Combining measures from several strata [cf. M&M 3§2.6, A&B3 §16.2, Rothman2002, Chapter 8]

Why not just add (sum, Σ) the 'a' frequencies across tables (strata), the 'b' frequencies across tables, ... the 'd' frequencies across tables, to make a single 2 x 2 table with entries

Σ a	**Σ b**
Σ b	**Σ d**

and use these 4 cell counts to perform the analyses?

e.g. 1 Batting Averages of Gehrig and Ruth

(see book "Innumeracy" by Paulos)

```
                         Gehrig          Ruth
1st half of season       .290    <       .300
2nd half of season       .390    <       .400
----------------------------------------------
Entire season            .357    >       .333 !!!
```

Explanation:

```
                         Gehrig          Ruth
1st half of season
          Hits             29              60
          AT BAT          100             200

2nd half of season
          hits             78              40
          AT BAT          200             100
----------------------------------------------
Entire season
          hits            107             100
          AT BAT          300             300
```

Two features, involving time, created this 'paradox'

-1- batting averages increased from 1st to 2nd half of season

-2- Ruth had greater proportion of his AT BAT's in 1st half than Gehrig

e.g. 2 Numbers of Applicants (n), and Admission rates (%) to Berkeley Graduate School

	Men		**Women**	
Faculty	**n**	**% admitted**	**n**	**% admitted**
A	825	62	108	82
B	560	63	25	68
C	325	37	593	34
D	417	33	375	35
E	191	28	393	27
F	373	6	341	7
Combined	2691	**44**	1835	**30**

(see early Chapter in text "Statistics" by Freedman et al)

Paradox: π(admission | male) > π(admission | male) overall, but, by an large, faculty by faculty, its the other way!!!

Explanation: Women are more likely than men to apply to the faculties that admit lower proportions of applicants.

Remedy: aggregate the within-strata comparisons [like vs. like], rather than make comparisons with aggregated raw data -- see next for classical ways of doing this; MH stands for "Mantel-Haenszel".

For *other examples*:-

1. See Moore and McCabe(3rd Ed) 2.6 (The Perils of Aggregation, including Simpson's paradox) They speak of 'lurking' variables; in epidemiology we speak of **'confounding'** variables.

2. See Rothman2002, p1 (death rates Panama vs. Sweden) and p2 (20-year mortality in female smokers and non-smokers in Whickham England)

Simpson's paradox is an **extreme form of confounding**. Some textbooks give made-up examples See web site for course 626 for several *real* examples.

Story 4: Does Smoking Improve Survival? in the EESEE Expansion Modules in the website for the text (link from course description) [also in Rothman2002, with finer age-categories]

http://WWW.WHFREEMAN.COM/STATISTICS/IPS/EESEE4/EESEES4.HTM

A survey concerned with thyroid and heart disease was conducted in 1972-74 in a district near Newcastle, United Kingdom by Tunbridge et al (1977). A follow-up study of the same subjects was conducted twenty years later by Vanderpump et al (1996). Here we explore data from the survey on the smoking habits of 1314 women who were classified as being a current smoker or as never having smoked at the time of the original survey. Of interest is whether or not they survived until the second survey.

Results

The following tables summarize the results of the experiment: *[note from JH.. We would not call it "an experiment"; mathematical statisticians call any process for generating data "an experiment"]*

Table 1: Relationship between smoking habits and 20-year survival in 1314 women (582 Smokers, 732 Non-Smokers)

Survival Status	Smoking Status: Smoker	Smoking Status: Non-Smoker	Compared...
Dead	139	230	
Alive	443	502	
Risk = $\frac{\#dead}{\#Total}$	$\frac{139}{582} = 23.9\%$	$\frac{230}{732} = 31.4\%$	Diff: **–7.5%** Ratio: **0.76**
Odds = $\frac{\#dead}{\#alive}$	$\frac{139}{732} = 0.314(:1)$	$\frac{230}{502} = 0.458(:1)$	Ratio: **0.68***

* shortcut: $\textbf{or} = \frac{a \times d}{b \times c} = \frac{139 \times 502}{230 \times 443} = \frac{69778}{101890} = \textbf{0.68}$

A message the tobacco companies would love us to believe!

