Ch. 2 Contingency Tables

Two-way contingency tables

Example : Physicians Health Study (5 year)

		Heart Attack YES	NO	Total
Group	Placebo	189	10,845	11,034
	Aspirin	104	10,933	11,037

(2×2 table)

Contingency table – Cells contain counts of outcomes. $I \times J$ table has I rows and J columns.

● A conditional distribution refers to prob. dist. of Y at fixed level of x.

Example.

		Y YES	NO	Total
X	Placebo	0.017	0.983	1.0
	Aspirin	0.009	0.991	1.0

Sample conditional dist. for placebo group is

$$0.017 = \frac{189}{11{,}034},\ 0.983 = \frac{10{,}845}{13{,}034}$$

Natural way to look at data when

Y = response variable
X = explanatory variable

Example. Diagnostic disease tests
Y = outcome of test : 1=positive, 2=negative
X = reality : 1=diseased, 2=not diseased
Test result

		Y 1	2	Total
X	1			
	2			

Sensitivity = $P(Y=1|X=1)$: Given that the subject has the disease, the prob. the diagnostic test is positive

Specificity = $P(Y=2|X=2)$: Given that the subject does not have the disease, the prob. the diagnostic test is negative

In practice, if you get positive result, more relevant to you is $P(X=1|Y=1)$. This may be low even if sensitivity and specificity are high (See pp 23-24 of Text for example of how this can happen when disease is relatively rare)

● What if X, Y both response variables?
$\{\pi_{ij}\} = \{P(X=x_i, Y=y_j)\}$ from the joint distribution of X and Y

	Y		
X	π_{11}	π_{12}	π_{1+}
	π_{21}	π_{22}	π_{2+}
	π_{+1}	π_{+2}	1.0

Marginal Prob.'s

Sample cell count $\{n_{ij}\}$

Sample cell proportion $\{p_{ij}\}$, $p_{ij} = \frac{n_{ij}}{n}$ $with\ n = \sum_i \sum_j n_{ij}$

Def. X and Y are <u>statistically independent</u> if true conditional dist. of Y is identical at each level of X

For example

	Y	
X	0.01	0.99
	0.01	0.99

Then, $\pi_{ij} = \pi_{i+}\pi_{+j}$, $all\ i,j$
i.e, $P(X=i, Y=j) = P(X=i)P(Y=j)$, $such\ as$

		Y: 1	2	Total
X	1	0.28	0.42	0.7
	2	0.12	0.18	0.3
		0.4	0.6	1.0

Comparing proportions in 2×2 Tables

		Y	
		S	F
X	1	π_1	$1-\pi_1$
	2	π_2	$1-\pi_2$

Conditional Distributions

$$\widehat{\pi_1}-\widehat{\pi_2}=p_1-p_2$$

$$SE(p_1-p_2)=\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}$$

Example

$$p_1=0.017,\ p_2=0.009, p_1-p_2=0.008$$

$$SE=\sqrt{\frac{0.017\times0.983}{11{,}034}+\frac{0.009\times0.991}{11{,}037}}=0.0015$$

95% C.I. for $\pi_1-\pi_2$ is

$$0.008\pm1.96(0.0015)=(0.005,\ 0.011)$$

Apparently $\pi_1-\pi_2>0(i.e.,\ \pi_1>\pi_2)$

Relative Risk $=\frac{\pi_1}{\pi_2}$

Example : sample $\frac{p_1}{p_2}=\frac{0.017}{0.009}=1.82$

Sample proportion of heart attacks was 82% higher for placebo group.

95% C.I. for $\log\left(\frac{\pi_1}{\pi_2}\right)$ is

$$\log\left(\frac{p_1}{p_2}\right)\pm1.96\sqrt{\frac{1-p_1}{n_1p_1}+\frac{1-p_2}{n_2p_2}}$$

95% C.I. is (1.43, 2.31)

Independence $\Leftrightarrow$ $\frac{\pi_1}{\pi_2}=1.0$

Odds Ratio

		Y	
		S	F
Group	1	π_1	$1-\pi_1$
	2	π_2	$1-\pi_2$

The odds the response is a S (success) instead of an F (failure) $=\frac{prob.(S)}{prob.(F)}$.

$$= \frac{\pi_1}{(1-\pi_1)} \quad in\ row1$$

$$= \frac{\pi_2}{(1-\pi_2)} \quad in\ row2$$

eg., if odds=3. S three times as likely as F.

if odds=1/3, F three times as likely as S.

