ON THE MEASURE OF A RANDOM SET

By H. E. Robbins

Post Graduate School, U. S. Naval Academy

1. Introduction. The following is perhaps the simplest non-trivial example of the type of problem to be considered in this paper. On the real number axis let N points x_i ($i = 1, 2, \cdots, N$) be chosen independently and by the same random process, so that the probability that x_i shall lie to the left of any point x is a given function of x,

$$\sigma(x) = \Pr(x_i < x). \tag{1}$$

With the points x_i as centers, N unit intervals are drawn. Let X denote the set-theoretical sum of the N intervals, and let $\mu(X)$ denote the linear measure of X. Then $\mu(X)$ will be a chance variable whose values may range from 1 to N, and whose probability distribution is completely determined by $\sigma(x)$. Let $\tau(u)$ denote the probability that $\mu(X)$ be less than u. Then by definition, the expected value of $\mu(X)$ is

$$E(\mu(X)) = \int_1^N u\, d\tau(u), \tag{2}$$

where

$$\tau(u) = \Pr(\mu(X) < u). \tag{3}$$

The problem is to transform the expression for $E(\mu(X))$ so that its value may be computed in terms of the given function $\sigma(x)$.

In order to do this, we observe that, since the x_i are independent,

$$\tau(u) = \int_{C(u)} \cdots \int d\sigma(x_1) \cdots d\sigma(x_N), \tag{4}$$

where the domain of integration $C(u)$ consists of all points $(x_1, \cdots, x_N)$ in Euclidean N-dimensional space such that the linear measure of the set-theoretical sum of N unit intervals with centers at the points x_i is less than u. Here, however, a difficulty arises. Due to the possible overlapping of the intervals, the geometrical description of the domain $C(u)$ is such as to make the explicit evaluation of the integral (4) a complicated matter.

The difficulty is even more serious in the analogous problem where instead of N unit intervals on the line we have N unit circles in the plane, with a given probability distribution for their centers (x_i, y_i). Again we seek the expected value of the measure of the set-theoretical sum of the N circles. The corresponding domain $C(u)$ in $2N$-dimensional space will now be very complicated.

It is the object of this paper to show how, in such cases as these, the expected value of $\mu(X)$ may be found without first finding the distribution function $\tau(u)$.

In fact, the theorem to be stated in (15) will in many important cases yield a comparatively simple formula for $E(\mu(X))$.

2. Expected value of $\mu(X)$. In order to state the problem in full generality, let us suppose that X is a random Lebesgue measurable subset of Euclidean n dimensional space E_n. By this we shall mean that in the space T of all possible values of X there is defined a probability measure $\rho(X)$ so that for every ρ-measurable subset S of T, the probability that X shall belong to S is given by the Lebesgue-Stieltjes integral

$$\text{(5)} \qquad \Pr (X \epsilon S) = \int_T C_S(X)\, d\rho(X),$$

where the integrand is the characteristic function of S,

$$\text{(6)} \qquad C_S(X) = \begin{cases} 1 & \text{for } X \epsilon S \\ 0 & \text{for } X \notin S. \end{cases}$$

In practice, the set X will be a function of a finite number of real parameters (e.g., the coordinates of the centers of the intervals or circles considered in the Introduction), $X = X(\alpha_1, \cdots, \alpha_r) = X(\alpha)$. There will be given a probability measure $\nu(\alpha)$ in the parameter space E_r, so that α will be a vector random variable in the ordinary sense. If A is any ν-measurable subset of E_r, then by definition,

$$\text{(7)} \qquad \Pr (\alpha \epsilon A) = \int_{E_r} C_A(\alpha)\, d\nu(\alpha).$$

Now for the set S' consisting of all X such that $X = X(\alpha)$ for α in A, we define $\rho(S') = \nu(A)$. Thus a ρ-measure is defined in the space T of X, which is the general situation considered in the preceding paragraph.

Returning to the general case described in the first paragraph of this section, we shall now prove the main theorem of this paper. To this end we define, for every point x of E_n and every set X of T, the function

$$\text{(8)} \qquad g(x, X) = \begin{cases} 1 & \text{for } x \epsilon X \\ 0 & \text{for } x \notin X. \end{cases}$$

Moreover, for every x in E_n we let $S(x)$ denote the set of all X in T which contain x. Then for every x in E_n we have from (6),

$$\text{(9)} \qquad g(x, X) = C_{S(x)}(X).$$

Let us denote the Lebesgue measure in E_n of the set X by $\mu(X)$. Assuming that the function $g(x, X)$ is a $\mu\rho$-measurable function of the pair (x, X) in the product space[1] of E_n with T, it follows from Fubini's theorem[1] that

$$\text{(10)} \qquad \int_{E_n \times T} g(x, X)\, d\mu\rho(x, X) = \int_{E_n} \int_T g(x, X)\, d\rho(X)\, d\mu(x).$$

[1] See S. Saks, *Theory of the Integral*, G. E. Stechert, N. Y., 1937, pp. 86, 87.