Table 2: Twenty-year survival status for 1314 women categorized by age and smoking habits at the time of the original survey.

Age Group (Years)	Survival Status	Smoking Status: Smoker	Smoking Status: Non-Smoker
18-44	Dead	19	13
	Alive	269	327
			(or =1.78)
44-64	Dead	78	52
	Alive	167	147
			(or =1.32)
>64	Dead	42	165
	Alive	7	28
			(or =1.02)

The odds ratio is > 1 in each age group!

Why the contradictory results?

Adjustment (compare like with like, i.e. Σ within-category estimates**)

Weighted averages [**explicit** weights (w's)]

Precision-based (inverse variance) Investigator-chosen ("Standardized")

Mean Difference $\Sigma\, w \times \bar{y}_{index} - \Sigma\, w \times \bar{y}_{ref}$

$= \Sigma\, w \times (\bar{y}_{index} - \bar{y}_{ref})$

Risk Difference $\Sigma\, w \times risk_{index} - \Sigma\, w \times risk_{ref}$

$= \Sigma\, w \times (risk_{index} - risk_{ref})$

Odds Ratio ("Woolf" method, precision based)

$\exp[\Sigma\, w \times logodds_{index} - \Sigma\, w \times logodds_{ref}]$

$= \exp[\Sigma\, w \times (\log [\text{odds ratio}])]$ ***all logs to base e***

where w = 1 / var[log [odds ratio]] = 1 / (1/a + 1/b + 1/c + 1/d)

Note: Computational formulae often constructed to minimize number of steps, and avoid division, and so may hide real structure of the estimator.

e.g. 8.1 in Rothman p147, for risk diff. (precision weighting)

Var[risk diff] proportional to $1/N_0 + 1/N_1 = (N_0+N_1)/(N_0 N_1)$

So that the denominator contribution, i.e., the weight, is

$w = 1/Var = (N_0 N_1)/(N_0+N_1) = (N_0 N_1)/T$

and numerator contribution is

$(risk_{index} - risk_{ref}) \times w$

$= (a/N_1 - b/N_0) \times w = (a/N_1 - b/N_0) \times (N_0 N_1)/T$

$= (a N_0 /T - b N_1) / T$ (after some algebra)

Ratio** Estimators ("M-H") [**implicit** precision weighting]

Risk Ratio

$$\frac{\Sigma\, \#cases_{index} \times DENOM_{ref} / DENOM_{total}}{\Sigma\, \#cases_{ref} \times DENOM_{index} / DENOM_{total}}$$

Rate Ratio same, except that denominators are amounts of ***person-time***, not persons.

Odds Ratio [case control study]

$$\frac{\Sigma\, \#cases_{index} \times \text{"denom"}_{ref} / \text{"size"}_{total}}{\Sigma\, \#cases_{ref} \times \text{"denom"}_{index} / \text{"size"}_{total}}$$

same as risk and rate ratio above except that "denominators" are partial (pseudo) ones estimated from a denominator series*("controls"); "size"_{total} refers to the size of (stratum-specific) case series and denominator series combined. *MODERN way to view case-control studies.

