

Data Science Unit 2

Methods for Describing
Sets of Data
Chapter 2

Learning Objectives

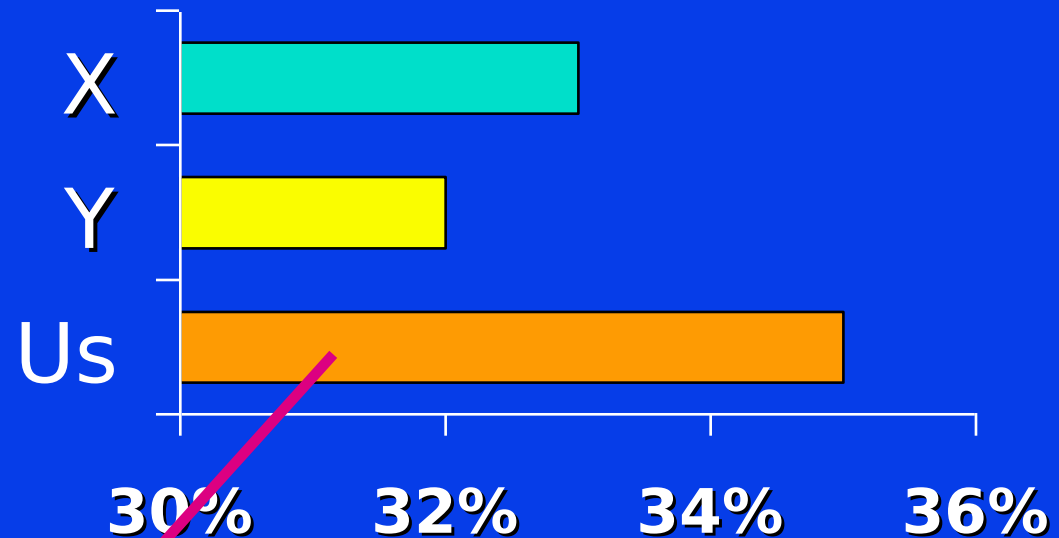
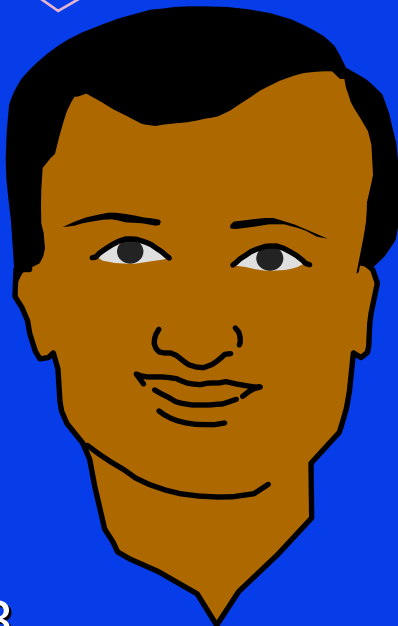
© 2003 Pearson Prentice
Hall

1. Describe Qualitative Data Graphically
2. Describe Numerical Data Graphically
3. Create & Interpret Graphical Displays
4. Explain Numerical Data Properties
5. Describe Summary Measures
6. Analyze Numerical Data Using
Summary Measures

Thinking Challenge

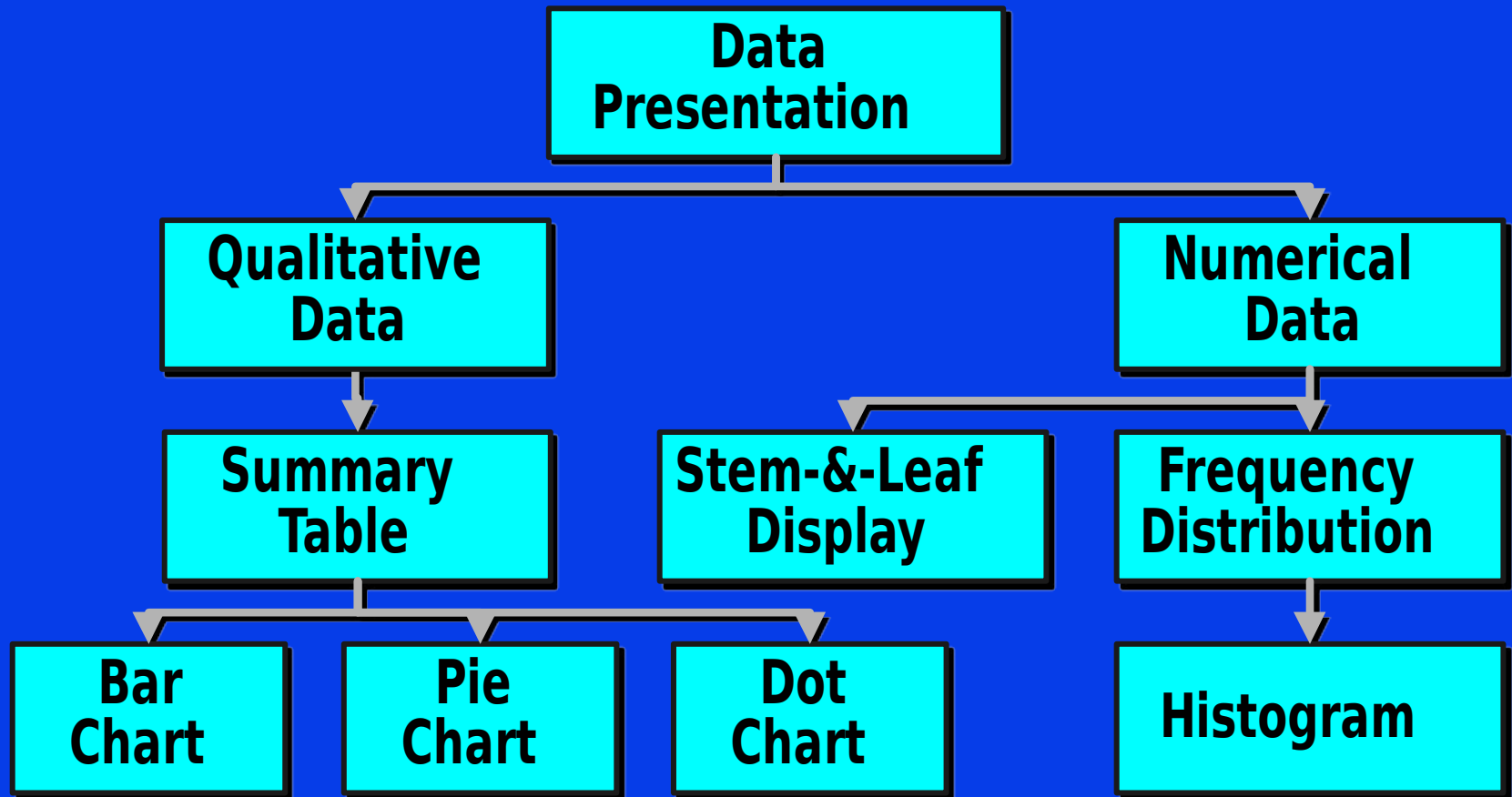
© 2003 Pearson Prentice
Hall

**Our market
share far exceeds
all competitors! -
VP**



Data Presentation

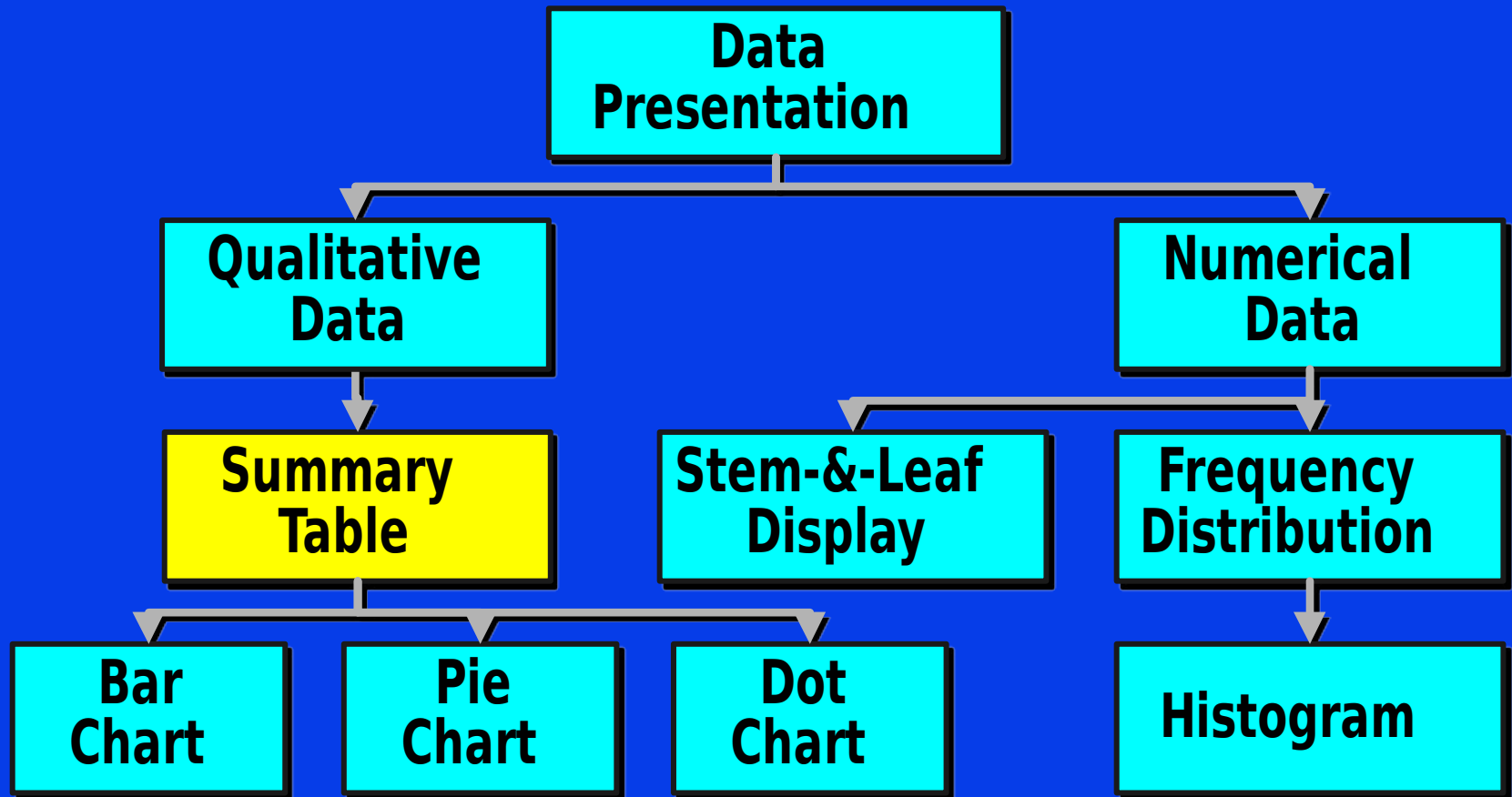
© 2003 Pearson Prentice
Hall



Presenting Qualitative Data

Data Presentation

© 2003 Pearson Prentice
Hall



Summary Table

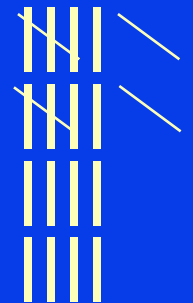
© 2003 Pearson Prentice
Hall

1. Lists Categories & No. Elements in Category
2. Obtained by Tallying Responses in Category
3. May Show Frequencies (Counts), % or Both

Row Is
Categor
y

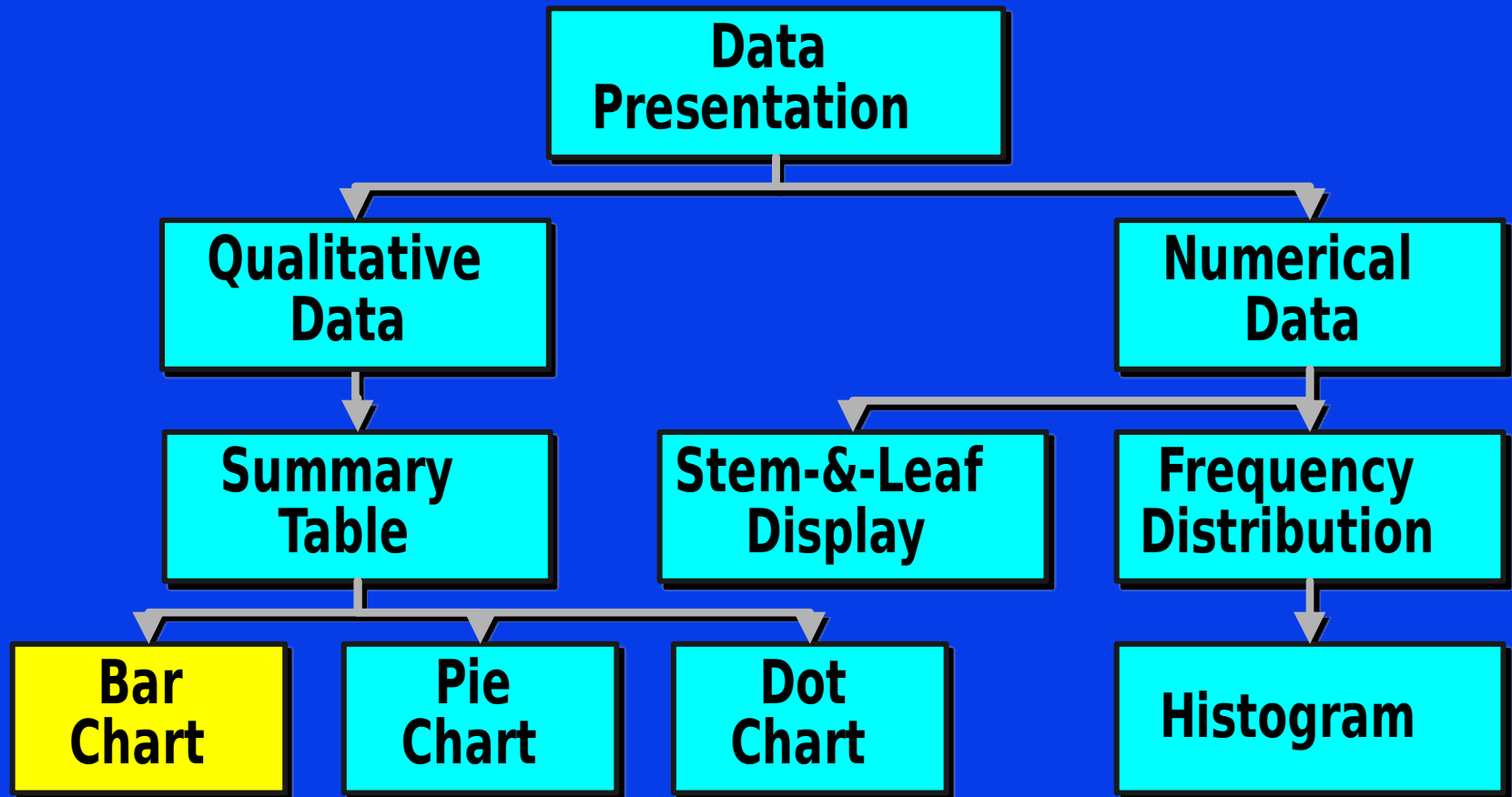
Major	Count
Accounting	130
Economics	20
Management	50
Total	200

Tally:



Data Presentation

© 2003 Pearson Prentice
Hall



Bar Chart

© 2003 Pearson Prentice
Hall

Horizontal
Bars for
Categoric
al
Variables

Major

Mgmt.

Econ.

Acct.

