CS395T
Computational Statistics with Application to Bioinformatics

Prof. William H. Press
Spring Term, 2011
The University of Texas at Austin

Lecture 18

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

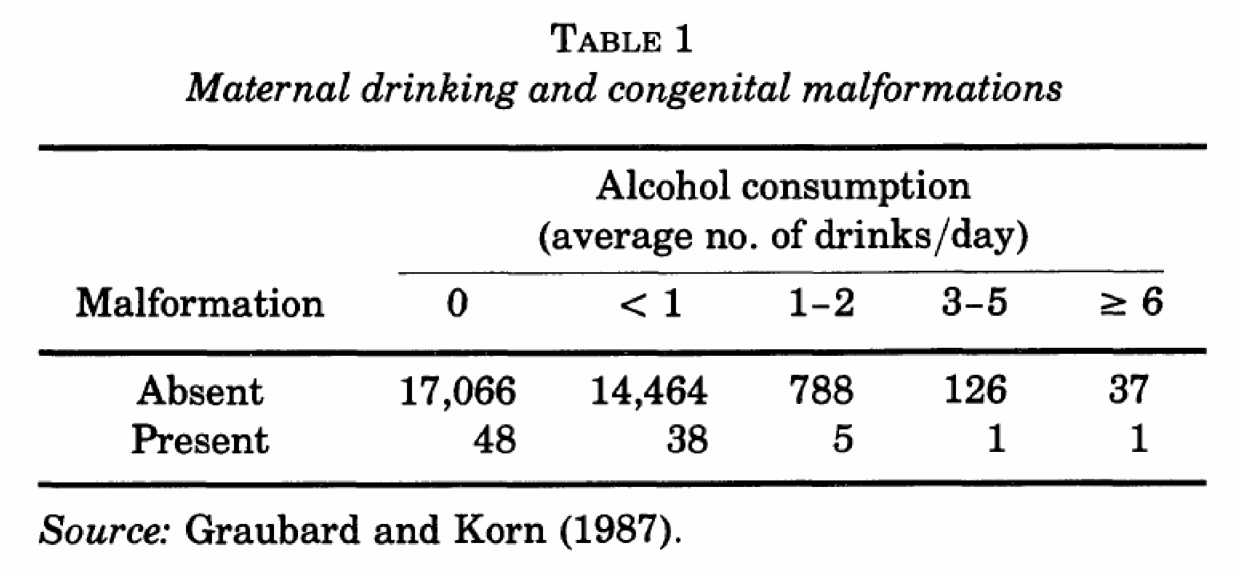


TABLE 1

Maternal drinking and congenital malformations

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

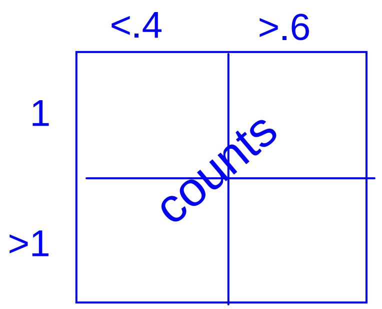
Source: Graubard and Korn (1987).

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.piso < 0.4) (g.piso > 0.6)];
crosstab(rowcon * (1:2)', colcon * (1:2)')
ans =
          48          2386           689
         220         13369          3982
```

annoying counts of “other”

```
table = contingencytable(rowcon,colcon)
```

my improved function (below)

```
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:

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

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

the statistic is:

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

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

null hypothesis:

expected value of N_{ij}

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

•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!)

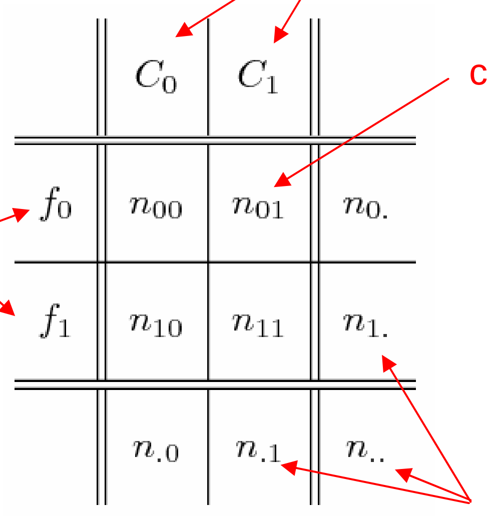
```
nhtable = sum(table,2)*sum(table,1)/sum(sum(table))
nhtable =
  1.0e+004 *
    0.2372    0.0703
    1.3383    0.3968
chis = sum(sum((table-nhtable).^2./nhtable))
chis =
    0.4369
p = chi2cdf(chis,1)
p =
    0.4914
```

