Opinionated Lessons in Statistics

by Bill Press

#32 Contingency Tables: A First Look

Contigency Tables, a.k.a. Cross-Tabulation

TABLE 1
Maternal drinking and congenital malformations

Malformation	Alcohol consumption (average no. of drinks/day)				
	0	< 1	1–2	3–5	≥ 6
Absent	17,066	14,464	788	126	37
Present	48	38	5	1	1

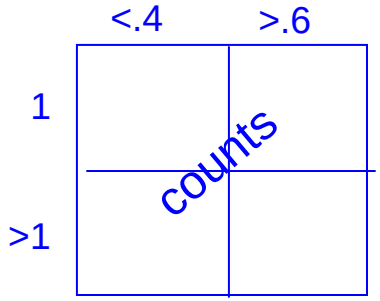
Source: Graubard and Korn (1987).

Is alcohol implicated in malformations?

This kind of data is often used to set public policy, so it is important that we be able to assess its statistical significance.

Contingency Tables (a.k.a. cross-tabulation)

Ask: Is a gene is more likely to be single-exon if it is AT-rich?




```
rowcon = [(g.ne == 1) (g.ne > 1)];
colcon = [(g.atf < 0.4) (g.atf > 0.6)];

table = contingencytable(rowcon,colcon)
table =
        2386         689
       13369        3982
```

(fewer genes AT rich than CG rich)

```
sum(table, 1)
ans =
       15755        4671
```

column marginals

```
ptable = table ./ repmat(sum(table,1),[2 1])
ptable =
    0.1514    0.1475
    0.8486    0.8525
```

So can we claim that these are statistically identical?
Or is the effect here also "significant but small"?

my contingency table function:

```
function table = contingencytable(rowcons, colcons)
nrow = size(rowcons,2);
ncol = size(colcons,2);
table = squeeze(sum( repmat(rowcons,[1 1 ncol]) .*  ...
        permute(repmat(colcons,[1 1 nrow]),[1 3 2]),1 ));
```

Chi-square (or Pearson) statistic for contingency tables

notation:

$$N_{i\cdot} = \sum_j N_{ij} \qquad N_{\cdot j} = \sum_i N_{ij}$$

$$N = \sum_i N_{i\cdot} = \sum_j N_{\cdot j}$$

null hypothesis:

$$\frac{n_{ij}}{N_{\cdot j}} = \frac{N_{i\cdot}}{N} \quad \rightarrow \quad n_{ij} = \frac{N_{i\cdot} N_{\cdot j}}{N}$$

expected value of N_{ij}

the statistic is:

$$\chi^2 = \sum_{i,j} \frac{(N_{ij} - n_{ij})^2}{n_{ij}}$$

```
table =
        2386         689
       13369        3982
```

- Are the conditions for valid chi-square distribution satisfied? Yes, because number of counts in all bins is large.
- If they were small, we *couldn't* use fix-the-moments trick, because small number of bins (no CLT). This occurs often in biomedical data.
- So what then? (We will return to this!)

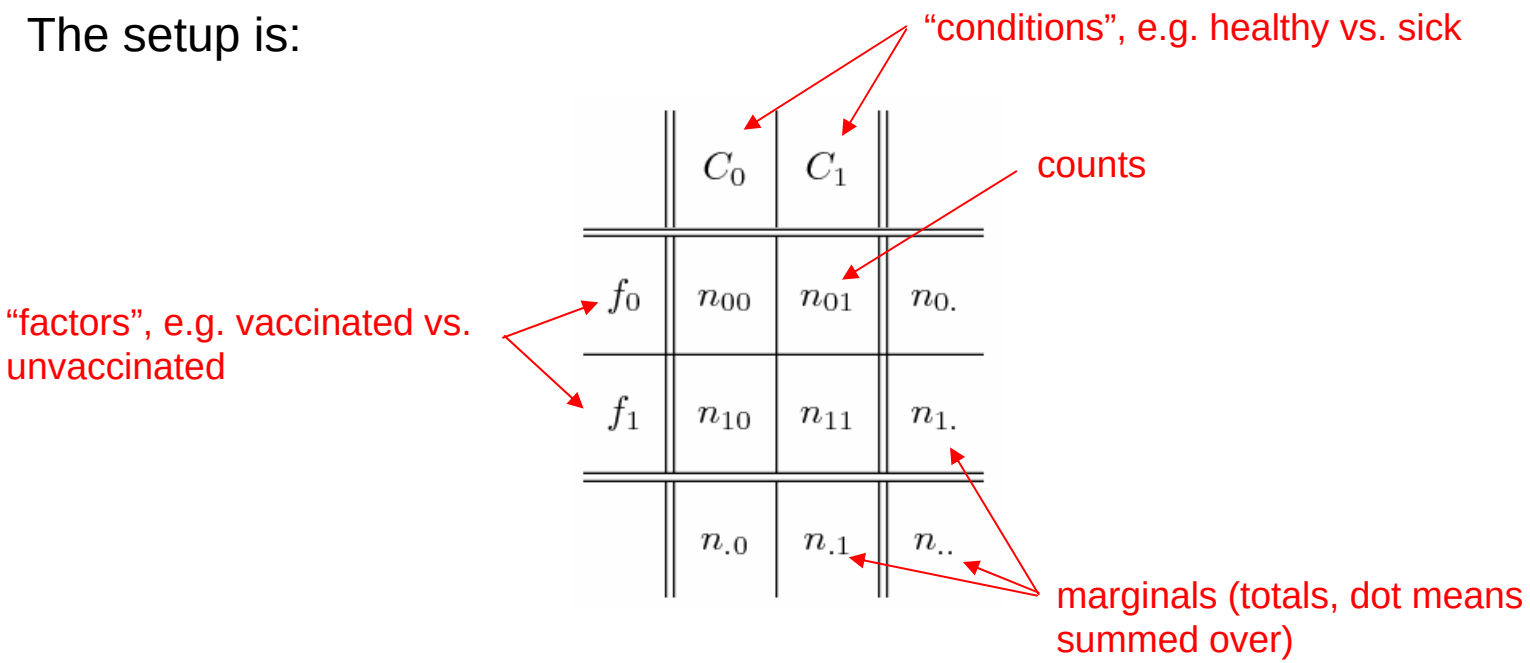
```
nhtable = sum(table,2)*sum(table,1)/sum(sum(table))
nhtable =
  1.0e+004 *
    0.2372    0.0703
    1.3383    0.3968
chis = sum(sum((table-nhtable).^2./nhtable))
chis =
    0.4369
p = chi2cdf(chis,1)
p =
    0.4914
```

