Lecture 5: Measures of association and hypothesis testing for categorical data

BTBI30081

統計應用方法Applied Methods in Statistics

2025/3/19

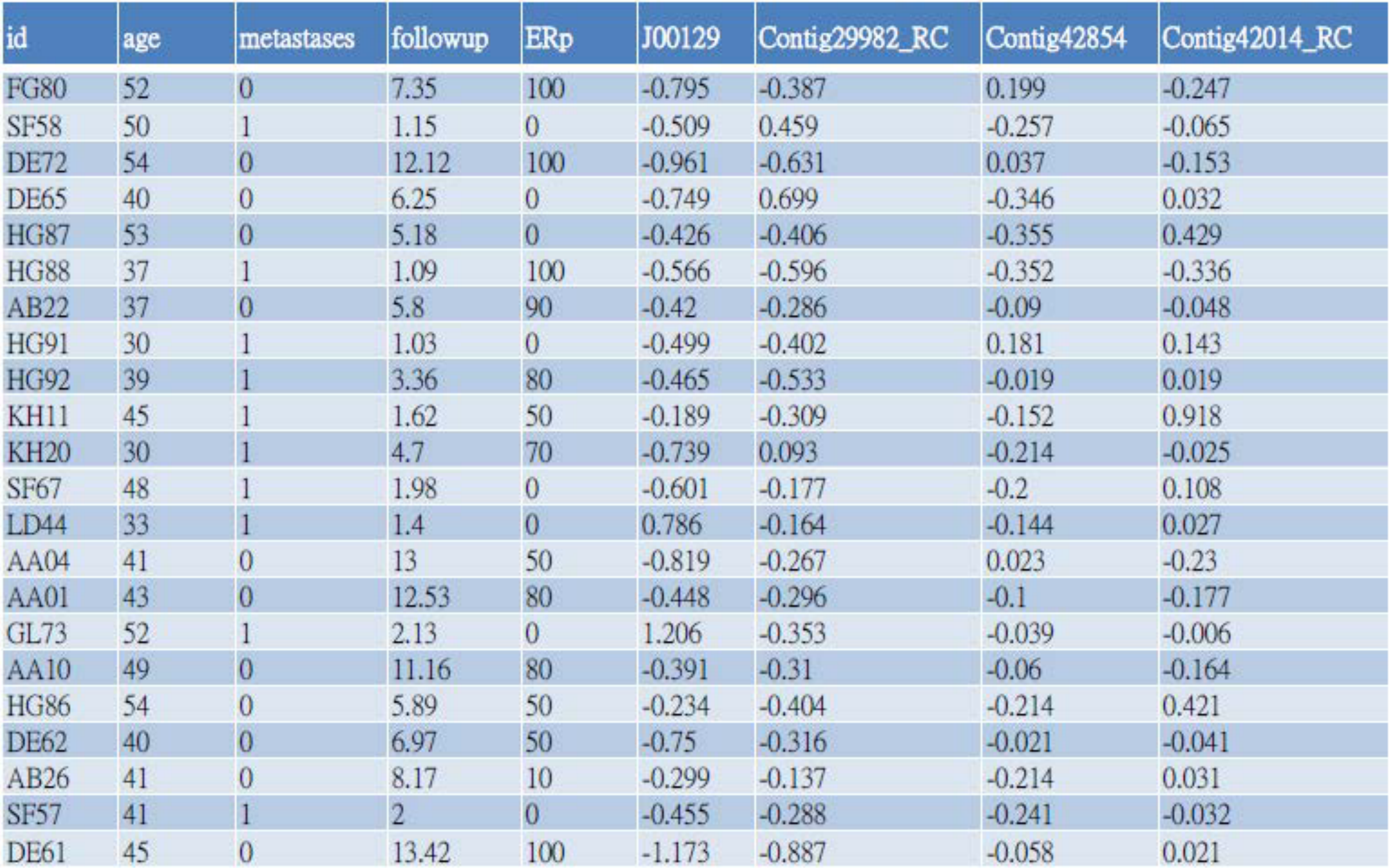
Example: Gene expression microarray data

