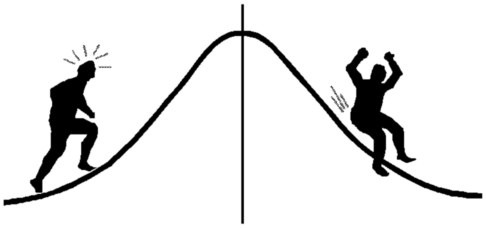

BIOMETRICS INFORMATION

(You're 95% likely to need this information)

PAMPHLET NO. # 36 DATE: January 6, 1992

SUBJECT: Contingency Tables and Log-linear Models

Contingency tables or log-linear models usually provide the most suitable methods for analysis of count data. This pamphlet will demonstrate these methods by way of an example. First, a "traditional" analysis of counts by a 3-dimensional contingency table will be demonstrated. This analytical method requires multiple runs of `PROC FREQ` (in SAS) and many hand calculations (Sections 1, 2 and 3 below). Then it will be shown that these many steps can be accomplished more quickly and comprehensively by the use of log-linear models using `PROC CATMOD`.

The example is based on made-up data using an experimental design where 4 insect traps have been set out in an area. The insects caught were identified by species and sex. SAS programs for many of the analyses and results presented are summarized in the Appendix.

Note that the main objective of this pamphlet is to introduce log-linear models by demonstrating some of their advantages over the traditional contingency table approach. Thus the emphasis is on **how** the analyses are done, and not with **why** they are done this way. A full understanding of the contents, especially for those unfamiliar with the topic, will require working through the analyses presented and studying some of the references. The references that I found most useful for understanding contingency tables are Sokal and Rohlf (1981) and Everitt (1977). Standard texts like Snedecor and Cochran (1980), Steel and Torrie (1980) and Zar (1974) would also be useful. Log-linear models have an extensive literature and are directly discussed, for instance, by Fienberg (1981) and Freeman (1987).

1. Two-way table

Suppose that the four traps caught three species of insects (ignoring sex) with the following results.

	Trap:	1	2	3	4	Total
Species	#1	62	93	77	51	283
	#2	49	53	40	30	172
	#3	71	109	79	38	297
						752

This set of data lends itself directly to a 2-dimensional (r x c) contingency table analysis. Two equivalent test statistics that I will present are the χ^2 and G statistics. Sokal and Rohlf (Chap. 17) describe these statistics. The G-statistic output from `PROC FREQ` is called the `Likelihood Ratio Chi-Square` on the printout. Both of these statistics test the null hypothesis (H_o) that there is no interaction between species and trap. If H_o is true then this would mean, for instance, that *the proportion of species trapped is constant for the four traps.* In other words, the observed proportions of the row totals: 283:172:297 are about the same as the proportions observed for each trap. The null hypothesis could also be stated by discussing ratios within the columns. For example, that the proportions 62:93:77:51 for species 1 about the same as for species 2 and 3.

BRITISH COLUMBIA

Ministry of Forests
Research Program

For this data table, $\chi^2 = G^2 = 6.8$†, df = 6, p = 0.34. This suggests that there is no reason to reject the null hypothesis, and no reason to believe that the traps were not equally effective at capturing all three species.

Pay close attention to the null hypothesis. Contingency table hypotheses always relate to *ratios*, and not the differences between means as in ANOVA.

2. One-way tables (main effects)

While it is indeed interesting and worthwhile to test whether the four traps have been equally effective at catching the three species of insects, a more interesting question might be: Is there a difference in abundance between the three species? To test this, we must also assume that the traps are catching the same proportion of insects from each species available in the population surrounding the traps.

If the species were equally abundant then each species should account for one third of the total insects caught (this is H_o):

					Test Statistics	
		Observed	Expected	Expected		
i.e.	Species	Count	Ratio	Count	$\chi^2 = (O\text{-}E)^2/E$	G = 2(O*ln(O/E))‡
	1	283	1	$250\,{}^{2}/_{3}$	4.17	68.67
	2	172	1	$250\,{}^{2}/_{3}$	24.69	-129.56
	3	297	1	$250\,{}^{2}/_{3}$	8.56	100.75
	Sum:	752	3	752	37.42	39.86

df = 2, $p \leq 0.0001$

The final test statistics are $\chi^2 = 37.42$ and G = 39.86, both with 2 df. Given an H_o of equal abundance both statistics are highly unlikely values ($p \leq 0.0001$), so that the null hypothesis is rejected. It is reasonably clear from the individual contributions to the chi-square statistic that species 2 occurs less frequently than expected, while species 1 and 3 occur about equally often.

