Ch. 3 Supplement
Associations in Categorical Data

Two-Way Tables

1

Objectives

- Given a two-way table of counts for two ***categorical variables***:
  - *Find the marginal distributions of the variables*
  - *Find a conditional distribution of the variables*
  - *Display the distributions as bar charts*
- In this lesson, we will study the relationship between two ***categorical variables*** using
  - *Counts*
  - *Marginal percents*
  - *Conditional percents*

2

Objectives

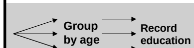
- *Relationships between categorical variables*
- *Simpson's paradox*

3

Two-way tables

An experiment has a ***two-way,*** or block, design if two ***categorical*** factors are studied with several levels of each factor.

Two-way tables organize data about two categorical variables obtained from a two-way, or block, design. (*There are now two ways to group the data.*)



First factor: *age*

Second factor: *education*

Years of school completed, by age (thousands of persons)

	Age group		
Education	25 to 34	35 to 54	55 and over
Did not complete high school	4,459	9,174	14,226
Completed high school	11,562	26,455	20,060
College, 1 to 3 years	10,693	22,647	11,125
College, 4 or more years	11,071	23,160	10,597

Two-Way Tables

- Data are cross-tabulated to form a two-way table with a *row variable* and *column variable*
- The count of observations falling into each combination of categories is cross-tabulated into each table cell
- Counts are totaled to create marginal totals

5

Overview

- ***Two way table*** - Presenting the data
- Describing each variable separately (***Marginal Distribution***)
- Describing the relation between the two variables (***Conditional Distribution***)

6

Two types of categorical variables:

1. Those that are inherently categorical.
   Example: eye color, gender, city.

2. Those that are obtained by grouping quantitative variables into classes.
   Example: age groups 25-34, 35-54, 55 and over.

7

Marginal distributions

We can look at each categorical variable in a two-way table separately by studying the row totals and the column totals. They represent the ***marginal distributions,*** expressed in counts or percentages (they are written as if in a margin).

Years of school completed, by age (thousands of persons)

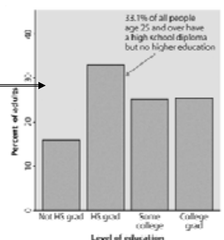
Education	Age group: 25 to 34	35 to 54	55 and over	Total
Did not complete high school	4,459	9,174	14,226	27,859
Completed high school	11,562	26,455	20,060	58,077
College, 1 to 3 years	10,693	22,647	11,125	44,465
College, 4 or more years	11,071	23,160	10,597	44,828
Total	37,786	81,435	56,008	175,230

2000 US census

The ***marginal distributions*** can then be displayed on separate bar graphs, typically expressed as percents instead of raw counts. Each graph represents only one of the two variables, completely ignoring the second one.

Years of school completed, by age (thousands of persons)

Education	Age group: 25 to 34	35 to 54	55 and over	Total
Did not complete high school	4,459	9,174	14,226	27,859
Completed high school	11,562	26,455	20,060	58,077
College, 1 to 3 years	10,693	22,647	11,125	44,465
College, 4 or more years	11,071	23,160	10,597	44,828
Total	37,786	81,435	56,008	175,230



Example in Class

9

Parental smoking

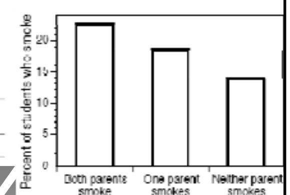
Does parental smoking influence the smoking habits of their high school children?

Summary two-way table:
High school students were asked whether they smoke and whether their parents smoke.

	Student smokes	Student does not smoke	*Total*
Both parents smoke	400	1380	1780
One parent smokes	416	1823	2239
Neither parent smokes	188	1168	1356
Total	1004	4371	5375

Marginal distribution for the categorical variable "***parental smoking***":
The row totals are used and re-expressed as a percent from the grand total.

Neither parent smokes	One parent smokes	Both parents smoke
13.9%	18.6%	22.5%



The percentages are then displayed in a bar graph.

10

Relationships between categorical variables

The cells of a ***two-way table represent the intersection*** of a given level of one categorical factor with a given level of the other categorical factor.

