Conditional Relative Frequencies and Association

The table of data in the previous examples provides more information than marginal frequencies. The data inside the table provides information about the relationship, or association of the two categorical variables. These are called conditional relative frequencies.

EXAMPLE:
Use the data from the previous examples to calculate the percentages of females at each of the four colleges and the percentages of males at each of the four colleges. Use the data to calculate the gender percentages of students at each college.

	Female	*Male*	*TOTAL*
College A	3832	4228	8060
College B	6765	5590	12355
College C	2889	3388	6277
College D	5580	5612	11192
TOTAL	19066	18818	37884

To calculate these conditional relative frequencies, computation will take place in only one column or row from the table. To calculate the percentage of all the female students that attend College A, divide the number of female students at College A by the total number of female students. Remember, these quantities are relative to an individual row or column total, not the grand total.

For the females:

$$\text{Relative frequency of female students attending College A} = \frac{3832}{19066} = 20.1\%.$$

$$\text{Relative frequency of female students attending College B} = \frac{6765}{19066} = 35.5\%.$$

$$\text{Relative frequency of female students attending College C} = \frac{2889}{19066} = 15.2\%.$$

$$\text{Relative frequency of female students attending College D} = \frac{5580}{19066} = 29.3\%.$$

For the males:

$$\text{Relative frequency of male students attending College A} = \frac{4228}{18818} = 22.5\%.$$

$$\text{Relative frequency of male students attending College B} = \frac{5590}{18818} = 29.7\%.$$

$$\text{Relative frequency of male students attending College C} = \frac{3388}{18818} = 18.0\%.$$

$$\text{Relative frequency of male students attending College D} = \frac{5612}{18818} = 29.8\%.$$

To calculate the percentage of students at College B who are female, divide the number of female students at College B by the total number of students at College B.

For students at College A:

$$\text{Relative frequency of College A students who are female} = \frac{3832}{8060} = 47.5\%.$$

$$\text{Relative frequency of College A students who are male} = \frac{4228}{8060} = 52.5\%.$$

For students at College B:

$$\text{Relative frequency of College B students who are female} = \frac{6765}{12355} = 54.8\%.$$

$$\text{Relative frequency of College B students who are male} = \frac{5590}{12355} = 45.2\%.$$

For students at College C:

$$\text{Relative frequency of College C students who are female} = \frac{2889}{6277} = 46.0\%.$$

$$\text{Relative frequency of College C students who are male} = \frac{3388}{6277} = 54.0\%.$$

For students at College D:

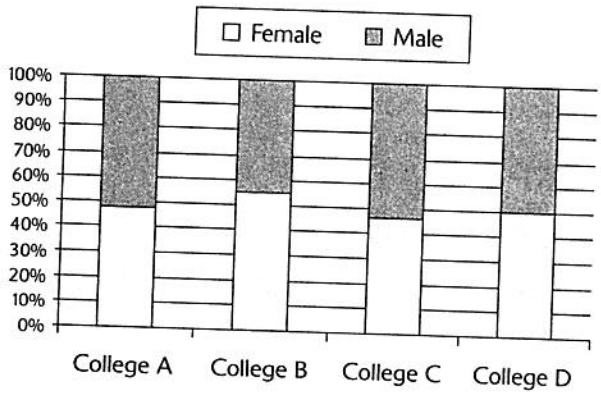
$$\text{Relative frequency of College D students who are female} = \frac{5580}{11192} = 49.9\%.$$

$$\text{Relative frequency of College D students who are male} = \frac{5612}{11192} = 50.1\%.$$

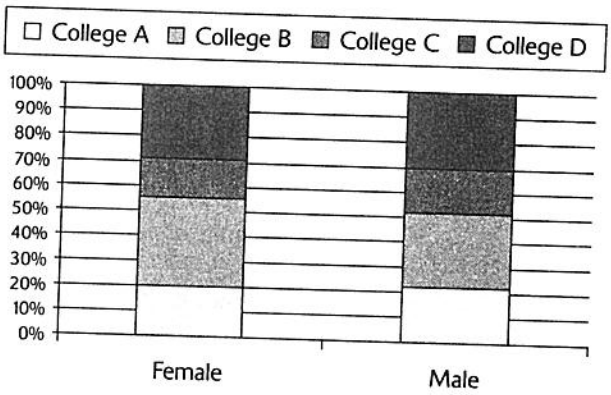
Different types of bar charts can illustrate these percentages. Stacked bar charts and clustered bar charts show these conditional relative frequencies.

