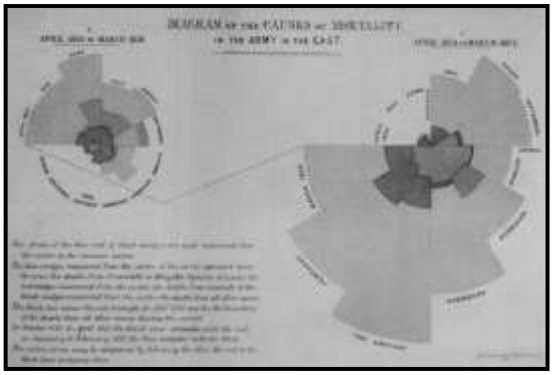
Chapter 3 Summary
Displaying and Describing Categorical Data

What did you learn?
We can summarize categorical data by counting the number of cases in each category, sometimes expressed as percents. Data can be displayed in a bar or circle graph. Categorical relationships can be viewed in a contingency table.

- We examine the marginal distribution of each variable
- We also look at conditional distribution of a variable within each category of the other variable
- We can display these conditional and marginal distributions in bar or circle graphs
- If the conditional distributions of one variable are about the same for every category of the other, the variables are independent

Three Rules of Data Analysis

1. Make a picture – organizes patterns and helps clear thinking about relationships
2. Make a picture – show important features of data and patterns in the data
3. Make a picture – tell others about the data



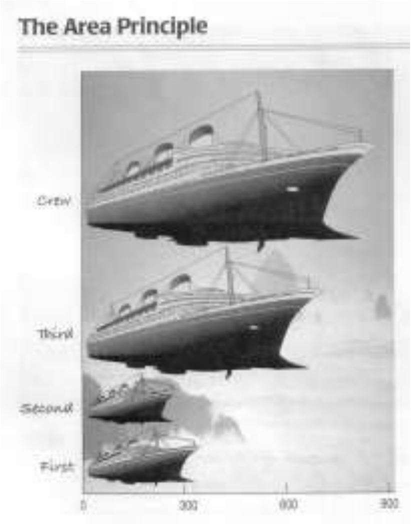
Florence Nightingale used data displays to show more soldiers in the Crimean War (1856) died of illness and infection than of battle wounds

Titanic data

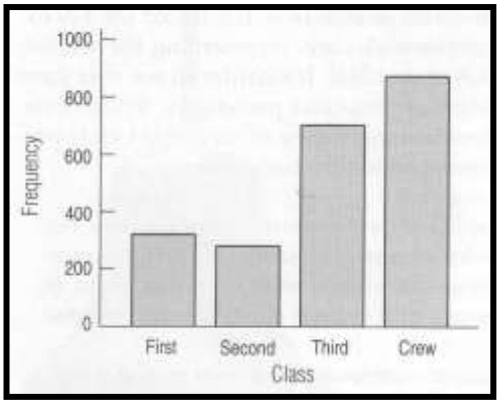
Class	Passenger Count	Percent
First	325	14.77%
Second	285	12.95%
Third	706	32.08%
Crew	885	40.21%

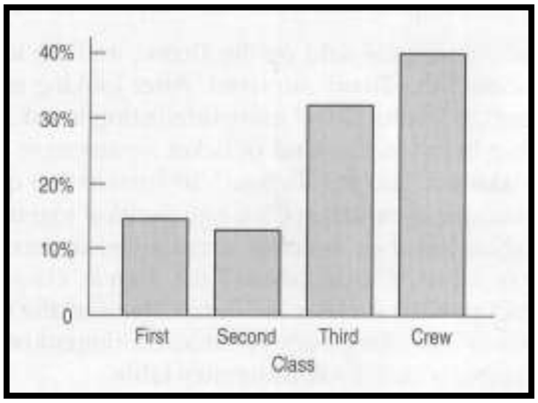
Frequency table	A frequency table lists the categories in a categorical variable and gives the count of observations for each category.
Proportion	A comparison of numbers. For our use, dividing the counts by the total number of cases.
Percentages	Multiplying proportions by 100 to express as percentages, or ratios compared to a total of 100.B41
Relative frequency table	A frequency table lists the categories in a categorical variable and gives the percentage of observations for each category.
Distribution	The distribution of a variable gives the possible values of the variable and the relative frequency of each value.

Area principle	In a statistical display, each data value should be represented by the same amount of area.

