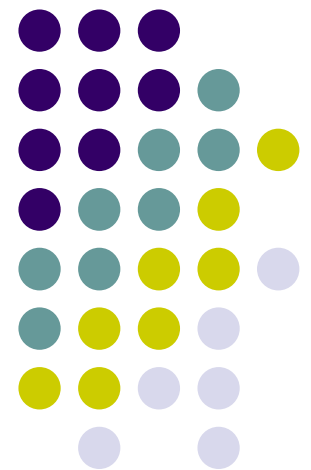


Econ 3790: Business and Economics Statistics

Instructor: Yogesh Uppal
Email: yuppal@ysu.edu



Chapter 2

Summarizing Qualitative Data



- Frequency distribution
 - Relative frequency distribution
 - Bar graph
 - Pie chart
-
- Objective is to provide insights about the data that cannot be quickly obtained by looking at the original data



Distribution Tables

- Frequency distribution is a tabular summary of the data showing the frequency (or number) of items in each of several non-overlapping classes
- Relative frequency distribution looks the same, but contains proportion of items in each class



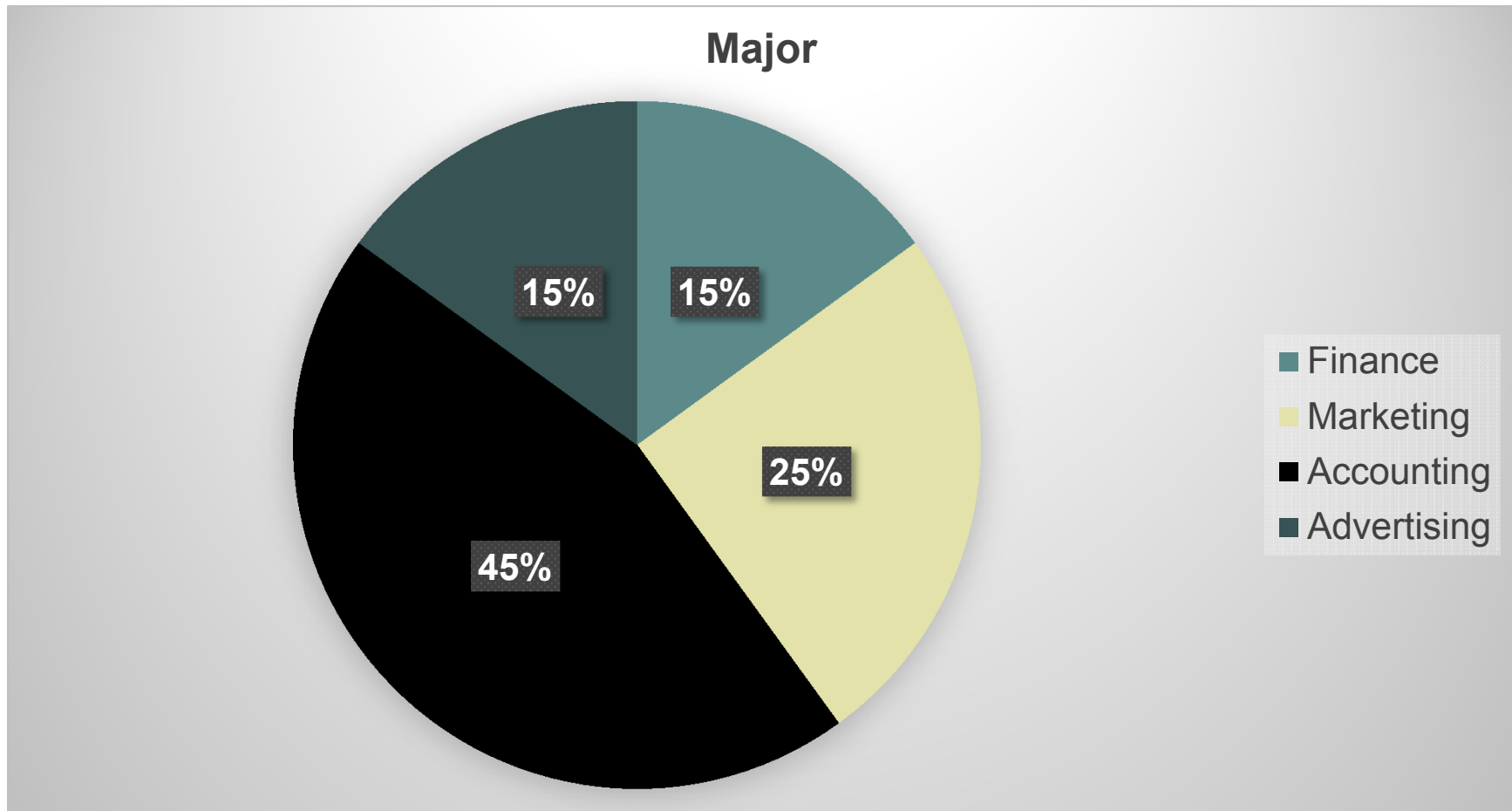
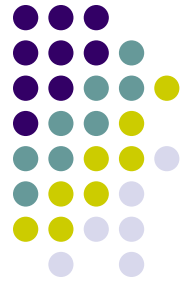
Example 1: What's your major?