A priori hypotheses other than equal abundance can also be tested in this manner. Suppose it had been expected that half as many of species 2 would be caught when compared to either species 1 or 3. Then:

				Test Statistics	
	Observed	Expected	Expected		
Species	Count	Ratio	Count	$\chi^2 = (O\text{-}E)^2/E$	G = 2(O*ln(O/E))
1	283	2	300.8	1.05	-34.53
2	172	1	150.4	3.10	46.16
3	297	2	300.8	0.05	-7.55
Sum:	752	5	752.0	4.20	4.09

df = 2, $0.10 \leq p \leq 0.25$

These results indicate that a 2:1:2 ratio of species abundance is indeed a reasonable hypothesis. Main effect differences between traps could also be tested in a similar way.

† The calculations required to obtain these statistics are described in many textbooks. See the references noted in the introduction.

‡ These are the general formulae for the χ^2 and G-statistics. Note that O stands for Observed values while E stands for Expected values and ln is the natural logarithm.

3. Three-way table

The first table was pooled over sex. The analysis **implicitly assumed** that the interaction of trap x species was the same for each sex. If not, then pooling would be unjustified and incorrect conclusions could have been reached. Suppose that the two-way table can be split into two tables by the sex of the insects as follows:

Sex = Female

	Trap:	1	2	3	4	Total
Species	#1	32	45	34	20	131
	#2	23	24	17	17	81
	#3	46	68	62	28	204
						416

Sex = Male

	Trap:	1	2	3	4	Total
Species	#1	30	48	43	31	152
	#2	26	29	23	13	91
	#3	25	41	17	10	93
						336

The legitimacy of pooling can be checked by the following calculations and χ^2-test (not directly achievable with `PROC FREQ`). The null hypothesis, H_o, is that the subtables are similar enough to be pooled. This can also be stated as the null hypothesis that there is no interaction between species, sex and trap.

	df	χ^2	Prob	G	Prob
Sex = Female	6	5.140	0.53	5.068	0.54
Sex = Male	6	11.141	0.08	11.240	0.08
Total	12	16.281		16.308	
Pooled	6	6.757	0.34	6.798	0.34
Homogeneity of Subtables	6	9.524	p = 0.15	9.510	p = 0.15

The test statistics for homogeneity of the subtables (obtained by subtraction of the pooled statistic from the total statistic) are χ^2 = 9.52 and G = 9.51 with 6 df. These values are quite likely if the subtables are similar and there is no 3-way interaction. Thus it is not necessary to change the previous analysis.

The species by sex interaction might be of particular interest. Suppose that we consider traps to be replicate samples, then we might want to pool the data from the four traps into one 2x3 contingency table. This would test if the species by sex interaction is reasonably the same for each trap and is another test of the 3-way interaction between species, sex, and trap. Calculations to answer this question are:

	df	χ^2	Prob	G	Prob
Trap = 1	2	4.314	0.12	4.351	0.11
Trap = 2	2	5.873	0.053	5.910	0.052
Trap = 3	2	23.545	0.001	24.586	0.001
Trap = 4	2	10.505	0.005	10.786	0.005
Total	8	44.146		45.633	
Pooled	2	35.516	0.001	36.124	0.001
Homogeneity of Subtables	6	8.630	p = 0.20	9.509	p = 0.15

The final statistics are $\chi^2 = 8.63$ and G = 9.51 with 6 df. Again these are likely values given the null hypothesis of similar species by sex interaction within each trap. Thus the pooled table can be used and there is a species by sex interaction because the pooled statistics, $\chi^2 = 35.516$ and G = 36.124, are quite large for df = 2 (p = 0.001). To see what is causing the interaction let's look closely at the pooled table:

	Species =	1	2	3	Total
Female	Obs. Count	131	81	204	416
	(χ^2)	(4.2)	(2.1)	(9.6)	
	ratio	0.463	0.471	0.687	0.553
Male	Obs. Count	152	91	93	336
	(χ^2)	(5.2)	(2.6)	(11.9)	
	ratio	0.537	0.529	0.313	0.447

The two cells for species 3 contribute most (9.6 + 11.9) to overall χ^2-statistic suggesting that it is somehow different from the other two species. Examination of the sex ratios for each species suggests that they are about the same for species 1 and 2 with more males than females, but that this is reversed for species 3 with more females than males. To check the reasonableness of this interpretation, the following two subtables are analyzed.