- Data from a study using gene expression profiling to predict breast cancer outcomes (http://www.nature.com/nature/journal/v415/n6871/full/415530a.html)
- 78 breast cancer: 44 remained disease-free for an interval of at least five years after their initial diagnosis (good prognosis group), while 34 patients had developed distant metastases within five years (poor prognosis group)

samplexprs.csv

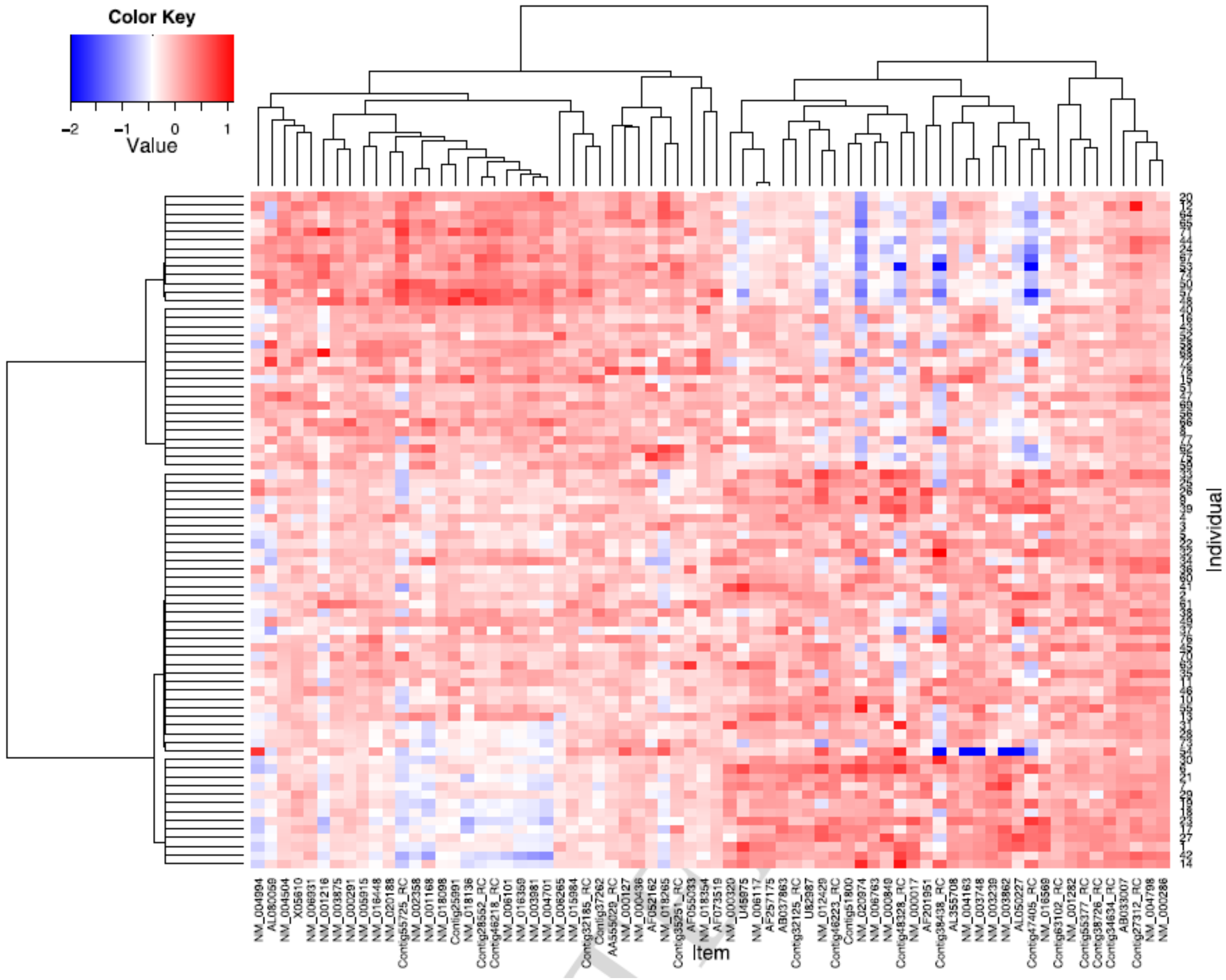
Variable	Description
id	An unique identification number
age	Age at diagnosis of breast cancer (year)
metastases	Developing distant metastases: 0=no (good prognosis group), 1=yes (poor prognosis group)
followup	Follow-up time (year)
ERp	ER-α expression level
J00129	log$_{10}$ gene expression intensity ratios
Contig29982_RC	log$_{10}$ gene expression intensity ratios

