Statistical Analysis.

Karl Pearson (1904): "On the Theory of Contingency and Its Relation to Association and Normal Correlation." [Draper's Company Research Memoirs Biometric Series I.]

Contingency Table { Cross Tabulation / Cross Tab.

◇ A matrix $\in [0,1]^{m \times n}$ displaying bi-variate (in general, multivariate) frequency distribution of the variables.

	Left-Handed	Right-Handed	Marginal Total
Male	2	3	5
Female	1	4	5
Marginal Total	3	7	10 Grand Total

Q: Are female less likely to be left-handed than right-handed?

◇ Statistical Tests:
- Pearson's chi-squared test
- G-test
- Fisher's Exact test
- Barnard's test.

Normalized Table $\in [0,1]^{2 \times 2}$

$$\bar{s}^2 = \frac{1}{m(m-1)} \sum_{i \neq j} s_{ij}^2$$

$$A = \begin{bmatrix} \frac{1}{5} & \frac{3}{10} \\ \frac{1}{10} & \frac{2}{5} \end{bmatrix} \qquad A^T = \begin{bmatrix} 1/5 & 1/10 \\ 3/10 & 2/5 \end{bmatrix}$$

$$S = AA^T = \begin{bmatrix} 13/100 & 14/100 \\ 14/100 & 17/100 \end{bmatrix} \qquad \bar{s}^2 = \frac{28/100}{2} = 14/100$$

◇ Note, if there was a perfect correlation $A = \cancel{\begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}}\ \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$

and $AA^T = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ and $\bar{s}^2 = \frac{2}{2} = 1$

◇ Note, also, if there was a perfect anticorrelation

$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ then $AA^T = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ $\bar{s}^2 = \frac{0}{2} = 0.$

DARWIN'S FINCHES:

Seymour, Baltra, Isabella, Fernandina, Santiago, Rabida, Pinzon, Santa Cruz, Santa Fe, San Cristobal, Espanola, Florena, Genovesa, Marchena, Pinta, Darwin, Wolf

Finch \ Islands	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
Large Ground Finch	0	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1
Medium Ground F.	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	0	0
Small Ground F.	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0
Cactus Ground F.	1	1	1	0	1	1	1	1	1	1	0	1	0	1	1	0	0
Sharp Beaked G.F.	0	0	1	1	1	0	0	1	0	1	0	1	1	0	1	1	1
Large Cactus G.F.	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0
Large Tree F.	0	0	1	1	1	1	1	1	1	0	0	1	0	1	1	0	0
Medium Tree F.	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Small Tree F.	0	0	1	1	1	1	1	1	1	1	0	1	0	0	1	0	0
Vegetarian F.	0	0	1	1	1	1	1	1	1	1	0	1	0	1	1	0	0
Woodpecker F.	0	0	1	1	1	0	1	1	0	1	0	0	0	0	0	0	0
Mangrove F.	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Werbler F.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

13

Marginal Bounds (Row Sums)

Finch	Row Sums	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
									Island									
Large ground finch	15	0	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1
Medium ground finch	13	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	0	0
Small ground finch	14	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0
Sharp-beaked ground finch	10	0	0	1	1	1	0	0	1	0	1	0	1	1	0	1	1	1
Cactus ground finch	12	1	1	1	0	1	1	1	1	1	1	0	1	0	1	1	0	0
Large cactus ground finch	2	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0
Large tree finch	10	0	0	1	1	1	1	1	1	1	0	0	1	0	1	1	0	0
Medium tree finch	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Small tree finch	10	0	0	1	1	1	1	1	1	1	1	0	1	0	0	1	0	0
Vegetarian finch	11	0	0	1	1	1	1	1	1	1	1	0	1	0	1	1	0	0
Woodpecker finch	6	0	0	1	1	1	0	1	1	0	1	0	0	0	0	0	0	0
Mangrove finch	2	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Warbler finch	17	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Marginal Bounds → column sums		4	4	11	10	10	8	9	10	8	9	3	10	4	7	9	3	3

NOTE: Island name code: A = Seymour, B = Baltra, C = Isabella, D = Fernandina, E = Santiago, F = Rábida, G = Pinzón, H = Santa Cruz, I = Santa Fe, J = San Cristóbal, K = Española, L = Floreana, M = Genovesa, N = Marchena, O = Pinta, P = Darwin, Q = Wolf.

43

Darwin's Data.

$A = 13 \times 17$ matrix over $\{0,1\}$

$= \{0,1\}^{13 \times 17}$

Grand Total $= 123$

Row Sums

~~A(i,·)~~ $A(1\cdot), A(2\cdot), \ldots A(13\cdot)$

Column Sums

$A(\cdot 1), A(\cdot 2), \ldots, A(\cdot 17)$.

Problem:

Data: Row Sums / Column Sums

$A(i\cdot) \quad i = 1, \ldots, m$

$A(\cdot j) \quad j = 1, \ldots, n$

$\sum_i A(i\cdot) = \sum_j A(\cdot j) = T$

Find:

(i) Compute k distinct 0-1 A matrices satisfying the marginals given by row sums and Column sums.