The ***marginal distributions*** summarize each categorical variable independently. But the two-way table actually describes the relationship between both categorical variables.

Because counts can be misleading (for instance, one level of one factor might be much less represented than the other levels), we prefer to ***calculate percents or proportions*** for the corresponding cells. These make up the ***conditional distributions.***

11

Conditional distributions

The counts or percents within the table represent the ***conditional distributions.*** Comparing the conditional distributions allows us to describe the "relationship" between both categorical variables.

Years of school completed, by age (thousands of persons)

Education	Age group: 25 to 34	35 to 54	55 and over	Total
Did not complete high school	4,459	9,174	14,226	27,859
Completed high school	11,562	26,455	20,060	58,077
College, 1 to 3 years	10,693	22,647	11,125	44,465
College, 4 or more years	11,071	23,160	10,597	44,828
Total	37,786	81,435	56,008	175,230

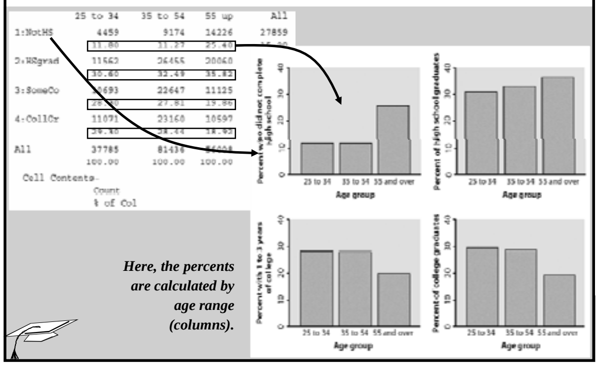
	25 to 34	35 to 54	55 up	All
1:NotHS	4459	9174	14226	27859
	11.80	11.27	25.40	15.90
2:HSgrad	11562	26455	20060	58077
	30.60	32.49	35.82	33.14
3:SomeCo	10693	22647	11125	44465
	28.30	27.81	19.86	25.38
4:CollGr	11071	23160	10597	44828
	29.30	28.44	18.92	25.58
All	37785	81436	56008	175229
	100.00	100.00	100.00	100.00

Cell Contents--
Count
% of Col

Here, the percents are calculated by age range (columns).

$$29.30\% = \frac{11071}{37785} = \frac{\text{cell total}}{\text{column total}}$$

The conditional distributions can be ***graphically compared*** using side-by-side bar graphs of one variable for each value of the other variable.



Here, the percents are calculated by age range (columns).

Music and wine purchase decision

What is the relationship between type of music played in supermarkets and type of wine purchased?

Wine	Music: None	French	Italian	Total
French	30	39	30	99
Italian	11	1	19	31
Other	43	35	35	113
Total	84	75	84	243

We want to compare the conditional distributions of the response variable (***wine purchased***) for each value of the explanatory variable (***music played***). Therefore, we calculate column percents.

Calculations: When no music was played, there were 84 bottles of wine sold. Of these, 30 were French wine.
30/84 = 0.357 ➔ **35.7%** of the wine sold was French when no music was played.

$\frac{30}{84} = 35.7\%$

$= \frac{\text{cell total}}{\text{column total}}$

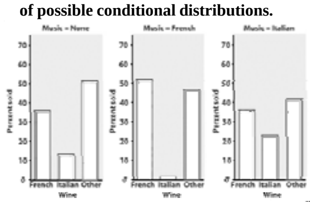
Column percents for wine and music

Wine	Music: None	French	Italian	Total
French	35.7	52.0	35.7	40.7
Italian	13.1	1.3	22.6	12.8
Other	51.9	46.7	41.7	46.5
Total	100.0	100.0	100.0	100.0

We calculate the column conditional percents similarly for each of the nine cells in the table.

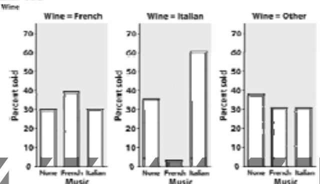
For every two-way table, there are two sets of possible conditional distributions.

Wine	Music: None	French	Italian	Total
French	30	39	30	99
Italian	11	1	19	31
Other	43	35	35	113
Total	84	75	84	243



Wine purchased for each kind of music played (column percents)

Does background music in supermarkets influence customer purchasing decisions?



