- Correspondence analysis ← LNp.5-13~17, CA identifies association (i.e., interaction in Poisson GLM) between variables
  - ➢ Cannot directly apply to 3-way table ← ∵ CA is based on singular value decomposition, which applied on matrix ↑ 2-way
  - ➢ Can combine two of the factors, say X_1 and X_2, into a factor with $I\times J$ levels and apply correspondence analysis on the 2-way table formed by the new factor and X_3

    new factor: $x_1=1, x_2=1$; $x_1=1, x_2=2$; X_3: 1, 2

    interested in $\pi_{B|ij}$'s in binomial GLM ← cf.
  - ➢ **Q**: which two factors should be chosen to merge?
  - **Ans**: pick up the two whose association is least interesting to us ← ∵ after merging X_1 & X_2, becomes difficult to study $X_1{:}X_2$ in the main effects of the new factor
- Simpson's paradox [5/12]
  - ➢ example:

X_1 (i): age; X_2 (j): smoker; X_3 (k): dead or alive

3-way table. there exists a significant interaction (association) btwn X_1 & X_2 ✱

X_1 (i)	X_2 (j) smoker	X_3 (k) dead (1)	alive (2)	y_{ij+}	$\hat\pi_{j\vert i+}$	⇒	smoker	dead	alive $\hat\pi_{B\vert ij}$	
1 age=35-44 younger	yes 1	14	95	∨ 109	(.47)	⇒	yes	.13	.87	1
	no 2	7	114	o 121	(.53)		no	.06	.94 ≥	1
	smoker	dead	alive		≥	different base rates	smoker	dead	≫ alive	
2 age=65-74 older	yes 1	29	7	∨ 36	(.22)	⇒	yes	.81	.19	1
	no 2	101	28	o 129	(.78)		no	.78	.22 ≥	1

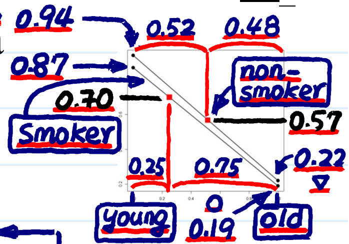
higher proportion of young people in smoker ($\hat\pi_{i|j+}$): 109/145 =.75; 121/250 =.48 ← $\hat\pi_{i|j+}$

2-way table

marginal total over age: smoker	dead (y_{+jk})	alive	y_{+j+}	$\hat\pi_{+j+}$	⇒	smoker	dead	alive $\hat\pi_{B\vert +j}$	
yes	43	102	∨ 145	(.37)	⇒	yes	.30	.70	1
no	108	142	o 250	(.63)		no	.43	.57 ≤	1

conditional indep. ⇏ marginal indep (LNp5-24) ← cf.

- *marginal association* added over X_1 is different from the *conditional association* observed within each category of X_1
- ➢ **Q**: Why it occurs? Why the table of y_{+jk} gives a contradictory result to the tables of $y_{jk|i}$?



- If (young & smoker 0.87) $\dfrac{y_{112}}{y_{11+}}=\dfrac{y_{112}}{y_{111}+y_{112}}<\dfrac{y_{122}}{y_{121}+y_{122}}=\dfrac{y_{122}}{y_{12+}}$ (young & non-smoker 0.94)

  and (old & smoker 0.19) $\dfrac{y_{212}}{y_{21+}}=\dfrac{y_{212}}{y_{211}+y_{212}}<\dfrac{y_{222}}{y_{221}+y_{222}}=\dfrac{y_{222}}{y_{22+}}$ (old & non-smoker 0.22)

  sum over age ⇒ $\boxed{j=1}$ $\dfrac{y_{112}+y_{212}}{y_{111}+y_{112}+y_{211}+y_{212}} \overset{?}{<} \dfrac{y_{122}+y_{222}}{y_{121}+y_{122}+y_{221}+y_{222}}$ $\boxed{j=2}$

  - Note. $\dfrac{y_{1j2}+y_{2j2}}{y_{1j+}+y_{2j+}}=\dfrac{y_{1j2}}{y_{1j+}}\underbrace{\dfrac{y_{1j+}}{y_{1j+}+y_{2j+}}}_{w_j\ (0\le w_j\le 1)}+\dfrac{y_{2j2}}{y_{2j+}}\underbrace{\left(1-\dfrac{y_{1j+}}{y_{1j+}+y_{2j+}}\right)}_{1-w_j}$

    $j=1$: ∨, $j=2$: △; $j=1$: o, $j=2$: ▽ — a weighted average

    $w_j=0.75$ for j=smoker; $w_j=0.48$ for j=non-smoker

- Note that smoker are more concentrated in the younger age group and younger people are more likely to live longer

Recall. SS4, hypergeometric approach for 2×2 table (LNp.5-11)

	x_1	x_2	x_3
1st	1	1	1
2nd	2	2	1
3rd	2	1	2
4th	1	2	1
⋮	⋮	⋮	⋮

→ fitted model 1 use X_1, X_2, X_3 ⇒ analysis of 3-way table.
→ fitted model 2 use X_2, X_3 ⇒ analysis of 2-way table