K = 1; 100; 10000.

There are 67,149,106,137,567,626 distinct tables.

(ii) Compute one A matrix that maximizes $\bar{S}^2$.

$$A = (x_{ij})$$

Feasibility

$$\sum_{j=1}^{n} x_{ij} : A(i\cdot) = i^{th} \text{ row sum} \quad i = 1, .., m$$

$$\sum_{i=1}^{m} x_{ij} : A(\cdot j) = j^{th} \text{ column sum} \quad j = 1, .., n$$

$$x_{ij} \in \{0, 1\}$$

Note: Feasibility is in P.
Based on Gale-Ryser Condition:
Using a trick called "Majorization"

$\boxed{\text{Optimization}}$

Note
$$s_{ij} = \sum_k x_{ik} x'_{kj} = \sum_k x_{ik} x_{jk}$$

$$= \sum_k 1_{x_{ik} = x_{jk} = 1}.$$

$$= \sum_k \mathbb{1}_{x_{ik} = x_{jk} = 1}.$$

We wish to maximize $\frac{1}{m(m-1)} \sum_{i \neq j} s_{ij}^2$

simplify to $\sum_{i \neq j} s_{ij}^2 \rightarrow \sum_{i \neq j} w_{ij} s_{ij}$

$$w_{ij} = \cancel{Am}\, A(i\cdot) A(j\cdot)$$

maximize
$$\sum_{i \neq j} A(i\cdot) A(j\cdot) \sum_k \mathbb{1}_{x_{ik} = x_{jk} = 1}$$

$$= \sum_{i \neq j} \sum_k A(i\cdot) A(j\cdot) \mathbb{1}_{x_{ik} = x_{jk} = 1}$$

subject to
$$\sum_{j=1}^{n} x_{ij} = A(i\cdot); \quad \sum_{i=1}^{m} x_{ij} = A(\cdot j);$$

$$x_{ij} \in \{0, 1\}.$$

A General Problem:

$c_1, c_2, \ldots, c_m$ are species.
$x_1, x_2, \ldots, x_n$ are islands.
We say c_i and c_j are symbionts in islands $X' \subseteq \{x_1, \ldots, x_n\}$
iff $x_{ik} \Leftrightarrow x_{jk} \ \forall x_k \in X'$
Both c_i and c_j cohabit islands X'
We say c_i and c_j are antibionts in islands $X' \subseteq \{x_1, \ldots, x_n\}$
iff $\bar{x}_{ik} \vee \bar{x}_{jk} \ \forall x_k \in X'$
$\boxed{x_{ik} \Rightarrow \bar{x}_{jk} \wedge x_{jk} \Rightarrow \bar{x}_{ik}}$
Neither c_i nor c_j cohabits any island in X'
Row Sum $A(i\cdot)$ = # islands occupied by a species i
Column Sum $A(\cdot j)$ = # species occupying island j

Data: m, n;
Symbiont relation $\{(c_i, c_j, X'_{ij})\}$
Antibiont relation $\{(c_i, c_j, X'_{ij})\}$
Row Sum
Column Sum
} CONSTRAINTS

Desiderata: Find a matrix $\in \{0,1\}^{m \times n}$ satisfying all the constraints.

NP Completeness

ONE-IN-THREE POSITIVE 3-SAT

For each clause C_i: Create two species, c_i & c_i'
$(c_i, c_i', \{x_i \mid x_i \in C_i\}) \leftarrow$ Add to Antibiont rel^n.
Row Sum $(c_i) = 1$; RowSum $(c_i') = 2$.
If $x_k \in C_i \cap C_j$
$(c_i, c_j, \{x_k\}) \leftarrow$ Add to symbiont rel^n
Column Sum $(x_k) = |\{C_i \mid x_k \in C_i\}|$

Example:

$$\overset{c_1}{(x \vee y \vee z)} \wedge \overset{c_2}{(x \vee u \vee z)} \wedge \overset{c_3}{(w \vee u \vee v)}$$

	x	u	v	w	y	z	
c_1	1				0	0	1
c_1'	0				1	1	2
c_2	1	0				0	1
c_2'	0	1				1	2
c_3		0	0	1			1
c_3'		1	1	0			2
	2	2	1	1	1	2	
	c_1, c_2	c_2, c_3	c_3	c_3	c_1	c_1, c_2	

Anti $\begin{cases} (c_1, c_1', \{x,y,z\}) \\ (c_2, c_2', \{x,u,z\}) \\ (c_3, c_3', \{w,u,v\}) \end{cases}$

Sym $\begin{cases} (c_1, c_2, \{x,z\}) \\ (c_2, c_3, \{u\}) \end{cases}$

LINEAR PROGRAMMING.