odds=3 $\Rightarrow$ $P(S)=3/4,\ P(F)=1/4$

$$P(S)=\frac{odds}{1+odds}$$

odds=1/3 $\Rightarrow$ $P(S)=\dfrac{1/3}{1+1/3}=1/4$

Def <u>odds ratio</u>

$$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$$

****

- θ can be computed using joint probabilities or either set of conditional probabilities (show ?)
- The odds ratio is appropriate when row totals are fixed, column totals are fixed, or neither set of marginal totals are fixed

****

Example

		Heart Attack YES	NO	Total
Group	Placebo	189	10,845	11,034
	Aspirin	104	10,933	11,037

Sample proportion

		Y YES	NO	Total
X	Placebo	p_1 (0.017)	$1-p_1$ (0.983)	1.0
	Aspirin	p_2 (0.009)	$1-p_2$ (0.991)	1.0

Sample odds = 0.017/0.9829=189/10,845=0.0174, placebo

= 104/10,933=0.0095, aspirin

Sample odds ratio

$$\hat{\theta}=\frac{0.0174}{0.0095}=1.83$$

The odds of a heart attack for placebo group was 1.83 time odds for aspirin group(i.e., 83% higher)

<u>Properties of odds ratio</u>

- each odds ≥ 0 and $\theta \geq 0$
- $\theta = 1$ when $\pi_1 = \pi_2$; i.e, response independent of group.
- The farther θ falls from 1, the stronger the association (For $Y=$ lung cancer, some studies have $\theta \approx 10$ for $X=$ smoking, $\theta \approx 2$ for X=passive smoking)
- If rows interchanged, or if columns interchanged, $\theta \rightarrow \frac{1}{\theta}.$

  eg. $\theta = 3$, $\theta = 1/3$ represent same strength of association but in opposite directions
- For counts

S	F
n_{11}	n_{12}
n_{21}	n_{22}

$$\hat{\theta}=\frac{n_{11}/n_{12}}{n_{21}/n_{22}}=\frac{n_{11}n_{22}}{n_{12}n_{21}}$$ = <u>cross-product ratio</u>

(Yule, 1900) (Strongly criticized by K Pearson)
- Treat X, Y symmetrically

		Group	
		Placebo	Aspirin
Heart Attack	Yes	189	104
	No	10,845	10,933

=> $\hat{\theta}=1.83$
- $\theta = 1 \Leftrightarrow \log\theta = 0$

  log odds ratio is symmetric about 0

  eg., $\theta = 2 \Rightarrow \log\theta = 0.7$

  $\theta = 1/2 \Rightarrow \log\theta = -0.7$

● Sampling dist. of $\hat{\theta}$ is skewed to right $\approx$ normal only of very large n

Note : we use “natural logs”(LN on most calculators). This is the log with $e = 2.718...$

● Sampling dist. of $\log\hat{\theta}$ is closer to normal, so construct C.I. for $\log\theta$ and then exponentiate endpoints to get C.I. for θ. Large-sample(asymptotic) standard error of $\log\hat{\theta}$ is

$$SE(\log\hat{\theta}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

C.I. for $\log\theta$ is

$$\log\hat{\theta} \pm Z_{\alpha/2} \times SE(\log\hat{\theta}) \overset{let}{=} (L,\ U)$$

C.I. for θ is $(e^L,\ e^U)$.

Example

$\hat{\theta} = \dfrac{189 \times 10{,}933}{104 \times 10{,}845} = 1.83$, $\log\hat{\theta} = 0.605$

$SE(\log\hat{\theta}) = \sqrt{\dfrac{1}{189} + \dfrac{1}{10{,}933} + \dfrac{1}{104} + \dfrac{1}{10{,}845}} = 0.123$

95% C.I. for $\log\theta$ is

$$0.605 \pm 1.96(0.123) = (0.365,\ 0.846)$$

95% C.I. for θ is

$$(e^{0.365},\ e^{0.846}) = (1.44,\ 2.33) > 1$$

Apparently $\theta > 1$

Note

● $\hat{\theta}$ is no midpint of C.I. because of skewness to right.