Major	Frequency
Finance	6
Marketing	10
Accounting	18
Advertising	6

Example 1: What's your major?



Example 1: What's your major?



Summarizing Quantitative Data



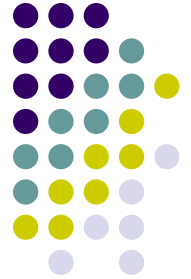
- Frequency Distribution
- Relative Frequency Distribution
- Dot Plot
- Histogram
- Cumulative Distributions

Example 1: Go Penguins



Month	Opponents	Rushing TDs
Sep	SLIPPERY ROCK	4
Sep	NORTHEASTERN	4
Sep	at Liberty	1
Sep	at Pittsburgh	0
Oct	ILLINOIS STATE	1
Oct	at Indiana State	4
Oct	WESTERN ILLINOIS	2
Oct	MISSOURI STATE	4
Oct	at Northern Iowa	0
Nov	at Southern Illinois	0
Nov	WESTERN KENTUCKY	3

Example 1: Go Penguins



Rushing TDs	Frequency	Relative Frequency
0	3	0.27
1	2	0.18
2	1	0.09
3	1	0.09
4	4	0.37
	Total=11	Total=1.00

Example 2: Rental Market in Youngstown



- Suppose you were moving to Youngstown, and you wanted to get an idea of what the rental market for an apartment (having more than 1 room) is like
- I have the following sample of rental prices

Example: Rental Market in Youngstown



- Sample of 28 rental listings from craigslist:

330	470	540	595
350	470	550	599
390	480	570	610
410	490	580	650
410	500	585	660
430	520	590	700
440	520	595	740

Frequency Distribution

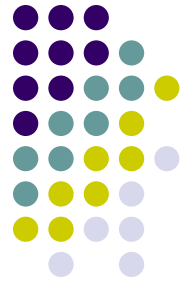


- To deal with large datasets
 - Divide data in different classes
 - Select a width for the classes

Frequency Distribution (Cont'd)



- Guidelines for Selecting Number of Classes
 - Use between 5 and 20 classes
 - Datasets with a larger number of elements usually require a larger number of classes
 - Smaller datasets usually require fewer classes



Frequency Distribution

- Guidelines for Selecting Width of Classes
 - Use classes of equal width
 - Approximate Class Width =

$$\frac{\text{Largest Data Value} - \text{Smallest Data Value}}{\text{Number of Classes}}$$



Frequency Distribution

- For our rental data, if we choose six classes:
 - Class Width = $(750-330)/6 = 70$

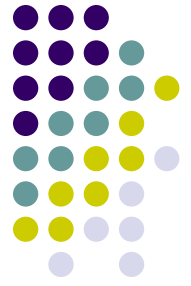
Rent (\$)	Frequency
330-399	3
400-469	4
470-539	7
540-609	9
610-679	3
680-749	2
Total	28



Relative Frequency

- To calculate relative frequency, just divide the class frequency by the total Frequency

Rent (\$)	Frequency	Relative Frequency
330-399	3	0.11
400-469	4	0.14
470-539	7	0.25
540-609	9	0.32
610-679	3	0.11
680-749	2	0.07
Total	28	1.00



