Multi-way frequency analysis

- A study of gender and eyewear-wearing finds the following frequencies:

Gender	Contacts	Glasses	None
Female	121	32	129
Male	42	37	85

- Is there association between eyewear and gender?
- Normally answer this with chisquare test (based on observed and expected frequencies from null hypothesis of no association).
- Two categorical variables and a frequency.
- We assess in way that generalizes to more categorical variables.

Data format

Data file like this:

```
female contacts 121
female glasses   32
female none     129
male   contacts  42
male   glasses   37
male   none      85
```

as the two categorical variables (gender, type of eyewear) and frequency (number of observations in that category combination).

Some code, using PROC CATMOD

```
data lens;
  infile "lenswear.dat";
  input sex $ lenswear $ frequency;

proc catmod;
  weight frequency;
  model sex*lenswear=_response_;
  loglin sex lenswear sex*lenswear;
```

In PROC CATMOD, specify frequency, then SAS black magic to get right thing, then model (on LOGLIN line!).

Maximum likelihood analysis

```
                    Maximum Likelihood Analysis

            Maximum likelihood computations converged.

             Maximum Likelihood Analysis of Variance

        Source                DF    Chi-Square    Pr > ChiSq
        --------------------------------------------------
        sex                    1         16.10        <.0001
        lenswear               2         64.63        <.0001
        sex*lenswear           2         17.16        0.0002

        Likelihood Ratio       0             .             .
```

- Conclude from `sex*lenswear` line that interaction is significant.
- That is, frequency depends on the sex-lenswear *combination* (not just on either variable singly).
- Or, there is association between sex and lenswear (as usual chisquare test concludes).

Understanding the association

Analysis of Maximum Likelihood Estimates

Parameter		Estimate	Standard Error
sex	female	0.2217	0.0552
lenswear	contacts	0.1146	0.0757
	glasses	-0.6138	0.0889
sex*lenswear	female contacts	0.3074	0.0757
	female glasses	-0.2943	0.0889

Estimates over each variable sum to 0, so complete table as over.

All the estimates

Parameter		Estimate	Error
sex	female	0.2217	0.0552
	male	-0.2217	
lenswear	contacts	0.1146	0.0757
	glasses	-0.6138	0.0889
	none	0.4992	
sex*lenswear	female contacts	0.3074	0.0757
	female glasses	-0.2943	0.0889
	female none	-0.0131	
	male contacts	-0.3074	
	male glasses	0.2943	
	male none	0.0131	

- Look for large (plus or minus) estimates.
- Females more likely to wear contacts and males glasses than expected (if no association).
- Overall, more females in study, and people less likely to wear glasses than other types of eyewear (and most likely to wear none).

Another example: reading, gender and occupation

		Preferred reading		
Profession	Sex	Scifi	Spy	Total
Politician	Male	15	15	30
	Female	10	15	25
	Total	25	30	55
Administrator	Male	10	30	40
	Female	5	10	15
	Total	15	40	55
Bellydancer	Male	5	5	10
	Female	10	25	35
	Total	15	30	45

Altogether 80 males and 75 females.

Data into SAS

This time there are 3 categorical variables (profession, sex, preferred reading) and a frequency. Arrange with one frequency on each line (without totals):

```
politician male scifi 15
politician male spy 15
politician female scifi 10
politician female spy 15
administrator male scifi 10
administrator male spy 30
administrator female scifi 5
administrator female spy 10
bellydancer male scifi 5
bellydancer male spy 5
bellydancer female scifi 10
bellydancer female spy 25
```

The code

```
data small;
  infile "multiway.dat";
  input profession $ sex $ readtype $ freq;

proc catmod;
  weight freq;
  model profession*sex*readtype=_response_;
  loglin profession sex readtype profession*sex
    profession*readtype sex*readtype
    profession*sex*readtype;
```

Loglin line could have been written `profession|sex|readtype` (include main effects and all interactions between variables), but done this way for a reason.