- RMD_example 5.1



id	age	metastases	followup	ERp	J00129	Contig29982_RC	Contig42854	Contig42014_RC
FG80	52	0	7.35	100	-0.795	-0.387	0.199	-0.247
SF58	50	1	1.15	0	-0.509	0.459	-0.257	-0.065
DE72	54	0	12.12	100	-0.961	-0.631	0.037	-0.153
DE65	40	0	6.25	0	-0.749	0.699	-0.346	0.032
HG87	53	0	5.18	0	-0.426	-0.406	-0.355	0.429
HG88	37	1	1.09	100	-0.566	-0.596	-0.352	-0.336
AB22	37	0	5.8	90	-0.42	-0.286	-0.09	-0.048
HG91	30	1	1.03	0	-0.499	-0.402	0.181	0.143
HG92	39	1	3.36	80	-0.465	-0.533	-0.019	0.019
KH11	45	1	1.62	50	-0.189	-0.309	-0.152	0.918
KH20	30	1	4.7	70	-0.739	0.093	-0.214	-0.025
SF67	48	1	1.98	0	-0.601	-0.177	-0.2	0.108
LD44	33	1	1.4	0	0.786	-0.164	-0.144	0.027
AA04	41	0	13	50	-0.819	-0.267	0.023	-0.23
AA01	43	0	12.53	80	-0.448	-0.296	-0.1	-0.177
GL73	52	1	2.13	0	1.206	-0.353	-0.039	-0.006
AA10	49	0	11.16	80	-0.391	-0.31	-0.06	-0.164
HG86	54	0	5.89	50	-0.234	-0.404	-0.214	0.421
DE62	40	0	6.97	50	-0.75	-0.316	-0.021	-0.041
AB26	41	0	8.17	10	-0.299	-0.137	-0.214	0.031
SF57	41	1	2	0	-0.455	-0.288	-0.241	-0.032
DE61	45	0	13.42	100	-1.173	-0.887	-0.058	0.021

Example: Gene expression microarray data (samplexprs.csv)



Heatmap for gene expression microarray data (samplexprs.csv)

Binomial test

- Compare the population proportion to a specified number
- Perform a test for testing whether or not the population proportion of ER negative ($p = \Pr(\mathrm{ERs} = 0)$) is equal to 0.4

	ERs - (0)	 + (1)	 total
number	22	56	78
prop.	0.28	0.72	1

ERs= $\begin{cases} 0, & \mathrm{ERp} \leq 10 \\ 1, & \mathrm{ERp} > 10 \end{cases}$

- $Ho: p = 0.4 \quad Ha: p \neq 0.4$

(RMD_example 5.2)

Prospective study
(cohort study 世代研究)

	metastases (outcome)		
ERs (risk)	good (0)	poor (1)	total
- (0)	9	13	22
+ (1)	35	21	56
total	44	34	78

$$\text{ERs} = \begin{cases} 0, & \text{ERp} \leq 10 \\ 1, & \text{ERp} > 10 \end{cases}$$

- Start with
  - 22 ERs negative patients
  - 56 ERs positive patients
- After a period of time, identify the numbers of patients who are poor or good groups.

Question: Does ER positive increase the likelihood of good prognosis?

- Good prognosis rates

ERs	
-	9/22 = 0.409
+	35/56 = 0.625
Total	44/78 = 0.564

- Calculate a risk ratio or “relative risk”

$$\mathrm{RR} = \frac{\Pr(\mathrm{good}|\mathrm{ERs}-)}{\Pr(\mathrm{good}|\mathrm{ERs}+)} = \frac{p_1}{p_2}$$

p_1 can be estimated by 9/22

p_2 can be estimated by 35/56

estimate of $\mathrm{RR} = \widehat{\mathrm{RR}} = \frac{9/22}{35/56} = 0.655$

- 34 percent increase in good prognosis!