Bar
Length
Shows
Frequency
or %
Equal Bar
Widths

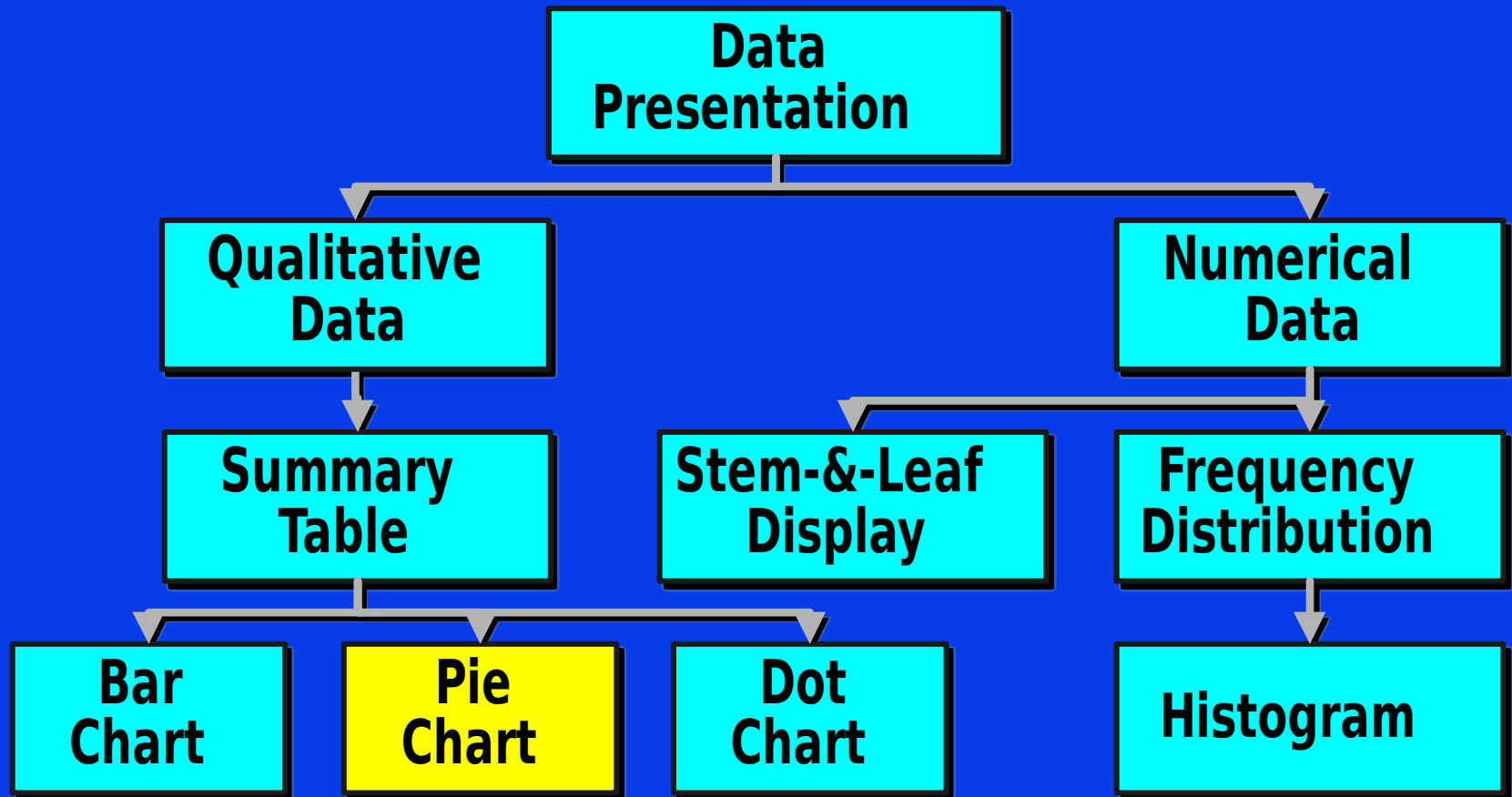
1/2 to 1
Bar Width

Zero Point → **0** **50** **100** **150**

Percent Used Also → **Frequency**

Data Presentation

© 2003 Pearson Prentice
Hall



Pie Chart

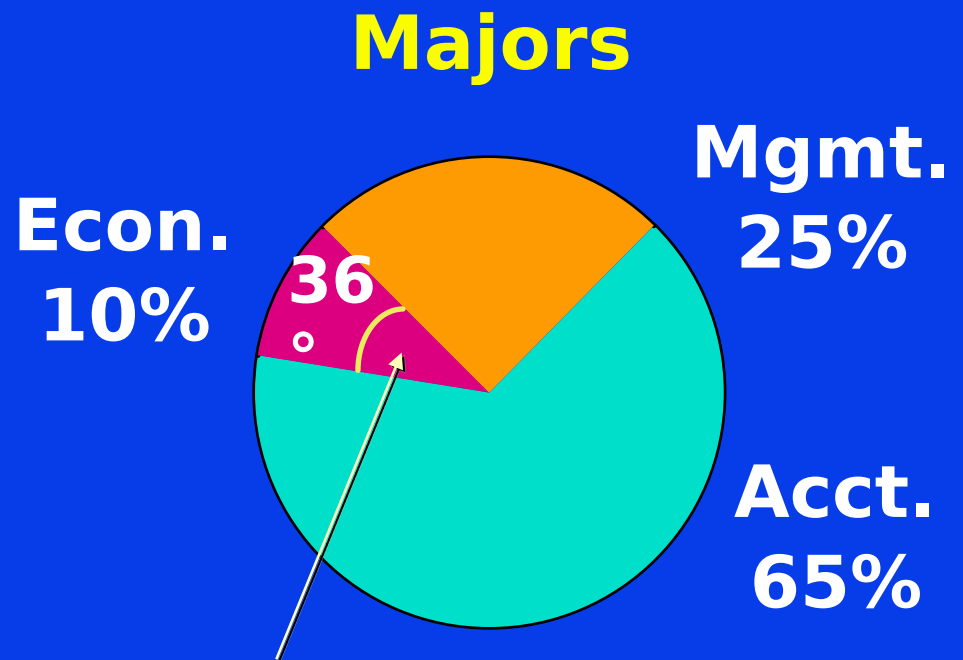
© 2003 Pearson Prentice
Hall

1. Shows
Breakdown of Total
Quantity
into Categories

2. Useful for
Showing Relative
Differences

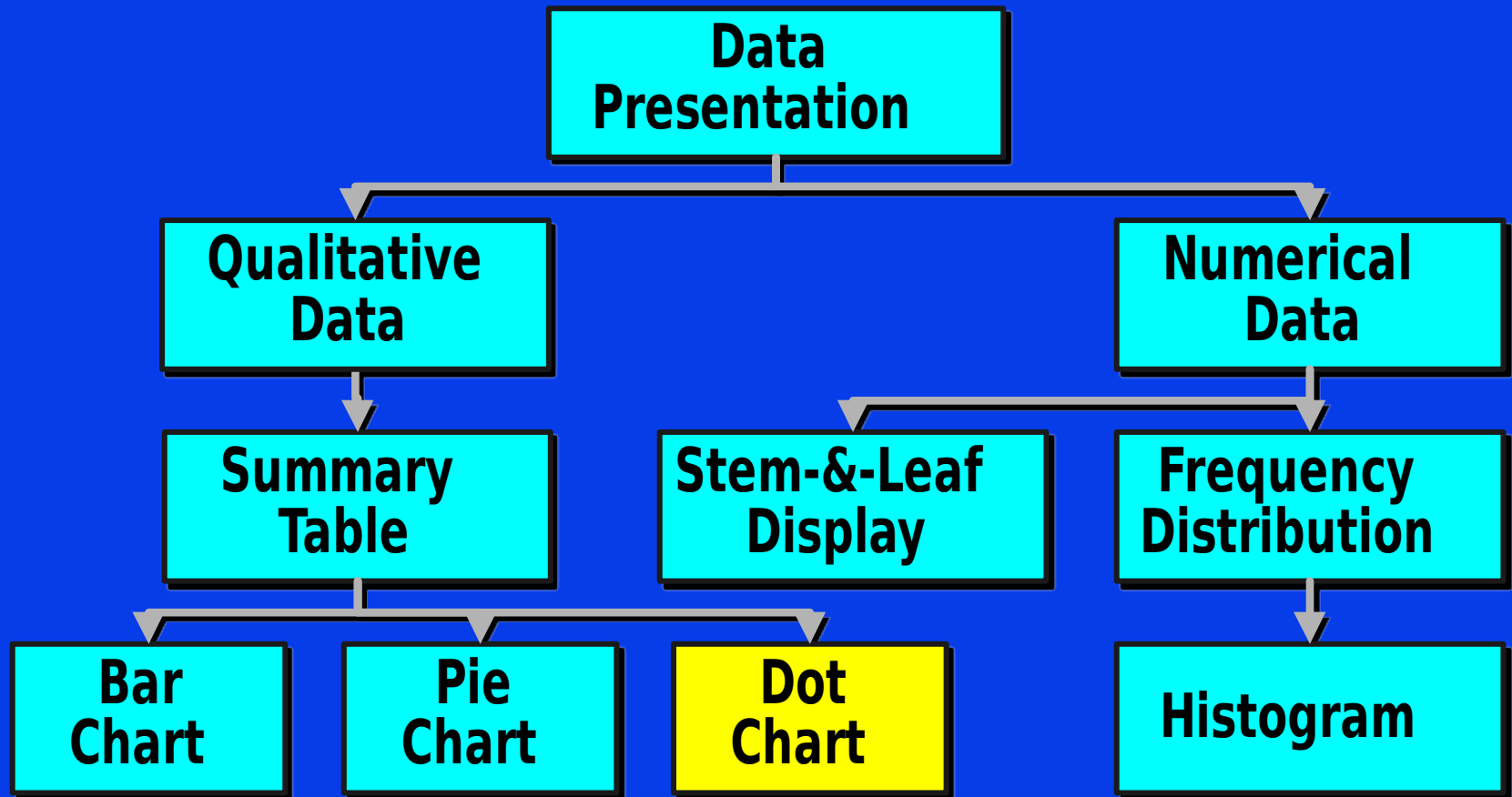
3. Angle Size

■ $(360^\circ)(\text{Percent}) = (360^\circ)(10\%) = 36^\circ$



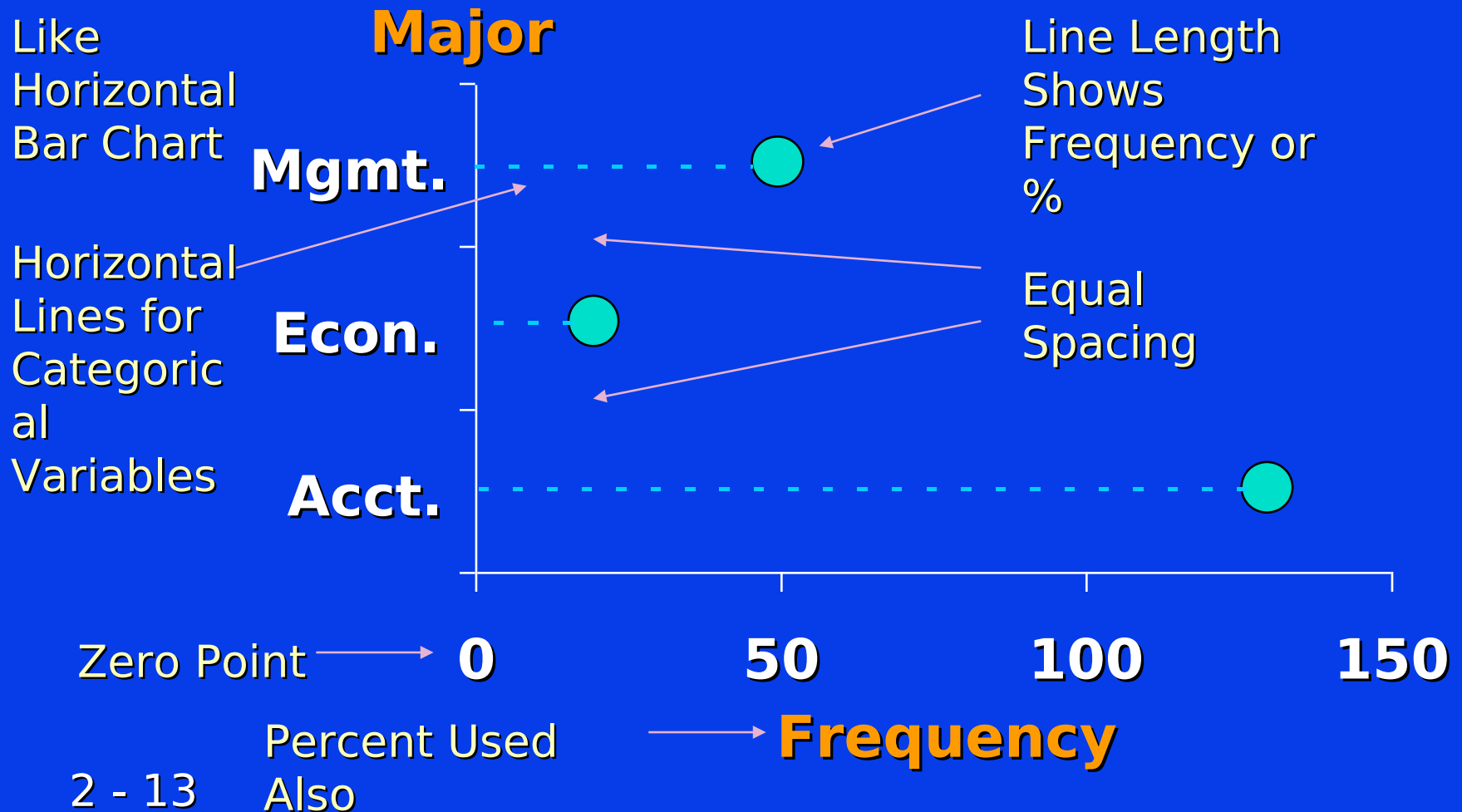
Data Presentation

© 2003 Pearson Prentice
Hall



Dot Chart

© 2003 Pearson Prentice
Hall



Graphical Excellence

© 2003 Pearson Prentice
Hall

1. Greatest Number of Ideas
2. Shortest Time to Understand
3. Least 'Ink' (Fill & Lines)
4. Smallest Space

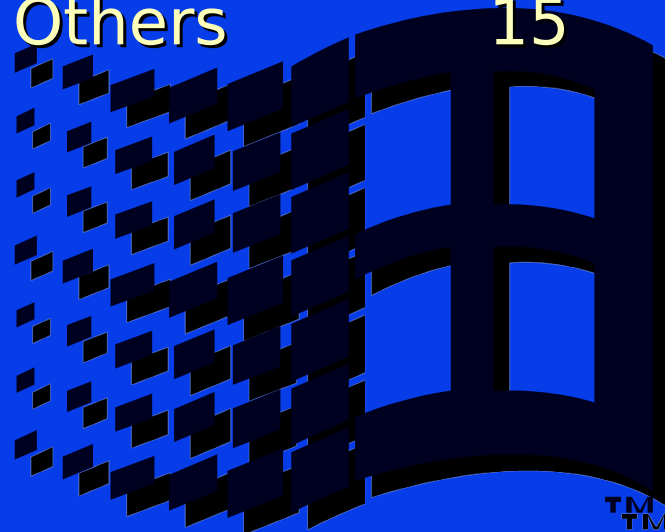


Thinking Challenge

© 2003 Pearson Prentice
Hall

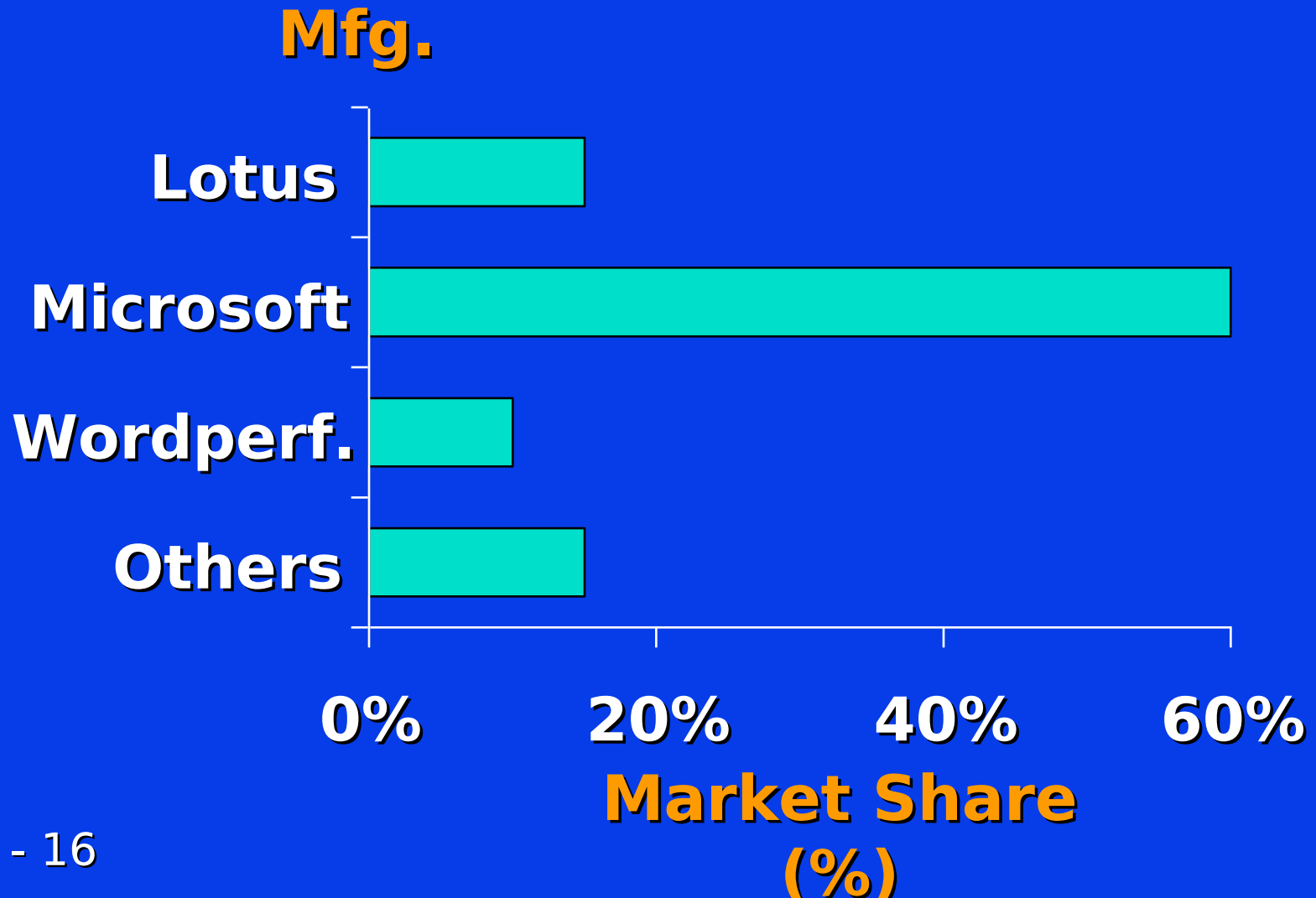
You're an analyst for IRI. You want to show the market shares held by Windows program manufacturers in 1992. Construct a **bar** chart, **pie** chart, & **dot** chart to describe the data.

