STATISTICS BY EXAMPLE
Weighing Chances

Prepared and edited by the
Joint Committee on the Curriculum in Statistics and Probability
of the American Statistical Association
and the National Council of Teachers of Mathematics

Frederick Mosteller, *Chairman*
Harvard University

Richard S. Pieters
Phillips Academy, Andover

William H. Kruskal
The University of Chicago

Gerald R. Rising
State University of New York at Buffalo

Richard F. Link
Artronic Information Systems, Inc.

with the assistance of

Roger Carlson
University of Missouri at Kansas City

Martha Zelinka
Weston High School, Weston, Massachusetts

ADDISON-WESLEY PUBLISHING COMPANY
Reading, Massachusetts · Menlo Park, California · London · Don Mills, Ontario

Independence of Amoebas

JOEL E. COHEN

INTRODUCTION

What effects are caused by the presence of two different kinds of amoebas in the intestinal tracts of men? One kind may cause disease and the other not, but the presence of the second kind may also prevent the disease usually caused by the first. A simple 2 × 2 table yields a chi-square test that helps to decide whether infection by one kind leads to infection by the other.

An epidemic of severe intestinal disease occurred among the workers in a woodworking plant in South Bend, Indiana. Doctors attributed the illnesses to unusually large numbers of an amoeba, called *Entamoeba histolytica*, growing in the men. After investigating alternative possibilities, public health officials concluded that the amoebae entered the plant's water supply through a leaky water main which passed near a leaky sewage pipe.

In order to see how many of the men in the plant who had not fallen sick had been infected with the amoeba unawares, the public health officials chose a random sample of 138 apparently well workers in the plant. The results are summarized in Table 1. They found the amoeba in the intestinal contents of 70 of the men. They distinguished two races of *Entamoeba histolytica*, large and small. (In the large race, the average diameter at one stage in the life cycle is

The author is at Harvard University, Cambridge, Massachusetts.

greater than 10 microns, while in the small race, the average diameter at the same stage is less than 10 microns.) They found the large race only in 35 men, the small race only in 23 men, and both races in 12 men.

Table 1. Observed frequencies of infections with the large and small races of Entamoeba histolytica among 138 randomly chosen workers in a South Bend (Indiana) woodworking plant

	Large race		
Small race	Present	Absent	Totals
Present	12	23	35
Absent	35	68	103
Totals	47	91	138

Since the small race of the amoeba is believed not to cause disease, we can ask some interesting questions of their data. Does infection with one race confer resistance to infection with the other? Or does infection with one race increase susceptibility to the other? In short, do the two races interact in infecting man?

We recall that if two events A and B are independent, then the probability that both occur is the product $P(A)P(B)$. If we have several trials, n, then the expected frequency with which both occur is $nP(A)P(B)$. Let $\bar{A}$ be the event complementary to A and $\bar{B}$ the event complementary to B.

In a 2 × 2 table, if the events A and B are independent, then so are the pairs of events $\bar{A}$ and B, A and $\bar{B}$, and $\bar{A}$ and $\bar{B}$, as you are asked to prove in Exercise 6. And so we find if A and B are independent, the expected counts in a 2 × 2 table are

	B	$\bar{B}$	Total
A	$nP(A)P(B)$	$nP(A)P(\bar{B})$	$nP(A)$
$\bar{A}$	$nP(\bar{A})P(B)$	$nP(\bar{A})P(\bar{B})$	$nP(\bar{A})$
Total	$nP(B)$	$nP(\bar{B})$	n

Note that the upper left-hand corner cell

$$nP(A)P(B) = \frac{(\text{row total})(\text{column total})}{n}$$

and that the formula on the right can be used to get every cell entry in the 2×2 table.

In the numerical tables the row and column totals are estimates of $nP(A)$ and the other marginal totals.

Let us return to the data of Table 1. If the presence or absence of the large race made no difference to the frequency with which the small race was present or absent, then among the total 47 individuals in which the large race was found, the small race should be present in the fraction 35/138, that is, in $47 \times 35/138 = 11.9$ individuals. The expected numbers of individuals with or without each kind of infection, assuming no interactions between the infections, are presented in Table 2.

Table 2. Predicted frequencies of infections with the large and small races of *Entamoeba histolytica* among 138 randomly chosen workers, assuming the same total prevalence of each race and no interaction between the races

	Large race		
Small race	Present	Absent	Totals
Present	11.9	23.1	35.0
Absent	35.1	67.9	103.0
Totals	47.0	91.0	138.0

The observed frequencies and those expected assuming no interaction are so close that the data provide no grounds for supposing that infection with one race of the amoeba makes an individual more or less susceptible to infection with another.

Exercises

1. Show arithmetically that if the expected count for the cell with 23 men is also made from the formula (row total)(column total)/(grand total), the expected values in the row must add to the row total exactly.

2. Generalize the result of Exercise 1 to the table with original cell entries

a	b
c	d

3. In the example, given the row and column totals,

what cell entries, if any, would have come closer to the expected values than the ones found?

4. What would be the expected value of chi-square for a 2 × 2 table if the rows and columns were independent?

5. Computing chi-square for the present numerical example does not seem worthwhile. Why not?

6. Prove that if A and B are independent, so are $\bar{A}$ and B.

7. Explain the practical importance of the discovery of independence between the races of amoeba. Explain what would have been the importance of finding a strong dependence.

Reference

[1] Charles A. LeMaistre *et al.*, "Studies of a waterborne outbreak of amebiasis, South Bend, Indiana", *Am. J. Hyg.* 64 (1956): 30-45.