● If any $n_{ij} = 0$, $\hat{\theta} = 0$ or ∞ , any better estimate and SE results by replacing $\{n_{ij}\}$ *by* $\{n_{ij} + 0.5\}$

● When π_1 and π_2 are close to 0,

$$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} \approx \frac{\pi_1}{\pi_2} \text{ (the relative risk)}$$

SAS analysis of odds ratio

```
data aspirin;
input group $ mi$ count;
cards;
placebo yes 189
placebo no 10845
aspirin yes 104
aspirin no 10933
;
proc freq order=data;
weight count;
tables group*mi/measures;
run;
```

```
                FREQ procedure
             Table:group * mi

group        mi
Frequency   |
Percent     |
Col pct     |
Row pct     |yes     |no      |   Sum
-----------+--------+--------+
placebo     |    189 |  10845 |  11034
            |   0.86 |  49.14 |  49.99
            |   1.71 |  98.29 |
            |  64.51 |  49.80 |
-----------+--------+--------+
aspirin     |    104 |  10933 |  11037
            |   0.47 |  49.54 |  50.01
            |   0.94 |  99.06 |
            |  35.49 |  50.20 |
-----------+--------+--------+
Sum              293    21778    22071
                1.33    98.67   100.00
```

```
          Estimates of the Relative Risk (Row1/Row2)
 Type of Study              Value        95% Confidence Bounds
------------------------------------------------------------

 Case-Control               1.8321        1.4400        2.3308 <= (θ^L, θ^U)
 Cohort (Col1 Risk)         1.8178        1.4330        2.3059
 Cohort (Col2 Risk)         0.9922        0.9892        0.9953

                         Sample size = 22071
```

ratio of "No" proportions

ratio of "Yes" proportions

Example : Case-Control study in London Hospitals (Doll and Hill, 1950)

$X=$ smoked $\geq$ 1 cigarette per day for at least 1 year?

$Y=$ Lung cancer

		Lung Cancer YES	Lung Cancer NO
X	Yes	688	650
	No	21	59
	Total	709	709

Case-Control studies are "retrospective". Binomial sampling model applies to X (sampled within levels of Y), not to Y.

Cannot estimate $P(Y=Yes|x)$

or $\pi_1 - \pi_2 = P(Y=Yes|X=Yes) - P(Y=Yes|X=No)$

or π_1/π_2

However, we can estimate $P(X|Y)$, so can estimate θ.

$$\begin{aligned}\hat{\theta} &= \frac{\hat{P}(X=Yes|Y=Yes)/\hat{P}(X=No|Y=Yes)}{\hat{P}(X=Yes|Y=No)/\hat{P}(X=No|Y=No)} \\ &= \frac{(688/709)/(21/709)}{(650/709)/(59/709)} \\ &= \frac{688\times 59}{650\times 21} = 3.0\end{aligned}$$

odds of lung cancer for smokers were 3.0 times odds for non-smokers.

In fact, if $P(Y=Yes|X)$ is near 0, then $\theta \approx \pi_1/\pi_2 = relative\ risk$, and can conclude that prob. of lung cancer is ≈ 3.0 times as high for smokers as for non-smoker

Chi-squared Test of Independence

Example Job satisfaction and Income

	Job satisfaction: Very Dissat	Little Dissat	Moderate Satisfied	Very Satisfied	Total
< 5,000	2	4	13	3	22
5,000~15,000	2	6	22	4	34
15,000~25,000	0	1	15	8	24
> 25,000	0	3	13	8	24
Total	4	14	63	23	104

H_0 : X and Y are indep.
H_a : X and Y are dependent.

H_0 means $P(X=i, Y=j) = P(X=i)P(Y=j)$

$$\pi_{ij} = \pi_{i+}\pi_{+j}$$

Expected frequency $\mu_{ij} = n\pi_{ij}$
= mean of distribution of cell count n_{ij}
= $n\pi_{i+}\pi_{+j}$ under H_0

ML estimates $\widehat{\mu_{ij}} = n\hat{\pi}_{i+}\hat{\pi}_{+j} = n\left(\frac{n_{i+}}{n}\right)\left(\frac{n_{+j}}{n}\right) = \frac{n_{i+}n_{+j}}{n}$ called estimated expected frequencies.

Test statistic

$$X^2 = \sum_{all\ cells} \frac{(n_{ij} - \widehat{\mu_{ij}})^2}{\widehat{\mu_{ij}}}$$

called Pearson Chi-Squared statistic(Karl Pearson, 1900)
X^2 has large-sample chi-squared dist. with $df = (I-1)(J-1)$, where I=number of rows and J=number of columns.
$p-$ value=$P(X^2 \geq x^2 observed) =$ right-tail prob. (Appendix 3)

<u>Example</u>: Job satisfaction and Income

$$X^2 = 11.5,\ df = (4-1)(4-1) = 9$$

Evidence against H_0 is weak. plausible that job satisfaction and income are independent.

