A Practical Algorithm for Exact Inference on Tables

Jeffrey W. Miller
Matthew T. Harrison

Brown University
Applied Mathematics

April 16, 2011

NESS 2011, Univ. of Connecticut

Motivation

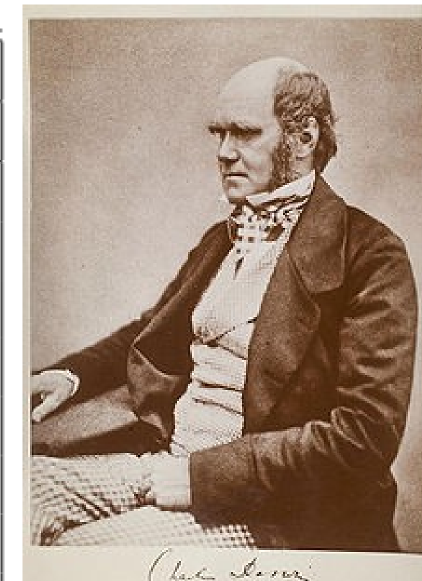
Suppose you are an ecologist, and observe...

Darwin's Finches on the Galápagos Islands

	Island																
Finch	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 finch	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	0	0
Small ground finch	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0
Sharp-beaked ground finch	0	0	1	1	1	0	0	1	0	1	0	1	1	0	1	1	1
Cactus ground finch	1	1	1	0	1	1	1	1	1	1	0	1	0	1	1	0	0
Large cactus ground finch	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0
Large tree finch	0	0	1	1	1	1	1	1	1	0	0	1	0	1	1	0	0
Medium tree finch	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Small tree finch	0	0	1	1	1	1	1	1	1	1	0	1	0	0	1	0	0
Vegetarian finch	0	0	1	1	1	1	1	1	1	1	0	1	0	1	1	0	0
Woodpecker finch	0	0	1	1	1	0	1	1	0	1	0	0	0	0	0	0	0
Mangrove finch	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Warblerfinch	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Islands: 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.

Chen et al. 2005



Motivation

This arrangement of finches seems unusual to you. To analyze it, you formulate a statistical test:

- Null hypothesis: uniform distribution given fixed row and column sums
- Test statistic: $$\bar{S}^2 = \frac{1}{m(m-1)} \sum_{i \neq j} c_{ij}^2,$$

  where m is the number of species, $\mathbf{C} = (c_{ij}) = \mathbf{A}\mathbf{A}^T$

  and $\mathbf{A}$ is the occurrence matrix.
- Estimate the p-value

How to sample?

Motivation

Fixing the row and column sums makes the distribution very complicated.

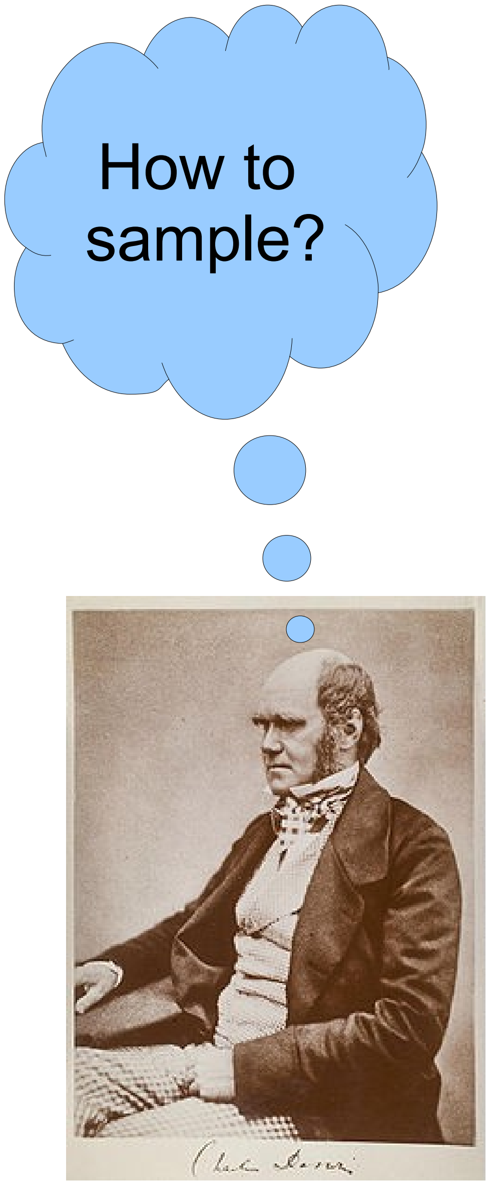
Trivial example:

Row sums = (2, 2, 1, 1), Column sums = (3, 2, 1)

There are 8 such binary matrices:

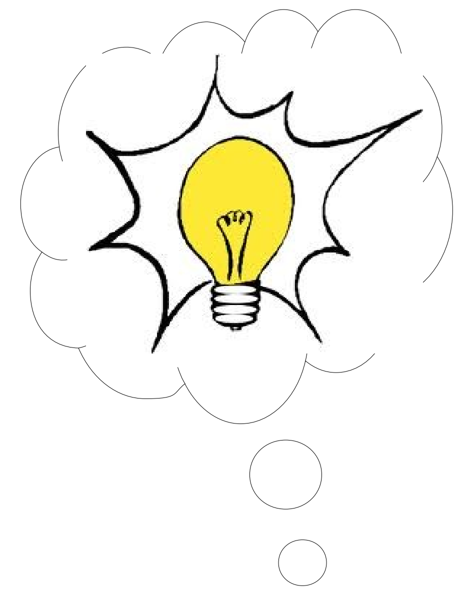
$$\begin{matrix}1&1&0\\0&1&1\\1&0&0\\1&0&0\end{matrix}\qquad\begin{matrix}1&0&1\\1&1&0\\1&0&0\\0&1&0\end{matrix}\qquad\begin{matrix}1&0&1\\1&1&0\\0&1&0\\1&0&0\end{matrix}\qquad\begin{matrix}0&1&1\\1&1&0\\1&0&0\\1&0&0\end{matrix}$$

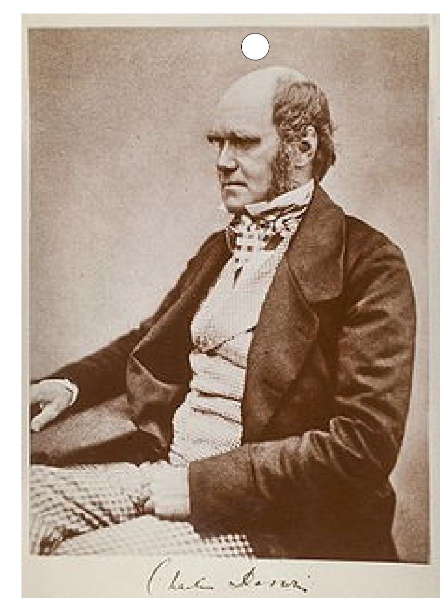
$$\begin{matrix}1&1&0\\1&1&0\\1&0&0\\0&0&1\end{matrix}\qquad\begin{matrix}1&1&0\\1&1&0\\0&0&1\\1&0&0\end{matrix}\qquad\begin{matrix}1&1&0\\1&0&1\\1&0&0\\0&1&0\end{matrix}\qquad\begin{matrix}1&1&0\\1&0&1\\0&1&0\\1&0&0\end{matrix}$$



Background

- This example is typical in ecology.
- Existing sampling methods:
  - Heuristic approximations
  - Markov chain Monte Carlo (MCMC)
  - Sequential Importance Sampling (SIS)
- Issue: Existing theory does not provide adequate guarantees on convergence rates.
- **We offer an exact algorithm that is tractable in many cases.**





Previous Work

- Heuristic approximations

  Connor and Simberloff (1979), Patterson and Atmar (1986), Gilpin and Diamond (1982), Coleman et al. (1982)

- Markov chain Monte Carlo (MCMC)