Standard Form:

- A Linear Function to be maximized:

$$f(x_1, x_2) = c_1 x_1 + c_2 x_2 = [c_1 \;\; c_2]\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = C^T x$$

- PROBLEM CONSTRAINTS:

$$\begin{aligned} a_{11}\, x_1 + a_{12} x_2 &\le b_1 \\ a_{21}\, x_1 + a_{22} x_2 &\le b_2 \\ a_{31}\, x_1 + a_{32} x_2 &\le b_3 \end{aligned} \Leftrightarrow \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \le \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

$$\Leftrightarrow \; Ax \le b$$

- Non-negative Variables

$$\begin{matrix} x_1 \ge 0 \\ x_2 \ge 0 \end{matrix} \Leftrightarrow \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \ge \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Leftrightarrow x \ge 0$$

Matrix Form

$$\max \{ c^T x \mid Ax \le b \;\wedge\; x \ge 0 \} \quad \leftarrow \text{Primal}$$

Symmetric Dual Form

$$\min \{ b^T y \mid A^T y \ge c \;\wedge\; y \ge 0 \} \quad \leftarrow \text{Dual}$$

Strong Duality Thm: If primal has an optimal soln x^*, then the dual has an optimal soln y^*, and

$$c^T x^* = b^T y^*.$$

Algorithms: Classical: Dantzig's Simplex Algorithm (Exptime) Fast in practice.

Criss-Cross Algorithm.

Modern: $\left.\begin{array}{l}\text{Ellipsoid Algorithm (Khachiyan)} \\ \text{Projective Algorithm (Karmakar)} \\ \text{Path-following Algorithm}\end{array}\right\}$ Interior Point $\in$ P.

Integer Linear Programming (ILP):

If all of the unknown variables are required to be integers then the problem is called an ILP.

0-1 Integer Programming or Binary Integer Programming (BIP) is a special case of ILP, where variables are required to be 0 or 1.

Mixed Integer Programming (MIP)

Only some of the unknown variables are required to be integers.

3-SAT.

For every variable x_i put

$$0 \leq x_i \leq 1 \qquad\qquad x_i \in \{0, 1\}$$

For every clause $c_j = (x_{i_1} \vee x_{i_2} \vee \bar{x}_{i_3})$ put

$$x_{i_1} + x_{i_2} + (1 - x_{i_3}) > 0 .$$

unnegated literals

Negated literals.

SEMIDEFINITE PROGRAMMING: (SDP)

$X =$ $n\times n$ matrix

X = Positive Semidefinite (psd) if $\forall_{v\in\mathbb{R}^n}$ $v^T X v \geq 0$

X = Positive Definite (pd) if $\forall_{v\in\mathbb{R}^n}$ $v^T X v > 0$.

$X \succeq 0$ $\leftarrow$ X is symmetric & positive Definite.

$$A \bullet B = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} B_{ij} = \mathrm{Tr}(A^T B)$$

SDP: minimize $C \bullet X$

s.t. $A_i \bullet X = b_i$ $\quad i = 1, \dots, m$

$X \succeq 0$.

$$A_i = \begin{pmatrix} a_{i1} & & & 0 \\ & a_{i2} & & \\ & & \ddots & \\ 0 & & & a_{in} \end{pmatrix} \quad i = 1, \dots, m \qquad C = \begin{pmatrix} c_1 & & & 0 \\ & c_2 & & \\ & & \ddots & \\ 0 & & & c_n \end{pmatrix}$$

LP $\equiv$ SDP:

minimize $C \bullet X$

s.t. $A_i X = b_i$ $\quad i = 1, \dots, m$

$X_{ij} = 0$ $\quad i = 1, \dots, m;\ j = i+1, \dots, n$

$X \succeq 0$

$$X = \begin{pmatrix} x_1 & & & 0 \\ & x_2 & & \\ & & \ddots & \\ 0 & & & x_n \end{pmatrix}$$

Quadratically Constrained Quadratic Prog.
QCQP.

$$\underset{x}{\text{minimize}} \quad x^T Q x + q_0^T x + c_0$$

$$\text{s.t.} \quad x^T Q_i x + q_i^T + c_i \le 0 \qquad i=1,\cdots,m.$$

Factor each Q_i

$$Q_i = M_i^T M_i$$

Note

$$\begin{pmatrix} I & M_i x \\ x^T M_i^T & -c_i \, q_i^T x \end{pmatrix} \succeq 0 \iff x^T Q_i x + q_i^T x + c_i \le 0$$

$$\underset{x,\theta}{\text{minimize}} \quad \theta \qquad \text{s.t.}$$

$$\begin{pmatrix} I & M_0 x \\ x^T M_0^T & -c_0 - q_0^T x + \theta \end{pmatrix} \succeq 0$$

$$\begin{pmatrix} I & M_i x \\ x^T M_i^T & -c_i - q_i^T x \end{pmatrix} \succeq 0 \quad i=1,\cdots,m.$$

SDP can be solved in Polytime using interior point method.