Three or more categorical variables

2.3 PARTIAL ASSOCIATION IN STRATIFIED 2x2 TABLES

- An important part of most studies, especially observational studies, is the choice of control variables.
- In studying the effect of *X on Y, one should* control any covariate that can influence that relationship.
- This involves using some mechanism to hold the covariate constant. Otherwise, an observed effect of *X on Y may actually reflect effects of that covariate on both X and Y.*
- *The relationship between X and Y then shows confounding.*
- *Experimental* studies can remove effects of confounding covariates by randomly assigning subjects to different levels of *X, but this is not possible with observational* studies.

Confounding example

- Study: effects of passive smoking with lung cancer
- A cross-sectional study might compare lung cancer rates between nonsmokers whose spouses smoke and nonsmokers whose spouses do not smoke.
- The study should attempt to control for age, socioeconomic status, or other factors that might relate both to spouse smoking and to developing lung cancer.
- Otherwise, results will have limited usefulness.
- Spouses of nonsmokers may tend to be younger than spouses of smokers, and younger people are less likely to have lung cancer.
- Then a lower proportion of lung cancer cases among spouses of nonsmokers may merely reflect their lower average age.

- the analysis of the association between categorical variables *X and Y while controlling for a possibly confounding variable Z.*
- *For simplicity, the examples refer to a single control variable.*
- *In later* chapters we treat more general cases and discuss the use of models to perform statistical control.

2.3.1 Partial Tables

- We control for *Z by studying the XY relationship at fixed levels of Z.*
- Two-way cross-sectional slices of the three-way contingency table cross classify *X and Y at separate categories of Z.*
- *These cross sections are called partial tables.*
- *They display the XY relationship while removing the effect of Z by holding its value constant.*

marginal table

- The two-way contingency table obtained by combining the partial tables is called the *XY marginal table.*
- *Each cell count in the marginal table is a sum of* counts from the same location in the partial tables.
- The marginal table, rather than controlling *Z, ignores it.*
- *The marginal table contains no information* about *Z.*
- *It is simply a two-way table relating X and Y but may reflect* the effects of *Z on X and Y.*

