Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://amstat.tandfonline.com/loi/uasa20

The Geometry of a Two by Two Contingency Table

Stephen E. Fienberg & John P. Gilbert

To cite this article: Stephen E. Fienberg & John P. Gilbert (1970) The Geometry of a Two by Two Contingency Table, Journal of the American Statistical Association, 65:330, 694-701

To link to this article: https://doi.org/10.1080/01621459.1970.10481117

© Journal of the American Statistical Association
June 1970, Volume 65, Number 330
Theory and Methods Section

The Geometry of a Two by Two Contingency Table

STEPHEN E. FIENBERG and JOHN P. GILBERT*

In this article, we discuss ideas about two by two contingency tables in terms of the geometry of the three dimensional simplex. In particular, we derive the loci of points corresponding to certain classes of two by two tables. We relate the geometry to a discussion of the mean square contingency, which is also associated with rejection regions for the chi-square goodness of fit test for independence.

1. INTRODUCTION

Any contingency table can be normalized to have entries that add to one. Thus there is a natural one to one correspondence between the set of two by two tables whose entries add to one and the points of a three-dimensional simplex (tetrahedron). In this article, we discuss some statistical ideas about two by two tables in terms of the geometry of the tetrahedron. In particular, we derive the loci of (a) all points corresponding to tables whose rows and columns are independent, (b) all points corresponding to tables with a given degree of association, and (c) all points corresponding to tables with a fixed set of marginal totals.

An important aspect of the discussion is our ability to illustrate the results by means of two dimensional figures. It is quite easy to construct a three dimensional wire and string model of the tetrahedron, which can be used as a teaching aid, to demonstrate the geometrical interpretation of various testing and estimation procedures.

The corresponding loci for $r \times c$ and multidimensional contingency tables can also be derived using similar techniques [3]. In addition, Fienberg [4] uses these geometric models to provide a general proof for the convergence of the iterative proportional fitting procedure of Deming and Stephan [2], also discussed in [9].

In the final section, using the geometrical structure of the three-dimensional simplex developed here, we discuss the relationship between surfaces of constant association and surfaces of constant mean square contingency and note that the latter can be used to determine rejection regions for the chi-square goodness of fit test for independence.

2. SURFACE OF INDEPENDENCE

We propose to illustrate the three-dimensional simplex by means of three-dimensional areal or normalized barycentric coordinates (for an exposition on such geometry see [1 or 12]). We choose the tetrahedron of reference (see Figure 1) so that $A_1 = (1, 0, 0, 0)$, $A_2 = (0, 1, 0, 0,)$ $A_3 = (0, 0, 1, 0)$, and $A_4 = (0, 0, 0, 1)$ correspond respectively to the tables

* Stephen E. Fienberg is assistant professor of statistics and mathematical biology, University of Chicago. John P. Gilbert is staff statistician, Harvard Computing Center. The authors are grateful to Bradley Efron, Jerome T. Holland, Paul Holland and Frederick Mosteller for helpful suggestions. Research for this article was supported in part at Harvard University by a National Science Foundation grant (GS-341) and at University of Chicago under sponsorship of the Statistics Branch, Office of Naval Research and the Alfred P. Sloan Foundation.

1	0	0	1	0	0	0	0
0	0	0	0	1	0	0	1

. (2.1)

The general point $P=(p_{11}, p_{12}, p_{21}, p_{22})$ corresponds to the general 2×2 table

p_{11}	p_{12}
p_{21}	p_{22}

. (2.2)

There is a one to one correspondence between points in the tetrahedron and population 2×2 tables, although for sample 2×2 tables, the correspondence is with all points which have rational coordinates.

