TOPIC 5

Exploring Categorical Data: Frequency Tables

- Marginal Frequencies for Two-Way Tables
- Conditional Relative Frequencies and Association

While many variables such as age, income, and years of education are quantitative or numerical in nature, others such as gender, race, brand preference, mode of transportation, and type of occupation are qualitative or categorical. Quantitative variables, too, are sometimes grouped into categorical classes.

MARGINAL FREQUENCIES FOR TWO-WAY TABLES

Qualitative data often encompass two categorical variables that may or may not have a dependent relationship. These data can be displayed in a *two-way contingency table.*

EXAMPLE 5.1

A 4-year study, reported in *The New York Times*, on men more than 70 years old analyzed blood cholesterol and noted how many men with different cholesterol levels suffered nonfatal or fatal heart attacks.

	Low cholesterol	Medium cholesterol	High cholesterol
Nonfatal heart attacks	29	17	18
Fatal heart attacks	19	20	9

Severity of heart attacks is the *row variable*, while cholesterol level is the *column variable.*

One method of analyzing these data involves first calculating the totals for each row and each column:

	Low cholesterol	Medium cholesterol	High cholesterol	Total
Nonfatal heart attacks	29	17	18	64
Fatal heart attacks	19	20	9	48
Total	48	37	27	112

These totals are placed in the right and bottom margins of the table and thus are called *marginal frequencies.*

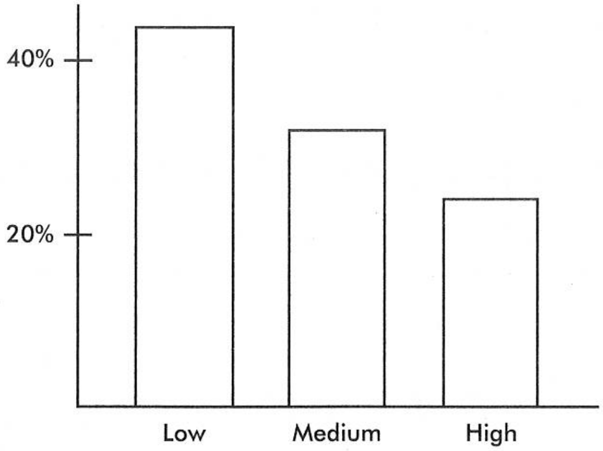
These marginal frequencies are often put in the form of proportions or percentages. The *marginal distribution* of the cholesterol level is

Low: $\frac{48}{112} = .429 = 42.9\%$

Medium: $\frac{37}{112} = .330 = 33.0\%$

High: $\frac{27}{112} = .241 = 24.1\%$.

This distribution can also be displayed in a bar graph as follows:



Cholesterol levels
(of elderly men who suffered heart attacks)

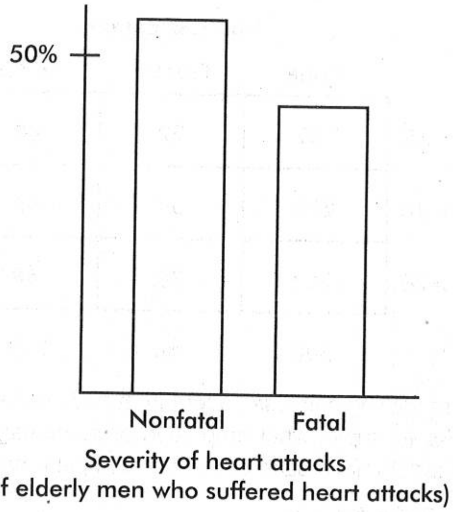
Similarly we can determine the marginal distribution for the severity of heart attacks:

Nonfatal: $\frac{64}{112} = .571 = 57.1\%$

Fatal: $\frac{48}{112} = .429 = 42.9\%$.

(continued)

The representative bar graph is



CONDITIONAL RELATIVE FREQUENCIES AND ASSOCIATION

The marginal distributions described and calculated above do not describe or measure the relationship between the two categorical variables. For this we must consider the information in the body of the table, not just the sums in the margins.