From (5) and (9) it follows that

$$\int_T g(x, X)\, d\rho(X) = \Pr\,(X \epsilon S(x)) = \Pr\,(x \epsilon X). \tag{11}$$

Again by Fubini's theorem we have

$$\int_{E_n \times T} g(x, X)\, d\mu\rho(x, X) = \int_T \int_{E_n} g(x, X)\, d\mu(x)\, d\rho(X). \tag{12}$$

But from (8),

$$\int_{E_n} g(x, X)\, d\mu(.c) = \int_X d\mu(x) = \mu(X). \tag{13}$$

Now from (10), (11), (12), and (13) we have

$$\int_{E_n} \Pr\,(x \epsilon X)\, d\mu(x) = \int_T \mu(X)\, d\rho(X). \tag{14}$$

But the latter integral is equal to $E(\mu(X))$. Hence we have the relation

$$E(\mu(X)) = \int_{E_n} \Pr\,(x \epsilon X)\, d\mu(x). \tag{15}$$

This is our fundamental result. We may state it as a

THEOREM: *Let X be a random Lebesgue measurable subset of E_n, with measure $\mu(X)$. For any point x of E_n let $p(x) = \Pr\,(x \epsilon X)$. Then, assuming that the function $g(x, X)$ defined by (8) is a measurable function of the pair (x, X), the expected value of the measure of X will be given by the Lebesgue integral of the function $p(x)$ over E_n.*

3. Higher moments of $\mu(X)$. We may generalize the result (15) to obtain similar expressions for the higher moments of $\mu(X)$. For the second moment we have the expression

$$E(\mu^2(X)) = \int_T \mu^2(X)\, d\rho(X). \tag{16}$$

Now from (13),

$$\begin{aligned} \mu^2(X) = \mu(X)\cdot\mu(X) &= \int_{E_n} g(x, X)\, d\mu(x) \cdot \int_{E_n} g(y, X)\, d\mu(y) \\ &= \int_{E_n} \int_{E_n} g(x, X)\cdot g(y, X)\, d\mu(x)\, d\mu(y). \end{aligned} \tag{17}$$

Let

$$g(x, y, X) = g(x, X)\cdot g(y, X) \begin{array}{l} = 1 \text{ if } X \text{ contains both } x \text{ and } y \\ = 0 \text{ otherwise.} \end{array} \tag{18}$$

Then from (16), (17), and (18), we have as before by Fubini's theorem,

$$(19)\qquad \begin{aligned} E(\mu^2(X)) &= \int_T \int_{E_n} \int_{E_n} g(x, y, X)\, d\mu(x)\, d\mu(y)\, d\rho(X) \\ &= \int_{E_n} \int_{E_n} \int_T g(x, y, X)\, d\rho(X)\, d\mu(x)\, d\mu(y). \end{aligned}$$

But from (5) and (18) it follows that

$$(20)\qquad \int_T g(x, y, X)\, d\rho(X) = \Pr\ (x \epsilon X \text{ and } y \epsilon X).$$

The latter probability may be denoted by $p(x, y)$. This function will be defined over the Cartesian product, E_{2n}, of E_n with itself. Let $\mu(x, y)$ denote Lebesgue measure in E_{2n}. Then from (19) we have

$$(21)\qquad E(\mu^2(X)) = \int_{E_{2n}} p(x, y)\, d\mu(x, y),$$

where

$$(22)\qquad p(x, y) = \Pr\ (x \epsilon X \text{ and } y \epsilon X).$$

The formula for the mth moment of $\mu(X)$ will clearly be

$$(23)\qquad \text{Exp}\ (\mu^m(X)) = \int_{E_{mn}} p(x_1, x_2, \cdots, x_m)\, d\mu(x_1, x_2, \cdots, x_m),$$

where $\mu(x_1, x_2, \cdots, x_m)$ denotes Lebesgue measure in E_{mn} and where

$$(24)\qquad p(x_1, x_2, \cdots, x_m) = \Pr\ (x_1 \epsilon X \text{ and } x_2 \epsilon X \cdots \text{ and } x_m \epsilon X).$$

In the next section we shall apply formulas (15) and (21) to a specific problem.

