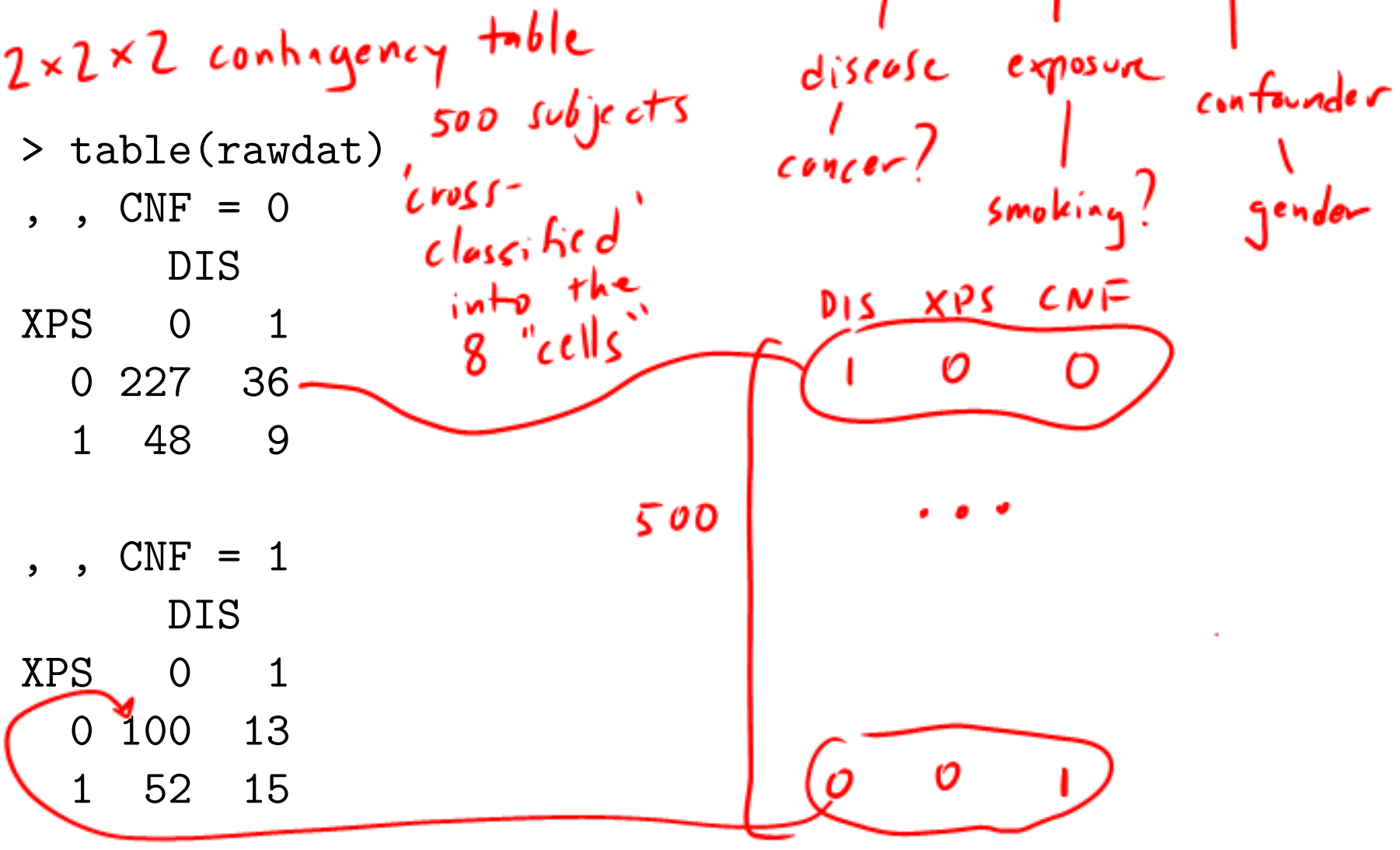
STATISTICS 538, Lecture #10

Log-Linear Models

November 24, 2010

Consider these data: binary variables DIS, XPS, CNF measured on 500 subjects




```
> table(rawdat)
, , CNF = 0
      DIS
XPS    0   1
  0 227  36
  1  48   9

, , CNF = 1
      DIS
XPS    0   1
  0 100  13
  1  52  15
```

Possibly sampled as cohort: $n = 500$ planned in advance

joint sampling (DIS, XPS, CNF) but only interested in (DIS | XPS, CNF)

OR

(maybe even decided on (XPS,CNF) proportions in advance)

```
> glm(DIS ~ XPS + CNF, family = binomial, data = rawdat)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.9048     0.1736 -10.970   <2e-16 ***
XPS           0.4657     0.2816   1.654   0.0982 .
CNF           0.0222     0.2690   0.083   0.9342
```

$\hat{\beta}$

$$\frac{\text{Odds}(DIS=1 \mid XPS=1, CNF)}{\text{Odds}(DIS=1 \mid XPS=0, CNF)} = e^{\beta}$$

Possibly **case-control** sampling: decided in advance to recruit 427 controls (DIS=0), 73 cases (DIS=1)

i.e. sampling XPS | DIS
not DIS | XPS

efficient study design for a rare disease

```
Call:
> glm(XPS ~ DIS + CNF, family = binomial, data = rawdat)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.6031     0.1547 -10.361  < 2e-16 ***
DIS           0.4657     0.2816   1.654   0!0982 .
CNF           1.0050     0.2131   4.717 2.40e-06 ***
```

same as before!

not surprising? either way going after (XPS, DIS) odds ratio conditioned on CNF

Possibly just recruited 'as many as possible' subjects over fixed time period

i.e. $n = 500$ usually fixed in advance

```
> agdata <- as.data.frame(table(rawdat))

> agdata
  XPS DIS CNF Freq
1   0   0   0  227
2   1   0   0   48
3   0   1   0   36
4   1   1   0    9
5   0   0   1  100
6   1   0   1   52
7   0   1   1   13
8   1   1   1   15
```

Fit model to explain these counts given all three variables?