EXAMPLE:
Use the results from the previous example and construct stacked bar charts and clustered bar charts by college and gender.

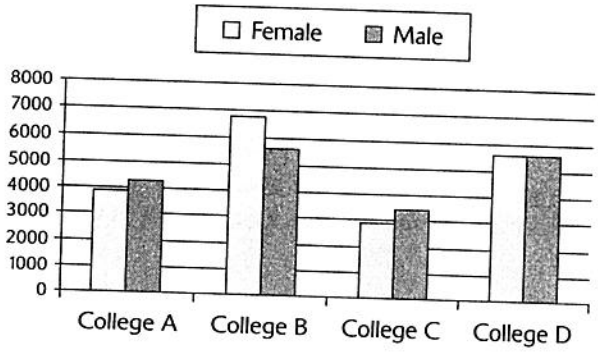
Stacked bar chart by College



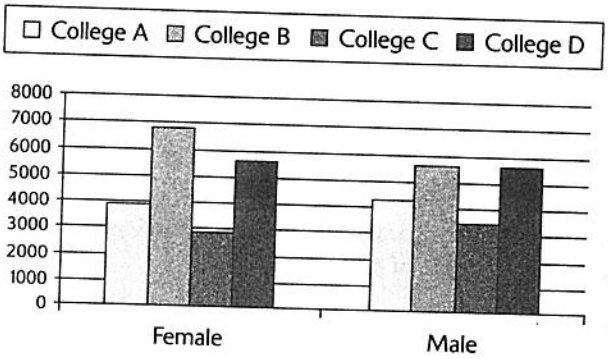
Stacked bar chart by Gender



Clustered column chart by College



Clustered column chart by Gender



Although two-way contingency tables are useful, be careful of hidden variables that can change, or even reverse categorical percentages. This phenomenon is known as **Simpson's Paradox** and is illustrated in the next example.

EXAMPLE:
You are trying to decide which of two vegetables to serve at a banquet. A random sample of 600 people is divided into two groups. Three hundred fourteen people were asked whether they like broccoli and 286 people were asked whether they like spinach. Use the table of responses here to determine which vegetable is more popular.

	Yes	*No*
Broccoli	148	166
Spinach	153	133

Calculate the percentage of those who said they like spinach and calculate the percentage of those who said they like broccoli. Should you serve the one that received the higher percentage?

Calculate the percentage of those who like broccoli and spinach by dividing the number of "yes" responses by the total.

$$\text{Percentage who like broccoli} = \frac{148}{148+166} = \frac{148}{314} = 0.471 = 47.1\%.$$

$$\text{Percentage who like spinach} = \frac{153}{153+133} = \frac{153}{286} = 0.535 = 53.5\%.$$

From these results, you might conclude that because a substantially higher percentage of people like spinach, you should serve spinach at your banquet.

You would be making a mistake if you don't consider any *hidden variables* that could change, or reverse your conclusion. If the original sample of 600 people was made up of children and adults and separate data was collected for each, evaluating each subset of data could be useful in helping make the correct decision.

The following table of responses is broken down into two subsets: Children and Adults. Note that corresponding data adds up to the original total.

Children		
	Yes	***No***
Broccoli	110	155
Spinach	43	88

Adults		
	Yes	***No***
Broccoli	38	11
Spinach	110	45

Now, separately for children and adults, calculate the percentage of those who like broccoli and spinach by dividing the number of "yes" responses by the total.

$$\text{Percentage of CHILDREN who like broccoli} = \frac{110}{110+155} = \frac{110}{265} = 0.415 = 41.5\%.$$

$$\text{Percentage of CHILDREN who like spinach} = \frac{43}{43+88} = \frac{43}{131} = 0.328 = 32.8\%.$$

$$\text{Percentage of ADULTS who like broccoli} = \frac{38}{38+11} = \frac{38}{49} = 0.776 = 77.6\%.$$

$$\text{Percentage of ADULTS who like spinach} = \frac{110}{110+45} = \frac{110}{155} = 0.710 = 71.0\%.$$

The results are interesting. When you consider the adults and children separately, each group favored broccoli by about 8%. But, if you consider both subgroups together as one group, spinach is favored by about 6%. This is **Simpson's Paradox** and results from the unbalanced nature of the subgroup responses. Overall, the yes/no ratio was about the same. But when the data is separated, adults have a much higher approval rating of both vegetables and children have a much lower approval rating of both vegetables. Combining all the information into one chart *hid* the effect of age on vegetable preference.

Statistical results, such as these, can be used to mislead consumers by not showing enough detail.