<u>Note</u>

- Chi-squared dist. has $\mu = df$, $\sigma = \sqrt{2df}$, more bell-shaped as df ↑.
- Likelihood-ratio test statistic

$$G^2 = 2\sum_{i,j} n_{ij} \log\left(\frac{n_{ij}}{\widehat{\mu_{ij}}}\right)$$
$$= -2\log\left[\frac{\text{maximize } likelihood\ when\ H_0\ is\ true}{\text{maximize } likelihood\ generally}\right]$$

G^2 also is approximated χ^2 with $df = (I-1)(J-1)$

<u>Example</u> : Revisit Job satisfaction

$$G^2 = 13.47,\ df = 9,\ p-value = .14$$

● df for χ^2 Test = No. parameters in general - No. parameters under H_0

eg) indep. $\pi_{ij} = \pi_{i+}\pi_{+j}$

$$\sum_{i,j}\pi_{ij} = 1$$

$$\begin{aligned} df &= (IJ-1) - [(I-1) + (J-1)] \\ &= (I-1)(J-1) \end{aligned}$$

$$\sum_j \pi_{+j} = 1$$

$$\sum_i \pi_{i+} = 1$$

(Fisher(1922), not Pearson, 1900)

● $X^2 = G^2 = 0$ when all $n_{ij} = \widehat{\mu_{ij}}$

● As $n\uparrow$, $X^2 \to \chi^2$ faster than $G^2 \to \chi^2$, usually close if most $\widehat{\mu_{ij}} \geq 5$.

● These tests treat X, Y as nominal. Reorder rows and columns, X^2 and G^2 are unchanged.

● For ordinal test, see sec 2.5 We re-analyze with ordinal model in Ch.6 (more powerful, much smaller $p-$ value)

<u>Standardized (Adjusted) Residuals</u>

$$r_{ij} = \frac{n_{ij} - \widehat{\mu_{ij}}}{\sqrt{\widehat{\mu_{ij}}(1-p_{i+})(1-p_{+j})}}$$

under H_0 : indep. , $r_{ij} \approx$ std. normal $N(0,\ 1)$

so, $|r_{ij}| > 2$ or 3 represents cell that provides strong evidence against H_0

<u>Example</u> Job satisfaction

$$n_{44} = 8,\ \widehat{\mu_{44}} = \frac{24 \times 23}{104} = 5.31$$

$$r_{44} = \frac{8 - 5.31}{\sqrt{5.31(1-24/104)(1-23/104)}} = 1.51$$

None of cells show much evidence of association

<u>Example</u> General Social Survey Data

		Religiosity			
		Very	Mod.	Slightly	Not
Gender	Female	170 (3.2)	340 (1.0)	174 (-1.1)	95 (-3.5)
	Male	98 (-3.2)	266 (-1.0)	161 (1.1)	123 (3.5)

$X^2 = 20.6,\ G^2 = 20.7,\ df = 3,\ p-value = 0.000$

- SAS (PROC GENMOD) also provides "Pearson Residuals"(label reschi)

$$e_{ij} = \frac{n_{ij} - \widehat{\mu_{ij}}}{\sqrt{\widehat{\mu_{ij}}}}$$

which are simpler nut less variable than $N(0,\ 1)$ ($\sum e_i^2 = X^2$)

Partitioning Chi-squared

$\chi_a^2 + \chi_b^2 = \chi_{a+b}^2$ for indep. chi-squared stat's

Example: Job satisfication and income (Revisited)

$G^2 = 13.47,\ X^2 = 11.52,\ df = 9$

Compare income levels an job satisfaction

	Job satisfaction				
	Very Dissat	Little Dissat	Moderate Satisfied	Very Satisfied	Total
< 5,000	2	4	13	3	22
5,000~15,000	2	6	22	4	34
15,000~25,000	0	1	15	8	24
> 25,000	0	3	13	8	24
Total	4	14	63	23	104