Mfg. Mkt. Share (%)	
Lotus	15
Microsoft	60
WordPerfect	10
Others	15



Bar Chart Solution*

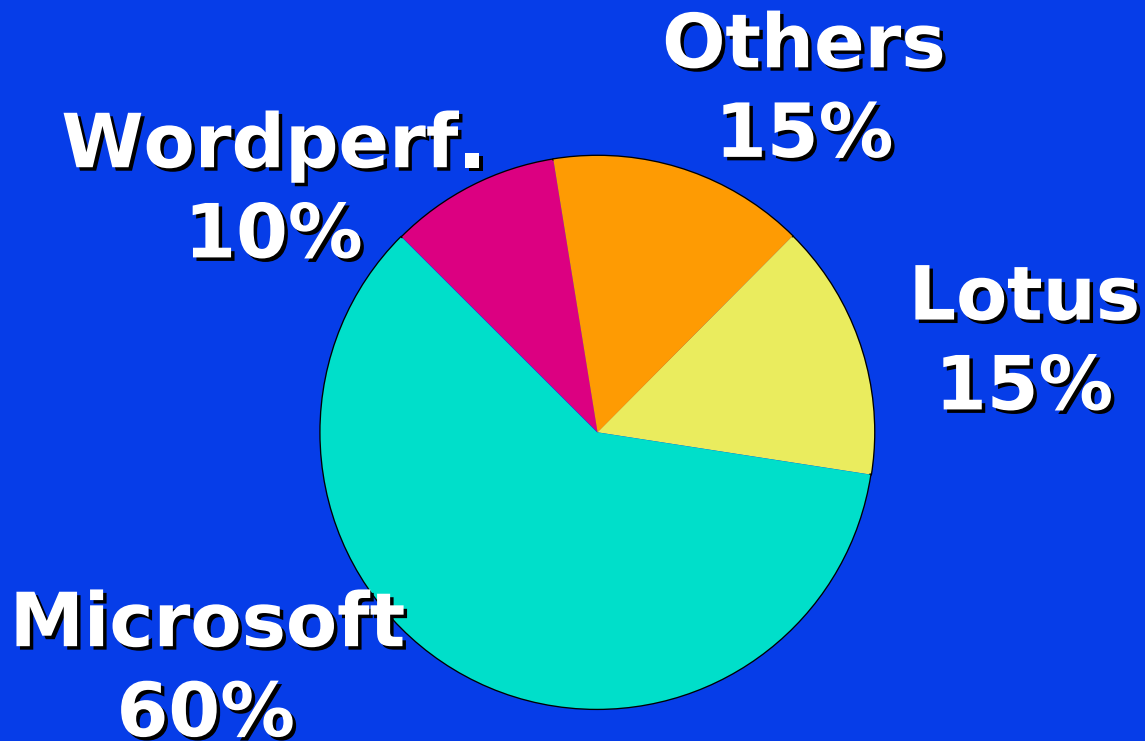
© 2003 Pearson Prentice
Hall



Pie Chart Solution*

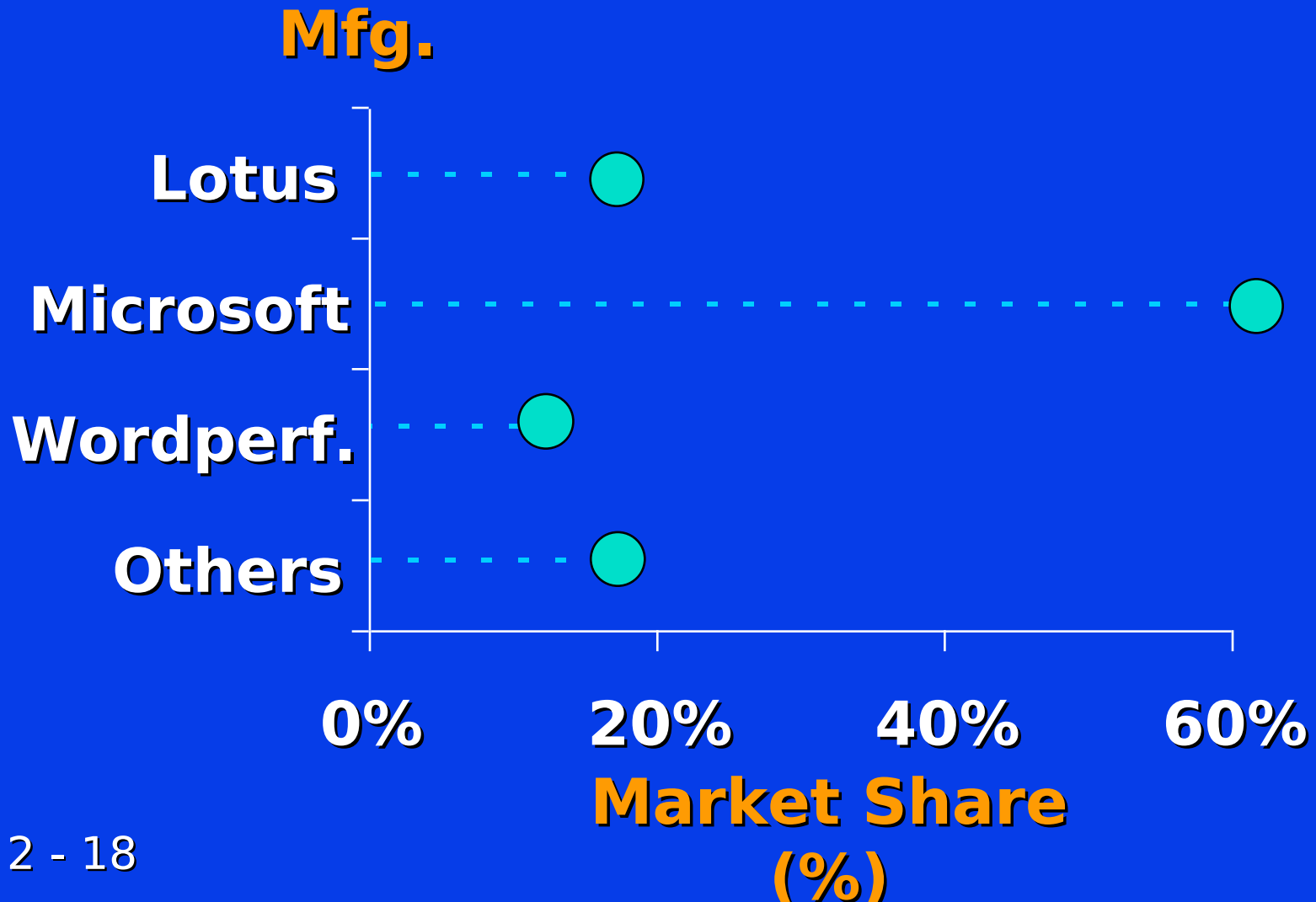
© 2003 Pearson Prentice
Hall

Market Share



Dot Chart Solution*

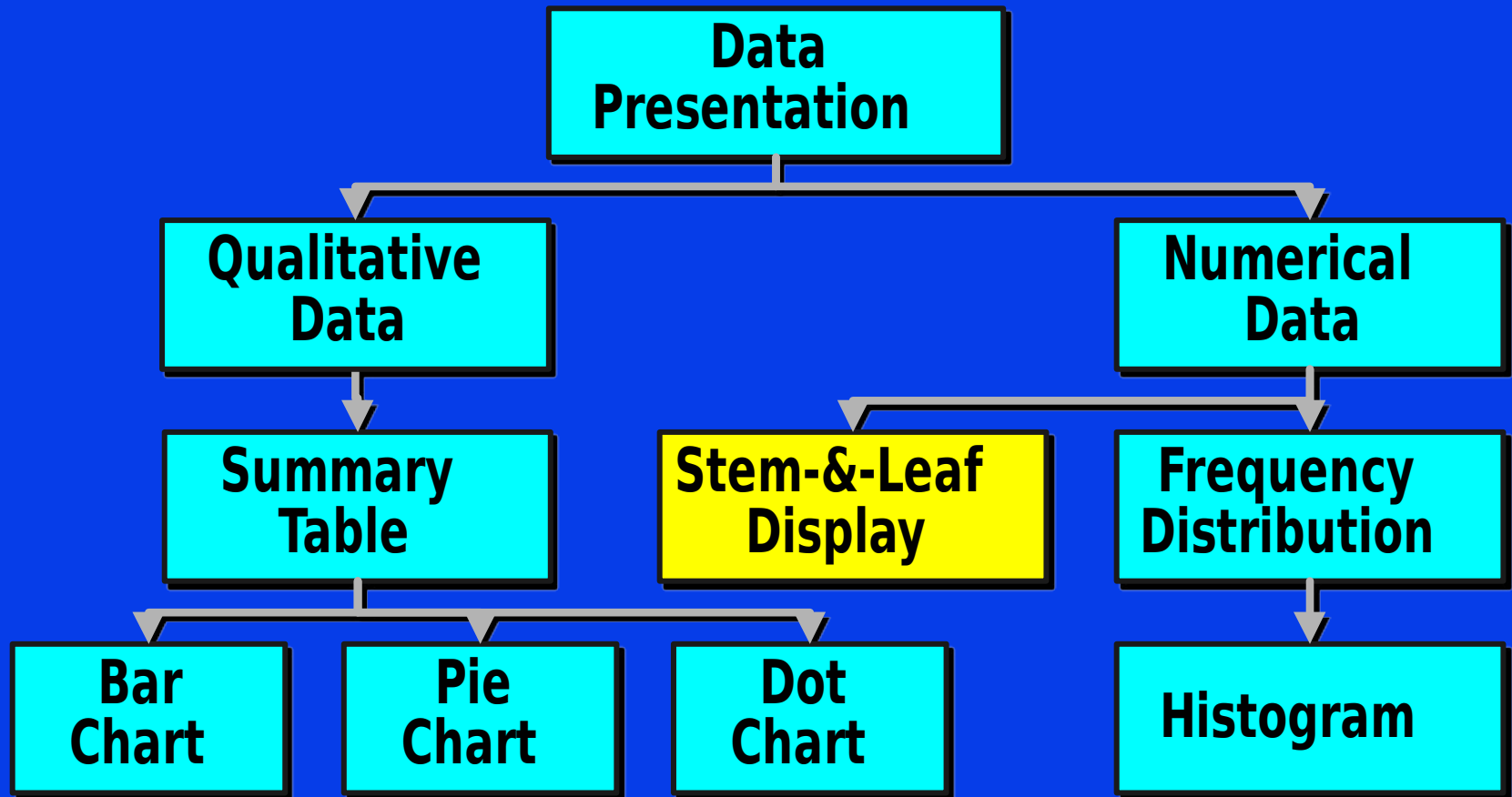
© 2003 Pearson Prentice
Hall



Presenting Numerical Data

Data Presentation

© 2003 Pearson Prentice
Hall

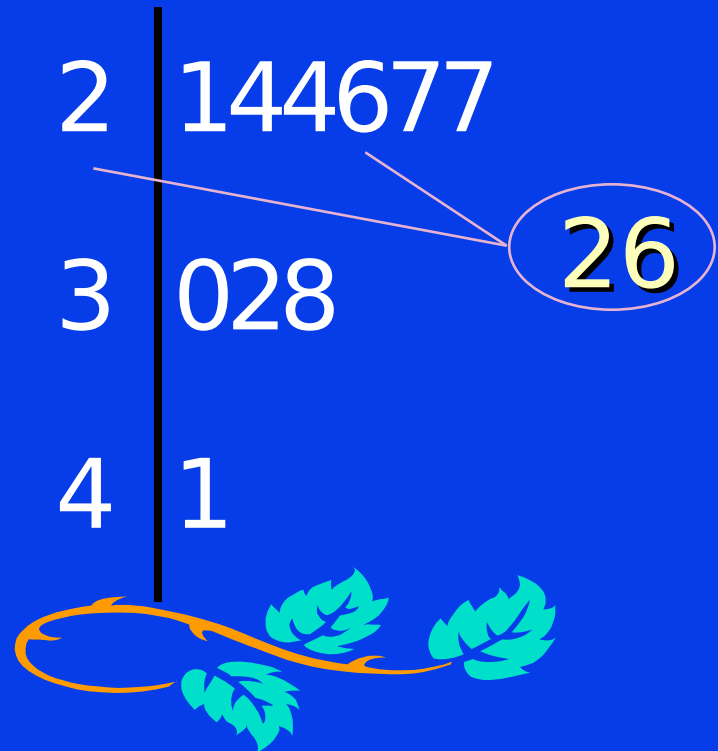


Stem-and-Leaf Display

© 2003 Pearson Prentice Hall

1. Divide Each Observation into Stem Value and Leaf Value

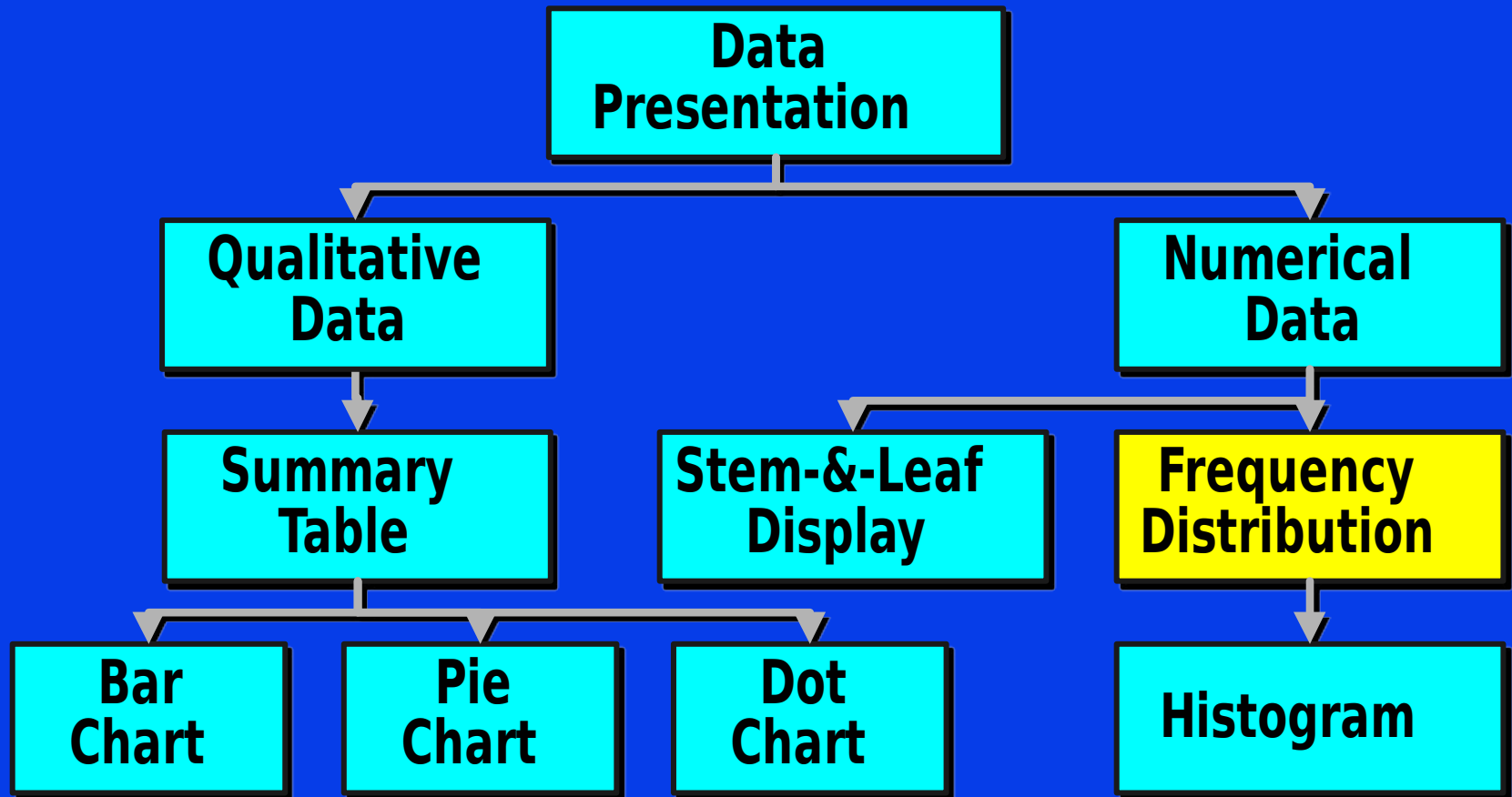
- Stem Value Defines Class
- Leaf Value Defines Frequency



2. Data: 21, 24, 24, **26**, 27, 27, 30, 32, 38, 41
 2 - 21

Data Presentation

© 2003 Pearson Prentice
Hall



Frequency Distribution Table

© 2003 Pearson Prentice
Hall

Raw Data: 24, 26, 24, 21, 27, 27, 30, 41,
32, 38

Class	Frequency
15 but < 25	3
25 but < 35	5
35 but < 45	2

Frequency Distribution Table Steps

© 2003 Pearson Prentice
Hall

1. Determine Range
2. Select Number of Classes
 - *Usually* Between 5 & 15 Inclusive
3. Compute Class Intervals (Width)
4. Determine Class Boundaries (Limits)
5. Compute Class Midpoints