conditional associations and marginal associations

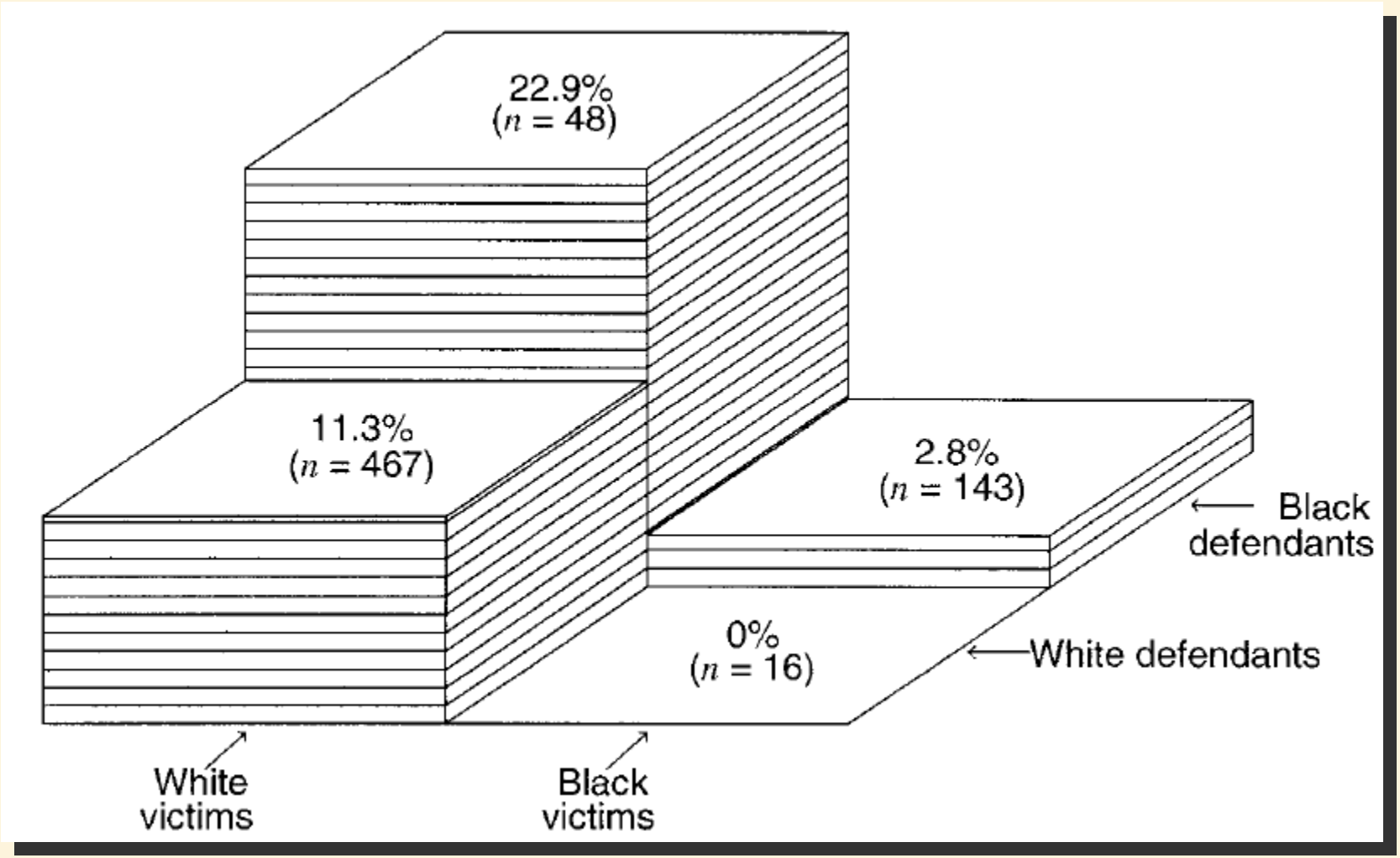
- The associations in partial tables are called *conditional associations, because* they refer to the effect of *X on Y conditional on fixing Z at some* level.
- Conditional associations in partial tables can be quite different from associations in marginal tables.
- In fact, it can be misleading to analyze only marginal tables of a multiway contingency table.

2.3.2 Death Penalty Example

TABLE 2.6 Death Penalty Verdict by Defendant's Race and Victims' Race

Victims' Race	Defendant's Race	Death Penalty		Percent Yes
		Yes	No	
White	White	53	414	11.3
	Black	11	37	22.9
Black	White	0	16	0.0
	Black	4	139	2.8
Total	White	53	430	11.0
	Black	15	176	7.9

It studied effects of racial characteristics on whether persons convicted of homicide received the death penalty.



Percent receiving death penalty.

- Variables:
  - Y=death penalty verdict, having the categories (yes, no),
  - X=race of defendant
  - Z=race of victims (white, black).
- Study: the effect of defendant's race on the death penalty verdict, treating victims' race as a control variable.

Conditional and marginal

```
DATA deathPenalty;
   input Z $ X $ y1 y2;
   Y="Yes"; w=y1; output;
   Y="No "; w=y2; output;
   drop y1 y2;
   cards;
White White  53 414
White Black  11 37
Black White   0 16
Black Black   4 139
;
```

```
proc sort; by Z;

proc freq; weight w;
   by Z;
   tables X*Y /nopercent
   nocol;
   run;

proc freq; weight w;
   tables X*Y /nopercent
   nocol;
   run;
```

---------- Z=White ----------

The FREQ Procedure

Table of X by Y

X	Y		
Frequency Row Pct	No	Yes	Total
Black	37 77.08	11 22.92	48
White	414 88.65	53 11.35	467
Total	451	64	515

---------- Z=Black ----------

The FREQ Procedure

Table of X by Y

X	Y		
Frequency Row Pct	No	Yes	Total
Black	139 97.20	4 2.80	143
White	16 100.00	0 0.00	16
Total	155	4	159

Table of X by Y

X	Y		
Frequency Row Pct	No	Yes	Total
Black	176 92.15	15 7.85	191
White	430 89.03	53 10.97	483
Total	606	68	674

Conditional Y*X given Z

- When the victims were white, the death penalty was imposed 22.9%-11.3%=11.6% more often for black defendants than for white defendants.
- When the victims were black, the death penalty was imposed 2.8% more often for black defendants than for white defendants.