Recall. In LM (check ✱, LNp.5-30)
true model: $Y=Z_1\beta_1+Z_2\beta_2+\varepsilon$
fitted model: $Y=Z_1\beta_1+\varepsilon'$
⇒ $\hat\beta_1$ biased if $Z_1^TZ_2\neq 0$

- Mantel-Haenszel (MH) test for $2\times2\times K$ table
  conditioned on X_3 (check □ in LNp 5-32) $X_1 \to$ ↑ X_2 ↑ X_3
  - ➢ Designed to test *independence* in 2×2 tables across K categories

level permutation in X_1 (or X_2): $\Delta \to \frac{1}{\Delta}$

➤ Recall. association of a 2×2 table can be completely characterized/measured by its odds-ratio Δ (Δ: positive, $1/\Delta$: negative) | negative, positive

- $\Delta = 1 \Leftrightarrow$ independence
- $\Delta > 1 \Leftrightarrow$ positive association
- $\Delta < 1 \Leftrightarrow$ negative association

Note. $\Delta = e^{\beta_{12}}$ ← interaction effect of X_1 & X_2 (* in LNp 5-26) positive: $\beta_{12} > 0$; negative: $\beta_{12} < 0$

➤ Null and alternative hypotheses of MH test

Δ_k: odds ratio of X_1 & X_2 at $X_3 = k$

- (a) H_0: $\Delta_1 = \Delta_2 = \ldots = \Delta_K = 1$ (conditional independence) ← conditioned on X_3
- H_1^*: at least one $\Delta_k \neq 1$ (different association) or ← Alternative. Breslow-Day test
- H_1: $\Delta_1 = \Delta_2 = \ldots = \Delta_K \neq 1$ (uniform association)

MH test works better under this alternative → or define a common odds ratio

- The test works better when the odds ratios of the K 2×2 tables (2×2×K table) do not vary greatly, e.g., the null of the GoF test for uniform association, $\Delta_1 = \Delta_2 = \ldots = \Delta_K$, does not rejected

rationale: e.g. X_3: a block factor, X_1, X_2: variables of interest, with a common association across blocks

➤ Procedure of the MH test

row & column | Note. information from marginal totals is not required in testing indep. (consider SS4, LNp 5-11)

- Suppose the marginal totals of each 2×2 table carry no information (e.g., fixed in advance) or are conditioned.
  - ▫ under H_0, can assume a hyper-geometric distribution for y_{11k} in each 2×2 table → $y_{11k} \overset{\text{indep}}{\sim}$ (LNp 5-11) (*) $HG(y_{1+k}, y_{+1k}, y_{+2k})$

    $E(y_{11k}) = y_{1+k} \times \frac{y_{+1k}}{y_{+1k}+y_{+2k}}$

What if we do not want to perform K tests & want to accumulate information in one test

$\Rightarrow$ y_{11k} is sufficient for testing independence of kth table

- MH statistic combine information of y_{11k}'s from K tables:

y_{11+} ← Does it actually use only marginal 2×2 table? Ans. No. Check

$$\frac{\left(\left|\sum_k [y_{11k} - E(y_{11k})]\right| - 1/2\right)^2}{\sum_k Var(y_{11k})} \overset{a}{\sim} \chi^2_1$$

← continuity correction (increasing p-value, more conservative, check LNp 5-7)

← $\dim(H_1) - \dim(H_0)$

∵ y_{11k}'s are independent.

If some $\Delta_k \gg 1$ & some $\Delta_k \ll 1$ then some $y_{11k} \gg E(y_{11k})$ & some $y_{11k} \ll E(y_{11k})$ they might cancel out & MH stat. becomes small (not rejected)

where $E(y_{11k})$ and $Var(y_{11k})$ are calculated under the H_0 → Check (*) in LNp. 5-32

- ▫ can calculate an exact p-value for smaller dataset using hypergeometric distribution ← # of possible values of y_{11k}'s are not large (check LNp 5-12)

  $\Rightarrow$ useful when data is sparse, under which the χ^2 approximations based on asymptotic thm is questionable

  many small y_{ijk}'s

➤ MH test is sometimes called Cochran-Mantel-Haenszel test because a version without the 1/2 is published earlier by Cochran (1954).

❖ **Reading**: Faraway (2006, 1st ed.), 4.4

5/14

Ordinal Variables

(LNp. 2-3 ~4) → ① continuous (interval) ② discrete interval ③ ordinal ④ nominal

ⓐ In analysis, cannot arbitrarily rearrange the order of categories

- Some variables have a nature ordering → ⓑ the order matters for the objective of analysis

  between categories distance = ?

  Otherwise, we can find order(s) in many (almost any) variables, but some not matter

  ➤ e.g., education: HS, BA, MA; political ideology: VL, SL, M, SC, VC

cf. • ordinal response 隨 ← LNp 5-28 • ordinal covariate 規