6. Count Observations & Assign to Classes

Frequency Distribution Table Example

© 2003 Pearson Prentice
Hall

Raw Data: 24, 26, 24, 21, 27, 27, 30, 41, 32, 38

Class		Midpoint	Frequency
Width {	15 but < 25	20	3
	25 but < 35	30	5
	35 but < 45	40	2

Boundaries

(Upper + Lower Boundaries) / 2

2 - 25

Relative Frequency & % Distribution Tables

© 2003 Pearson Prentice Hall

Relative Frequency Distribution

Class	Prop.
15 but < 25	.3
25 but < 35	.5
35 but < 45	.2

Percentage Distribution

Class	%
15 but < 25	30.0
25 but < 35	50.0
35 but < 45	20.0

Cumulative Percentage Distribution Table

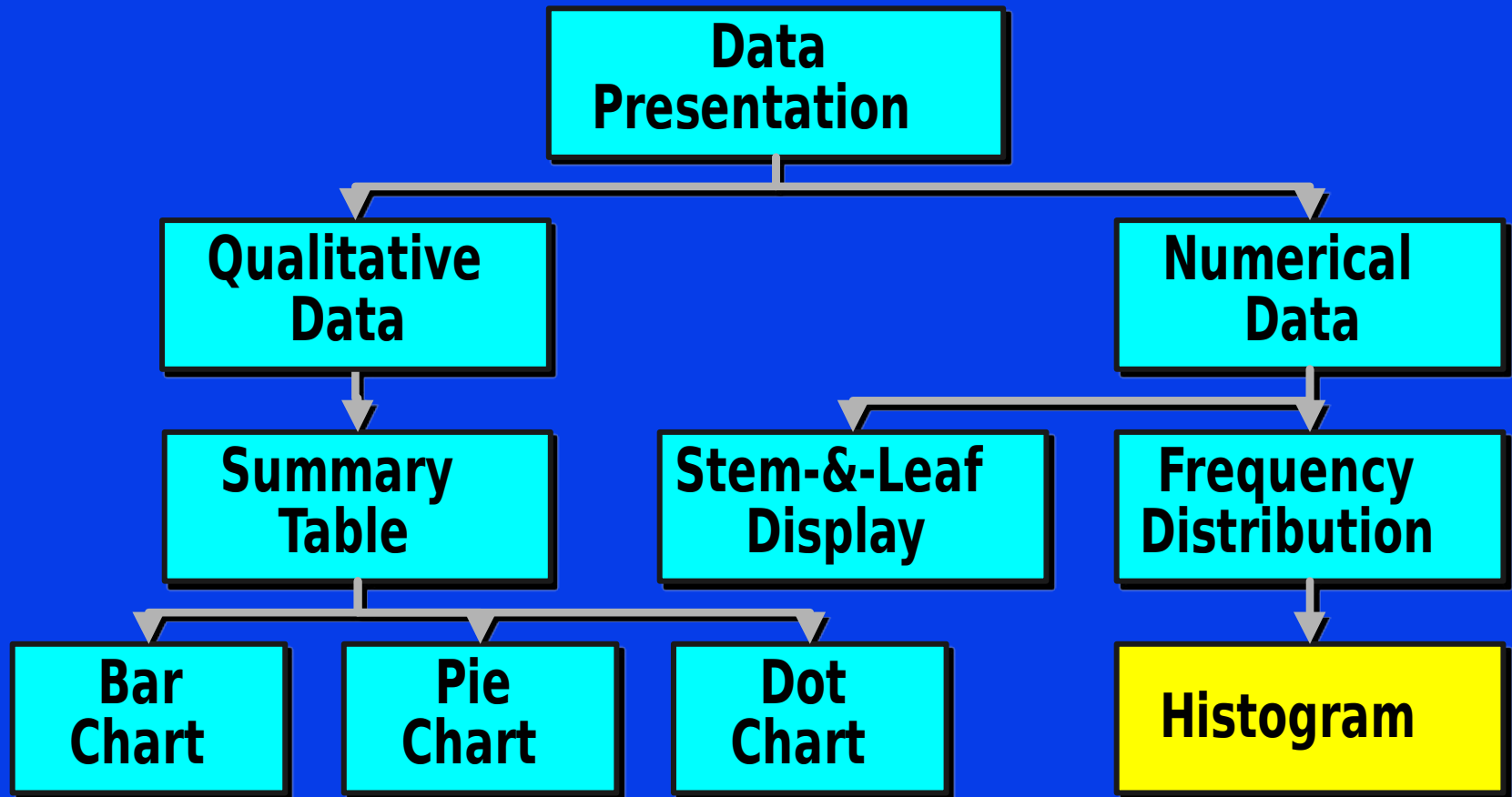
© 2003 Pearson Prentice
Hall

Raw Data: 24, 26, 24, 21, 27, 27, 30, 41,
32, 38

Lower Class Boundar y	Class	Cumulative Percentage	Percentage Less than Lower Class Boundary
	15 but < 25	0.0	
	25 but < 35	30.0	
	35 but < 45	80.0	30% + 50%
	45 but < 55	100.0	80% + 20%

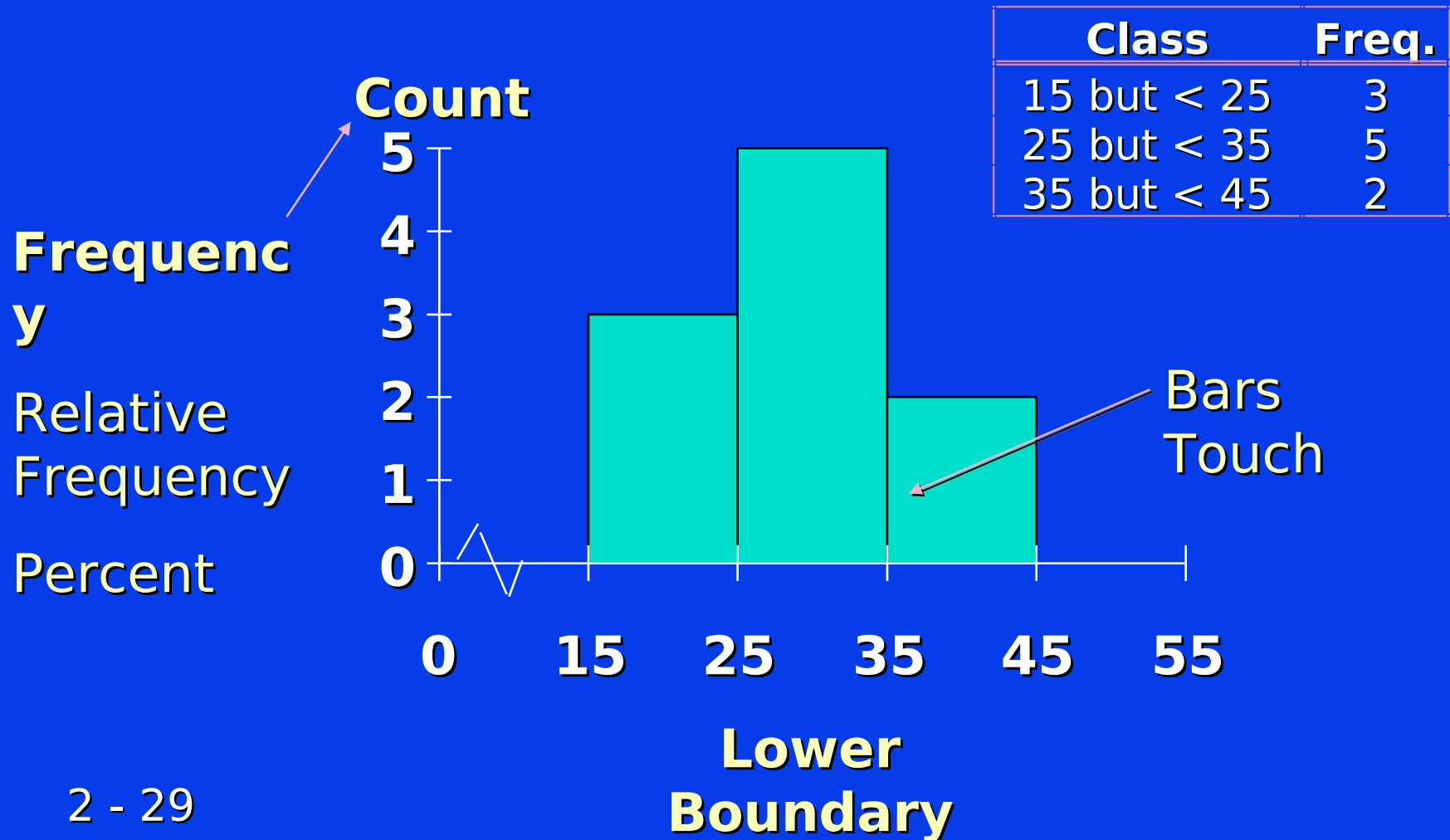
Data Presentation

© 2003 Pearson Prentice
Hall



Histogram

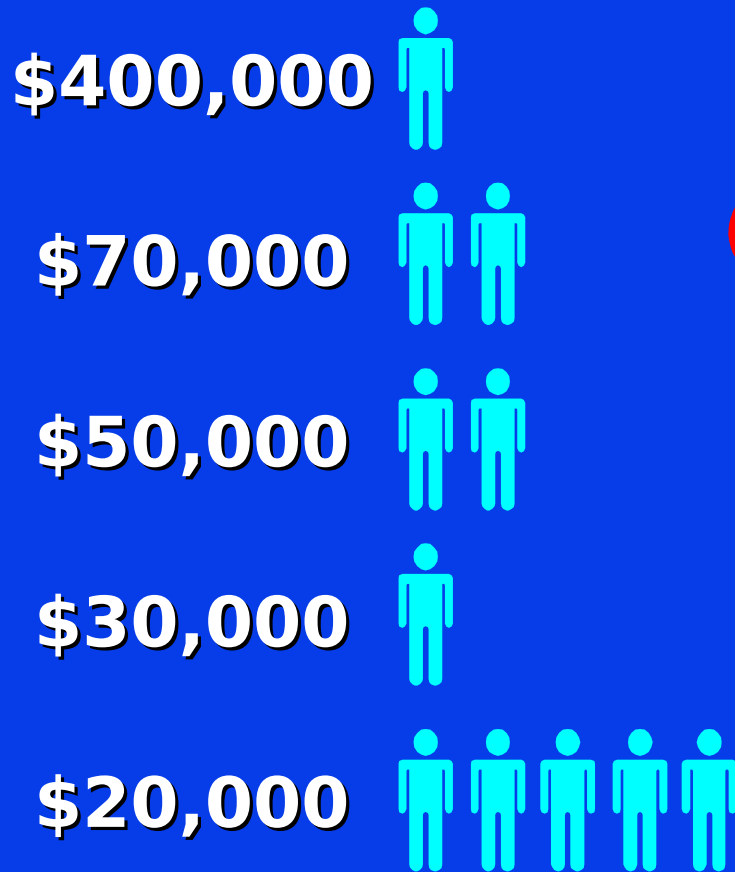
© 2003 Pearson Prentice
Hall



Numerical Data Properties

Thinking Challenge

© 2003 Pearson Prentice
Hall



Strike!

... employees cite low pay -- most workers earn only \$20,000.

... President claims average pay is \$70,000!