Music played for each kind of wine purchased (raw percents)

15

Case Study

Age and Education

(*Statistical Abstract of the United States*, 2001)

Data from the U.S. Census Bureau (2000)

Level of education by age

16

Case Study

Age and Education

Education	*Age Group* 25 to 34	35 to 54	55+	TOTAL
Did not complete HS	4,459	9,174	14,226	27,859
Completed HS	11,562	26,455	20,060	58,077
College, 1 to 3 years of college	10,693	22,647	11,125	44,465
College, 4 or more years of college	11,071	23,160	10,597	44,828
TOTAL	37,785	81,436	56,008	175,229

17

Case Study

Age and Education

TABLE 4.6 Years of school completed, by age (thousands of persons)

Variables

Education	Age group: 25 to 34	35 to 54	55 and over	Total
Did not complete high school	4,459	9,174	14,226	27,859
Completed high school	11,562	26,455	20,060	58,077
College, 1 to 3 years	10,693	22,647	11,125	44,465
College, 4 or more years	11,071	23,160	10,597	44,828
Total	37,786	81,435	56,008	175,230

Marginal distributions

18

Case Study

Age and Education

TABLE 4.6 Years of school completed, by age (thousands of persons)

Variables → Age group

Education	25 to 34	35 to 54	55 and over	Total
Did not complete high school	4,459	9,174	14,226	27,858
Completed high school	11,562	26,455	20,060	58,077
College, 1 to 3 years	10,693	22,647	11,125	44,465
College, 4 or more years	11,071	23,160	10,597	44,828
Total	37,786	81,435	56,008	175,230

Marginal totals

19

Marginal Percents

- It is more informative to display counts as *percents*
- ***Marginal percents***

$$\text{marginal percent} = \frac{\text{marginal total}}{\text{table total}} \times 100\%$$

- Use a bar graph to display marginal percents (optional)

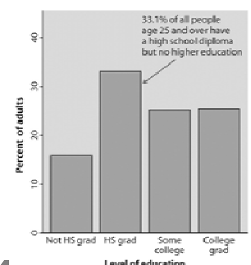
20

Case Study

Age and Education

Row Marginal Distribution

Did not graduate HS	(27,859 ÷ 175,230) × 100% = **15.9%**
Did graduate HS	(58,077 / 175,230) × 100% = **33.1%**
Finished 1-3 yrs college	(44,465 / 175,230) × 100% = **25.4%**
Finished ≥4 yrs college	(44,828 / 175,230) × 100% = **25.6%**



21

Conditional Percents

- Relationships are described with conditional percents
- There are two types of conditional percents:
  - Column percents
  - Row percents

22

Row Conditional Percent Column Conditional Percent

$$\text{column percent for cell} = \frac{\text{cell count}}{\text{column total}} \times 100\%$$

$$\text{row percent for cell} = \frac{\text{cell count}}{\text{row total}} \times 100\%$$

To know which one to use, ask

"What comparison is most relevant?"

23

Case Study

Age and Education

Compare the 25-34 age group to the 35-54 age group in % completing college:

Education	25 to 34	35 to 54	55 and over
Did not complete high school	4,459	9,174	14,226
Completed high school	11,562	26,455	20,060
College, 1 to 3 years	10,693	22,647	11,125
College, 4 or more years	11,071	23,160	10,597
Total	37,786	81,435	56,008

Change the counts to column percents (important):

24

Case Study

Age and Education

If we compute the percent completing college for all of the age groups, this gives conditional distribution (column percents) completing college by age:

Age:	25-34	35-54	55 and over
Percent with ≥ 4 yrs college:	**29.3%**	**28.4%**	**18.9%**

25

Association

- If the conditional distributions are nearly the same, then we say that there is **not** an association between the row and column variables
- If there are significant differences in the conditional distributions, then we say that there **is** an association between the row and column variables

26

Simpson's Paradox

- Simpson's paradox occurs when an association between two variables is reversed upon observing a third variable.
- Simpson's paradox ≡ a **lurking variable** creates a **reversal** in the direction of the association
- To uncover Simpson's Paradox, divide data into **subgroups** based on the lurking variable

27

Simpson's Paradox

Beware of lurking variables

An association or comparison that holds for all of several groups can reverse direction when the data are combined to form a single group. This reversal is called ***Simpson's paradox.***

Example: Hospital death rates

	Hospital A	Hospital B
Died	63	16
Survived	2037	784
Total	2100	800
% Surv.	97.0%	98.0%

On the surface, ***Hospital B*** would seem to have a better record.