EXAMPLE 5.2

Is hair loss pattern related to body mass index? One study (*Journal of the American Medical Association*, February 24, 1993, page 1000) of 769 men showed the following numbers:

		Hair loss pattern		
		None	Frontal	Vertex
Body mass index	<25	137	22	40
	25–28	218	34	67
	>28	153	30	68

(continued)

The analysis first involves finding the row and column totals as we did before.

		Hair loss pattern			
		None	Frontal	Vertex	
Body mass index	<25	137	22	40	199
	25–28	218	34	67	319
	>28	153	30	68	251
		508	86	175	769

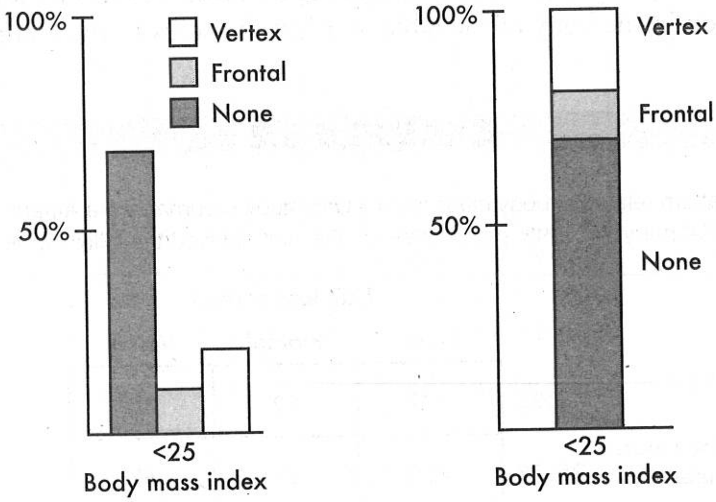
We are interested in predicting hair loss pattern from body mass index, and so we look at each row separately. For example, what proportion or percentage of the 199 men with a body mass index less than 25 have each of the hair loss patterns?

None: $\frac{137}{199} = .688 = 68.8\%$

Frontal: $\frac{22}{199} = .111 = 11.1\%$

Vertex: $\frac{40}{199} = .201 = 20.1\%.$

These *conditional relative frequencies* can be displayed either with groupings of bars or by a segmented bar chart where each segment has a length corresponding to its relative frequency:



Similarly, the conditional relative frequencies for the 319 men with a body mass index between 25 and 28 are

None: $\frac{218}{319} = .683 = 68.3\%$

Frontal: $\frac{34}{319} = .107 = 10.7\%$

Vertex: $\frac{67}{319} = .210 = 21.0\%.$

(continued)

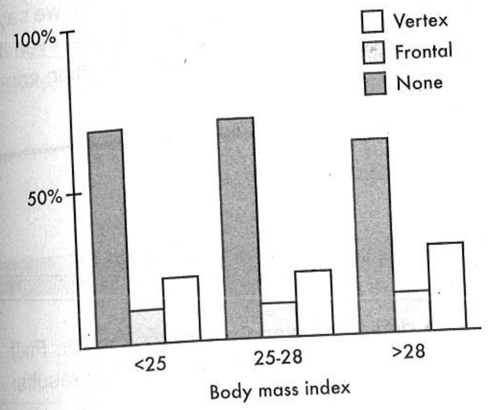
For the 251 men with a body mass index of more than 28 we have

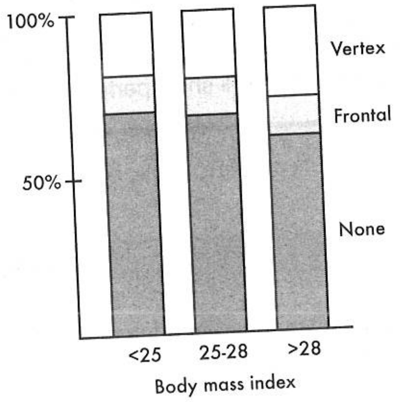
None: $\frac{153}{251} = .610 = 61.0\%$

Frontal: $\frac{30}{251} = .120 = 12.0\%$

Vertex: $\frac{68}{251} = .271 = 27.1\%.$

Both of the following bar charts give good visual pictures:





Segmented bar charts indicate a slight relationship between higher vertex pattern baldness and a body mass index of more than 28.