4. Let a, p, B be given positive numbers such that $(B + a)p \leq a$ and $a \leq B$. We shall define the random linear point set X as follows. N intervals, each of length a, are chosen independently on the number axis. The probability density function for the center of the ith interval will be assumed to be constant and equal to p/a in the interval $-a/2 \leq x \leq B + (a/2)$; it may be arbitrary outside this interval. The set X is now defined as the intersection of the fixed interval I: $0 \leq x \leq B$ with the variable set-theoretical sum of the N intervals. The hypothesis of (15) is clearly satisfied. The probability that any point x in the interval I shall be contained in the ith interval of length a is clearly $(p/a)a = p$. From this it follows that

$$(25)\qquad \Pr\ (x \epsilon X) = p(x) = \begin{cases} 1 - (1 - p)^N \text{ for } 0 \leq x \leq B \\ 0 \text{ elsewhere.} \end{cases}$$

From (15) it follows that

$$(26)\qquad E(\mu(X)) = \int_0^B p(x)\, dx = B(1 - (1 - p)^N).$$

(The same formula holds in the case where the N intervals of length a are replaced by N circles of area a and I by a plane domain of area B, provided that for every point of the domain the probability of being contained in the ith circle is equal to a constant p. A similar remark holds for spheres in space.)

To evaluate $E(\mu^2(X))$ in the linear case we make use of the identity

$$\text{(27)} \quad \Pr(A \text{ and } B) = \Pr(A) + \Pr(B) + \Pr(\text{neither } A \text{ nor } B) - 1,$$

which holds for any two events A and B. It follows from (27) and (25) that if x and y are any two points of I, then

$$\begin{aligned} p(x, y) &= \Pr(x \epsilon X \text{ and } y \epsilon X) \\ \text{(28)} \qquad &= \Pr(x \epsilon X) + \Pr(y \epsilon X) + \Pr(x \notin X \text{ and } y \notin X) - 1 \\ &= 1 - 2(1 - p)^N + \Pr(x \notin X \text{ and } y \notin X). \end{aligned}$$

Let

$$\text{(29)} \quad h(x, y) = \Pr(x \notin X \text{ and } y \notin X).$$

Then

$$\text{(30)} \quad h(x, y) = \begin{cases} [1 - (p/a)2a]^N = (1 - 2p)^N, & \text{for } |y - x| \geq a \\ [1 - (p/a)(a + |y - x|)]^N = \left(\dfrac{a - ap - p|y - x|}{a}\right)^N, & \\ & \text{for } |y - x| < a. \end{cases}$$

Now from (21), (28), and (29) we have

$$\begin{aligned} E(\mu^2(X)) &= \int_0^B \int_0^B p(x, y)\, dy\, dx \\ \text{(31)} \qquad &= \int_0^B \int_0^B [1 - 2(1 - p)^N + h(x, y)]\, dy\, dx \\ &= B^2[1 - 2(1 - p)^N] + 2 \int_0^B \int_x^B h(x, y)\, dy\, dx. \end{aligned}$$

When the latter integral is evaluated the result is

$$\begin{aligned} E(\mu^2(X)) &= B^2[1 - 2(1 - p)^N] + (B - a)^2(1 - 2p)^N \\ \text{(32)} \qquad &\quad + \frac{2aB(1 - p)^{N+1}}{(N + 1)p} - \frac{2a(B - a)(1 - 2p)^{N+1}}{(N + 1)p} \\ &\quad - \frac{2a^2}{(N + 1)(N + 2)p^2}[(1 - p)^{N+2} - (1 - 2p)^{N+2}]. \end{aligned}$$

Combining this with (26), we find for the variance of $\mu(X)$ the expression

$$\begin{aligned} \sigma^2 &= E(\mu^2(X)) - [E(\mu(X))]^2 \\ \text{(33)} \qquad &= (B - a)^2(1 - 2p)^N - B^2(1 - p)^{2N} + \frac{2aB(1 - p)^{N+1}}{(N + 1)p} \\ &\quad - \frac{2a(B - a)(1 - 2p)^{N+1}}{(N + 1)p} - \frac{2a^2}{(N + 1)(N + 2)p^2}[(1 - p)^{N+2} - (1 - 2p)^{N+2}]. \end{aligned}$$