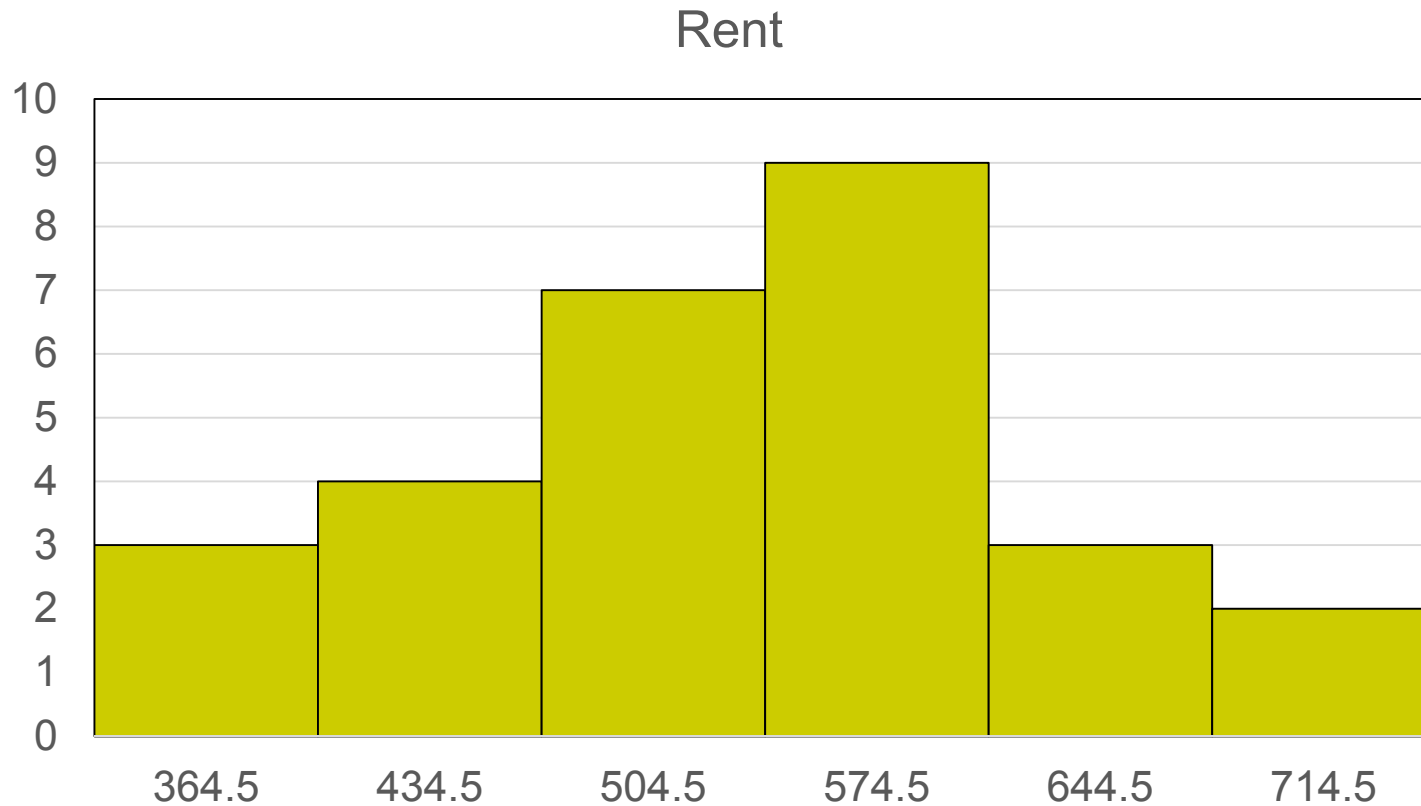
Relative Frequency

- Insights gained from Relative Frequency Distribution:
 - 32% of rents are between \$539 and \$609
 - Only 7% of rents are above \$680