EXAMPLE 5.3

A study was made to compare year in high school with preference for vanilla or chocolate ice cream with the following results:

	Vanilla	Chocolate
Freshman	20	10
Sophomore	24	12
Junior	18	9
Senior	22	11

(continued)

What are the conditional relative frequencies for each class?

Freshmen:	$\frac{20}{30}$ = .667 prefer vanilla and $\frac{10}{30}$ = .333 prefer chocolate.
Sophomores:	$\frac{24}{36}$ = .667 prefer vanilla and $\frac{12}{36}$ = .333 prefer chocolate.
Juniors:	$\frac{18}{27}$ = .667 prefer vanilla and $\frac{9}{27}$ = .333 prefer chocolate.
Seniors:	$\frac{22}{33}$ = .667 prefer vanilla and $\frac{11}{33}$ = .333 prefer chocolate.

In such a case, where all the conditional relative frequency distributions are identical, we say that the two variables show *perfect independence.* (However, it should be noted that even if the two variables are completely independent, the chance is very slim that a resulting contingency table will show perfect independence.)

EXAMPLE 5.4

Suppose you need heart surgery and are trying to decide between two surgeons, Dr. Fixit and Dr. Patch. You find out that each operated 250 times last year with the following results:

	Dr. F	Dr. P
Died	60	50
Survived	190	200

Whom should you go to? Among Dr. Fixit's 250 patients 190 survived, for a survival rate of $\frac{190}{250}$ = .76 or 76%, while among Dr. Patch's 250 patients 200 survived, for a survival rate of $\frac{200}{250}$ = .80 or 80%. Your choice seems clear.

However, everything may not be so clear-cut. Suppose that on further investigation you determine that the surgeons operated on patients who were in either good or poor condition with the following results:

Good condition

	Dr. F	Dr. P
Died	8	17
Survived	60	120

Poor condition

	Dr. F	Dr. P
Died	52	33
Survived	130	80

Note that adding corresponding boxes from these two tables gives the original table above.

How do the surgeons compare when operating on patients in good health? Dr. Fixit's 68 patients in good condition have a survival rate of $\frac{60}{68}$ = .882 or 88.2%, while Dr. Patch's 137 patients in good condition have a survival rate of $\frac{120}{137}$ = .876 or 87.6%. Similarly, we note that Dr. Fixit's 182 patients in poor condition have a survival rate of $\frac{130}{182}$ = .714 or 71.4%, while Dr. Patch's 113 patients in poor condition have a survival rate of $\frac{80}{113}$ = .708 or 70.8%.

Thus Dr. Fixit does better with patients in good condition (88.2% versus Dr. Patch's 87.6%) and also does better with patients in poor condition (71.4% versus Dr. P's 70.8%). However, Dr. Fixit has a lower overall patient survival rate (76% versus Dr. Patch's 80%)! How can this be?

This problem is an example *of Simpson's paradox*, where a comparison can be reversed when more than one group is combined to form a single group. The effect of another variable, sometimes called a *lurking variable*, is masked when the groups are combined. In this particular example, closer scrutiny reveals that Dr. Fixit operates on many more patients in poor condition than Dr. Patch, and these patients in poor condition are precisely the ones with lower survival rates. Thus even though Dr. Fixit does better with all patients, his overall rating is lower. Our original table hid the effect of the lurking variable related to the condition of the patients.

Summary

- Two-way contingency tables are useful in showing relationships between two categorical variables.
- The row and column totals lead to calculations of the marginal distributions.
- Focusing on single rows or columns leads to calculations of conditional distributions.
- Segmented bar charts are a useful visual tool to show conditional distributions.
- Simpson's paradox occurs when the results from a combined grouping seem to contradict the results from the individual groups.