  Brualdi (1980), Roberts and Stone (1990), Manly (1995), Chen, Diaconis, Holmes, and Liu (2005), (*and many more!*)

- Sequential Importance Sampling (SIS)

  Chen, Diaconis, Holmes, and Liu (2005), Harrison (to appear)

Theory

- There exists a **recursion** that efficiently counts the number of binary matrices with given margins (when the column sums are not too large).
- Main idea: **exploit symmetries**

Theorem

$$\bar{N}(\mathbf{p},\mathbf{r}) = \sum_{\mathbf{s}\in C^{\mathbf{r}}(p_1)} \binom{\mathbf{r}}{\mathbf{s}} \bar{N}(L\mathbf{p},\mathbf{r}\backslash\mathbf{s})$$

- O($mn^b(\log n)^3$) time for *m* x *n* matrices, assuming bounded margins and $b \geq$ max column sum.
- After counting, **sampling** is *easy and fast*.

Example #1: Darwin's finches

We compare results with

Chen, Diaconis, Holmes, and Liu (JASA 2005).

Results for Darwin's Finch Data, using $\bar{S}^2$

Method	# samples	p-value	Sampling time
Exact	1,000,000	$(4.67 \pm .22) \times 10^{-4}$	31 minutes
SIS	1,000,000	$(3.96 \pm .36) \times 10^{-4}$	18 minutes
MCMC	15,000,000	$(3.56 \pm .68) \times 10^{-4}$	18 minutes

Exact 95% Confidence Interval: (4.26, 5.11) x 10^{-4}

Number of matrices with margins as observed:

- Exact number: 67,149,106,137,567,626 (1.4 sec)
- Chen's estimate: 6.7150×10^{16}

