Multiway Frequency Tables

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.

The data file

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

- This is *not tidy!*
- Two variables are gender and *eyewear*, and those numbers all frequencies.

```
my_url <- "http://www.utsc.utoronto.ca/~butler/d29/eyewear.txt"
(eyewear <- read_delim(my_url, " "))
```

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

Tidying the data

```
eyewear %>%
  pivot_longer(contacts:none, names_to="eyewear",
               values_to="frequency") -> eyes
eyes
```

gender	eyewear	frequency
female	contacts	121
female	glasses	32
female	none	129
male	contacts	42
male	glasses	37
male	none	85

Making tidy data back into a table

- use `spread`
- or this (we use it again later):

```
xt <- xtabs(frequency ~ gender + eyewear, data = eyes)
xt
```

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

Modelling

- Predict frequency from other factors and combos.
- `glm` with `poisson` family.

```r
eyes.1 <- glm(frequency ~ gender * eyewear,
  data = eyes,
  family = "poisson"
)
```

- Called **log-linear model**.

What can we get rid of?

```r
tidy(drop1(eyes.1, test = "Chisq"))
```

```
## Warning in tidy.anova(drop1(eyes.1, test =
## "Chisq")): The following column names in ANOVA
## output were not recognized or transformed: Deviance,
## LRT
```

term	df	Deviance	AIC	LRT	p.value
	NA	0.00000	47.95815	NA	NA
gender:eyewear	2	17.82863	61.78678	17.82863	0.0001345

nothing!

Conclusions

- `drop1` says what we can remove at this step. Significant = must stay.
- Cannot remove anything.
- Frequency depends on `gender-wear` *combination*, cannot be simplified further.
- Gender and eyewear are *associated*.
- Stop here.

prop.table

Original table:

```
xt
```

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

Calculate eg. row proportions like this:

```
prop.table(xt, margin = 1)
```

```
##         eyewear
## gender    contacts   glasses      none
##   female 0.4290780 0.1134752 0.4574468
##   male   0.2560976 0.2256098 0.5182927
```

Comments

- `margin` says what to make add to 1.
- More females wear contacts and more males wear glasses.

No association

- Suppose table had been as shown below:

```
my_url <- "http://www.utsc.utoronto.ca/~butler/d29/eyewear2.txt"
eyewear2 <- read_table(my_url)
eyes2 <- eyewear2 %>% gather(eyewear, frequency, contacts:none)
xt2 <- xtabs(frequency ~ gender + eyewear, data = eyes2)
xt2
```

```
##          eyewear
## gender   contacts glasses none
##   female      150      30  120
##   male         75      16   62
```

```
prop.table(xt2, margin = 1)
```

```
##          eyewear
## gender     contacts   glasses      none
##   female 0.5000000 0.1000000 0.4000000
##   male   0.4901961 0.1045752 0.4052288
```

Comments

- Females and males wear contacts and glasses *in same proportions*
    - though more females and more contact-wearers.
- No *association* between gender and eyewear.

Analysis for revised data

```r
eyes.2 <- glm(frequency ~ gender * eyewear,
  data = eyes2,
  family = "poisson"
)
tidy(drop1(eyes.2, test = "Chisq"))
```

```
## Warning in tidy.anova(drop1(eyes.2, test =
## "Chisq")): The following column names in ANOVA
## output were not recognized or transformed: Deviance,
## LRT
```

term	df	Deviance	AIC	LRT	p.value
	NA	0.0000000	47.46718	NA	NA
gender:eyewear	2	0.0473227	43.51450	0.0473227	0.9766164

No longer any association. Take out interaction.

No interaction

```r
eyes.3 <- update(eyes.2, . ~ . - gender:eyewear)
tidy(drop1(eyes.3, test = "Chisq"))
```

```
## Warning in tidy.anova(drop1(eyes.3, test =
## "Chisq")): The following column names in ANOVA
## output were not recognized or transformed: Deviance,
## LRT
```

term	df	Deviance	AIC	LRT	p.value
	NA	0.0473227	43.51450	NA	NA
gender	1	48.6238913	90.09107	48.57657	0
eyewear	2	138.1304141	177.59759	138.08309	0

- More females (gender effect)
- more contact-wearers (eyewear effect)
- no association (no interaction).