Odds Ratio [cohort/prevalence study]

$$\frac{\Sigma\, \#cases_{index} \times \#\text{"rest"}_{ref} / total}{\Sigma\, \#cases_{ref} \times \#\text{"rest"}_{index} / total}$$

Not that common to use this measure, since odds ratio more cumbersome to explain, and less 'natural'. Might use it to maintain comparability with results of a log-odds (logistic) regression. If #case a small fraction

**NOTE ON RATIO ESTIMATORS: Even though one could (if all denominators were obligingly non-zero) rewrite the ratio estimator as a weighted average of ratios, this would run counter to Mantel's express wishes.. to calculate just one ratio at the end, i.e. a ratio of two sums, rather than a sum of ratios. The main reason is statistical stability: imagine a (simpler, non-comparative) situation where one wished to estimate the overall sex ratio in small day-care facilities: would you average the ratios from each facility, or take a single ratio of the total number of males to the total number of females? The caveat does not apply to absolute differences, where the difference of two weighted averages (same set of weights for both) is the same as the weighed average of the differences.

Matched-pairs: the limiting case of finely stratified data

Examples: pair-matched case-control studies; Mother -> infant transmission of HIV in twins in relation to order of delivery; & others...
[see 607 notes for Ch 9]
ALSO: **Case-crossover studies** (***self-matched*** case-control studies)
eg" Redelmeier: auto accidents, while on/off cell phone when driving

> e.g. ***Response of <u>same</u> subject in each of 2 conditions (self-paired)***
>
> ***Responses of matched pair, one in 1 condition, 1 in other***
>
> ***Δ's in paired responses on interval scale, reduced to <u>sign</u> of Δ***

The 4 possibilities for 2 pair-members are:
(using ***generic*** 2 x 2 table: 2nd row might be a 'denominator series' of 1 per case)

Outcome	Category of Determinant: Index	Reference	Total
Yes	a = 1 or 0	b =1 or 0	1
No	c = 0 or 1	d = 0 or 1	1
	1	1	T=2

The contributions to or_{MH} from the 4 possibilities are ...

Outcome*	Determinant: Index	Ref	Tot	$\frac{a \times d}{T}$	$\frac{b \times c}{T}$	No. Pairs
Yes	1	1	1			
No	0	0	1			
	1	1	2	0	0	"A"
Yes	0	0	1			
No	1	1	1			
	1	1	2	0	0	"D"
Yes	1	0	1			
No	0	1	1			
	1	1	2	1/2	0	"B"
Yes	0	1	1			
No	1	0	1			
	1	1	2	0	1/2	"C"

$$\textbf{Odds Ratio estimator} = \frac{A \times 0 + B \times 1/2 + C \times 0 + D \times 0}{A \times 0 + B \times 0 + C \times 1/2 + D \times 0} = \frac{B}{C}$$

Tabular format for displaying matched pair-data

COHORT STUDY

Result in One PAIR Member	Result in Other PAIR Member: + ve	– ve	Total # PAIRS
+ ve	A	B	
– ve	C	D	
			n

CASE-CTL STUDY

Exposure in "Case"	Exposure in "Control": + ve	– ve	Total PAIRS
+ ve	A	B	
– ve	C	D	
			n

* In matched (self- or other) case-control study, the "denominator series" is not limited to 1 "probe-for-exposure" per case... could ask about "usual" exposure (e.g. % time usually exposed) or sample several "person -moments" ['controls'] per case. i.e. the 2nd row total could be > 2.

Standardization of Rates [proportion-type and incidence-type] [**explicit, investigator-selected** weights]

- Usual to first calculate standardized rate for index category (of the determinant) and standardized rate for reference category (of the determinant) separately, then compare the standardized rates.
- If one uses the confounder distribution in one of the two compared determinant categories as the common set of weights, then the standardized rate in this category remains unchanged from the crude rate in this category.