But once patient condition is taken into account, we see that, in fact, ***Hospital A*** has a better record for both patient conditions (good and poor).

Patients in good condition

	Hospital A	Hospital B
Died	6	8
Survived	594	592
Total	600	600
% surv.	99.0%	98.7%

Patients in poor condition

	Hospital A	Hospital B
Died	57	8
Survived	1443	192
Total	1500	200
% surv.	96.2%	96.0%

Here, patient condition was the lurking variable.

Discrimination? (Simpson's Paradox)

Consider college acceptance rates by sex.

	Accepted	Not accepted	*Total*
Men	198	162	360
Women	88	112	200
Total	286	274	***560***

198 of 360 (**55%**) of men accepted
88 of 200 (**44%**) of women accepted

Is this discrimination?

29

Discrimination? (Simpson's Paradox)

- Or is there a lurking variable that explains the association?
- To evaluate this, split applications according to the lurking variable "***School applied to***"
  - Business School (240 applicants)
  - Art School (320 applicants)

30

Discrimination? (Simpson's Paradox)

BUSINESS SCHOOL

	Accepted	Not accepted	*Total*
Men	18	102	**120**
Women	24	96	**120**
Total	**42**	**198**	***240***

18 of 120 men (**15%**) of men were accepted to B-school
24 of 120 (**20%**) of women were accepted to B-school
A higher percentage of **women** were accepted

31

Discrimination? (Simpson's Paradox)

ART SCHOOL

	Accepted	Not accepted	*Total*
Men	180	60	**240**
Women	64	16	**80**
Total	**244**	**76**	***320***

180 of 240 men (**75%**) of men were accepted
64 of 80 (**80%**) of women were accepted
A higher percentage of **women** were accepted.

32

Discrimination? (Simpson's Paradox)

- Within each school, a higher percentage of women were accepted than men. (There was not any discrimination against women.)
- This is an example of Simpson's Paradox.
  - When the **lurking variable** (***School applied to***) was ignored, the data suggest discrimination against women.
  - When the School applied to was considered, the association is reversed.

33

Example 4.19, p. 241

TABLE 4.6 Years of school completed, by age, 2000 (thousands of persons)

	Age group			
Education	25 to 34	35 to 54	55+	Total
Did not complete high school	4,474	9,155	14,224	27,853
Completed high school	11,546	26,481	20,060	58,087
1 to 3 years of college	10,700	22,618	11,127	44,445
4 or more years of college	11,066	23,183	10,596	44,845
Total	37,786	81,435	56,008	175,230

34

Example 4.20, p. 242: Marginal Distribution

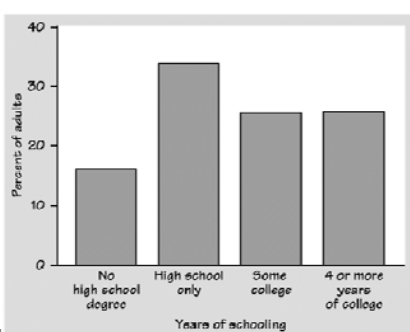


Figure 4.24, p. 243

35

Example 4.21, p. 244: How Common is College Education?