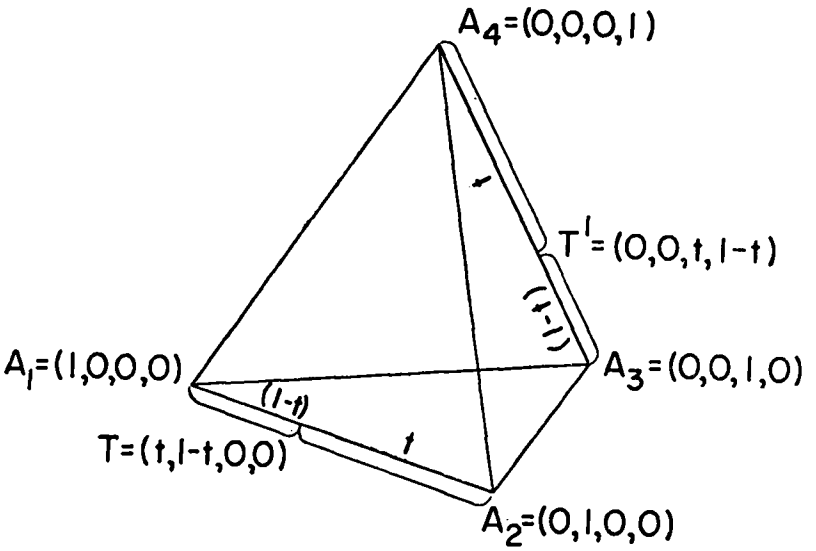
Any point T on the line A_1A_2 is determined by a number t such that $1\geq t\geq0$, and

$$\frac{t}{1-t}=\frac{\overline{A_2T}}{\overline{TA_1}}, \tag{2.3}$$

where $\overline{A_2T}$ and $\overline{TA_1}$ represent the distances between A_2 and T, and T and A_1, respectively. Thus T has coordinates $(t, 1-t, 0, 0)$. We can also choose a point T' on A_3A_4 such that it has coordinates $(0, 0, t, 1-t)$ and

$$\frac{t}{1-t}=\frac{\overline{A_4T'}}{\overline{T'A_3}}. \tag{2.4}$$

Figure 1. TETRAHEDRON OF REFERENCE



Now by definition, any point I on the line TT' within the tetrahedron corresponds to a second number s, such that $1 \geq s \geq 0$, and

$$\frac{s}{1-s} = \frac{\overline{T'I}}{\overline{IT}} \cdot \tag{2.5}$$

I then has coordinates $(st,\ (1-t)s,\ t(1-s),\ (1-s)(1-t))$ and corresponds to the table

st	$s(1-t)$
$(1-s)t$	$(1-s)(1-t)$

(2.6)

whose row and column marginal totals are independent. By allowing s and t to take on all possible values between 0 and 1, we can find all points which correspond to tables whose rows and columns are independent. The lines TT' defined by different values of $t(0 \leq t \leq 1)$ lie on this *surface of independence.*

Alternatively, we might have defined the points S on A_1A_3, S' on A_2A_4, and I' on SS' such that

$$\frac{\overline{SA_3}}{\overline{A_1S}} = \frac{s}{1-s} = \frac{\overline{S'A_4}}{\overline{A_2S'}} \quad \text{and} \quad \frac{t}{1-t} = \frac{\overline{I'S'}}{\overline{SI'}} \cdot \tag{2.7}$$

It is obvious that I' has coordinates $(st,\ s(1-t),\ (1-s)t,\ (1-s)(1-t))$, and that $I' = I$. Thus the lines SS' defined by different values of $s(0 \leq s \leq 1)$ also lie on the *surface of independence.* This surface is completely determined by either family of lines (see Figure 2). In fact the surface of independence is a section of a hyperbolic paraboloid (see [8]) and its saddle-point is at the center of the tetrahedron, $C = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. The hyperbolic paraboloid is a *doubly ruled* surface since its surface contains two families of straight lines or "rulings." The tables corresponding to points on any one of the lines TT' have the same column margins (totals), while the tables corresponding to points on any one of the lines SS' have the same row margins.

3. SURFACE OF CONSTANT ASSOCIATION

Association in a 2×2 table can be defined (and completely determined) by any one of a number of similar measures, see [6]. For the table,