Chest pain, being overweight and being a smoker

- In a hospital emergency department, 176 subjects who attended for acute chest pain took part in a study.
- Each subject had a normal or abnormal electrocardiogram reading (ECG), were overweight (as judged by BMI) or not, and were a smoker or not.
- How are these three variables related, or not?

The data

In modelling-friendly format:

```
ecg bmi smoke count
abnormal overweight yes 47
abnormal overweight no 10
abnormal normalweight yes 8
abnormal normalweight no 6
normal overweight yes 25
normal overweight no 15
normal normalweight yes 35
normal normalweight no 30
```

First step

```r
my_url <- "http://www.utsc.utoronto.ca/~butler/d29/ecg.txt"
chest <- read_delim(my_url, " ")
chest.1 <- glm(count ~ ecg * bmi * smoke,
  data = chest,
  family = "poisson"
)
tidy(drop1(chest.1, test = "Chisq"))
```

```
## Warning in tidy.anova(drop1(chest.1, test =
## "Chisq")): The following column names in ANOVA
## output were not recognized or transformed: Deviance,
## LRT
```

term	df	Deviance	AIC	LRT	p.value
	NA	0.000000	53.70730	NA	NA
ecg:bmi:smoke	1	1.388544	53.09584	1.388544	0.2386511

That 3-way interaction comes out

Removing the 3-way interaction

```
chest.2 <- update(chest.1, . ~ . - ecg:bmi:smoke)
tidy(drop1(chest.2, test = "Chisq"))
```

```
## Warning in tidy.anova(drop1(chest.2, test =
## "Chisq")): The following column names in ANOVA
## output were not recognized or transformed: Deviance,
## LRT
```

term	df	Deviance	AIC	LRT	p.value
	NA	1.388544	53.09584	NA	NA
ecg:bmi	1	29.019513	78.72681	27.630969	0.0000001
ecg:smoke	1	4.893513	54.60081	3.504968	0.0611850
bmi:smoke	1	4.468863	54.17616	3.080319	0.0792450

At $\alpha = 0.05$, `bmi:smoke` comes out.

Removing bmi:smoke

```
chest.3 <- update(chest.2, . ~ . - bmi:smoke)
tidy(drop1(chest.3, test = "Chisq"))
```

```
## Warning in tidy.anova(drop1(chest.3, test =
## "Chisq")): The following column names in ANOVA
## output were not recognized or transformed: Deviance,
## LRT
```

term	df	Deviance	AIC	LRT	p.value
	NA	4.468863	54.17616	NA	NA
ecg:bmi	1	36.562474	84.26977	32.09361	0.0000000
ecg:smoke	1	12.436473	60.14377	7.96761	0.0047622

`ecg:smoke` has become significant. So we have to stop.

Understanding the final model

- Thinking of ecg as “response” that might depend on anything else.
- What is associated with ecg? Both bmi on its own and smoke on its own, but *not* the combination of both.
- ecg:bmi table:

```
xtabs(count ~ ecg + bmi, data = chest)
```

```
##           bmi
## ecg        normalweight overweight
##   abnormal           14         57
##   normal             65         40
```

- Most normal weight people have a normal ECG, but a majority of overweight people have an *abnormal* ECG. That is, knowing about BMI says something about likely ECG.

ecg:smoke

- ecg:smoke table:

```
xtabs(count ~ ecg + smoke, data = chest)
```

```
##             smoke
## ecg           no yes
##   abnormal 16  55
##   normal   45  60
```

- Most nonsmokers have a normal ECG, but smokers are about 50–50 normal and abnormal ECG.
- Don't look at smoke:bmi table since not significant.

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

Use `status` as variable name for "on time/delayed".

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

Arranging the data

- Can only have single thing in columns, so we have to construct column names like this:

```
airport     aa_ontime aa_delayed aw_ontime aw_delayed
LosAngeles    497            62       694        117
Phoenix       221            12      4840        415
SanDiego      212            20       383         65
SanFrancisco  503           102       320        129
Seattle      1841           305       201         61
```

- Read in:

```
my_url <- "http://www.utsc.utoronto.ca/~butler/d29/airlines.tx
airlines <- read_table2(my_url)
```

Tidying

- Some tidying gets us the right layout, with frequencies all in one column and the airline and delayed/on time status separated out:

```r
airlines %>%
  gather(line.status, freq, contains("_")) %>%
  separate(line.status, c("airline", "status")) -> punctual
```

- See how this works by running it one line at a time.