	Job satisfaction			
	VD	LD	MS	VS
< 5,000	2	4	13	3
5,000~15,000	2	6	22	4

	Job satisfaction			
	VD	LD	MS	VS
15,000~25,000	0	1	15	8
> 25,000	0	3	13	8

	Job satisfaction			
	VD	LD	MS	VS
< 15,000	4	10	35	7
> 15,000	0	4	28	16

X^2	G^2	df
0.30	0.30	3
1.14	1.19	3
10.32	11.98	3
11.76	13.47	9

See Next SAS program and Output~!!

```
/* Partitioning Chi-squared*/
  data jobsatis;
  input income satis count @@;
  cards;
  3 1 2 3 2 4 3 3 13 3 4 3
  10 1 2 10 2 6 10 3 22 10 4 4
  20 1 0 20 2 1 20 3 15 20 4 8
  30 1 0 30 2 3 30 3 13 30 4 8
  ;
  run;

  proc freq data=jobsatis;
  weight count;
  tables income*satis/chisq expected nopercent norow nocol;
  run;

  data collapse1;
  input income satis count @@;
  cards;
  3 1 2 3 2 4 3 3 13 3 4 3
  10 1 2 10 2 6 10 3 22 10 4 4
  ;
  run;
  data collapse2;
  input income satis count @@;
  cards;
  20 1 0 20 2 1 20 3 15 20 4 8
  30 1 0 30 2 3 30 3 13 30 4 8
  ;
  run;
  data collapse3;
  input income $ satis count @@;
  cards;
  <15 1 4 <15 2 1 <15 3 35 <15 4 7
  >15 1 0 >15 2 4 >15 3 28 >15 4 16
  ;
  run;

  proc freq data=collapse1;
  weight count;
  tables income*satis/chisq expected nopercent norow nocol;
  run;
  proc freq data=collapse2;
  weight count;
  tables income*satis/chisq expected nopercent norow nocol;
  run;
  proc freq data=collapse3;
  weight count;
  tables income*satis/chisq expected nopercent norow nocol;
  run;
```

Note

- Job satisfaction appears to depend on whether incoem > or <15,000.
- G^2 exactly partions, X^2 does not
- Text gives guidelines on how to partition so separate components indep., which is needed for G^2 to partition exactly

Small-sample test of indep.

2×2 case(Fisher, 1935)

	Y		
X	n_{11}	n_{12}	n_{1+}
	n_{21}	n_{22}	n_{2+}
	n_{+1}	n_{+2}	n

Exact null dist. of $\{n_{ij}\}$, based on fixed row and column tables, is

$$P(n_{11}|n_{1+},n_{2+},n_{+1},n_{+2})=\frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{+1}-n_{11}}}{\binom{n}{n_{+1}}};\ \text{Hypergeometric dist}$$

where $\binom{a}{b}=\dfrac{a!}{b!(a-b)!}$

Example : Tea tasting (Fisher)

		Guess		Total
		Milk	Tea	
Pour	Milk	?		4
First	Tea			4
Total		4	4	8

$n_{11}=0,1,2,3,4$

For $n_{11}=4$ has prob.

$$P(4)=\frac{\binom{4}{4}\binom{4}{0}}{\binom{8}{4}}=\frac{\left(\frac{4!}{4!0!}\right)\left(\frac{4!}{0!4!}\right)}{\left(\frac{8!}{4!4!}\right)}=\frac{4!4!}{8!}=1/70=0.014$$

For $n_{11}=3$

$$P(3)=\frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}}=16/70=0.229$$

n_{11}	$P(n_{11})$
0	0.014
1	0.229
2	0.514
3	0.229
4	0.014

IF observed table is given by

		Guess		Total
		Milk	Tea	
Pour	Milk	3	1	4
First	Tea	1	3	4
Total		4	4	8

For 2×2 tables,

$$H_0 : indep. \quad \Leftrightarrow \quad H_0 : \theta = 1 \text{ for } \theta = odds\ \ ratio$$

For $H_0 : \theta = 1 \quad Vs. \quad H_a : \theta > 1$

$$p-\text{ value= } P(\hat{\theta} \geq \widehat{\theta_{obs}}) = 0.229 + 0.014 = 0.243$$

Not much evidence against H_0.

Test using hypergeometic called <u>Fisher's exact test</u>

For $H_a : \theta \neq 1$, $p-$ value=two-tail prob. of outcomes no more likely than observed.