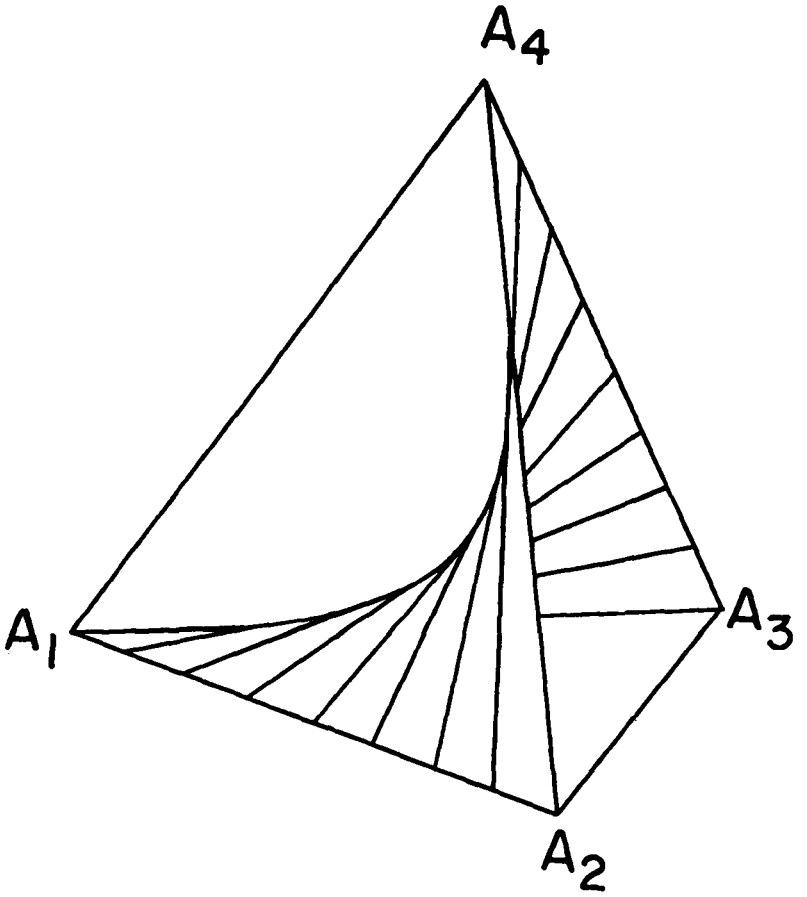
p_{11}	p_{12}
p_{21}	p_{22}

(3.1)

we will define association by the coefficient

$$\alpha = \frac{p_{11}p_{22}}{p_{12}p_{21}}, \tag{3.2}$$

Figure 2. SURFACE OF INDEPENDENCE DEFINED BY FAMILY OF LINES TT'



where $\infty \geq \alpha \geq 0$. The use of α to measure association has been employed by, among others, Goodman [5], Lindley [10], and Mosteller [11]. Note that the row and column marginal totals of a table are independent if and only if $\alpha = 1$. A *positive* association ($\alpha > 1$) where $\alpha = 7$, for example, can be considered as an equal but opposite departure from independence when compared with a negative association ($\alpha < 1$) where $\alpha = 1/7$.

Note that α can be transformed into an equivalent measure

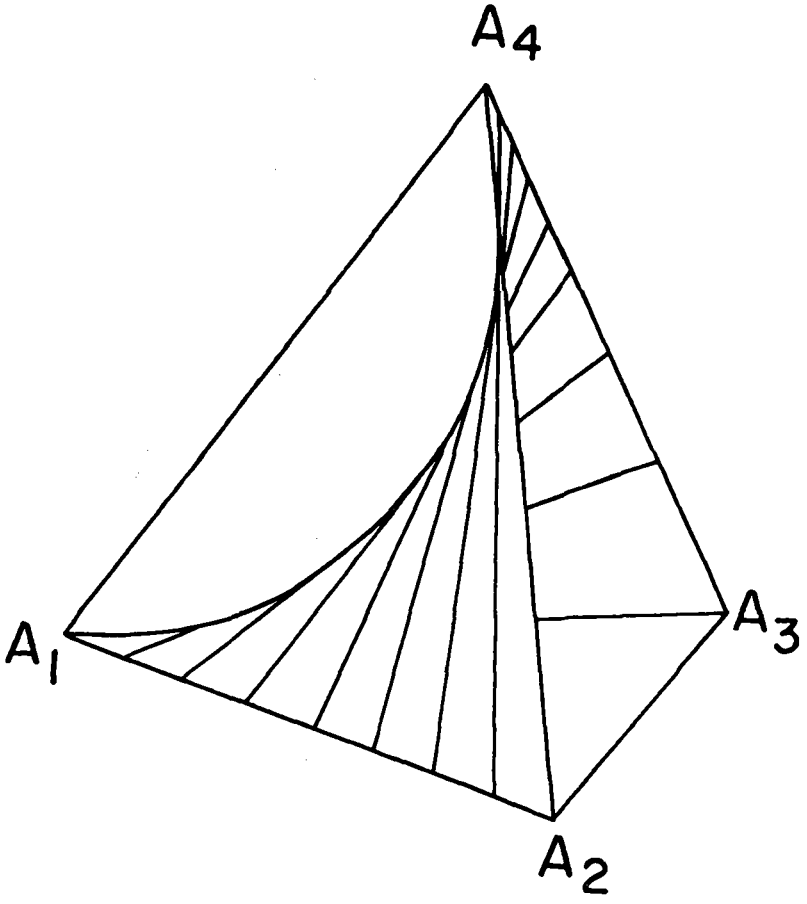
$$\alpha^* = \frac{2}{\pi} \arctan(\ln \alpha)$$