- *Controlling for victims' race by keeping it fixed, the* death penalty was imposed more often on black defendants than on white defendants.

Marginal Y*X

- Overall, 11.0% of white defendants and 7.9% of black defendants received the death penalty.
- *Ignoring victims' race, the death penalty was imposed less often on* black defendants than on white defendants.
- The association reverses direction compared to the partial tables.

Why does the association change so much when we ignore versus control victims' race?

- This relates to the nature of the association between victims' race and each of the other variables.
- First, the association between victims' race and defendant's race is extremely strong.

(I) The marginal table relating these variables has odds ratio (467*143)/(48*16)=87.0

So whites are tending to kill whites

Marginal table Z*X

```
proc freq;
  weight w;
  tables X*Z
/nopercent nocol norow;
  run;
```

X / Z	Black	White	Total
Black	143	48	191
White	16	467	483
Total	159	515	674

Marginal table Z*Y

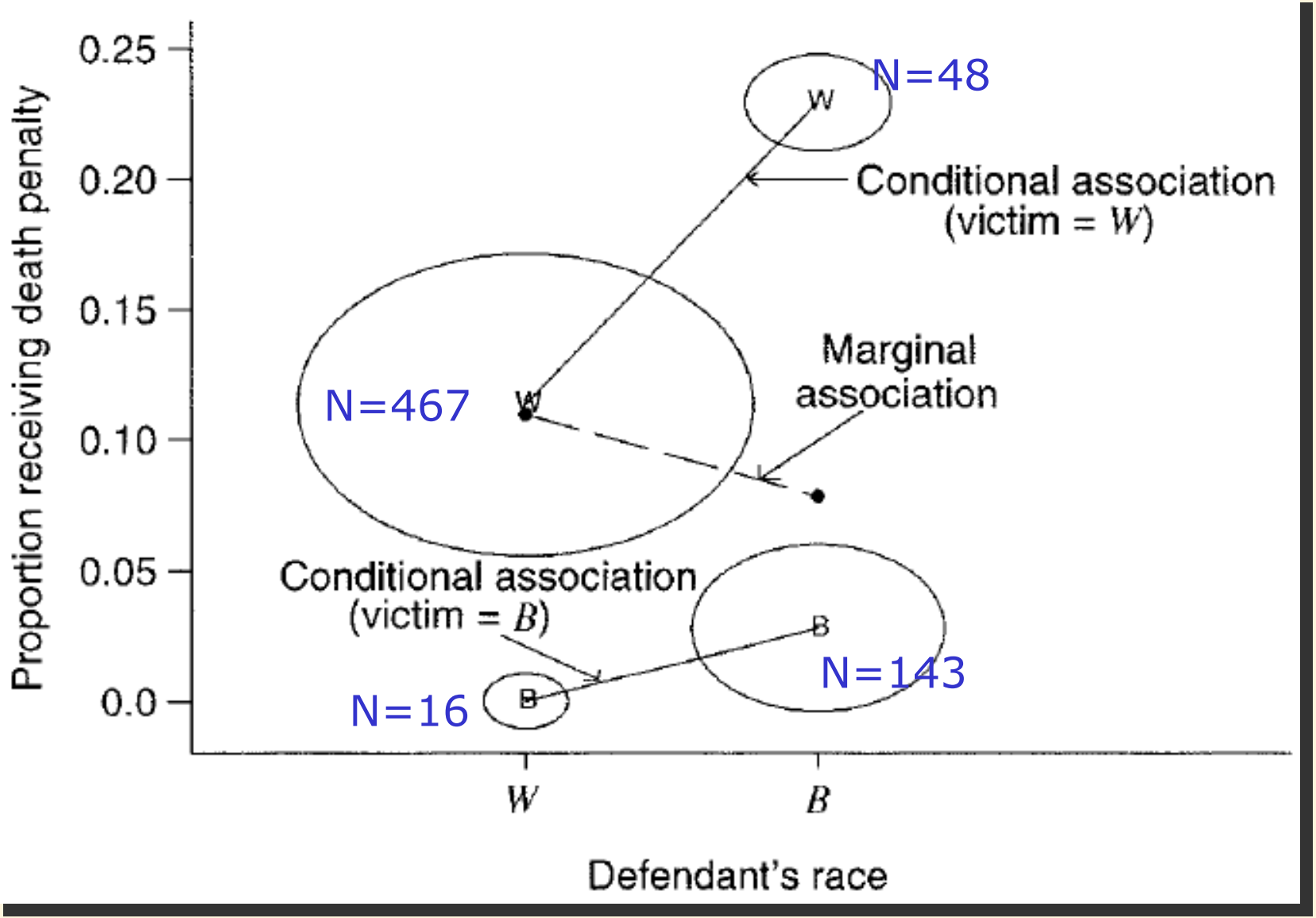
(II) regardless of defendant's race, the death penalty was much more likely when the victims were white than when the victims were black.