Fitting Poisson model (with interactions) to cell counts

using log link

```
> glm(Freq ~ .^2, family = poisson, data = agdata)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.43340    0.06564  82.777  < 2e-16 ***
XPS1        -1.60314    0.15473 -10.361  < 2e-16 ***
DIS1        -1.90485    0.17365 -10.970  < 2e-16 ***
CNF1        -0.84767    0.11790  -7.190 6.49e-13 ***
XPS1:DIS1    0.46570    0.28159   1.654   0.0982 .
XPS1:CNF1    1.00502    0.21307   4.717 2.40e-06 ***
DIS1:CNF1    0.02220    0.26902   0.083   0.9342
```

... + β XPS × DIS

What does a 'log-linear' model for counts imply?

Say $Y_{abc} = \#\{A = a, B = b, C = c\}$, and model

Freq

A, B, C binary

$$\log E(Y_{abc}) = \beta_0 + \beta_a a + \beta_b b + \beta_c c + \beta_{ab} ab + \beta_{ac} ac + \beta_{bc} bc$$

$$Pr(A = 1 | B = b, C = c) = \frac{E(Y_{1bc})}{E(Y_{0bc}) + E(Y_{1bc})}$$

$$= \frac{e^{\beta_a + \beta_{ab} b + \beta_{ac} c + \text{stuff}}}{e^{0 + \text{stuff}} + e^{\beta_a + \beta_{ab} b + \beta_{ac} c + \text{stuff}}}$$

← all terms not involving a

$$= \frac{e^{\beta_a + \beta_{ab} b + \beta_{ac} c}}{1 + e^{\beta_a + \beta_{ab} b + \beta_{ac} c}} = \text{expit}\{\beta_a + \beta_{ab} b + \beta_{ac} c\}$$

So, log-linear model for cell counts embeds logistic regression models for one variable given the others

reflects the point of agreement between cohort & case-control analyses

- Note that β_{ab} is both the main effect of B in $(A|B, C)$ and the main effect of A in $(B|A, C)$
- The 'embedding' is very general - any number of categorical variables, with multinomial logit models arising for variables with more than two levels.
- The embedding scales up to higher-order interactions, e.g., an $A \times B \times C$ term in the log linear model induces a $B \times C$ interaction in the $(A|B, C)$ model.

$A, B, C, D, \ldots$

e.g. $A = \{0, 1, 2\}$, $B = \{0, 1\}$

$I\{B=1\}$

$\beta_{ab,2} I\{A=2\} B$

$$\beta_{a,1} I\{A=1\} + \beta_{a,2} I\{A=2\} + \beta_b B + \beta_{ab,1} I\{A=1\} B + \ldots$$

Applied practice somewhat 'loose,' fit log-linear models without too much worry about actual sampling scheme

But ... beware of the 'minimal model' concept

For instance, note that our toy dataset involves:

```
> xtabs(Freq~CNF+XPS, data=agdata)
```

CNF	XPS 0	XPS 1	
0	263	57	320
1	113	67	180
	376	124	500

for quantities (i.e., sample sizes) that are "fixed by design," fitted values should match exactly

Now, what if the data arose via a random size of size 500

fixed in advance or "by design"

```
> ft1 <- glm(Freq ~ 1, family=poisson, data=agdata)

> fitted(ft1)
   1    2    3    4    5    6    7    8
62.5 62.5 62.5 62.5 62.5 62.5 62.5 62.5

> xtabs(fitted(ft1)~CNF+XPS, data=agdata)

      XPS
CNF    0   1
  0  125 125 | 250
  1  125 125 | 250
          ----|----
          500
```

minimal model, but not very interesting - postulates

$Pr(DIS=1 \mid XPS, CNF) = \frac{1}{2}$

i.e. fitted values reproduce the a priori chosen sample size - regarded as a good thing

What if the data arose via random samples of size ~~180~~ 320 (CNF=0) and ~~120~~ 180 (CNF=1)?

```
> ft2 <- glm(Freq ~ CNF, family=poisson, data=agdata)

> fitted(ft2)
 1  2  3  4  5  6  7  8
80 80 80 80 45 45 45 45

> xtabs(fitted(ft2) ~ CNF+XPS, data=agdata)

      XPS
CNF    0   1
  0  160 160 | 320
  1   90  90 | 180
             | 500
```

still not interesting

$Pr(DIS=1 \mid XPS, CNF) = \frac{1}{2}$

reproduces the "fixed by design" totals

What if the data arose via stratified sampling for each (CNF,XPS) combo, sizes 263, 57,113, 67

```
> ft3 <- glm(Freq ~ CNF + XPS + CNF:XPS, family=poisson,
             data=agdata)
```

$Pr(DIS=1 \mid XPS, CNF) = \frac{1}{2}$

still not interesting!

```
> fitted(ft3)
    1     2     3     4     5     6     7     8
131.5  28.5 131.5  28.5  56.5  33.5  56.5  33.5
```

```
> xtabs(fitted(ft3) ~ CNF+XPS, data=agdata)

   XPS
CNF   0   1
  0 263  57
  1 113  67
```

reproduces the fixed stuff ✓