- $\begin{cases} \mathrm{RR} = 1 \rightarrow \text{no association} \\ \mathrm{RR} > 1 \rightarrow \text{positive association} \\ \mathrm{RR} < 1 \rightarrow \text{negative association} \end{cases}$
- hypothesis testing:

$$\begin{cases} \mathrm{Ho}: \mathrm{RR} = 1 \\ \mathrm{Ha}: \mathrm{RR} \neq 1 \end{cases} \rightarrow \begin{cases} \mathrm{Ho}: p_1 = p_2 \\ \mathrm{Ha}: p_1 \neq p_2 \end{cases}$$

- RMD_example 5.3

x^2 test / Fisher's exact test

	metastases		
ERs	good (0)	poor (1)	total
- (0)	9 (*a*)	13 (*b*)	22
+ (1)	35 (*c*)	21 (*d*)	56
total	44	34	78

- Whether "metastases" is independent of "ERs", test the association between "metastases" and "ERs"

1. x^2 test if $a, b, c, d \geq 5$
2. Fisher's exact test if any $a, b, c, d < 5$

- Perform a x^2 test for testing whether or not the difference of the population proportions of being good prognosis between ERs- (p_1) and ERs+ (p_2) is equal to 0.
- Ho : $p_1 = p_2$ Ha : $p_1 \neq p_2$
- If $p < 0.05$ (significant), the probability of good prognosis for patients with ER negative as compared to patients with ER positive is 0.655. In other word, there is a possible 34% increase in being good prognosis when ER positive.
- RMD_example 5.4

Retrospective study (case-control study 病例對照研究)

	metastases		
ERs	good (0)	poor (1)	total
- (0)	9 (*a*)	13 (*b*)	22
+ (1)	35 (*c*)	21 (*d*)	56
total	44	34	78

- If, in fact, start with
  - 44 controls (good prognosis)
  - 34 cases (poor prognosis)
- Then, see how many controls with ER negative and how many cases with ER negative

- In case-control study, we cannot estimate $\Pr(\text{good}|\text{ERs}-)$, therefore, we cannot estimate RR.
- In case-control study, we can estimate $\Pr(\text{ERs}-|\text{good})$.
- The odds of good prognosis for ER negative is $\frac{p_1}{1-p_1}$.

  The odds of good prognosis for ER positive is $\frac{p_2}{1-p_2}$.

  The odds ratio = OR = $\frac{p_1/(1-p_1)}{p_2/(1-p_2)}$.
- The odds ratio can be estimated by

$$\widehat{\text{OR}} = \frac{ad}{bc} = \frac{9 \times 21}{13 \times 35} = 0.415$$

- The estimate of OR is good for both cohort and case-control study.
- When $\Pr(\text{good})$ is small, the odds ratio is approximately equal to the relative risk.

- $\begin{cases} \text{OR} = 1 \rightarrow \text{no association} \\ \text{OR} > 1 \rightarrow \text{positive association} \\ \text{OR} < 1 \rightarrow \text{negative association} \end{cases}$
- hypothesis testing:

$$\begin{cases} \text{Ho: OR} = 1 \\ \text{Ha: OR} \neq 1 \end{cases} \rightarrow \begin{cases} \text{Ho:}\, p_1 = p_2 \\ \text{Ha:}\, p_1 \neq p_2 \end{cases}$$

- RMD_example 5.3
- Use x^2 test / Fisher's exact test for hypothesis testing
- If p < 0.05 (significant), there is a possible 60% increase in the odds of being good prognosis when ER positive.

Notes on x^2 test / Fisher's exact test

- In cohort studies, ERs negative patients and ERs positive patients need to be independent.
- In case-control studies, controls (good prognosis) and cases (poor prognosis) need to be independent.