Histogram



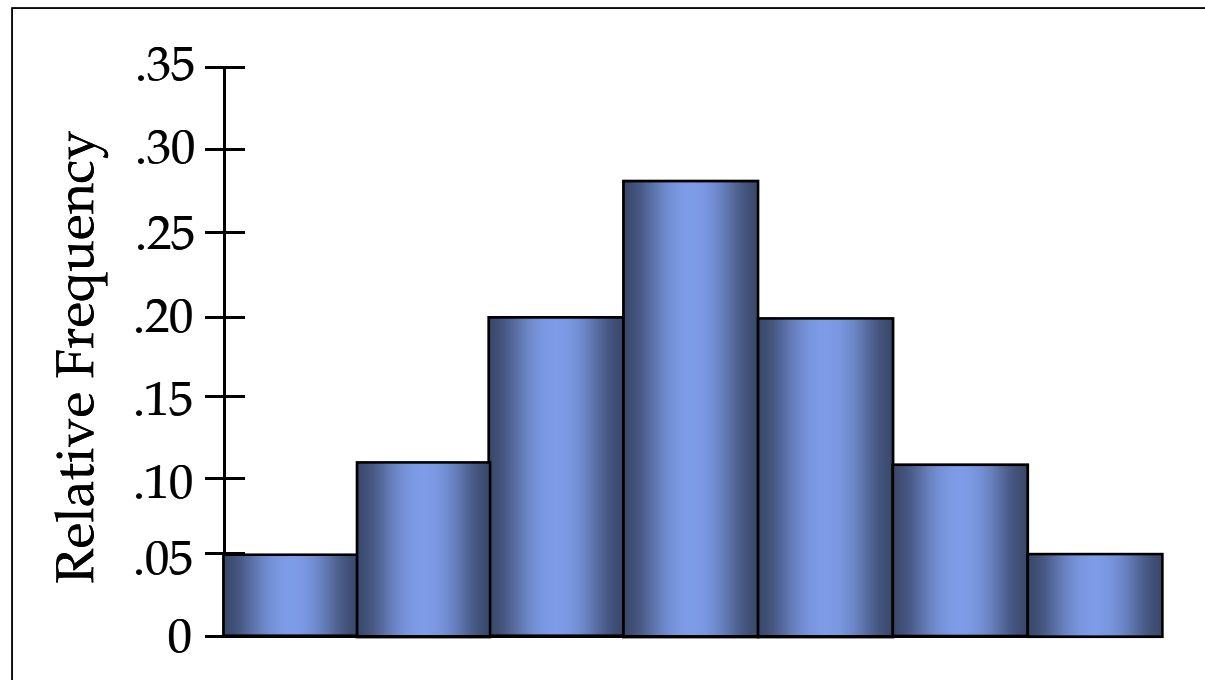
Histogram of Youngstown Rental Prices



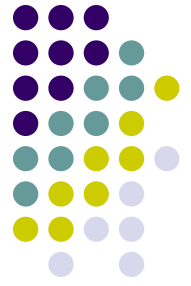


Describing a Histogram

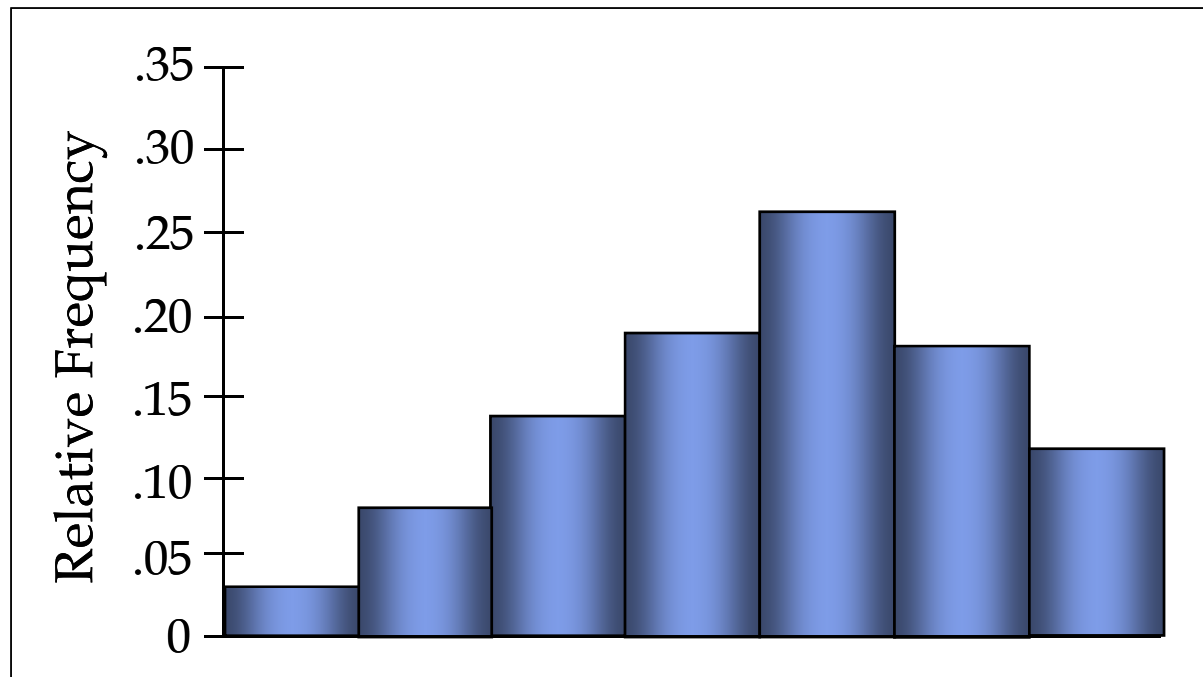
- Symmetric
 - Left tail is the mirror image of the right tail
 - Example: heights and weights of people

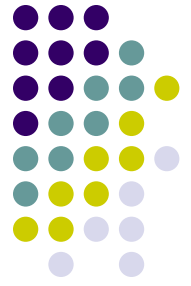


Describing a Histogram



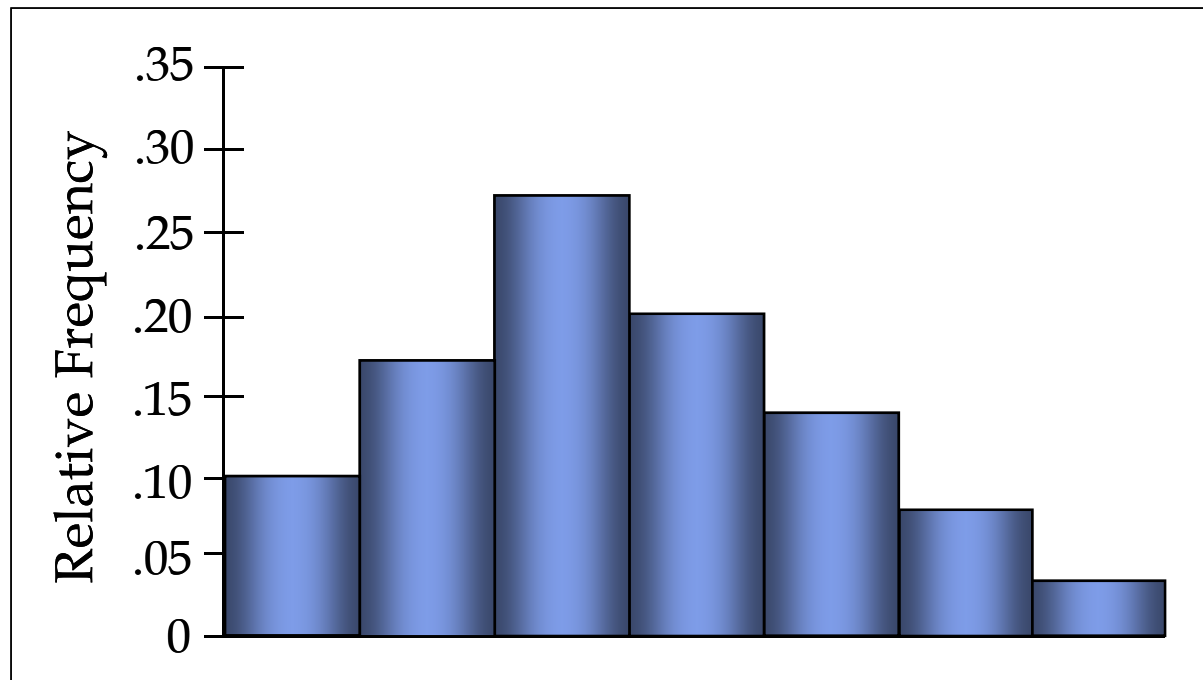
- Moderately Left or Negatively Skewed
 - A longer tail to the left
 - Example: exam scores