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:



	C_0	C_1	
f_0	n_{00}	n_{01}	$n_{0.}$
f_1	n_{10}	n_{11}	$n_{1.}$
	$n_{.0}$	$n_{.1}$	$n_{..}$

“conditions”, e.g. healthy vs. sick

counts

“factors”, e.g. vaccinated vs. unvaccinated

marginals (totals, dot means summed over)

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 now see that it depends (though often only slightly) on the experimental protocol, not just on the counts!

Protocol 1: Retrospective analysis or “case/control study”

	C_0	C_1	
f_0	n_{00}	n_{01}	$n_{0.}$
f_1	n_{10}	n_{11}	$n_{1.}$
	$n_{.0}$	$n_{.1}$	$n_{..}$

C_1 already has the disease. We retrospectively look at their factors.

	C_0	C_1	
q	$\text{bin}_q(n_{.0}, n_{00})$	$\text{bin}_q(n_{.1}, n_{01})$	$n_{0.}$
$(1-q)$	$\surd$	$\surd$	$n_{1.}$
	$n_{.0}$ (fixed)	$n_{.1}$ (fixed)	$n_{..}$ (fixed)

In the null hypothesis, both columns share row probabilities q and (1-q). But we don’t know q. It’s a “nuisance parameter”.

$$\begin{aligned}
P(\text{table}) &= \text{bin}_q(n_{.0}, n_{00})\text{bin}_q(n_{.1}, n_{01}) \\
&= \binom{n_{.0}}{n_{00}} q^{n_{00}}(1-q)^{n_{10}} \binom{n_{.1}}{n_{01}} q^{n_{01}}(1-q)^{n_{11}} \\
&= \frac{n_{.0}!n_{.1}!}{n_{00}!n_{01}!n_{10}!n_{11}!} q^{n_{0.}}(1-q)^{n_{1.}} \\
&= \text{bin}_q(n_{..}, n_{0.}) \times \frac{n_{0.}!\,n_{1.}!\,n_{.0}!\,n_{.1}!}{n_{..}!\,n_{00}!\,n_{01}!\,n_{10}!\,n_{11}!} \\
&\equiv \text{bin}_q(n_{..}, n_{0.}) \times \text{hyper}(n_{00}; n_{..}, n_{.0}, n_{0.}) \\
&= P(n_{0.}\,|\,n_{..}, q) \times P(\text{table}\,|\,n_{0.}, n_{..})
\end{aligned}$$

$$\text{hyper}(n_{00}; n_{..}, n_{.0}, n_{0.}) = \frac{\binom{n_{.0}}{n_{00}}\binom{n_{.1}}{n_{01}}}{\binom{n_{..}}{n_{0.}}}$$

Digression on 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?

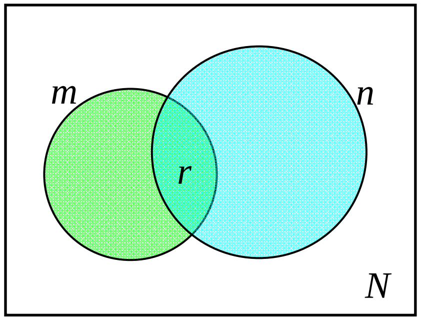
choose overlap

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



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

Yes, it is symmetrical on m and n!

The model problem for 2x2 contingency tables is a slightly different variant

The Texas Legislature has *m* Republicans and *n* Democrats.
A committee of size *r* is chosen randomly [not realistic!]

What is the probability distribution of *i*, the number of Democrats on the committee?