Standard Notation

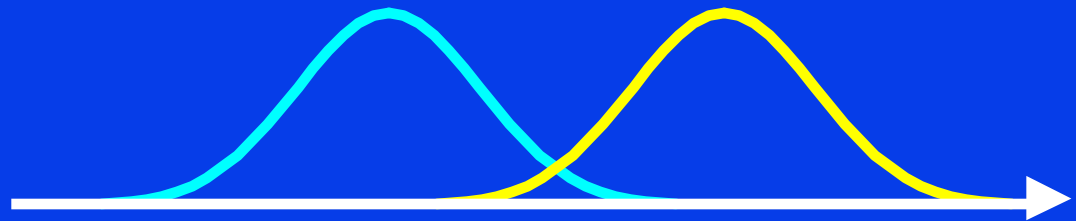
© 2003 Pearson Prentice
Hall

Measure	Sample	Population
Mean	$\bar{X}$	μ
Stand. Dev.	S	σ
Variance	S^2	σ^2
Size	n	N

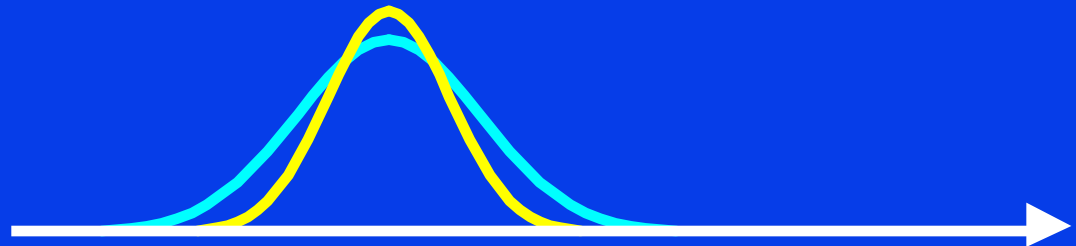
Numerical Data Properties

© 2003 Pearson Prentice
Hall

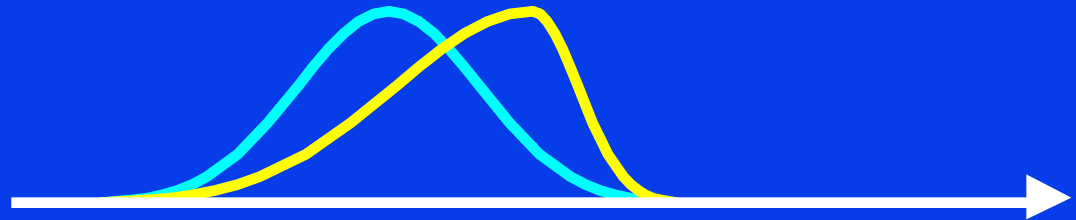
**Central
Tendency
(Location)**



**Variation
(Dispersion)**

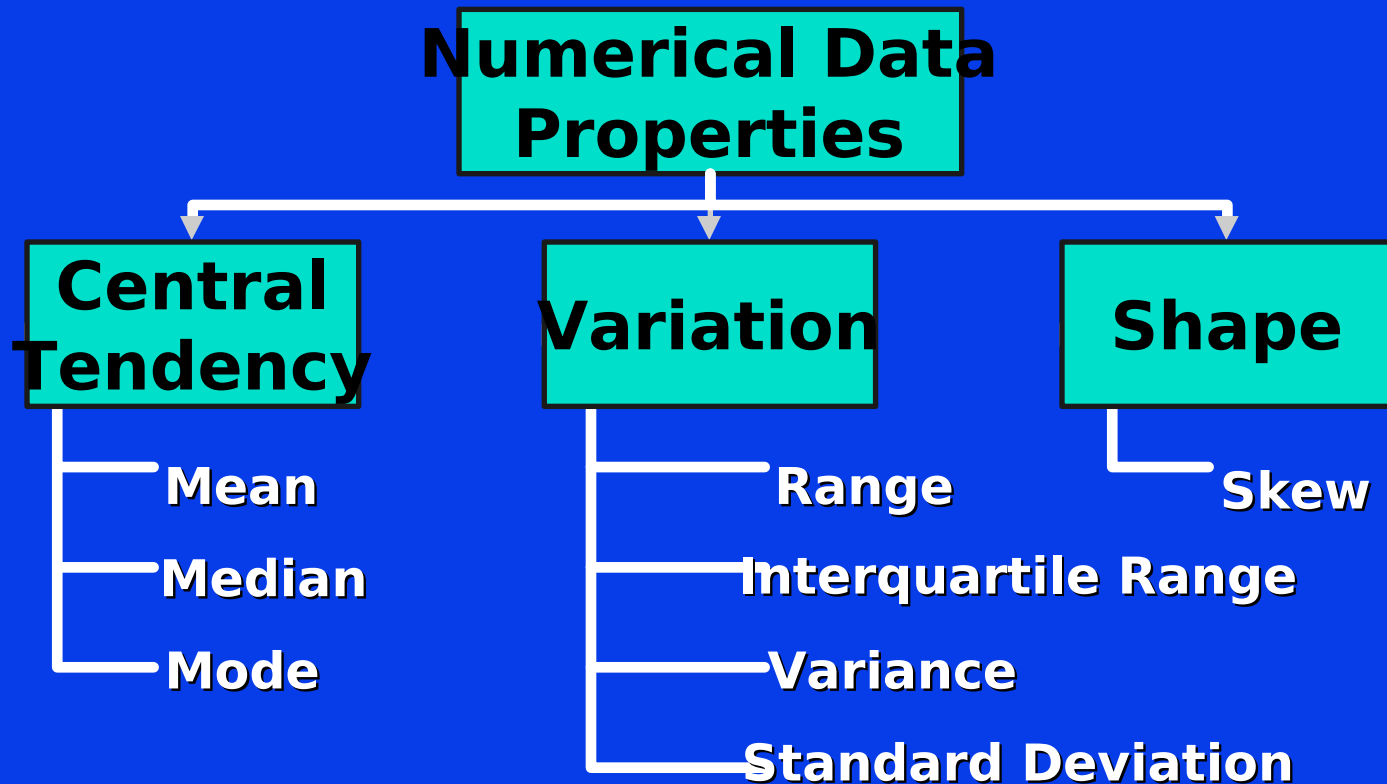


Shape



Numerical Data Properties & Measures

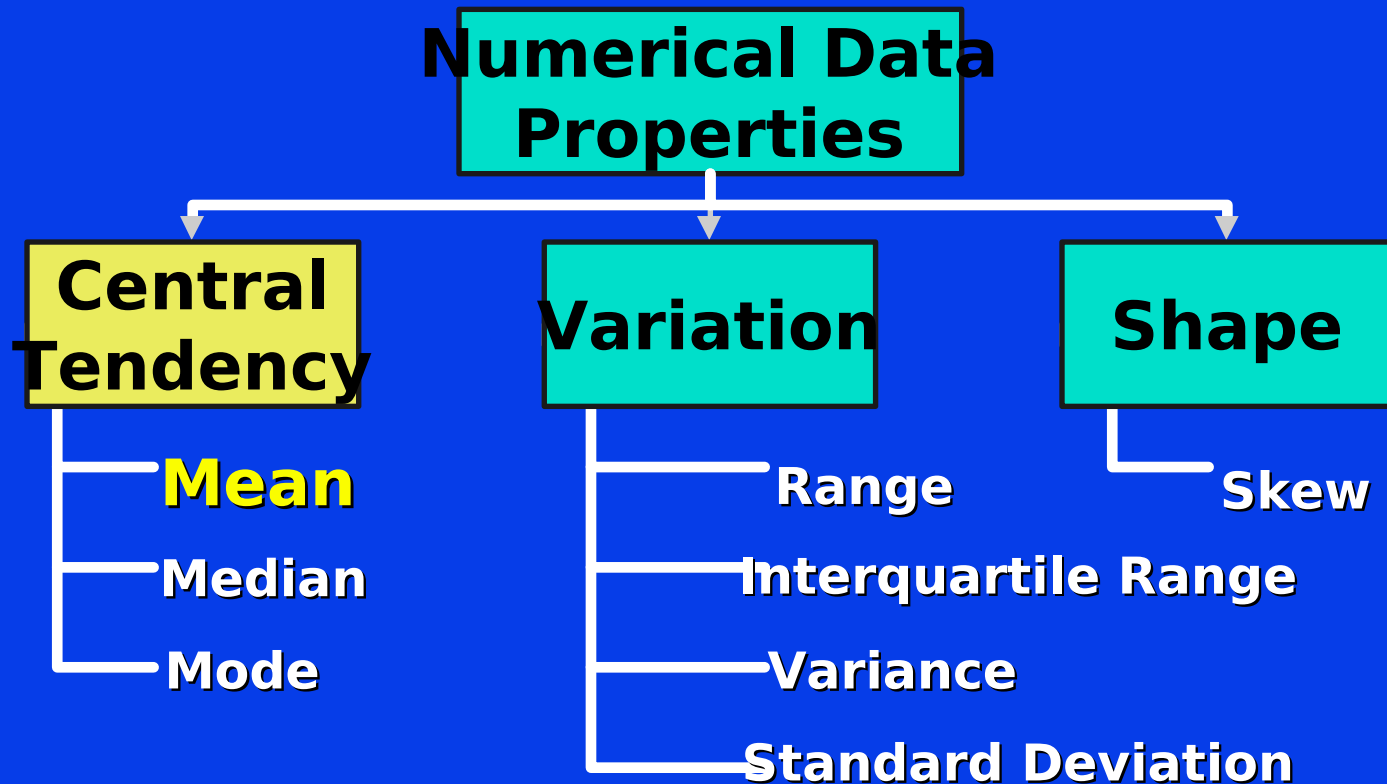
© 2003 Pearson Prentice
Hall



Central Tendency

Numerical Data Properties & Measures

© 2003 Pearson Prentice
Hall



Mean

© 2003 Pearson Prentice
Hall

1. Measure of Central Tendency
2. Most Common Measure
3. Acts as 'Balance Point'
4. Affected by Extreme Values ('Outliers')
5. Formula (Sample Mean)

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Mean Example

© 2003 Pearson Prentice
Hall

Raw Data: 10.3 4.9 8.9 11.7 6.3

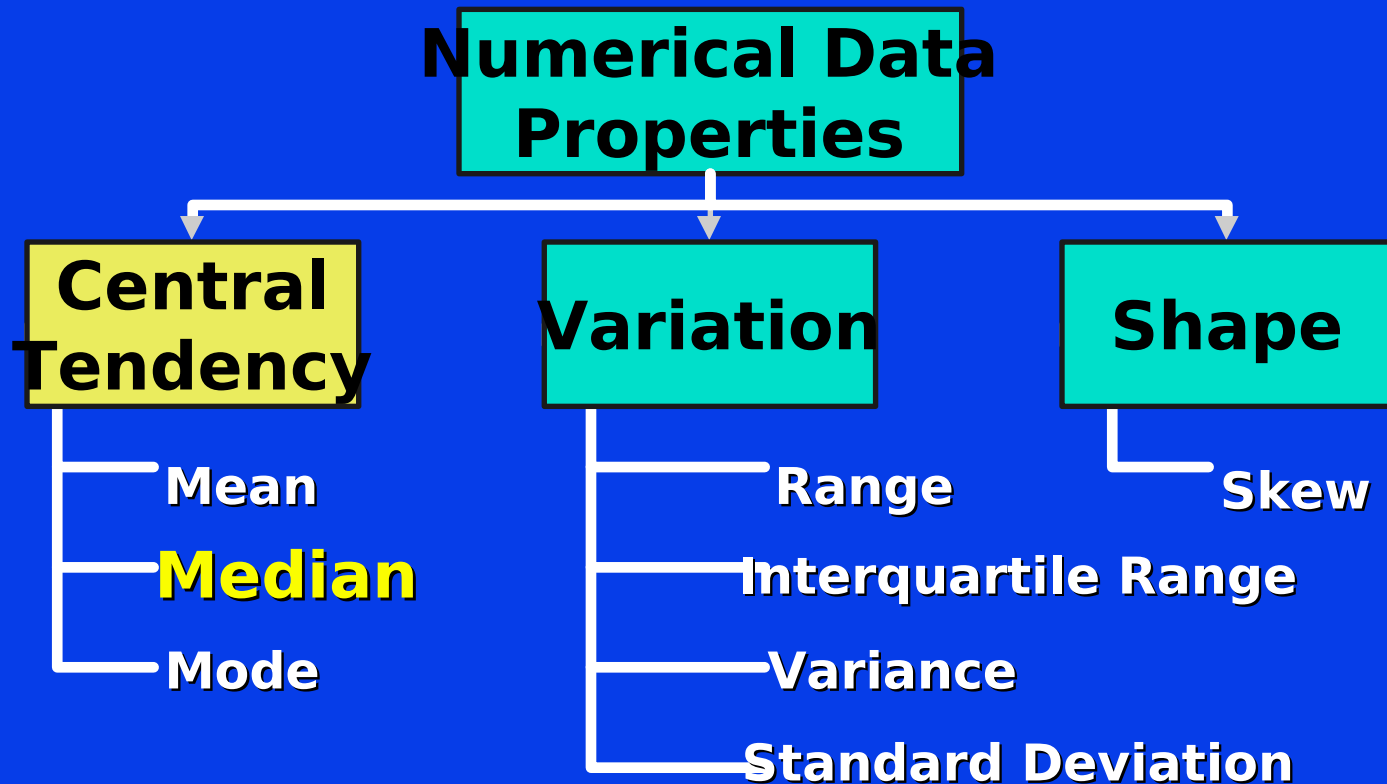
$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{X_1 + X_2 + X_3 + X_4 + X_5 + X_6}{6}$$

$$= \frac{10.3 + 4.9 + 8.9 + 11.7 + 6.3 + 7.7}{6}$$

$$= 8.30$$

Numerical Data Properties & Measures

© 2003 Pearson Prentice
Hall



Median

© 2003 Pearson Prentice
Hall

1. Measure of Central Tendency
2. Middle Value In Ordered Sequence
 - If Odd n , Middle Value of Sequence
 - If Even n , Average of 2 Middle Values
3. Position of Median in Sequence

$$\text{Positioning Point} = \frac{n+1}{2}$$

4. Not Affected by Extreme Values

Median Example

Odd-Sized Sample

© 2003 Pearson Prentice
Hall

Raw Data:	24.1	22.6	21.5	23.7	22.6
Ordered:	21.5	22.6	22.6	23.7	24.1
Position:	1	2	3	4	5



$$\text{Positioning Point} = \frac{n+1}{2} = \frac{5+1}{2} = 3.0$$

Median = 22.6

Median Example

Even-Sized Sample

© 2003 Pearson Prentice Hall

Raw Data: 10.3 4.9 8.9 11.7 6.3
7.7

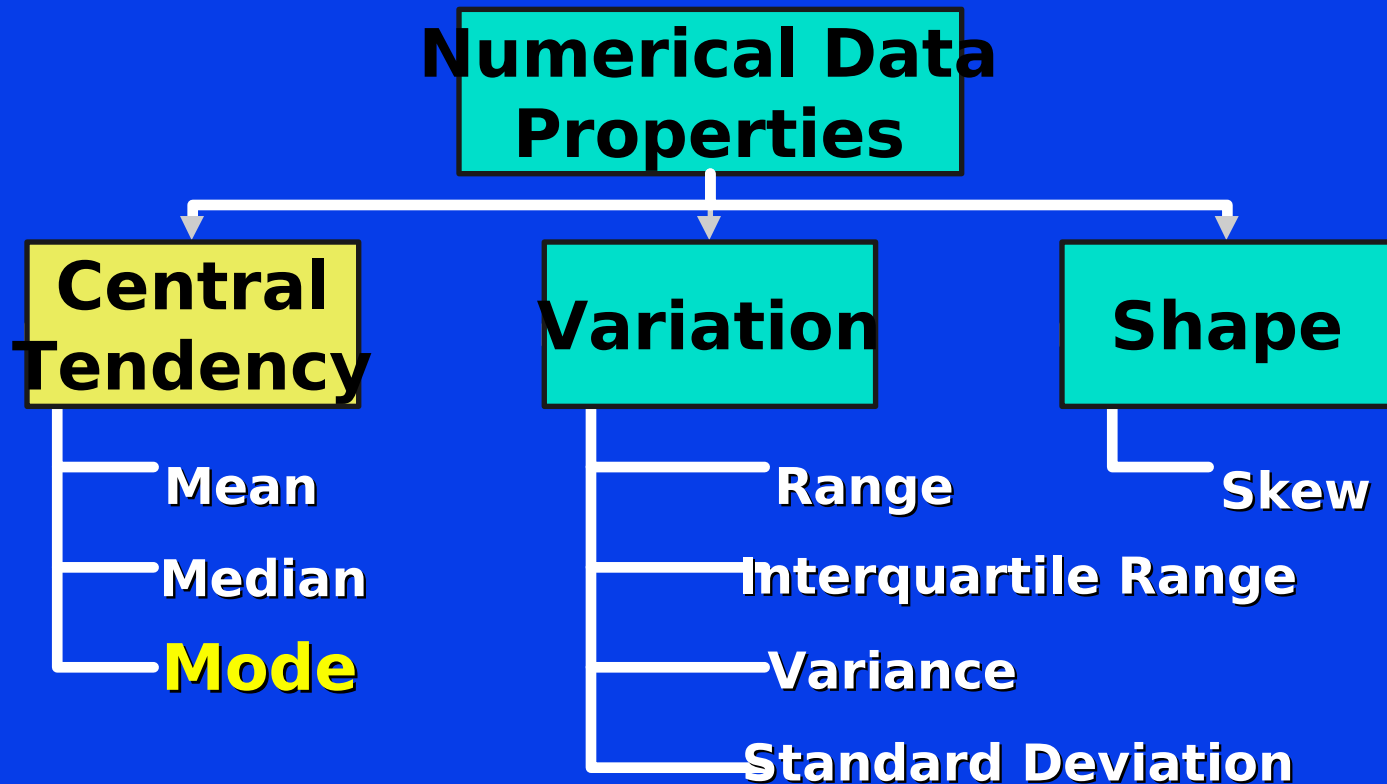
Ordered: 4.9 6.3 **7.7** **8.9** 10.3
11.7