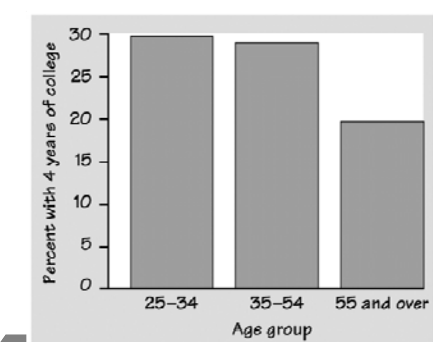


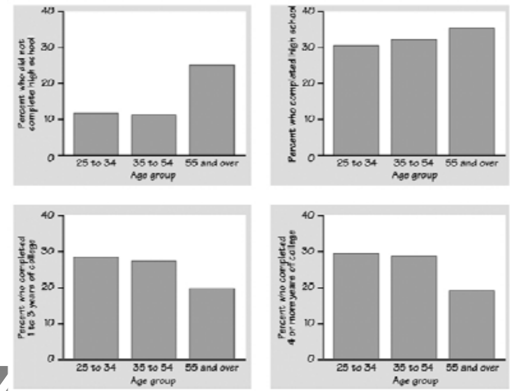
Figure 4.25, p. 244

36

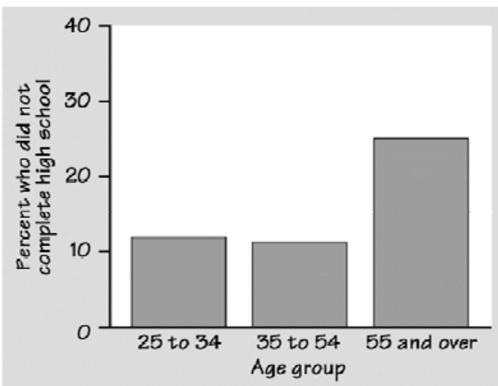
TABLE OF EDU BY AGE

EDU / Frequency / Col Pct	AGE 25-34	35-54	55 over	Total
NoHS	4474 11.84	9155 11.24	14224 25.40	27853
HSonly	11546 30.56	26481 32.52	20060 35.82	58087
SomeColl	10700 28.32	22618 27.77	11127 19.87	44445
Coll4yrs	11066 29.29	23183 28.47	10596 18.92	44845
Total	37786	81435	56008	175230

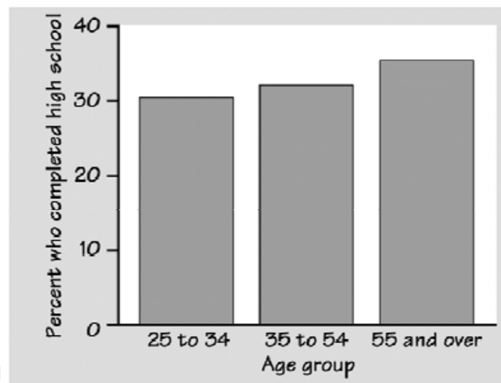
37



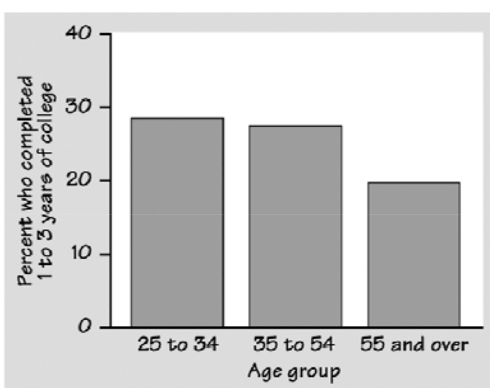
38



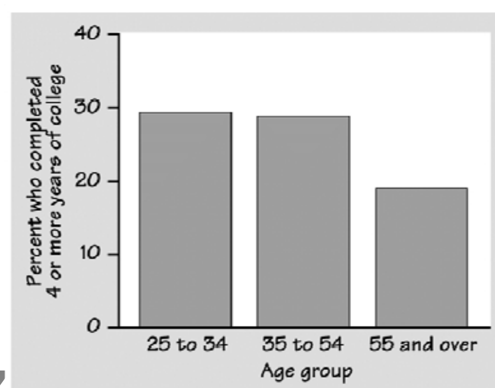
39



40



41



42

The *conditional distributions* and *Plot*

A sample of 500 persons is questioned regarding political affiliation and attitude toward a proposed national health care plan. The responses are cross classified according to the political affiliation and opinion categories and displayed in the following 2×3 two way table (also called ***contingency table***)

	Attitude			
Affiliation	favor	Indifferent	opposed	***Total***
Democrat	138	83	64	285
Republican	64	67	84	215
Total	202	150	148	500