Example #1: Darwin's finches

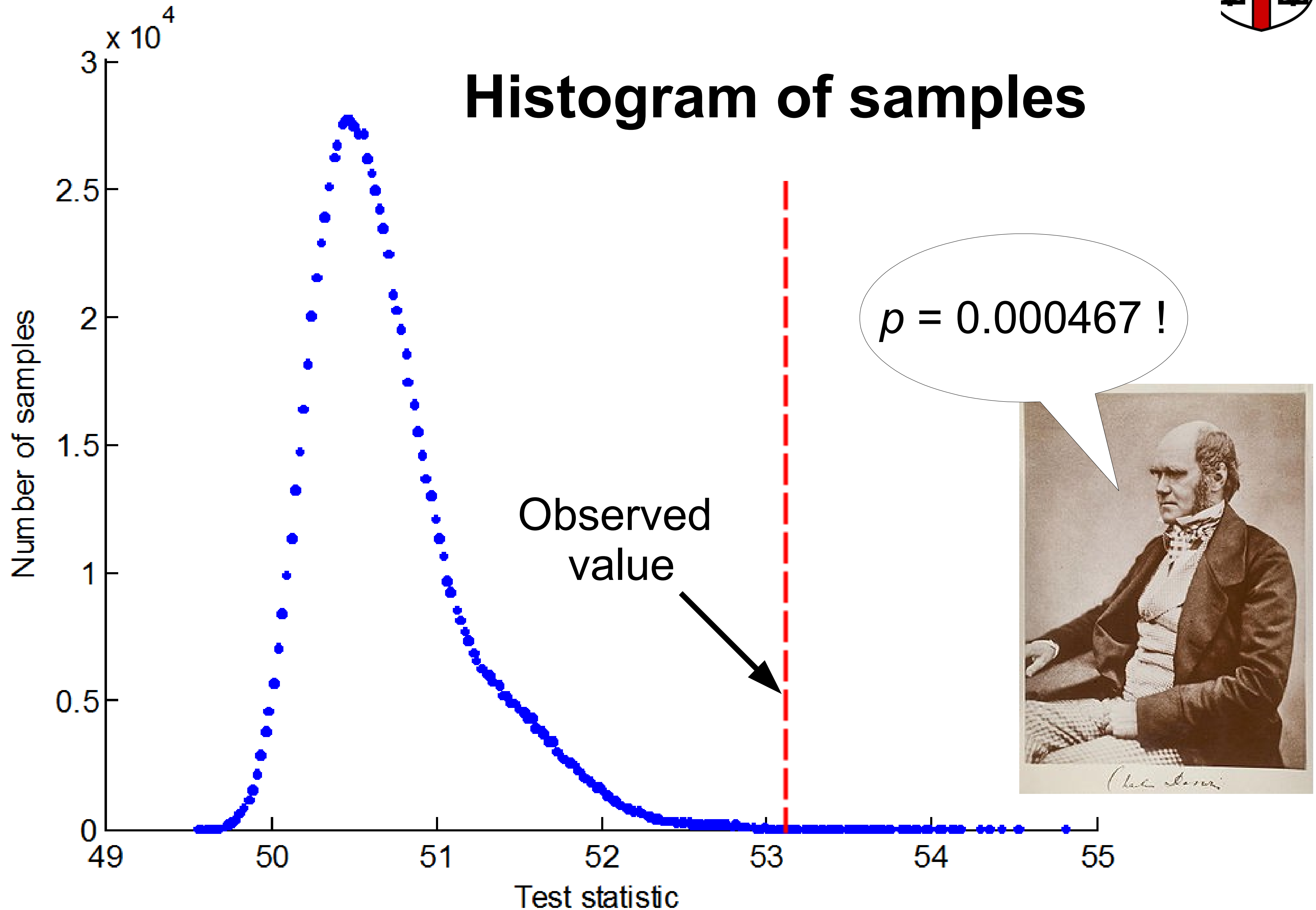


Table 1: 26 Mammalian Species in 28 Mountain Ranges in the American Southwest

	Range																											
Species	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
A	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0
B	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0
C	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1
D	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	0	0	0	0
E	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	0	0	0	0
F	1	1	1	1	1	0	0	0	1	1	1	1	1	0	1	1	0	1	1	1	1	0	1	0	1	0	0	0
G	1	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
H	1	1	1	1	1	1	0	1	1	0	1	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
I	1	1	1	1	1	1	1	1	0	1	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
J	1	1	1	1	1	1	1	1	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L	1	1	1	1	1	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0
M	1	1	1	1	1	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
N	1	1	1	1	0	1	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
O	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P	1	1	0	1	1	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	1	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R	1	1	1	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
U	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Z	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Patterson and Atmar, 1986

Example #2: Montane mammals

We compare results with Patterson and Atmar, 1986.

Results for Montane Mammal Data, using S_n

Method	# samples	p-value	mean of S_n	std. dev.	min	max
Exact	1,000,000	$0.0322 \pm .00018$	80.71	9.697	44	132
Heuristic	1,000	9×10^{-20}	227.9	18.135	180	287