  See the worked example comparing death rates in Quebec males in 1971 and 1991 in the document ***"Direct" and "Indirect" Standardization:2 sides of same coin?(.pdf)*** under "Material from previous years" in the c626 web page. this is an interesting local case of natural confounding: relative to that 20 years earlier, the crude mortality rate in 1991 was 1.00. yet, in every age category, the rate in 1991 was at least 10% lower, and in many age-groups, more than 20% lower than in 1971 (in the table, the rate ratios in bold are 71/91, so take their reciprocals to see the rate ratios 91/71)
- Read Rothman's comment (p159) about the ***uniformity of effect*** (eg a constant rate ratio across age groups in the Que example). Why in his last sentence in that paragraph does he seem to "allow" a weighted average of very different rate ratios, if they were derived from standardization, but NOT if they were derived from (precision-weighted) pooling?
- Rothman (p161) emphasizes how "silly" the term "indirect" standardization used with standardized mortality ratio, is. He correctly points out that "the calculations for any rate standardization, "direct" or "indirect", are basically the same". He leaves it as an **exercise** (Q4 page 166) to work out **what the weights are in the so-called "indirect" standardization used to compute an SMR (or SIR)**.

  Hint: write the SMR (with Σ denoting sum over strata) as

$$\text{SMR} = \frac{\text{Total \# cases observed}}{\text{Total \# cases expected}}$$

$$= \frac{\Sigma \text{ \# observed}}{\Sigma \text{ \# expected}} \qquad (*)$$

$$= \frac{\Sigma \text{ observed \#}}{\Sigma \text{ ref. rate} \times \text{exposed PT}}$$

$$= \frac{\Sigma \text{ observed rate} \times \text{exposed PT}}{\Sigma \text{ ref. rate} \times \text{exposed PT}}$$

$$= \frac{\Sigma \text{ observed rate} \times w}{\Sigma \text{ ref. rate} \times w}, \qquad \textbf{with } w \textbf{ = exposed PT}$$

If one starts again from (*), one can show that the **SMR can also be represented as *a weighted average of rate ratios*** [as was mentioned in footnote to Quebec table*]

$$\text{SMR} = \frac{\Sigma \text{ \# observed}}{\Sigma \text{ \# expected}}$$

$$= \frac{\Sigma \text{ obs. rate} \times \text{exposed PT}}{\Sigma \text{ \# expected}}$$

$$= \frac{\Sigma \frac{\text{obs. rate}}{\text{ref. rate}} \times \text{ref. rate} \times \text{exposed PT}}{\Sigma \text{ \# expected}} \qquad \text{(divide \& mult. by ref rate)}$$

$$= \frac{\Sigma \frac{\text{obs. rate}}{\text{ref. rate}} \times \text{\# expected}}{\Sigma \text{ \# expected}} \qquad = \textbf{weighted ave. of rate ratios}$$

*cf. Liddell FD. The measurement of occupational mortality. ***Br J Ind Med***. 1960 Jul;17:228-33.

Table 2: Twenty-year survival status for 1314 women categorized by age and smoking habits at the time of the original survey.
Worked out calculations (see same calculations on spreadsheet) for... (* r1 r2 are row totals; c1 c2 are column totals) *See Rothman Ch 8*

Mantel-Haenszel summary odds ratio, or_{MH} and
Mantel Haenszel (Chi-Square) Test of $OR_1 = OR_2 = OR_3 = 1$

Woolf: exp[weighted average of *ln* or 's]
Var[weighted ave] = 1/ {Sum of Weights}

Age Group (Years)	Surv Status	Smoking Status: Smoker	Smoking Status: Non-Smoker	**Calculations for Summary odds ratio:** n	a d / n	b c / n	Calculations for Test Statistic*: $E[a \mid H_0]$ = $\frac{r1 \cdot c1}{n}$	$Var[a \mid H_0]$ = $\frac{r1 \cdot r2 \cdot c1 \cdot c2}{n^2 \cdot \{n-1\}}$	**Calculations for Woolf's Method:** *ln* or (1)	Var[*ln* or] = 1/a + 1/b + 1/c + 1/d	Weight (2) = $\frac{1}{Var[\ ln\ or]}$	W × *ln* or = (1) × (2)
18-44	Dead	**19**	13									
	Alive	269	327									
			(or =1.78)	**628**	**9.89**	**5.59**	**14.7**	*7.6*	**0.575**	**0.1363**	**7.335**	**4.218**
44-64	Dead	**78**	52									
	Alive	167	147									
			(or =1.32)	**444**	**25.82**	**19.56**	**71.7**	*22.8*	**0.278**	**0.0448**	**22.30**	**6.199**
>64	Dead	**42**	165									
	Alive	7	28									
			(or =1.02)	**242**	**4.86**	**4.77**	**41.9**	*4.9*	**0.018**	**0.2084**	**4.798**	**0.086**
		139	Sum	*1314*	**40.57**	**29.92**	**128.3**	*35.2*			**34.433**	**10.503**