43

Calculate the conditional frequencies for attitude, given affiliation

	Attitude			
Affiliation	favor	Indifferent	opposed	***Total***
Democrat	138 **48.4%**	83 **29.1%**	64 **22.5%**	285
Republican	64 **29.8%**	67 **31.2%**	84 **39.0%**	215
Total	202	150	148	500

Given a Democrat:

Favor: 138 out of 285 = 48.4%

Indifferent: 83 out of 285 = 29.1%

Opposed: 64 out of 285 = 22.5%

Given a Republican:

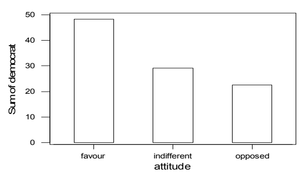
Favor: 64 out of 215 =29.8%

Indifferent: 67 out of 215 = 31.2%

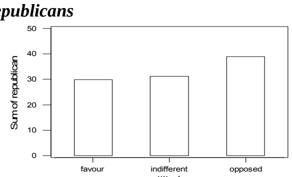
Opposed: 84 out of 215 = 39.0%

44

democrats



republicans



Calculate the conditional distribution of political affiliation given attitude:

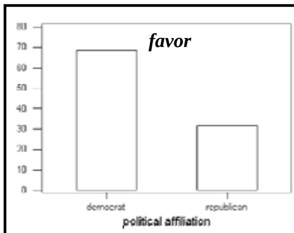
	Attitude			
Affiliation	favor	Indifferent	opposed	***Total***
Democrat	138 **68.3%**	83 **55.3%**	64 **43.2%**	285
Republican	64 **31.7%**	67 **44.7%**	84 **56.7%**	215
Total	202	150	148	500

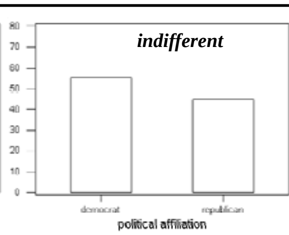
Given favor

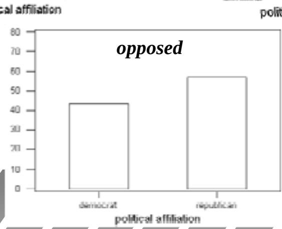
democrat: 138 out of 202 = 68.3%

republican: 64 out of 202 = 31.7%

46







47

Example

- A business school conducted a survey of companies in its state. A questionnaire was mailed to 200 small companies, 200 medium-sized companies, and 200 large companies. The rate of non-response is important in deciding how reliable survey results are.
- A 3×2 contingency table (but we use only percentages).
- Here are the data on response to this survey:

	Response	No response	***Total***
Small	125	75	200
Medium	81	119	200
Large	40	160	200

48

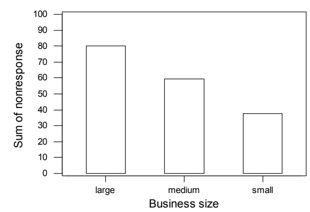
A. What was the overall percent of non-response?
***Answer:* (75 + 119 + 160) /600 = 0.59 → 59%**

B. Calculate the percent of no response for each type of business. Describe how non-response is related to size of business.
Answer:
small: 75 / 200 = 0.375 → **37.5%**
medium: 119 / 200 = 0.595 → **59.5%**
large: 160 / 200 = 0.80 → **80%**
The larger the business, the less likely it is to respond

	Response	No response	*Total*
Small	125	75	200
Medium	81	119	200
Large	40	160	200

49

C. Draw a bar graph to compare the non-response percents for the three size categories.
Answer:



50

D. Using the total number of responses as a base, compute the percent of responses that come from each of small, medium and large businesses
Answer:

	Response
Small	125 50.8%
Medium	81 32.9%
Large	40 16.3%

Total = **246**

	Response	No response	*Total*
Small	125	75	200
Medium	81	119	200
Large	40	160	200

51

E. In preparing an analysis of the survey results, do you think it would be reasonable to proceed as if the responses represented companies of each size equally?

Answer:
No. Over half of respondents were small businesses, while less than 17% of responses came from large businesses.

52