Positioning Point = $\frac{n+1}{2} = \frac{6+1}{2} = 3.5$

$$\text{Median} = \frac{7.7 + 8.9}{2} = 8.30$$

Numerical Data Properties & Measures

© 2003 Pearson Prentice
Hall



Mode

1. Measure of Central Tendency
2. Value That Occurs Most Often
3. Not Affected by Extreme Values
4. May Be No Mode or Several Modes
5. May Be Used for Numerical & Categorical Data

Mode Example

© 2003 Pearson Prentice
Hall

No Mode

Raw Data: 10.3 4.9 8.9 11.7 6.3 7.7

One Mode

Raw Data: 6.3 **4.9** 8.9 6.3 **4.9** **4.9**

More Than 1 Mode

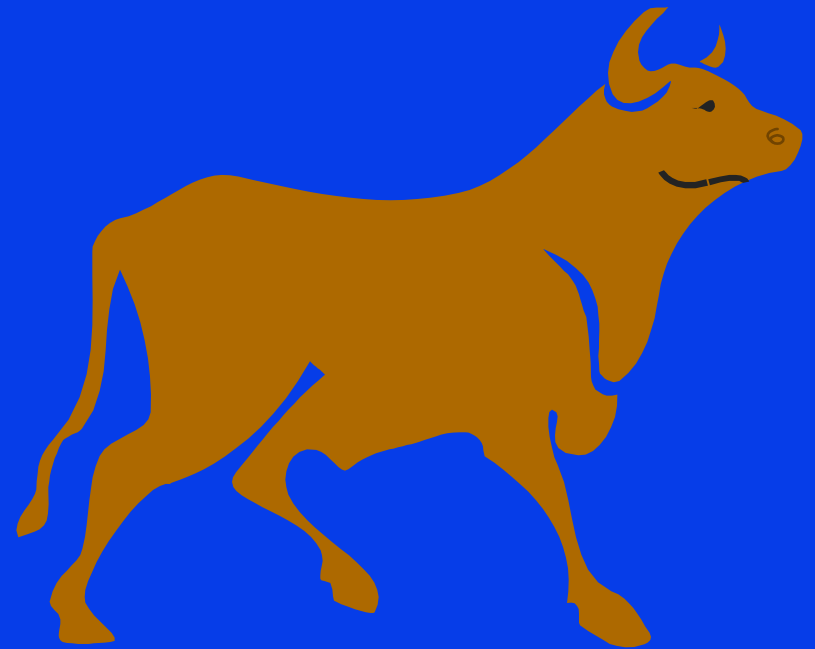
Raw Data: 21 **28** **28** 41 **43** **43**

Thinking Challenge

© 2003 Pearson Prentice
Hall

You're a financial analyst for Prudential-Bache Securities. You have collected the following closing stock prices of new stock issues: **17, 16, 21, 18, 13, 16, 12, 11.**

Describe the stock prices in terms of **central tendency.**



Central Tendency Solution*

© 2003 Pearson Prentice
Hall

Mean

$$\begin{aligned}\bar{X} &= \frac{\sum_{i=1}^n X_i}{n} = \frac{X_1 + X_2 + \cdots + X_8}{8} \\ &= \frac{17 + 16 + 21 + 18 + 13 + 16 + 12 + 11}{8} \\ &= 15.5\end{aligned}$$

Central Tendency Solution*

© 2003 Pearson Prentice
Hall

Median

Raw Data:	17	16	21	18	13	16	12	11
Ordered:	11	12	13	16	16	17	18	21
Position:	1	2	3	4	5	6	7	8

$$\text{Positioning Point} = \frac{n+1}{2} = \frac{8+1}{2} = 4.5$$

$$\text{Median} = \frac{16+16}{2} = 16$$

Central Tendency Solution*

© 2003 Pearson Prentice
Hall

Mode

Raw Data: 17 16 21 18 13 16 12
 11

Ordered: 11 12 13 **16 16** 17 18
 21

Summary of Central Tendency Measures

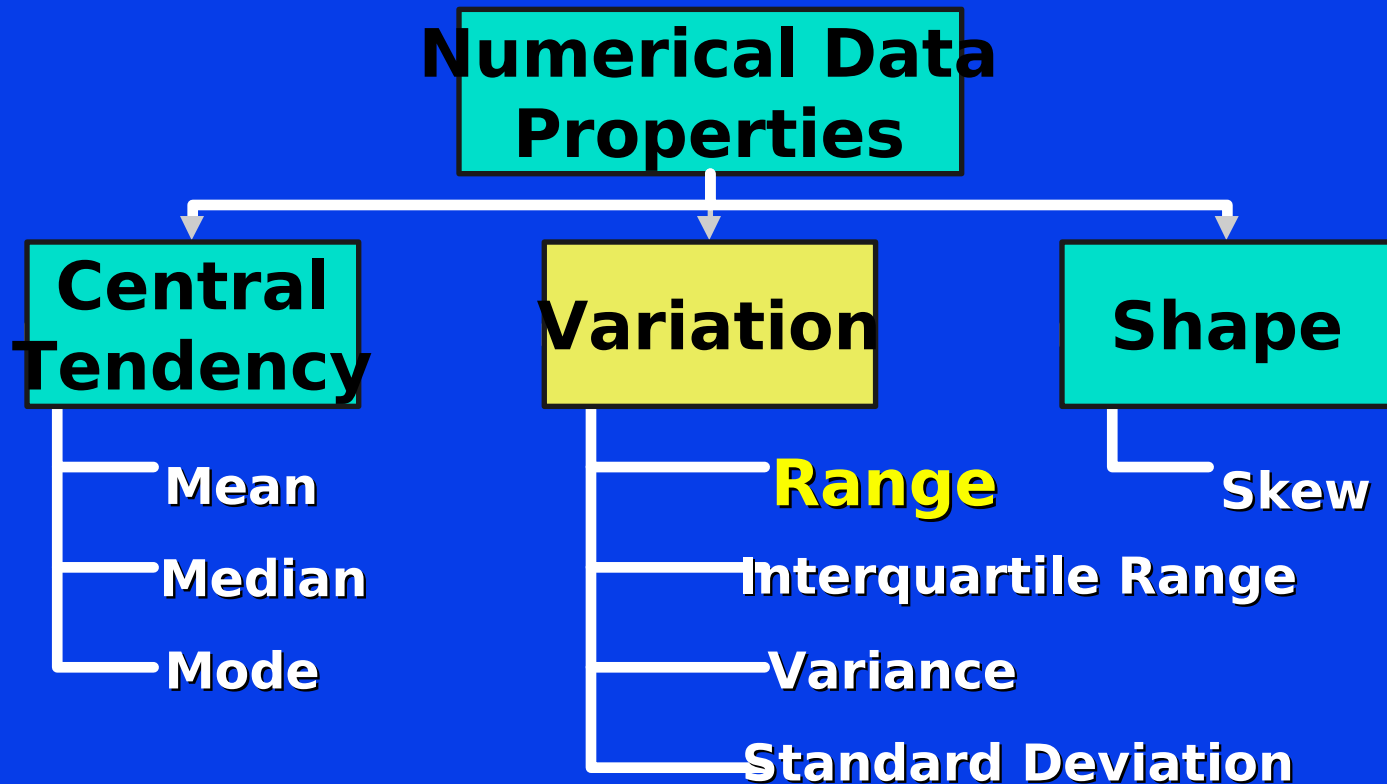
© 2003 Pearson Prentice
Hall

Measure	Equation	Description
Mean	$\Sigma X_i / n$	Balance Point
Median	$\frac{(n+1)\text{Position}}{2}$	Middle Value When Ordered
Mode	none	Most Frequent

Variation

Numerical Data Properties & Measures

© 2003 Pearson Prentice
Hall



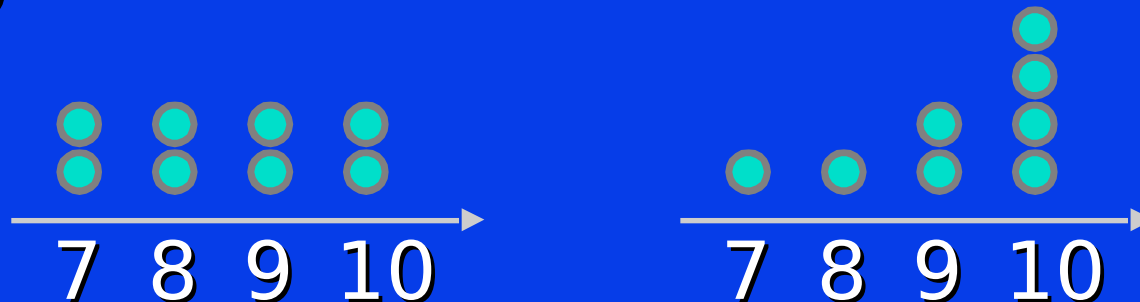
Range

© 2003 Pearson Prentice
Hall

1. Measure of Dispersion
2. Difference Between Largest & Smallest Observations

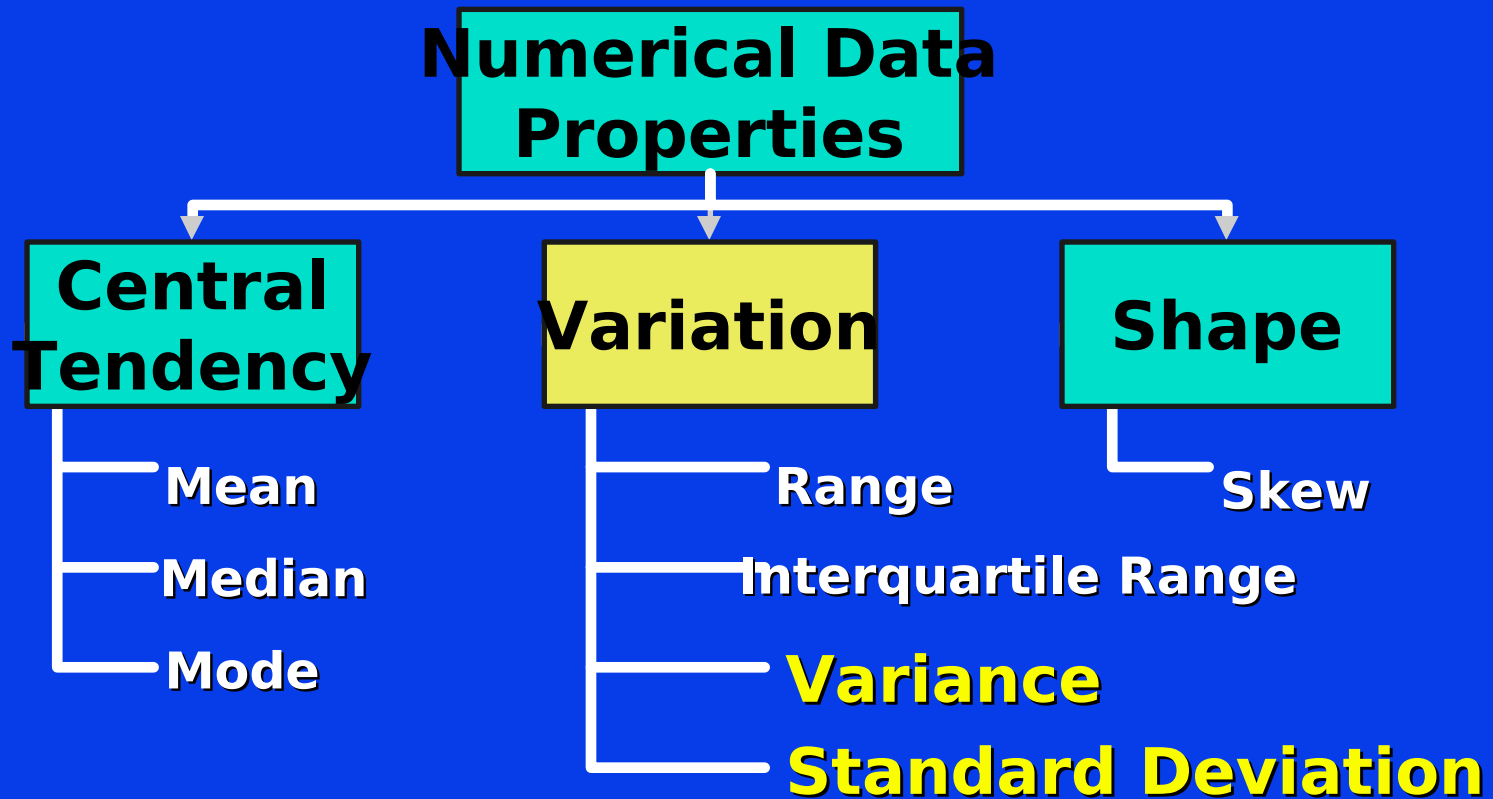
$$\text{Range} = X_{\text{largest}} - X_{\text{smallest}}$$

3. Ignores How Data Are Distributed



Numerical Data Properties & Measures

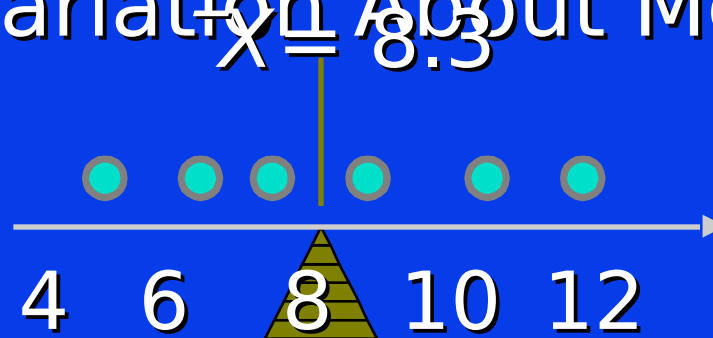
© 2003 Pearson Prentice
Hall



Variance & Standard Deviation

© 2003 Pearson Prentice Hall

1. Measures of Dispersion
2. Most Common Measures
3. Consider How Data Are Distributed
4. Show Variation About Mean ($\bar{X}$ or μ)



Sample Variance Formula

© 2003 Pearson Prentice Hall

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

$n - 1$ in denominator! (Use **N** if **Population Variance**)

$$= \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1}$$



Sample Standard Deviation Formula

© 2003 Pearson Prentice Hall

$$S = \sqrt{S^2}$$

$$= \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

$$= \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}}$$

Variance Example

© 2003 Pearson Prentice Hall

Raw Data: 10.3 4.9 8.9 11.7 6.3

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} \quad \text{where } \bar{x} = \frac{\sum_{i=1}^n x_i}{n} = 8.3$$

$$s^2 = \frac{(10.3 - 8.3)^2 + (4.9 - 8.3)^2 + \dots + (11.7 - 8.3)^2}{6 - 1}$$

$$= 6.368$$

Thinking Challenge

© 2003 Pearson Prentice
Hall

You're a financial analyst for Prudential-Bache Securities. You have collected the following closing stock prices of new stock issues: **17, 16, 21, 18, 13, 16, 12, 11.**

What are the **variance** and **standard deviation** of the stock prices?