Matched-pair study

- Samples are not independent.
- Matched pairs (e.g., case-control pair matched on age in case-control studies)

	poor	good
ERs	0	1
	1	0
	0	1
	1	1
	1	1
	1	1
	0	1
	0	0
	0	0
	0	1
	1	1
	1	0
	0	1
	1	1
	1	1
	1	1
	1	1
	1	1
	1	1
	1	1
	1	1
	1	1
	1	0
	0	1
	0	0

Example: In a case-control studies, 25 poor prognosis (cases) match 25 good prognosis (controls) on age

- If the data are displayed in a way for regular x^2 tests

metastases

ERs	good (0)	poor (1)	total
- (0)	6	9	15
+ (1)	19	16	35
total	25	25	50

  - Cases and controls are not independent!
  - x^2 tests are not valid!

	poor	good
ERs	0	1
	1	0
	0	1
	1	1
	1	1
	1	1
	0	1
	0	0
	0	0
	0	1
	1	1
	1	0
	0	1
	1	1
	1	1
	1	1
	1	1
	1	1
	1	1
	1	1
	1	1
	1	1
	1	0
	0	1
	0	0

Example: In a case-control studies, 25 poor prognosis (cases) match 25 good prognosis (controls) on age

- The data are displayed in a different type of table

		cases (poor)		
		ERs-	ERs+	total
controls (good)	ERs-	3 (*a*)	3 (*b*)	6
	ERs+	6 (*c*)	13 (*d*)	19
	total	9	16	

a, *d*: concordant pairs = same exposure

b, *c*: discordant pairs = different exposure

- The concordant pairs give us no information about differences. We focus on the discordant pairs.
- The estimated odds ratio of being good prognosis for ERs- versus ERs+ is

$$\widehat{\mathrm{OR}} = \frac{b}{c} = \frac{3}{6} = 0.5$$

McNemar's test

		cases (poor)		
		ERs-	ERs+	total
controls	ERs-	3 (*a*)	3 (*b*)	6
(good)	ERs+	6 (*c*)	13 (*d*)	19
	total	9	16	

- “cases” and “controls” are not independent-- use McNemar’s test to test the association
- Matched case-control study, and “paired” study

- Perform a McNemar's test for testing if there is association between prognosis and ER status?

$$\begin{cases}\text{Ho: no association}\\ \text{Ha: have association}\end{cases}$$

- If $p < 0.05$ (significant), conclude that there appears to be increased probability of being good prognosis for ER positive.
- RMD_example 5.5

Measure of agreement

- Example:
  1. 2 physicians diagnose the same patients. Do physicians agree on diagnosis?
  2. Expression of J00129 and Contig29982_RC on the same patient. Do expressions of two genes agree?
- Expression agreement:

		Contig29982_RC > -0.5		
		no	yes	total
J00129 > -0.5	no	21	30	51
	yes	6	21	27
	total	27	51	78

Question: Is there agreement? How much?

1. Hypothesis test:

   Ho: no agreement between J00129 and Contig29982_RC (i.e., no association between gene and expression status)

   → use McNemar's test

2. The proportion of agreement = $\frac{21+21}{78} = 53.8\%$

   disadvantage:

   - very strongly influenced by the distribution of positive and negative
   - it's possible that there will be a high agreement by chance alone.

3. Kappa coefficient: measure of agreement excluding by chance alone

$$\kappa = \frac{p_0 - p_e}{1 - p_e},$$

where p_0 is the observed proportion of agreement, and p_e is the proportion expected by chance.

- R can calculate κ
- some guidelines:

$$\begin{cases} 0.8 \le \kappa\text{: almost perfect agreement} \\ 0.6 \le \kappa < 0.8\text{: substantial agreement} \\ 0.4 \le \kappa < 0.6\text{: moderate agreement} \\ 0.2 \le \kappa < 0.4\text{: fair agreement} \\ 0 \le \kappa < 0.2\text{: slight agreement} \\ \kappa < 0\text{: agreement is same as random} \end{cases}$$

Cohen's kappa test

- Ho: $\kappa = 0$ Ha: $\kappa \neq 0$
- If not significant, the extent of agreement is same as random.
- RMD_example 5.6

When there are more than one table

$\text{age} \leq 45$

		metastases		
		good	poor	total
ERs	-	4	6	10
	+	12	19	31
	total	16	25	41

$OR_1 = 1.056$

$\text{age} > 45$

		metastases		
		good	poor	total
ERs	-	5	7	12
	+	23	2	25
	total	28	9	37

$OR_2 = 0.062$

Interaction (交互作用)