MH Odds Ratio **1.36**
(40.57/29.92)

weighted ave. of *ln* or ' s 10.503/ 34.433 = **0.305**
exp[weighted ave. of *ln* or ' s] exp[**0.305**] = **1.36**

$$\mathbf{or_{MH}} = \frac{\sum a\,d / n}{\sum b\,c / n} = \frac{40.57}{29.92} = \mathbf{1.36}\ ; \qquad X^2_{MH\ (1\ df)} = \frac{\{\sum a - \sum E[\,a \mid H_0\,]\}^2}{\sum Var[\,a \mid H_0\,]} = \frac{\{139 - 128.3\}^2}{35.2} = 3.24 \qquad X_{MH} = 1.80$$

(Miettinen) **Test-based** 100(1 - α)% **CI** for OR: $\mathbf{or_{MH}}^{1 \pm z_{\alpha/2} / X_{MH}} = 1.36^{1 \pm 1.96/1.80}$ = **0.97 to 1.89** (95% CI)

(Woolf) 100(1 - α)% **CI** for OR: exp[{weighted ave. of *ln* or's} ± $z_{\alpha/2}$ Sqrt[1/34.433] = exp[0.305 ± 1.96×0.170] = **0.97 to 1.89**

Via SAS

```
data sasuser.simpson;
input age $ i_smoke i_dead
number;

lines;
18-44 1  1   19
18-44 1  0  269
18-44 0  1   13
18-44 0  0  327
44-64 1  1   78
44-64 1  0  167
44-64 0  1   52
44-64 0  0  147
64-   1  1   42
64-   1  0    7
64-   0  1  165
64-   0  0   28
;

run;
options ls = 75 ps = 50; run;
proc freq data=sasuser.simpson;

tables age * i_smoke * i_dead /
   nocol norow nopercent cmh expected;

weight number;
/* weight indicates multiples */

run;
```