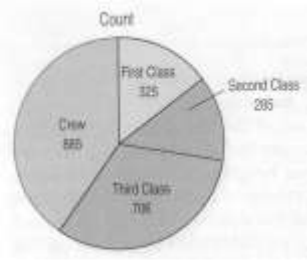


Bar chart	Bar charts show a bar representing the count of each category in a categorical variable.
Relative frequency bar chart	A bar chart that shows the proportion of people in each category rather than counts.





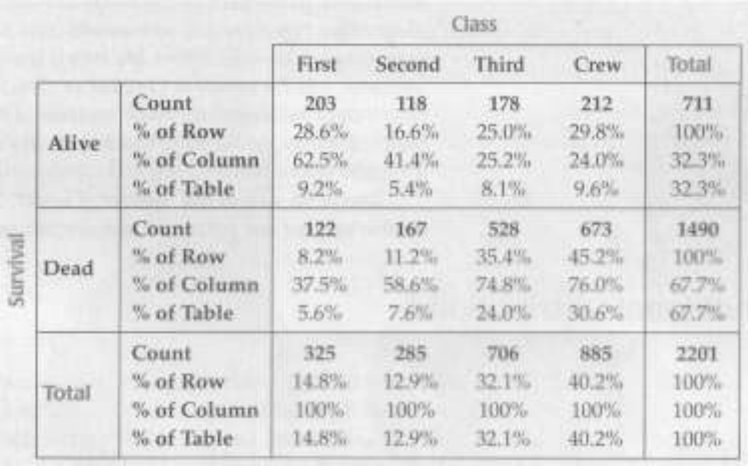
Pie chart	Pie charts show how a "whole" divides into categories by showing a wedge of a circle whose area corresponds to the proportion in each category.



Categorical data condition	Before making a bar or pie chart, be sure that the categorical data is in counts or percentages of individuals.
Contingency table	A contingency table displays counts and, sometimes, percentages of individuals falling into named categories on two or more variables. The table categorizes the individuals on all variables at once, to reveal possible patterns in one variable that may be contingent on the category of the other.

Chapter 3 Summary Continued

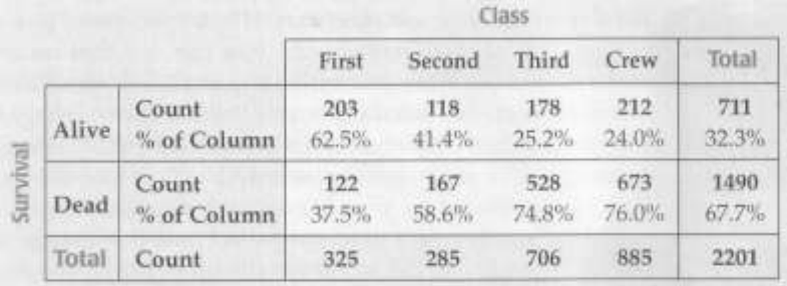
Marginal distribution	In a contingency table, the distribution of either variable alone is called the marginal distribution. The counts or percentages are the totals found in the margins (last row or column) of the table.



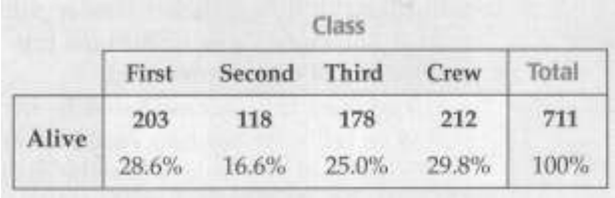
			Class				
			First	Second	Third	Crew	Total
Survival	Alive	Count	203	118	178	212	711
		% of Row	28.6%	16.6%	25.0%	29.8%	100%
		% of Column	62.5%	41.4%	25.2%	24.0%	32.3%
		% of Table	9.2%	5.4%	8.1%	9.6%	32.3%
	Dead	Count	122	167	528	673	1490
		% of Row	8.2%	11.2%	35.4%	45.2%	100%
		% of Column	37.5%	58.6%	74.8%	76.0%	67.7%
		% of Table	5.6%	7.6%	24.0%	30.6%	67.7%
	Total	Count	325	285	706	885	2201
		% of Row	14.8%	12.9%	32.1%	40.2%	100%
		% of Column	100%	100%	100%	100%	100%
		% of Table	14.8%	12.9%	32.1%	40.2%	100%

Conditional distribution	The distribution of a variable restricting the *Who* to consider only a smaller group of individuals is called a conditional distribution.