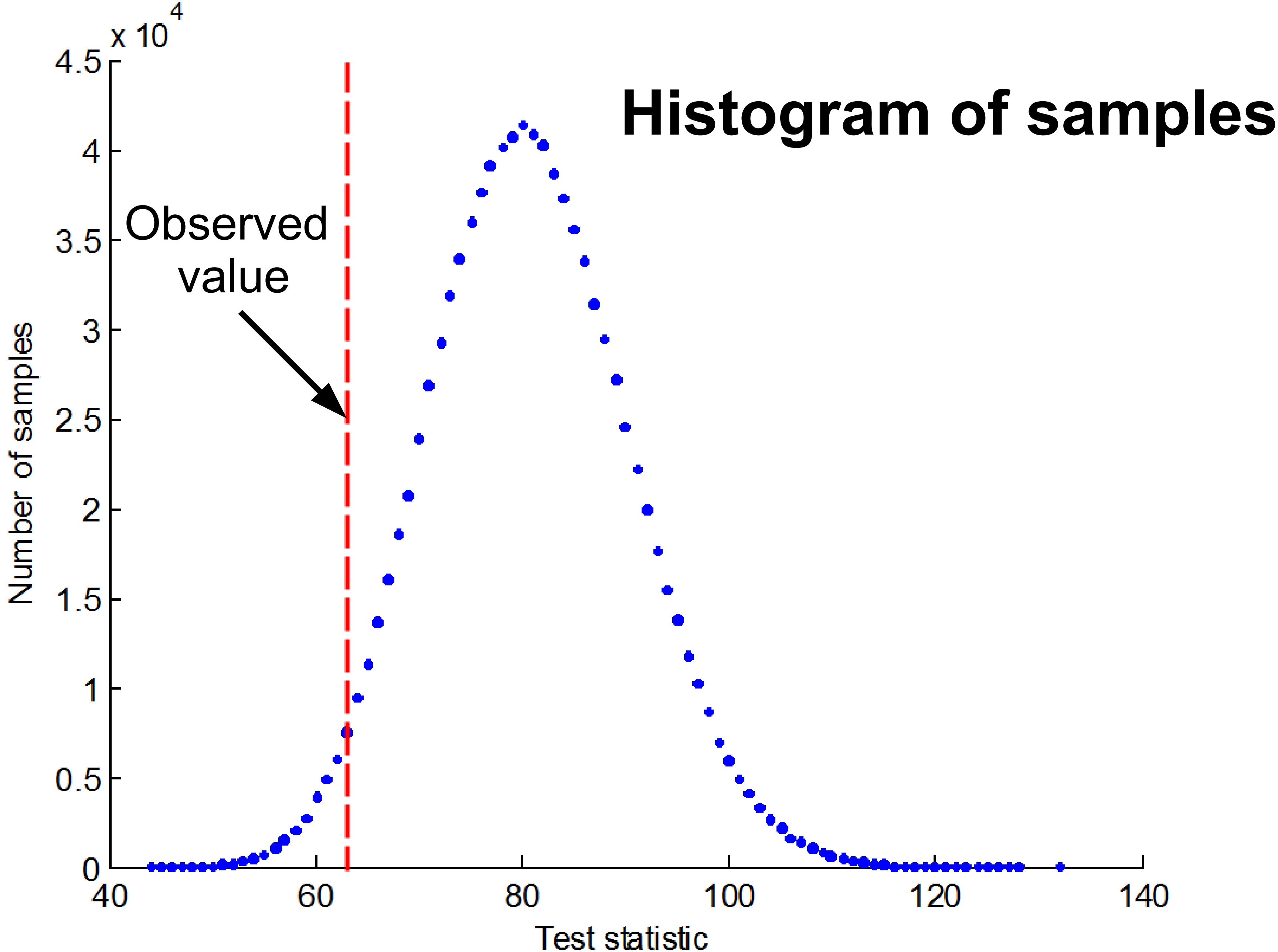
Test statistic S_n = "nested subset statistic" (Observed value: 63)

Number of matrices with margins as observed:

2,663,296,694,330,271,332,856,672,902,543,209,853,700

~ 2.6×10^{39} (computed in 32 minutes)

Example #2: Montane mammals



20 Lizard Species on 25 Islands in the Gulf of California

	Island																								
Lizard	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	1	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	1	0	1	0
2	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1
3	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	1	1	0	0
4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	1	0	1	0
5	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
6	1	0	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	0	1	0	0	1	1	1	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	1	1	1	1
8	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	0
9	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	0
11	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0
13	1	0	1	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	1	1	1	1
14	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	1	1	1	1	0	0	0	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0
17	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	1	0	0	1	0	1	1
18	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1
20	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Manly, 1995

Example #3: Island lizards

We compare results with Manly, 1995.