which is symmetric around 0 (the value for independence), positive for *positive* association, negative for *negative* association, and ranges between -1 and $+1$.

We now define a *surface of constant* α, in the same way we defined the *surface of independence*, by choosing T and T^* such that

$$\frac{\overline{TA_2}}{\overline{A_1T}} = \frac{t}{1-t} = \alpha \cdot \frac{\overline{T^*A_4}}{\overline{A_3T^*}} \qquad 0 \leq t \leq 1, \tag{3.3}$$

Figure 3. SURFACE OF CONSTANT $\alpha(\alpha = 3)$ DEFINED BY FAMILY OF LINES TT^*



or by choosing S and S^* such that

$$\frac{\overline{SA_3}}{\overline{A_1S}} = \frac{s}{1-s} = \alpha \cdot \frac{\overline{S^*A_4}}{\overline{A_2S^*}} \qquad 0 \leq s \leq 1. \tag{3.4}$$

The *surfaces of constant* α are completely determined by either one of these two families of straight lines (see Figure 3). These surfaces intersect the *surface of independence* along the four edges of the tetrahedron: A_1A_2, A_1A_3, A_2A_4, and A_3A_4. In fact, the *surfaces of constant* α are sections of hyperboloids of one sheet [8] which are also doubly ruled surfaces. Note that the plane, the hyperbolic paraboloid, and the hyperboloid of one sheet are the *only* doubly ruled surfaces.

A 2×2 table exhibits perfect positive association when all the probability is in the (1, 1) and (2, 2) cells, and $\alpha = \infty$. The locus of such tables is the line segment A_1A_4, which is the limit of the *surfaces of constant* α as α tends to ∞. Similarly, a table with perfect negative association has all its probability in the (1, 2) and (2, 1) cells, and the locus of all such tables is the line segment A_2A_3, which is the limit of the *surfaces of constant* α at α tends to zero.

Finally, the lines TT^* are not the loci of points corresponding to tables with constant column margins, as were the lines TT'. Similarly the lines SS^* do not correspond to tables with constant row margins.

4. TABLES WITH FIXED MARGINS

Tables with both row and column marginal totals fixed are of interest to statisticians because of their use in Fisher's exact test. Therefore, we shall derive the locus of all such tables.

Consider two general points in the tetrahedron,

$$P_1 = (a_1,\ s - a_1,\ t - a_1,\ 1 - s - t + a_1)$$

and

$$P_2 = (a_2,\ s - a_2,\ t - a_2,\ 1 - s - t + a_2),$$

which correspond to two tables with row marginal totals $(t,\ 1-t)$, and column marginal totals $(s,\ 1-s)$. The direction numbers of the line P_1P_2 are given by the vector

$$(1,\ -1,\ -1,\ 1), \tag{4.1}$$

which is independent of a_1 and a_2. Thus all points with the marginal totals $(s,\ 1-s)$ and $(t,\ 1-t)$ lie on the straight line P_1P_2. Moreover, it is clear that P_1P_2 is orthogonal to the lines A_1A_4 and A_2A_3, which have direction numbers

$$(1,\ 0,\ 0,\ -1) \quad \text{and} \quad (0,\ 1,\ -1,\ 0), \tag{4.2}$$

respectively. Thus all tables with the same set of margins correspond to points on a straight line orthogonal to A_1A_4 and A_2A_3. Note that this line of constant margins is parallel to the line connecting the mid points of A_1A_4 and A_2A_3 and, hence, is *not* the line through P perpendicular to the *surface of independence*, unless the margins are all equal.