Assessing what to take out

From the “maximum likelihood analysis of variance”:

```
           Maximum Likelihood Analysis of Variance

  Source                        DF    Chi-Square    Pr > ChiSq

  profession                     2          3.46        0.1777
  sex                            1          0.01        0.9256
  readtype                       1          7.61        0.0058
  profession*sex                 2         17.58        0.0002
  profession*readtype            2          2.62        0.2691
  sex*readtype                   1          0.66        0.4168
  profession*sex*readtype        2          1.89        0.3894

  Likelihood Ratio               0           .            .
```

- Model fits perfectly (see Likelihood Ratio line)
- As ANOVA, remove 3-way interaction.
- Change `loglin` line to this:

```
loglin profession sex readtype profession*sex
  profession*readtype sex*readtype;
```

Output from this

```
Maximum Likelihood Analysis of Variance

Source                    DF    Chi-Square    Pr > ChiSq

profession                 2          3.58        0.1674
sex                        1          0.00        0.9453
readtype                   1         13.02        0.0003
profession*sex             2         23.00        <.0001
profession*readtype        2          4.32        0.1155
sex*readtype               1          0.62        0.4321

Likelihood Ratio           2          1.85        0.3969
```

- Bottom line: “is there evidence of lack of fit?” Answer no: model fits OK.
- Now look at two-way interactions and take out non-significant ones.
- Code for that:

```
loglin profession sex readtype profession*sex;
```

Output

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
profession	2	2.90	0.2348
sex	1	0.03	0.8686
readtype	1	12.68	0.0004
profession*sex	2	22.79	<.0001
Likelihood Ratio	5	6.56	0.2557

- Model still fits OK (last line).
- Two-way interaction significant: stays.
- Main effects involving profession and sex have to stay.
- Main effect involving reading type significant, so stays.
- Done. Now interpret the estimates.

The maximum likelihood estimates

with missing ones filled in:

Parameter		Estimate	Standard Error	Chi-Square	Pr > ChiSq
profession	administ	0.0526	0.1257	0.18	0.6753
	bellydan	-0.2169	0.1374	2.49	0.1144
sex	female	0.0149	0.0903	0.03	0.8686
readtype	scifi	-0.2989	0.0839	12.68	0.0004
	spy	0.2989			
profession*sex	administ female	-0.5053	0.1257	16.17	<.0001
	bellydan female	0.6114	0.1374	19.82	<.0001
	politician female	-0.1061			
	administ male	0.5053			
	bellydan male	-0.6114			
	politician male	0.1061			

- Readtype: people overall prefer spy novels
- Interaction: bellydancers tend to be female and administrators male (more so than even split of males/females would suggest).

A different way to read the data

- Entering the words into the data file is repetitive. Start with data as laid out in table (in `freq.dat`):

```
15 15
10 15
10 30
5 10
5 5
10 25
```

- Then use “loops” to associate with variables:

```
data myfreq;
  infile "freq.dat";
  do profession="politician   ","administrator","bellydancer";
    do sex="male  ","female";
      do readtype="scifi","spy";
        input freq @@;
        output;
      end;
    end;
  end;
```

- Resulting data set and PROC CATMOD as before.

Simpson's paradox: the airlines example

	Alaska Airlines		America West	
Airport	On time	Delayed	On time	Delayed
Los Angeles	497	62	694	117
Phoenix	221	12	4840	415
San Diego	212	20	383	65
San Francisco	503	102	320	129
Seattle	1841	305	201	61
Total	3274	501	6438	787

- Alaska: 13.3% flights delayed ($501/(3274+501)$).
- America West: 10.9% ($787/(6438+787)$).
- America West more punctual, right?

Percentage delayed by airport

Airport	Alaska	America West
Los Angeles	11.4	14.4
Phoenix	5.2	7.9
San Diego	8.6	14.5
San Francisco	16.9	28.7
Seattle	14.2	23.2
Total	13.3	10.9

- America West better overall, yet *worse at every single airport*!
- Can PROC CATMOD explain?
- 3 categorical variables (airline, airport, on time/delayed), frequency.

Data for SAS

```
losangeles alaska ontime 497
losangeles alaska delayed 62
losangeles aw ontime 694
losangeles aw delayed 117
phoenix alaska ontime 221
phoenix alaska delayed 12
phoenix aw ontime 4840
phoenix aw delayed 415
...
sanfran alaska ontime 503
sanfran alaska delayed 102
sanfran aw ontime 320
sanfran aw delayed 129
seattle alaska ontime 1841
seattle alaska delayed 305
seattle aw ontime 201
seattle aw delayed 61
```

Code

```
data airline;
  infile "airport.dat";
  input airport $ airline $ status $ freq;

proc catmod;
  weight freq;
  model airport*airline*status=_response_;
  loglin airport|airline|status;
```

Or write out all the effects on the `loglin` line.

