

Chapter 4 § 3

Relations in Categorical Data

Definitions:

Two-way table – (of counts) describe the relationship between two categorical variables.

Row variable – values that are labeled in the rows of the table.

Column variable - values that are labeled in the columns of the table.

Marginal distribution – values that appear at the right and bottom margins of the two-way table.

Conditional distribution – the value of the row variable for one specific value of the column variable.

Simpson's Paradox – a comparison between two variables that holds for each individual value of third variable can be changed or even reversed when the data for all values of the third variable are combined.

Marginal Distributions

The distribution of a categorical variable just says how often each outcome occurred.

Education	Age Group			Total
	25 to 34	35 to 54	55 and over	
Did not complete high school	5325	9152	16035	30512
Completed high school	14061	24070	18320	56451
1 to 3 years of college	11659	19926	9662	41247
4 or more years of college	10342	19878	8005	38225
	41388	73028	52022	166438

Distributions of education alone are often called marginal distribution because they appear at the right and bottom margins of the two-way table.

Percents are often easier to grasp than counts.

By taking the total count of the complete high school, 30,512, and dividing it by the grand total of 166,438, we will get the percentage $\approx 18.3\%$.

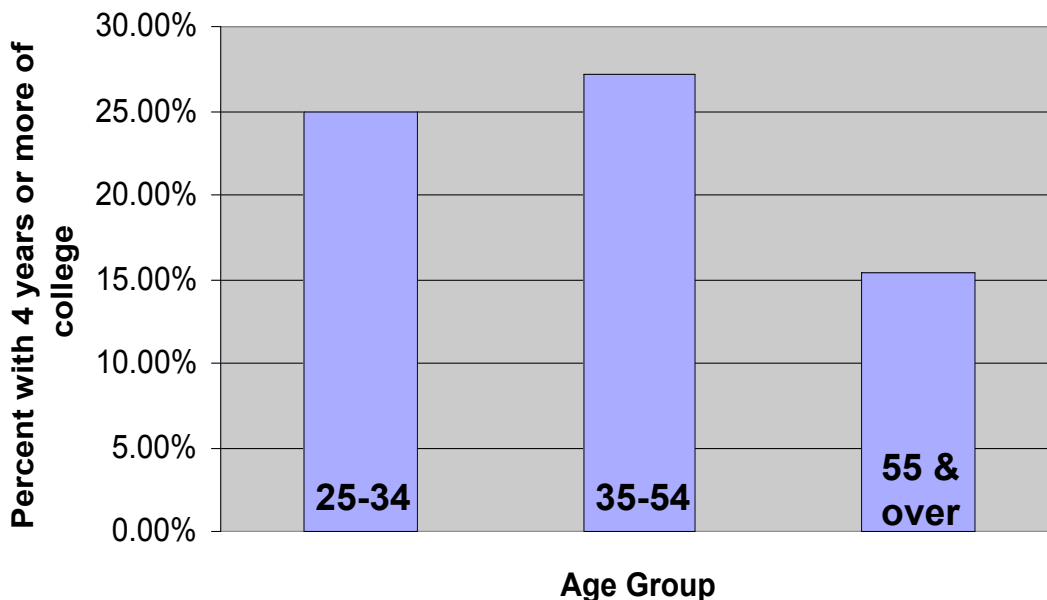
Education	Did not finish high school	completed high school	1-3 years of college	4 or more years of college
Percent	18.3	33.9	24.8	23

Describing relationships

The marginal distributions of age and of education separately do not tell us how the two variables are related. That information is in the body of the table.

To describe relationships among categorical variables, calculate appropriate percents from the counts given. We use percents because counts are often hard to compare.

Education	Age Group		
	25 to 34	35 to 54	55 and over
Did not complete high school	12.87%	12.53%	30.82%
Completed high school	33.97%	32.96%	35.22%
1 to 3 years of college	28.17%	27.29%	18.57%
4 or more years of college	24.99%	27.22%	15.39%



The bar chart compares the sizes of different items. Notice we draw the bars with space between them.

Simpson's Paradox

As in the case with quantitative variables, the effects of lurking variables can change or even reverse relationships between two categorical variables.

	Hospital A	Hospital B
Died	63	16
Survived	2037	784
Total	2100	800

The evidence seems clear. Hospital A loses 3% (63/2100) of its surgery patients, and Hospital B loses only 2% (16/800). It seems that you should choose Hospital B if you need surgery.

	Good Condition		Poor Condition	
	Hospital A	Hospital B	Hospital A	Hospital B
Died	6	8	57	8
Survived	594	592	1443	192
Total	600	600	1500	200

The patient's condition is a lurking variable when we compare the death rates at the two hospitals. Hospital A beats Hospital B for patients in good condition: only 1% (6/600) died in Hospital A, compared with 1.3% (8/600) in Hospital B. And Hospital A wins again for patients in poor condition, losing 3.8% (57/1500) to Hospital B's 4% (8/200). When we ignore the lurking variable, Hospital B seems safer, even though Hospital A does better for both classes of patients.

The lurking variables in Simpson's paradox are categorical. That is, they break the individuals into groups, as when surgery patients are classified as "good condition" or "poor condition." Simpson's paradox is just an extreme form of the fact that observed associations can be misleading when there are lurking variables.