Variation Solution*

© 2003 Pearson Prentice
Hall

Sample Variance

Raw Data: 17 16 21 18 13 16 12

$$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} \quad \text{where } \bar{x} = \frac{\sum_{i=1}^n x_i}{n} = 15.5$$

$$S^2 = \frac{(17 - 15.5)^2 + (16 - 15.5)^2 + \dots + (12 - 15.5)^2}{8 - 1}$$

$$= 11.14$$

Variation Solution*

© 2003 Pearson Prentice
Hall

Sample Standard Deviation

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}} = \sqrt{1114} = 3.34$$

Summary of Variation Measures

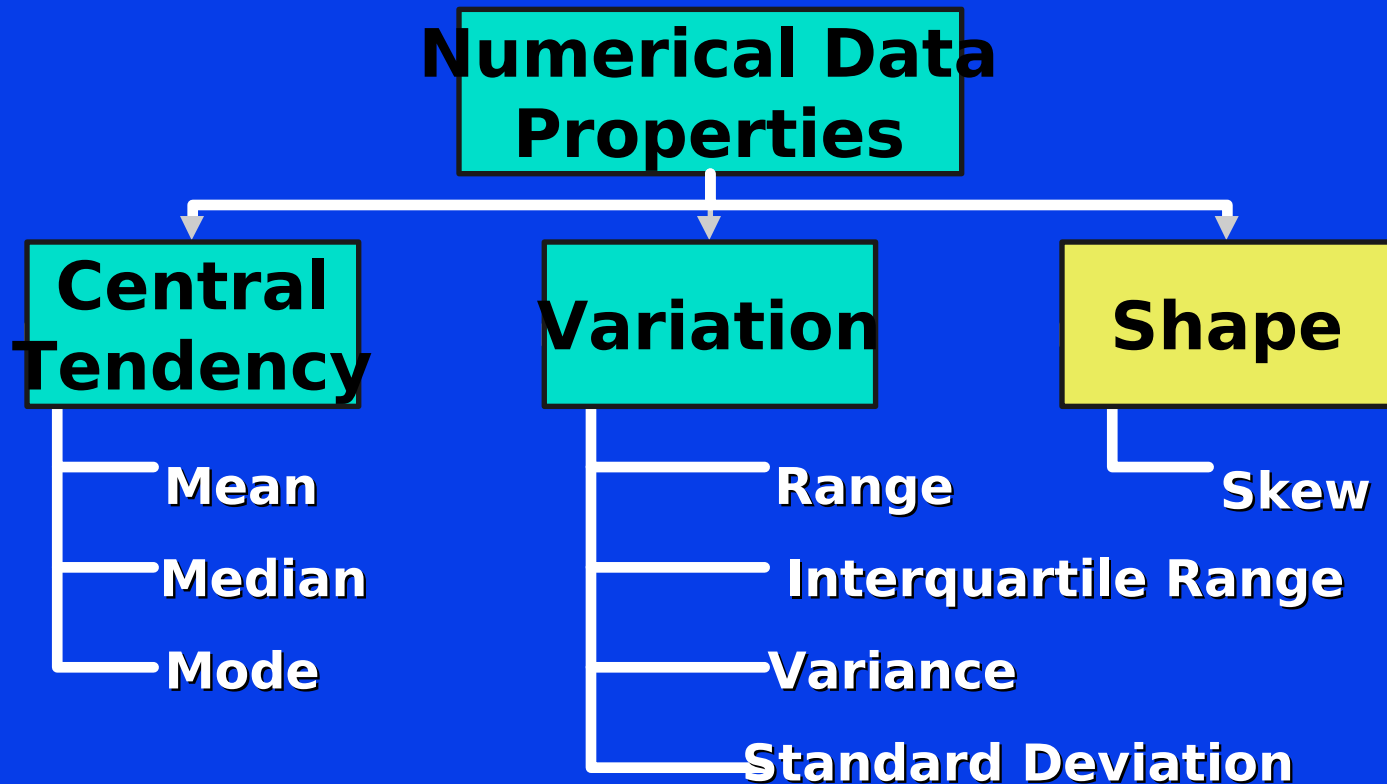
© 2003 Pearson Prentice Hall

Measure	Equation	Description
Range	$X_{\text{largest}} - X_{\text{smallest}}$	Total Spread
Interquartile Range	$Q_3 - Q_1$	Spread of Middle 50%
Standard Deviation (Sample)	$\sqrt{\frac{\sum (X_i - \bar{X})^2}{n - 1}}$	Dispersion about Sample Mean
Standard Deviation (Population)	$\sqrt{\frac{\sum (X_i - \mu_X)^2}{N}}$	Dispersion about Population Mean
Variance (Sample)	$\frac{\sum (X_i - \bar{X})^2}{n - 1}$	Squared Dispersion about Sample Mean

Shape

Numerical Data Properties & Measures

© 2003 Pearson Prentice
Hall



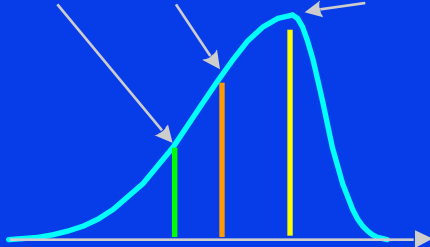
Shape

© 2003 Pearson Prentice
Hall

1. Describes How Data Are Distributed
2. Measures of Shape
 - Skew = Symmetry

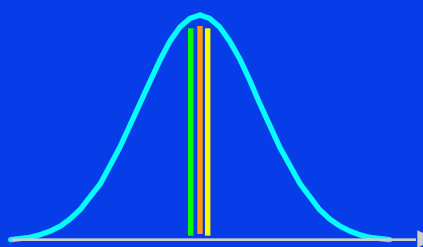
Left-Skewed

Mean Median Mode



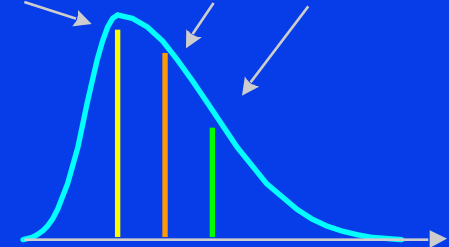
Symmetric

Mean=Median=Mode



Right-Skewed

Mode Median Mean

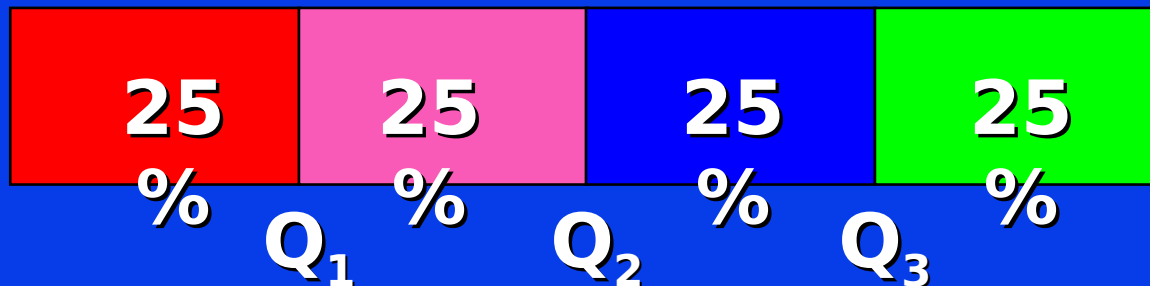


Quartiles & Box Plots

Quartiles

© 2003 Pearson Prentice
Hall

1. Measure of *Noncentral* Tendency
2. Split Ordered Data into 4 Quarters



3. Position of i -th Quartile
Positioning Point of $Q_i = \frac{i \cdot a + 1f}{4}$

Quartile (Q_1) Example

© 2003 Pearson Prentice
Hall

Raw Data: 10.3 4.9 8.9 11.7 6.3
7.7

Ordered: 4.9 **6.3** 7.7 8.9 10.3
11.7



Position:
 $Q_1 \text{ Position} = \frac{1 \cdot a + 1 \cdot f}{4} = \frac{1 \cdot 4 + 1 \cdot 1}{4} = \frac{5}{4} = 1.25 \approx 2$

$$Q_1 = 6.3$$

Quartile (Q_2) Example

© 2003 Pearson Prentice
Hall

Raw Data: 10.3 4.9 8.9 11.7 6.3
7.7

Ordered: 4.9 6.3 **7.7** **8.9** 10.3
11.7

Position: $Q_2 \text{ Position} = \frac{2 \cdot a + 1}{4} = \frac{2 \cdot 7 + 1}{4} = 3.5$

$$Q_2 = \frac{7.7 + 8.9}{2} = 8.3$$

Quartile (Q_3) Example

© 2003 Pearson Prentice
Hall

Raw Data: 10.3 4.9 8.9 11.7 6.3
7.7

Ordered: 4.9 6.3 7.7 8.9 **10.3**
11.7

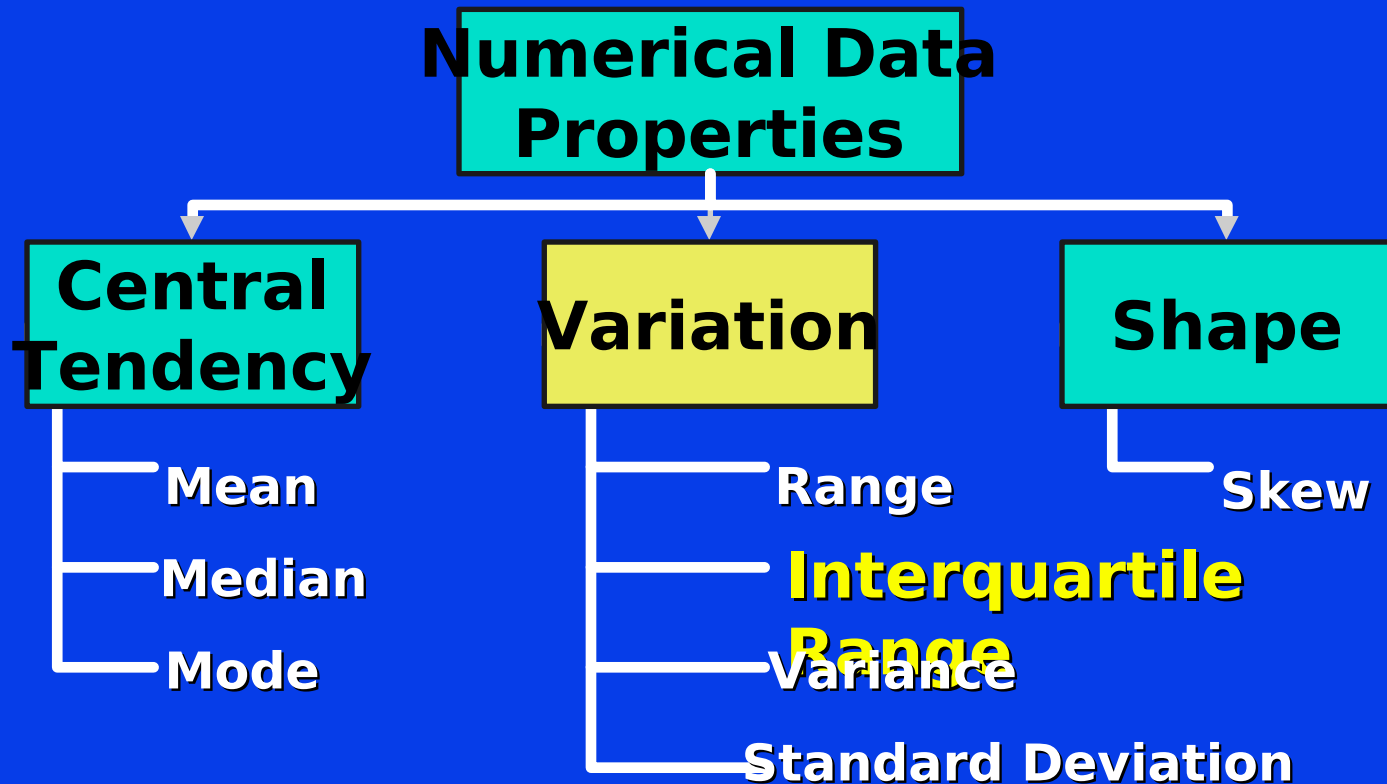
Position: $Q_3 \text{ Position} = \frac{3 \cdot a + 1}{4} = \frac{3 \cdot 4 + 1}{4} = 5.25 \approx 5$

(Note: In the original image, a pink diamond points to the value 8.9 in the ordered data list, which is the 5th position.)

$$Q_3 = 10.3$$

Numerical Data Properties & Measures

© 2003 Pearson Prentice
Hall



Interquartile Range

© 2003 Pearson Prentice
Hall

1. Measure of Dispersion
2. Also Called Midspread
3. Difference Between Third & First Quartiles

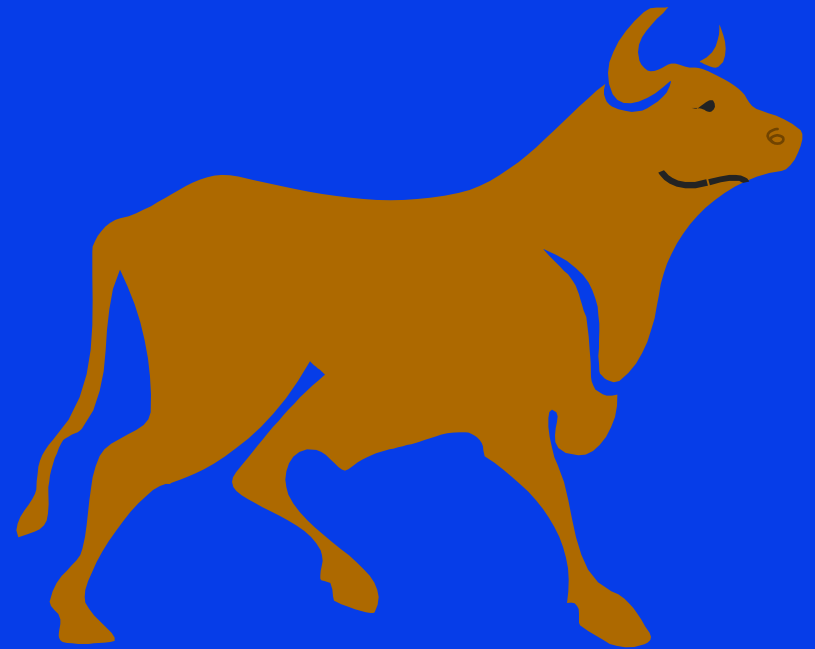
$$\text{Interquartile Range} = Q_3 - Q_1$$

4. Spread in Middle 50%
5. Not Affected by Extreme Values

Thinking Challenge

© 2003 Pearson Prentice
Hall

You're a financial analyst for Prudential-Bache Securities. You have collected the following closing stock prices of new stock issues: **17, 16, 21, 18, 13, 16, 12, 11.** What are the **quartiles, Q_1 and Q_3 ,** and the **interquartile range?**