Y / Z	Black	White	Total
No	155	451	606
Yes	4	64	68
Total	159	515	674

Killing whites is more likely to result in the death penalty.

the marginal association should show a greater tendency than the conditional associations for white defendants to receive the death penalty.

Proportion receiving death penalty by defendant's race, controlling and ignoring victims' race.



Simpson's paradox (Simpson 1951, Yule 1903)

- The result that a marginal association can have a different direction from each conditional association is called *Simpson's paradox*
- It applies to quantitative as well as categorical variables.
- Statisticians commonly use it to caution against imputing causal effects from an association of *X with Y.*
- *For instance, when doctors started to observe strong odds* ratios between smoking and lung cancer, Statisticians such as R. A. Fisher warned that some variable (e.g., a genetic factor) could exist such that the association would disappear under the relevant control.
- However, with a very strong *XY* association, a very strong association must exist between the confounding variable *Z and both X and Y in order for the effect to disappear or change* under the control.

2.3.3 Conditional and Marginal Odds Ratios for 2 x 2 x *K* tables

- for 2x2x*K* tables, where *K* denotes the number of categories of a control variable
- Let $\{\mu_{ijk}\}$ denote cell expected frequencies for some sampling model, such as binomial, multinomial, or Poisson sampling.
- Within a fixed category *k of Z, the odds ratio*

$$\theta_{XY(k)} = \frac{\mu_{11k}\mu_{22k}}{\mu_{12k}\mu_{21k}} \quad \text{or sample OR} \quad \hat{\theta}_{XY(k)} = \frac{n_{11k}n_{22k}}{n_{12k}n_{21k}}$$

  describes conditional *XY association in partial table k.*
- The odds ratios for the *K partial tables are called XY conditional odds ratios.*

- *XY marginal odds ratio* *sample OR*

$$\theta_{XY} = \frac{\mu_{11+}\mu_{22+}}{\mu_{12+}\mu_{21+}} \qquad \hat{\theta}_{XY} = \frac{n_{11+}n_{22+}}{n_{12+}n_{21+}}$$

TABLE 2.6 Death Penalty Verdict by Defendant's Race and Victims' Race

Victims' Race	Defendant's Race	Death Penalty Yes	Death Penalty No	Percent Yes
White	White	53	414	11.3
	Black	11	37	22.9
Black	White	0	16	0.0
	Black	4	139	2.8
Total	White	53	430	11.0
	Black	15	176	7.9

association between defendant's race and the death penalty

- Given victims' race is white

$$\hat{\theta}_{XY(1)} = \frac{53 \times 37}{414 \times 11} = 0.43$$

The sample odds for white defendants receiving the death penalty were 43% of the sample odds for black defendants.