It is easy to show by simple algebraic manipulation that a specific value of α and a pair of margins completely determine a 2×2 table. The geometric equivalent of this statement is as follows. First we go to the point on the *surface of independence* whose corresponding table has the given margins. Then we move along the straight line through the point orthogonal to A_1A_4 and A_2A_3 until we intersect the surface corresponding to α. This point of intersection then corresponds to the 2×2 table determined by α and the pair of margins.

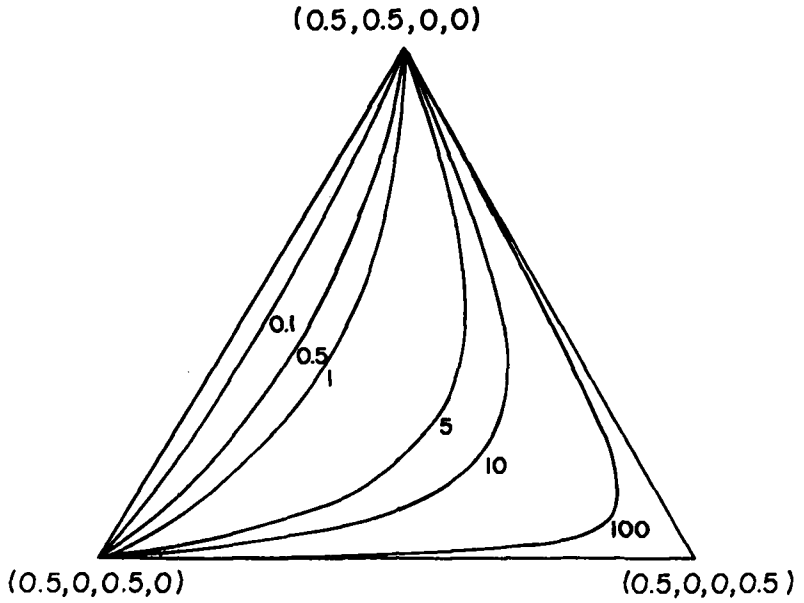
5. ASSOCIATION AND CHI SQUARE

The fact that a test of independence having very desirable properties may be based on the chi-square statistic,

$$X^2 = \frac{(X_{11}X_{22} - X_{12}X_{22})^2 \cdot N}{(X_{11} + X_{12})(X_{21} + X_{22})(X_{11} + X_{21})(X_{12} + X_{22})}, \tag{5.1}$$

where X_{ij} is the observed cell count in the $(i,\ j)$ cell and the sum of the X_{ij} is N, does not necessarily mean that this statistic, or some function of it, is a suitable measure of the degree of association (see [6] and [7]). One can be convinced of the inadequacies of chi-square-like measure by examining the geom-

Figure 4. CONTOURS OF CONSTANT α FOR $p_{11}=0.5$ ($\alpha=100, 10, 5, 1, 0.5, 0.1$)



etry of their contours in the tetrahedron. One widely used chi-square-like measure is

$$\hat{\phi}^2 = X^2/N, \tag{5.2}$$

the sample mean square contingency which was first proposed by Karl Pearson, and which can be considered as a sample estimate of

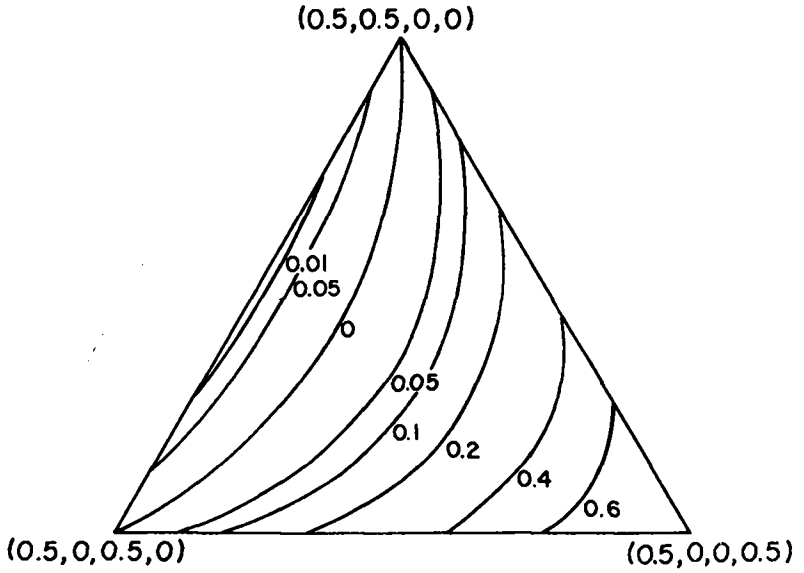
$$\phi^2 = \frac{(p_{11}p_{22} - p_{12}p_{21})^2}{(p_{11} + p_{12})(p_{21} + p_{22})(p_{11} + p_{21})(p_{12} + p_{22})}. \tag{5.3}$$