Results for Island Lizards, using S_d

Method	# samples	p-value
Exact	1,000,000	$(1.34 \pm .12) \times 10^{-4}$
MCMC	1,000,000	$(5.0 \pm .4) \times 10^{-4}$

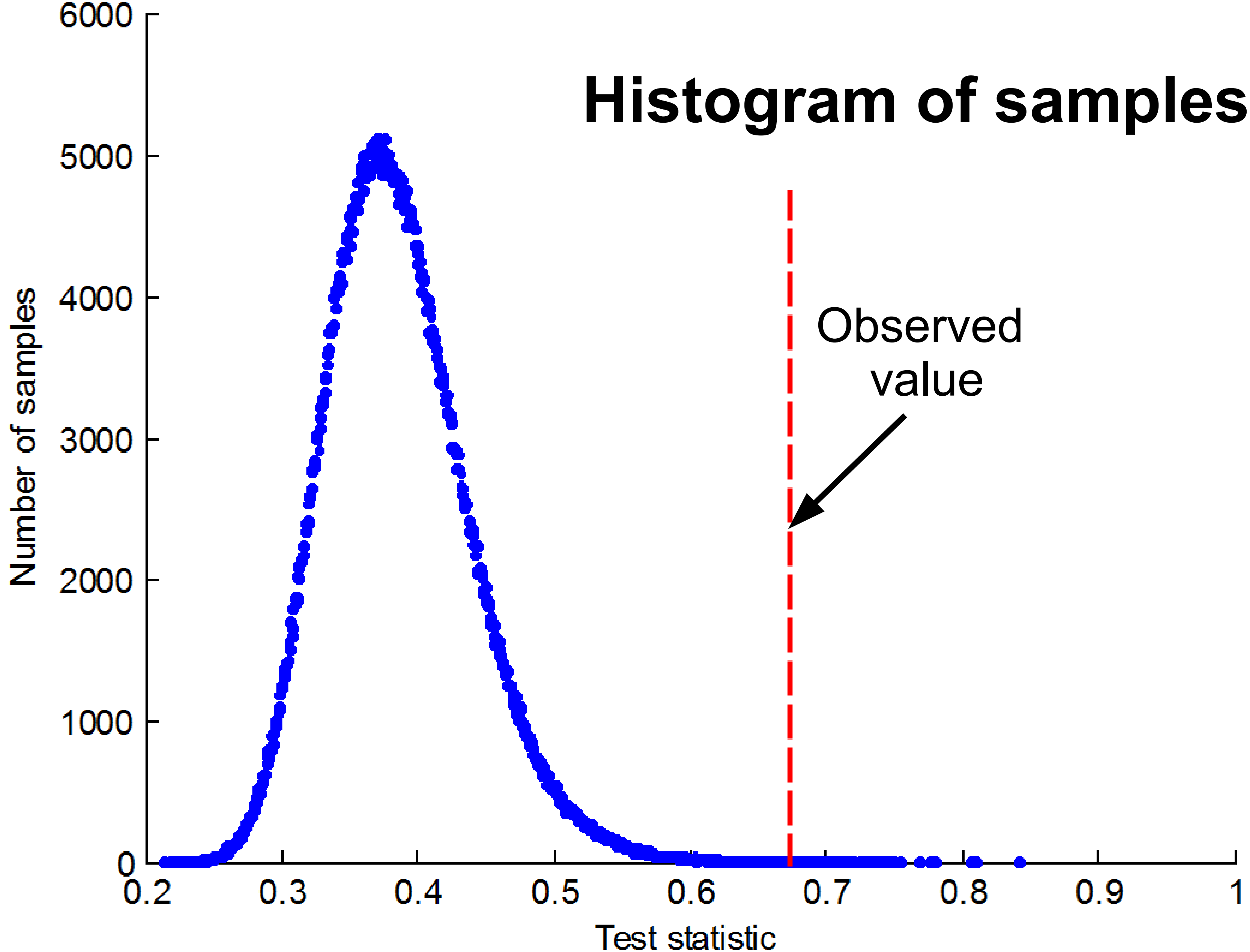
Test statistic S_d = “deviation from expected co-occurrences”

Number of matrices with margins as observed:

55,838,420,515,731,001,979,319,625,577,023,858,901,579,264

~ 5.5×10^{43} (computed in 11 minutes)

Example #3: Island lizards



Summary

- Uniform distribution on binary matrices with given row/column sums
- Existing theory does not provide adequate guarantees on convergence rates.
- **We offer an exact algorithm that is tractable in many cases.**

Contact info:

Jeff Miller

jeffrey_miller@brown.edu

Thank you for listening!