- Given victims' race is black

$$\hat{\theta}_{XY(2)} = \frac{0 \times 139}{16 \times 4} = 0$$

- Estimation of the marginal odds ratio

$$\hat{\theta}_{XY} = \frac{53 \times 176}{430 \times 15} = 1.45$$

- The sample odds of the death penalty were 45% higher for white defendants than for black defendants.

2.3.4 Marginal versus Conditional Independence

- If X and Y are independent in partial table *k*, then X and Y are called *conditionally independent* at level *k* of Z.

$$P(Y=j \mid X=i, Z=k) = P(Y=j \mid Z=k), \qquad \text{for all } i, j.$$

- X and Y are said to be conditionally independent given Z when they are conditionally independent at every level of Z
- Then, given Z, Y does not depend on X.
- conditional independence is then equivalent to

$$\pi_{ijk} = \pi_{i+k}\pi_{+jk}/\pi_{++k} \qquad \text{for all } i, j, \text{ and } k.$$

- summing over *k* on both sides yields

$$\pi_{ij+} = \sum_k \left(\pi_{i+k}\pi_{+jk}/\pi_{++k}\right).$$

$$\pi_{ij+} = \sum_k \left(\pi_{i+k} \pi_{+jk} / \pi_{++k} \right).$$

- Marginal Independence $\pi_{ij+} = \pi_{i++} \pi_{+j+}$
- Obviously, Conditional Independence Does Not Imply Marginal Independence

TABLE 2.7 Expected Frequencies Showing That Conditional Independence Does Not Imply Marginal Independence

Clinic	Treatment	Response: Success	Response: Failure
1	A	18	12
	B	12	8
2	A	2	8
	B	8	32
Total	A	20	20
	B	20	40

$$\theta_{XY(1)} = \frac{18 \times 8}{12 \times 12} = 1.0, \qquad \theta_{XY(2)} = \frac{2 \times 32}{8 \times 8} = 1.0.$$

- Given the clinic, response and treatment are conditionally independent.
- Ignore the clinical, response and treatment are not marginally independent.

$$\theta_{XY} = (20 \times 40)/(20 \times 20) = 2.0$$

- Ignoring the clinic, why are the odds of a success for treatment A twice those for treatment B?
- The conditional XZ and YZ odds ratios give a clue.
- The odds ratio between Z and either X or Y, at each fixed category of the other variable, equals 6.0. For instance, the XZ odds ratio at the first category of Y equals?18*8/12*2=6.0.
- The conditional odds (given response) of receiving treatment A at clinic 1 are six times those at clinic 2, and the conditional odds (given treatment) of success at clinic 1 are six times those at clinic 2.
- Clinic 1 tends to use treatment A more often, and clinic 1 also tends to have more successes.
- For instance, if patients at clinic 1 tended to be younger and in better health than those at clinic 2, perhaps they had a better success rate regardless of the treatment received.

2.3.5 Homogeneous Association

A $2 \times 2 \times K$ table has *homogeneous XY association* when

$$\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)}.$$

- Then the effect of X on Y is the same at each category of Z.
- Conditional independence of X and Y is the special case in which each $\theta_{XY(k)} = 1.0.$
- Under homogeneous XY association, homogeneity also holds for the other associations. (symmetric)
- When it occurs, there is said to be no interaction between two variables in their effects on the other variable.

Summary: Contingency Tables

- Sampling schemes:
  - the overall n is not fixed, $n_{ij} \sim Poisson(\mu_{ij})$
  - n is fixed, $n_{ij} \sim Mult(n, \{\pi_{ij}\})$
  - Row total is fixed, product multinomial

    $$(n_{i1}, \cdots, n_{iJ}) \sim Mult(n_{i+}, \{\pi_{1|i}, \cdots, \pi_{J|i}\})$$

    such as
    - Stratified random sampling (strata defined by X)
    - An experiment where X=treatment group
    - Interested in $P(Y|X)$ and not $P(X)$
  - Hypergeometric sampling

- Measure of association
  - Difference in proportions
  - Relative risk
  - Odd ratio
- Independence

3-way tables

- Conditional and Marginal Odds Ratios for 2 x 2 x K tables (*Simpson's paradox*)
- Marginal versus Conditional Independence
- Homogeneous Association