The data frame `punctual`

airport	airline	status	freq
LosAngeles	aa	ontime	497
Phoenix	aa	ontime	221
SanDiego	aa	ontime	212
SanFrancisco	aa	ontime	503
Seattle	aa	ontime	1841
LosAngeles	aa	delayed	62
Phoenix	aa	delayed	12
SanDiego	aa	delayed	20
SanFrancisco	aa	delayed	102
Seattle	aa	delayed	305
LosAngeles	aw	ontime	694
Phoenix	aw	ontime	4840
SanDiego	aw	ontime	383
SanFrancisco	aw	ontime	320
Seattle	aw	ontime	201
LosAngeles	aw	delayed	117
Phoenix	aw	delayed	415
SanDiego	aw	delayed	65
SanFrancisco	aw	delayed	129
Seattle	aw	delayed	61

Proportions delayed by airline

- Two-step process: get appropriate subtable:

```
xt <- xtabs(freq ~ airline + status, data = punctual)
xt
```

```
##        status
## airline delayed ontime
##      aa     501   3274
##      aw     787   6438
```

- and then calculate appropriate proportions:

```
prop.table(xt, margin = 1)
```

```
##        status
## airline   delayed    ontime
##      aa 0.1327152 0.8672848
##      aw 0.1089273 0.8910727
```

- More of Alaska Airlines' flights delayed (13.3% vs. 10.9%).

Proportion delayed by airport, for each airline

```r
xt <- xtabs(freq ~ airline + status + airport, data = punctual)
xp <- prop.table(xt, margin = c(1, 3))
ftable(xp,
  row.vars = c("airport", "airline"),
  col.vars = "status"
)
```

```
##                         status    delayed     ontime
## airport      airline                                
## LosAngeles   aa                0.11091234 0.88908766
##              aw                0.14426634 0.85573366
## Phoenix      aa                0.05150215 0.94849785
##              aw                0.07897241 0.92102759
## SanDiego     aa                0.08620690 0.91379310
##              aw                0.14508929 0.85491071
## SanFrancisco aa                0.16859504 0.83140496
##              aw                0.28730512 0.71269488
## Seattle      aa                0.14212488 0.85787512
##              aw                0.23282443 0.76717557
```

Simpson's Paradox

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 more punctual overall,
- but worse at *every single* airport!
- How is that possible?
- Log-linear analysis sheds some light.

Model 1 and output

```
punctual.1 <- glm(freq ~ airport * airline * status,
  data = punctual, family = "poisson"
)
tidy(drop1(punctual.1, test = "Chisq"))
```

```
## Warning in tidy.anova(drop1(punctual.1, test =
## "Chisq")): The following column names in ANOVA
## output were not recognized or transformed: Deviance,
## LRT
```

term	df	Deviance	AIC	LRT	p.value
	NA	0.000000	183.4348	NA	NA
airport:airline:status	4	3.216569	178.6513	3.216569	0.5222589

Remove 3-way interaction

```
punctual.2 <- update(punctual.1, ~ . - airport:airline:status)
tidy(drop1(punctual.2, test = "Chisq"))
```

```
## Warning in tidy.anova(drop1(punctual.2, test =
## "Chisq")): The following column names in ANOVA
## output were not recognized or transformed: Deviance,
## LRT
```

term	df	Deviance	AIC	LRT	p.value
	NA	3.216569	178.6513	NA	NA
airport:airline	4	6432.454138	6599.8889	6429.23757	0
airport:status	4	240.107798	407.5426	236.89123	0
airline:status	1	45.465141	218.8999	42.24857	0

Stop here.

Understanding the significance

- airline:status:

```
xt <- xtabs(freq ~ airline + status, data = punctual)
prop.table(xt, margin = 1)
```

```
##         status
## airline    delayed    ontime
##      aa 0.1327152 0.8672848
##      aw 0.1089273 0.8910727
```

- More of Alaska Airlines' flights delayed overall.
- Saw this before.