See for SAS 'trick' to produce Tables in an orientation that gives the ratios of interest (use PROC FORMAT to associate another values with each actual value; then use the ORDER=FORMATTED option in PROC FREQ)

```
        TABLE 1 OF I_SMOKE BY I_DEAD
          CONTROLLING FOR AGE=18-44

   I_SMOKE     I_DEAD

Frequency|
Expected |       0|       1|  Total
   ------+--------+--------+
       0 |    327 |     13 |   340
         | 322.68 | 17.325 |
   ------+--------+--------+
       1 |    269 |     19 |   288
         | 273.32 | 14.675 |
   ------+--------+--------+
   Total      596       32    628
```

```
        TABLE 2 OF I_SMOKE BY I_DEAD
          CONTROLLING FOR AGE=44-64

   I_SMOKE     I_DEAD

Frequency|
Expected |       0|       1|  Total
   ------+--------+--------+
       0 |    147 |     52 |   199
         | 140.73 | 58.266 |
   ------+--------+--------+
       1 |    167 |     78 |   245
         | 173.27 | 71.734 |
   ------+--------+--------+
   Total      314      130    444
```

```
        TABLE 3 OF I_SMOKE BY I_DEAD
           CONTROLLING FOR AGE=64-

   I_SMOKE     I_DEAD

Frequency|
Expected |       0|       1|   Total
   ------+--------+--------+
       0 |     28 |    165 |   193
         | 27.913 | 165.09 |
   ------+--------+--------+
       1 |      7 |     42 |    49
         | 7.0868 | 41.913 |
   ------+--------+--------+
   Total       35      207    242
```

```
SUMMARY STATISTICS FOR

        I_SMOKE BY I_DEAD
       CONTROLLING FOR AGE

Cochran-Mantel-Haenszel Statistics
(Based on Table Scores)

Alt. Hypothesis        DF Value  Prob
-------------------------------------
Nonzero Correlation     1 3.239 0.072
Row Mean Scores Differ  1 3.239 0.072
General Association     1 3.239 0.072


Estimates of Common Relative Risk
(Row1/Row2)

                                         95%
Type of Study  Method     Estimate Conf Bounds
-------------------------------------
Case-Control Mantel-Haenszel 1.357 0.973 1.892
(Odds Ratio)   Logit         1.357 0.971 1.894

Cohort       Mantel-Haenszel 1.047 0.996 1.101
(Col1 Risk)    Logit         1.034 0.998 1.072

Cohort       Mantel-Haenszel 0.864 0.738 1.013
(Col2 Risk)    Logit         0.953 0.849 1.071

Confidence bounds for M-H estimates
are test-based.

Breslow-Day Test for Homogeneity of
the Odds Ratios

Chi-Square = 0.950 DF = 2 Prob = 0.622


Total Sample Size = 1314
```

Via Stata

```
clear
input str5 age  i_smoke i_dead  number
          18_44   1        1      19
          18_44   1        0     269
          18_44   0        1      13
          18_44   0        0     327
          44_64   1        1      78
          44_64   1        0     167
          44_64   0        1      52
          44_64   0        0     147
          64_     1        1      42
          64_     1        0       7
          64_     0        1     165
          64_     0        0      28
end
cc i_dead i_smoke [freq=number], by(age)
```

```
         age |      OR     [95% CI]    M-H Weight
      -------+------------------------------------------
       18_44 |    1.78   .87   3.61       5.57    (Cornfield)
       44_64 |    1.32   .87   1.99      19.56    (Cornfield)
         64_ |    1.02   .42   2.43       4.77    (Cornfield)
      -------+------------------------------------------
       Crude |     .68   .53    .88               (Cornfield)
M-H combined |    1.36   .97   1.90
-------------+------------------------------------------
Test of homogeneity (M-H)  chi2(2) = 0.95  Pr>chi2 = 0.6234

             Test that combined OR = 1:
                        Mantel-Haenszel chi2(1) =       3.24
                                        Pr>chi2 =     0.0719
```

Also available...

```
cc  i_dead i_smoke [freq=number], by(age) woolf

cc  i_dead i_smoke [freq=number], by(age) tb

*tb = "test-based"
```

Aggregating Odds Ratio (OR)'s ...Woolf's Method

Recall: data from single 2x2 table: $or = \frac{ad}{bc}$

$$SE[\ ln\ (or)\ ] = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

data from several (K) 2x2 tables: (Σ: summation over strata)

$$ln\ (or_{Woolf}) = \frac{\sum w_k\ ln\ (or_k)}{\sum w_k} \quad \text{(weighted average)}$$

$$\text{with } w_k = \frac{1}{Var[\ ln\ [or_k]\ ]} \quad \text{(weight 1 / variance)}$$

(note: $Var = SE^2$)

$$SE[ln\ (or_{Woolf})] = \sqrt{\frac{1}{\sum w_k}} = \sqrt{\frac{Var^*}{K}} \quad \text{[see drivation \#]}$$

(Var^* : harmonic mean of K Var's)

$$CI[\ OR\ ] = exp\{\ CI[\ ln\ (OR)\ ]\ \}$$

 # Derivation: $Var[\Sigma\{w \times ln\} / \Sigma w] = (1/\Sigma w])^2 \times \Sigma\{w^2 \times Var[ln]\}$

$= (1/\Sigma w])^2 \times \Sigma\{1/w\} = 1/\Sigma w$ [since $w = 1/var[ln]$]

See worked example in Spreadsheet (under Resources Ch 9)
[Robins-Breslow-Greenland SE for ln or_{MH} not programmed]

References: A&B Ch 4.8 and 16, Schlesselman, KKM, Rothman...