	Species =	1	2
Female	obs:	131	81
	(χ^2)	(0.0)	(0.0)
	ratio:	0.463	0.471
Male	obs:	152	91
	(χ^2)	(0.0)	(0.0)
	ratio:	0.537	0.529

$\chi^2 = 0.03$, df = 1, p = 0.87

Species =	1 & 2	3
Female	212	204
	(6.3)	(9.6)
	0.466	0.687
Male	243	93
	(7.8)	(11.9)
	0.534	0.313

$\chi^2 = 35.5$, df = 1, $p \leq 0.0001$

These results confirm the interpretation that species 1 & 2 have similar sex ratios but that their common sex ratio is different from that of species 3.

When analyzing many subtables like this, it is important to be careful about the probability levels. One suggestion is to use the critical chi-squared value from the full table for all of the subtables [Miller (1966) quoted in Fleiss (1981)]. Other options are described by Jones (1984). Although his discussion is in terms of means, his comments are directly applicable to multiple contingency tables. These comments mean that the probability values given for the above subtables should not be taken too seriously.

4. Log-linear models

So far, we have found a species x sex interaction, no species x trap interaction and species and trap main effects. This has required conducting many statistical tests and computer runs. A global view of all the possible interactions could have been determined by one log-linear model. These models are an extension of contingency tables and are better able to handle several factors and more complex hypotheses. The `CATMOD` procedure in Version 6.03 of SAS (on the PC) produced the following results:

Source	df	Chi-square	Prob
Trap	3	43.94	≤0.0001
Species	2	25.95	≤0.0001
Sex	1	6.17	0.013
Species x Sex	2	34.60	≤0.0001
Trap x Species	6	7.70	0.26
Trap x Sex	3	1.02	0.80
Trap x Species x Sex	6	8.99	0.17
Goodness of fit	0	0.00	1.0

This table summarizes all the possible factorial effects we could have looked for in the data. The goodness of fit test (indicated on the SAS printout by `Likelihood Ratio`) is perfect since the model is "saturated" which means that each cell value is predicted by its observed value. Since trap was used as a block it is reassuring to see no trap interactions. Based on the high probability values for trap x species, trap x sex, and trap x species x sex the next step is to try a model without those interactions. Note that the tests for individual terms in the model are called **Wald Tests** and are not considered to be reliable. Thus, it is important to refit the model that might be the best-fitting one. In this case, the results are:

Source	df	Chi-square	Prob
Trap	3	48.24	≤0.0001
Species	2	30.69	≤0.0001
Sex	1	5.09	0.024
Species x Sex	2	34.74	≤0.0001
Goodness of Fit	15	17.03	0.32

This would be the final model of choice for this example, since the Goodness of Fit Test, $\chi^2 = 17.03$, with df = 15 has a high probability (p = 0.32) suggesting that this model fits the data well (H_o for a Goodness of Fit test is that the model does adequately fit the data). The specific form of the species by sex interaction can be studied as we did previously with contingency tables.

Note that the chi-square values are only approximately additive and constant, unlike the sums of squares in ANOVA. The two log-linear models had χ^2-values of 34.6 and 34.7 for the interaction of species and sex while the first χ^2-test calculated was 35.5 (top of page 4). Nevertheless the values are similar, which is expected in this case because of the large sample sizes. The methods discussed here are large sample methods and are less reliable with small sample sizes. More complete discussions of these methods can be found in Everitt (1977) and Fienberg (1981).

5. Some topics not covered

1) Continuity Correction in 2x2 tables
2) Fisher's Exact Test
3) Discussion of Pearson's χ^2 statistic and the log-likelihood ratio statistic G. (See Sokal and Rohlf, Chap. 17)
4) Problems with small sample sizes and/or many cells with zero counts.

References:

Everitt, B.S. 1977. The analysis of contingency tables. Chapman and Hall, London, Great Britain.