- The associations between metastases and ERs are different in different age groups, i.e., $\mathrm{OR}_1 \neq \mathrm{OR}_2$
- Should show the OR for each age group.
- RMD_example 5.7

Simpson's paradox

OR=0.905 Lung cancer

	No	Yes	Total
Non-smokers	176	64	240
Smokers	158	52	210
Total	334	116	450

Males

$OR_M=1.992$ Lung cancer

	No	Yes	Total
Non-smokers	36	4	40
Smokers	131	29	160
Total	167	33	200

Females

Lung cancer $OR_F=1.988$

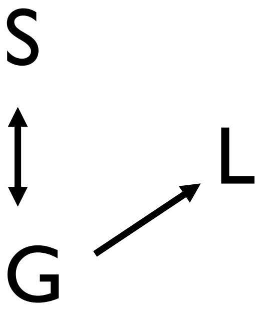
	No	Yes	Total
Non-smokers	140	60	200
Smokers	27	23	50
Total	167	83	250

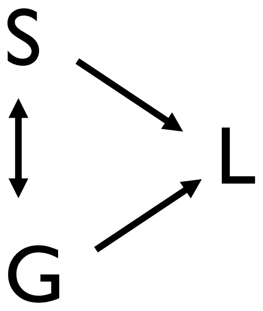
Why?

- Most smokers are males.
- But, it is a disease more prevalent in females.

Confounding (干擾)

POTENTIAL CONFOUNDER:





⟶ causal

⟷ associated

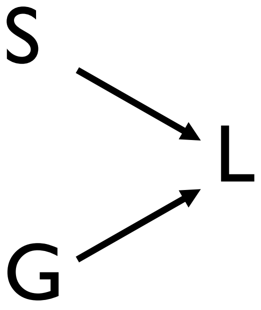
S=Smoke (risk)

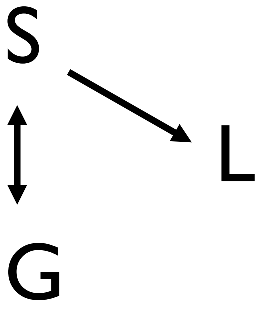
L=Lung cancer (outcome)

G=gender (confounder)

(干擾因子)

NOT A POTENTIAL CONFOUNDER:





Confounding

- The association between risk (smoking) and outcome (lung cancer) is the same in different confounder (gender) groups, but is different from the combined one (i.e., combining males and females), i.e., $\mathrm{OR}_\mathrm{M} = \mathrm{OR}_\mathrm{F} \neq \mathrm{OR}$

∴假設所有年齡層的odds ratio相同

① 得到整體的、代表性的估計值.

② 同時控制混淆因子（ex、年齡）

③ 方便比較

↗ 每層OR差異不大：Mantel-Haenszel 得到加權平均的共同odds ratio

↘ 每層OR差異很大：Breslow-Day test 檢定各層OR是否統一

Mantel-Haenszel test

- If the population is stratified (by “gender”), we then use Mantel-Haenszel test to test the association between “smoking” and “lung cancer” **after adjusting for “gender”**.
- Under Mantel-Haenszel test, we assume that the odds ratio between “smoking” and “lung cancer” for “males” is the same as the odds ratio for “females”, i.e., $\mathrm{OR}_\mathrm{M} = \mathrm{OR}_\mathrm{F}$

Mantel-Haenszel test (cont'd)

- We can use Breslow-Day test for homogeneity of the odds ratios.
- Use Mantel-Haenszel odds ratio (relative risk) to estimate the common odds ratio (relative risk): Take a weighted average of $\mathrm{OR}_{\mathrm{M}}, \mathrm{OR}_{\mathrm{F}}$ with weights $r_{\mathrm{M}}, r_{\mathrm{F}}$:

$$\frac{r_{\mathrm{M}}\mathrm{OR}_{\mathrm{M}} + r_{\mathrm{F}}\mathrm{OR}_{\mathrm{F}}}{r_{\mathrm{M}} + r_{\mathrm{F}}}$$

- RMD_example 5.8