d.f. = 4 – 2 – 2 + 1

wow, can't get less significant than this! No evidence of an association between single-exon and AT- vs. CG-rich.

When counts are small, some subtle issues show up. Let's look closely.

The setup is:



The null hypothesis is: "Conditions and factors are unrelated."

To do a p-value test we must:

1. Invent a statistic that measures deviation from the null hypothesis.
2. Compute that statistic for our data.
3. Find the distribution of that statistic over the (unseen) population.

That's the hard part! What is the "population" of contingency tables? We'll soon see that it depends (maybe only slightly?) on the experimental protocol, not just on the counts!

Let's review the hypergeometric distribution

What is the (null hypothesis) probability of a car race finishing with 2 Ferraris, 2 Renaults, and 1 Honda in the top 5 if each team has 6 cars in the race and the race consists of only those teams?

Hypergeometric probabilities have product of "chooses" in the numerator, and a denominator "choose" with sums of numerator arguments.

$$\frac{\binom{A}{a}\binom{B}{b}\binom{C}{c}}{\binom{A+B+C}{a+b+c}} = \frac{\binom{6}{2}\binom{6}{2}\binom{6}{1}}{\binom{18}{5}} = 0.1576$$

Out of N genes, m are associated with disease 1 and n with disease 2. What is the (null hypothesis) probability of finding r genes overlap?

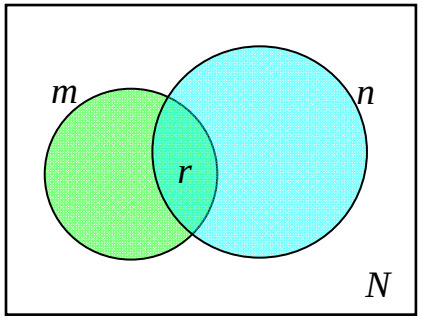
choose overlap

choose rest of 2nd set

choose 1st set

$$\frac{\binom{N}{m}\binom{m}{r}\binom{N-m}{n-r}}{\binom{N}{m}\binom{N}{n}} = \frac{\binom{m}{r}\binom{N-m}{n-r}}{\binom{N}{n}}$$

choose each set independently



$$= \frac{m!n!(N-m)!(N-n)!}{r!(m-r)!(n-r)!(N-m-n+r)!N!} \equiv \mathrm{hyper}(r; N, m, n)$$

Yes, it is symmetrical on m and n!

And now, review the multinomial distribution

On each i.i.d. try, exactly one of *K* outcomes occurs, with probabilities

$$p_1, p_2, \ldots, p_K \qquad \sum_{i=1}^{K} p_i = 1$$

For *N* tries, the probability of seeing exactly the outcome

$$n_1, n_2, \ldots, n_K \qquad \sum_{i=1}^{K} n_i = N$$

is

$$P(n_1, \ldots, n_K | N, p_1, \ldots, p_K) = \underbrace{\frac{N!}{n_1! \cdots n_K!}}_{\text{number of equivalent arrangements}} \overbrace{p_1^{n_1} p_2^{n_2} \cdots p_K^{n_K}}^{\text{probability of one specific outcome}}$$

N=26: `abcde fgh ijklmnop q rs tuvwxyz`

N! arrangements

`(12345)(123)(12345678)(1)(12)(1234567)`

partition into the observed n_i's

$n_1 = 5$ $n_2 = 3$ $n_6 = 7$