Describing a Histogram

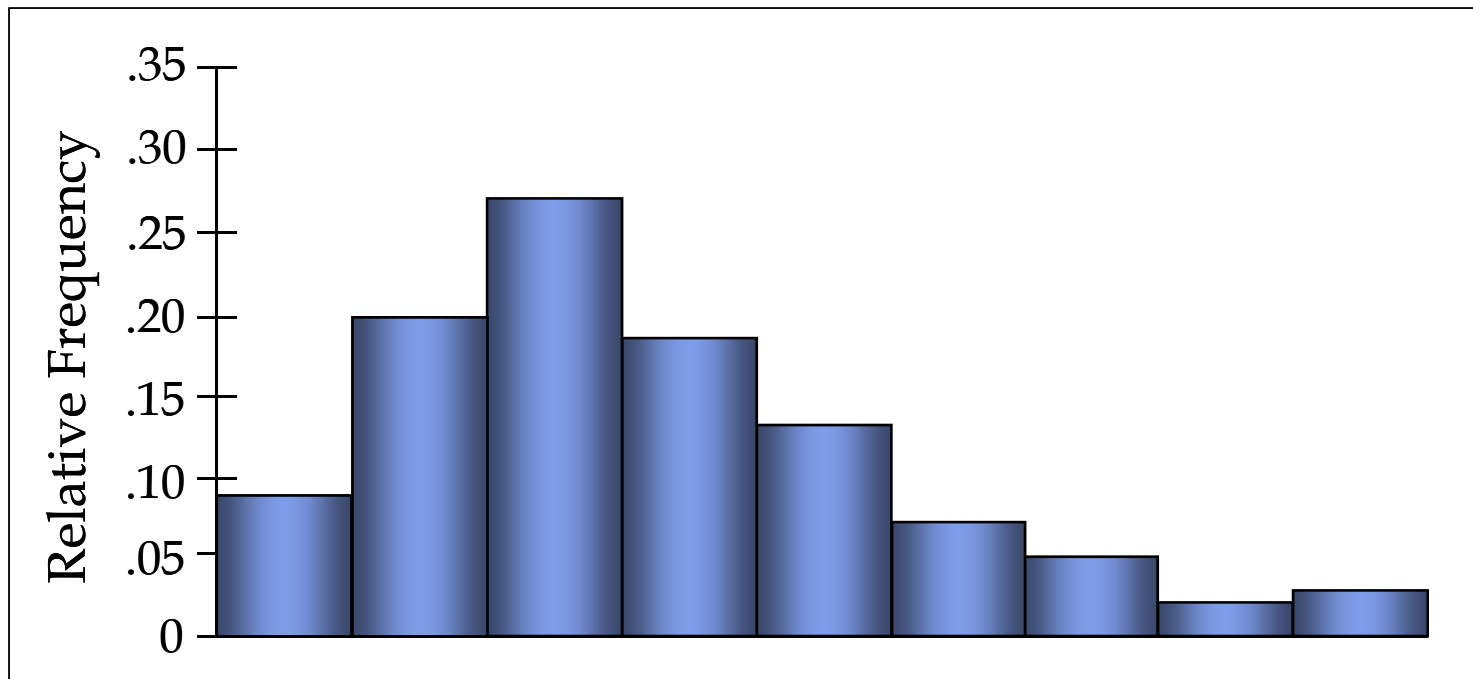
- Moderately Right or Positively Skewed
 - A longer tail to the right
 - Example: hourly wages

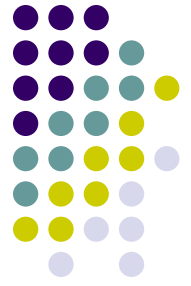




Describing a Histogram

- Highly Right or Positively Skewed
 - A very long tail to the right
 - Example: executive salaries





Cumulative Distributions

- Cumulative frequency distribution:
 - shows the number of items with values less than or equal to a particular value (or the upper limit of each class when we divide the data in classes)
- Cumulative relative frequency distribution:
 - shows the proportion of items with values less than or equal to a particular value (or the upper limit of each class when we divide the data in classes)
 - Usually only used with quantitative data!

Example 1: Go Penguins (Cont'd)



Rushing TDs	Frequency	Relative Frequency	Cumulative Fre.	Cumulative Relative Fre.
0	3	0.27	3	0.27
1	2	0.18	5	0.45
2	1	0.09	6	0.54
3	1	0.09	7	0.63
4	4	0.37	11	1
Total	11	1		



Cumulative Distributions

- Youngstown Rental Prices

Rent (\$)	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
330-399	3	0.11	3	0.11
400-469	4	0.14	7	0.25
470-539	7	0.25	14	0.50
540-609	9	0.32	23	0.82
610-679	3	0.11	26	0.93
680-749	2	0.07	28	1.00
Total	28	1.00		

Crosstabulations and Scatter Diagrams



- So far, we have focused on methods that are used to summarize data for one variables at a time
- Often, we are really interested in the relationship between two variables
- Crosstabs and scatter diagrams are two methods for summarizing data for two (or more) variables simultaneously

Crosstabs



- A crosstab is a tabular summary of data for two variables
- Crosstabs can be used with any combination of qualitative and quantitative variables
- The left and top margins define the classes for the two variables