<u>Example</u>

$$p-\text{ value=}P(0) + P(1) + P(3) + P(4) = 0.486$$

<u>Note</u>

● Fisher's Exact test extends to $I\times J$ tables ($p-$ value=0.23 for job satisfaction and income)

● If make conclusion, eg, rejecting H_0 if $p \leq \alpha = 0.05$, actual $P(Type\ \ I\ error) < 0.05$ because of discreteness (see Text).

```
options ls=100 ps=200;
data fisher;
 input poured guess count;
 cards;
 1 1 3
 1 2 1
 2 1 1
 2 2 3
 ;
 run;
proc freq;
  weight count;
  table poured*guess / relrisk chisq;
  exact fisher or / alpha=.05;
run;
```

```
                 FREQ Procedure
             Table:poured * guess
       poured      guess
       Frequency  |
       Percent    |
       Row Pct    |
       Col Pct    |       1|       2|  Total
       -----------+--------+--------+
                1 |      3 |      1 |      4
                  |  37.50 |  12.50 |  50.00
                  |  75.00 |  25.00 |
                  |  75.00 |  25.00 |
       -----------+--------+--------+
                2 |      1 |      3 |      4
                  |  12.50 |  37.50 |  50.00
                  |  25.00 |  75.00 |
                  |  25.00 |  75.00 |
       -----------+--------+--------+
       Total             4        4        8
                     50.00    50.00   100.00
```

Statistics for Table of poured by guess

Statistic	df	Value	Prob	
Chi-square (X^2)	1	2.0000	0.1573	
Likelihood Ratio Chi-Square (G^2)	1	2.0930	0.1480	
Continuity Adj. Chi-Square (Yates)	1	0.5000	0.4795	(approximates Fisher's 2-side test)
Mantel-Haenszel Chi-Square	1	1.7500	0.1859	
Phi Coefficient		0.5000		
Contingency Coefficient		0.4472		
Cramer's V		0.5000		

WARNING: 100% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Fisher's Exact Test

Cell (1,1) Frequency (F)	3	
Left-sided Pr <= F	0.9857	
Right-sided Pr >= F	0.2429	$\boxed{H_a : \theta < 1}$
Table Probability (P)	0.2286	$\boxed{H_a : \theta > 1}$
Two-sided Pr <= P	0.4857 (2 x 0.2429)	$\boxed{H_a : \theta \neq 1}$

Estimates of the Relative Risk (Row1/Row2) (from option 'relrisk')

Type of Study	Value	95% Confidence Limits ← (delta method)		
Case-Control (Odds Ratio)	9.0000	0.3666	220.9270	$= \hat{\theta} e^{\pm 1.96 \hat{\sigma}(\log \hat{\theta})} = 9 e^{\pm 1.96 \sqrt{2 \cdot 1/3 + 2 \cdot 1/1}}$
Cohort (Col1 Risk)	3.0000	0.5013	17.9539	
Cohort (Col2 Risk)	0.3333	0.0557	1.9949	

Odds Ratio (Case-Control Study)

Odds Ratio	9.0000
Asymptotic Conf Limit	
95% Lower Conf Limit	0.3666
95% Upper Conf Limit	220.9270

```
Exact Conf. Limits
95%  Lower Conf Limit       0.2117
95%  Upper Conf Limit     626.2435

      Samplesize = 8
```

Tree-way Contingency Tables

Example : FL death penalty court cases

Victim's Race	defendant's Race	Death Penalty Yes	No	%Yes
White	White	53	414	11.3
	Black	11	37	22.9
Black	White	0	16	0.0
	Black	4	139	2.8

Let $Y=$ death penalty(Response var.)
$X=$ defendant's Race(Explanatory)
$Z=$ Victim's Race(Control var.)

The partial tables are

53	414
11	37

$Z=White$

0	16
4	139

$Z=Black$

They control(hold constant) Z
The conditional odds ratios are

For $Z=White$, $\hat{\theta}_{XY(1)}=\dfrac{53\times 37}{414\times 11}=0.43$

$Z=Black$, $\hat{\theta}_{XY(2)}=0.00$ (0.94 after add 0.5 to cells)

Controlling for Victim's race, odds of receiving death penalty were lower for white defendants than for black defendants

Add partial tables (XY marginal table)

		Death Penalty Yes	No
defendant's Race	White	53	430
	Black	15	176

$\hat{\theta}_{XY}=1.45$

Ignoring victim's race, odds of death penalty higher for white defendant's

Simpson's Paradox : All partial tables show reverse association from that in marginal table.

Cause ?

Moral ? can be dangerous to “collapse” contingency tables.

Def X and Y are conditionally independent given Z, if they are independent in each partial table

In $2\times 2\times K$ table,

$$\theta_{XY(1)} = \cdots = \theta_{XY(K)} = 1.0$$

Note The conditional independence does not imply that X and Y are marginally indep.

For Example,

Clinic(Z)	Treatment(X)	Response(Y) S	F	θ
1	A	18	12	1.0
	B	12	8	
2	A	2	8	1.0
	B	8	32	
Marginal	A	20	20	2.0
	B	20	40	