Understanding the significance (2)

- `airport:status`:

```
xt <- xtabs(freq ~ airport + status, data = punctual)
prop.table(xt, margin = 1)
```

```
##                status
## airport            delayed     ontime
##   LosAngeles   0.13065693 0.86934307
##   Phoenix      0.07780612 0.92219388
##   SanDiego     0.12500000 0.87500000
##   SanFrancisco 0.21916509 0.78083491
##   Seattle      0.15199336 0.84800664
```

- Flights into San Francisco (and maybe Seattle) are often late, and flights into Phoenix are usually on time.
- Considerable variation among airports.

Understanding the significance (3)

- airport:airline:

```
xt <- xtabs(freq ~ airport + airline, data = punctual)
prop.table(xt, margin = 2)
```

```
##                airline
## airport                 aa         aw
##   LosAngeles    0.14807947 0.11224913
##   Phoenix       0.06172185 0.72733564
##   SanDiego      0.06145695 0.06200692
##   SanFrancisco 0.16026490 0.06214533
##   Seattle       0.56847682 0.03626298
```

- What fraction of each airline's flights are to each airport.
- Most of Alaska Airlines' flights to Seattle and San Francisco.
- Most of America West's flights to Phoenix.

The resolution

- Most of America West's flights to Phoenix, where it is easy to be on time.
- Most of Alaska Airlines' flights to San Francisco and Seattle, where it is difficult to be on time.
- Overall comparison looks bad for Alaska because of this.
- But, *comparing like with like*, if you compare each airline's performance *to the same airport*, Alaska does better.
- Aggregating over the very different airports was a (big) mistake: that was the cause of the Simpson's paradox.
- Alaska Airlines is *more* punctual when you do the proper comparison.

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).
- Log-linear model finds any associations at all.

The data

after tidying:

```
stage operation xray survival freq
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
```

Reading in data

```r
my_url <- "http://www.utsc.utoronto.ca/~butler/d29/cancer.txt"
cancer <- read_delim(my_url, " ")
cancer %>% slice(1:6)
```

stage	operation	xray	survival	freq
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

Model 1

hopefully looking familiar by now:

```r
cancer.1 <- glm(freq ~ stage * operation * xray * survival,
  data = cancer, family = "poisson"
)
```

Output 1

See what we can remove:

```r
tidy(drop1(cancer.1, test = "Chisq"))
```

```
## Warning in tidy.anova(drop1(cancer.1, test =
## "Chisq")): The following column names in ANOVA
## output were not recognized or transformed: Deviance,
## LRT
```

term	df	Deviance	AIC	LRT	p.value
	NA	0.0000000	98.12961	NA	NA
stage:operation:xray:survival	1	0.6026558	96.73227	0.6026558	0.4375665

Non-significant interaction can come out.

Model 2

```
cancer.2 <- update(cancer.1, . ~ . - stage:operation:xray:survival)
tidy(drop1(cancer.2, test = "Chisq"))
```

```
## Warning in tidy.anova(drop1(cancer.2, test =
## "Chisq")): The following column names in ANOVA
## output were not recognized or transformed: Deviance,
## LRT
```

term	df	Deviance	AIC	LRT	p.value
	NA	0.6026558	96.73227	NA	NA
stage:operation:xray	1	2.3575888	96.48720	1.7549331	0.1852578
stage:operation:survival	1	1.1773024	95.30692	0.5746466	0.4484184
stage:xray:survival	1	0.9557671	95.08538	0.3531113	0.5523571
operation:xray:survival	1	1.2337838	95.36340	0.6311281	0.4269418

Least significant term is `stage:xray:survival`: remove.

Take out stage:xray:survival

```
cancer.3 <- update(cancer.2, . ~ . - stage:xray:survival)
tidy(drop1(cancer.3, test = "Chisq"))
```

```
## Warning in tidy.anova(drop1(cancer.3, test =
## "Chisq")): The following column names in ANOVA
## output were not recognized or transformed: Deviance,
## LRT
```

term	df	Deviance	AIC	LRT	p.value
	NA	0.9557671	95.08538	NA	NA
stage:operation:xray	1	3.0866591	95.21627	2.1308920	0.1443567
stage:operation:survival	1	1.5660529	93.69567	0.6102858	0.4346802
operation:xray:survival	1	1.5512410	93.68085	0.5954739	0.4403102

`operation:xray:survival` comes out next.