Summary Risk Ratio and Summary Rate Ratio

See Rothman pp 147- (Risk Ratio) and pp153- (Rate Ratio)

Berkeley Data: M:F Comparative parameters Odds Ratio (OR), Risk Ratio (RR) and Risk Difference (RΔ)

(Using KKM table 17.16 notation)

	E	$\bar{E}$	
D	a	b	m_1
$\bar{D}$	c	d	m_0
	n_1	n_0	n

Faculty	Admitted?	Men	Women	All	for RΔ: a/n_1	b/n_0	RΔ	for OR: $\frac{a \cdot d}{b \cdot c}$	$\frac{a \cdot d}{n}$	$\frac{b \cdot c}{n}$	for RR: $\frac{a \cdot n_0}{b \cdot n_1}$	$\frac{a \cdot n_0}{n}$	$\frac{b \cdot n_1}{n}$	for RΔ: var(RΔ)*	w = 1/var	w•RΔ
A	Y	512	89	601	0.62	0.82	−0.20	0.35	10.4	29.9	0.75	59.3	78.7	1.63E-3	614	−125
	N	313	19	332												
	All	825	108	933												
B	Y	353	17	370	0.63	0.68	−0.05	0.80	4.8	6.0	0.93	15.1	16.3	9.12E-3	110	−5
	N	207	8	215												
	All	560	25	585												
C	Y	120	202	322	0.37	0.34	+0.03	1.13	51.1	45.1	1.08	77.5	71.5	1.10E-3	913	26
	N	205	391	596												
	All	325	593	918												
D	Y	138	131	269	0.33	0.35	−0.02	0.92	42.5	46.1	0.95	65.3	69.0	1.14E-3	879	−16
	N	279	244	523												
	All	417	375	792												
E	Y	53	94	147	0.28	0.24	+0.04	1.22	27.1	22.2	1.16	35.7	30.7	1.51E-3	661	25
	N	138	299	437												
	All	191	393	584												
F	Y	22	24	101	0.06	0.07	−0.01	0.83	9.8	11.8	0.84	10.5	12.5	3.41E-4	2935	-33
	N	351	317	668												
	All	373	341	769												
All	*Y*	*1198*	*557*	*1755*	*0.44*	*0.30*	*+0.14*	*1.84*			*1.47*					
	N	*1493*	*1278*	*2771*												
	All	*373*	*341*	*4526*												
								∑:	145.8	161.1		263.4	278.7		6113	−129

$$OR_{MH} = \frac{145.8}{161.1} = \mathbf{0.91}$$

$$RR_{MH} = \frac{263.4}{278.7} = \mathbf{0.94}$$

$$R\Delta_w = \frac{\sum w \cdot R\Delta}{\sum w} = \frac{-129}{6113} = \mathbf{-0.02}$$

* var(RΔ) = Sum of 2 binomial variances

Test of equal M:F admission rates; **Confidence Intervals** for OR_{MH} (Berkeley data, KKM and A&B notation; cf. Rothman'02,Table 8.4, p152)

