances. By comparing known distributions with those acquired statistically—in a variety of ways—one tries to guess the shape of a domain. This amounts to finding the best possible approximation to a step function by functions of the form $\Sigma(\pm)B_{mn}$. These methods are not justified for a small number of observations, yet could be tested by computer. In any case, one is motivated to seek a better understanding of the function theory of the B_S.

References

1. T. S. Ferguson, *Mathematical statistics, a decision theoretic approach*, Academic Press, New York and London, 1967.

2. W. F. Pohl, *The probability of linking of random closed curves*, Geometry Symposium, Utrecht, Lecture Notes in Math., vol. 894, Springer Verlag, Berlin, Heidelberg and New York. 1980.

3. G. Polya, *Patterns of plausible inference*, Princeton Univ. Press, Princeton, N. J., 1954, p. 84.

4. L. A. Santalo, *Integral geometry and geometric probability*, Encyclopedia of Mathematics, vol. 1, Addison-Wesley, Reading, Mass., 1976.

5. P. Waksman, *The associated function of a plane polygon*, Ph.D. Dissertation, Univ. of Minnesota, 1983.

6. _____, *Plane polygons and a conjecture of Blaschke's*, J. Appl. Probab. (to appear).

Department of Mathematics, University of Southern California, Los Angeles, California 90089 - 1113