Alternative form for data

- Data file:

```
497 62 694 117
221 12 4840 415
212 20 383 65
503 102 320 129
1841 305 201 61
```

- Code to read this:

```
data myfreq;
  infile "freq2.dat";
  do airport="losangeles  ","phoenix","sandiego",
    "sanfrancisco","seattle";
    do airline="alaska     ","americawest";
      do status="ontime ","delayed";
        input freq @@;
        output;
      end;
    end;
  end;
```

Output

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
airport	4	185.99	<.0001
airline	1	118.66	<.0001
airport*airline	4	1138.97	<.0001
status	1	1487.23	<.0001
airport*status	4	99.56	<.0001
airline*status	1	29.09	<.0001
airport*airline*status	4	3.26	0.5156
Likelihood Ratio	0	.	.

- Complicated model fits perfectly (not interesting)
- 3-way interaction non-significant: remove.
- Change loglin line to:

```
loglin airport|airline|status @ 2;
```

(include all interactions $\leq$ 2-way).

Output now

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
airport	4	231.19	<.0001
airline	1	163.72	<.0001
airport*airline	4	3225.58	<.0001
status	1	2700.13	<.0001
airport*status	4	246.27	<.0001
airline*status	1	41.74	<.0001
Likelihood Ratio	4	3.22	0.5223

- Model fits OK (no evidence of lack of fit).
- All 2-way interactions significant: stop here.

Airline by status, adding missing ones

```
                    Analysis of Maximum Likelihood Estimates

                                                   Standard          Chi-   Prob >
 Parameter                             Estimate       Error        Square    ChiSq

....
 airline*status  alaska delayed         -0.1361      0.0211         41.74   <.0001
                 alaska ontime           0.1361
                 aw delayed              0.1361
                 aw ontime              -0.1361
```

- Alaska *more* likely to be on time and America West *more* likely to be delayed, allowing for effects of other variables.
- This in contrast to overall %'s.
- Other interactions shed some light.

Airport by airline

Analysis of Maximum Likelihood Estimates

Parameter		Estimate	Standard Error	Chi-Square	Prob > ChiSq
....					
airport*airline	losangel alaska	-0.0164	0.0261	0.39	0.5303
	phoenix alaska	-1.4049	0.0302	2165.96	<.0001
	sandiego alaska	-0.1618	0.0348	21.57	<.0001
	sanfran alaska	0.3461	0.0287	145.07	<.0001
	seattle alaska	1.2539			

- America West figures negatives of Alaska figures.
- Frequency less than expected for AA into Phoenix (AA flies less often into Phoenix).
- Frequency more than expected for AA into San Francisco and Seattle (AA flies more often into San Francisco and Seattle).
- Conversely, America West flies more into Phoenix and less into San Francisco and Seattle.

Airport by status

Analysis of Maximum Likelihood Estimates

Parameter		Estimate	Standard Error	Chi-Square	Pr > ChiSq
airport*status	losangel delayed	-0.0335	0.0360	0.87	0.3520
	phoenix delayed	-0.4110	0.0305	181.94	<.0001
	sandiego delayed	-0.0762	0.0487	2.44	0.1180
	sanfran delayed	0.3268	0.0343	90.68	<.0001
	seattle delayed	0.1929			

- On-time estimates negatives of delayed figures.
- Fewer flights to Phoenix are delayed (than to other places).
- More flights to San Francisco and Seattle delayed.

Resolution of this Simpson's paradox

- Alaska Airlines flies mostly into San Francisco and Seattle, while America West flies mostly into Phoenix (airport by airline)
- Flights into Phoenix are more likely to be on time, while flights into San Francisco and Seattle are more likely to be delayed.
- In "overall % late", AA gets penalized for flying into airports where hard to be on time.
- When you allow for who flies where, AA comes out more punctual (as seen in airport-by-airport statistics).

Ovarian cancer: a four-way table

- Retrospective study of ovarian cancer done in 1973.
- Information about 299 women operated on for ovarian cancer 10 years previously.
- Recorded:
  - stage of cancer (early or advanced)
  - type of operation (radical or limited)
  - X-ray treatment received (yes or no)
  - 10-year survival (yes or no)
- Survival looks like response (suggests logistic regression). PROC CATMOD finds any associations at all.