When $\phi^2=0$, α as defined by (3.2) equals 1, and conversely. However, contours of constant ϕ^2 do not generally correspond to pairs of values of α because ϕ^2 is not independent of the marginal totals, while α is (see [11]).

Figure 4 contains contours of constant α ($\alpha=100$, 10, 5, 1, 0.5, 0.1) for p_{11} equal to 0.5, and Figure 5 contains contours of constant ϕ^2 ($\phi^2=0.6$, 0.4, 0.2, 0.1, 0.05, 0, 0.05, 0.1) for the same value of p_{11}. It is quite clear that near $\phi^2=0$, contours of constant ϕ^2 will closely approximate contours of constant α, except for points very near the faces of the tetrahedron. Such approximations become less satisfactory as ϕ^2 becomes larger.

We can also use the contours of constant ϕ^2 to determine the rejection region of the chi-square test for independence, for fixed values of N. For example, when $N=40$ and $X_{11}=20$, the 0.05 rejection region for the chi-square test statistics is approximately given by the region outside the two contours for $\phi^2=0.1$ which are illustrated in Figure 5. These rejection regions, for the Pearson chi-square test statistic (without a correction for continuity), differ from the

Figure 5. CONTOURS OF CONSTANT MEAN SQUARE CONTINGENCY FOR $p_{11}=0.5$ ($\phi^2=0.6, 0.4, 0.2, 0.1, 0.05, 0, 0.05, 0.1$)



rejection regions for the corresponding likelihood ratio chi-square test statistic when the sample size is small.

REFERENCES

[1] Coxeter, H. S. M., *Introduction to Geometry*, New York: John Wiley & Sons, 1961.

[2] Deming, W. E. and Stephan, F. F., "On a Least Squares Adjustment of a Sample Frequency Table When the Expected Marginal Totals Are Known," *Annals of Mathematical Statistics*, 11 (1940), 427–44.

[3] Fienberg, S. E., "The Geometry of an $r\times c$ Contingency Table," *Annals of Mathematical Statistics*, 39 (1968), 1186–90.

[4] ———, "An Iterative Procedure for Estimation in Contingency Tables," *Annals of Mathematical Statistics*, 41 (1970).

[5] Goodman, L. A., "Simultaneous Confidence Intervals for Cross-Product Ratios in Contingency Tables," *Journal of the Royal Statistical Society*, Series B, 26 (1964), 86–102.

[6] ——— and Kruskal, W., "Measures of Association for Cross Classifications," *Journal of the American Statistical Association*, 49 (1954), 732–64.

[7] ———, "Measures of Association for Cross Classifications, II: Further Discussion and References," *Journal of the American Statistical Association*, 54 (1959), 123–63.

[8] Hilbert, D. and Cohen-Vossen, S., *Geometry and the Imagination*, New York: Chelsea Publishing Co., 1952.

[9] Ireland, C. T. and Kullback, S., "Contingency Tables with Given Marginals," *Biometrika*, 55 (1968), 179–88.

[10] Lindley, D. V., "The Bayesian Analysis of Contingency Tables," *Annals of Mathematical Statistics*, 35 (1964), 1622–43.

[11] Mosteller, F., "Association and Estimation in Contingency Tables," *Journal of the American Statistical Association*, 63 (1968), 1–28.

[12] ——— and Tukey, J. W., "Data Analysis Including Statistics," in Gardner Lindzey and Elliot Anderson, eds., *Handbook of Social Psychology*, Reading, Mass.: Addison-Wesley, (Rev. ed.), 1968.