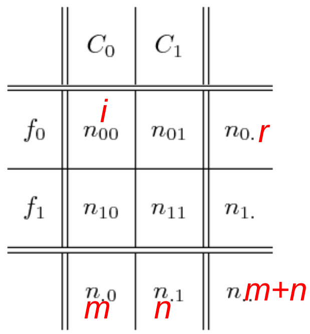
number of ways to choose *i* Democrats

number of ways to choose *r-i* Republicans

$$= \frac{\binom{n}{i}\binom{m}{r-i}}{\binom{m+n}{r}}$$

total number ways to choose the committee

$$p(i) = \mathrm{hyper}(i; m+n, n, r)$$

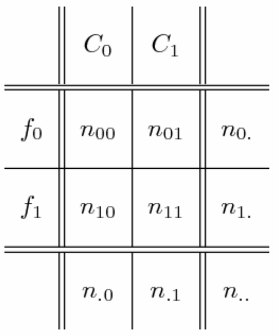


	C_0	C_1	
f_0	i n_{00}	n_{01}	$n_{0\cdot}$ r
f_1	n_{10}	n_{11}	$n_{1\cdot}$
	$n_{\cdot 0}$ m	$n_{\cdot 1}$ n	$n_{\cdot\cdot}$ m+n

$$\mathrm{hyper}(n_{00}; n_{\cdot\cdot}, n_{\cdot 0}, n_{0\cdot}) = \frac{\binom{n_{\cdot 0}}{n_{00}}\binom{n_{\cdot 1}}{n_{01}}}{\binom{n_{\cdot\cdot}}{n_{0\cdot}}}$$

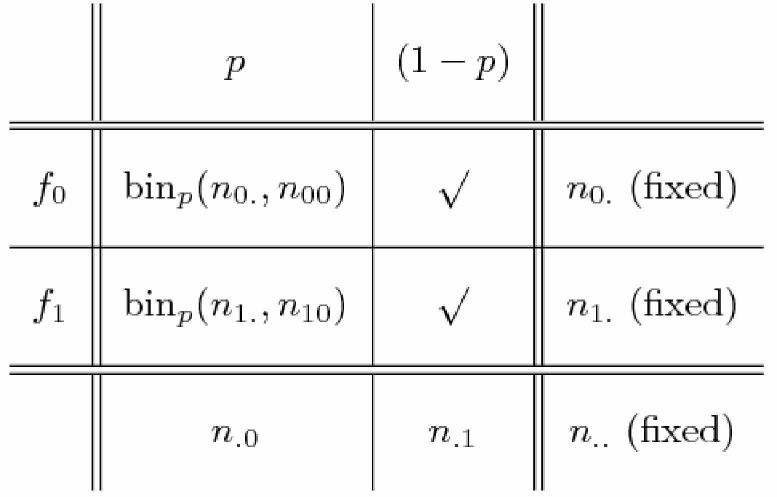
Protocol 2: Prospective experiment or “longitudinal study”

Identify samples with the factors, then watch to see who gets the disease



	C_0	C_1	
f_0	n_{00}	n_{01}	$n_{0\cdot}$
f_1	n_{10}	n_{11}	$n_{1\cdot}$
	$n_{\cdot 0}$	$n_{\cdot 1}$	$n_{\cdot\cdot}$

In the null hypothesis, both rows share row probabilities p and (1-p). But we don’t know p. It’s now the nuisance parameter.



	p	$(1-p)$	
f_0	$\text{bin}_p(n_{0\cdot}, n_{00})$	$\surd$	$n_{0\cdot}$ (fixed)
f_1	$\text{bin}_p(n_{1\cdot}, n_{10})$	$\surd$	$n_{1\cdot}$ (fixed)
	$n_{\cdot 0}$	$n_{\cdot 1}$	$n_{\cdot\cdot}$ (fixed)

$$\begin{aligned} P(\text{table}) &= \text{bin}_p(n_{0\cdot}, n_{00})\text{bin}_q(n_{1\cdot}, n_{10}) \\ &= \text{bin}_p(n_{\cdot\cdot}, n_{\cdot 0}) \times \frac{n_{0\cdot}!\, n_{1\cdot}!\, n_{\cdot 0}!\, n_{\cdot 1}!}{n_{\cdot\cdot}!\, n_{00}!\, n_{01}!\, n_{10}!\, n_{11}!} \\ &= P(n_{\cdot 0} \,|\, n_{\cdot\cdot}, p) \times P(\text{table} \,|\, n_{\cdot 0}, n_{\cdot\cdot}) \end{aligned}$$