TEST $\pi_M = \pi_F$

Faculty	Admitted?	Men	Women	All	$E[a\|H_o]$	$Var[aH_o]$
A	Y	512	89	601	531.4	21.9
	N	313	19	332		
	All	825	108	933		
B	Y	353	17	370	354.2	5.6
	N	207	8	215		
	All	560	25	585		
C	Y	120	202	322	114.0	47.9
	N	205	391	596		
	All	325	593	918		
D	Y	138	131	269	141.6	44.3
	N	279	244	523		
	All	417	375	792		
E	Y	53	94	147	48.1	24.3
	N	138	299	437		
	All	191	393	584		
F	Y	22	24	101	24.0	10.8
	N	351	317	668		
	All	373	341	769		
All	*Y*	*1198*	*557*	*1755*		
	N	*1493*	*1278*	*2771*		
	All	*373*	*341*	*4526*		
				∑:	1213.4	154.7

$$\frac{\{\sum a - \sum a|H_o\}^2}{\sum Var[a|H_o]} = \frac{\{1198 - 1213.4\}^2}{154.7} = \mathbf{1.52} \quad [\#]$$

This MH X^2 of 1.52 is "NS" in the χ^2 1df distribution

[#] see Rothman2002, p162

CI for OR_{MH} [notation from A&B p461]

(Method of Robins, Breslow & Greenland 1986 *)

Faculty	$\frac{a+d}{n}$ (P)	$\frac{b+c}{n}$ (Q)	$\frac{a \bullet d}{n}$ (R)	$\frac{b \bullet c}{n}$ (S)	P•R	P•S	Q•R	Q•S
A	0.57	0.43	10.4	29.9	5.9	17.0	4.5	12.9
B	0.62	0.38	4.8	6.0	3.0	3.7	1.8	2.3
C	0.56	0.44	51.1	45.1	28.5	25.1	22.7	20.0
D	0.48	0.52	42.5	46.1	20.5	22.3	22.0	23.9
E	0.60	0.40	27.1	22.2	16.4	13.4	10.8	8.8
F	0.47	0.53	9.8	11.8	4.6	5.6	5.1	6.2
			145.8 (R_+)	161.1 (S_+)	78.9	87.1	66.9	74.1

Rothnan2002, p152 uses different notation

A&B {R,S, P,Q} -> Rothman{G,H, P,Q}

$$Var[\ln OR_{MH}] = \frac{\sum P \bullet R}{2R_+^2} + \frac{\sum[P \bullet S + Q \bullet R]}{2R_+ \bullet S_+} + \frac{\sum Q \bullet S}{2S_+^2}$$

$$= \frac{78.9}{2 \bullet 145.8^2} + \frac{87.1 + 66.9]}{2 \bullet 14.5.8 \bullet 161.1} + \frac{74.1}{2 \bullet 161.1^2} = 0.0066$$

continued at top of next column ...

CI OR_{MH} continued...

$\ln OR_{MH} = \ln 0.91 = -0.10$

$Var[\ln OR_{MH}] = 0.0066$

$SE[\ln OR_{MH}] = \sqrt{Var} = \mathbf{0.08}$

$CI[\ln OR_{MH}] = -0.10 \pm z \bullet 0.08$

$= -0.26$ to 0.06 (95%)

$CI[OR_{MH}] =$

$\exp[-0.26]$ to $\exp[0.06]$

= **0.77** to **1.06**

CI [OR_{MH}] "test-based" (Miettinen 1976)

Chi-MH = $|\ln or_{MH}|$ / $SE[\ln or_{MH}]$ ===>

$SE[\ln or_{MH}] = |\ln or_{MH}|$ / Chi-MH {0.10/√1.52= 0.08}

$CI[\ln OR_{MH}] = \ln or \pm z\ SE[\ln or_{MH}]$

$CI[OR_{MH}]$ = CI [exp[ln or_{MH}]]

= exp[CI for ln] = $or_{MH}^{[1 \pm z/\text{Chi-MH}]}$

= $or_{MH}^{[1 \pm 1.96/\sqrt{1.52}]}$ in our example

CI [RΔ] ... *(continued from last column, previous page)*

SE[RΔ] = $\sqrt{1/\sum w}$ = 0.013

CI [RΔ] $= -0.02 \pm z \times 0.013$