The data

for SAS purposes:

```
early radical no no 10
early radical no yes 41
early radical yes no 17
early radical yes yes 64
early limited no no 1
early limited no yes 13
early limited yes no 3
early limited yes yes 9
advanced radical no no 38
advanced radical no yes 6
advanced radical yes no 64
advanced radical yes yes 11
advanced limited no no 3
advanced limited no yes 1
advanced limited yes no 13
advanced limited yes yes 5
```

Stage, type, x-ray, survival, frequency.

The code

hopefully looking familiar by now:

```
data cancer;
  infile "cancer.dat";
  input stage $ operation $ xray $ survival $ count;

proc catmod;
  weight count;
  model stage*operation*xray*survival=_response_;
  loglin stage|operation|xray|survival;
```

Alternative data entry

- Data like this:
  ```
  10 41 17 64 1 13 3 9
  38 6 64 11 3 1 13 5
  ```
- All values for each stage first. Within each stage, all values for kind of operation; within these, all values for X-ray, then all values for survival:

```
data freq;
  infile "freq3.dat";
  do stage="early    ","advanced";
    do operation="radical","limited";
      do xray="no ","yes";
        do survival="no ","yes";
          input count @@;
          output;
        end;
      end;
    end;
  end;
```

Output #1

```
            Maximum Likelihood Analysis of Variance
Source                         DF    Chi-Square     Pr > ChiSq
operation*xray                  1          0.80         0.3712
stage*operation*xray            1          1.33         0.2495
survival                        1          0.15         0.6979
stage*survival                  1         40.09         <.0001
operation*survival              1          1.69         0.1930
stage*operation*survival        1          0.11         0.7425
xray*survival                   1          0.48         0.4871
stage*xray*survival             1          0.87         0.3502
operation*xray*survival         1          0.48         0.4874
stage*operat*xray*surviv        1          0.57         0.4499

Likelihood Ratio                0           .            .
```

- Four-way interaction and all 3-way interactions not significant: remove all, and check resulting model for fit.
- Change loglin line to this:
  `loglin stage|operation|xray|survival @ 2;`
  that is, keep main effects and interactions up to 2-way.

Output #2

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
stage	1	0.27	0.6033
operation	1	102.15	<.0001
stage*operation	1	0.59	0.4415
xray	1	10.01	0.0016
stage*xray	1	0.62	0.4324
operation*xray	1	0.01	0.9326
survival	1	0.23	0.6294
stage*survival	1	99.45	<.0001
operation*survival	1	2.06	0.1511
xray*survival	1	0.09	0.7696
Likelihood Ratio	5	7.17	0.2084

- Model still fits all right.
- Only significant 2-way interaction is stage by survival.
- Take out others and check fit again.
- Change loglin line to

```
loglin stage operation xray survival stage*survival;
```

Output #3

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
stage	1	1.50	0.2202
operation	1	110.28	<.0001
xray	1	17.46	<.0001
survival	1	0.55	0.4584
stage*survival	1	100.74	<.0001
Likelihood Ratio	10	10.99	0.3583

- Model fit still OK (no evidence of lack of fit)
- Stage and survival main effects have to stay.
- Operation and X-ray main effects are significant, so they stay.
- Done. Interpret maximum likelihood estimates.

Maximum likelihood estimates

Analysis of Maximum Likelihood Estimates

Parameter		Estimate	Standard Error	Chi-Square	Pr > ChiSq
stage	advanced	-0.0930	0.0759	1.50	0.2202
operation	limited	-0.8271	0.0788	110.28	<.0001
xray	no	-0.2492	0.0596	17.46	<.0001
survival	no	0.0562	0.0759	0.55	0.4584
stage*survival	advanced no	0.7613	0.0759	100.74	<.0001

- Stage by survival interaction: stage of cancer and survival associated. Higher frequency with being in advanced stage and not surviving: advanced stage associated with non-survival.
- Fewer women had the limited operation (more had the radical one)
- Fewer woman had no X-ray treatment (more did have X-ray treatment).
- Interaction with “response” (survival) usually of most interest.

General procedure

- Start with “complete model” including all possible interactions.
- Look at highest-order interaction(s) remaining, remove if non-significant.
- If an interaction significant, keep also everything contained within that interaction. Eg. A*B interaction significant, keep A and B main effects.
- Continue until everything either significant or must be kept.
- Then look at maximum likelihood estimates (can fill in those not shown) and interpret according to whether $+$ or $-$.
- Main effects not usually very interesting.
- Interactions with “response” usually of most interest: show association with response.