Remove operation:xray:survival

```r
cancer.4 <- update(cancer.3, . ~ . - operation:xray:survival)
tidy(drop1(cancer.4, test = "Chisq"))
```

```
## Warning in tidy.anova(drop1(cancer.4, test =
## "Chisq")): The following column names in ANOVA
## output were not recognized or transformed: Deviance,
## LRT
```

term	df	Deviance	AIC	LRT	p.value
	NA	1.551241	93.68085	NA	NA
xray:survival	1	1.697682	91.82729	0.1464406	0.7019603
stage:operation:xray	1	6.841961	96.97157	5.2907197	0.0214394
stage:operation:survival	1	1.931103	92.06072	0.3798619	0.5376771

Comments

- `stage:operation:xray` has now become significant, so won't remove that.
- Shows value of removing terms one at a time.
- There are no higher-order interactions containing both `xray` and `survival`, so now we get to test (and remove) `xray:survival`.

Remove `xray:survival`

```r
cancer.5 <- update(cancer.4, . ~ . - xray:survival)
tidy(drop1(cancer.5, test = "Chisq"))
```

```
## Warning in tidy.anova(drop1(cancer.5, test =
## "Chisq")): The following column names in ANOVA
## output were not recognized or transformed: Deviance,
## LRT
```

term	df	Deviance	AIC	LRT	p.value
	NA	1.697682	91.82729	NA	NA
stage:operation:xray	1	6.927690	95.05730	5.2300086	0.0222004
stage:operation:survival	1	2.024220	90.15383	0.3265384	0.5677045

Remove stage:operation:survival

```
cancer.6 <- update(cancer.5, . ~ . - stage:operation:survival)
tidy(drop1(cancer.6, test = "Chisq"))
```

```
## Warning in tidy.anova(drop1(cancer.6, test =
## "Chisq")): The following column names in ANOVA
## output were not recognized or transformed: Deviance,
## LRT
```

term	df	Deviance	AIC	LRT	p.value
	NA	2.024220	90.15383	NA	NA
stage:survival	1	135.197636	221.32725	133.173416	0.0000000
operation:survival	1	4.115730	90.24534	2.091510	0.1481196
stage:operation:xray	1	7.254229	93.38384	5.230009	0.0222004

Last step?

Remove operation:survival.

```r
cancer.7 <- update(cancer.6, . ~ . - operation:survival)
tidy(drop1(cancer.7, test = "Chisq"))
```

```
## Warning in tidy.anova(drop1(cancer.7, test =
## "Chisq")): The following column names in ANOVA
## output were not recognized or transformed: Deviance,
## LRT
```

term	df	Deviance	AIC	LRT	p.value
	NA	4.115730	90.24534	NA	NA
stage:survival	1	136.729112	220.85872	132.613382	0.0000000
stage:operation:xray	1	9.345738	93.47535	5.230009	0.0222004

Finally done!

Conclusions

- What matters is things associated with `survival` (`survival` is "response").
- Only significant such term is `stage:survival`:

```
xt <- xtabs(freq ~ stage + survival, data = cancer)
prop.table(xt, margin = 1)
```

```
##           survival
## stage             no       yes
##   advanced 0.8368794 0.1631206
##   early    0.1962025 0.8037975
```

- Most people in early stage of cancer survived, and most people in advanced stage did not survive.
- This true *regardless* of type of operation or whether or not X-ray treatment was received. These things have no impact on survival.

What about that other interaction?

```
xt <- xtabs(freq ~ operation + xray + stage, data = cancer)
ftable(prop.table(xt, margin = 3))
```

```
##                stage   advanced      early
## operation xray                            
## limited   no         0.02836879 0.08860759
##           yes        0.12765957 0.07594937
## radical   no         0.31205674 0.32278481
##           yes        0.53191489 0.51265823
```

- Out of the people at each stage of cancer (since `margin=3` and `stage` was listed 3rd).
- The association is between `stage` and `xray` *only for those who had the limited operation*.
- For those who had the radical operation, there was no association between `stage` and `xray`.
- This is of less interest than associations with `survival`.

General procedure

- Start with “complete model” including all possible interactions.
- `drop1` gives highest-order interaction(s) remaining, remove least non-significant.
- Repeat as necessary until everything significant.
- Look at subtables of significant interactions.
- Main effects not usually very interesting.
- Interactions with “response” usually of most interest: show association with response.