Quartile Solution*

© 2003 Pearson Prentice
Hall

Q_1

Raw Data: 17 16 21 18 13 16 12
11

Ordered: 11 12 13 16 16 17 18
21

$$Q_1 \text{ Position} = \frac{1 \cdot a + 1 \cdot f}{2} = \frac{1 \cdot 12 + 1 \cdot 13}{2} = 12.5$$

$$Q_1 = 12.5$$

Quartile Solution*

© 2003 Pearson Prentice
Hall

Q_3

Raw Data: 17 16 21 18 13 16 12
11

Ordered: 11 12 13 16 16 17 **18**
21

$$Q_3 \text{ Position} = \frac{3 \cdot a + 1f}{4} = \frac{3 \cdot 16 + 1 \cdot 18}{4} = \frac{66}{4} = 6.75 \approx 7$$

$$Q_3 = 18$$

2 - 75

Interquartile Range Solution*

© 2003 Pearson Prentice Hall

Interquartile Range

Raw Data: 17 16 21 18 13 16 12
11

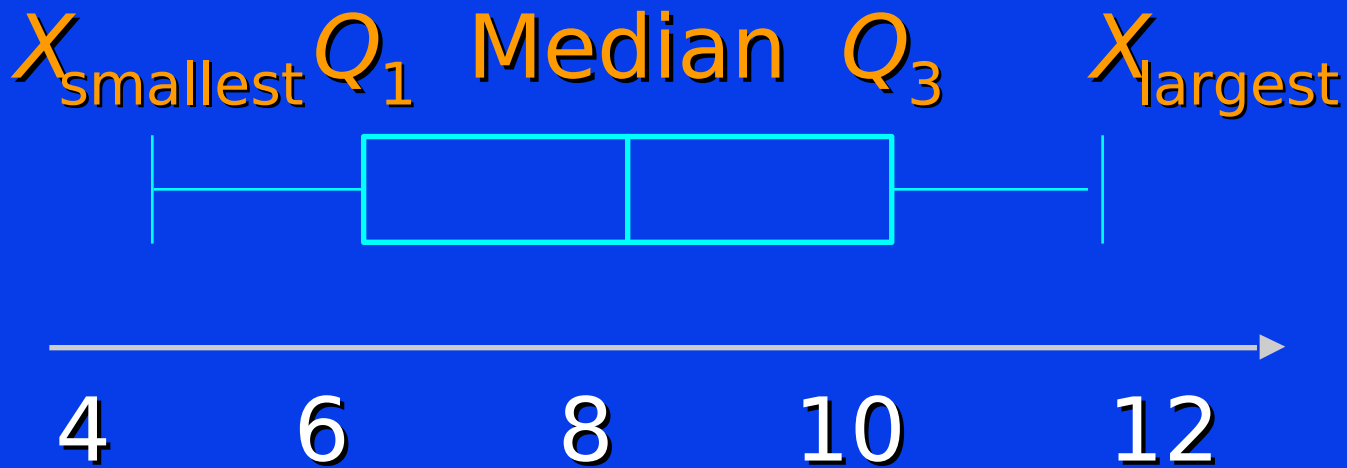
Ordered: 11 12 13 16 16 17 18
21

Position
Interquartile Range = $Q_3 - Q_1 = 18.5 - 12.5 = 6$

Box Plot

© 2003 Pearson Prentice
Hall

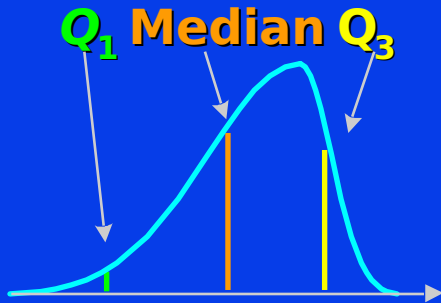
1. Graphical Display of Data Using 5-Number Summary



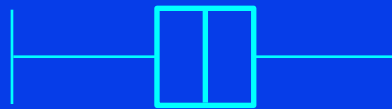
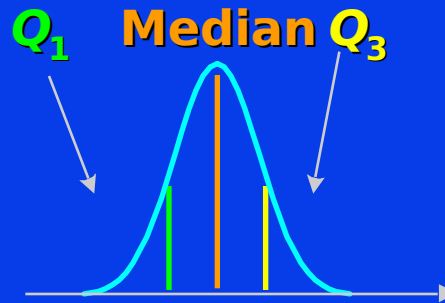
Shape & Box Plot

© 2003 Pearson Prentice
Hall

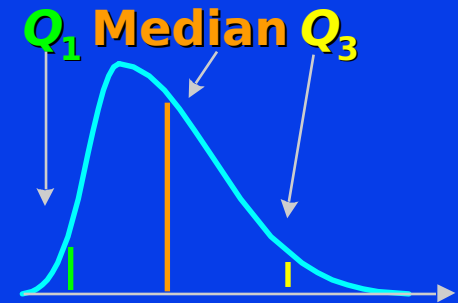
Left-Skewed



Symmetric



Right-Skewed



Distorting the Truth with Descriptive Techniques

Errors in Presenting Data

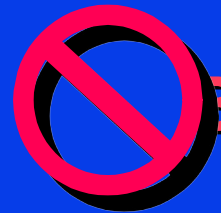
© 2003 Pearson Prentice
Hall

1. Using 'Chart Junk'
2. No Relative Basis in Comparing Data Batches
3. Compressing the Vertical Axis
4. No Zero Point on the Vertical Axis



'Chart Junk'

© 2003 Pearson Prentice
Hall



Bad Presentation

Minimum Wage



1960:



\$1.00



1970:
\$1.60

1980:

\$3.10

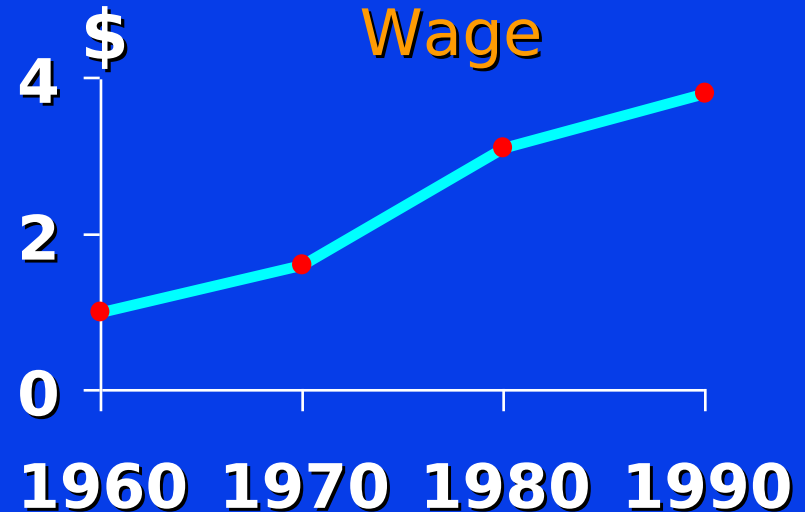


1990:

\$3.80

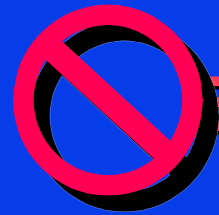
Good Presentation

Minimum Wage



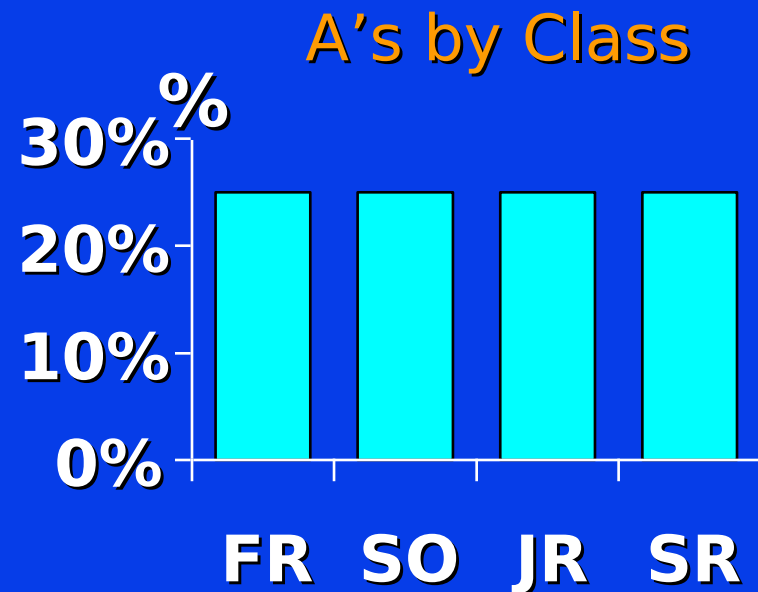
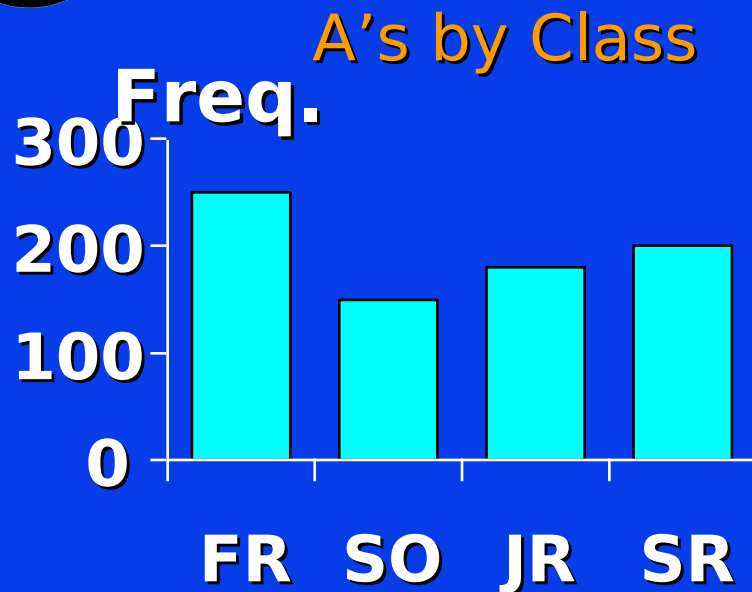
No Relative Basis

© 2003 Pearson Prentice
Hall



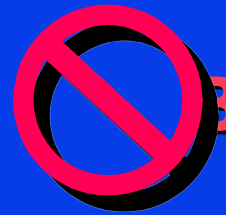
Bad Presentation

Good Presentation

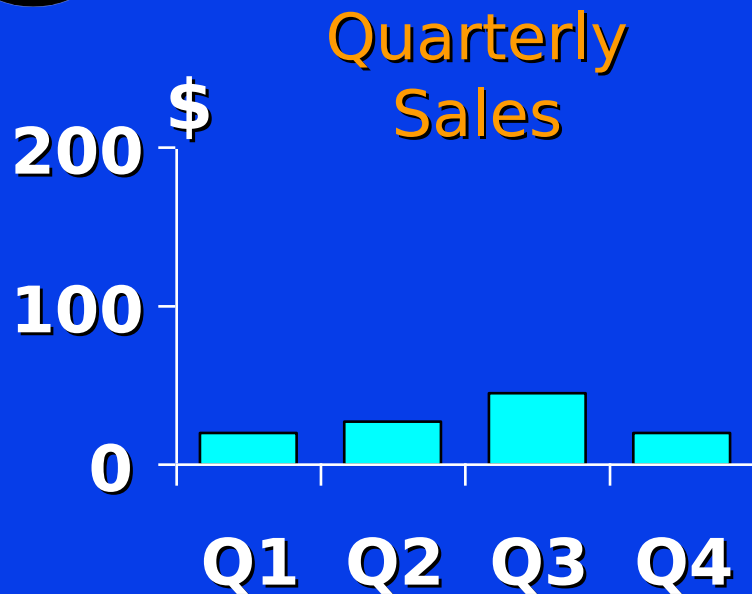


Compressing Vertical Axis

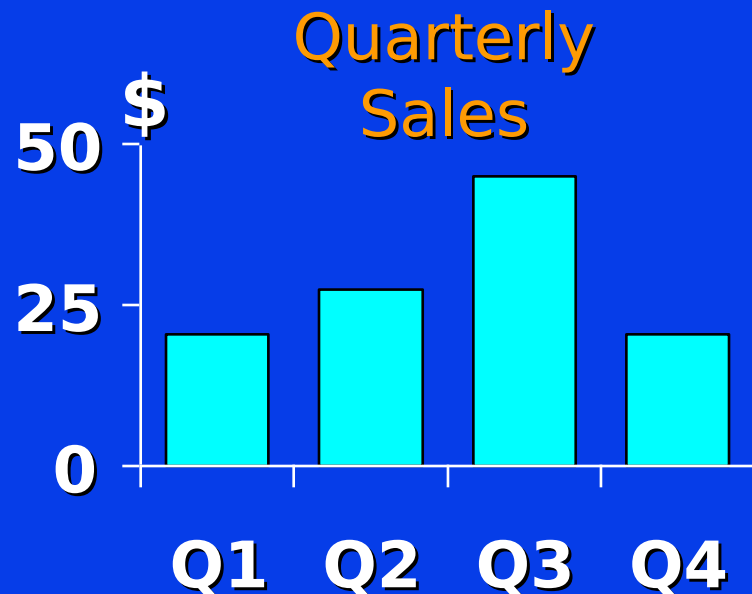
© 2003 Pearson Prentice Hall



Bad Presentation

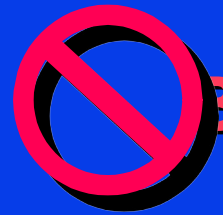


Good Presentation

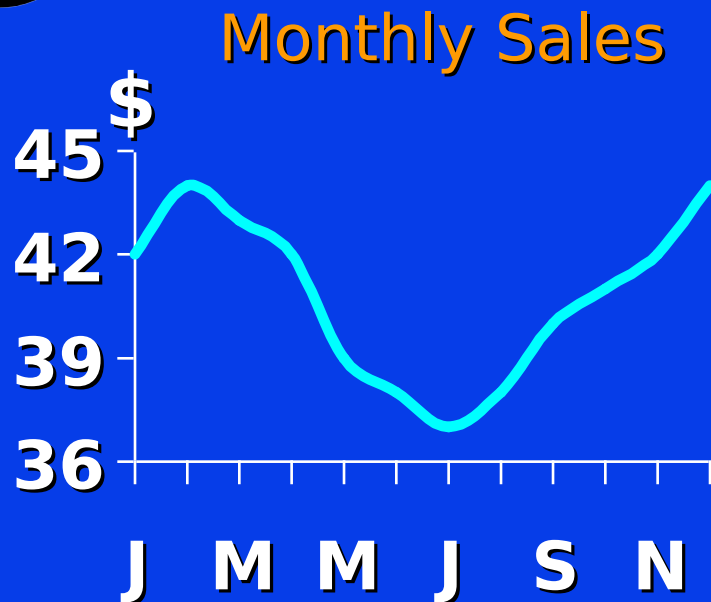


No Zero Point on Vertical Axis

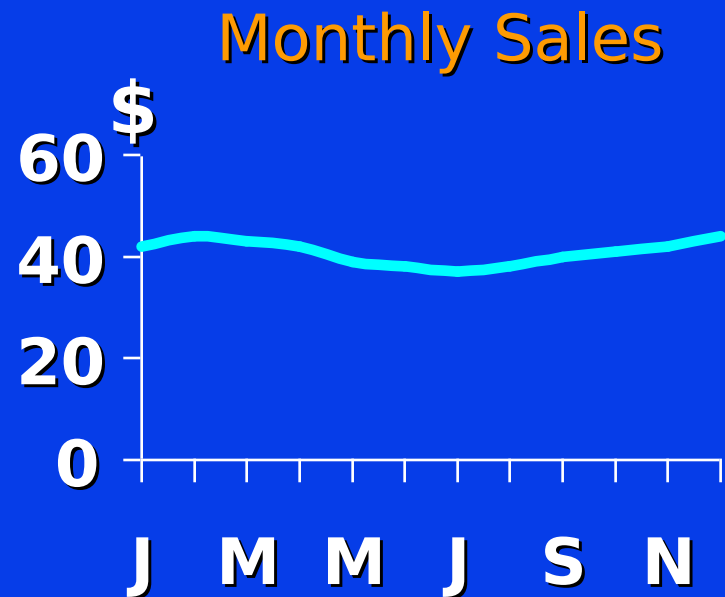
© 2003 Pearson Prentice
Hall



Bad Presentation



Good Presentation



Conclusion

© 2003 Pearson Prentice
Hall

1. Described Qualitative Data Graphically
2. Described Numerical Data Graphically
3. Created & Interpreted Graphical Displays
4. Explained Numerical Data Properties
5. Described Summary Measures
6. Analyzed Numerical Data Using Summary Measures