Example: Data on MLB Teams



- Data from the 2002 Major League Baseball season
- Two variables:
 - Number of wins
 - Average stadium attendance

Crosstab



Frequency distribution for the wins variable

Wins	<u>Average Stadium Attendance</u>			Total
	10000-21000	21001-32000	32001-43000	
55-64	3	1	0	4
65-74	0	4	3	7
75-84	5	2	1	8
85-94	0	2	2	4
95-104	0	3	4	7
Total	8	12	10	30

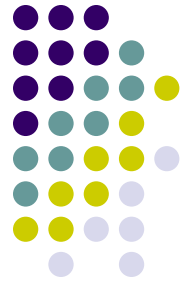
Frequency distribution for the attendance variable

Crosstabs: Row or Column Percentages



- Converting the entries in the table into row percentages or column percentages can provide additional insight about the relationship between the two variables

Crosstab: Row Percentages



Wins	<u>Average Stadium Attendance</u>			Total
	10000-21000	21001-32000	32001-43000	
55-64	0.75	0.25	0.00	1.00
65-74	0.00	0.57	0.43	1.00
75-84	0.63	0.25	0.13	1.00
85-94	0.00	0.50	0.50	1.00
95-104	0.00	0.43	0.57	1.00
Total	0.27	0.40	0.33	1.00

Crosstab: Column Percentages



Wins	<u>Average Stadium Attendance</u>			Total
	10000-21000	21001-32000	32001-43000	
55-64	0.38	0.08	0.00	0.13
65-74	0.00	0.33	0.30	0.23
75-84	0.63	0.17	0.10	0.27
85-94	0.00	0.17	0.20	0.13
95-104	0.00	0.25	0.40	0.23
Total	1.00	1.00	1.00	1.00

Crosstab: Simpson's Paradox



- Data in two or more crosstabulations are often aggregated to produce a summary crosstab
- We must be careful in drawing conclusions about the relationship between the two variables in the aggregated crosstab
- Simpsons' Paradox:
 - In some cases, the conclusions based upon an aggregated crosstab can be completely reversed if we look at the unaggregated data



Crosstab: Simpsons Paradox

Frequency distribution
for the wins variable

Wins	Average Stadium Attendance			Total
	10000-21000	21001-32000	32001-43000	
55-64	0	1	3	4
65-74	0	2	2	4
75-84	4	0	3	7
85-94	2	3	2	7
95-104	2	4	2	8
Total	8	10	12	30

Frequency distribution
for the attendance variable

Scatter Diagram and Trendline

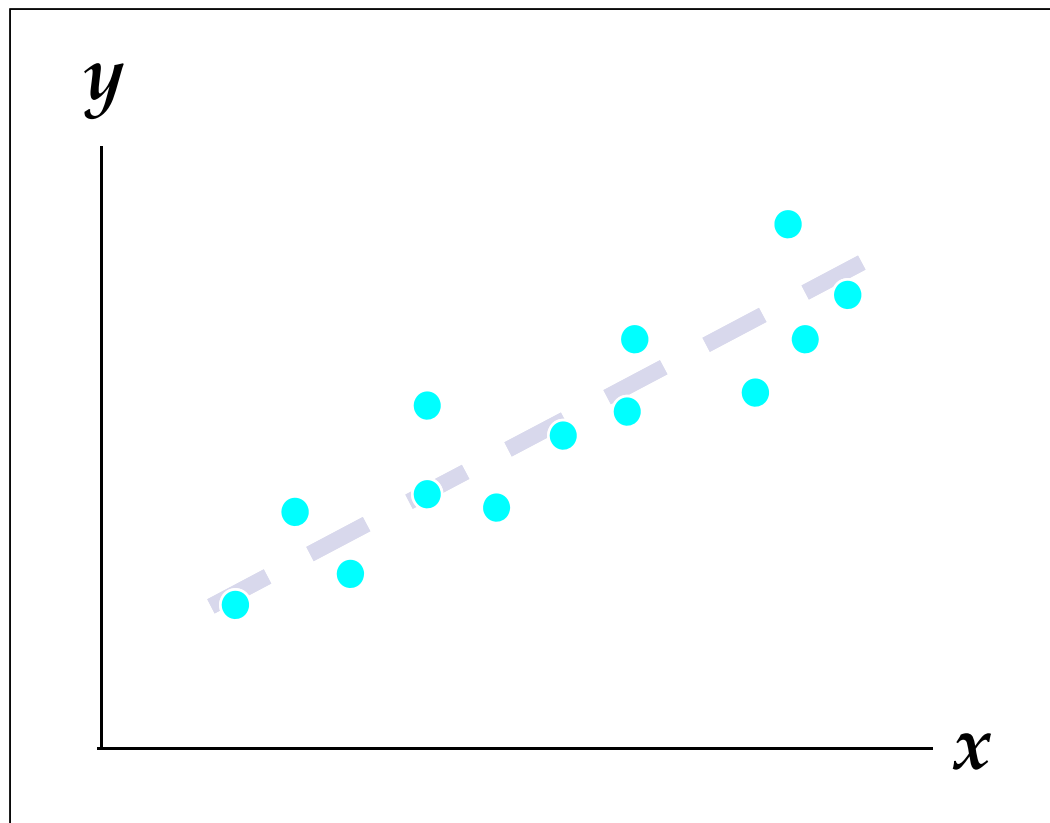


- A scatter diagram, or scatter plot, is a graphical presentation of the relationship between two quantitative variables
- One variable is shown on the horizontal axis and the other is shown on the vertical axis
- The general pattern of the plotted lines suggest the overall relationship between the variables
- A trendline is an approximation of the relationship