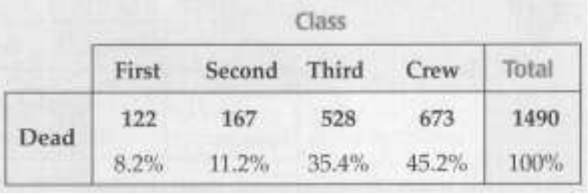
Did the chance of surviving the *Titanic* depend on ticket class?



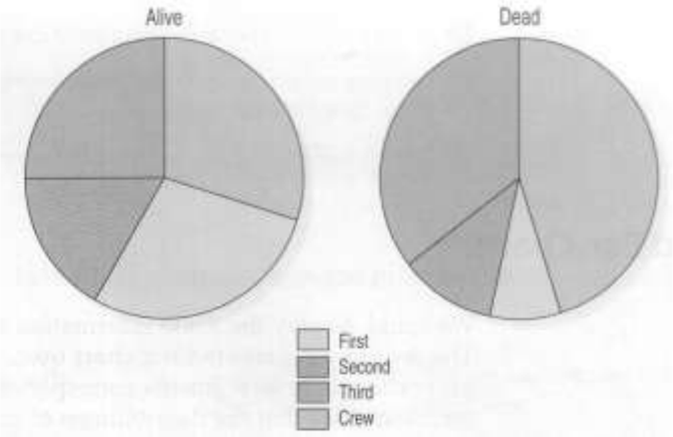
			Class				
			First	Second	Third	Crew	Total
Survival	Alive	Count	203	118	178	212	711
		% of Column	62.5%	41.4%	25.2%	24.0%	32.3%
	Dead	Count	122	167	528	673	1490
		% of Column	37.5%	58.6%	74.8%	76.0%	67.7%
	Total	Count	325	285	706	885	2201



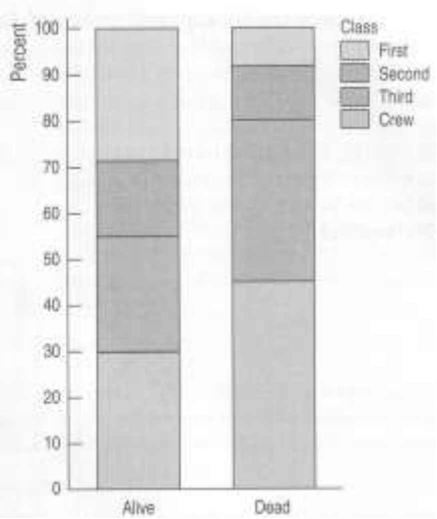
	Class				
	First	Second	Third	Crew	Total
Alive	203	118	178	212	711
	28.6%	16.6%	25.0%	29.8%	100%



	Class				
	First	Second	Third	Crew	Total
Dead	122	167	528	673	1490
	8.2%	11.2%	35.4%	45.2%	100%



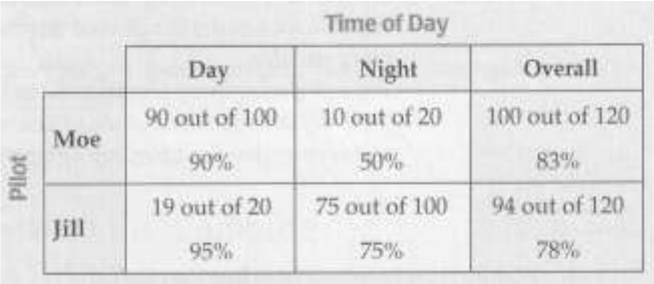
Independence	Variables are said to be independent if the conditional distribution of one variable is the same for each category of the other.
Segmented bar chart	A bar chart where each bar is treated as the "whole" and divides the bar proportionally into segments corresponding to the percentage in each group.



What can go wrong?

- Don’t violate the area principle.
- Keep it honest.
- Don’t confuse similar-sounding percentages.
- Don’t forget to look at the variables separately, too.
- Be sure to use enough individuals.
- Don’t overstate your case.
- Don’t use unfair or silly averages.

Simpson's paradox	When averages are taken across different groups, they can appear to contradict the overall averages. This is known as "Simpson's paradox."



		Time of Day		
		Day	Night	Overall
Pilot	Moe	90 out of 100 90%	10 out of 20 50%	100 out of 120 83%
	Jill	19 out of 20 95%	75 out of 100 75%	94 out of 120 78%

It’s always better to compare percentages or other averages within each level of the other variable.