Fienberg, S.E. 1981. The analysis of cross-classified categorical data. 2nd ed. The MIT Press, Cambridge, Mass.

Fleiss, J.L. 1981. Statistical methods for rates and proportions, 2nd ed. John Wiley and Sons. Toronto. (see page 141).

Freeman, H. 1987. Applied categorical data analysis. Marcel Dekker, Inc., New York and Basil.

Jones, D. 1984. Use, misuse and role of multiple-comparison procedures in ecological and agricultural entomology. Environ. Entomol. 13:635-649.

Miller, R.G. 1966. Simultaneous statistical inference. McGraw-Hill, N.Y. (see Section 6.2).

SAS Institute Inc. 1988. SAS User's guide: Statistics, Version 6.03 edition. SAS Institute Inc., Gary, N.C.

Snedecor, G.W. and W.G. Cochran. 1980. Statistical methods, 7th ed. The Iowa State Univ. Press Ames, Iowa.

Sokal, R.R. and F.J. Rohlf. 1981. Biometry. W.H. Freeman and Co., San Francisco. (see Chap 17).

Steel, R.G.D. and J.H. Torrie. 1980. Principles and procedures of statistics: a biometrical approach, 2nd ed. McGraw-Hill Book Co., New York, New York.

Zar, J.H. 1974. Biostatistical analysis. Prentice-Hall Inc., Englewood Cliffs, N.J.

Contact: Wendy Bergerud
387-5676

Appendix: Example SAS Program

```
title 'Example Contingency Tables Analysis' ;
options pagesize=60 linesize=78;
proc format; value sex 1 = 'female' 2 ='male' ;
        value spp 1 = 'spp 1'  2 ='spp 2'  3 ='spp 3' ;
        value spt 1,2 ='spp 1 & 2'     3 ='spp 3' ;
run;
```

```
data example;
 do sex = 1,2;
   do species = 1 to 3;
     do trap = 1 to 4;
     input count @ @;
     output;
 end; end; end;
format  sex sex.  species spp. ;
cards;
32 45 34 20
23 24 17 17
46 68 62 28
30 48 43 31
26 29 23 13
25 41 17 10
run;
```

/* FREQUENCY TABLE ANALYSES */

```
title3 'Two-way table of species by trap';
title4 '(table pooled over sex)';
proc freq data=example; weight count;
 tables species*trap / norow     nopercent cellchi2 chisq;
run;
title3 'Two-way tables of species by trap';
title4 '(separate tables for each sex)';
proc freq data=example; weight count;
 by sex;
 tables species*trap / norow     nopercent cellchi2 chisq;
run;
title3 'Two-way table of species by sex';
title4 '(pooled over traps for all species)';
proc freq data=example; weight count;
 tables sex*species /     nocol nopercent cellchi2 chisq;
run;
proc sort data=example; by trap;
title3 'Two-way tables of species by sex';
title4 '(individual tables for each trap)';
proc freq data=example; weight count;
 by trap;
 tables sex*species /     nocol nopercent cellchi2 chisq;
run;
title3 'Two-way table of species by sex';
title4 '(pooled over traps and without species 3)';
proc freq data=example; weight count;
 where species = 1 or species = 2;
 tables sex*species /     nocol nopercent cellchi2 chisq;
run;
title3 'Two-way table of species by sex';
title4 '(pooled over traps and species 1 & 2 pooled)';
proc freq data=example; weight count;
 format species spt.;
 tables sex*species /     nocol nopercent cellchi2 chisq;
run;
```

/* LOG-LINEAR MODEL ANALYSES */

```
title3 'Three-way table of sex*species*trap';
title4 'using proc catmod';
proc catmod data=example;
 weight count;
 model sex*species*trap = _response_ / ml nogls noprofile noparm noresponse;
title5 'saturated model: with three-way interaction specified';
 loglin trap|species|sex;
run;
title5 'without three-way interaction specified';
 loglin trap species sex species*sex trap*species trap*sex;
run;
title5 'without trap interactions';
 loglin trap species sex species*sex ;
run;
quit;
```

———————————NEW PROBLEM———————————

Run the SAS program given and compare with results given in the text. Also, test for a trap main effect using traditional contingency table methods (should get χ^2 = 49.7 and G = 51.15 , df = 3).