Scatter Diagram

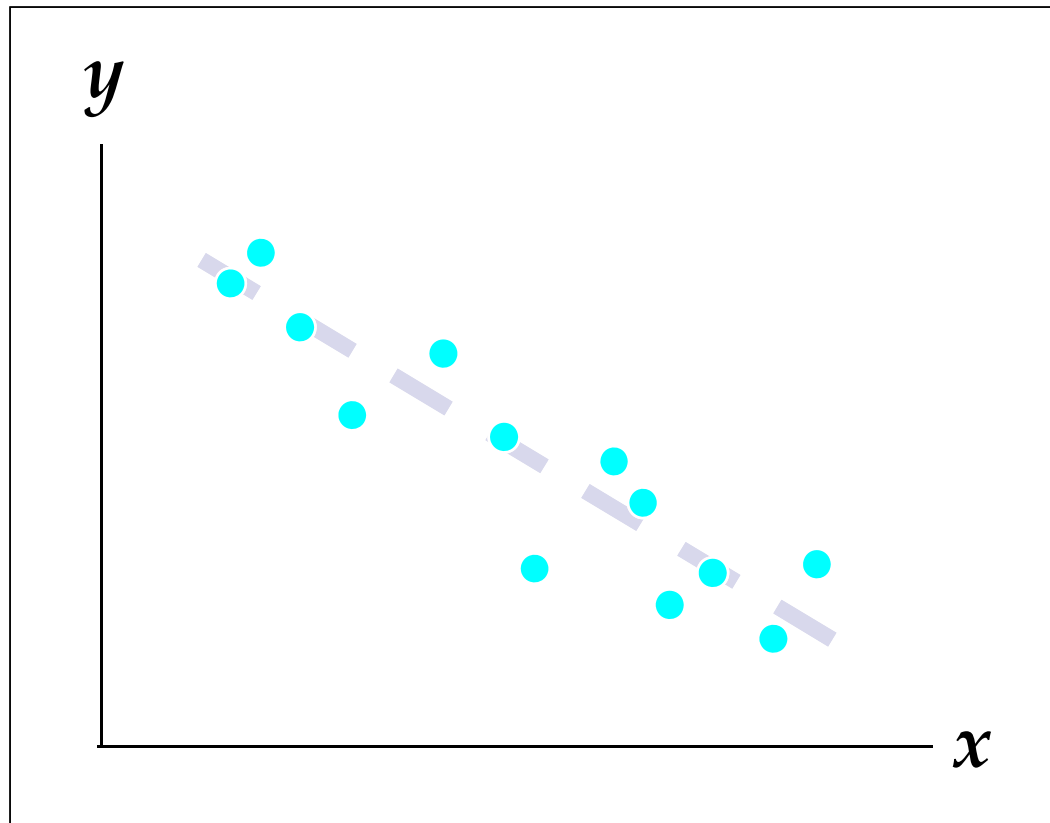
- A Positive Relationship:





Scatter Diagram

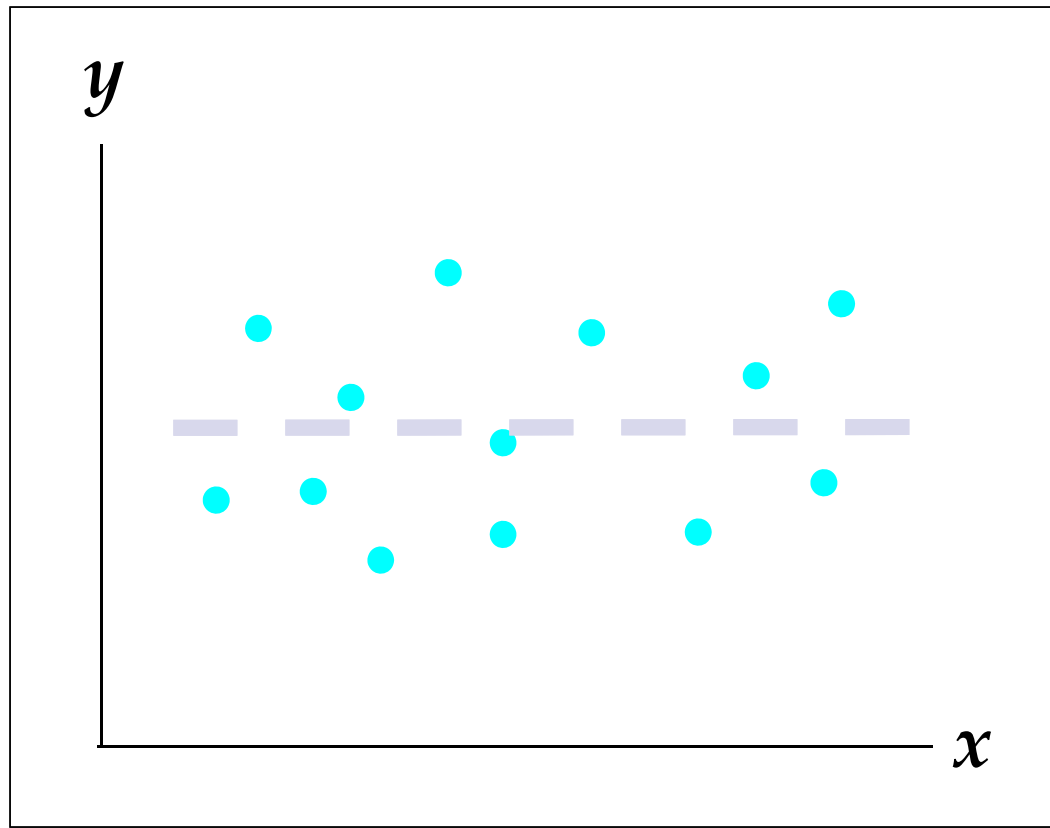
- A Negative Relationship



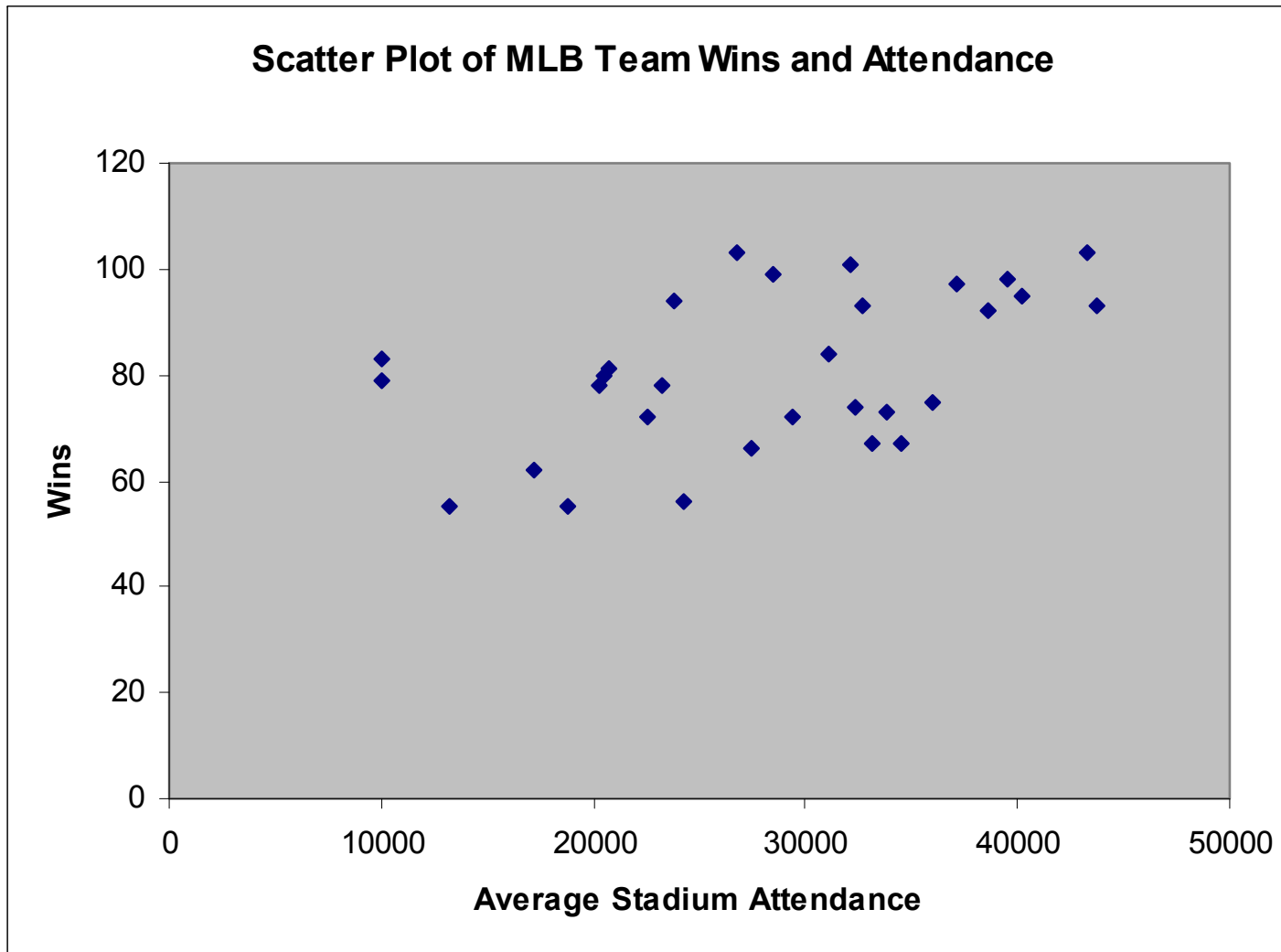
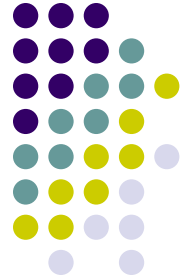


Scatter Diagram

- No Apparent Relationship



Example: MLB Team Wins and Attendance



Tabular and Graphical Descriptive Statistics

